

Xiangqi Kong

Explainable AI for Time-Series Forecasting of Airport Passenger Flow

Faculty of Engineering and Applied Sciences

PhD
Academic Year: 2022–2025

Supervisor: Dr Yang Xing
Associate Supervisor: Prof Antonios Tsourdos
January 2026



Faculty of Engineering and Applied Sciences

PhD

Academic Year: 2022–2025

Xiangqi Kong

Explainable AI for Time-Series Forecasting of Airport Passenger Flow

Supervisor: Dr Yang Xing
Associate Supervisor: Prof Antonios Tsourdos
January 2026

© Cranfield University 2025. All rights reserved. No part of this publication may be reproduced without the written permission of the copyright owner.

Academic Integrity Declaration

I declare that:

- the thesis submitted has been written by me alone.
- the thesis submitted has not been previously submitted to this university or any other.
- that all content, including primary and/or secondary data, is true to the best of my knowledge.
- that all quotations and references have been duly acknowledged according to the requirements of academic research.

I understand that to knowingly submit work in violation of the above statement will be considered by examiners as academic misconduct.

Abstract

Accurate forecasting of airport passenger flows is critical for efficient airport operations, supporting decisions such as gate and stand allocation, staffing, disruption management, and long-term infrastructure planning. Current forecasting practice faces two persistent limitations. First, long-horizon passenger flow is affected by non-stationarity, horizon variability, and sudden exogenous shocks (e.g., policy changes and major events), which can cause rapid performance degradation and brittle behaviour outside the training regime. Second, although modern neural forecasters can be highly accurate in certain settings, they are typically opaque, and their post-hoc explanations are often scenario dependent and rarely presented in a form that supports auditing, comparison across time, or operational reuse. As a result, decision-makers may receive forecasts without defensible answers to what drove a change, which inputs mattered at which times, and whether an explanation is trustworthy.

This thesis addresses these limitations by combining data construction, explanation evaluation, and model design. The thesis is organised on a paper basis. Chapter 1 introduces the research questions and overall methodology. Chapter 2 performs a controlled evaluation of post-hoc explainability methods using synthetic data with attribution ground truth and real-world airport passenger flow datasets, and demonstrates that explanation quality can vary strongly by scenario and is only weakly coupled to forecast accuracy. Chapter 3 addresses practical shortcomings of saliency outputs by introducing sparse, normalised, and time-aligned post-processing, producing explanations that are easier to interpret and can be reused downstream (e.g., for data augmentation). Chapter 4 tackles opacity by developing an intrinsically interpretable forecasting model based on a structured neural decomposition of trend, seasonality, and residual components, allowing transparent diagnostics of model behaviour. Chapter 5 targets long-horizon robustness under horizon variability and shocks through variable-horizon patch-autoregressive training, semantic anchoring, and scenario-

based sensitivity analysis. Chapter 6 synthesises findings and outlines future directions. The Appendix A and Appendix B document dataset construction (Freedom of Information Act (FOIA) throughput and custom Global Database of Events, Language, and Tone (GDELT) indices) and review explainable interfaces for human–autonomy teaming.

The main contributions are fivefold. First, a high-resolution airport passenger-flow dataset is constructed by integrating FOIA throughput records with tailored GDELT-based external covariates. Second, a systematic evaluation framework for post-hoc explainability is developed, demonstrating strong scenario dependence and a weak coupling between forecast accuracy and post hoc explanation quality. Third, a principled post-processing approach is introduced to produce sparse, normalised, and time-aligned explanations, improving interpretability and enabling downstream reuse. Fourth, an intrinsically interpretable forecasting model is proposed via structured neural decomposition, allowing transparent attribution of trend, seasonality, and residual dynamics. Fifth, robustness for variable-horizon forecasting is improved through patch-autoregressive with semantic anchoring and scenario-based sensitivity analysis.

Keywords:
airport passenger flow; time-series forecasting; interpretability; post-hoc attribution; saliency maps; feature importance; neural decomposition; trend-seasonality modelling; long-horizon prediction; horizon variability; exogenous shocks; robustness diagnostics; sensitivity analysis.

Acknowledgements

This thesis was conducted as part of an EPSRC-SAAB full funded project, and I am grateful for the financial support that made this work possible.

First, I would like to thank my primary supervisor, Dr Yang Xing, for his guidance and support throughout my PhD journey. He guided me into the road of research, helping me progress from initial uncertainty to a better understanding of the field. He has always been available when needed, and his expertise has been essential to this work. He consistently guided me to view problems from a positive perspective and encouraged me to think more deeply, and this influence extended beyond my work. Some teachers have a lasting influence on their students' lives, and he is that kind of teacher. I would also thank my second supervisor, Prof Antonios Tsourdos, for his valuable feedback and support that helped shape this research.

Further more, I acknowledge the support from Braithwaite Jarryd and Wikander Andreas, who provided the flexibility necessary for this research and offered valuable insights during our collaborative meetings. Adolfo Perrusquia and Ip-Shing Fan provided many constructive comments during the review process. Thanks to George Yazigi for providing invaluable assistance with high-performance computing resources, and his prompt responses to technical issues helped reduce experimental time significantly. Thanks to Adrian Cole for reaching out when I was too shy to initiate conversations, which gave me the courage to connect with others. I am also grateful to Ziyue Wang and Xinyue Cui for their support and friendship, especially when I first arrived and everything was new and challenging. I really value the time we spent together.

Finally, I sincerely thank my parents for their unconditional love, support, and trust. All they want is that I be healthy and happy, and their understanding and encouragement have been constant. I am also deeply grateful to my partner for his love, companionship, and support throughout this journey. We have spent every day together, which made this

journey feel much less lonely. His understanding, patience, and encouragement meant a lot to me.

Contents

Academic Integrity Declaration i

Abstract iii

Acknowledgements v

Contents vii

List of Figures xiii

List of Tables xvii

List of Abbreviations xix

1 Introduction and Research Framework 1
- 1.1 Airport Passenger Flow Forecasting as a High-Stakes Decision Problem . 2
- 1.2 Challenges of Explainability in Time-Series 3
- 1.3 Research Gaps and Motivation 4
- 1.4 Research Questions 5
- 1.5 Thesis Contributions 6
- 1.6 Thesis Structure 7
- 1.7 Publications and Manuscripts in Preparation 9

References 11

2 Post-hoc Explainability Evaluation 15

Abstract 17
- 2.1 Introduction 18
- 2.2 Related Work 20
  - 2.2.1 Time Series Forecasting in Transportation Research 20
  - 2.2.2 Post-hoc Explainability Methods for Time Series Models 21
  - 2.2.3 Evaluation of Explainability Methods 23
  - 2.2.4 Positioning of This Work 24
- 2.3 Problem Setting and Evaluation Framework 24
  - 2.3.1 Multivariate Time Series Forecasting for Airport Passenger Flow . 24
  - 2.3.2 Explainability Objectives in Time Series Forecasting 25
  - 2.3.3 Dual Evaluation Strategy 25
- 2.4 Experimental Design 27
  - 2.4.1 Forecasting Models 28
  - 2.4.2 Explainability Methods 29

2.4.3 Datasets for Evaluation . . . 31
2.4.3.1 Synthetic Datasets . . . 31
2.4.3.2 Real-World Datasets . . . 35
2.4.4 Evaluation Metrics and Baseline Strategy . . . 36
2.4.4.1 Oracle-based Metrics (Synthetic Datasets) . . . 37
2.4.4.2 Behavioural Metrics (Real-World Datasets) . . . 38
2.4.4.3 Baseline Definitions and Evaluation Setup . . . 39
2.5 Results on Synthetic Data . . . 40
2.5.1 Overall Performance Across Temporal Patterns . . . 40
2.5.2 Model and Explainer Interactions . . . 43
2.5.3 Relationship Between Prediction Accuracy and Explainability Quality . . . 45
2.5.4 Robustness Analysis Across Random Seeds . . . 47
2.5.5 Batch Experiment Failure Analysis . . . 49
2.5.6 Summary of Synthetic Data Findings . . . 50
2.6 Results on Real World Airport Data . . . 52
2.6.1 Deletion-based Evaluation in Airport Passenger Flow . . . 53
2.6.2 Model Randomisation and Attribution Stability Diagnostics . . . 54
2.6.3 Relationship between Prediction Accuracy and Explainability Quality on Real-world Data . . . 56
2.7 Discussion and Practical Implications . . . 58
2.7.1 Implications of Synthetic and Real-world Results . . . 58
2.7.2 Practical Insights on Explainability . . . 58
2.7.3 On SHAP Instability and Numerical Reliability . . . 59
2.8 Conclusion . . . 60

References **61**

3 Sparse Saliency Methods **67**

Abstract **69**
3.1 INTRODUCTION . . . 69
3.1.1 Motivation . . . 69
3.1.2 Contribution . . . 71
3.2 Related Works . . . 71
3.2.1 Time Series Modelling . . . 71
3.2.2 Saliency Methods for Time Series . . . 73
3.3 Methodology . . . 74
3.3.1 Overall Framework . . . 74
3.3.2 Sparse Normalised Saliency . . . 75
3.4 Experiments and Analysis . . . 77
3.4.1 Experiment Setup . . . 77
3.4.1.1 Data Preparation . . . 78
3.4.1.2 Sparse Saliency Integration . . . 78
3.4.1.3 Measurement metrics . . . 79
3.4.2 Results Analysis . . . 79
3.5 Conclusions and Future Work . . . 83

References **85**

4 Interpretable Decomposition Forecasting **89**

Abstract **91**
4.1 Introduction . . . 92
4.2 Related Work . . . 95
4.2.1 Classical Decomposition and the Need for Interpretability . . . 95
4.2.2 Decomposition-Based Neural Forecasting . . . 96
4.2.3 Frequency-Aware and Harmonic modelling . . . 97
4.3 Methodology . . . 99
4.3.1 Problem Formulation . . . 101
4.3.2 Input Normalization . . . 101
4.3.3 Trend Module: Compact Smooth Prior . . . 101
4.3.4 Seasonal Module: Channel-Aware Encoder and Harmonic Decoder 102
4.3.4.1 Channel-Aware Encoder. . . . 103
4.3.4.2 Harmonic Decoder. . . . 105
4.3.5 Residual Module . . . 105
4.3.6 Fusion Module: Adaptive Channel-Wise Gate . . . 106
4.4 Experimental Setup . . . 106
4.4.1 Real-World Forecasting Benchmarks . . . 106
4.4.2 Synthetic Datasets for Decomposition Evaluation . . . 107
4.4.3 Training and Evaluation Setup . . . 109
4.4.3.1 Learning Objectives. . . . 109
4.4.3.2 Evaluation Metrics. . . . 109
4.4.3.3 Baselines. . . . 110
4.4.3.4 Implementation Details. . . . 110
4.5 Experiment Results and Discussion . . . 110
4.5.1 Forecasting Accuracy on Real-World Datasets . . . 110
4.5.2 Impact of Input Lookback Length . . . 111
4.5.3 Ablation Study . . . 113
4.5.4 Forecasting Performance on Synthetic Datasets . . . 113
4.5.4.1 Overall Accuracy. . . . 115
4.5.4.2 Computational Efficiency. . . . 115
4.5.4.3 Case Study on Complex Dataset. . . . 116
4.5.5 Interpretability Evaluation . . . 117
4.5.5.1 Decomposition on Real-World Data . . . 117
4.5.5.2 Ground-Truth Evaluation on Synthetic Data . . . 118
4.5.6 Decomposition–Forecast Quality Relationship Exploration . . . 120
4.5.7 Airport Case Study (Top 20 Busiest U.S. Airports) . . . 122
4.6 Limitations and Future Work . . . 124
4.7 Conclusion . . . 125

References **127**

5 Variable-Horizon Forecasting **137**

Abstract **139**
5.1 Introduction . . . 139
5.2 Related Work . . . 142
5.2.1 Long-Horizon Time-Series Forecasting . . . 142
5.2.2 Autoregressive Latent Dynamics . . . 143
5.2.3 Language Models in Time-Series Modelling . . . 143
5.3 Problem Setup and Notation . . . 144
5.3.1 Forecasting Task Definition . . . 144
5.3.2 Horizon Sets . . . 145
5.3.3 Evaluation Metrics . . . 145
5.4 Methodology . . . 146
5.4.1 Patchification and Tokenisation . . . 147
5.4.2 Patch-Autoregressive Latent Dynamics . . . 148
5.4.3 Variable-Horizon Training Protocol . . . 149
5.4.4 Semantic Anchoring via Frozen Language Models . . . 151
5.4.5 Concept-Based Explanation Mechanism . . . 153
5.4.6 Latent Conditioning on Exogenous Information . . . 154
5.5 Datasets and Experimental Setup . . . 155
5.5.1 Dataset Statistics . . . 156
5.5.2 Implementation Details . . . 158
5.6 Results . . . 159
5.6.1 Multi-Horizon Forecasting Performance . . . 159
5.6.2 Patch-AR vs Single-Step AR Ablation . . . 160
5.6.3 Concept Ablation Study . . . 161
5.6.4 Concept-Based Explanations . . . 163
5.6.5 Diagnostic Evaluation of Explanation Quality . . . 165
5.6.6 Robustness under Exogenous Shocks via Scenario-Based Sensitivity Analysis . . . 168
5.6.6.1 Accuracy with and without exogenous shocks . . . 170
5.6.6.2 Computational Complexity Analysis . . . 171
5.6.6.3 Directional Response under Scenario Perturbations . . . 173
5.6.7 Additional Accuracy-Related Results . . . 175
5.6.7.1 Single-Horizon Sanity Runs . . . 175
5.7 Analysis and Discussion . . . 176
5.7.1 Why Variable-Horizon Training Works . . . 176
5.7.2 Role of Semantic Anchoring . . . 176
5.7.3 Robustness under Exogenous Shocks . . . 177
5.8 Conclusion . . . 177

References **179**

6 Conclusions and Future Work **183**
6.1 Comparison of Explainability Paradigms . . . 183
6.2 Key Lessons for Explainable Time-Series Forecasting . . . 185
6.2.1 Implications for Real-World Deployment . . . 186
6.3 Limitations . . . 187
6.4 Future Research Directions . . . 187

References **191**

A Airport Passenger Dataset Construction and Analysis **193**

Abstract **195**
A.1 Introduction . . . 195
A.2 Related Work . . . 198
A.2.1 Aviation Demand Shocks and Recovery Dynamics . . . 198
A.2.2 Information Effects and Anticipatory Behaviour in Transportation 199
A.2.3 International Exposure, Airport Structure, and Vulnerability . . . 200
A.3 Data and Empirical Strategy . . . 201
A.3.1 Dataset Construction . . . 202
A.3.1.1 Airport Passenger Throughput . . . 202
A.3.1.2 Information Shock Measures from GDELT . . . 203
A.3.2 Empirical Design and Identification Strategy . . . 204
A.3.2.1 Timing and Sensitivity to Information Shocks . . . 205
A.3.2.2 Exposure-Interaction Model and Recovery Analysis . . 205
A.4 Results . . . 205
A.4.1 Timing of Information Shocks and Demand Responses . . . 206
A.4.2 Differential Sensitivity to Information by International Exposure . 207
A.4.3 Recovery Dynamics: Persistent but Not Permanent Heterogeneity 209
A.4.4 Robustness Checks . . . 211
A.5 Discussion . . . 212
A.5.1 Information and Anticipatory Demand Behaviour . . . 214
A.5.2 International Exposure and Recovery Heterogeneity . . . 214
A.5.3 The Role of Control-Related Information . . . 215
A.5.4 Airport Size and Structural Trade-offs . . . 215
A.5.5 Policy Implications . . . 215
A.5.6 Limitations and Future Research . . . 216
A.6 Conclusion . . . 216

References **221**

B Explainable Interface Survey **225**

Abstract **227**
B.1 Introduction . . . 227
B.1.1 Background . . . 227
B.1.2 Contribution . . . 229
B.1.3 Project Organisation . . . 230
B.2 High-Level Architectures of Explainable Interface in HAT . . . 232
B.2.1 Clarification of Explainable Interface . . . 232
B.2.2 The Need of Explainable Interface . . . 234
B.2.2.1 Regulation and Guidelines . . . 234
B.2.2.2 Expectation of Different Stakeholders . . . 235
B.2.3 Architecture for Explainable Interface in HAT . . . 236
B.3 Frameworks and Techniques for Model Explainability . . . 238

B.3.1 Pre-modelling Explainability . . . 239
B.3.2 Inherent Interpretability of Model by Design . . . 240
B.3.2.1 Using Transparent Model Structures . . . 240
B.3.2.2 Imposing Interpretability Constraints . . . 241
B.3.2.3 Combining different interpretable modules . . . 242
B.3.3 Post-modelling Explainability . . . 244
B.3.3.1 Feature-based Explanation . . . 244
B.3.3.2 Example-based Explanation . . . 245
B.3.3.3 Concept-based Explanation . . . 245
B.3.4 Causal Inference and Counterfactual Explanation . . . 246
B.3.5 Explainable Reinforcement Learning . . . 248
B.3.6 LLMs for Explanation Generation . . . 249
B.4 Human-Centered Explainable Interface in HAT . . . 250
B.4.1 Human Behaviours modelling in HAT . . . 251
B.4.2 Explainable Interface Design Principles . . . 254
B.4.2.1 Considerations in Explainable Interface Design . . . 254
B.4.2.2 Extended User Experience Framework for Explainable Interface . . . 258
B.5 Evaluation of Explainable Interface in HAT . . . 259
B.5.1 Model Performance Centred Evaluation Methods . . . 261
B.5.1.1 Inherent Interpretability Level Evaluation for Functional Model . . . 261
B.5.1.2 Quantitative Evaluation for Post-hoc Explainers . . . 261
B.5.2 Human-Centered Evaluation Methods . . . 264
B.5.3 Evaluation Methods towards Teaming Goals . . . 266
B.6 Challenges and Future Directions . . . 267
B.6.1 Challenges . . . 267
B.6.2 Directions . . . 269
B.7 Conclusion . . . 270
B.8 Explainable Interface Application Scenarios in HAT: Case Studies . . . 270
B.8.1 Healthcare . . . 271
B.8.1.1 Clinical Decision Support Systems . . . 271
B.8.1.2 Surgical Robotics . . . 273
B.8.1.3 Challenges fo Explainable Interface in Healthcare . . . 275
B.8.2 Transportation . . . 277
B.8.2.1 Autonomous Vehicles . . . 277
B.8.2.2 Air Transportation Systems . . . 279
B.8.2.3 Challenges for Explainable Interface in Transportation . 281
B.8.3 Explainable Interface and Digital Twin . . . 281
B.8.3.1 Relationship of Physical System, Digital Twin, and XAI 281
B.8.3.2 XAI Methods for Digital Twin . . . 282
B.8.3.3 Digital Twin as an Explainable Interface . . . 282
B.8.4 Challenges for XAI and Digital Twin Integration . . . 282

References **285**

C Ethical Approval **331**

List of Figures

1.1 Thesis structure and chapter connections. 8

2.1 Dual evaluation framework for explainability. Synthetic datasets enable oracle-based evaluation against predefined attribution targets, while real-world datasets are assessed using behavioural metrics that probe model sensitivity and attribution robustness. 27
2.2 Oracle-score heatmap on synthetic datasets using the regenerated analysis output (all seeds), following the unified normalization protocol defined in Section 2.5. 41
2.3 Model–explainer interaction heatmap on synthetic datasets, showing composite explainability scores across forecasting models and dataset types under the unified normalization protocol defined in Section 2.5. 45
2.4 Aggregate relationship between prediction error (mean test MSE) and dataset-specific oracle metrics for seed 42. Each point is one model–explainer combination; each panel (trend, frequency-single, var-coupling) has $n = 26$ points. 46
2.5 Cross-seed variability of synthetic oracle metrics (seeds 42–44). Panels correspond to the five synthetic dataset types and use each dataset's primary oracle metric. Within each panel, models are on the x-axis, explainers are colour-coded, and each box summarises three seed values for one model–explainer pair (with jittered seed points). 49
2.6 Overview of behavioural explainability metrics on real-world airport data, aggregated across the four representative datasets. Rows correspond to forecasting models and columns to explainers. Left: mean deletion AUC (more negative = larger performance degradation under deletion). Middle: mean model-randomisation similarity (cosine similarity between original and randomised-model attributions; values near zero indicate stronger sensitivity to learnt parameters). Right: median $\log_{10}(1 + \text{noise stability})$ under input perturbations. N/A indicates an inapplicable or undefined case: TCN–DeepLift is omitted for the same Captum limitation as in the synthetic benchmark; XGBoost randomisation is not defined in this pipeline; and some neural-model LIME randomisation entries are undefined due to degenerate randomised-model attribution vectors (NaN cosine similarity). 56

3.1 Illustration of the proposed framework integrating saliency tensors into the forecasting model lifecycle. The diagram highlights the process of normalising and sparsifying saliency tensors to enhance predictive accuracy and interpretability. 76
3.2 Comparison of GRU, XGBoost, and XGBoost+ predictions against ground truth for boarding areas A–G and Other (targets 1–8). 81

3.3 Readability comparison of saliency visualisations (heatmaps of the same tensors): (A) aggregated original saliency tensor $\mathbf{A}$; (B) aggregated sparse saliency tensor $\mathbf{A}'$; (C) monthly passenger flow for San Francisco Airport boarding areas. 83

4.1 Overall architecture of iDecPred. The model normalises inputs with Reversible Instance Normalisation (ReVIN), decomposes them into trend (spline + affine), seasonal (channel-aware encoder with harmonic decoder), and residual (Multi-Layer Perceptron (MLP)) components, and fuses their forecasts via a channel-wise attention gate to produce final predictions with interpretable components. 100

4.2 Comparison of paradigms for multivariate time series forecasting (Channel-Independent (CI), Channel-Mixing (CM), Graph-Based (GB), Channel-as-Token (CaT)) and the proposed Low-rank Channel Interaction (LCI). Notation: B batch size; L input (look-back) length; H ($\equiv H_{\text{out}}$) forecasting horizon; C number of channels; d embedding dimension. For CM, W denotes dense channel-mixing weights; for GB, A denotes the adjacency with neighborhood $\mathcal{N}(i)$; for CaT, $\alpha_{cc'}$ are attention weights; for LCI, U, V are low-rank channel factors and gate_c is an adaptive gating coefficient. . 102

4.3 Synthetic Data Generation Pipeline. Latent trend/seasonal/noise bases are projected by W_k and summed to form $X(t)$. Optional local, causal, and calendar effects as well as anomalies and missingness can be injected to increase difficulty. 107

4.4 iDecPred performance on forecast horizons 96 and 720 with varying look-back lengths $L \in \{96, 192, 336, 720\}$ on the Electricity and Traffic. 112

4.5 Average radar chart of forecasting performance across 16 synthetic datasets under the $96-96$ setting. Curves closer to the boundary indicate better performance. 115

4.6 Efficiency–accuracy trade-offs on the C_862 dataset (L=96,H=96). Bubble size denotes parameter count. 116

4.7 Efficiency–accuracy trade-offs on the C_862 dataset (L=96,H=720). Bubble size denotes parameter count. 117

4.8 iDecPred decomposition on the Electricity dataset. The model jointly outputs trend, seasonal and final forecast components, enabling interpretable inspection on real-world data without access to ground-truth decompositions. 118

4.9 Decomposition–forecast relationship on C_7. Top row: iDecPred; bottom row: N-BEATS. (a) seasonal MAE vs. MAE_Y (iDecPred); (b) trend MAE vs. MAE_Y (iDecPred); (c) seasonal MAE vs. MAE_Y (N-BEATS); (d) trend MAE vs. MAE_Y (N-BEATS). Each point is a $\langle \text{window}, \text{channel} \rangle$; color encodes local density; red line is an OLS fit. Legends report n, Pearson r, and Spearman ρ. 121

5.1 Overview of the implemented pipeline. Trainable forecasting path: patchification; reprogramming with a concept bank; a fixed "Predict next values:" prefix plus reprogrammed patches fed to frozen GPT-2 as continuous embeddings (patch positions only); optional amplitude residual from raw patch tokens (Section 5.4.2); encoding to an initial latent state; patch-level autoregressive rollout to horizon H. Interpretability: reprogramming attention yields per-patch concept scores (Section 5.4.5); GPT-2 does not decode future numeric values. 147
5.2 Pearson correlation heatmap for the top-twenty busiest US airports: daily TSA-reported throughput per airport and two z-scored national exogenous indices (policy-related aviation restriction and public-health stress; construction in Appendix A). Airport–index associations appear in the heatmap cells linking each IATA code with *Policy (z)* or *Health stress (z)*. The daily panel spans 30 December 2018 to 31 May 2025. 157
5.3 STL decomposition of Atlanta (ATL) daily throughput: original series, trend, weekly seasonal component, and residuals. Weekly seasonality, longer-run trend (including COVID-19-era shifts), and irregular residual variation are visible for this representative hub. 158
5.4 Concept ablation (relative MSE and MAE change vs defined concepts). Both MSE and MAE show the relative change when using learned concepts compared to defined concepts. 162
5.5 Concept-based explanations for one input window. Top: the selected context window (one variable) with patch boundaries. Bottom: heatmap of concept scores over patches; concepts are selected by highest activation in this window (max over patches). 165
5.6 Multi-variable concept score heatmap at aligned patch indices. Rows are variables (airports) and columns are patch indices over the same context window; colors show the selected concept score across variables for each patch index (z-scored for visualisation). 165
5.7 Concept-score anchoring to patch statistics (median Spearman ρ heatmap) 168
5.8 Forecasting performance comparison across temporal resolutions. Panel (a) shows Daily resolution with MAE and MSE for models with and without exogenous features. Panel (b) shows Hourly resolution with MAE and MSE. Hourly models achieve lower absolute error while daily models show larger relative improvements from exogenous features. 171
5.9 Delta sign proportions for daily and hourly models on the test split. Each bar shows the proportion of forecast steps with negative, zero, and positive deltas under different scenarios (zero, lastx, constant). 174
5.10 Interactive scenario-based sensitivity analysis interface for airport passenger flow forecasting. The interface allows users to configure scenarios, visualize forecast comparisons, and examine sensitivity metrics across different shock conditions. 175

A.1 Data construction and empirical framework. This figure illustrates the data sources (airport throughput, GDELT news articles, policy timeline), data processing steps (normalisation, filtering), construction of information indices (health and control), airport structural attributes (international exposure and size), and the empirical framework (event study, exposure-interaction fixed effects models, and recovery analysis). 201
A.2 Geographic distribution of study airports: international exposure and recovery capacity. Colour indicates international passenger share in 2019 (redder = higher exposure), and marker size indicates recovery capacity (measured as average normalised passenger volume over days 30–180 post-shock). 203
A.3 News-based shock decomposition. Health-related information rises earlier and more gradually, while control-related information increases more visibly in early March. 204
A.4 Standardized health-related and control-related aviation information indices over time. Health-related information rises earlier and exhibits a pronounced increase in early March 2020, while control-related information remains relatively low through February and increases more visibly from early March. Both indices reach their peak around mid-March, though their temporal evolution differs in terms of onset and growth dynamics. . 206
A.5 Airport-level sensitivity to information shocks. Panel (a) relates sensitivity estimates to international passenger share, and panel (b) relates sensitivity to recovery performance. Across airports, sensitivity coefficients exhibit substantial heterogeneity, ranging from strongly negative to near zero. . . 208
A.6 Recovery trajectories by international exposure group. While both groups experience sharp initial declines, recovery paths diverge substantially, with lower-exposure airports recovering faster and reaching higher fractions of baseline passenger flow. High international exposure airports show persistently lower recovery throughout the recovery period. 210

B.1 Design Approach for Explainable Interface enhanced Human-Autonomy Teaming. 231
B.2 Clarification of Model Explainability, Explanations, Explainable Interfaces. 234
B.3 Explainable Interface in Human-Autonomy Teaming System. 237
B.4 A Comprehensive Framework for Model Explainability in Machine Learning. 239
B.5 Interactive Dynamics between Humans and Autonomous Systems through an Explainable Interface. 251
B.6 Extended User Experience Framework for Explainable Interface: Contrasting Task-oriented and Experience-oriented Approaches in Human-Autonomy Teaming. 259
B.7 Evaluation Framework of Explainable Interface for Human-Autonomy Teaming. 261
B.8 Explainable Interface in level 1 and level 2 surgical robotics 275
B.9 Explainable Interface Requirements for Human Vehicle Interaction in Conditional Driving Automation 279

List of Tables

2.1 Overview of forecasting models and the compatible explainability methods evaluated in this study. 29
2.2 Synthetic datasets used for oracle-based evaluation. Each dataset isolates a specific temporal or multivariate structure and provides corresponding ground-truth relevance used to assess attribution quality. 32
2.3 Summary of real-world datasets used for evaluation: key statistics for each split and temporal window. 36
2.4 Explainer performance relative to a random baseline. Success rate is the percentage of runs that beat random baseline. Average improvement is a coarse cross-dataset summary over primary metrics and should be interpreted cautiously because metric scales and directions differ across datasets (e.g., entropy is lower-better for regime switching). Best/worst datasets are based on success rate and improvement magnitude. 43
2.5 Seed-wise correlations between prediction accuracy (test MSE) and explainability metrics for three representative model–explainer combinations (seeds 42–44). 47
2.6 Deletion AUC on real-world airport data (more negative = larger degradation under deletion). Values are mean and standard deviation across the four representative datasets (ATL, LAX, Top 20 busiest airports, and SFO boarding-area departures), computed from per-run model–explainer–dataset deletion AUC means. 54
2.7 Prediction accuracy vs explainability metrics on real-world airport data. Correlations use test MSE and each behavioural metric over successful runs. Pearson and Spearman coefficients are descriptive summaries of linear and monotonic association. Sample sizes: deletion AUC and log noise stability, $n = 64$; randomisation similarity, $n = 36$ (neural models only; XGBoost omitted because randomisation is undefined). Noise stability is transformed as $\log_{10}(1 + \text{noise stability})$ due to heavy SHAP tails. 57

3.1 Comparison of models with and without exogenous variables. 80
3.2 Performance comparison for different sparse saliency tensor settings (mask percentiles) and XGBoost estimator counts. 82

4.1 Statistics of real-world benchmark datasets. 106
4.2 Multivariate forecasting results on prediction horizon 96. The input sequence length is fixed at 96 for all models. Lower MSE/MAE is better. . . 112
4.3 Ablation study results of the proposed model on seven benchmark datasets after removing individual components, including the Channel-Aware Encoder (CAE), Cross-Channel Mixer, Embedding Layer, and Fusion Block 114

4.4 Ground-truth evaluation on synthetic datasets. Metrics include Pearson correlation, trend/seasonal MAE, and final MAE. iDecPred* denotes weighted trend/seasonal reconstructions from the same iDecPred model. Boldface indicates the best result. 119
4.5 Airport benchmark panels: scale and summary statistics. Targets are TSA FOIA-released throughput for the twenty busiest U.S. hubs [41]; exogenous shocks follow the GDELT-based construction in Appendix A [12]. 122
4.6 Airport case study results for the 96–96 setting (Top 20 busiest U.S. airports). 123
4.7 Best-performing configurations (lowest MSE) on the airport case study. . 124

5.1 Flattened curated vocabulary after merging all core and optional domain groups with deduplication (M=249). Airport-focused runs in this chapter use M=185 (core plus airport/passenger only). 153
5.2 Top-20 busiest U.S. airports passenger flow forecasting performance at different granularities. Results are reported for a single variable-horizon Patch-AR model with defined concepts trained per granularity, compared with iDecPred baseline. 159
5.3 Patch-AR vs Single-Step AR (Daily) 160
5.4 Patch-AR vs Single-Step AR (Hourly) 160
5.5 Full Concept Ablation (Daily, All Horizons) 163
5.6 Full Concept Ablation (Hourly, All Horizons) 163
5.7 Stability Across Reruns and Noise 166
5.8 Concept-Score Anchoring to Patch Statistics (Selected Concepts) 167
5.9 Window-Level Significance (Selected Concept-Statistic Pairs) 167
5.10 Forecasting accuracy comparison across temporal resolutions (96-step horizon): no exogenous shock inputs vs. both contemporaneous raw shock channels. 170
5.11 Computational efficiency comparison across resolutions and models. Inference time is for a single 96-step prediction. Parameter counts are in millions (M). 173
5.12 Variable-horizon versus single-horizon training at matched horizons. Variable-horizon results are taken from Table 5.2. 176

A.1 Exposure–information interaction fixed effects estimates 209
A.2 Robustness checks: Stability of exposure–information interaction effects across alternative specifications 213

B.1 Overview of Relevant Surveys on XAI and HAT, with Main Topic for Contribution 231
B.2 Design of Explainable Interface for Different Autonomy Levels 272

List of Abbreviations

Adam	Adaptive Moment Estimation
AIC	Akaike Information Criterion
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
ATL	Hartsfield-Jackson Atlanta International Airport
AUC	Area Under the Curve
CI	Confidence Interval
CNN	Convolutional Neural Network
CUDA	Compute Unified Device Architecture
CV	Coefficient of Variation
DEN	Denver International Airport
DFW	Dallas/Fort Worth International Airport
DL	DeepLift
EI	Explainable Interface
EPSRC	Engineering and Physical Sciences Research Council
ES-RNN	Exponential Smoothing Recurrent Neural Network
FE	Fixed Effects
FEAS	Faculty of Engineering and Applied Sciences
FEDformer	Frequency Enhanced Decomposed Transformer
FITS	Frequency Interpolation for Time Series Forecasting
FOIA	Freedom of Information Act
GAN	Generative Adversarial Network
GDELT	Global Database of Events, Language, and Tone

GNN	Graph Neural Network
Grad-CAM	Gradient-weighted Class Activation Mapping
GPT-2	Generative Pre-trained Transformer 2
GRU	Gated Recurrent Unit
GS	GradientSHAP
GPU	Graphics Processing Unit
HAT	Human-Autonomy Teaming
HCI	Human-Computer Interaction
iDecPred	Interpretable Decomposition Prediction
IG	Integrated Gradients
i.i.d.	Independent and identically distributed
iTransformer	Inverted Transformer for Time Series Forecasting
ITS	Intelligent Transportation Systems
JFK	John F. Kennedy International Airport
LAX	Los Angeles International Airport
LGA	LaGuardia Airport
LLM	Large Language Model
LIME	Local Interpretable Model-agnostic Explanations
LSTNet	Long- and Short-term Time-series Network
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MIA	Miami International Airport
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
MSTL	Multiple Seasonal-Trend Decomposition using Loess
NFM	Neural Fourier Model
NLP	Natural Language Processing
ODE	Ordinary Differential Equation

OLS	Ordinary Least Squares
ORD	O'Hare International Airport
Patch-AR	Patch-Autoregressive
PatchTST	Patch Time Series Transformer
PCA	Principal Component Analysis
PHX	Phoenix Sky Harbor International Airport
ReVIN	Reversible Instance Normalization
ReLU	Rectified Linear Unit
RISE	Randomized Input Sampling for Explanation
RNN	Recurrent Neural Network
RMSE	Root Mean Squared Error
RobustSTL	Robust Seasonal-Trend Decomposition Algorithm
SFO	San Francisco International Airport
SHAP	Shapley Additive Explanations
SOTA	State of the Art
SSN	State Space Network
STEM	Science, Technology, Engineering and Mathematics
STL	Seasonal-Trend Decomposition using Loess
TCAV	Testing with Concept Activation Vectors
TCN	Temporal Convolutional Network
TBATS	Trigonometric Box-Cox ARIMA Trend Seasonal
TFT	Temporal Fusion Transformer
TSA	Transportation Security Administration
VAR	Vector Autoregression
WFTNet	Wavelet Frequency Transformation Network
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting

1

Introduction and Research Framework

Airport passenger flow forecasting is a core capability for modern air transportation systems, supporting operational decisions such as transfer passenger flow management [5], gate and stand allocation based on flight schedules [10], and terminal capacity expansion planning [11]. These high-stakes decisions increasingly rely on multivariate time-series forecasting models that ingest heterogeneous signals (e.g., historical passenger flows across multiple terminals or routes, calendar effects, and exogenous covariates) [14, 9]. As forecasting pipelines become more complex and more consequential, practitioners and regulators face a familiar tension: the strongest predictive models are often less transparent, yet regulatory frameworks increasingly mandate transparency and explainability. The European Union Aviation Safety Agency (EASA) advocates for human-centric AI approaches that prioritise transparency [3], the UK Civil Aviation Authority (CAA) has established strategies for regulating AI and advanced automation in aerospace [13], and the EU Artificial Intelligence Act mandates transparency obligations for high-risk AI systems [2]. These requirements reflect the critical need for trust, auditing, and actionable diagnosis in high-stakes decision-making contexts.

This thesis studies explainable multivariate time-series forecasting with a focus on airport passenger flow and related high-dimensional forecasting settings. The main task is formulated as a regression problem: given historical observations of one or more time series and optionally exogenous covariates, the objective is to predict future values (or multi-step trajectories) of the target series over a specified horizon. The goal is not only to

improve predictive accuracy, but to produce explanations that are reliable enough to support scientific understanding and operational decision-making. In this thesis, explainability refers to the ability to articulate why a model produces a given forecast or which inputs and structures drive it; the specific form of explanation varies by chapter. Chapter 2 and Chapter 3 focus on post-hoc explainability: attribution of predictions to input time steps and variables (e.g., via saliency or Shapley-style methods), with Chapter 3 improving the usability of such attributions through sparse, aligned saliency. Chapter 4 offers intrinsic explainability by design: the proposed model (iDecPred) decomposes the series into trend, seasonal, and residual components and forecasts each explicitly, so that explanations are the component-level outputs themselves. Chapter 5 provides semantic and scenario-based explainability: concept-level descriptions of patch dynamics via semantic anchoring, and robustness diagnostics under alternative exogenous scenarios.

1.1 Airport Passenger Flow Forecasting as a High-Stakes Decision Problem

In this thesis, the term airport passenger flow denotes the volume of passengers passing through an airport or a designated part of it, such as a terminal or boarding area, per unit of time. It is typically measured as throughput, for example in hourly or daily passenger counts. Airport passenger flow exhibits strong temporal structure, including seasonality, long-term trends, and irregular shocks, as well as cross-series interactions driven by shared factors such as terminal usage, airline schedules, and macro-level events [6]. Forecasts are required at multiple lead times. Short-term forecasts support operational control over horizons ranging from hours to days, while longer-term forecasts inform planning and budgeting over weeks to months. In real deployments, these horizons are rarely independent. Decision-makers require coherent forecasts across horizons and need to understand when and why forecasts change under alternative assumptions about future conditions.

From a decision perspective, forecast errors can be costly: under-forecasting can lead to congestion and service degradation, while over-forecasting can lead to overstaffing, wasted capacity, and inefficient capital allocation [1]. In this context, explainable forecasting is not merely desirable but essential. Operational and planning decisions are made by humans who must understand and trust the basis of a forecast before committing resources; keeping humans in the loop is a requirement in these settings, and explainability is what makes such oversight effective. Black-box predictions are less likely to be adopted or correctly acted upon. When forecasts go wrong, stakeholders need to diagnose whether failures stem from data quality, model assumptions, or external shocks, so that processes can be corrected and future forecasts improved. In addition, aviation is a heavily regulated sector where transparency, auditability, and human-in-the-loop oversight are increasingly mandated for high-impact AI systems, as reflected in the regulatory references above. For these reasons, this thesis treats explainability as a core requirement for airport passenger flow forecasting, not an optional add-on. Explainability can reduce the risks above by enabling model diagnostics (e.g., detecting spurious reliance on proxy features) and improving communication between model developers and operational stakeholders.

1.2 Challenges of Explainability in Time-Series

Applying explainability methods to time-series forecasting introduces challenges that are less prominent in static prediction tasks:

- **Temporal dependence and autocorrelation.** Inputs are not independent; explanations must reflect temporal dynamics rather than treating time steps as interchangeable features.
- **High dimensionality from multivariate signals and windowing.** Sliding-window construction can multiply the effective input dimension, producing dense and noisy attribution maps that are difficult for humans to interpret.

- **Model and explainer mismatch.** Many popular explainers were designed for vision or tabular settings and may behave differently under forecasting objectives and sequential architectures.
- **Lack of ground-truth explanations in real data.** Unlike synthetic benchmarks where generating processes can define attribution ground truth, real-world airport data rarely provides an oracle notion of "correct" feature importance.
- **Faithfulness versus usefulness.** A visually plausible explanation may not be faithful to the model, and a faithful explanation may not improve decision quality. This creates a three-way tension between accuracy, faithfulness, and decision utility.

These challenges motivate careful empirical evaluation and the development of explanation strategies tailored to forecasting settings, where the explanation object is not a single class label but a multi-step future trajectory.

1.3 Research Gaps and Motivation

Despite rapid progress in explainable AI, three gaps remain particularly important for multivariate time-series forecasting in transportation:

- **Limited evidence-based guidance for post-hoc explainers in forecasting.** In practice, explainers are often selected by convenience, yet their quality can vary substantially across architectures, temporal patterns, and datasets.
- **Human interpretability of saliency in multivariate forecasting.** Standard saliency outputs are often dense and noisy, especially under windowed inputs, making it difficult to translate attribution into actionable insights.
- **Need for structured, intrinsically and semantically interpretable forecasting models.** Post-hoc explanations alone may not provide the kind of temporal,

component-wise, or scenario-level (semantic) understanding required in high-stakes deployment, particularly for long-horizon forecasts and exogenous shocks.

This thesis aims to address these research gaps through three complementary avenues: (1) conducting large-scale, systematic evaluations of post-hoc explainers in the context of multivariate time-series forecasting; (2) developing improved post-processing techniques for saliency maps, enabling more interpretable and actionable explanations that benefit both human users and downstream tasks; and (3) proposing intrinsically interpretable forecasting model architectures and diagnostic tools tailored for long-horizon and scenario-based airport passenger flow prediction.

1.4 Research Questions

Guided by the above gaps, the thesis investigates the following research questions:

- **RQ1:** How do widely used post-hoc explainability methods behave for multivariate time-series forecasting, and how sensitive is explanation quality to temporal patterns, forecasting architectures, and explainer choice?
- **RQ2:** How can saliency-based explanations be post-processed or aggregated over time, features, or semantic groupings to improve human interpretability in multivariate forecasting, and can such explanations be used to improve downstream model performance or to guide model simplification?
- **RQ3:** Beyond token-level (time-step or feature-level) explanations, what forms of structured or higher-level explanations can be provided for time-series forecasts, and how can these be made reliable and actionable for decision-makers?
- **RQ4:** Can we design forecasting models that are intrinsically interpretable by construction (e.g., through explicit trend/seasonal/residual or concept components) while maintaining accuracy and computational efficiency for long-horizon multivariate forecasting?

- **RQ5:** How can forecast explanations support horizon-generalizable forecasting under exogenous-shock scenarios, especially using concept-based and scenario-driven interpretability frameworks?

1.5 Thesis Contributions

The main contributions of this thesis can be summarised as follows:

- **An hourly airport passenger flow dataset with external features.** An integrated hourly passenger-flow dataset was constructed from Freedom of Information Act (FOIA)-released airport throughput data [12] and custom Global Database of Events, Language, and Tone (GDELT)-based indices capturing external information signals [4]. The external features contain two standardised indices built from aviation-related news: (i) health-related information, measuring the intensity of coverage linking epidemiological risk (e.g., COVID-19, pandemic, outbreak) to aviation; and (ii) control-related information, measuring coverage of travel restrictions, border controls, and operational measures affecting air travel. The detailed construction process is described in Appendix A. These covariates enable multivariate forecasting with explicit external drivers and facilitate robustness-oriented analyses under evolving information environments.

- **Evidence-based evaluation of post-hoc explainers for airport passenger flow forecasting.** A systematic empirical evaluation of widely used explainers across multiple forecasting architectures was conducted, using both synthetic datasets with attribution ground truth and real-world airport passenger flow datasets. The experiments show that no single explainer consistently dominates across scenarios and that predictive accuracy and explanation quality can be weakly correlated.

- **Sparse, normalised saliency for improved interpretability and practical benefit.** A flexible post-processing approach is proposed to generate sparse saliency maps

(with optional window aggregation/alignment) to better match human intuition in multivariate time series. Saliency-guided data augmentation is further explored, and improvements are demonstrated when transferring saliency-derived information to simpler models.

- **Forecasting future values of each component and the overall series via decomposition.** A framework is proposed that predicts the future trajectories of trend, seasonality, and residual elements separately, as well as the recombined total forecast. This enables interpretable forecasting by providing visibility into how each decomposed part contributes to the final prediction, moving beyond post-hoc explanations.

- **Variable-horizon patch-autoregressive forecasting with semantic anchoring and robustness diagnostics.** A unified framework is presented for forecasting across multiple horizons with a single model via patch-level rollouts; semantic anchoring is introduced to improve interpretability of latent dynamics, and robustness is analysed under exogenous shocks using scenario-based sensitivity analysis.

1.6 Thesis Structure

This thesis is organised on a paper basis. Figure 1.1 provides an overview of the thesis structure and the logical connections between chapters; the narrative progresses from post-hoc explanation evaluation, through more usable and intrinsically interpretable forecasting models, to deployment-relevant diagnostics.

- **Chapter 2** evaluates post-hoc explainability methods for multivariate time-series forecasting using synthetic and real-world airport passenger flow datasets.

- **Chapter 3** introduces sparse saliency post-processing to improve explanation readability and usability.

- **Chapter 4** presents iDecPred, an intrinsically interpretable neural decomposition model for long-horizon multivariate forecasting.

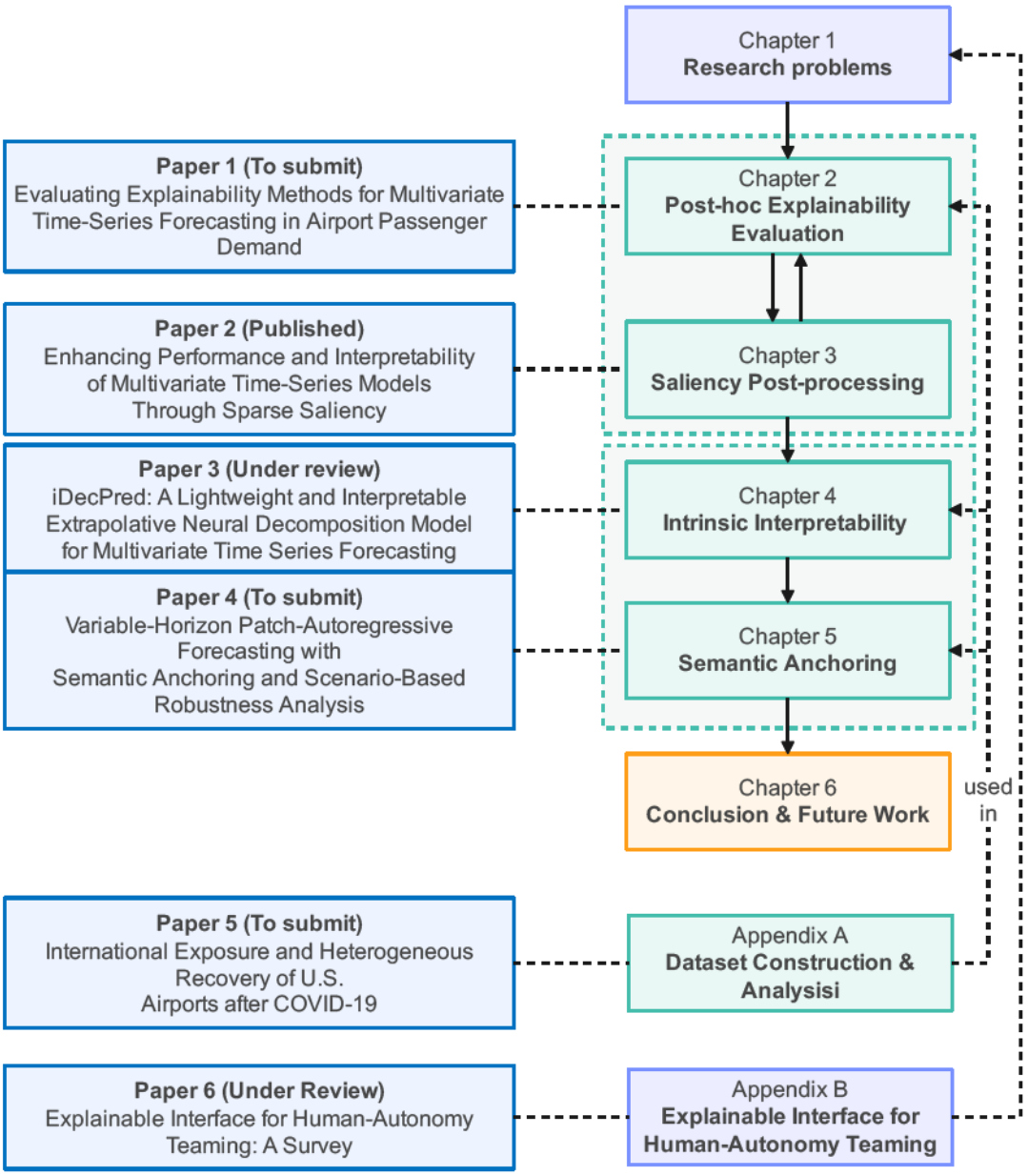


Figure 1.1: Thesis structure and chapter connections.

- **Chapter 5** addresses deployment-relevant variable-horizon forecasting with semantic anchoring and scenario-based robustness analysis.
- **Chapter 6** summarises the thesis contributions and discusses limitations and future research directions.

The chapters form a progressive build-up: the empirical explainer evaluation in Chapter 2 motivates the need for more human-readable and reusable explanations in Chapter 3; limitations of purely post-hoc attribution motivate the intrinsically interpretable decomposition approach in Chapter 4; and the variable-horizon deployment setting motivates the methods developed in Chapter 5, which are synthesised in Chapter 6. The airport passenger flow dataset and GDELT-based external features used in several of these chapters are documented in Appendix A.

Appendix A documents the construction of the airport passenger flow dataset and the GDELT-based external features used throughout the thesis. The core chapters focus on developing explainability methods at the model level. Appendix B complements this by addressing explainable interfaces for human-autonomy teaming (HAT) from a conceptual and design perspective [8]. It clarifies the distinction between model explainability, explanations, and explainable interfaces, and surveys design and evaluation considerations for HAT. Thus Appendix B provides the broader deployment context within which the time-series explainability methods developed in this thesis can be situated when explanations are presented to operators or stakeholders in high-stakes settings.

1.7 Publications and Manuscripts in Preparation

This thesis is submitted to Cranfield University in partial fulfilment of the requirements for the degree of Doctor of Philosophy. The work presented is original and was carried out during my period of study at Cranfield University. This thesis has not been submitted previously, in whole or in part, for any other degree or qualification. Publications and manuscripts arising from this research are listed below.

- Evaluating Explainability Methods for Multivariate Time-Series Forecasting in Airport Passenger Demand. Xiangqi Kong, Yang Xing, Zeyu Liu, Antonios Tsourdos, and Andreas Wikander (to submit)

- Enhancing Performance and Interpretability of Multivariate Time-Series Model Through Sparse Saliency. Xiangqi Kong, Yang Xing, Zeyu Liu, Antonios Tsourdos, and Andreas Wikander *In Proceedings of the IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, Edmonton, Canada, September 24–27, 2024. IEEE, 2024. DOI: 10.1109/ITSC58415.2024.10920008. [7] (published)

- iDecPred: A Lightweight and Interpretable Extrapolative Neural Decomposition Model for Multivariate Time Series Forecasting. Xiangqi Kong, Yang Xing, Zeyu Liu, Antonios Tsourdos, and Andreas Wikander (IEEE Transactions on Neural Networks and Learning Systems under review)

- Variable-Horizon Patch-Autoregressive Forecasting with Semantic Anchoring and Scenario-Based Robustness Analysis. Xiangqi Kong, Yang Xing, Zeyu Liu, Antonios Tsourdos, and Andreas Wikander (to submit)

- International Exposure and Heterogeneous Recovery of U.S. Airports after COVID-19. Xiangqi Kong, Yang Xing, Antonios Tsourdos, and Andreas Wikander (to submit)

- Explainable Interface for Human-Autonomy Teaming: A Survey. Xiangqi Kong, Yang Xing, Antonios Tsourdos, Ziyue Wang, Weisi Guo, Adolfo Perrusquia, and Andreas Wikander. arXiv:2405.02583, 2024. [8] (ACM Computing Surveys under review)

References

[1] ACI EUROPE. Monitoring of passenger flows and mitigation of queues and crowds at airports. Technical report, Airports Council International Europe, Brussels, Belgium, 2022. Guidance document.

[2] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 (artificial intelligence act), 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ%3AL_202401689. Official Journal of the European Union; see Article 13 for transparency obligations.

[3] European Union Aviation Safety Agency (EASA). Artificial intelligence roadmap 2.0: Human-centric approach to AI in aviation. Technical report, EASA, May 2023. URL https://www.easa.europa.eu/sites/default/files/dfu/easa_ai_roadmap_2_0_2023.pdf. Version 2.0.

[4] GDELT Project. A global database of society. https://www.gdeltproject.org/, 2022. Accessed: 2026-01-13.

[5] Xiaojia Guo, Yael Grushka-Cockayne, and Bert De Reyck. Forecasting airport transfer passenger flow using real-time data and machine learning. *Manufacturing & Service Operations Management*, 24(6):3193–3214, 2022. doi: 10.1287/msom.2021.0975.

[6] David H. Hopfe, Kiljae Lee, and Chunyan Yu. Short-term forecasting airport passenger flow during periods of volatility: Comparative investigation of time series vs. neural network models. *Journal of Air Transport Management*, 115:102525, 2024. doi: 10.1016/j.jairtraman.2023.102525.

[7] Xiangqi Kong, Yang Xing, Zeyu Liu, Antonios Tsourdos, and Andreas Wikander. Enhancing performance and interpretability of multivariate time-series model through sparse saliency. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, pages 3265–3270. IEEE, 2024. doi: 10.1109/itsc58415.2024.10920008.

[8] Xiangqi Kong, Yang Xing, Antonios Tsourdos, Ziyue Wang, Weisi Guo, Adolfo Perrusquia, and Andreas Wikander. Explainable interface for human-autonomy teaming: A survey. *arXiv preprint arXiv:2405.02583*, 2024.

[9] Hongtao Li, Kun Jin, Shaolong Sun, Xiaoyan Jia, and Yongwu Li. Metro passenger flow forecasting through multi-source time-series fusion: An ensemble deep learning approach. *Applied Soft Computing*, 120:108644, 2022. doi: 10.1016/j.asoc.2022.108644.

[10] Lin Lin, Xiaochen Liu, Xiaohua Liu, Tao Zhang, and Yang Cao. A prediction model to forecast passenger flow based on flight arrangement in airport terminals. *Energy and Built Environment*, 4(6):680–688, 2023. doi: 10.1016/j.enbenv.2022.06.006.

[11] Erma Suryani, Shuo-Yan Chou, and Chih-Hsien Chen. Air passenger demand forecasting and passenger terminal capacity expansion: A system dynamics framework. *Expert Systems with Applications*, 37(3):2324–2339, 2010. doi: 10.1016/j.eswa.2009.07.041.

[12] Transportation Security Administration. FOIA electronic reading room. `https://www.tsa.gov/foia/readingroom`, 2022. Accessed: 2026-01-13.

[13] UK Civil Aviation Authority. The CAA's response to emerging AI-enabled automation: Part A: Strategy for regulating AI and advanced automation in aerospace (CAP3064A). Technical report, Civil Aviation Authority, 2024. URL `https://www.caa.co.uk/publication/download/23437`. Version 1.3.2; first published November 2024.

[14] Renju Aleyamma Zachariah, Sahil Sharma, and Vijay Kumar. Systematic review of passenger demand forecasting in aviation industry. *Multimedia Tools and Applications*, 82(30):46483–46519, 2023. doi: 10.1007/s11042-023-15552-1.

2

Post-hoc Explainability Evaluation

PAPER I – Post-hoc Explainability Evaluation

Evaluating Explainability Methods for Multivariate Time-Series Forecasting in Airport Passenger Flow

Abstract

Explainability is increasingly important for multivariate time series forecasting in transportation research, where complex models are used to analyse flow patterns and temporal dynamics. However, most Explainable Artificial Intelligence (XAI) methods have been developed for other domains, and their behaviour in time series forecasting tasks, particularly for airport passenger flow, remains insufficiently understood. This chapter presents a systematic empirical evaluation of post-hoc explainability methods for multivariate time series forecasting in airport passenger flow, focusing on attribution for a fixed forecast target.

Five widely used explainability methods, Integrated Gradients (IG), DeepLift (DL), GradientSHAP (GS), Local Interpretable Model-agnostic Explanations (LIME), and Shapley Additive Explanations (SHAP), are evaluated across six forecasting models, including recurrent, convolutional, attention based, and tree based architectures. The evaluation is conducted on nine datasets, comprising five synthetic datasets with known attribution ground truth. To address the lack of ground truth explanations in real world data, a dual evaluation setting is adopted that combines oracle based assessment on synthetic data with behavioural evaluation on real world data.

The results indicate that explainability performance is strongly influenced by temporal pattern characteristics, forecasting architecture, and explainer choice, and that no single explainability method consistently performs well across all scenarios. Associations between prediction accuracy and explainability diagnostics are metric-dependent: on synthetic data, combination-level correlations between test MSE and oracle scores are generally weak, whereas on real-world data test MSE can co-vary strongly with deletion-based behavioural scores while relating weakly to other diagnostics. Evaluation on real-world data also

highlights limitations of commonly used behavioural metrics when applied to time series forecasting. Together, these findings provide empirical, evidence based guidance for the selection and interpretation of post-hoc explainability methods in transportation time series forecasting studies.

2.1 Introduction

In this chapter, explainability is taken to mean post-hoc attribution: the assignment of importance scores to input time steps and variables that indicate their contribution to a specific forecast. Instead of proposing a new forecasting model, this chapter focuses on evaluating how well existing post-hoc explainers (Integrated Gradients, DeepLift, GradientSHAP, LIME, and SHAP) provide such attributions when applied to trained forecasting models, and under what conditions their outputs align with ground truth or with model behaviour. The explainability offered here is therefore that of attribution maps over the input window, interpretable as temporal and variable-wise importance for a chosen forecast target.

Forecasting airport passenger flow is a fundamental task in air transportation research, with applications in capacity planning, congestion analysis, and long term infrastructure assessment [33]. Recent advances in multivariate time series forecasting have led to the widespread adoption of data driven models capable of capturing complex temporal dependencies and interactions across multiple passenger flow streams [6]. Deep learning based approaches, including recurrent, convolutional, and attention based architectures, have demonstrated strong predictive performance on airport passenger flow data, with comparative studies showing their advantages over traditional time series methods [16]. However, the increasing complexity of these models has also reduced their interpretability, limiting their usefulness for analysis and raising concerns regarding transparency and trust in transportation forecasting studies and regulatory contexts.

Explainability has therefore emerged as an important consideration in transportation

related forecasting research, particularly when models are used to analyse flow drivers or assess the impact of external factors such as disruptions, policy changes, or long term trends [1]. A wide range of post-hoc XAI methods have been proposed to attribute model predictions to input features, including gradient based and model agnostic approaches [9, 25, 27, 36, 11]. While these methods have been extensively studied in domains such as computer vision and natural language processing, their behaviour in time series forecasting remains less well understood [3]. Existing work on time series explainability has focused primarily on classification tasks [35], whereas multivariate time series forecasting poses distinct challenges due to temporal dependence, feature redundancy, and interactions across variables and time steps.

A central challenge in this context is the evaluation of explainability methods for real world airport passenger flow data, where ground truth explanations are unavailable. In the absence of ground truth, prior studies often rely on behavioural evaluation metrics, such as feature deletion or model randomisation tests, to assess whether explanations are consistent with model behaviour [30, 7].

In this chapter, a systematic empirical evaluation of post-hoc explainability methods for multivariate time series forecasting in airport passenger flow is presented. Multiple explainability techniques are evaluated across diverse forecasting architectures using both synthetic datasets with known attribution ground truth and real-world airport passenger flow datasets. To address the absence of ground truth in real-world data, a dual evaluation framework is adopted that combines oracle-based evaluation on synthetic data with behavioural evaluation on real-world data.

The main contributions of this work are twofold. First, a large-scale empirical evaluation of post-hoc explainability methods for multivariate time series forecasting is conducted under a fixed-target protocol, together with a dual evaluation framework that separates explanation correctness, assessed on synthetic data, from behavioural properties observed on real-world data, enabling controlled and principled comparison across models, explainers, and datasets. Second, based on this framework, the strengths and limitations of existing

explainability methods are systematically characterised, and evidence-based guidance is provided for their selection and interpretation in transportation time series forecasting studies.

2.2 Related Work

2.2.1 Time Series Forecasting in Transportation Research

Time series forecasting has been widely studied in transportation research, with applications including traffic flow prediction, passenger flow analysis, and transportation system planning. In the aviation domain, airport passenger flow forecasting has been used to support studies on congestion patterns, capacity assessment, and long-term infrastructure development. Early work primarily relied on statistical models such as ARIMA, SARIMA, SARIMAX [16], which provide interpretability but are often limited in their ability to model nonlinear temporal dynamics during periods of volatility.

In recent years, transportation forecasting has increasingly transitioned toward data driven methodologies. Deep learning architectures such as recurrent neural networks (RNNs), temporal convolutional networks (TCNs), and attention based models including Transformers have demonstrated strong predictive performance by capturing long range temporal dependencies and complex interactions across multiple passenger flow streams [13, 21]. Representative models including Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), TCNs, Transformer based variants, as well as approaches incorporating Generative Adversarial Networks (GANs) and Graph Neural Networks (GNNs) have been widely applied to transportation forecasting tasks such as airport passenger flow prediction [26, 19]. Despite these methodological advances, most existing studies primarily emphasize predictive accuracy, while the interpretability of learned models and their decision making mechanisms remains relatively underexplored.

2.2.2 Post-hoc Explainability Methods for Time Series Models

Post-hoc explainability has become a predominant paradigm for interpreting complex time series models. Although the choice between intrinsically interpretable models and post-hoc explanation methods remains debated in the literature [29, 24], the increasing reliance on highly expressive deep learning architectures and the need to explain already-deployed black-box systems make post-hoc approaches practically unavoidable in many real-world time series applications, as it enables explanation without modifying the underlying forecasting or classification architecture. Existing approaches can be broadly categorized into attribution-based methods, segment- or pattern-based explanations, counterfactual explanations, and post-hoc intervention frameworks.

Attribution-based methods assign importance scores to input features or time steps to explain individual predictions. Representative techniques include LIME [28], which fits a local surrogate model around a target instance, and SHAP [23], which provides a unified Shapley-value-based framework with theoretical guarantees such as consistency and local accuracy. Integrated Gradients [32] further extends gradient-based attribution by enforcing axiomatic properties, making it a common baseline for deep time series models. To better adapt attribution to sequential data, TimeSHAP [5] introduces sequence-aware perturbations and produces explanations at multiple granularities, including feature-, time-step-, and event-level importance. Despite their flexibility, these methods typically operate at the level of individual time points, which may limit their semantic interpretability for long or highly structured sequences.

To address the limitations of pointwise attribution, recent work has shifted toward segment- or pattern-level explanations. ShapeX [17] proposes a shapelet-driven post-hoc explanation framework for time series classification, where explanations are expressed in terms of salient subsequences rather than isolated time steps, improving causal fidelity and interpretability. Complementary to explanation generation, WAE [7] introduces a window-based quantitative evaluation metric for attribution-based explanations in time series forecasting. By measuring attribution errors at the window level, WAE enables systematic

comparison of explainability methods and highlights the importance of temporally coherent explanation units.

Another line of work focuses on counterfactual explanations, which aim to answer how a prediction would change under minimal and plausible modifications of the input sequence. Native-Guide [10] generates counterfactuals by leveraging nearest-neighbor instances from alternative classes, ensuring realism and feasibility. Counterfactual explanations for multivariate time series [4] further extends counterfactual reasoning to multivariate time series by explicitly accounting for inter-variable dependencies and temporal coherence. While counterfactual explanations provide actionable insights, they are often computationally expensive and sensitive to data distribution assumptions.

Beyond explanation, recent studies explore post-hoc intervention mechanisms that directly improve model outputs. PIR [22] introduces an instance-aware post-hoc revision framework for time series forecasting, which identifies unreliable predictions and refines them using both local contextual adjustment and global retrieval of similar historical instances. Unlike traditional explainability methods that focus on interpretation alone, PIR demonstrates that post-hoc analysis can be leveraged to enhance instance-level reliability and forecasting performance.

Overall, existing post-hoc explainability methods offer diverse perspectives on interpreting time series models, yet challenges remain in balancing semantic interpretability, faithfulness, and practical utility, particularly for long-horizon and instance-specific forecasting scenarios.

In transportation time series explainability domain, early work focuses on feature- and temporal-level explanations in traffic forecasting. LSTM-based predictors are often paired with saliency methods such as Integrated Gradients or attention mechanisms to highlight key time steps and exogenous factors affecting traffic patterns [26, 31]. Graph-based transportation models further incorporate spatial dependencies, and explainability methods are adapted accordingly: for example, GNNExplainer and related perturbation-based approaches are used to identify critical links and nodes in spatio-temporal traffic

graphs [40, 20]. Counterfactual frameworks generate minimal changes to input signals (e.g., sensor data) that would alter a predicted congestion event, helping planners understand the sensitivity of forecasts to specific temporal patterns [38].

2.2.3 Evaluation of Explainability Methods

Evaluating explainability methods for time series forecasting models remains challenging due to the absence of ground-truth explanations in real-world data. In practical applications, the underlying data-generating processes and factors driving model predictions are unknown, making direct assessment of explanation correctness difficult.

To address this challenge, sanity check-based evaluations verify whether explanations are sensitive to learned model parameters. Adebayo et al. [2] introduced model parameter randomisation tests, comparing explanations from trained models against those from randomly initialised models. If explanations remain unchanged, the method may not reflect the model's learned representations.

Explanation fidelity is often evaluated using perturbation-based behavioural tests [12]. Hooker et al. [14] proposed the Remove-And-Retrain (ROAR) framework, which removes important features and retrains the model, providing a more reliable fidelity measure than direct deletion by mitigating distribution shift effects.

For time series settings, Ismail et al. [18] demonstrated that many attribution methods developed for static data fail to disentangle temporal and feature-wise importance in multivariate time series. They introduced controlled synthetic benchmarks with predefined attribution targets, enabling principled evaluation along temporal and feature dimensions.

Robustness to input perturbations is another important evaluation criterion. Formal metrics such as explanation fidelity and sensitivity [39] quantify the consistency between explanations and model behavior. Although these behavioural tests do not provide direct ground-truth validation, they offer complementary perspectives on faithfulness, robustness, and reliability for evaluating post-hoc explainability methods in multivariate time series forecasting.

2.2.4 Positioning of This Work

This chapter builds on existing research on time series explainability and its evaluation [18, 35] by providing a systematic empirical evaluation of post-hoc explainability methods for multivariate time series forecasting in airport passenger flow. By explicitly distinguishing between explanation correctness on synthetic data and behavioural properties on real-world data, this work provides a clearer assessment of the strengths and limitations of post-hoc explainability methods for transportation time series forecasting.

2.3 Problem Setting and Evaluation Framework

This study focuses on evaluating post-hoc explainability methods for multivariate time series forecasting in airport passenger flow. The objective of this section is to formalise the forecasting problem, define the explainability objectives, and describe the evaluation framework adopted in this work.

2.3.1 Multivariate Time Series Forecasting for Airport Passenger Flow

Let $X \in \mathbb{R}^{T \times V}$ denote a multivariate time series, where T is the input sequence length and V is the number of variables. Given a historical input window $X_{t-T+1:t}$, the forecasting task aims to predict the next H steps:

$$\hat{Y}_{t+1:t+H} = f_\theta(X_{t-T+1:t}), \quad \hat{Y}_{t+1:t+H} \in \mathbb{R}^{H \times V}. \tag{2.1}$$

In this implementation, $T = t_{\text{in}}$ and $H =$ horizon are used. The values of $(t_{\text{in}}, \text{horizon})$ are dataset-dependent; Section 2.4.3 gives the full description of the real-world and synthetic datasets used in this chapter, including the input length and forecast horizon for each (see Table 2.3 for real-world data).

2.3.2 Explainability Objectives in Time Series Forecasting

Explainability in multivariate time series forecasting aims to identify which components of the input sequence contribute most to a given prediction. In the context of airport passenger flow, explanations should capture both temporal relevance and variable importance. Specifically, the following objectives are considered:

- **Temporal attribution:** identify influential time steps within the input window.
- **Variable attribution:** identify the relative importance of different passenger-flow related variables.
- **Time–variable interactions:** identify which variable at which time step is most influential for a specific forecast target.

Post-hoc explainability methods in our framework produce an attribution tensor $A \in \mathbb{R}^{T \times V}$ for a chosen scalar target output (by default, a specific horizon index and variable index). This supports joint quantitative analysis across both the temporal and variable dimensions.

2.3.3 Dual Evaluation Strategy

Evaluating explainability methods for real world transportation data is challenging due to the absence of ground truth explanations. Prior studies often rely on behavioural evaluation metrics that assess explanation consistency with model outputs rather than correctness. However, many behavioural metrics were originally developed for other data modalities and may not be well suited to time series forecasting, where temporal autocorrelation and feature redundancy are prevalent, complicating the interpretation of attribution-based evaluations.

To address these challenges, a dual evaluation strategy is adopted that combines oracle based evaluation on synthetic datasets with behavioural evaluation on real world airport passenger datasets.

Synthetic datasets are constructed with predefined temporal and variable importance patterns, enabling direct comparison between attribution results and known ground truth. This allows assessment of explanation correctness under controlled conditions.

Real world datasets do not provide ground truth explanations. For these datasets, explainability methods are evaluated using behavioural metrics that examine how model performance responds to input perturbations or model modifications. While such metrics do not establish explanation correctness, they provide insight into the consistency and robustness of explanations in practical forecasting settings.

By combining these two evaluation perspectives, this evaluation framework aims to clarify both the strengths and limitations of existing explainability methods when applied to multivariate time series forecasting in airport passenger flow. Figure 2.1 summarises the evaluation setup.

Formally, for an input window $X \in \mathbb{R}^{T \times V}$ and a forecasting model f_θ that outputs $\hat{Y} = f_\theta(X) \in \mathbb{R}^{H \times V}$, a scalar forecast target is defined

$$s_{h,v}(X) = \hat{Y}_{h,v}, \tag{2.2}$$

where $h \in \{0, \ldots, H-1\}$ and $v \in \{0, \ldots, V-1\}$. An explainer produces an attribution map for this target,

$$A^{(h,v)}(X) \in \mathbb{R}^{T \times V}. \tag{2.3}$$

To evaluate attributions along the temporal and variable axes, normalised attribution mass distributions are derived:

$$P_A(t) = \frac{\sum_{v=1}^{V} \left| A_{t,v}^{(h,v)}(X) \right|}{\sum_{t'=1}^{T} \sum_{v'=1}^{V} \left| A_{t',v'}^{(h,v)}(X) \right|}, \tag{2.4}$$

$$P_A(v) = \frac{\sum_{t=1}^{T} \left| A_{t,v}^{(h,v)}(X) \right|}{\sum_{t'=1}^{T} \sum_{v'=1}^{V} \left| A_{t',v'}^{(h,v)}(X) \right|}. \tag{2.5}$$

For synthetic datasets, the data generator provides oracle information $\mathcal{O}$ (e.g., time

importance, variable importance, dominant frequencies, or lag structure). Oracle-based evaluation measures alignment between $A^{(h,v)}(X)$ (or the induced distributions P_A) and $\mathscr{O}$ using task-appropriate metrics. For real-world datasets, where $\mathscr{O}$ is unavailable, behavioural evaluation measures how model performance and attribution maps change under input perturbations or model randomisation. Details on synthetic data generation and oracle definitions are described in Section 2.4.3.1, and metric definitions and implementation settings are provided in Section 2.4.4.

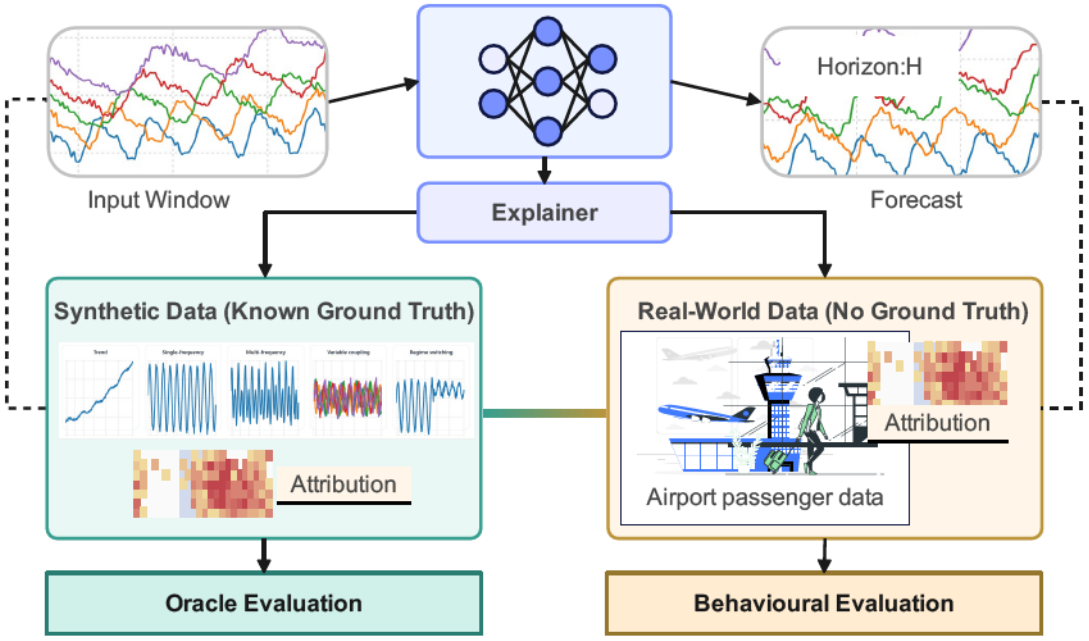


Figure 2.1: Dual evaluation framework for explainability. Synthetic datasets enable oracle-based evaluation against predefined attribution targets, while real-world datasets are assessed using behavioural metrics that probe model sensitivity and attribution robustness.

2.4 Experimental Design

This section describes the forecasting models, explainability methods, datasets, and evaluation metrics used in the empirical study. The experimental design is intended to support a systematic and reproducible comparison of explainability methods for multivariate time series forecasting in airport passenger flow.

2.4.1 Forecasting Models

Explainability methods are evaluated across six forecasting models that represent commonly used modelling paradigms for time series forecasting, including recurrent, convolution based, attention based, state space based, and tree based approaches.

The neural network models include Long Short Term Memory networks and Gated Recurrent Units, which model temporal dependencies through recurrent structures; Temporal Convolutional Networks, which capture temporal patterns using dilated convolutions; Transformer based models that rely on self attention mechanisms; and State Space Networks (SSN), which model temporal dynamics through latent state transitions. In addition, Extreme Gradient Boosting (XGBoost) is included as a representative tree based model commonly used for structured forecasting tasks.

All neural network models are implemented using consistent architectural configurations to improve comparability. Models are trained on standardised input data and evaluated on held out test sets. XGBoost models use a multi output regression formulation for multivariate forecasting. Detailed architectural settings, training procedures, and hyperparameter choices are provided in the appendix.

Unless otherwise stated, neural models use hidden dimension 64, 2 layers, and dropout 0.1. LSTM and GRU use a sequence encoder followed by a linear decoder from the last hidden state to horizon $\times V$. The Transformer uses $d_{\text{model}} = 64$, $n_{\text{head}} = 4$, two encoder layers, and a feedforward dimension of 256, with learnable horizon query embeddings. The State Space Network (SSN) uses $d_{\text{model}} = 64$, $d_{\text{state}} = 16$, and two stacked state-space blocks. The TCN uses three temporal blocks with channels $[64, 64, 64]$, kernel size 3, and dropout 0.2. For neural models, Mean Squared Error (MSE) is optimised with Adaptive Moment Estimation (Adam) (learning rate 0.001) for up to 50 epochs with batch size 32, selecting the best checkpoint by validation loss. These choices are motivated by comparability and standard practice: the learning rate 0.001 is the default for Adam in common deep-learning frameworks and is widely used for time-series and sequence models, balancing convergence speed and stability; 50 epochs provides an upper bound

that is sufficient for validation loss to plateau on our dataset sizes while keeping training tractable for the large number of model–explainer–dataset combinations; and batch size 32 is a standard choice that balances gradient noise and memory use for the input dimensions and model sizes considered here. Model selection by validation loss effectively acts as early stopping, so the best checkpoint may occur before the final epoch.

For Extreme Gradient Boosting (XGBoost), inputs are flattened into lag features of dimension $T \cdot V$ and regressors are trained using squared error loss (default settings: 100 estimators, maximum tree depth 6, learning rate 0.1). By default, a separate regressor is trained for each (h, v) output; an optional multi-output training mode is available to accelerate training.

Table 2.1: Overview of forecasting models and the compatible explainability methods evaluated in this study.

Model Category	Model(s)	Explainers
Neural	LSTM, GRU, TCN, Transformer, SSN	IG, DL, GS, LIME, SHAP
Tree-based	XGBoost	SHAP, LIME

2.4.2 Explainability Methods

Five post-hoc explainability methods are considered: IG, DL, GS, LIME, and SHAP. These methods cover both gradient based and model agnostic approaches and are widely used in explainability research.

Gradient based methods compute attributions by propagating gradients or activation differences through the model, while model agnostic methods rely on input perturbations and surrogate models to estimate feature importance. All methods produce attribution scores for each variable and time step, enabling joint temporal and variable level analysis.

GradientSHAP (GS) differs from SHAP in its computational approach: GradientSHAP is a gradient-based approximation of SHAP values that uses gradients of the model output with respect to inputs, making it applicable only to differentiable models (neural networks). In contrast, SHAP uses different implementations depending on the model type: for tree-

based models like XGBoost, the built-in TreeSHAP algorithm is used (which exploits the tree structure for exact Shapley value computation), while for neural models, the model-agnostic kernel SHAP explainer is used to estimate Shapley values through input perturbations.

For gradient based methods, a zero baseline is used given the standardised input data. For LIME, a time series specific masking strategy is employed to generate perturbed samples. For XGBoost models, the built-in TreeSHAP implementation provided by the XGBoost library is used, which computes exact Shapley values efficiently by leveraging the tree structure.

All explainers produce attribution maps over the input window with shape (T,V) for a scalar prediction target (h,v). While forecasting models output a full prediction matrix $\hat{Y} \in \mathbb{R}^{H \times V}$, computing attributions for all possible targets would increase the computational cost of explanation and evaluation by a factor of $H \cdot V$, making large-scale comparisons across models, explainers, datasets, and random seeds impractical. To ensure tractability and avoid post-hoc target selection, a deterministic explanation target is fixed across all experiments: the first forecast horizon ($h = 0$) and the first variable ($v = 0$). This choice corresponds to the most near-term prediction, which is often of primary operational interest, and ensures consistent, reproducible comparisons without cherry-picking targets that may favor specific explainers.

For IG, 50 integration steps and a zero baseline are used unless otherwise specified. DL also uses a zero baseline unless otherwise specified. For GS, five samples are used; when a baseline is not provided, the per-sample mean over time is used as a single baseline.

For LIME, a time-series-specific masking strategy is implemented with a weighted local linear surrogate (Ridge regression with $\alpha = 0.1$). For each explained sample, 50 perturbed binary masks are generated, sampled i.i.d. (independent and identically distributed) from a Bernoulli(0.5) distribution, and the all-ones mask is included as the unperturbed sample. Distances are computed as the Hamming distance to the all-ones mask and converted to weights by an exponential kernel with width $0.75\sqrt{d}$, where d is the number of interpretable

features. Masked inputs are created by setting masked time–variable regions to zero. Three masking granularities are supported: masking time segments jointly across variables; masking variable–segment pairs (default); and masking individual time–variable pairs. When segment-based masking is used, the segment length is 5 time steps.

For SHAP, different implementations are used depending on the model architecture. For XGBoost, the built-in TreeSHAP algorithm is used (via XGBoost's native contribution output), which computes exact Shapley values by exploiting the tree structure and is computationally efficient. For neural models, the model-agnostic kernel SHAP explainer is used with a background set constructed from the input distribution (mean, zero, and randomly selected samples), using 100 background samples by default.

2.4.3 Datasets for Evaluation

The evaluation includes both synthetic datasets and real-world airport passenger datasets. All datasets are normalised using standard preprocessing procedures and divided into training, validation, and test sets.

2.4.3.1 Synthetic Datasets

Synthetic datasets are designed to represent common temporal patterns in transportation time series, including trends, periodic behaviour, variable coupling, and regime switching. Each synthetic dataset is constructed with known ground truth attribution patterns, enabling oracle-based evaluation. Five synthetic dataset types are generated to cover trend, single-frequency, multi-frequency, variable-coupling, and regime-switching patterns. Table 2.2 summarises each type and its oracle; the generation procedure and oracle definitions are detailed below.

All synthetic datasets produce input–output pairs (X, Y) with $X \in \mathbb{R}^{T \times V}$ and $Y \in \mathbb{R}^{H \times V}$. By default, $T = 50$, $H = 10$, and $V = 5$ are used. The alignment is temporal: X is the input window over time steps $t = 0, \ldots, T-1$ and variables $v = 1, \ldots, V$, and Y is the target window over the immediately following H time steps, indexed by $h = 0, \ldots, H-1$. Thus

$X_{t,v}$ and $Y_{h,v}$ refer to the same variable v at consecutive segments of time; the forecasting task is to predict Y given X. Unless otherwise stated, $n = 1000$ samples per dataset are generated, Gaussian noise with standard deviation 0.1 is used, and the random seed is fixed to 42 for reproducibility. Each synthetic dataset is split 70%/10%/20% into train, validation, and test sets.

Table 2.2: Synthetic datasets used for oracle-based evaluation. Each dataset isolates a specific temporal or multivariate structure and provides corresponding ground-truth relevance used to assess attribution quality.

Dataset type	Targeted structure	Oracle relevance definition
Trend	Variable-specific trends with recency-weighted temporal relevance	Variable importance and temporal importance profile
Single-frequency	Single dominant periodic component	Dominant frequency and relevant variable; uniform temporal relevance
Multi-frequency	Multiple periodic components across variables	Dominant frequencies and relative variable relevance of signal components
Variable coupling	Cross-variable lag dependence (VAR dynamics)	Lag-dependent coupling structure and lag-importance profile
Regime switching	Regime-dependent relevance with structural changes	Regime-specific variable relevance and recency-weighted temporal relevance

Trend dataset. The trend dataset is designed so that different variables exhibit different trend strengths and more recent input time steps are more informative. A non-negative variable-importance vector $w \in \mathbb{R}^V$ is sampled from an exponential distribution and normalised so that $\sum_v w_v = 1$. A recency profile $r \in \mathbb{R}^T$ is defined by $r_t \propto \exp(\lambda_t)$ with λ_t a linear ramp from -2 to 0 over $t = 0, \ldots, T-1$, then normalised so that $\sum_t r_t = 1$. Both w and r are drawn once per dataset and fixed across samples.

The vector w enters the generative model through the trend coefficients. For each

sample, $\beta \in \mathbb{R}^V$ is drawn by scaling a standard Gaussian vector element-wise by w: $\beta_v = \zeta_v w_v$ with $\zeta_v \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Thus w sets the scale of β; variables with larger w_v tend to have larger $|\beta_v|$ and contribute more to both input and forecast. The observations are

$$X_{t,v} = \beta_v \frac{t+1}{T} + \varepsilon_{t,v}, \quad t = 0,\dots,T-1, \qquad Y_{h,v} = \beta_v \frac{T+h+1}{T} + \eta_{h,v}, \quad h = 0,\dots,H-1, \tag{2.6}$$

where $\varepsilon_{t,v}$ and $\eta_{h,v}$ are independent Gaussian noise. So variable importance is present implicitly as $\beta = \zeta \odot w$. The recency profile r is not used to generate X or Y; it is stored in the oracle as the ground-truth temporal importance for evaluation, indicating which input time steps are more informative for predicting Y under this trend model (later steps are more informative for linear extrapolation). The profile r is used only when evaluating whether an explainer's temporal attributions align with this ground truth. The oracle thus contains w (variable importance) and r (time-step importance) as the reference for attribution evaluation.

Single-frequency dataset. The single-frequency dataset is designed to contain a dominant periodic component in one variable. A sinusoid is generated with a fixed dominant frequency f (default $f = 0.1$ cycles per time unit), random phase, and random amplitude. The first variable contains the sinusoid plus noise, while all other variables are independent noise. The oracle identifies the dominant frequency and the responsible variable, and uses a uniform temporal relevance profile.

Multi-frequency dataset. The multi-frequency dataset extends the single-frequency case by placing two sinusoidal components in two variables (default frequencies $f_1 = 0.1$ and $f_2 = 0.2$), each with independently sampled phase and amplitude. Remaining variables are noise-only. The oracle reports the two dominant frequencies and the intended relative importance of the two signal variables (default weights 0.6 and 0.4).

Variable-coupling dataset. The variable-coupling dataset is generated from a vector autoregressive process of order p (default $p = 1$), intended to test recovery of cross-variable dependencies and lag structure. One coupling matrix per lag, $\{A^{(1)},\dots,A^{(p)}\}$, is sampled

with entries drawn from a zero-mean Gaussian distribution and scaled by an exponentially decaying lag-weight profile so that more recent lags are stronger. If necessary, the matrices are rescaled to ensure stability of the autoregressive dynamics.

For each sample, a sequence of length $T + H + p$ is generated using the recursion

$$s_t = \sum_{\ell=1}^{p} A^{(\ell)} s_{t-\ell} + \xi_t, \tag{2.7}$$

with Gaussian noise ξ_t, and then take X and Y as consecutive windows from the generated sequence.

The oracle includes the full set of lag matrices $\{A^{(\ell)}\}$ and a lag-importance profile computed from the magnitudes of the coefficients (normalised per predicted variable across lags).

Regime-switching dataset. The regime-switching dataset is designed to simulate structural changes. A Markov regime sequence of length $T + H$ is sampled with a transition matrix that favours staying in the same regime (default: 0.7 self-transition probability). Each regime has its own variable-importance distribution over V variables, sampled randomly and normalised to sum to one.

Given the regime at time t, the value of variable v is generated from a noisy regime-dependent signal:

$$z_{t,v} = w_v^{(k_t)} (1 + \delta_{t,v}) + \varepsilon_{t,v}, \tag{2.8}$$

where k_t is the regime index, $w^{(k)}$ is the variable-importance vector for regime k, and $\delta_{t,v}$ and $\varepsilon_{t,v}$ are Gaussian noise terms. The input X uses the first T time steps and the forecast Y uses the next H time steps. A recency profile over the input window is additionally defined, analogous to the trend dataset, to support time-focused oracle evaluation. The oracle includes the regime-specific variable-importance vectors and the input-window recency profile.

2.4.3.2 Real-World Datasets

The real-world datasets consist of airport passenger flow data from multiple major United States airports, covering different spatial aggregation levels and multiple years. No ground truth explanations are available for these datasets.

For the main real-world evaluation (Section 2.6), four representative datasets are selected based on diversity in scale and aggregation level.

1. Airport-level multivariate series (LAX and ATL). The dataset comprises 974 daily observations for each of five major U.S. airports (ATL, DEN, DFW, LAX, ORD), covering 15 February 2020 to 15 October 2022 [15]. For the present evaluation, only the LAX and ATL series are used. Each observation is airport-level (one row per airport per day). The 12 variables are: pax (daily airport passenger flow from aggregated TSA security checkpoint throughput); retail, groc, park, transit, work, resident (Google Community Mobility Report metrics for the county containing the airport); srch_cov, srch_flght, srch_airprt (Google Trends search intensity for "COVID", "flight", and "airport"); and cov_death, cov_case (county-level daily COVID-19 deaths and confirmed cases from the Johns Hopkins CSSE COVID-19 Data Repository). Temporal granularity is daily.

2. Top 20 busiest U.S. airports (aggregated multi-airport). A single multivariate series of 2,345 daily observations from 30 December 2018 to 31 May 2025 [34]. Each of the 20 variables is the total passenger count (enplanements or equivalent throughput) for one of the 20 busiest U.S. airports (ATL, BOS, CLT, DEN, DFW, DTW, EWR, FLL, IAH, JFK, LAS, LAX, LGA, MCO, MIA, MSP, ORD, PHX, SEA, SFO). Each column is the total passenger count for one airport. Temporal granularity is daily.

3. SFO boarding-area departures. A single multivariate series of 315 monthly observations from 1 July 1999 to 1 September 2025 for San Francisco International Airport (SFO) [8]. Each of the 8 variables is the monthly departure count for each of the eight boarding areas (A through G plus "Other"). The spatial granularity is boarding-area level. Temporal granularity is monthly.

For real-world datasets, a contiguous split along time is used with adaptive ratios

based on dataset length: if $n_{\text{total}} < 500$, 60%/20%/20% (train/val/test) is used; otherwise, 70%/10%/20% is used. Table 2.3 summarises main statistics for each dataset.

Table 2.3: Summary of real-world datasets used for evaluation: key statistics for each split and temporal window.

Dataset	Vars	Rows	Train	Val	Test	(T,H)	Granularity
LAX	12	974	681	97	196	(76, 20)	Daily
ATL	12	974	681	97	196	(76, 20)	Daily
Top 20 busiest	20	2,345	1,641	235	469	(96, 12)	Daily
SFO boarding	8	315	189	63	63	(49, 13)	Monthly

Notes: Samples are created via a rolling window: each sample uses an input window of length T and predicts the subsequent horizon of length H (i.e., $n_{\text{split}} - T - H + 1$ samples per split). The nominal setting $(T,H) = (96,12)$ requires at least 108 rows per split. When the smallest split has fewer than $T+H$ rows, the data loader reduces the window to fit: the usable length is determined from the smallest split, and 80% of it is allocated to T and 20% to H. In this table, LAX and ATL therefore use $(T,H) = (76,20)$ (the validation split, 97 rows, is limiting); the Top 20 busiest series uses the nominal $(96,12)$; and the SFO boarding-area monthly series uses $(49,13)$ under its shorter split lengths.

2.4.4 Evaluation Metrics and Baseline Strategy

A dual evaluation framework aligned with Section 2.3.3 is adopted, distinguishing between oracle-based metrics on synthetic datasets and behavioural metrics on real-world datasets. The former enables direct comparison with known ground-truth attribution structures, while the latter evaluates consistency with model sensitivity and robustness in the absence of oracle supervision. In both settings, random attribution baselines are incorporated to distinguish meaningful explanations from spurious or structure-agnostic patterns. All default configurations used in the main experiments are explicitly stated.

2.4.4.1 Oracle-based Metrics (Synthetic Datasets)

Let $A^{(h,v)}(X) \in \mathbb{R}^{T\times V}$ denote the attribution map. Marginal importance distributions over time and variables are constructed based on absolute attribution mass:

$$P_A(t) = \frac{\sum_{v=1}^{V} |A_{t,v}|}{\sum_{t',v'} |A_{t',v'}|}, \qquad P_A(v) = \frac{\sum_{t=1}^{T} |A_{t,v}|}{\sum_{t',v'} |A_{t',v'}|}. \tag{2.9}$$

When the oracle specifies a temporal importance profile $P_{\mathcal{O}}(t)$, temporal alignment is quantified using the Kullback–Leibler divergence:

$$D_{\mathrm{KL}}(P_A(t) \,\|\, P_{\mathcal{O}}(t)) = \sum_{t=1}^{T} P_A(t) \log \frac{P_A(t)}{P_{\mathcal{O}}(t)}, \tag{2.10}$$

together with the entropy of the attribution distribution,

$$H(P_A(t)) = -\sum_{t=1}^{T} P_A(t) \log P_A(t), \tag{2.11}$$

which captures the concentration or dispersion of temporal importance.

When the oracle provides a variable-importance vector $P_{\mathcal{O}}(v)$, (i) the Spearman rank correlation between $P_A(v)$ and $P_{\mathcal{O}}(v)$ is reported, and (ii) top-k precision is reported, defined as the fraction of oracle top-k variables recovered by the attribution ranking. Throughout all experiments, $k = 3$.

For frequency-based synthetic datasets, where the oracle specifies a dominant frequency f^*, attribution mass is aggregated over variables, the discrete Fourier transform of $\sum_v |A_{t,v}|$ is computed, and a frequency saliency score is defined as the proportion of total spectral power concentrated at the frequency bin closest to f^*.

For variable-coupling datasets, the oracle provides a joint lag-variable importance structure. Temporal indices are mapped to lag space (most recent timestep corresponding to lag 0) to obtain an attribution-derived lag-variable representation. Evaluation is performed using (i) Pearson correlation between flattened attribution and oracle arrays, and (ii) top-k recovery of lag-variable pairs, with $k = 3$.

2.4.4.2 Behavioural Metrics (Real-World Datasets)

In the absence of ground-truth attributions, explainers are evaluated using three complementary behavioural probes assessing faithfulness, model dependence, and robustness.

Deletion test. The goal is to measure whether inputs flagged as important by the attribution are also important for the trained model's prediction under a simple perturbation protocol. For each proportion $p \in 0.1, 0.2, \ldots, 1.0$, the attribution map is flattened to a vector over all (t, v) positions in the input window, entries are ranked by $|A^{(h,v)}(X)|$, and the top p fraction of positions is selected (by absolute attribution magnitude). $X^{(p)}$ is formed by zeroing those selected entries and leaving all other time–variable values unchanged, which is equivalent to applying a multiplicative mask with zeros at removed positions and ones elsewhere. This evaluation masking is unrelated to the reference baseline inputs used internally by gradient-based explainers (Integrated Gradients, DeepLift, etc.) to define path integrals; here only the forward pass input to f_θ is altered.

Let $\mathscr{L}$ denote the mean squared error. The curve is defined as

$$\mathrm{Perf}(p) = 1 - \frac{\mathscr{L}(f_\theta(X^{(p)}), Y)}{\mathscr{L}(f_\theta(X), Y) + \varepsilon}, \tag{2.12}$$

where ε ensures numerical stability. When removing attributed-important mass increases error, $\mathrm{Perf}(p)$ becomes more negative. The *deletion AUC* is the area under the $\mathrm{Perf}(p)$ curve computed by trapezoidal integration over the fixed grid $\{0.1, \ldots, 1.0\}$; more negative values indicate stronger cumulative performance degradation under progressive deletion, which is consistent with higher behavioural faithfulness of the map to the model, subject to redundancy among correlated inputs (see Section 2.6.1).

Model randomisation test. For neural models, parameters are fully randomised by sampling each weight from $\mathscr{N}(0, 0.01)$, and attributions are recomputed. Similarity between original and randomised attributions is measured via cosine similarity after flattening and ℓ_2 normalisation. Results are averaged over $N = 5$ independent randomisations. This test is not applicable to non-neural models, which are excluded from this evaluation.

Noise stability test. Robustness under small input perturbations is assessed by sampling $X^{(m)} = X + \eta^{(m)}$, where $\eta^{(m)} \sim \mathcal{N}(0, \sigma^2)$. The stability score is the average ℓ_2 distance between original and perturbed attributions:

$$\frac{1}{M} \sum_{m=1}^{M} \|A^{(h,v)}(X) - A^{(h,v)}(X^{(m)})\|_2. \tag{2.13}$$

In all experiments, $\sigma = 0.01$ and $M = 10$, with lower values indicating greater robustness.

2.4.4.3 Baseline Definitions and Evaluation Setup

This section specifies the baseline configurations used in evaluation and explanation, and clarifies the distinction between two conceptually different baselines. The random baseline serves as an external comparator for benchmarking attribution quality, whereas the attribution baseline is an internal reference input used by certain explainers.

Random baseline (evaluation comparator). To contextualise both oracle-based and behavioural scores, the same metrics are computed using random attribution maps $R \in \mathbb{R}^{T \times V}$. In the main experiments, R is generated by i.i.d. sampling from Uniform($[-1, 1]$), with a fixed global seed and a deterministic sample-level offset to ensure reproducibility. This same default is used for both synthetic and real-world evaluations. Alternative variants (Gaussian $\mathcal{N}(0, 1)$ and shuffled-uniform) are implemented but used only for supplementary analysis.

Attribution baseline selection. The attribution baseline is an explainer-internal reference input used only by gradient-based methods to compute attribution differences; it is distinct from the random baseline, which is a separate comparator used for metric-level benchmarking. For gradient-based explainers (Integrated Gradients, DeepLift, GradientSHAP), a zero reference baseline (same shape as the input, all entries set to zero) is used in all main experiments because inputs are standardised and this choice provides a neutral and consistent reference across datasets and models. Two alternative attribution baselines are implemented for robustness analysis: (i) a mean baseline defined as the training-data

average input window, and (ii) a representative baseline corresponding to a domain-specific reference instance. These alternatives are not used in primary results.

Perturbation-based methods (LIME and SHAP) do not require such baselines. LIME relies on local perturbations, while SHAP operates on background distributions; any provided baseline is ignored for LIME and only used in SHAP when explicitly specified as background data.

Summary of default configuration. The main experiments consistently use: top-$k = 3$, $N = 5$ randomisations, $M = 10$ perturbations, $\sigma = 0.01$, uniform random baseline maps for benchmark comparison, and a zero attribution baseline (reference input) for gradient-based explainers. All alternative configurations are implemented but restricted to supplementary or robustness analyses.

2.5 Results on Synthetic Data

This section presents the evaluation results on synthetic datasets with known ground truth attribution patterns. The purpose of these experiments is to assess the correctness of explainability methods under controlled conditions and to examine how their performance varies across forecasting models and temporal patterns.

Unless otherwise stated in this section, synthetic score visualisations use a unified normalization and aggregation protocol: for each dataset type, each oracle metric is min–max normalized to $[0,1]$ over all valid runs in that dataset slice, lower-better metrics are inverted, and explainer scores are computed as weighted averages of these normalized metrics. This protocol is used consistently for Figure 2.2 and Figure 2.3.

2.5.1 Overall Performance Across Temporal Patterns

The evaluation results show notable differences in explainability performance across temporal patterns. No single explainability method consistently performs best across all synthetic datasets. Instead, performance is strongly influenced by the characteristics of the

underlying temporal structure. Unless otherwise stated, the primary dataset-specific oracle metric is: Trend = variable-importance Spearman correlation, Frequency (single/multi) = frequency saliency, Variable-coupling = lag–variable correlation, and Regime-switching = temporal entropy (lower is better).

Figure 2.2 visualises the regenerated oracle-score heatmap (all seeds) under the unified normalization protocol defined at the start of this section.

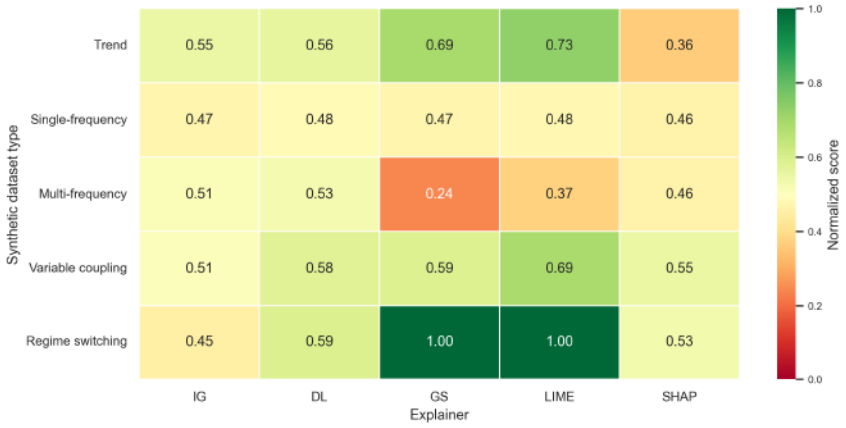


Figure 2.2: Oracle-score heatmap on synthetic datasets using the regenerated analysis output (all seeds), following the unified normalization protocol defined in Section 2.5.

Under this normalized composite scoring, the leading explainer varies by temporal pattern: LIME performs best on trend (0.733) and var-coupling (0.690), DeepLift performs best on frequency-single (0.483) and frequency-multi (0.529), and regime-switching yields a tie between GradientSHAP and LIME (both 1.000). These aggregate rankings show that explainability effectiveness is pattern-dependent and that no single method dominates all synthetic structures.

The dataset-specific primary oracle metrics partly reinforce this conclusion while also showing where the composite view changes the ranking. On trend datasets, DeepLift achieves the strongest mean Spearman correlation (0.306), followed by IG (0.278), whereas SHAP performs poorly overall with negative mean Spearman (-0.016), indicating weak alignment with the ground-truth variable relevance. On frequency datasets, SHAP is strongest on frequency-single (0.017) and DeepLift is strongest on frequency-multi (0.019), whereas GradientSHAP consistently fails to recover the dominant frequency pattern and

yields zero saliency scores across all 30 runs. For variable-coupling datasets, LIME is clearly strongest (mean lag–variable correlation 0.229), while GradientSHAP again fails to recover the target structure and produces zero lag–variable correlation across all 15 runs. For regime-switching datasets, both LIME and GradientSHAP attain the minimum mean entropy (0.000), but such near-zero entropy should be interpreted cautiously because it may reflect trivially concentrated attribution mass rather than meaningful alignment with regime boundaries.

To guard against trivial alignment, each explainer is additionally compared against a random attribution baseline evaluated with the same oracle metrics. Using i.i.d. uniform random attributions on $[-1, 1]$, the primary-metric baseline levels are near chance (mean baseline across runs: Trend Spearman ≈ -0.023, Frequency saliency ≈ 0.001, Var Coupling lag–variable correlation ≈ -0.001, Regime Switching temporal entropy ≈ 3.877). Most explainers outperform these baselines on their corresponding primary metric, but failures do occur for some method–model pairs. Notably, GradientSHAP on frequency datasets produces zero saliency scores across all models and seeds (0/30 runs better than random), and SHAP on trend shows negative mean correlations (mean Spearman $= -0.016$) with 9 out of 18 model–seed combinations performing worse than random. These comparisons help identify explainers that can produce uninformative attributions despite nontrivial-looking heatmaps.

Table 2.4 summarises synthetic performance from the updated aggregated outputs, with baseline-relative statistics. This table reports each explainer's performance relative to a random attribution baseline; it is a validity check against randomness rather than a direct explainer-to-explainer ranking.

Table 2.4: Explainer performance relative to a random baseline. Success rate is the percentage of runs that beat random baseline. Average improvement is a coarse cross-dataset summary over primary metrics and should be interpreted cautiously because metric scales and directions differ across datasets (e.g., entropy is lower-better for regime switching). Best/worst datasets are based on success rate and improvement magnitude.

Explainer	Success Rate	Avg Improvement	Best / Worst Dataset
Integrated Gradients	86.7%	0.463	Regime Switching / Var Coupling
DeepLift	88.3%	0.581	Regime Switching / Var Coupling
GradientSHAP	46.7%	0.780	Regime Switching / Frequency Single
LIME	52.2%	0.863	Regime Switching / Frequency Single
SHAP	83.3%	0.471	Regime Switching / Trend

2.5.2 Model and Explainer Interactions

The interaction between forecasting model architecture and explainability method significantly influences attribution quality. Different models exhibit distinct patterns of explainability performance, and the optimal explainer choice varies substantially across model types. Figure 2.3 presents faceted model–explainer heatmaps of dataset-specific composite explainability scores. N/A marks an inapplicable pair, not missing data: XGBoost is not differentiable with respect to inputs, so IG, DeepLift, and GradientSHAP are omitted and only LIME and SHAP are evaluated.

Several interaction patterns are evident. On trend datasets, the composite scores in Figure 2.3 favour LIME much more strongly than the primary-metric view alone would suggest: LIME is the top explainer for every compatible model, with the highest score obtained by SSN–LIME (0.750) and similarly strong values for GRU, TCN, and XGBoost (all about 0.731–0.732). Integrated Gradients remains competitive for selected architectures such as SSN (0.738).

On frequency datasets, the preferred explainer depends more strongly on the forecasting model. Transformer–DeepLift is the strongest combination in both panels (0.572 for Frequency Single and 0.628 for Frequency Multi), whereas LIME is preferred for

most recurrent and convolutional models on Frequency Single and IG becomes highly competitive on SSN and Transformer for Frequency Multi. SHAP remains relevant for XGBoost and some recurrent settings, but it is not the dominant explainer across the frequency panels in the composite interaction view.

In contrast, variable-coupling datasets show a much more stable pattern, with LIME consistently ranking first for every compatible model. The strongest cells lie between 0.671 and 0.711, with LSTM–LIME, TCN–LIME, and Transformer–LIME all performing well, indicating that sparse lag-structure recovery is comparatively robust to the underlying forecasting architecture under this composite score.

The regime-switching setting further illustrates that strong heatmap values do not always imply equally informative explanations. In Figure 2.3, both LIME and GradientSHAP attain the ceiling score for many model combinations, and all available XGBoost cells for this dataset are also maximal under LIME. As discussed in the preceding subsection, however, near-zero entropy may arise either from meaningful focus on regime boundaries or from trivially concentrated attribution mass.

Overall, the results show that explainer selection should be conditioned jointly on forecasting architecture and dataset structure. No single method dominates across all model classes, and conclusions based on model-averaged rankings alone would obscure the architecture-specific interaction patterns visible in Figure 2.3.

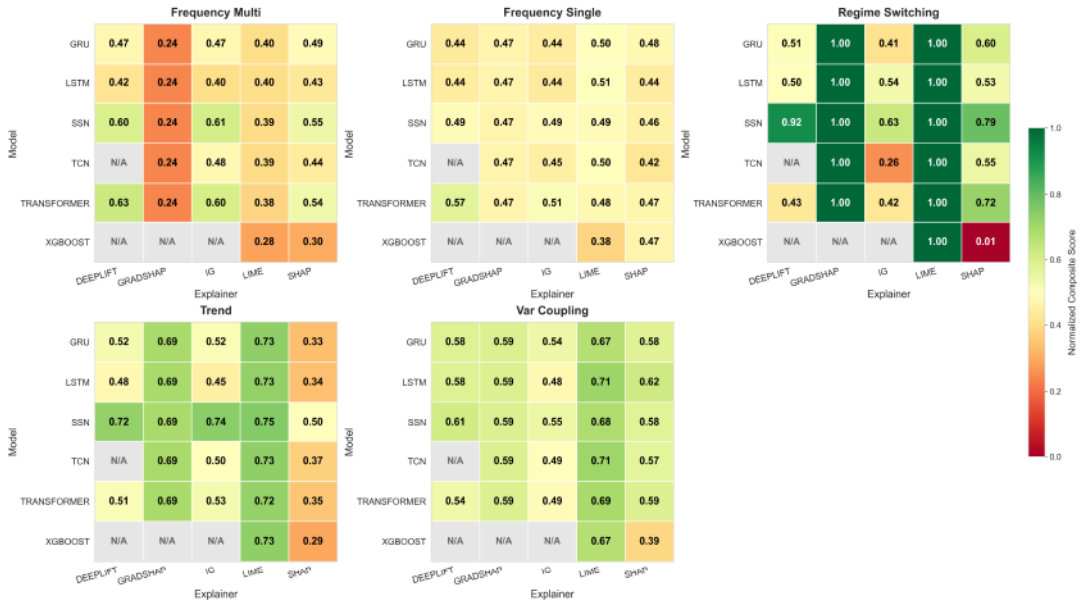


Figure 2.3: Model–explainer interaction heatmap on synthetic datasets, showing composite explainability scores across forecasting models and dataset types under the unified normalization protocol defined in Section 2.5.

2.5.3 Relationship Between Prediction Accuracy and Explainability Quality

Whether forecasting accuracy is associated with explainability quality is examined by correlating prediction error (test MSE) with dataset-specific oracle metrics. Two complementary views are used. **(i) Aggregate across combinations:** one point per model–explainer pair, using mean test MSE and mean oracle metric on seed 42 only. Different seeds correspond to different synthetic realizations; holding seed fixed avoids confounding cross-dataset patterns with changes in data generation. **(ii) Within fixed combinations:** per-test-sample points pooled across seeds 42–44 (100 test samples per run, three seeds), for three representative pairs (one per dataset type: SSN–DeepLift on trend, Transformer–DeepLift on frequency-single, TCN–LIME on variable-coupling) to illustrate behaviour at finer granularity. Table 2.5 reports the same data seed-by-seed.

At the aggregate level (seed 42), estimated linear associations between test MSE and the oracle metric are small in magnitude. None reach statistical significance at $\alpha = 0.05$, although the Spearman trend panel is borderline ($\rho = 0.352$, $p = 0.078$). These

results are nevertheless consistent with no clear linear relationship in these panels: trend (variable-importance Spearman; Pearson $r = 0.287$, $p = 0.155$; Spearman $\rho = 0.352$, $p = 0.078$; $n = 26$); frequency-single (frequency saliency; $r = 0.076$, $p = 0.714$; $\rho = 0.150$, $p = 0.466$; $n = 26$); variable-coupling (lag–variable correlation; $r = 0.169$, $p = 0.409$; $\rho = 0.142$, $p = 0.490$; $n = 26$).

The pooled per-sample analysis uses a larger effective n and a different unit of analysis than the aggregate panels, so correlation magnitudes and significance need not match the combination-level scatterplots. For SSN–DeepLift on trend, small but statistically significant linear associations are obtained (Pearson $r = 0.144$, $p = 0.013$; Spearman $\rho = 0.148$, $p = 0.010$; $n = 300$). For Transformer–DeepLift on frequency-single, estimates remain small and non-significant ($r = 0.063$, $p = 0.278$; $\rho = 0.043$, $p = 0.457$; $n = 300$). For TCN–LIME on variable-coupling, Pearson $r = 0.131$ ($p = 0.023$) is significant while Spearman $\rho = 0.112$ is borderline ($p = 0.052$; $n = 300$).

Figure 2.4 visualises the aggregate (combination-level) analysis for seed 42: each panel plots one point per model–explainer pair (dataset-specific oracle metric versus mean test MSE), coloured by model type, with $n = 26$ per panel.

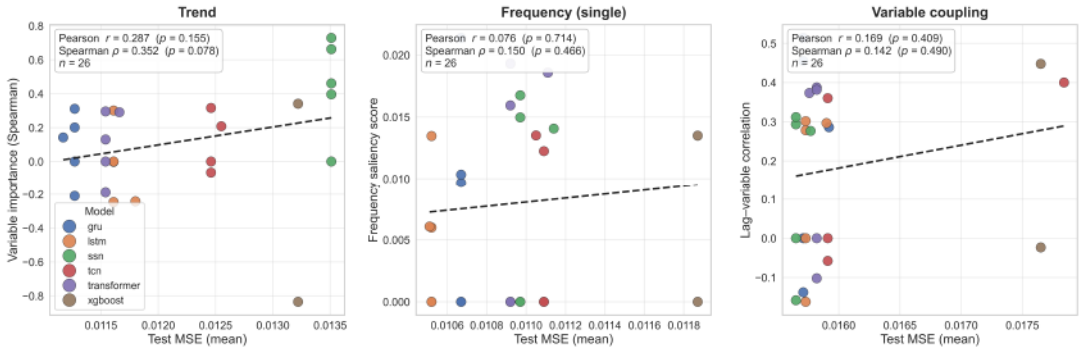


Figure 2.4: Aggregate relationship between prediction error (mean test MSE) and dataset-specific oracle metrics for seed 42. Each point is one model–explainer combination; each panel (trend, frequency-single, var-coupling) has $n = 26$ points.

Table 2.5: Seed-wise correlations between prediction accuracy (test MSE) and explainability metrics for three representative model–explainer combinations (seeds 42–44).

Comb.	Data	Seed	n	Pearson r (p)	Spearman ρ (p)
SSN–DeepLift	Trend	42	100	0.139 (0.168)	0.137 (0.174)
		43	100	0.160 (0.112)	0.171 (0.088)
		44	100	–0.011 (0.912)	0.011 (0.917)
Transformer–DeepLift	Freq. Single	42	100	0.024 (0.812)	–0.007 (0.945)
		43	100	0.109 (0.279)	0.062 (0.541)
		44	100	0.037 (0.714)	–0.013 (0.902)
TCN–LIME	Var Coupling	42	100	0.006 (0.950)	–0.007 (0.945)
		43	100	–0.090 (0.374)	–0.194 (0.053)
		44	100	–0.067 (0.508)	–0.051 (0.617)

Overall, test MSE and oracle explainability scores do not show a strong or robust linear coupling in the aggregate panels ($n = 26$), and the combination-level evidence is consistent with treating prediction quality and explanation quality as largely distinct dimensions. Pooled per-sample analyses can detect very small linear trends with larger n, but effect sizes remain modest and seed-wise diagnostics (Table 2.5) show that associations are not stable across seeds for these illustrative pairs. High prediction accuracy therefore does not guarantee reliable explanations, and models with higher error may still yield attributions that align well with ground-truth structure under the chosen metrics.

These findings emphasise the importance of evaluating explainability independently of prediction performance and suggest that explainability should be treated as a distinct model property rather than a byproduct of predictive accuracy.

2.5.4 Robustness Analysis Across Random Seeds

Each model–explainer pair is compared across seeds 42–44 in the multi-seed synthetic aggregation. Figure 2.5 has one panel per dataset type (primary oracle metric per panel): models on the horizontal axis, explainers colour-coded, each box summarising three seed-level values. With only three seeds per box, the figure should be interpreted as a descriptive

stability check rather than as a distributional estimate.

Read robustness together with metric meaning. On trend, frequency, and variable-coupling, values near zero usually mean failure to recover oracle structure, so stable near-zero boxes indicate stable underperformance, not good robustness. On regime-switching, temporal entropy falls when attribution concentrates; entropy near zero can reflect meaningful alignment with regimes or trivial mass collapse. In those cases, interpretation should also use cross-seed dispersion.

Trend. Strong pairs coexist with unstable gradient-based behaviour: cross-seed standard deviations range from 0.000 (GradientSHAP, identically zero here) to 0.737 (LSTM–Integrated Gradients). SSN–DeepLift combines high mean Spearman (0.774) with low dispersion (std 0.068); SSN–Integrated Gradients is still strong (mean 0.729) but less stable (std 0.135). LSTM–Integrated Gradients and Transformer–DeepLift vary strongly across seeds, so high trend attribution is not uniformly robust within gradient families.

Frequency. Top performers fluctuate while some failures are stably poor. Frequency-single: Transformer–DeepLift has the highest mean saliency (0.028) but large std (0.020); SHAP–GRU is lower on average (0.021) but tightly clustered (std 0.001). Frequency-multi: Transformer–DeepLift again leads (mean 0.032, std 0.013); SHAP and SSN combinations are lower but steadier. GradientSHAP is zero on both frequency types for all models and seeds, which reflects systematic failure rather than robust success.

Variable-coupling. LIME stays competitive (mean lag–variable correlation about 0.20–0.26; std roughly 0.22–0.37) but remains seed-sensitive. Transformer–DeepLift is the starkest unstable case (mean 0.139, std 0.509), so the correlation sign can flip across seeds: strong average recovery need not imply stable recovery.

Regime-switching. LIME shows entropy 0.000 for every model and seed; GradientSHAP does so for many pairs. Those zeros denote very narrow temporal mass, not automatically superior explanations. DeepLift–SSN contrasts as the lowest non-zero mean entropy among gradient examples (0.292, std 0.442); Integrated Gradients tends toward higher entropy and more seed spread.

In sum, robustness is dataset- and method-specific. Trend: SSN–DeepLift is the clearest stable high performer; LSTM–Integrated Gradients stays seed-sensitive. Frequency: best average (Transformer–DeepLift) tracks high variance; SHAP–GRU trades mean for stability. Variable-coupling: LIME is the strongest family on average, with non-negligible dispersion. Regime-switching: LIME and GradientSHAP stably yield near-zero entropy, indicating robust temporal concentration but not guaranteed explanation quality. Prefer pairs with both strong mean oracle scores and limited cross-seed spread, rather than high means or apparent stability alone.

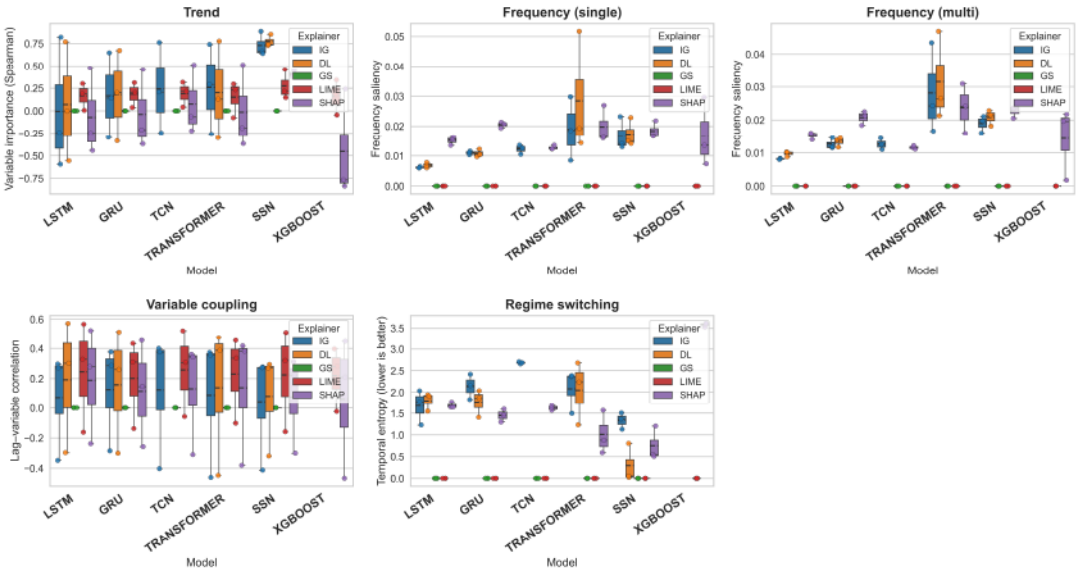


Figure 2.5: Cross-seed variability of synthetic oracle metrics (seeds 42–44). Panels correspond to the five synthetic dataset types and use each dataset's primary oracle metric. Within each panel, models are on the x-axis, explainers are colour-coded, and each box summarises three seed values for one model–explainer pair (with jittered seed points).

2.5.5 Batch Experiment Failure Analysis

Before summarising the synthetic-data findings, failure cases observed during the full synthetic batch run are documented (configuration: synthetic, all models, all explainers, seed=42). The batch comprised 150 planned model/explainer/dataset combinations. Of these, 130 succeeded (86.7%) and 20 failed (13.3%). All failures correspond to known incompatibilities or library limitations and were excluded from downstream aggregation; the following subsection (Summary of Synthetic Data Findings) applies to the 130 successful

combinations.

Failure Type A: TCN + DeepLift (5 failures). All five failures occurred for the TCN model when using DeepLift, across all five synthetic dataset types. The root cause is a DeepLift limitation: Captum's DeepLift requires per-module input/output attributes tracked via hooks, and can fail when a module (e.g., `nn.ReLU`) is used multiple times in a single forward pass. In the TCN implementation, ReLU modules within temporal blocks are invoked multiple times (including the residual branch), triggering the error.

Failure Type B: XGBoost + gradient-based explainers (15 failures). Fifteen failures occurred for XGBoost when attempting to use gradient-based explainers (IG, DeepLift, GradientSHAP) across all five synthetic dataset types. This incompatibility is fundamental: tree-based models are not differentiable with respect to inputs, so gradient-based attribution methods are not applicable. The codebase explicitly checks this condition and raises an error before running attribution.

2.5.6 Summary of Synthetic Data Findings

The evaluation on synthetic datasets with known ground truth attribution patterns yields the following synthesis; it applies to the **130 successful** model–explainer–dataset combinations. **20** planned combinations failed and were excluded from aggregation (Section 2.5.5).

Overall performance and primary metrics (Figure 2.2 and Section 2.5.2). Explainability is strongly shaped by temporal pattern, forecasting architecture, and explainer choice; no single method wins every scenario. When averaged across models and seeds using the **primary** oracle metric, DeepLift leads on trend (Spearman), SHAP on frequency-single, DeepLift on frequency-multi, and LIME on variable-coupling. For regime-switching, GradientSHAP and LIME attain mean temporal entropy $= 0.0$, which signals strong temporal concentration but may also reflect trivially collapsed mass, so careful interpretation is needed. Separately, under the **unified normalized composite** scores visualised in Figure 2.2 and Figure 2.3, rankings need not coincide with primary-metric means: for example, composite cells on trend favour LIME strongly for many compatible models (Sec-

tion 2.5.2), whereas primary-metric means emphasise DeepLift and Integrated Gradients. Reporting both views avoids overstating a single "best" explainer.

Model–explainer interactions (Figure 2.3). Cell-level heatmaps make architecture conditioning explicit: Transformer–DeepLift is strongest on both frequency panels under the composite score, whereas variable-coupling shows LIME ranked first for every compatible model. The regime-switching panel again shows ceiling scores for LIME and GradientSHAP for many cells; as above, read those values alongside metric semantics rather than as automatic superiority.

Robustness across seeds (Figure 2.5). Cross-seed stability is dataset- and pair-dependent, not a property of one explainer family uniformly. Illustratively, SSN–DeepLift combines high trend Spearman with relatively tight dispersion, while LSTM–Integrated Gradients and Transformer–DeepLift can vary sharply across seeds; on frequency, strong mean performance can co-occur with large seed-level spread (e.g., Transformer–DeepLift), and GradientSHAP remains identically zero on frequency throughout. LIME stays competitive on variable-coupling but exhibits non-negligible dispersion across seeds. On regime-switching, repeated near-zero entropy for LIME and GradientSHAP indicates **stable temporal concentration**, which is not the same as guaranteed explanation quality. Multi-seed evaluation is therefore essential, and preferred combinations should jointly show strong mean oracle alignment and limited cross-seed spread where applicable.

Prediction error versus explainability (Section 2.5.3; Figure 2.4). Combination-level aggregates for a fixed seed (42) show weak linear associations between test MSE and the primary oracle metrics in the three panels reported ($n = 26$ each); none reach significance at $\alpha = 0.05$, with the Spearman trend panel borderline ($p = 0.078$). Pooled per-sample analyses use a different unit of analysis and larger effective n, so correlation magnitudes and significance need not match the aggregate scatterplots; they can reveal small significant trends with modest effect sizes. Seed-wise diagnostics for representative pairs (Table 2.5) further show that associations are not stable across seeds. High test MSE therefore does not preclude good oracle scores, and low MSE does not guarantee faithful

explanations under the chosen metrics.

Comparison to random attribution baselines (Table 2.4). Relative to i.i.d. uniform random attributions evaluated with the same oracle metrics, most explainers usually outperform the baseline on the corresponding primary metric, but systematic failures appear for specific pairs: GradientSHAP collapses to zero saliency on frequency and variable-coupling (never beating random on those primary metrics in our runs), and SHAP underperforms on trend with frequent worse-than-random outcomes. These checks surface methods that can yield uninformative maps despite visually nontrivial heatmaps.

In short, explainability should be assessed with both primary and composite oracle views where relevant, across multiple seeds, and without equating regime entropy near zero with explanation quality without additional scrutiny; prediction accuracy remains an imperfect proxy for explanation correctness in this setting.

2.6 Results on Real World Airport Data

This section presents the evaluation results on real world airport passenger flow datasets. Unlike synthetic data, these datasets do not provide ground truth explanations, so explainability methods are assessed using behavioural evaluation metrics. The goal of this analysis is to examine how explainability methods behave in practical transportation forecasting settings and to identify consistent patterns and limitations.

Experimental configuration. Based on the findings from synthetic data evaluation (Section 2.5), the top-performing explainability methods are selected for real-world evaluation: DeepLift, LIME, and SHAP. These three explainers reflect the updated synthetic aggregation (LIME and SHAP are strong on several dataset types, while DeepLift remains competitive on selected model–dataset pairs) and match the focused real-world benchmark configuration. All six forecasting models (LSTM, GRU, TCN, Transformer, SSN, and XGBoost) are evaluated on four representative real-world datasets: LAX (12 variables), ATL (12 variables), top 20 busiest airports (20 variables), and SFO boarding-area departures (8

variables). The focused benchmark records 64 successful model–explainer–dataset runs in the real-world aggregation; the four omitted combinations are TCN with DeepLift, which fails under the same Captum limitation documented for synthetic runs (Section 2.5.5). The selected datasets span different scales (8, 12, and 20 variables) and aggregation levels (airport-level and boarding-area-level).

2.6.1 Deletion-based Evaluation in Airport Passenger Flow

In airport passenger flow forecasting, input variables and time steps are often highly correlated due to shared seasonal patterns, network effects, and aggregation across multiple passenger flow streams. As a result, deletion-based evaluation in this setting does not solely reflect explanation correctness, but also captures structural redundancy inherent in transportation passenger flow time series.

Across all evaluated real-world airport datasets, deletion AUC values exhibit substantial variation across model–explainer combinations (see Table 2.6). Aggregating across the four representative datasets, DeepLift tends to yield more negative deletion AUC on neural models (e.g., GRU: -2.723 ± 4.071, LSTM: -2.200 ± 3.648), whereas LIME and SHAP typically produce smaller performance degradation (e.g., SSN–LIME: -0.111 ± 0.127; SSN–SHAP: -0.189 ± 0.160). For XGBoost, SHAP exhibits substantially larger deletion impact (-2.258 ± 4.102) than LIME (-0.471 ± 1.081). The large standard deviations indicate strong dataset dependence, consistent with redundancy and scale differences across airport passenger flow series.

Importantly, smaller performance degradation under deletion does not necessarily indicate weaker explanations in this context. In airport passenger flow data, multiple time steps and variables may convey overlapping information (e.g., adjacent temporal windows during peak seasons or correlated flows across terminals). Consequently, deletion AUC reflects a combination of attribution selectivity and model redundancy, rather than explanation quality alone. Therefore, deletion-based metrics on real-world airport data should be interpreted comparatively rather than absolutely. In practice, deletion AUC is

most informative when used to compare explainers within the same forecasting model and dataset, rather than as a standalone indicator of explanation reliability.

Table 2.6: Deletion AUC on real-world airport data (more negative = larger degradation under deletion). Values are mean and standard deviation across the four representative datasets (ATL, LAX, Top 20 busiest airports, and SFO boarding-area departures), computed from per-run model–explainer–dataset deletion AUC means.

Model–Explainer	Mean Deletion AUC	Std Dev
GRU–DeepLift	**-2.723**	4.071
GRU–LIME	-0.523	0.904
GRU–SHAP	-0.532	0.903
LSTM–DeepLift	-2.200	3.648
LSTM–LIME	-0.242	0.489
LSTM–SHAP	-0.191	0.393
SSN–DeepLift	-1.651	1.788
SSN–LIME	-0.111	0.127
SSN–SHAP	-0.189	0.160
TCN–LIME	-0.277	0.442
TCN–SHAP	-0.366	0.460
Transformer–DeepLift	-1.795	2.782
Transformer–LIME	-0.607	1.268
Transformer–SHAP	-0.640	1.077
XGBoost–LIME	-0.471	1.081
XGBoost–SHAP	-2.258	4.102

2.6.2 Model Randomisation and Attribution Stability Diagnostics

Model randomisation and input-noise stability are best interpreted here as diagnostic rather than ranking metrics. Figure 2.6 summarises both in a comparative format: the middle panel reports mean model-randomisation similarity, and the right panel reports median log-scale instability under input perturbations.

Model randomisation provides limited discrimination on real-world airport data. The middle panel shows that mean attribution similarity between original and randomised models remains close to zero for most neural model–explainer pairs, with only modest

positive values for a few GRU and Transformer cases. This weak separation suggests that randomisation adds little comparative insight in this domain. A plausible reason is that transportation time series are strongly autocorrelated and smooth, so even randomised models can respond similarly to structured input patterns. In addition, some neural-model LIME cells are shown as N/A because randomised-model attributions can collapse to degenerate (near-zero) vectors, making cosine similarity numerically undefined and producing NaN aggregates. Model randomisation is therefore treated here as a basic sanity check rather than as a primary ranking criterion.

Input-noise stability is more informative as a numerical reliability diagnostic. In the right panel of Figure 2.6, DeepLift tends to yield the smallest median log-scale instability, whereas SHAP remains unstable for several architectures even after aggregation across datasets. LIME often falls near zero under the current implementation, which should be interpreted cautiously rather than read as automatic superiority. Because individual SHAP runs can reach raw instability on the order of 10^{14}, the figure reports the **median** of $\log_{10}(1 + \text{noise stability})$ across the four datasets for each model–explainer pair. This aggregation reduces the influence of single extreme runs while preserving the broader contrast between numerically stable and unstable settings.

In several configurations, extremely large instability values are observed, indicating numerical sensitivity rather than meaningful changes in attribution structure. In practice, such instability should be interpreted as a warning about numerical reliability, rather than as evidence of stronger explanatory power.

For transportation forecasting studies, unstable attributions can hinder longitudinal analysis and policy interpretation, even when deletion-based metrics appear favourable. Stability analysis therefore serves primarily as a diagnostic tool to identify explainability configurations that require cautious interpretation.

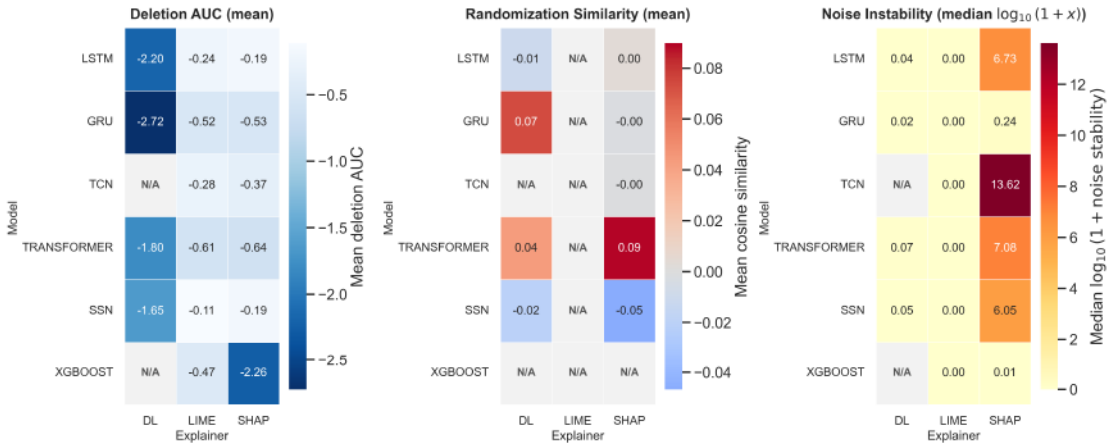


Figure 2.6: Overview of behavioural explainability metrics on real-world airport data, aggregated across the four representative datasets. Rows correspond to forecasting models and columns to explainers. Left: mean deletion AUC (more negative = larger performance degradation under deletion). Middle: mean model-randomisation similarity (cosine similarity between original and randomised-model attributions; values near zero indicate stronger sensitivity to learnt parameters). Right: median $\log_{10}(1+\text{noise stability})$ under input perturbations. N/A indicates an inapplicable or undefined case: TCN–DeepLift is omitted for the same Captum limitation as in the synthetic benchmark; XGBoost randomisation is not defined in this pipeline; and some neural-model LIME randomisation entries are undefined due to degenerate randomised-model attribution vectors (NaN cosine similarity).

2.6.3 Relationship between Prediction Accuracy and Explainability Quality on Real-world Data

The relationship between forecasting accuracy and explainability performance is further examined using the three behavioural metrics defined for real-world data: deletion AUC, model-randomisation similarity, and log-transformed noise stability. Here, prediction accuracy is quantified by test MSE, rather than by correlation between predicted and observed time-series values. To characterise run-level association between test MSE and each behavioural metric across successful model–explainer–dataset runs, both Pearson correlation (linear association) and Spearman rank correlation (monotonic association) are reported. These statistics are used descriptively only: they do not assess forecast accuracy on their own, do not imply causality, and, in the absence of ground-truth explanations, do not establish explanation correctness. Correlation analysis nevertheless helps indicate whether more accurate forecasting models also tend to exhibit systematically different

behavioural explainability scores (see Table 2.7).

Across model–explainer–dataset runs, test MSE shows a moderate positive correlation with deletion AUC (Pearson $r = 0.430$, $p < 0.001$; Spearman $\rho = 0.715$, $p = 3.1 \times 10^{-11}$; $n = 64$), indicating that lower test error tends to coincide with more negative deletion AUC (larger performance degradation under deletion). Correlations between test MSE and randomisation similarity are weak for neural models where this score is defined ($n = 36$). Correlations with log-scaled noise instability are not statistically significant at $\alpha = 0.05$ (Pearson $p = 0.084$; Spearman $p = 0.651$; $n = 64$; Table 2.7).

Taken together, these associations are mixed rather than uniform across explainability diagnostics. Test MSE and deletion AUC exhibit a substantial positive association in this benchmark, so runs with lower error also tend to show more negative deletion AUC (stronger degradation when attributed inputs are removed). This pattern should not be read as evidence of ground-truth explanation quality: deletion AUC probes consistency with the *trained* model under a fixed masking protocol and can be affected by redundancy and scale, and correlation across runs does not establish causality. By contrast, associations with randomisation similarity and log-scaled noise instability are weak or non-significant here, which cautions against treating a single accuracy–deletion relationship as a general summary of "explainability". As in the synthetic setting, dedicated behavioural and, where possible, oracle-aligned metrics are therefore recommended, rather than inferring explanation reliability from test MSE alone.

Table 2.7: Prediction accuracy vs explainability metrics on real-world airport data. Correlations use test MSE and each behavioural metric over successful runs. Pearson and Spearman coefficients are descriptive summaries of linear and monotonic association. Sample sizes: deletion AUC and log noise stability, $n = 64$; randomisation similarity, $n = 36$ (neural models only; XGBoost omitted because randomisation is undefined). Noise stability is transformed as $\log_{10}(1 + \text{noise stability})$ due to heavy SHAP tails.

Metric Pair	Pearson r	p-value	Spearman ρ	p-value
MSE–Deletion AUC	**0.430**	**0.00039**	**0.715**	3.1×10^{-11}
MSE–Randomization	0.092	0.595	-0.173	0.314
MSE–Noise Stability (log)	-0.218	0.084	-0.058	0.651

2.7 Discussion and Practical Implications

2.7.1 Implications of Synthetic and Real-world Results

Results from synthetic datasets indicate that explainability performance is strongly influenced by temporal pattern characteristics and forecasting architecture. Different explainability methods capture different aspects of temporal structure, and no single method consistently performs well across all scenarios.

Analysis on real-world airport passenger data further shows that explainability behaviour observed under controlled conditions does not always translate directly to operational datasets. Behavioural evaluation metrics provide insight into model sensitivity and robustness, but do not offer definitive measures of explanation correctness.

Together, these findings indicate that explainability in transportation time series forecasting should not be treated as a model-independent property, but rather interpreted in relation to data characteristics, forecasting architecture, and evaluation context.

2.7.2 Practical Insights on Explainability

By combining insights from synthetic and real-world evaluations, several practical considerations emerge for transportation forecasting studies.

First, explainer selection should be informed by dominant temporal patterns in the data. Under the primary oracle metrics reported for each synthetic dataset type (Section 2.5), DeepLift leads on trend (variable-importance Spearman), SHAP is strongest on frequency-single and DeepLift on frequency-multi, and LIME is strongest on variable-coupling; rankings under the separate unified composite score can differ, so practical choices should consider both views rather than a single aggregate label.

Second, real-world attribution results should be treated as exploratory analytical tools rather than definitive indicators of causal influence. In airport passenger flow forecasting, explanations are most informative when used to identify broad temporal regions or variable groups of interest, rather than precise time–variable pairs.

Third, explainability metrics should be selected and interpreted with care. Deletion-based evaluation provides limited but useful comparative information when model redundancy is acknowledged, while stability analysis is valuable for identifying numerically unreliable explanations. Model randomisation tests appear less informative in highly correlated transportation time series.

Overall, effective use of explainability in airport passenger flow forecasting requires alignment between forecasting models, explainer choice, data characteristics, and evaluation objectives.

2.7.3 On SHAP Instability and Numerical Reliability

The real-world stability results reveal extremely large instability values for some SHAP configurations. This behaviour can be attributed to the use of model-agnostic kernel SHAP in high-dimensional, strongly correlated time–variable spaces, where the surrogate regression becomes ill-conditioned.

Small perturbations to the input or background set can lead to disproportionately large attribution magnitude changes, inflating norm-based stability metrics. In this work, extreme instability is interpreted as a warning about numerical reliability rather than as evidence of stronger explanatory power.

Potential mitigation strategies include reducing the dimensionality of the interpretable space, improving background set summarization, applying stronger regularisation, or using model-specific SHAP variants when available. In practice, reporting robust summaries and log-scaled stability metrics can help prevent extreme values from dominating aggregate conclusions. More broadly, the discussion above reinforces that explainability is heterogeneous across models and metrics, and that no single diagnostic should dominate interpretation without cross-checks from complementary probes.

2.8 Conclusion

This chapter presented an empirical evaluation of post-hoc explainability methods for multivariate time series forecasting in airport passenger flow, focusing on attribution for a fixed forecast target. By examining multiple explainability techniques across diverse forecasting architectures, the study aimed to clarify how existing methods behave in transportation time series settings and to identify their strengths and limitations.

Using a dual evaluation strategy that combines synthetic datasets with known ground truth and real-world airport passenger flow data, the results indicate that explainability performance is strongly influenced by temporal pattern characteristics, forecasting architecture, and explainer choice. No single explainability method consistently performs well across all scenarios, and different methods are better suited to different temporal structures. The relationship between prediction accuracy and explainability diagnostics is not uniform: on synthetic data, combination-level associations between test MSE and oracle metrics are generally weak (Section 2.5.3), whereas on real-world data test MSE can co-vary strongly with deletion AUC yet weakly with randomisation similarity and log-scaled noise instability (Table 2.7). Explainability should therefore be assessed with dedicated metrics and not inferred from forecasting accuracy alone.

Evaluation on real-world data further reveals limitations of commonly used behavioural metrics for time series forecasting [37]. While such metrics provide useful insights into model sensitivity and robustness, they often offer limited discriminative power and can be difficult to interpret in the presence of strong temporal dependence and feature redundancy.

Overall, the findings of this study provide empirical guidance for the selection and interpretation of post-hoc explainability methods in transportation time series forecasting. They also highlight the need for future research on evaluation methodologies that are better aligned with the structural characteristics of multivariate temporal data.

References

[1] Ismail Abdulrashid, Reza Zanjirani Farahani, Shamkhal Mammadov, Mohamed Khalafalla, and Wen-Chyuan Chiang. Explainable artificial intelligence in transport logistics: Risk analysis for road accidents. *Transportation Research Part E: Logistics and Transportation Review*, 186:103563, 2024. doi: 10.1016/j.tre.2024.103563.

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, pages 9505–9515, 2018.

[3] Pierre-Daniel Arsenault, Shengrui Wang, and Jean-Marc Patenaude. A survey of explainable artificial intelligence (xai) in financial time series forecasting. *ACM Computing Surveys*, 57(10):1–37, 2025. doi: 10.1145/3729531.

[4] Emre Ates, Burak Aksar, Vitus J. Leung, and Ayse K. Coskun. Counterfactual explanations for multivariate time series. In *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*, pages 1–8. IEEE, 2021. doi: 10.1109/icapai49758.2021.9462056.

[5] João Bento, Pedro Saleiro, André F. Cruz, Mário A. T. Figueiredo, and Pedro Bizarro. TimeSHAP: Explaining recurrent models through sequence perturbations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2565–2573, 2021. doi: 10.1145/3447548.3467166.

[6] Nihad Brahimi, Huaping Zhang, and Zahid Razzaq. Explainable spatio-temporal inference network for car-sharing demand prediction. *ISPRS International Journal of Geo-Information*, 14(4):163, 2024. doi: 10.3390/ijgi14040163.

[7] Yueshan Chen and Sihai Zhang. Wae: An evaluation metric for attribution-based xai on time series forecasting. *Neurocomputing*, 622:129379, 2025. doi: 10.1016/j.neucom.2025.129379.

[8] City and County of San Francisco. Air traffic passenger statistics. `https://data.sfgov.org/Transportation/Air-Traffic-Passenger-Statistics/rkru-6vcg`, 2022. San Francisco International Airport Report on Monthly Passenger Traffic Statistics by Airline. DataSF Open Data Portal. Accessed: 2026-01-13.

[9] Jonathan Crabbé and Mihaela Van Der Schaar. Explaining time series predictions with dynamic masks. In *International Conference on Machine Learning*, pages 2166–2177. PMLR, 2021.

[10] Eamonn Delaney, Derek Greene, and Mark T. Keane. Instance-based Counterfactual Explanations for Time Series Classification. *arXiv preprint arXiv:2009.13211*, 2020.

[11] Joseph Enguehard. Learning perturbations to explain time series predictions. In *International Conference on Machine Learning*, pages 9329–9342. PMLR, 2023.

[12] Ruth Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437, 2017.

[13] Xiaojia Guo, Yael Grushka-Cockayne, and Bert De Reyck. Forecasting airport transfer passenger flow using real-time data and machine learning. *Manufacturing & Service Operations Management*, 24(6):3193–3214, 2022. doi: 10.1287/msom.2021.0975.

[14] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[15] David Hopfe. Airport data for artificial intelligence forecasting of air passenger throughput, 2023.

[16] David H. Hopfe, Kiljae Lee, and Chunyan Yu. Short-term forecasting airport passenger flow during periods of volatility: Comparative investigation of time series vs. neural network models. *Journal of Air Transport Management*, 115:102525, 2024. doi: 10.1016/j.jairtraman.2023.102525.

[17] Bosong Huang, Ming Jin, Yuxuan Liang, Johan Barthelemy, Debo Cheng, Qingsong Wen, Chenghao Liu, and Shirui Pan. ShapeX: Shapelet-Driven Post Hoc Explanations for Time Series Classification Models. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, Vancouver, Canada, 2025. Shapelet-driven post-hoc explanation framework for time series classification.

[18] Aya Abdelsalam Ismail, Mohamed Gunady, Héctor Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series prediction. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[19] Weiwei Jiang and Jiayun Luo. Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications*, 207:117921, 2022. doi: 10.1016/j.eswa.2022.117921.

[20] Lingbai Kong, Hanchen Yang, Wengen Li, Yichao Zhang, Jihong Guan, and Shuigeng Zhou. Traffexplainer: A framework toward GNN-based Interpretable Traffic Prediction. *IEEE Transactions on Artificial Intelligence*, 6(3):559–573, 2024. doi: 10.1109/tai.2024.3459857.

[21] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

[22] Zhiding Liu, Mingyue Cheng, Guanhao Zhao, Jiqian Yang, Qi Liu, and Enhong Chen. Improving Time Series Forecasting via Instance-aware Post-hoc Revision. In

Proceedings of the Conference on Neural Information Processing Systems (NeurIPS 2025), 2025. Model-agnostic framework for post-hoc revision of time series forecasts.

[23] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774, 2017.

[24] Christoph Molnar. *Interpretable Machine Learning*. Lulu.com, 2022. Available at https://christophm.github.io/interpretable-ml-book/.

[25] Qingyi Pan, Wenbo Hu, and Jun Zhu. Series saliency: Temporal interpretation for multivariate time series forecasting. *arXiv preprint arXiv:2012.09324*, 2020.

[26] Nicholas G Polson and Vadim O Sokolov. Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, 79:1–17, 2017. doi: 10.1016/j.trc.2017.02.024.

[27] Vikas C Raykar, Arindam Jati, Sumanta Mukherjee, Nupur Aggarwal, Kanthi Sarpatwar, Giridhar Ganapavarapu, and Roman Vaculin. Tsshap: Robust model agnostic feature-based explainability for time series forecasting. *arXiv preprint arXiv:2303.12316*, 2023.

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[29] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215, 2019.

[30] Udo Schlegel, Hiba Arnout, Mennatallah El-Assady, Daniela Oelke, and Daniel A Keim. Towards a rigorous evaluation of xai methods on time series. In *2019*

IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pages 4197–4201. IEEE, 2019. doi: 10.1109/iccvw.2019.00516.

[31] Xin Sun, Honglei Wang, and Shilong Mei. Explainable highway performance degradation prediction model based on LSTM. *Advanced Engineering Informatics*, 61:102539, 2024. doi: 10.1016/j.aei.2024.102539.

[32] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328, 2017.

[33] Erma Suryani, Shuo-Yan Chou, and Chih-Hsien Chen. Air passenger demand forecasting and passenger terminal capacity expansion: A system dynamics framework. *Expert Systems with Applications*, 37(3):2324–2339, 2010. doi: 10.1016/j.eswa.2009.07.041.

[34] Transportation Security Administration. Foia electronic reading room. `https://www.tsa.gov/foia/readingroom`, 2022. Accessed: 2026-01-13.

[35] Hugues Turbé, Mina Bjelogrlic, Christian Lovis, and Gianmarco Mengaldo. Evaluation of post-hoc interpretability methods in time-series classification. *Nature Machine Intelligence*, 5(3):250–260, 2023. doi: 10.1038/s42256-023-00620-w.

[36] Zhendong Wang, Ioanna Miliou, Isak Samsten, and Panagiotis Papapetrou. Counterfactual explanations for time series forecasting. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 1391–1396. IEEE, 2023. doi: 10.1109/icdm58522.2023.00180.

[37] Shashank Yadav and Vignesh Subbian. Failure modes of time series interpretability algorithms for critical care applications and potential solutions. *arXiv preprint arXiv:2506.19035*, 2025.

[38] Ying Yang, Kai Du, Xingyuan Dai, and Jianwu Fang. Counterfactual Graph Transformer for Traffic Flow Prediction. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 521–527. IEEE, 2023. doi: 10.1109/itsc57777.2023.10422111.

[39] Chih-Kuan Yeh, Been Kim, Sercan Ö. Arik, Chun-Liang Li, and Tomas Pfister. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[40] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating Explanations for Graph Neural Networks. *Advances in Neural Information Processing Systems*, 32, 2019.

3

Sparse Saliency Methods

PAPER II – Sparse Saliency Methods

Enhancing Performance and Interpretability of Multivariate Time-Series Models Through Sparse Saliency

Abstract

Explainable time-series modelling is essential for modern Intelligent Transportation Systems (ITS), but balancing accuracy with interpretability in multivariate time-series forecasting remains a significant challenge. This difficulty arises from the need to discern the importance of features and their temporal dynamics. Autocorrelations and data processing techniques such as sliding windows increase feature dimensions, complicating the use of traditional post-hoc explanation methods. To address these issues, a flexible post-processing approach is proposed that generates sparse, normalised saliency values based on existing saliency generation methods such as GradientSHAP. An optional window aggregation and alignment strategy is also introduced to match the saliency with the original time series data, thereby aligning aggregated saliency with the original monthly series for visual comparison. Additionally, it is explored how sparse saliency can be used for data augmentation to improve model performance. Using naturalistic data from San Francisco airport, this approach is demonstrated in the context of ITS time-series prediction and explanation. Evaluation results show that integrating sparse saliency from high-performing models not only increases XGBoost performance by 10.92% but also reduces model complexity; post-processing yields sparser saliency heatmaps than dense GradientSHAP outputs.

3.1 INTRODUCTION

3.1.1 Motivation

Multivariate time series forecasting plays a crucial role in ITS, which includes domains such as traffic management, autonomous driving, and fault diagnostics [31]. Traffic

prediction plays a pivotal role in traffic planning and network optimisation, enabling the anticipation of future trends and states of interconnected variables critical for informed decision-making and strategic planning. Moreover, the analysis of large-scale time series data from ITS often necessitates human intervention [20], either in real-time or post-hoc, to adhere to ethical and regulatory policies [8]. To mitigate the decision-making risks inherent in black-box models such as Transformers and recurrent neural networks, researchers are focusing on two primary approaches: enhancing model transparency [15] and utilising post-hoc explanations to interpret model predictions [30]. In this chapter, explainability refers to the same post-hoc attribution perspective as in Chapter 2: importance scores over input time steps and variables (saliency) that explain a model's forecast. We do not introduce a new explainability method from scratch; instead, we take existing saliency outputs from black-box forecasters and improve their interpretability and usability through sparse normalisation and optional time-axis alignment. Thus the explainability offered here remains attribution-based, but in a form that is sparser, easier to read, and reusable for saliency-guided data augmentation and model simplification.

Among the post-hoc techniques, extracting and analysing feature importance through saliency values have gained prominence [10][26][4]. However, the understandability of these techniques heavily depends on the performance of the underlying models and the explanation methods used [1][27][24]. Common attribution methods such as Gradient-SHAP (GS) [18] often struggle with interpreting time series data, producing dense and noisy outputs that obscure relevant information, especially when the data has underlying temporal or latent patterns [22]. Furthermore, the rising dimensionality in multivariate time series introduces complexities due to autocorrelation and the application of data processing techniques, such as sliding window methods. These factors complicate the extraction of meaningful insights and make it difficult for humans to derive clear understandings from saliency maps.

In response to these challenges, this study introduces a method that enhances current saliency techniques by integrating sparse and normalisation strategies, thereby reducing

complexity and minimising minor disturbances. Additionally, it aggregates fine-grained saliency maps over training windows so they can be compared with the original time series. Furthermore, it is explored whether sparse saliency values from advanced models can enhance the predictive capabilities of simpler models like XGBoost, which are commonly used in multivariate time series analysis but may underperform compared to more sophisticated neural network architectures.

3.1.2 Contribution

This study makes several key contributions:

- Sparser Saliency Maps: existing saliency (e.g., GradientSHAP) is post-processed to produce sparser tensors $\mathbf{A}'$ (Section 3.3), which appear less cluttered in heatmap form than dense attributions.
- Performance Enhancement Validation: the effectiveness of sparse saliency in improving the performance of underperforming models is empirically validated. The final results demonstrate the efficiency of the proposed sparse saliency enhancement methodology.
- Reduction of Model Complexity: experiments show that integrating sparse saliency into the learning process can reduce the overall complexity of the XGBoost model (less estimators) while simultaneously enhancing its performance.

3.2 Related Works

3.2.1 Time Series Modelling

Researchers have proposed multivariate models to capture the complex interactions and correlations among different traffic variables. One such model is SARIMAX [3], which incorporates seasonal autoregressive integrated moving average principles to account

for seasonal patterns. However, traditional prediction methods may not be effective in situations where external exceptional events, such as pandemics, introduce significant uncertainty in forecasting. To address these challenges, machine learning models have been employed. LightGBM has shown superior performance compared to baseline models that rely solely on historical passenger data, time-related inputs, and quarterly Gross Domestic Product (GDP) by considering epidemic-related variables to predict future air passenger flow [25]. In bus passenger flow prediction tasks that integrate point-of-interest data, XGBoost has shown better training efficiency and prediction accuracy compared to other models [19].

Deep recurrent neural networks like Long Short-Term Memory (LSTM) capture long-term dependencies in sequential data [11]. Gated Recurrent Units (GRU) simplify the architecture by update and reset gates, making it computationally more efficient while still effectively capturing temporal dependencies [7]. For large datasets and long forecasting horizons, transformer-based models have proven to be effective due to their ability to model complex dependencies in time series data. The Temporal Fusion Transformer (TFT) integrates static and time-varying data to capture global and local temporal patterns, enhancing forecasting accuracy and interpretability with attention mechanism [15]. Autoformer improves upon traditional Transformer architectures by integrating an automatic decomposition mechanism [29]. In addition, the Timexer model excels in time series forecasting with exogenous variables, especially in handling low-quality exogenous data and capturing inflection points in the target variable. This model supports multi-level prediction and can forecast multiple future time steps. Techniques like inverted embedding and patch embedding are incorporated to enhance its performance [28]. Constructing auxiliary time series from the original time series, which incorporates inter-series relationships also improve performance. This approach, combined with a simple 2-layer MLP as the core predictor, significantly reduces complexity and parameter count [17].

3.2.2 Saliency Methods for Time Series

Saliency generation in time series data can be categorised into three main methods: gradient-based, perturbation-based, and attention-based. Gradient-based methods compute gradients of the model's output with respect to the input time series, using these gradients to create saliency maps that highlight influential segments or features. Attention mechanisms, exemplified by the self-attention in Transformers, generate saliency maps by computing attention weights for each timestep or feature, indicating their importance for model predictions. However, traditional saliency techniques like Gradient-weighted Class Activation Mapping (Grad-CAM) [23], Shapley Additive Explanations (SHAP) [18], originally designed for image and natural language processing, do not naturally suit time series data. These methods often fail in tasks where the detection of latent patterns—such as dominant frequencies or trend dependencies—is essential, which is common in applications like physiological signal analysis and fault diagnosis. Evaluations on time series classification tasks reveal that existing perturbation and gradient-based methods sometimes lack consistency and robustness [27][22].

Recently, several saliency generation methods specific to time series have been proposed, Series saliency method extracts "series images" from sliding windows of the time series and applies a saliency map segmentation based on the smallest destroying region principle. This approach considers both the feature and temporal dimensions, improving accuracy and interpretability [21]. Jonathan Crabbé et al. introduce Dynamic Masks, which employ perturbation masks tailored to time series data to capture temporal dependencies and identify relevant features for predictions. These masks are optimised for parsimony and legibility, improving feature importance identification [4]. Joseph Enguehard [5] proposes learning perturbations to explain time series predictions, where a trainable mask is used to perturb the input and identify influential features and their timing. This approach learns both the masks and associated perturbations, resulting in enhanced explanations. Additionally, the Time Interpret library extends Captum for temporal data explanation, providing saliency generation methods and evaluation tools [6]. The ContraLSP frame-

work combines contrastive learning and sparse gate techniques, introduces counterfactual samples and maintains the distribution of the data, resulting in improved explanation quality and alleviating distribution shift issues [16].

The core idea behind saliency guided training is to focus the model's attention on the most prominent regions of the input during the training process, instead of treating all parts equally [14]. The saliency guided training procedure involves iteratively masking features that have small and possibly noisy gradients, while also aiming to maximise the similarity of model outputs for both the masked and unmasked inputs. This approach has shown effectiveness in various domains, including images, language, and multivariate time series [12].

3.3 Methodology

3.3.1 Overall Framework

As depicted in Figure 3.1, this chapter proposes a novel framework that incorporates sparse saliency tensors to enhance interpretability and accuracy in time series prediction models. The process begins with dataset preparation, where the raw dataset tensor $\mathbf{D}$ undergoes a windowing technique to create sequences that capture temporal dependencies. The symbol seq_len denotes the number of time steps or observations included in each window extracted from $\mathbf{D}$. The i-th window is expressed as:

$$\mathbf{W}_i = \mathbf{D}[i : i + \text{seq_len}] \tag{1}$$

Following data preparation, the windowed training tensor $\mathbf{D}'_{\text{train}}$ is obtained, multiple forecasters $\{M_i\}$ are then trained using $\mathbf{D}'_{\text{train}}$, and the optimal model M is selected based on performance metrics. An explainer subsequently generates the original saliency tensor $\mathbf{A}$ from M. Let $\mathbf{A} \in \mathbb{R}^{N_w \times T \times F}$ denote the tensor of GradientSHAP attributions aligned with $\mathbf{D}'_{\text{train}}$, where N_w is the number of training windows, $T = \text{seq_len}$, and F is the number of

features (variables). Hereafter, $\mathbf{A}$ and $\mathbf{A}'$ are referred to as *saliency tensors* (real-valued attributions); a slice $\mathbf{A}_{w,:,:}$ for $w = 1,\ldots,N_w$ may be displayed as a time–feature heatmap. The auxiliary mask used inside E is not what is denoted by $\mathbf{A}'$.

The sparse saliency generation phase processes $\mathbf{A}$ through a proposed sparse and normalise function E, resulting in the sparse saliency tensor $\mathbf{A}'$ that emphasises the most impactful features. The data dimensions are then aggregated from $(N_w \times T \times F)$ to $(T \times F)$. The approach of utilising post-processing to introduce sparsity provides the framework with enhanced flexibility.

In this study, for dataset augmentation part, the sparse saliency tensor $\mathbf{A}'$ is simply integrated with the corresponding positions of the windowed training tensor $\mathbf{D}'_{\text{train}}$ using element-wise multiplication, denoted as:

$$\mathbf{D}''_{\text{train}} = \mathbf{A}' \odot \mathbf{D}'_{\text{train}} \tag{2}$$

This enhanced dataset serves in the final phase, model update, where M is retrained or updated with $\mathbf{D}''_{\text{train}}$, resulting in an enhanced model M' that incorporates refined insights for improved performance and interpretability.

3.3.2 Sparse Normalised Saliency

This section clarifies the sparse saliency generation phase and introduces the proposed function E, designed to enhance the clarity and readability of saliency tensors by integrating mask filtering with adaptive scaling techniques. Function E selectively emphasises the most impactful features, thereby producing sparser tensors $\mathbf{A}'$ than the original $\mathbf{A}$; heatmaps of $\mathbf{A}'$ retain non-zero attributions only where the mask $\mathbb{M}$ is true.

Given the original saliency tensor $\mathbf{A}$ generated by an explainer, which signifies the importance assigned to each feature by a model, the function aims to normalise these attributions using scaling and thresholding mechanisms controlled by the parameters percentile p, scalar α, and normalisation method. In the piecewise expressions below, i

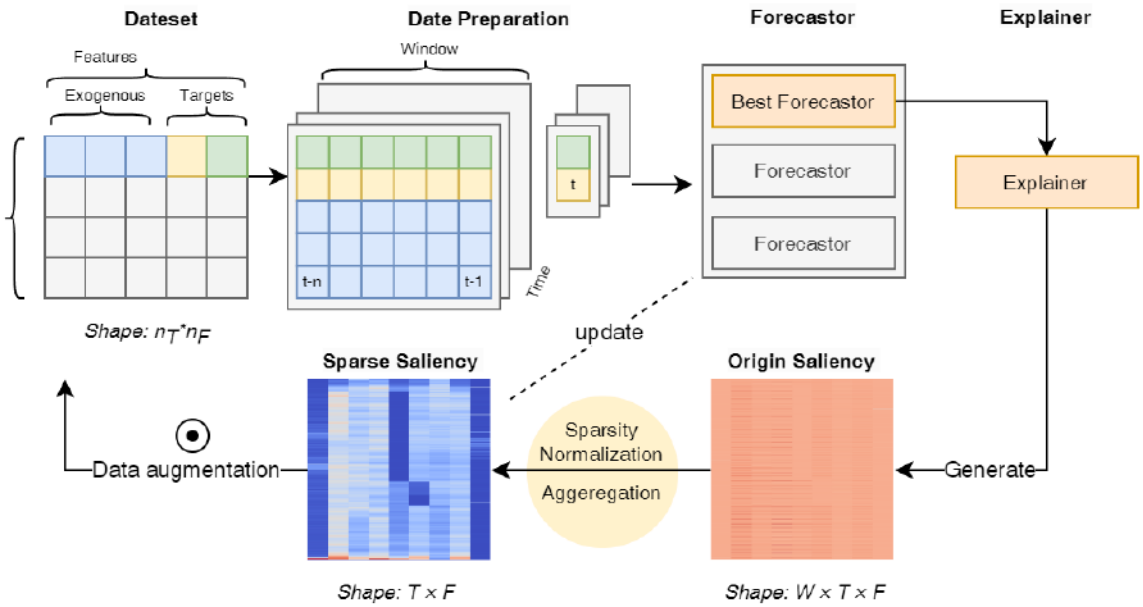


Figure 3.1: Illustration of the proposed framework integrating saliency tensors into the forecasting model lifecycle. The diagram highlights the process of normalising and sparsifying saliency tensors to enhance predictive accuracy and interpretability.

indexes elements of $\mathbf{A}$ and $\mathbf{A}'$ in a fixed flattening (consistent with $\mathbb{M}[i]$).

First, a threshold T is defined based on the given percentile p of the absolute values of attributions $\mathbf{A}$:

$$T = \text{quantile}\left(|\mathbf{A}|, \frac{p}{100}\right) \tag{3}$$

The scale factor S is computed using the standard deviation σ of the attributions $\mathbf{A}$ and the given scalar α, which influences how much the standard deviation of the attributions affects the scaling of the data:

$$S = \frac{\alpha}{\sigma(\mathbf{A})} \tag{4}$$

A mask $\mathbb{M}$ is created to identify elements in $\mathbf{A}$ that are greater than or equal to the threshold T:

$$\mathbb{M} = |\mathbf{A}| \geq T \tag{5}$$

Depending on the chosen method $\mathscr{M}$, scaled attributions $\mathbf{A}'$ are calculated. For the hyperbolic tangent method ($\mathscr{M}$=$tanh$):

$$[\mathbf{A}']_i = \begin{cases} \frac{e^{S\cdot[\mathbf{A}]_i} - e^{-S\cdot[\mathbf{A}]_i}}{e^{S\cdot[\mathbf{A}]_i} + e^{-S\cdot[\mathbf{A}]_i}} & \text{if } \mathbb{M}[i] = \text{True} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

When $\mathscr{M}$=$Softmax$:

$$[\mathbf{A}']_i = \begin{cases} \frac{\exp(S\cdot[\mathbf{A}]_i)}{\sum_{j\in J}\exp(S\cdot[\mathbf{A}]_j)} & \text{if } \mathbb{M}[i] = \text{True} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where J is the set of indices where $\mathbb{M}$ is True.

To represent the function E in a formalised manner, the steps can be summarised and combined into a sequence of operations. The entire process can be expressed in a formula:

$$E(\mathbf{A}, p, \alpha, \mathscr{M}) = \mathbf{A}' \tag{8}$$

In summary, by applying mask $\mathbb{M}$ and normalisation operations using the *tanh* or *softmax* methods, the sparsity of the original saliency tensor $\mathbf{A}$ derived from the model M is increased while the redundant information from dense attributions is reduced. Consequently, a smaller subset of time–feature cells carries non-zero attribution in $\mathbf{A}'$.

3.4 Experiments and Analysis

3.4.1 Experiment Setup

This section details the experimental setup, designed as follows:

3.4.1.1 Data Preparation

In this study, the Airport Passenger Flow Dataset obtained from San Francisco Airport [2] was used as the primary data source. This dataset allowed analysis of the monthly passenger flow across eight different boarding areas (A-G, Other). To forecast future traffic for a specific subsequent month, the data were segmented into 12-month windows and an 85% training and 15% testing split was utilised. To provide a comprehensive understanding of the variations in passenger flow influenced by external factors, pandemic-related search trends from Google Trends [9] were also incorporated as an exogenous variable, enriching the dataset. In addition, when incorporating the sparse saliency tensor $\mathbf{A}'$, element-wise multiplication with the windowed training tensor $\mathbf{D}'_{\text{train}}$ was used and scaling using the median of entries of $\mathbf{A}'$ was applied to the entire testing set. This approach allowed effective integration of $\mathbf{A}'$ into the analysis, ensuring consistency in the scaling process across both the training and testing datasets.

3.4.1.2 Sparse Saliency Integration

A suite of basic forecasters tailored to multivariate time series analysis was employed, including XGBoost, LSTM, GRU, and Transformer models. The model delivering the best performance was then used to generate the original saliency tensor $\mathbf{A}$ using GradientSHAP. Sparse normalisation techniques were applied with a mask percentile P of 0, 5, 10, 50, 90, a scalar α of 1, and a hyperbolic tangent activation function. The training dataset was augmented by element-wise multiplication with the sparsely normalised tensor $\mathbf{A}'$, whereas the testing set was scaled using the median of $\mathbf{A}'$ computed from the training split. This approach confined the saliency generation process to the training data, preventing data leakage and its potential impact on the results. Then, the modified datasets were utilised for model update. The effects of varying mask percentiles on the XGBoost model's performance were analysed, along with the performance variability across different estimators, which serve as indicators of the model's complexity. Based on preliminary empirical tests, use of the tanh function is recommended when there are many extremely

low saliency values, as it converges rapidly near zero and effectively filters out more noise.

3.4.1.3 Measurement metrics

The effectiveness of the forecasting models is quantitatively assessed using four standard error metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Symmetric Mean Absolute Percentage Error (sMAPE), and Coefficient of Determination (R2). These metrics provide a comprehensive evaluation of the models' accuracy in forecasting future values. Considering that different normalisations of $\mathbf{A}'$ may introduce varying scaling factors to the augmented training dataset in data preparation process, the need for consistent model evaluation across different scenarios is taken into account. Therefore, all values are descaled to their original scale before conducting the comparative analysis. This approach allows fair and accurate comparisons among the different models, regardless of the variations introduced by $\mathbf{A}'$.

3.4.2 Results Analysis

Based on the results presented in Table 3.1, "XGBoost+" refers to using a dataset enhanced by the GRU model, without exogenous variables at a 10% mask percentile. The more powerful Transformer model does not show a significant advantage with this relatively small dataset. On the contrary, the lightweight GRU model demonstrates the best performance on the original dataset, both with and without the inclusion of exogenous variables.

In Figure 3.2, the predictions of GRU and XGBoost models on the test dataset with exogenous variables are compared. It is evident that XGBoost initially exhibits a significant discrepancy with the actual values, particularly in boarding area D, as indicated by the red circle. By utilising the sparse saliency enhanced dataset generated by the GRU model with exogenous variables and a mask percentile $p = 10$, the performance of the XGBoost model with same estimators as before shows a substantial improvement. This update effectively resolves the prediction issue in boarding area D, addressing the earlier

Table 3.1: Comparison of models with and without exogenous variables.

	MAE	MSE	sMAPE	R2
only history data				
XGBoost	6.63E+04	9.59E+09	341.0656	0.2072
LSTM	7.68E+04	1.03E+10	337.1754	0.7716
GRU	6.96E+04	8.89E+09	321.5465	**0.8230**
Transformer	1.10E+05	2.21E+10	383.1288	0.5237
XGBoost+	**6.20E+04**	**8.21E+09**	**308.5541**	0.3485
with exogenous variables				
XGBoost	6.64E+04	9.63E+09	341.0782	0.2327
LSTM	7.52E+04	1.04E+10	339.0567	0.7872
GRU	6.73E+04	**8.01E+09**	324.1746	**0.8237**
Transformer	9.64E+04	1.63E+10	366.8931	0.5646
XGBoost+	**6.21E+04**	8.25E+09	**308.5801**	0.3486

discrepancy observed.

Table 3.2 presents results demonstrating the influence of different mask percentile values of sparse saliency on the performance of the XGBoost model. Notably, even when the number of XGBoost estimators was reduced to 20, optimal performance was achieved with a mask percentile of 10. This finding underscores the improvement in model performance through the use of an updated model with an appropriate mask percentile for sparse saliency. The MAE metric indicates that XGBoost+, with a score of 6.21e+04, outperformed the previous best performance achieved by GRU, which scored 6.73e+04, resulting in a 7.73% improvement for data with exogenous variables and a 10.92% improvement for data without exogenous variables. Conversely, when directly using the dense GradientSHAP saliency tensor without sparse normalisation, the model's performance decreased.

As illustrated in Table 3.2, an excessive mask percentile p, such as 90, results in significant information loss, negatively impacting outcomes. However, as depicted in Figure 3.3, these higher p yield sparser heatmaps; the figure places aggregated $\mathbf{A}$, aggregated $\mathbf{A}'$, and monthly passenger flows side by side for the same period so readers can compare saliency structure against the series.

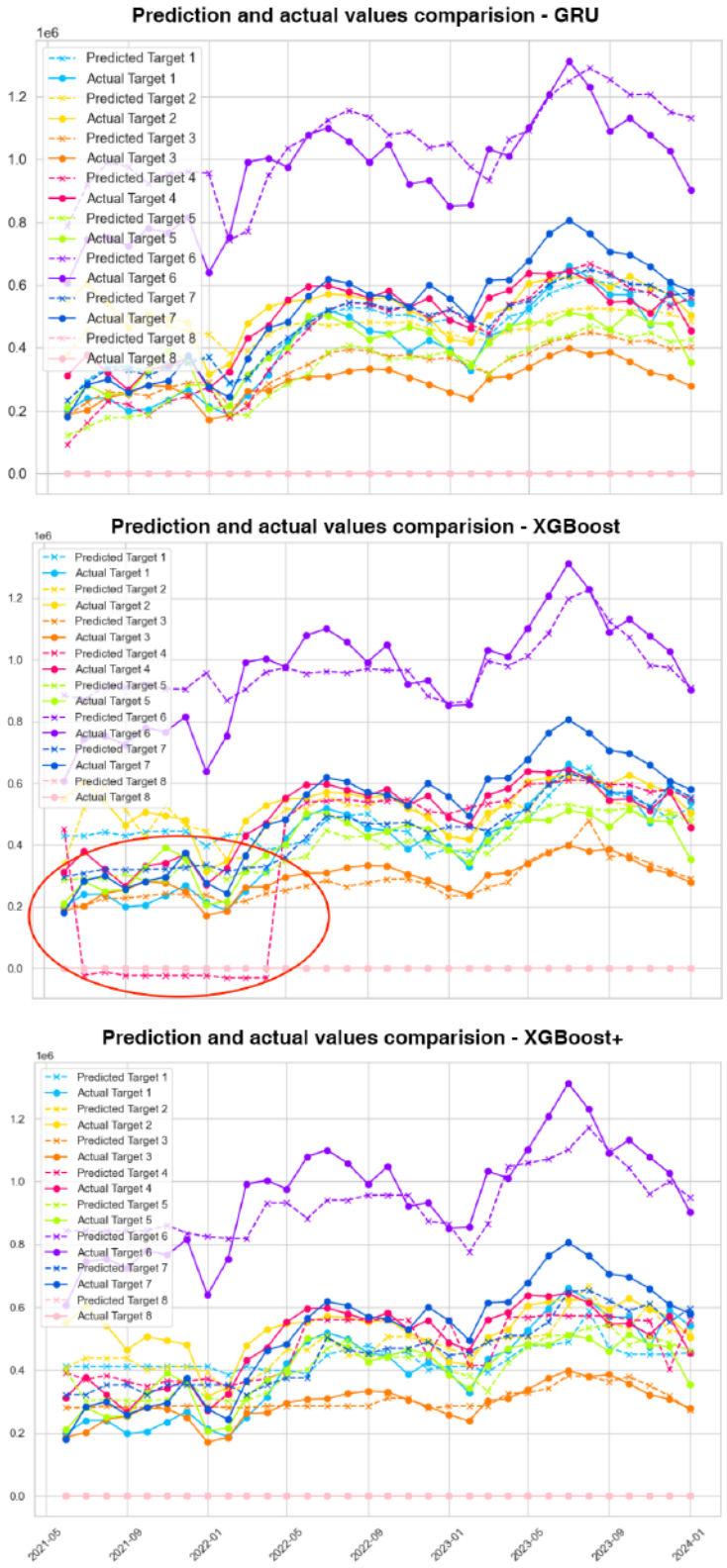


Figure 3.2: Comparison of GRU, XGBoost, and XGBoost+ predictions against ground truth for boarding areas A–G and Other (targets 1–8).

Table 3.2: Performance comparison for different sparse saliency tensor settings (mask percentiles) and XGBoost estimator counts.

mask Percentile	xgboost estimators	Descaled MAE	Descaled MSE	Descaled SMAPE	R2
0	10	6.57E+04	8.58E+09	314.0539	0.3186
0	**20**	**6.26E+04**	**8.38E+09**	**309.4975**	**0.3344**
0	40	7.01E+04	1.04E+10	320.8295	0.2339
0	60	7.21E+04	1.10E+10	324.9005	0.1864
0	80	7.26E+04	1.12E+10	325.7797	0.1690
0	100	7.32E+04	1.14E+10	326.9381	0.1548
5	10	6.57E+04	8.58E+09	314.0543	0.3186
5	**20**	**6.26E+04**	**8.38E+09**	**309.4980**	**0.3344**
5	40	7.01E+04	1.04E+10	320.8301	0.2339
5	60	7.21E+04	1.10E+10	324.9010	0.1864
5	80	7.27E+04	1.12E+10	326.0429	0.1672
5	100	7.33E+04	1.14E+10	327.0487	0.1528
10	10	6.58E+04	8.60E+09	314.1743	0.3179
10	**20**	**6.20E+04**	**8.21E+09**	**308.5541**	**0.3485**
10	40	6.99E+04	1.06E+10	320.7102	0.2303
10	60	7.19E+04	1.13E+10	324.5954	0.1647
10	80	7.28E+04	1.16E+10	326.0260	0.1459
10	100	7.39E+04	1.20E+10	327.8790	0.1233
50	**10**	**6.63E+04**	**9.03E+09**	**315.6055**	**0.2611**
50	20	6.91E+04	9.78E+09	320.2657	0.2250
50	40	7.74E+04	1.25E+10	336.0832	0.0589
50	60	8.14E+04	1.35E+10	342.9682	-0.0791
50	80	8.42E+04	1.46E+10	347.2121	-0.1596
50	100	8.51E+04	1.50E+10	347.7131	-0.1912
90	**10**	**1.03E+05**	**1.98E+10**	**368.6473**	**-0.4944**
90	20	1.12E+05	2.46E+10	387.9443	-0.7831
90	40	1.20E+05	3.22E+10	399.8461	-1.2336
90	60	1.26E+05	3.71E+10	405.6978	-1.7085
90	80	1.26E+05	3.91E+10	403.8361	-1.8246
90	100	1.26E+05	4.00E+10	403.9194	-1.9232

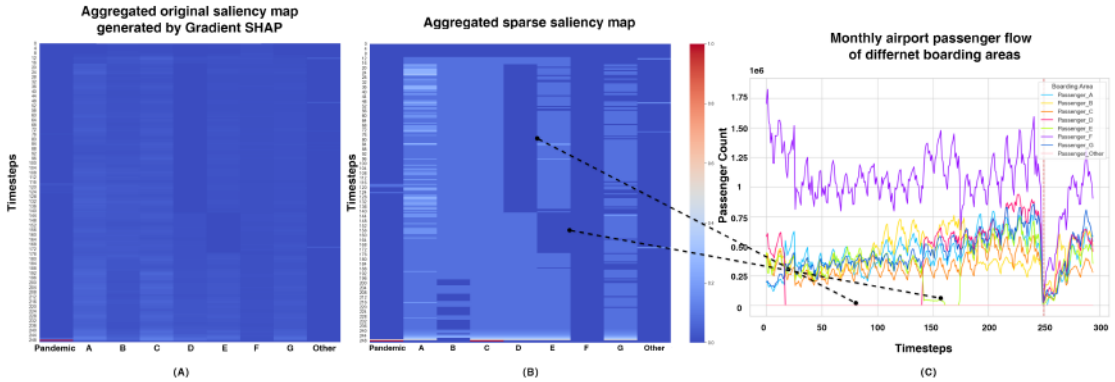


Figure 3.3: Readability comparison of saliency visualisations (heatmaps of the same tensors): (A) aggregated original saliency tensor $\mathbf{A}$; (B) aggregated sparse saliency tensor $\mathbf{A}'$; (C) monthly passenger flow for San Francisco Airport boarding areas.

3.5 Conclusions and Future Work

This chapter examined sparse normalisation of GradientSHAP saliency and its use in saliency-guided data augmentation for multivariate airport passenger-flow forecasting. Under the evaluated configurations, the processed saliency improved forecast accuracy and reduced XGBoost complexity. Figure 3.3 provides a qualitative comparison between aggregated dense and sparse saliency and the monthly flow series. The improved visual readability claims are based on qualitative inspection rather than a formal user study.

However, the universal applicability of this approach across various time series tasks remains to be explored. Future work should include a comparative analysis of the impact of saliency tensors generated by different explainers after undergoing sparsification and normalisation to gain further insights. Additionally, the selection of explainers and the development of improved data augmentation methods merit further investigation.

Systematic evaluation of explainable interfaces [13] would complement the qualitative comparisons presented here. As this study is based on a small-scale, coarse-grained dataset with monthly resolution and limited real-time requirements, future work should extend the evaluation to larger and finer-grained datasets to assess scalability and real-time applicability.

References

[1] C. Balestra, B. Li, and E. Müller. On the consistency and robustness of saliency explanations for time series classification. *arXiv preprint arXiv:2309.01457*, 2023.

[2] City and County of San Francisco. City and County of San Francisco Air Traffic Passenger Statistics (DataSF). `https://data.sfgov.org/Transportation/Air-Traffic-Passenger-Statistics/rkru-6vcg/about_data`, 2024. Accessed May 2024.

[3] M. Cools, E. Moons, and G. Wets. Investigating the variability in daily traffic counts through use of arimax and sarimax models: assessing the effect of holidays on two site locations. *Transportation Research Record*, 2136:57–66, 2009.

[4] Jonathan Crabbé and Mihaela Van Der Schaar. Explaining time series predictions with dynamic masks. In *International Conference on Machine Learning*, pages 2166–2177, 2021.

[5] J. Enguehard. Learning perturbations to explain time series predictions. In *International Conference on Machine Learning*, pages 9329–9342, 2023.

[6] J. Enguehard. Time interpret: a unified model interpretability library for time series. *arXiv preprint arXiv:2306.02968*, 2023.

[7] R. Fu, Z. Zhang, and L. Li. Using lstm and gru neural network methods for traffic flow prediction. In *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pages 324–328, 2016.

[8] Loveleen Gaur and Biswa Mohan Sahoo. Introduction to explainable AI and intelligent transportation. In *Explainable Artificial Intelligence for Intelligent Transport-*

ation Systems: Ethics and Applications, pages 1–25. Springer, Cham, 2022. doi: 10.1007/978-3-031-09644-0_1.

[9] Google LLC. Google Trends. `https://trends.google.com/trends`, 2024. Accessed May 2024.

[10] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16:45–74, 2024.

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

[12] A. Ismail, H. Corrada Bravo, and S. Feizi. Improving deep learning interpretability by saliency guided training. In *Advances in Neural Information Processing Systems*, volume 34, pages 26726–26739, 2021.

[13] Xiangqi Kong, Yang Xing, Antonios Tsourdos, Ziyue Wang, Weisi Guo, Adolfo Perrusquia, and Andreas Wikander. Explainable interface for human-autonomy teaming: A survey. *arXiv preprint arXiv:2405.02583*, 2024.

[14] A. Liang, J. Thomason, and E. Bıyık. Visarl: Visual reinforcement learning guided by human saliency. *arXiv preprint arXiv:2403.10940*, 2024.

[15] Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012.

[16] Z. Liu, Y. Zhang, T. Wang, Z. Wang, D. Luo, M. Du, M. Wu, Y. Wang, C. Chen, L. Fan, and Others. Explaining time series via contrastive and locally sparse perturbations. *arXiv preprint arXiv:2401.08552*, 2024.

[17] J. Lu, X. Han, Y. Sun, and S. Yang. Cats: Enhancing multivariate time series forecasting by constructing auxiliary time series as exogenous variables. *arXiv preprint arXiv:2403.01673*, 2024.

[18] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[19] Wanjun Lv, Yongbo Lv, Qi Ouyang, and Yuan Ren. A bus passenger flow prediction model fused with point-of-interest data based on extreme gradient boosting. *Applied Sciences*, 12(3):940, 2022. doi: 10.3390/app12030940.

[20] Ashley Nunes, Bryan Reimer, and Joseph F. Coughlin. People must retain control of autonomous vehicles. *Nature*, 556(7700):169–171, 2018. doi: 10.1038/d41586-018-04158-5.

[21] Q. Pan, W. Hu, and N. Chen. Two birds with one stone: Series saliency for accurate and interpretable multivariate time series forecasting. In *IJCAI*, pages 2884–2891, 2021.

[22] M. Schröder, A. Zamanian, and N. Ahmidi. Post-hoc saliency methods fail to capture latent feature importance in time series data. In *International Workshop on Trustworthy Machine Learning for Healthcare*, pages 106–121, 2023.

[23] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[24] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling lime and shap: Adversarial attacks on post-hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

[25] H. Tang, J. Yu, B. Lin, Y. Geng, Z. Wang, X. Chen, L. Yang, T. Lin, and F. Xiao. Airport terminal passenger forecast under the impact of covid-19 outbreaks: A case study from china. *Journal of Building Engineering*, 65:105740, 2023.

[26] Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David Duvenaud, and Anna Goldenberg. What went wrong and when? instance-wise feature importance for time-series black-box models. In *Advances in Neural Information Processing Systems*, volume 33, pages 799–809, 2020.

[27] H. Turbé, M. Bjelogrlic, C. Lovis, and G. Mengaldo. Evaluation of post-hoc interpretability methods in time-series classification. *Nature Machine Intelligence*, 5: 250–260, 2023.

[28] Y. Wang, H. Wu, J. Dong, Y. Liu, Y. Qiu, H. Zhang, J. Wang, and M. Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. *arXiv preprint arXiv:2402.19072*, 2024.

[29] Haixu Wu, Jialong Xu, Jieming Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pages 22419–22430, 2021.

[30] Yibing Xie, Nichakorn Pongsakornsathien, Alessandro Gardi, and Roberto Sabatini. Explanation of machine-learning solutions in air-traffic management. *Aerospace*, 8 (8):224, 2021. doi: 10.3390/aerospace8080224.

[31] Yang Xing, Zhongxu Hu, Peng Hang, and Chen Lv. Learning from the dark side: A parallel time series modelling framework for forecasting and fault detection on intelligent vehicles. *IEEE Transactions on Intelligent Vehicles*, 9(2):3205–3219, 2024. doi: 10.1109/TIV.2023.3342648.

4

Interpretable Decomposition Forecasting

iDecPred: A Lightweight and Interpretable Extrapolative Neural Decomposition Model for Multivariate Time Series Forecasting

Abstract

Forecasting multivariate time series in high-stakes domains requires models that are accurate, interpretable, and computationally efficient. While recent neural architectures deliver strong predictive performance, they often rely on opaque latent representations and lack structured decomposition. This chapter introduces Interpretable Decomposition Prediction (iDecPred), a lightweight and interpretable neural decomposition framework that formulates forecasting as an extrapolative process. The model explicitly disentangles temporal dynamics into three components: (i) a spline-based trend extrapolator with affine augmentation for long-horizon forecasting, (ii) a channel-aware seasonal encoder with harmonic decoders to capture heterogeneous and non-stationary periodicities, and (iii) a residual predictor for irregular variations. These are adaptively fused through a channel-wise gating mechanism to produce interpretable forecasts. Experiments on seven real-world benchmarks (ETTh1/2, ETTm1/2, Electricity, Traffic, and Weather) and 16 synthetic datasets with ground-truth decompositions demonstrate that iDecPred achieves state-of-the-art or competitive accuracy while using up to 10$\times$ fewer parameters and offering 10–20$\times$ lower inference latency on high dimension datasets (862 channels) compared to transformer-based baselines. Beyond accuracy, iDecPred directly outputs trend, seasonal, and residual components during forecasting, enabling transparent diagnostics and faithful component-wise interpretation without post-hoc analysis. By bridging classical decomposition principles with modern neural design, iDecPred establishes a compact yet effective paradigm for interpretable long-horizon time series forecasting.

4.1 Introduction

In this chapter, explainability is understood as intrinsic interpretability by decomposition: the model (iDecPred) explicitly decomposes each series into trend, seasonal, and residual components and produces a forecast for each. Explanations are not obtained via a separate post-hoc explainer but arise directly from the model outputs, which include the predicted trajectories of each component and offer interpretable insights into how their dynamics shape the final prediction.

Multivariate time series forecasting is a fundamental problem with wide ranging applications in energy systems [36], finance [2], transportation [53], and industrial automation [54]. In these domains, accurate and interpretable predictions are essential to ensure operational reliability, enable risk management, and support scientific discovery [21]. For example, in power systems, decomposing the load series into trend, seasonal, and residual parts improves both robustness and interpretability under extreme events, thereby supporting capacity planning and emergency scheduling [23]. In transportation, spatiotemporal-aware trend-seasonality decomposition improves traffic flow prediction accuracy while enabling interpretable diagnostics at the network level, which is crucial for congestion management and infrastructure planning [46, 7, 42]. Explicitly separating long-term trends, seasonal patterns, and short-term noise enhances transparency and yields semantically meaningful representations [39].

Classical statistical approaches such as Autoregressive Integrated Moving Average (ARIMA) [1] and exponential smoothing methods [27] perform explicit trend-seasonal decomposition, but their reliance on strong assumptions such as linearity or stationarity limits their ability to model complex real-world signals. Fourier transform-based methods are effective for interpolation but cannot reliably extrapolate into the future [13]. Neural Ordinary Differential Equation (ODE) variants can better capture smooth temporal dynamics but often incur excessive computational cost [19]. Recent neural forecasting architectures, such as Recurrent Neural Networks (RNN) and Transformer-based variants achieve strong predictive accuracy but usually depend on opaque latent representations

without imposing a structured decomposition of temporal dynamics. Even in synthetic signals with simple known periodicity, leading architectures such as iTransformer [26] may not recover meaningful seasonal components [56].

Some works have incorporated decomposition ideas into neural forecasting by explicitly modelling trend, seasonal, or residual components. For example, Neural Basis Expansion Analysis for Interpretable Time Series Forecasting (N-BEATS) [30] employs forward and backward basis expansions to approximate trend and seasonality. While it achieves strong accuracy on stationary and smooth time series, its reliance on global fully connected blocks limits adaptability to complex or non-stationary dynamics. Autoformer [50] and Frequency Enhanced Decomposed Transformer (FEDformer) [64] adopt simple moving average to extract trend and embed decomposition blocks within Transformer architectures; however, their designs mainly aim to improve sequence modelling efficiency rather than to ensure faithful component-wise forecasts.

Other popular architectures such as Patch Time Series Transformer (PatchTST) [29] and Inverted Transformer for Time Series Forecasting (iTransformer) [26] achieve strong accuracy through patch representations or variable-wise attention. However, they remain largely black-box predictors that do not guarantee interpretable or extrapolative decomposition. Moreover, their reliance on large Transformer backbones results in higher parameter counts and greater computational demands.

Despite recent progress, existing models still lack efficient, accurate, and robust mechanisms for forecast-oriented decomposition that predict future trend, seasonal, and residual components instead of reconstructing past signals [13, 3, 30, 49]. This chapter introduces iDecPred, a neural framework that forecasts each component through a spline-based trend extrapolator, a channel-aware Fourier seasonal decoder, and a lightweight residual predictor. Integrated with a compact attention gate, iDecPred combines classical decomposition principles with modern deep learning to achieve interpretable, accurate, and efficient forecasting.

This work makes the following key contributions:

- **Forecast-oriented decomposition framework.** iDecPred is proposed as a lightweight neural model that explicitly decomposes multivariate time series into trend, seasonal, and residual components and forecasts them jointly. Unlike existing methods, iDecPred is designed for extrapolative forecasting rather than interpolation, thereby aligning decomposition with the predictive task.

- **Novel architectural modules.** This chapter introduces (i) a spline-based trend extrapolator with affine augmentation for long-horizon forecasting, (ii) a novel channel-aware seasonal module with harmonic decoders to capture heterogeneous and non-stationary periodicities, and (iii) a lightweight residual predictor for irregular variations. These are fused through a channel-wise adaptive gate, ensuring interpretability and efficiency.

- **Comprehensive empirical validation.** Extensive experiments on seven real-world benchmarks and sixteen synthetic datasets with ground-truth decompositions demonstrate that iDecPred achieves state-of-the-art (SOTA) or competitive accuracy while requiring up to $10\times$ fewer parameters and reduced training time and inference latency compared with existing baselines.

- **Interpretability and diagnostic insights.** Beyond accuracy, iDecPred directly outputs interpretable trend, seasonal, and residual components during forecasting, enabling transparent diagnostics and structured evaluation of decomposition quality without post-hoc explanation.

The remainder of this chapter is organised as follows. Section 4.2 reviews related work in time series decomposition and frequency-aware neural forecasting. Section 4.3 presents the architecture of iDecPred. Section 4.4 describes the experimental setup, followed by results and analysis in Section 4.5. Section 4.6 discusses limitations and future directions, and Section 4.7 concludes the chapter.

4.2 Related Work

This section reviews previous research on time series decomposition, decomposition-based forecasting, frequency-aware modelling, and evaluation frameworks. These areas form the intellectual foundation of iDecPred.

4.2.1 Classical Decomposition and the Need for Interpretability

Classical time series analysis is based on the assumption that signals can be decomposed into structured components such as long-term trends, seasonal cycles, and stochastic residuals [15]. Statistical models including ARIMA [1] and Holt–Winters exponential smoothing [27, 48] explicitly capture these components within additive or multiplicative frameworks. Seasonal-Trend Decomposition using Loess (STL) [8, 15] extends this idea through nonparametric smoothing, while variants such as Multiple Seasonal-Trend Decomposition using Loess (MSTL) [4] and Robust Seasonal-Trend Decomposition Algorithm (RobustSTL) [47] enhance adaptability to multiple seasonalities and improve robustness to outliers. Spectral methods, including Fourier and wavelet transforms [24], provide complementary perspectives on periodic behaviour, although they rely on fixed bases and generally remain decoupled from end-to-end forecasting pipelines. Taken together, these methods offer interpretability but rely on restrictive assumptions of linearity, stationarity, or fixed periodicity, which limit their scalability for modern multivariate forecasting.

Equally critical is the question of evaluation. Beyond measuring aggregate forecast errors, reliable decomposition requires assessing whether extracted components align with underlying structures and whether residuals exhibit expected statistical properties [15, 5]. Synthetic benchmarks with known ground truth [4], residual diagnostics (e.g., autocorrelation tests), and domain-informed validation like known daily or weekly cycles are all used in practice. Recent work proposes hybrid evaluation frameworks that combine quantitative diagnostics with visual inspection and domain knowledge [49, 32]. Within this perspective, interpretability is assessed not only through predictive accuracy but also

through the fidelity of the decomposed components.

Although classical statistical approaches naturally provide interpretable components, most neural forecasting architectures behave as black boxes, which hinders their adoption in high-stakes domains where understanding the drivers of predictions is essential [21, 62]. Post-hoc explanation techniques such as Shapley Additive Explanations (SHAP) [28], Local Interpretable Model-agnostic Explanations (LIME) [35], and their time-series adaptations [33, 60] have been proposed. However, these methods are often computationally expensive and sensitive to input perturbations, which can undermine their reliability in practice [55, 37].

In summary, while classical decomposition methods offer transparency through explicit modelling of trend, seasonality, and residuals, they rely on restrictive assumptions that limit scalability to complex, multivariate settings. Conversely, modern neural forecasters achieve strong accuracy but often sacrifice interpretability. Bridging these paradigms calls for models that retain the structural clarity of classical decomposition while leveraging the representational power of deep learning.

4.2.2 Decomposition-Based Neural Forecasting

Recent neural forecasting models revisit classical decomposition ideas, aiming to couple the interpretability of trend–seasonal separation with the representational power of deep learning. These architectures can be broadly grouped into static decomposition, parametric decomposition, and learnable decomposition.

Several models apply fixed filters to separate trend and seasonal components before forecasting. For example, DLinear [59] uses a moving-average filter to extract trend and feeds the residual into linear layers. Similarly, FEDformer, an extension of Autoformer, relies on moving averages for trend removal while leveraging Fourier attention to capture seasonalities. These methods inherit interpretability from explicit filtering, but because the decomposition is static and not optimised jointly with forecasting, they often lack adaptability to domain-specific dynamics.

Architectures such as N-BEATS [30] model signals via combinations of polynomial and harmonic bases. While interpretable variants of N-BEATS explicitly parameterize trend and seasonal blocks, they generally sacrifice predictive performance compared to their unconstrained counterparts. It demonstrates interpretability by imposing structure, but at the expense of flexibility.

A more principled line of work integrates decomposition into the architecture itself. ETSformer [49] uses Exponential Smoothing Attention and Frequency Attention to output level, growth, and seasonal components, achieving interpretable representations with near-linear complexity. Leddam [57] further replaces manual filters with learnable convolutional kernels and dual-attention mechanisms, enhancing adaptability in multivariate contexts. However, most models do not enforce explicit constraints such as orthogonality or semantic labeling, which limits the interpretability of their outputs.

In summary, current neural models often draw inspiration from decomposition but stop short of producing explicit, component-wise forecasts. Static filters lack adaptability, basis expansions trade off flexibility, and end-to-end learnable approaches rarely enforce interpretability guarantees. This motivates new frameworks that achieve both accuracy and semantically meaningful decomposition in multivariate, non-stationary settings.

4.2.3 Frequency-Aware and Harmonic modelling

Many real-world sequences exhibit strong periodic structure, motivating spectral modelling as a natural extension of decomposition. Existing approaches can be organised into parametric spectral models, neural architectures with frequency bias, and wavelet or multi-resolution frameworks.

Prophet [40] and Trigonometric Box-Cox ARIMA Trend Seasonal (TBATS) [27] use truncated Fourier series or trigonometric state-space models to represent seasonality. While they offer interpretability, their applicability is constrained by the need for manually specified periods and strong assumptions such as slowly evolving seasonality, and they are mainly designed for univariate data.

Early designs like Long- and Short-term Time-series Network (LSTNet) [18] and Exponential Smoothing Recurrent Neural Network (ES-RNN) [38] introduce recurrence structures that encode periodic lags or smoothed seasonal inputs. More recent models operate directly in the frequency domain: Frequency Interpolation for Time Series Forecasting (FITS) [54] interpolates signals in complex Fourier space, FEDformer attends over sampled Fourier modes, and Neural Fourier Model (NFM) [16] learn spectral coefficients end-to-end. While effective, their outputs remain latent rather than explicitly decomposed into interpretable components.

Waveformer [45] integrates the maximal overlap discrete wavelet transform with attention mechanisms to model trend and seasonal signals at multiple scales. Other designs combine Fourier and wavelet features in local windows, as in TimesNet [51] and Wavelet Frequency Transformation Network (WFTNet) [24], enhancing adaptability to nonstationarity. Meanwhile, new frequency-decomposition architectures (e.g., TimeKAN [14], FreEformer [58], Fredformer [31]) push toward cascaded, adaptive, and balanced spectral learning.

Overall, frequency-aware modelling has proven powerful for time series forecasting and beyond [20, 34, 61], even extending to large-scale language models [9]. Yet, most methods either retain frequency structure in latent space without interpretability, or depend on static bases that lack adaptability. The open challenge is to design architectures that combine adaptive spectral learning with explicit, component-wise decomposition, ensuring both predictive performance and interpretable outputs.

In conclusion, classical decomposition methods provide transparent decompositions into trend, seasonal, and residual components, but they rely on restrictive assumptions (e.g., linearity, stationarity, or fixed periodicity), which limits scalability. Neural forecasting models revisit decomposition ideas and embed them into architectures to improve flexibility, yet most rely on static filters, fixed bases, or loosely interpretable modules that do not yield explicit component-wise forecasts. Frequency-aware approaches achieve strong predictive performance and scalability by leveraging Fourier, wavelet, or multiresolution analysis,

but they often keep spectral features latent or depend on static bases without semantic interpretability.

Recent time series foundation models such as TEMPO [6], Moirai and Moirai-MoE [25], Time-MoE [52], UniTS [11] and TimeCMA [22] extend this trajectory via large-scale pretraining and prompt-based adaptation architectures. These models show strong zero-shot and few-shot generalizations across domains. Although models like TEMPO improve interpretability through explicit decomposition into trend, seasonal, and residual components, most prioritize scale and transferability, remaining opaque in their reasoning. Inference efficiency remains a bottleneck due to long-context attention and large parameter counts. Sparse mixture-of-expert designs mitigate some of this cost via expert routing but only partially reduce end-to-end computation.

Together, these lines of work highlight a persistent gap: the need for forecasting models that combine accuracy, scalability, and adaptive, component-wise interpretability. This motivates our proposed framework, which integrates decomposition and frequency learning into a unified, principled architecture.

4.3 Methodology

This section presents the iDecPred architecture. Figure 4.1 summarises the end-to-end pipeline: reversible instance normalisation (RevIN), three parallel forecast heads for trend, seasonal, and residual components, and a channel-wise gate that fuses their outputs into the final prediction.

Following the intrinsic notion outlined in the chapter introduction, iDecPred is interpretable because its forward pass exposes structured, directly inspectable quantities such as component forecasts and fusion weights (formalised below in Section 4.3.1), rather than relying on a separate post-hoc explainer. This approach provides auditability of how the model allocates mass across trend, seasonality, and residual pathways. However, this does not imply full transparency of the entire input-to-output mapping. Submodules

such as encoders, residual predictors, and the gating network remain learned and can be high-dimensional, and the decomposition is a model-induced factorisation that is not guaranteed to recover unknown latent data-generating components in real applications. Here, interpretability refers to faithfulness to the model's own factorised representation and the ability to diagnose component-wise behaviour, supported where possible by controlled synthetic experiments and forecast-based evaluation.

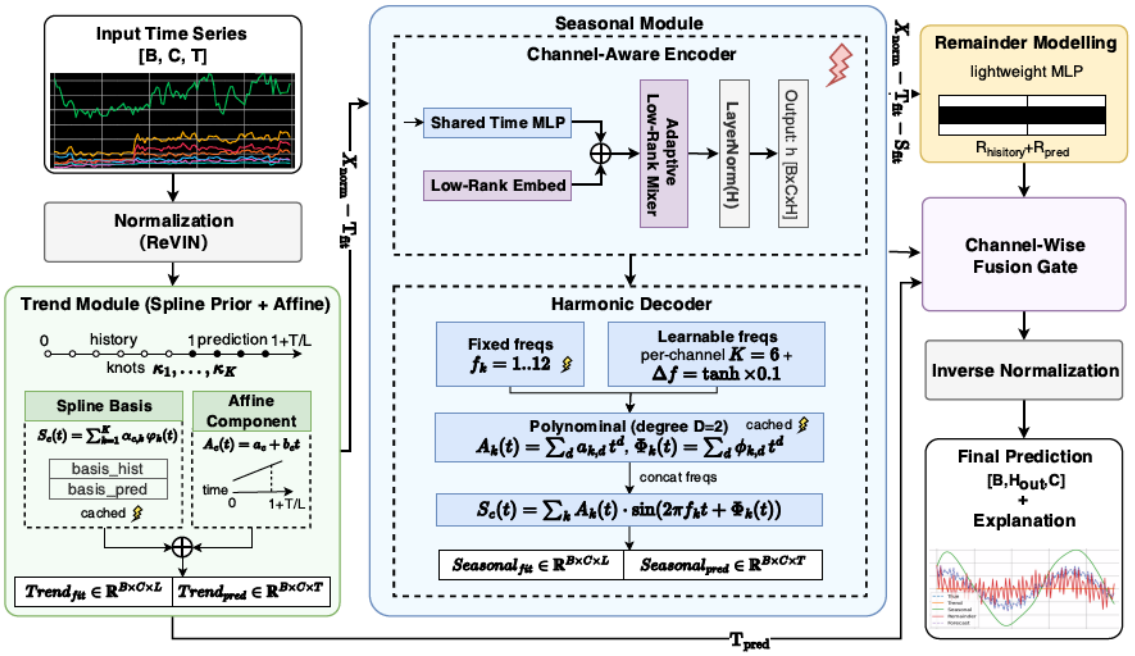


Figure 4.1: Overall architecture of iDecPred. The model normalises inputs with Reversible Instance Normalisation (ReVIN), decomposes them into trend (spline + affine), seasonal (channel-aware encoder with harmonic decoder), and residual (Multi-Layer Perceptron (MLP)) components, and fuses their forecasts via a channel-wise attention gate to produce final predictions with interpretable components.

iDecPred is proposed as a forecast-oriented neural decomposition model that explicitly separates multivariate time series into trend, seasonal, and residual components. The design emphasizes three principles: (i) interpretability, by aligning components with classical decomposition semantics and providing component-wise forecasts; (ii) adaptability, by learning decomposition and fusion jointly with the forecasting objective, enabling the model to capture dataset-specific and non-stationary dynamics; and (iii) efficiency, by employing lightweight, modular architectures that scale to high-dimensional and long-horizon forecasting tasks.

4.3.1 Problem Formulation

Let $\mathbf{X} \in \mathbb{R}^{B \times C \times L}$ denote a batch of multivariate time series with B samples, C channels, and input length L. Given a forecast horizon H, $\hat{\mathbf{Y}} \in \mathbb{R}^{B \times C \times H}$ is predicted. Instead of directly mapping $\mathbf{X} \mapsto \hat{\mathbf{Y}}$, iDecPred decomposes the prediction into three component forecasts:

$$\left\{\hat{\mathbf{Y}}^{\text{trend}}, \hat{\mathbf{Y}}^{\text{seasonal}}, \hat{\mathbf{Y}}^{\text{resid}}\right\}, \tag{4.1}$$

and then fuses them through a channel-wise gate:

$$\hat{\mathbf{Y}} = \sum_{m \in \{\text{trend}, \text{seasonal}, \text{resid}\}} \mathbf{W}^{(m)} \odot \hat{\mathbf{Y}}^{(m)}, \tag{4.2}$$

where $\mathbf{W}^{(m)} \in (0,1)^{B \times C \times H}$ are adaptive fusion weights.

4.3.2 Input Normalization

To mitigate per-channel scale variance and distributional shift, Reversible Instance Normalization (RevIN) [17] is employed. For each instance,

$$\mathbf{X}_{\text{norm}} = \frac{\mathbf{X} - \mu}{\sigma}, \quad \mu, \sigma \in \mathbb{R}^{B \times C \times 1}, \tag{4.3}$$

and the inverse transform is applied to map forecasts back to the original scale.

4.3.3 Trend Module: Compact Smooth Prior

The trend is modelled using a sum of compactly supported, twice-continuously differentiable (piecewise) radial kernels centred at K knots on a shared normalised time axis. Let the normalised history be $t \in [0,1]$ and the forecast horizon $t \in [1, 1+H/L]$. With knots $\{t_k\}_{k=1}^{K}$ placed uniformly on $[0,\, 1+H/L]$ and

$$b_k(t) = \left(1 - |t - t_k|^3\right)_+^3, \tag{4.4}$$

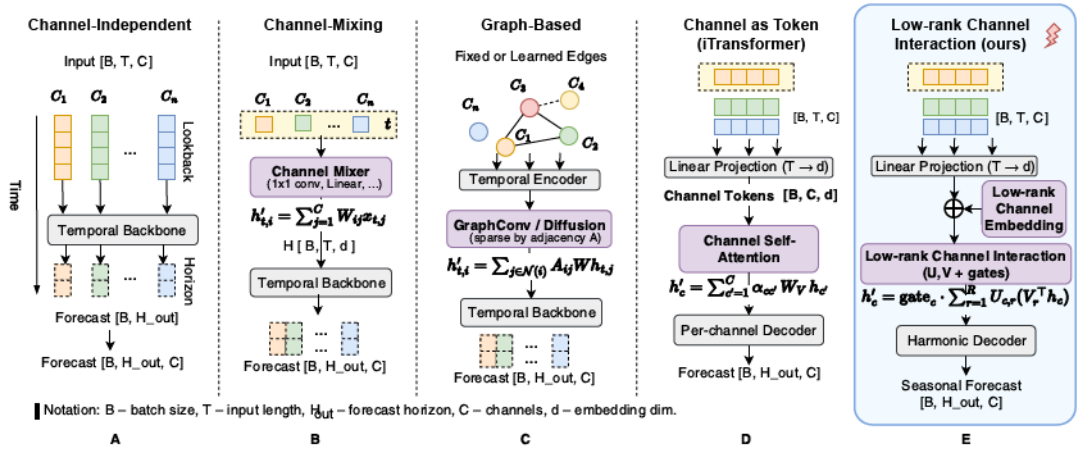


Figure 4.2: Comparison of paradigms for multivariate time series forecasting (Channel-Independent (CI), Channel-Mixing (CM), Graph-Based (GB), Channel-as-Token (CaT)) and the proposed Low-rank Channel Interaction (LCI). Notation: B batch size; L input (look-back) length; H ($\equiv H_{\text{out}}$) forecasting horizon; C number of channels; d embedding dimension. For CM, W denotes dense channel-mixing weights; for GB, A denotes the adjacency with neighborhood $\mathcal{N}(i)$; for CaT, $\alpha_{cc'}$ are attention weights; for LCI, U, V are low-rank channel factors and gate_c is an adaptive gating coefficient.

the per-channel trend is

$$\tau_c(t) = \sum_{k=1}^{K} \alpha_{c,k}\, b_k(t) + a_c + b_c\, t, \tag{4.5}$$

where $\alpha_{c,k}, a_c, b_c$ are learned per-channel parameters shared across history and horizon to ensure temporal continuity. The basis offers smoothness, locality (compact support), and interpretability via knot-wise coefficients.

4.3.4 Seasonal Module: Channel-Aware Encoder and Harmonic Decoder

Seasonal patterns are modelled by a Channel-Aware Encoder followed by a Fourier-based Harmonic Decoder. The encoder produces latent representations that preserve temporal dynamics while adaptively capturing cross-channel dependencies. The decoder reconstructs oscillatory components through harmonics with fixed and learnable frequencies.

4.3.4.1 Channel-Aware Encoder.

Let $\mathbf{X}_{\text{resid}} \in \mathbb{R}^{B\times C\times L}$ denote the detrended input (same axes as $\mathbf{X}$ in Section 4.3.1). A shared two-layer MLP is first applied per channel:

$$\tilde{\mathbf{h}}_c = f_{\text{shared}}(\mathbf{x}_c) \in \mathbb{R}^d, \tag{4.6}$$

yielding $\tilde{\mathbf{H}} \in \mathbb{R}^{B\times C\times d}$.

A low-rank channel embedding is added to inject channel-specific bias:

$$\mathbf{E} = \mathbf{E}_1\mathbf{E}_2, \qquad \mathbf{E}_1 \in \mathbb{R}^{C\times r},\ \mathbf{E}_2 \in \mathbb{R}^{r\times d},\ r \ll d, \tag{4.7}$$

and update each representation as

$$\mathbf{h}_c = \tilde{\mathbf{h}}_c + \mathbf{E}[c,:]. \tag{4.8}$$

Stacking over channels yields $\mathbf{H} = [\mathbf{h}_1;\ldots;\mathbf{h}_C] \in \mathbb{R}^{B\times C\times d}$.

To model inter-channel dependencies, channels are summarised by averaging over the embedding dimension,

$$\mathbf{s} = \frac{1}{d}\sum_{\ell=1}^{d} \mathbf{H}[:,:,\ell] \in \mathbb{R}^{B\times C}. \tag{4.9}$$

With $\mathbf{U} \in \mathbb{R}^{C\times r}$ and $\mathbf{V} \in \mathbb{R}^{r\times C}$, low-rank channel mixing is realised through *batched matrix multiplication* on channel–embedding slices, avoiding tensor mode subscript notation that readers may find unfamiliar. Concretely, let $\mathbf{H}_b \in \mathbb{R}^{C\times d}$ denote the b-th batch slice of $\mathbf{H}$ (rows index channels, columns the embedding). Then $\mathbf{U}^\top \in \mathbb{R}^{r\times C}$ maps channel space to rank r, and $\mathbf{V}^\top \in \mathbb{R}^{C\times r}$ maps back:

$$\begin{aligned} \mathbf{Z}_b &= \mathbf{U}^\top\mathbf{H}_b \in \mathbb{R}^{r\times d}, \\ \mathbf{G}_{r,b} &= g_r(\mathbf{s})_{b,:}\,\mathbf{1}_d^\top \in \mathbb{R}^{r\times d}, \\ \mathbf{P}_b &= \mathbf{V}^\top\big(\mathbf{Z}_b \odot \mathbf{G}_{r,b}\big) \in \mathbb{R}^{C\times d}, \end{aligned} \tag{4.10}$$

where $g_r(\mathbf{s}) = \text{softplus}(\text{fc}_r(\mathbf{s}) + \alpha_0) \in \mathbb{R}^{B\times r}$ and $\mathbf{1}_d$ is a d-vector of ones, and $b = 1,\ldots,B$ with $\{\mathbf{Z}_b\}$, $\{\mathbf{G}_{r,b}\}$, and $\{\mathbf{P}_b\}$ stacking to $\mathbf{Z}, \mathbf{G}_r \in \mathbb{R}^{B\times r\times d}$ and $\mathbf{P} \in \mathbb{R}^{B\times C\times d}$. The same contractions are written compactly as `einsum`:

$$\begin{aligned}\mathbf{Z} &= \text{einsum}('bcd, rc \to brd',\ \mathbf{H},\ \mathbf{U}^\top),\\ \mathbf{P} &= \text{einsum}('brd, cr \to bcd',\ \mathbf{Z}\odot\mathbf{G}_r,\ \mathbf{V}^\top).\end{aligned} \tag{4.11}$$

$$g_c(\mathbf{s}) = \text{softplus}\Big(\text{fc}_{c,2}(\text{GELU}(\text{fc}_{c,1}(\mathbf{s}))) + \alpha_0\Big) \in \mathbb{R}^{B\times C}. \tag{4.12}$$

To avoid batch-dependent diagonals, the batch mean is used $\bar{g}_c = \frac{1}{B}\sum_{b=1}^{B} g_c(\mathbf{s})_{b,:} \in \mathbb{R}^C$. The encoder output is

$$\mathbf{H}' = \text{LayerNorm}\Big(\mathbf{H} + \text{Dropout}(\mathbf{P}) \odot \bar{g}_c[:, \text{None}]\Big), \tag{4.13}$$

with $\bar{g}_c[:, \text{None}] \in \mathbb{R}^{C\times 1}$ broadcast across batch and embedding dimensions.

As illustrated in Figure 4.2, our Low-rank Channel Interaction (LCI) module differs from prior designs. Specifically, Channel-Independent (CI) and Channel-Mixing (CM) either ignore or rigidly mix channels, GB relies on fixed/learned adjacency structures, and Channel-as-Token (CaT) employs costly self-attention over channels. In contrast, LCI introduces low-rank channel embeddings with adaptive gating, achieving linear complexity while retaining interpretability.

Complexity. Each batched projection $\mathbf{U}^\top\mathbf{H}_b$ or $\mathbf{V}^\top(\cdot)$ costs $\mathcal{O}(Cdr)$ per batch index, hence $\mathcal{O}(BCdr)$ overall; the Hadamard gating adds $\mathcal{O}(BCr)$. Thus the overall per-layer complexity is $\mathcal{O}(BCdr)$, linear in both C and d for fixed r. Parameter count is $\mathcal{O}(Cr)$ for $(\mathbf{U},\mathbf{V})$, plus $\mathcal{O}(Cr + rd)$ if low-rank channel embeddings are included.

Interpretability. The gating mechanism provides explicit, data-dependent control over both channel importance and interaction structure: $g_c(\mathbf{s})$ scales each channel's contribution, while $g_r(\mathbf{s})$ dynamically modulates the low-rank mixer $(\mathbf{UV})$. The instantaneous inter-

channel influence is given by

$$\mathbf{M}(\mathbf{s}) = \mathrm{diag}\big(\bar{g}_c(\mathbf{s})\big)\,(\mathbf{UV}), \tag{4.14}$$

where the row-wise magnitude of $\mathbf{M}(\mathbf{s})$ reflects dominant channel contributors under each input.

4.3.4.2 Harmonic Decoder.

Each $\mathbf{h}'_c$ maps to amplitudes and phases of harmonics. $K_{\text{fix}} = 12$ fixed frequencies $\Omega_{\text{fix}} = \{1,\ldots,12\}$ (measured in cycles per normalised window) and $K_{\text{learn}} = 6$ learnable, channel-specific frequencies $\tilde{\Omega}_c$ are used. Time is normalised on the shared axis (history $[0,1]$, horizon $[1,1+H/L]$) and augmented with polynomial bases $\{t^d\}_{d=0}^{2}$. Linear projections from $\mathbf{h}'_c$ produce $\{A_{c,k,d}, \Phi_{c,k,d}\}$, and the forecast takes the form

$$\begin{aligned}\hat{y}_{c,\tau} = \sum_{k=1}^{K_{\text{fix}}+K_{\text{learn}}} \Big(\sum_{d=0}^{2} A_{c,k,d}\, t_\tau^d\Big)\, \sin\Big(2\pi\, \omega_{c,k}\, t_\tau \\ + \sum_{d=0}^{2} \Phi_{c,k,d}\, t_\tau^d\Big),\end{aligned} \tag{4.15}$$

with $\omega_{c,k} \in \Omega_{\text{fix}} \cup \tilde{\Omega}_c$. Optionally, a small MLP predicts $\Delta\omega_{c,m} \in [-0.1, 0.1]$ to refine learnable frequencies.

4.3.5 Residual Module

Irregular variations not captured by the trend or seasonal modules are modelled by a lightweight MLP. Specifically, the residual component is obtained from the detrended input after subtracting the reconstructed trend and seasonal parts:

$$\hat{\mathbf{Y}}^{\text{resid}} = \mathrm{MLP}\Big(\mathbf{X}_{\text{resid}} - \hat{\mathbf{X}}^{\text{trend}} - \hat{\mathbf{X}}^{\text{seasonal}}\Big). \tag{4.16}$$

This residual path is simple yet effective, avoiding overfitting while providing robustness to unpredictable fluctuations.

Table 4.1: Statistics of real-world benchmark datasets.

Dataset	ETTh	ETTm	Traffic	Electricity	Weather
Variates	7	7	862	321	21
Timesteps	17,420	69,680	17,544	26,304	52,696
Granularity	1h	15min	1h	1h	10min

4.3.6 Fusion Module: Adaptive Channel-Wise Gate

The three component forecasts are combined by a channel-wise fusion gate. Given the three component forecasts $\hat{\mathbf{Y}}^{\text{trend}}$, $\hat{\mathbf{Y}}^{\text{seasonal}}$, and $\hat{\mathbf{Y}}^{\text{resid}}$, a self-attention mechanism summarises cross-component interactions and outputs weights $\mathbf{W}^{(m)} \in (0,1)^{B\times C\times H}$ via a sigmoid, with $m \in \{\text{trend}, \text{seasonal}, \text{resid}\}$. The final prediction is

$$\hat{\mathbf{Y}} = \sum_{m\in\{\text{trend},\,\text{seasonal},\,\text{resid}\}} \mathbf{W}^{(m)} \odot \hat{\mathbf{Y}}^{(m)}. \tag{4.17}$$

Overall, the methodology of iDecPred consists of: (i) a compact cubic-kernel trend prior for interpretable extrapolation, (ii) a novel channel-aware seasonal module that combines fixed and learnable Fourier harmonics with input-adaptive gating, (iii) a lightweight residual MLP for irregular variations, and (iv) an adaptive channel-wise fusion gate for data-dependent integration. Together, these modules form a coherent, interpretable, and high-performing forecasting framework that is distinct from prior decomposition-based approaches.

4.4 Experimental Setup

4.4.1 Real-World Forecasting Benchmarks

To evaluate forecasting performance under practical conditions, iDecPred is benchmarked on widely used multivariate time series datasets that span diverse application domains, including energy, weather, and transportation. These datasets are standard in long-term forecasting research [63, 50], and their statistics are summarised in Table 4.1.

4.4.2 Synthetic Datasets for Decomposition Evaluation

Multivariate time series with an additive latent structure are generated to assess component-wise interpretability and disentanglement. As shown in Figure 4.3, this controlled generative design exposes ground-truth trend, seasonal, and noise factors, enabling direct, component-level evaluation and quantification of how decomposition fidelity relates to forecast error across channel dimensionalities and sequence lengths. For C channels and N time steps on a shared grid, the observed signal is

$$X(t) \;=\; W_1\, u_{\text{trend}}(t) \;+\; W_2\, u_{\text{seasonal}}(t) \;+\; W_3\, u_{\text{noise}}(t), \tag{4.18}$$

where $u_k(t) \in \mathbb{R}^{D_k}$ are latent bases for component $k \in \{\text{trend}, \text{seasonal}, \text{noise}\}$, and $W_k \in \mathbb{R}^{C \times D_k}$ are component-specific projection matrices. All components share the same time grid of length N and combine linearly as in (4.18). Here N counts time indices along the synthetic trajectory and is distinct from the model input look-back length L in Section 4.3.1.

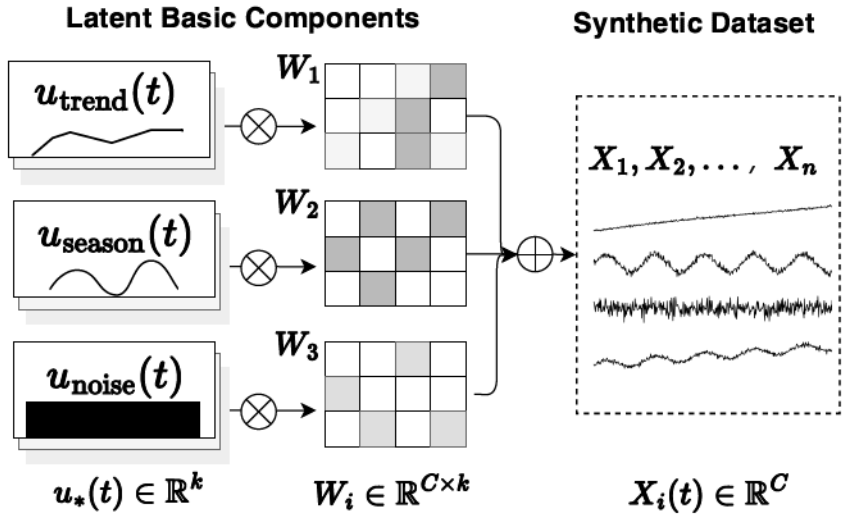


Figure 4.3: Synthetic Data Generation Pipeline. Latent trend/seasonal/noise bases are projected by W_k and summed to form $X(t)$. Optional local, causal, and calendar effects as well as anomalies and missingness can be injected to increase difficulty.

Latent components. Trend bases include linear, polynomial, piecewise, sigmoid, and random-walk processes (smoothed to avoid edge bias) [10, 8]. Seasonal bases include harmonic expansions and multi-sinusoids (with optional slow phase drift and amplitude

modulation), and square waves [27, 15]. Noise is Gaussian or AR(1) with controllable variance (heteroskedastic or periodically modulated). For identifiability, the seasonal bases are orthogonalised against the trend bases (Gram–Schmidt) and re-standardised (trend retains DC; seasonal is zero-mean) [8].

Structure matrices and entanglement. Each W_k is row-normalised (L1) to control the contribution per-channel and is sampled from simple structural families: block (disjoint groups), full (dense), sparse (Bernoulli mask), identity (one-to-one) or degenerate_full (all ones). Varying W_k modulates cross-channel correlation and component entanglement.

Optional perturbations. (i) channel-wise low-order polynomials, (ii) a single-lag causal pathway from a designated source channel, and (iii) calendar modulations (weekday/month/hour) are optionally added to Complex datasets. Point anomalies and missing values are also injected at controlled rates. These options stress-test robustness without changing the additive generative form (4.18).

Synthetic Dataset families. Four synthetic data families are instantiated: (i) *trend-only*, (ii) *seasonal-only*, (iii) *mixed-balanced* (with comparable signal-to-noise ratios across trend and seasonal components), and (iv) *complex* (combining trend and seasonal signals with optional local, causal, calendar, and anomaly effects). Each family is generated at multiple scales with channel size $C \in \{7, 21, 321, 862\}$ and sequence length up to $N = 20{,}000$, matching the dimensional ranges used in common multivariate forecasting benchmarks [51]. This multi-scale design enables evaluation of model scalability and cross-channel interpretability. Fixed random seeds are used to ensure exact reproducibility. In the following, the datasets are indicated by T_xx (trend-only), S_xx (seasonal-only), indicated by B_xx (mixed-balanced) and C_xx (complex), where the suffix "xx" indicates the number of channels.

Outputs. Each dataset release contains three components:

- the observed multivariate sequence $X \in \mathbb{R}^{C \times N}$;
- ground-truth sources, including X_{trend}, X_{seasonal}, and X_{noise}, and (when enabled) auxiliary terms such as X_{local}, X_{causal}, and X_{calendar};

- the corresponding projection matrices W_1, W_2, and W_3 that map latent bases to the observed channels.

This organization enables controlled evaluation of both reconstruction fidelity and component disentanglement.

4.4.3 Training and Evaluation Setup

4.4.3.1 Learning Objectives.

Following the xPatch framework [39], iDecPred is trained with a horizon-weighted MAE loss. For horizon step $i \in \{1, \ldots, H\}$ the weight is

$$w_i = -\arctan(i) + \frac{\pi}{4} + 1, \tag{4.19}$$

which emphasizes short-term accuracy while gradually down-weighting longer horizons. The objective is

$$\mathscr{L}_{\text{xpatch}} = \frac{1}{BHC} \sum_{b=1}^{B} \sum_{i=1}^{H} \sum_{c=1}^{C} w_i \left| \hat{Y}_{b,i,c} - Y_{b,i,c} \right|. \tag{4.20}$$

This preserves the robustness of MAE while incorporating horizon awareness, and is applied consistently for training, validation, and testing.

4.4.3.2 Evaluation Metrics.

Forecasting accuracy is assessed with standard Mean Squared Error (MSE) and Mean Absolute Error (MAE), averaged across output channels and prediction horizons unless otherwise noted.

To evaluate interpretability on synthetic datasets, the alignment between the predicted and ground-truth components is further measured. Specifically, the Pearson correlation is reported to assess shape alignment and the MAE to assess scale-invariant reconstruction accuracy. These metrics provide a structured evaluation of iDecPred's ability to disentangle and recover latent components.

4.4.3.3 Baselines.

iDecPred is benchmarked against six representative forecasting models on real-world datasets: TimeMixer [43], iTransformer [26], PatchTST [29], DLinear [59], TimesNet [51], and FEDformer [64]. On synthetic datasets, TimeXer [44] and N-BEATS [30] are additionally included, together with TimeMixer, iTransformer, and PatchTST.

4.4.3.4 Implementation Details.

Training follows the standard long-term forecasting protocol adopted in recent work [43, 51]. All models are trained with an input (look-back) length of $L = 96$ and evaluated on forecast horizons $H \in \{96, 192, 336, 720\}$. To examine robustness to temporal context, extended experiments are also conducted with longer look-back windows ($L \in \{192, 336, 720\}$) on the Electricity, and Traffic datasets. Unless otherwise specified, the reported results correspond to the 96–96 setting.

For iDecPred, the embedding dimension is fixed at $d_{\text{model}} = 128$ across all datasets.

All experiments are conducted on a single NVIDIA A100-40GB GPU with fixed random seeds to ensure reproducibility.

4.5 Experiment Results and Discussion

This section presents a comprehensive evaluation of iDecPred on both real-world and synthetic datasets. Forecasting accuracy, decomposition components alignment, frequency decomposition quality, and computational efficiency are assessed, and ablation studies are performed to isolate the impact of core architectural choices.

4.5.1 Forecasting Accuracy on Real-World Datasets

Table 4.2 reports multivariate forecasting results at look-back L=96 and horizon H=96. The best results are highlighted in bold, while the second-best are underlined. iDecPred

delivers the best or competitive performance in all seven benchmarks, achieving the highest score in 9 of 14 cells in the metrics.

On the ETT variants, iDecPred leads on both metrics for ETTm1 and ETTm2 and secures the best MAE on ETTh1 and ETTh2. On ETTh1, it attains the lowest MAE and is competitive in MSE (0.378), slightly behind TimeMixer (0.375). On ETTh2, it achieves the best MAE and the second-best MSE (0.291 vs. 0.289).

On Weather and Electricity, two datasets with higher seasonality and stationarity, iDecPred demonstrates strong performance. It achieves the best results on Weather in both MSE and MAE, outperforming the next best model by approximately 7% in MSE. On Electricity, it attains the lowest MAE and the second-best MSE, yielding a 3% MAE improvement over the next best baseline, iTransformer.

Traffic data remains the most challenging due to its low stationarity, high stochasticity, and high dimensionality. iTransformer achieves the best performance (0.395/0.268), while iDecPred remains competitive with an MSE of 0.456 and MAE of 0.289. Notably, iDecPred maintains a compact parameter footprint: it uses only about 0.65 million parameters. In contrast, iTransformer requires 6.4 million parameters under the same setting, highlighting the efficiency of our modular design for high-dimensional forecasting tasks.

Across diverse real-world benchmarks, iDecPred achieves state-of-the-art (SOTA) or second-best results, offering a strong balance between accuracy and parameter efficiency. Its decomposition-aware architecture performs particularly well at look-back length (L=96) and forecast horizon (H=96), generalizing effectively across datasets with varying seasonality, stationarity, and dimensionality.

4.5.2 Impact of Input Lookback Length

To evaluate how temporal context affects forecasting performance in high-dimensional settings, iDecPred's behaviour on the Electricity and Traffic datasets is analysed by varying the input (look-back) length $L \in \{96, 192, 336, 720\}$. Results across short- and long-horizon settings ($H = 96, 720$) are visualised in Figure 4.4.

Table 4.2: Multivariate forecasting results on prediction horizon 96. The input sequence length is fixed at 96 for all models. Lower MSE/MAE is better.

	iDecPred (ours)		TimeMixer (2024)		iTransformer (2024)		PatchTST (2023)		TimesNet (2023)		DLinear (2023)		FEDformer (2022)	
Dataset	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**
ETTh1	0.378	**0.388**	**0.375**	0.400	0.386	0.405	0.414	0.419	0.384	0.402	0.386	0.400	0.376	0.410
ETTh2	0.291	**0.337**	**0.289**	0.341	0.297	0.349	0.302	0.348	0.340	0.374	0.333	0.387	0.358	0.397
ETTm1	**0.314**	**0.340**	0.320	0.357	0.334	0.368	0.329	0.367	0.338	0.375	0.345	0.372	0.379	0.419
ETTm2	**0.174**	**0.251**	0.175	0.258	0.180	0.264	0.175	0.259	0.187	0.267	0.193	0.292	0.203	0.287
Weather	**0.151**	**0.189**	0.163	0.209	0.174	0.214	0.177	0.218	0.172	0.220	0.196	0.255	0.217	0.296
Electricity	0.149	**0.239**	0.153	0.247	**0.148**	0.240	0.181	0.270	0.168	0.272	0.197	0.282	0.193	0.308
Traffic	0.456	0.289	0.462	0.285	**0.395**	**0.268**	0.462	0.295	0.617	0.336	0.650	0.396	0.587	0.366

It is observed that increasing the input look-back length consistently improves performance across both datasets and forecasting horizons. On Electricity, for instance, at $H = 96$, the MSE decreases by 7.4% and the MAE by 2.1% as L increases from 96 to 720. At $H = 720$, the MSE drops by 8.6% and the MAE by 3.7%.

On Traffic, the gains are more pronounced. With a forecasting horizon of $H = 96$, extending L from 96 to 720 reduces the MSE from 0.464 to 0.391 and the MAE from 0.309 to 0.283, corresponding to relative improvements of 15.8% in MSE and 8.3% in MAE. Similar trends are observed for longer horizons on the Traffic dataset.

These results confirm that access to longer temporal context benefits the model's dynamic prediction capabilities, especially in complex, high-dimensional regimes like Traffic.

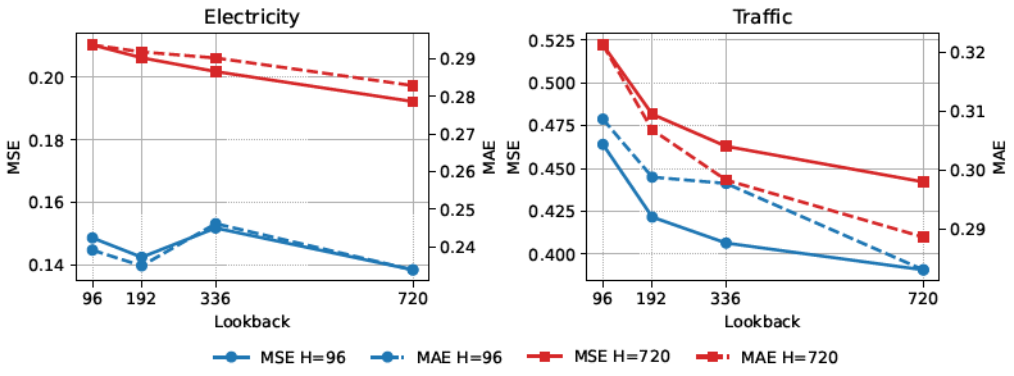


Figure 4.4: iDecPred performance on forecast horizons 96 and 720 with varying look-back lengths $L \in \{96, 192, 336, 720\}$ on the Electricity and Traffic.

Figure 4.4 summarises forecast error as a function of look-back length L under the

protocol in Section 4.4. In light of the iDecPred design, the pattern reflects a standard trade-off between richer temporal context and forward-pass cost: the seasonal and residual branches encode the full L-step history, so memory and arithmetic scale approximately linearly with L and the trained models are not parameter-matched across window lengths. The subsequent low-rank cross-channel mixing depends primarily on channel width and rank, without an additional cost component that scales comparably with further increases in L.

4.5.3 Ablation Study

To evaluate the contribution of individual components, ablation experiments are conducted in which the channel-aware encoder (CAE), cross-channel mixer, embedding layer, and fusion block are selectively removed. The results, shown in Table 4.3, demonstrate that the CAE and Mixer are the most essential modules: removing either one consistently causes noticeable and sometimes severe performance declines, especially on datasets with strong cross-channel dependencies such as Electricity and Traffic. The embedding layer is also important, particularly for ETTm1 and ETTm2, where its removal results in marked decreases in performance. In comparison, the fusion block has a relatively minor effect, with small changes in results when it is omitted. These findings confirm that the design choices of including the CAE and Mixer are crucial for achieving robust and accurate long-term forecasting.

4.5.4 Forecasting Performance on Synthetic Datasets

Forecasting performance on synthetic datasets with controlled additive structures is further evaluated, spanning multiple channel configurations and four representative flavors, as introduced in Section 4.4.2. In total, experiments are conducted on all 16 datasets derived from these configurations.

Table 4.3: Ablation study results of the proposed model on seven benchmark datasets after removing individual components, including the Channel-Aware Encoder (CAE), Cross-Channel Mixer, Embedding Layer, and Fusion Block

Horizon	Ablation	ETTh1		ETTh2		ETTm1		ETTm2		Weather		Electricity		Traffic	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
96–96	w/o CAE	0.38	0.389	0.292	0.336	0.311	0.338	0.172	0.249	0.164	0.201	0.165	0.249	0.501	0.307
	w/o mixer	0.379	0.389	0.295	0.343	0.31	0.338	0.174	0.251	0.162	0.198	0.163	0.248	0.493	0.303
	w/o embed	0.376	0.388	0.288	0.334	0.311	0.341	0.174	0.251	0.15	0.188	0.149	0.24	0.455	0.29
	w/o fusion	0.381	0.389	0.297	0.34	0.306	0.336	0.174	0.25	0.149	0.188	0.153	0.244	0.482	0.306
	full	0.378	0.388	0.291	0.337	0.314	0.34	0.174	0.251	0.151	0.189	0.149	0.239	0.456	0.289
96–192	w/o CAE	0.439	0.424	0.37	0.386	0.365	0.366	0.236	0.292	0.207	0.241	0.176	0.258	0.499	0.307
	w/o mixer	0.439	0.424	0.366	0.384	0.364	0.366	0.237	0.292	0.208	0.241	0.175	0.258	0.499	0.305
	w/o embed	0.447	0.424	0.363	0.382	0.365	0.365	0.243	0.294	0.2	0.236	0.167	0.255	0.472	0.297
	w/o fusion	0.438	0.421	0.369	0.387	0.36	0.364	0.239	0.294	0.2	0.237	0.169	0.258	0.494	0.31
	full	0.438	0.421	0.372	0.39	0.365	0.365	0.242	0.294	0.201	0.237	0.166	0.254	0.468	0.294
96–336	w/o CAE	0.479	0.444	0.419	0.422	0.396	0.387	0.296	0.331	0.264	0.283	0.192	0.275	0.512	0.311
	w/o mixer	0.481	0.446	0.412	0.421	0.399	0.389	0.295	0.331	0.264	0.284	0.192	0.275	0.511	0.31
	w/o embed	0.485	0.446	0.427	0.424	0.397	0.388	0.298	0.331	0.259	0.28	0.181	0.27	0.49	0.303
	w/o fusion	0.483	0.448	0.42	0.425	0.394	0.387	0.3	0.334	0.257	0.28	0.193	0.282	0.508	0.317
	full	0.486	0.447	0.426	0.426	0.397	0.388	0.297	0.332	0.259	0.282	0.182	0.272	0.482	0.303
96–720	w/o CAE	0.47	0.456	0.439	0.448	0.463	0.427	0.418	0.406	0.351	0.342	0.235	0.309	0.545	0.327
	w/o mixer	0.473	0.462	0.437	0.446	0.465	0.428	0.408	0.398	0.359	0.349	0.235	0.309	0.551	0.33
	w/o embed	0.492	0.472	0.431	0.442	0.558	0.482	0.417	0.404	0.35	0.341	0.213	0.295	0.526	0.322
	w/o fusion	0.485	0.464	0.434	0.442	0.467	0.428	0.395	0.39	0.341	0.336	0.225	0.307	0.552	0.338
	full	0.487	0.469	0.434	0.447	0.469	0.433	0.408	0.398	0.368	0.356	0.21	0.294	0.522	0.321
Performance drop (%) relative to full model															
w/o CAE		-4.22	-2.25	-0.83	-1.85	-2.45	-1.78	-1.17	**0.57**	**8.87**	**4.66**	**34.66**	**11.96**	**27.18**	**14.85**
w/o mixer		-3.24	-0.75	-2.91	-1.20	-1.81	-1.55	-2.70	-1.13	**10.34**	**5.57**	**32.28**	**11.43**	**26.19**	**13.46**
w/o embed		**2.34**	**1.12**	-3.85	-4.33	**17.97**	**11.59**	**3.02**	**1.01**	-6.03	-5.48	**1.26**	**0.17**	**3.10**	**1.36**
w/o fusion		-0.16	-0.39	-0.03	-1.19	-5.02	-2.75	-3.44	-1.81	-9.46	-6.66	**17.65**	**11.91**	**22.33**	**20.84**

4.5.4.1 Overall Accuracy.

Figure 4.5 illustrates the average radar plot of forecasting performance. Curves closer to the boundary indicate superior performance. Notably, iDecPred achieves the best overall accuracy, with the lowest average test MSE (0.133) and MAE (0.211), outperforming strong baselines such as PatchTST (MSE 0.143, MAE 0.230) and TimeMixer (MSE 0.155, MAE 0.235). In contrast, N-BEATS yields considerably higher errors (MSE 0.636, MAE 0.457): while its performance remains acceptable on datasets dominated by trend or seasonal components, it is notably weaker than other models on balanced and complex datasets.

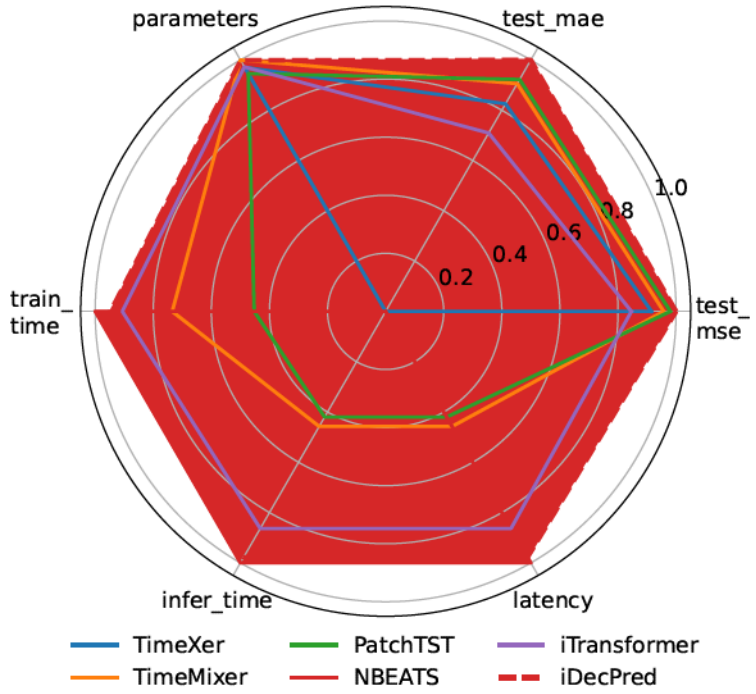


Figure 4.5: Average radar chart of forecasting performance across 16 synthetic datasets under the 96−96 setting. Curves closer to the boundary indicate better performance.

4.5.4.2 Computational Efficiency.

Besides accuracy, computational efficiency is further compared in terms of training cost, inference latency, and model compactness. iDecPred strikes an excellent balance: with only 2.7×10^5 parameters, it trains in 121s on average and achieves the lowest inference latency

(0.175ms per window), which is 10× faster than PatchTST and 5× faster than TimeMixer. This compact design makes iDecPred particularly suitable for real-time forecasting under resource-constrained environments.

4.5.4.3 Case Study on Complex Dataset.

Efficiency–accuracy trade-offs are further investigated on the most challenging C_862 dataset under both short-horizon (L=96, H=96) and long-horizon (L=96, H=720) settings. Results are summarised in Figure 4.6 and Figure 4.7.

For the short-horizon case, iDecPred yields the lowest MSE and MAE, while requiring only 4.4×10^5 parameters and achieving sub-millisecond inference latency (0.38ms). In contrast, large models such as TimeXer and PatchTST require 20× higher inference costs with inferior accuracy.

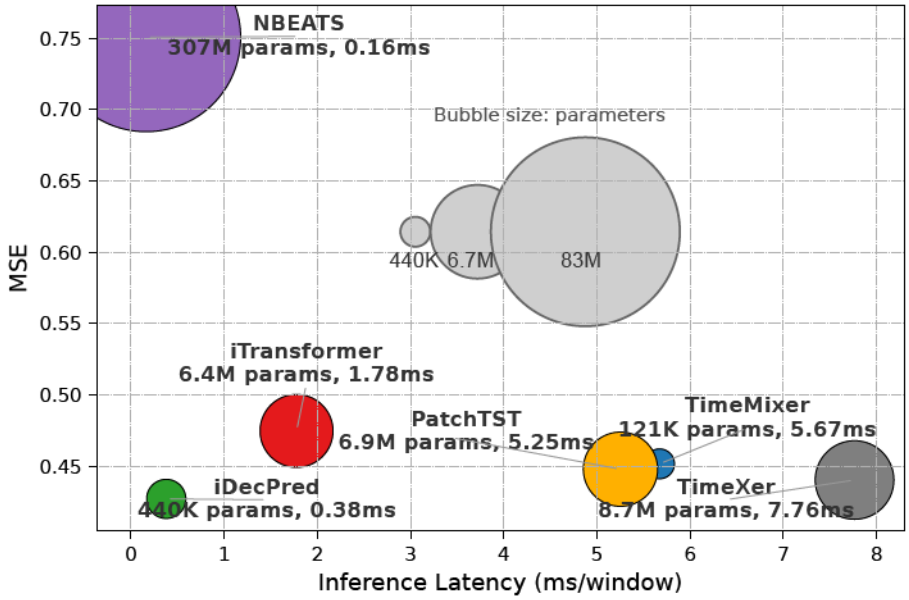


Figure 4.6: Efficiency–accuracy trade-offs on the C_862 dataset (L=96, H=96). Bubble size denotes parameter count.

Under the long-horizon setting (H=720), iDecPred attains MSE 0.498 and MAE 0.518, remaining competitive with PatchTST (0.483, 0.509) while outperforming iTransformer (0.519, 0.531). In addition, iDecPred achieves a 4× reduction in inference cost (1.14ms vs. 5.19ms). These results underscore the method's robustness across horizons and its

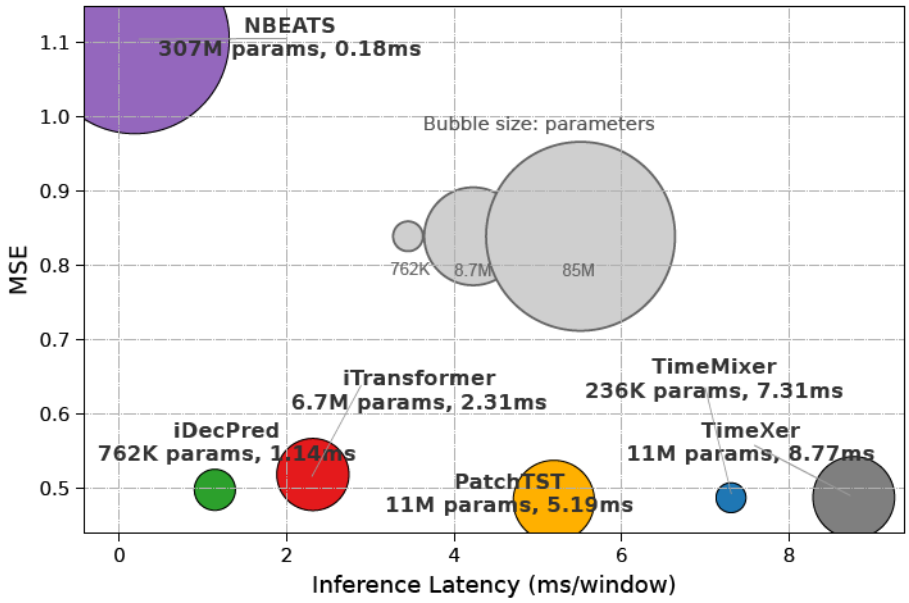


Figure 4.7: Efficiency–accuracy trade-offs on the C_862 dataset (L=96, H=720). Bubble size denotes parameter count.

scalability to large forecasting windows.

Across short and long horizons, iDecPred attains strong accuracy–latency trade-offs, outperforming heavier baselines while remaining robust.

4.5.5 Interpretability Evaluation

4.5.5.1 Decomposition on Real-World Data

The model outputs trend and seasonal components during prediction, enabling direct inspection without ground-truth decompositions. This is illustrated on Electricity (Figure 4.8).

Figure 4.8 provides a qualitative case illustration based on the same Electricity benchmark as the main forecasting experiments. The panels present, for selected channels and horizons, the model's predicted trend, seasonal, and combined trajectories alongside the observed series. The intention is not to demonstrate numerical recovery of unknown latent components, as no ground-truth decomposition is available in this setting. Instead, this illustration highlights that iDecPred's forward pass directly yields structured, component-level outputs, allowing an analyst to inspect these elements together with the final prediction. In

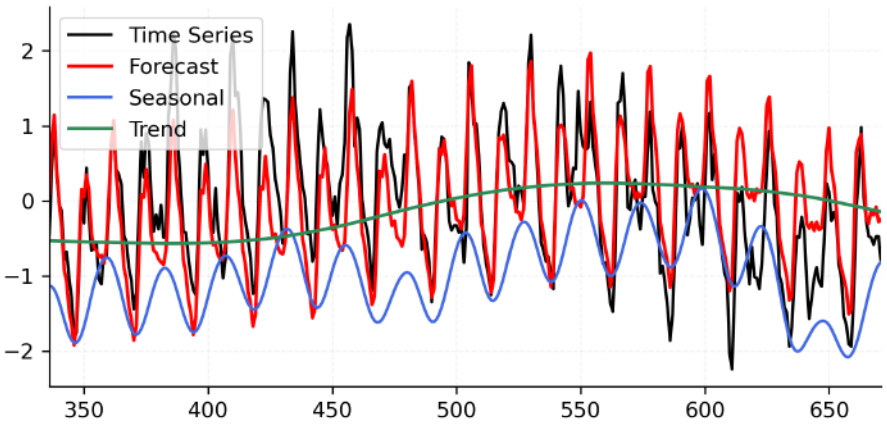


Figure 4.8: iDecPred decomposition on the Electricity dataset. The model jointly outputs trend, seasonal and final forecast components, enabling interpretable inspection on real-world data without access to ground-truth decompositions.

this visualization, the trend branch captures the smooth low-frequency structure, while the seasonal branch models higher-frequency, quasi-periodic variations. The plotted forecast shows the integrated result, as described in Section 4.3.1.

It documents qualitative behaviour on the Electricity benchmark used elsewhere in this chapter, without relying on post-hoc explainers. These serve as concrete examples of the decomposition interface promised by the architecture.

4.5.5.2 Ground-Truth Evaluation on Synthetic Data

Table 4.4 reports results across four synthetic dataset families: trend-only (T), seasonal-only (S), mixed-balanced (B), complex (C).

On the trend-only datasets, iDecPred achieves near-perfect alignment, with Pearson correlations approaching 1.0 and extremely low MAE, clearly outperforming N-BEATS.

On the seasonal-only datasets, iDecPred also achieves very high correlations (S_7: 0.955 vs. 0.895), indicating excellent phase alignment with the ground truth. However, its seasonal MAE is sometimes larger (S_7: 0.565 vs. 0.398), suggesting deviations in amplitude despite correct timing. This trade-off is consistent with the model design, where trend and seasonal components are fused through learnable weights that may attenuate seasonal magnitude.

On the mixed-balanced datasets, iDecPred again outperforms N-BEATS in trend modelling (e.g., B_21: MAE 0.263 vs. 0.740; correlation 0.483 vs. –0.010) and achieves higher seasonal correlations (0.872 vs. 0.733 in B_21), though with larger seasonal MAE. Nevertheless, in terms of the final reconstruction error, iDecPred consistently yields much lower MAE (e.g., B_7: 0.107 vs. 0.398).

On the complex datasets, which combine trend, seasonal, and additional local or causal variations, iDecPred maintains strong trend correlations (e.g., C_21: 0.917) but again shows larger seasonal MAE than N-BEATS (1.412 vs. 0.221). Importantly, the final MAE remains substantially lower for iDecPred (e.g., C_21: 0.530 vs. 0.916), demonstrating robustness under more challenging signal compositions.

Table 4.4: Ground-truth evaluation on synthetic datasets. Metrics include Pearson correlation, trend/seasonal MAE, and final MAE. iDecPred* denotes weighted trend/seasonal reconstructions from the same iDecPred model. Boldface indicates the best result.

Model	Dataset	Trend Pearson	Trend MAE	Seasonal Pearson	Seasonal MAE	Final MAE
iDecPred	C_7	**0.388**	0.911	**0.645**	1.374	**0.505**
iDecPred	B_7	**0.898**	**0.096**	0.924	1.208	**0.107**
iDecPred	S_7	/	0.156	0.955	0.565	0.149
iDecPred	T_7	**1.000**	**0.013**	/	1.566	**0.080**
iDecPred	C_21	0.255	0.917	0.526	1.412	**0.530**
iDecPred	B_21	**0.483**	**0.263**	**0.872**	1.687	**0.111**
iDecPred	S_21	/	0.471	0.914	**0.436**	0.160
iDecPred	T_21	**0.998**	**0.030**	/	1.109	**0.078**
iDecPred*	C_7	0.383	**0.388**	0.635	0.406	**0.505**
iDecPred*	B_7	0.771	0.768	**0.928**	0.405	**0.107**
iDecPred*	S_7	/	**0.063**	**0.969**	0.408	0.149
iDecPred*	T_7	0.855	1.381	/	0.009	**0.080**
iDecPred*	C_21	**0.278**	**0.228**	0.525	0.317	**0.530**
iDecPred*	B_21	0.399	1.085	0.811	0.614	**0.111**
iDecPred*	S_21	/	**0.182**	**0.922**	0.486	0.160
iDecPred*	T_21	0.723	0.901	/	0.173	**0.078**
N-BEATS	C_7	-0.426	1.503	0.628	**0.256**	1.601
N-BEATS	B_7	0.486	0.313	0.904	**0.231**	0.398
N-BEATS	S_7	/	0.387	0.895	**0.398**	**0.106**
N-BEATS	T_7	0.998	0.115	/	**0.003**	0.134
N-BEATS	C_21	-0.607	0.900	**0.561**	**0.221**	0.916
N-BEATS	B_21	-0.010	0.740	0.733	**0.324**	0.594
N-BEATS	S_21	/	0.557	0.753	0.570	**0.130**
N-BEATS	T_21	0.999	0.123	/	**0.002**	0.148

To further examine the seasonal discrepancy, weighted trend and seasonal reconstructions are also reported, denoted iDecPred*, which apply the same model outputs after modulation by the learnt fusion weights. These results show that iDecPred* preserves high correlation while significantly reducing seasonal MAE (e.g., C_7: 0.406 vs. 1.374; S_21: 0.486 vs. 0.436), confirming that the fusion mechanism maintains phase fidelity while correcting amplitude distortions.

Overall, iDecPred demonstrates clear superiority in trend modelling and stable phase alignment in seasonal prediction, while the weighted reconstructions further improve amplitude calibration, leading to more robust and accurate final forecasting performance than N-BEATS across both simple and complex synthetic datasets.

4.5.6 Decomposition–Forecast Quality Relationship Exploration

Beyond reporting raw forecasting errors, it is further investigated how the quality of intermediate trend and seasonal decompositions correlates with the final predictive performance, as illustrated in Figure 4.9.

For each $\langle \text{window}, \text{channel} \rangle$ on dataset C_7 (sampled $n = 20{,}000$ per panel), decomposition quality (MAE and Pearson's r against ground truth) is computed and paired with the windowed forecast error. Density-coloured scatter plots are visualised and an OLS line is fitted; legends report Pearson r and Spearman ρ (all $p = 0.00$, numerically $\ll 10^{-3}$). One dot corresponds to a single $\langle \text{window}, \text{channel} \rangle$. The x-axis is a decomposition metric (trend/seasonal MAE or r); the y-axis is MAE_Y for the same pair.

iDecPred. On C_7, seasonal MAE exhibits a moderate positive association with forecast MAE (Pearson $r = 0.528$, Spearman $\rho = 0.533$), while trend MAE shows a slightly weaker but still clear association (Pearson $r = 0.450$, Spearman $\rho = 0.403$). Thus, for our model both component fidelities matter, with seasonal reconstruction exerting a somewhat larger influence on the local forecast error.

Relative to iDecPred, N-BEATS shows an extremely strong coupling between seasonal MAE and forecast MAE ($r \approx 0.879$, $\rho \approx 0.886$), while the trend–forecast association is

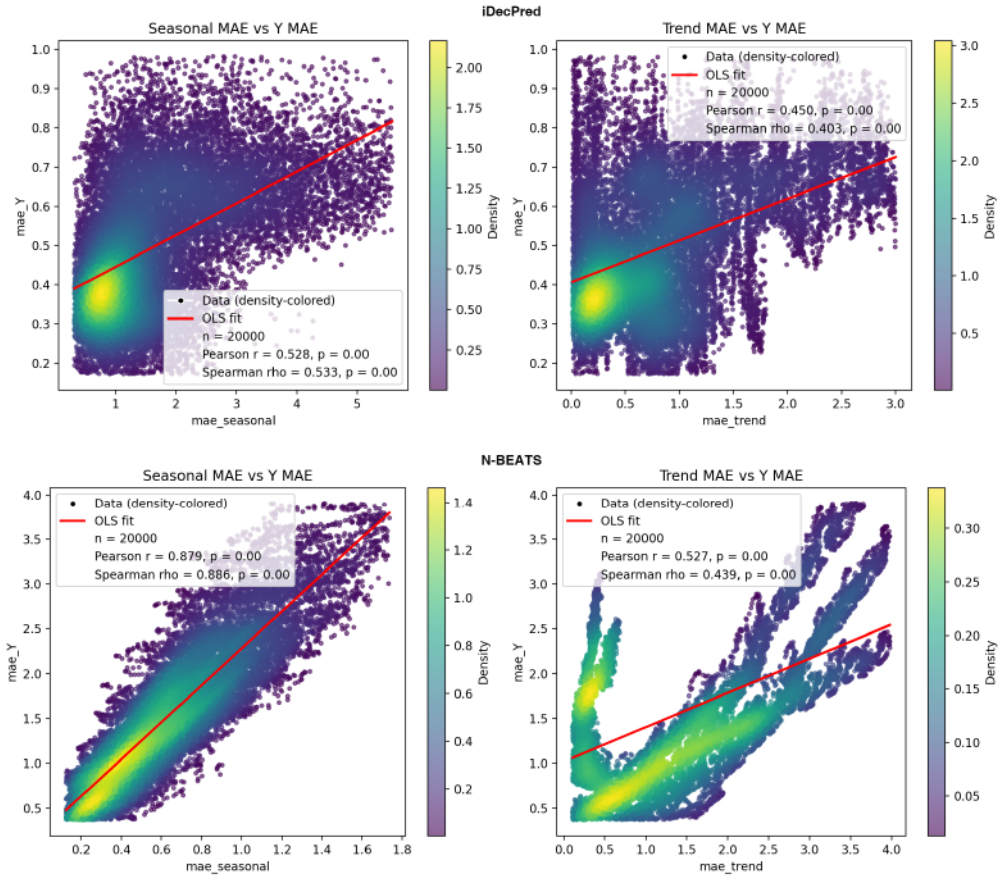


Figure 4.9: Decomposition–forecast relationship on C_7. Top row: iDecPred; bottom row: N-BEATS. (a) seasonal MAE vs. MAE_Y (iDecPred); (b) trend MAE vs. MAE_Y (iDecPred); (c) seasonal MAE vs. MAE_Y (N-BEATS); (d) trend MAE vs. MAE_Y (N-BEATS). Each point is a $\langle$window, channel$\rangle$; color encodes local density; red line is an OLS fit. Legends report n, Pearson r, and Spearman ρ.

comparable across models (both $\approx$0.45–0.53). This indicates N-BEATS' errors are far more sensitive to seasonal reconstruction fidelity on C_7, whereas trend fidelity contributes similarly for both methods.

Across models, (i) a positive association between decomposition quality and forecast error is observed and (ii) a consistently stronger influence of seasonal fidelity than trend fidelity is observed, with trend still contributing. On complex datasets where seasonality, trend, and other factors interact, improving component reconstruction, particularly seasonal fidelity (pronounced for N-BEATS and moderate for iDecPred), is associated with tangible error reductions. Taken together, the results support the rationale of decomposition-based forecasting structures and underscore the benefit of modules that enhance the realism of

predicted trend and seasonal components.

4.5.7 Airport Case Study (Top 20 Busiest U.S. Airports)

iDecPred is evaluated on a multivariate airport benchmark whose targets are passenger throughput series from Transportation Security Administration (TSA) releases published via the U.S. Freedom of Information Act electronic reading room [41]. From the available airport records, the twenty busiest U.S. hubs are retained as separate channels (C=20, IATA-coded). The same system is observed at hourly and daily resolutions: the daily panel aggregates the hourly targets, so both splits share an identical calendar span. Optional exogenous inputs are two national, standardised information indices derived from the GDELT event corpus [12]; keyword filtering, share construction, z-scoring, and temporal alignment to the throughput panels are documented in Appendix A. Table 4.5 records panel sizes and coarse descriptive statistics; the experimental grid is $L \in \{96, 192, 336, 720\}$, with $H \in \{96, 192, 336, 720\}$ for hourly runs and $H \in \{96, 192\}$ for daily runs.

Table 4.5: Airport benchmark panels: scale and summary statistics. Targets are TSA FOIA-released throughput for the twenty busiest U.S. hubs [41]; exogenous shocks follow the GDELT-based construction in Appendix A [12].

	Hourly	**Daily**
Timesteps	56,280	2,345
UTC span	30 Dec. 2018 – 31 May 2025	same
Sampling	hourly grid	day-aligned aggregates
Targets	C=20 IATA hubs; non-neg. counts	C=20 IATA hubs; non-neg. counts
Exogenous[a]	2 z-scored shocks (ban, health)	2 z-scored shocks (ban, health)
Missing	none	none
Pooled range[b]	$[0, \sim 10^4]$	$[0, \sim 1.4 \times 10^5]$
Zeros (targets)	$\approx 5\%$	$< 0.02\%$
Shock cadence	constant within UTC day	one value per date

[a]Ban, health, and both variants only (indices defined in Appendix A). [b]Min–max over all target channels and timesteps.

Four variants are compared: baseline (no external features), ban (control-related aviation-restriction information), health (health-related aviation information), and both. Table 4.6 reports the 96–96 results, and Table 4.7 summarises best-performing configurations. Notably, adding external features does *not* consistently improve forecasting accuracy and can be detrimental in some settings. This suggests that either (i) the current architecture

underutilizes exogenous signals during fusion, or (ii) the external indices are noisy/weakly aligned with the target dynamics at the forecasting granularity. Improving exogenous-feature utilization (e.g., dedicated exogenous encoders and stronger conditioning/fusion mechanisms) is left for future work.

Table 4.6: Airport case study results for the 96–96 setting (Top 20 busiest U.S. airports).

Setting	Variant	MSE	MAE	Params (M)	Latency (ms)	Peak Mem (MB)
Hourly	baseline	0.106	0.207	0.206	5.611	55.274
Hourly	ban	0.107	0.207	0.206	5.704	57.107
Hourly	health	0.107	0.207	0.206	5.550	57.107
Hourly	both	0.106	0.207	0.206	5.183	58.937
Daily	baseline	0.329	0.422	0.206	5.173	55.245
Daily	ban	0.323	0.418	0.206	4.539	57.078
Daily	health	0.323	0.418	0.206	4.821	57.078
Daily	both	0.324	0.418	0.206	7.327	58.908

Under the hourly 96–96 configuration, predictive performance remains remarkably consistent across all model variants, with MSE values ranging narrowly from 0.106 to 0.107 and MAE fixed at 0.207. This stability in error metrics indicates that the inclusion of the exogenous indices does not lead to noticeable gains in forecast accuracy for high-frequency, short-horizon airport time series. The models also display nearly identical complexity and computational footprint.

At the daily timescale, a slight improvement can be observed when external indices are incorporated into the model. In particular, introducing exogenous information leads to a reduction in MSE from 0.329 for the baseline to as low as 0.323, and a corresponding decrease in MAE from 0.422 to 0.418. This suggests a limited but tangible benefit from exogenous features when the temporal granularity is coarser.

The results presented in Table 4.7 further elucidate the importance of input window length. Extending the look-back period to $L = 720$ with a short forecast horizon ($H = 96$) yields the lowest error rates observed in this evaluation. When all exogenous features are utilized, the model achieves an hourly MSE of 0.086 and a daily MSE of 0.266. These findings indicate that substantial gains in forecasting accuracy for this dataset arise chiefly

from increased temporal context, whereas the improvements attributable to exogenous features are relatively modest when input length is short.

Table 4.7: Best-performing configurations (lowest MSE) on the airport case study.

Setting	Best configuration	MSE	MAE
Hourly	L=720, H=96 (both)	0.086	0.185
Daily	L=720, H=96 (both)	0.266	0.371

4.6 Limitations and Future Work

While iDecPred achieves competitive forecasting accuracy, efficiency, and interpretability, several limitations remain.

First, the spline-based trend module assumes relatively smooth temporal dynamics. While effective for signals with gradual evolution or piecewise-smooth regime changes, it may be less suitable for scenarios with multiple abrupt regime shifts or strong multiplicative interactions between latent components. The spline representation may also be less responsive to isolated sharp outliers. In such cases, the residual MLP and adaptive fusion gate help capture irregular fluctuations by increasing the contribution of the residual component. As observed in the synthetic stress tests, the performance advantage of iDecPred narrows in multi-regime-shift settings and can be surpassed by basis-expansion models such as N-BEATS [30] under fully multiplicative dynamics.

Second, the adaptive fusion mechanism that combines trend, seasonal, and residual forecasts introduces a trade-off between forecasting accuracy and exact component-wise reconstruction. Because the fusion gate redistributes signal energy across components to optimise the final prediction, the predicted trend and seasonal components may exhibit amplitude deviations when evaluated independently against ground truth. As observed in the synthetic experiments, iDecPred can achieve strong phase alignment but sometimes yields larger component-wise MAE than N-BEATS [30].

Third, the relative performance of iDecPred compared with strong baselines such as

TimeMixer [43] or iTransformer [26] can be dataset-dependent. While iDecPred performs well on many benchmarks, datasets with extremely high dimensionality, stochasticity, or weak temporal structure (e.g., Traffic) remain challenging for decomposition-based forecasting methods.

An important direction for future research is to improve component-level fidelity by introducing constraints on the fusion mechanism so that the amplitudes of trend and seasonal components are better preserved. Incorporating structure-aware regularisation tailored to different decomposition components, such as smoothness constraints for trends and periodic consistency for seasonal terms, may further improve the interpretability and stability of the recovered components. Another promising direction is to extend the decomposition framework to better handle structural changes and multiplicative interactions between components, which remain challenging for additive decomposition models. Extensions that incorporate exogenous variables may also broaden the applicability of the framework in real-world forecasting systems.

4.7 Conclusion

iDecPred is presented as a lightweight and interpretable neural decomposition framework for multivariate time series forecasting. Unlike conventional black-box neural models, iDecPred explicitly models future trend, seasonal, and residual components through modular neural architectures in an end-to-end learning framework and fuses them via a channel-wise gating mechanism. This design bridges classical time-series decomposition principles with modern neural forecasting, improving both predictive accuracy and model transparency.

Comprehensive evaluations on seven standard benchmarks, two real-world airport passenger flow datasets, and sixteen synthetic datasets show that iDecPred achieves state-of-the-art results on several datasets and competitive performance on others, while using fewer parameters, exhibiting lower inference latency and peak memory than transformer-based baselines. Additional experiments under regime shifts and multiplicative stress

scenarios further demonstrate the robustness of the proposed framework. On synthetic datasets, iDecPred demonstrates a strong balance between accuracy and efficiency, and its predicted components exhibit consistent alignment with ground-truth trend and seasonal patterns. These results highlight the potential of forecast-oriented decomposition as a practical design principle for interpretable time series forecasting.

Overall, iDecPred provides a compact and effective framework that integrates accuracy, efficiency, and interpretability, making it a promising foundation for trustworthy forecasting systems in domains such as energy, transportation, and environmental monitoring.

References

[1] Adebiyi A. Ariyo, Adewumi O. Adewumi, and Charles K. Ayo. Stock price prediction using the ARIMA model. In *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, pages 106–112, 2014. doi: 10.1109/UKSim.2014.67.

[2] Pierre-Daniel Arsenault, Shengrui Wang, and Jean-Marc Patenaude. A survey of explainable artificial intelligence (xai) in financial time series forecasting. *ACM Computing Surveys*, 57(10):1–37, 2025. doi: 10.1145/3729531.

[3] Kasun Bandara, Christoph Bergmeir, and Hansika Hewamalage. Lstm-msnet: Leveraging forecasts on sets of related time series with multiple seasonal patterns. *IEEE Transactions on Neural Networks and Learning Systems*, 32(4):1586–1599, 2020. doi: 10.1109/tnnls.2020.2985720.

[4] Kosala Bandara, Christoph Bergmeir, and Slawek Smyl. Forecasting across time series databases using a nonparametric approach to multiple seasonal time series decomposition. *Expert Systems with Applications*, 169:114481, 2021.

[5] George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1970.

[6] Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. Tempo: Prompt-based generative pre-trained transformer for time series forecasting. *arXiv preprint arXiv:2310.04948*, 2023.

[7] Lingxiao Cao, Bin Wang, Guiyuan Jiang, Yanwei Yu, and Junyu Dong. Spatiotemporal-aware trend-seasonality decomposition network for traffic flow fore-

casting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11463–11471, 2025. doi: 10.1609/aaai.v39i11.33247.

[8] Robert B. Cleveland, William S. Cleveland, J. McRae, and I. Terpenning. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3–73, 1990.

[9] Yihong Dong, Ge Li, Xue Jiang, Yongding Tao, Kechi Zhang, Hao Zhu, Huanyu Liu, Jiazheng Ding, Jia Li, Jinliang Deng, et al. Fanformer: Improving large language models through effective periodicity modelling. *arXiv preprint arXiv:2502.21309*, 2025.

[10] James Durbin and Siem Jan Koopman. *Time series analysis by state space methods.* Oxford University Press, 2 edition, 2012.

[11] Shanghua Gao, Teddy Koker, Owen Queen, Tom Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. Units: A unified multi-task time series model. *Advances in Neural Information Processing Systems*, 37:140589–140631, 2024.

[12] GDELT Project. A global database of society. `https://www.gdeltproject.org/`, 2022. Accessed: 2026-01-13.

[13] Luke B. Godfrey and Michael S. Gashler. Neural decomposition of time-series data for effective generalization. *IEEE Transactions on Neural Networks and Learning Systems*, 29(7):2973–2985, 2017. doi: 10.1109/tnnls.2017.2709324.

[14] Songtao Huang, Zhen Zhao, Can Li, and Lei Bai. Timekan: Kan-based frequency decomposition learning architecture for long-term time series forecasting. *arXiv preprint arXiv:2502.06910*, 2025.

[15] Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice.* OTexts, 3rd edition, 2021. URL `https://otexts.com/fpp3/`.

[16] Minjung Kim, Yusuke Hioka, and Michael Witbrock. Neural Fourier modelling: A highly compact approach to time-series analysis. *arXiv preprint arXiv:2410.04703*, 2024.

[17] Taesung Kim, Jinhee Kim, Ethan Toney, and Roland Memisevic. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=cGDAkQo1C0p`.

[18] Guan Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. modelling long- and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 95–104, 2018.

[19] Mikko Lehtimäki, Lassi Paunonen, and Marja-Leena Linne. Accelerating neural odes using model order reduction. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):519–531, 2022.

[20] Tianfu Li, Chuang Sun, Sinan Li, Zhiying Wang, Xuefeng Chen, and Ruqiang Yan. Explainable graph wavelet denoising network for intelligent fault diagnosis. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):8535–8548, 2022.

[21] Bryan Lim, Sercan Ö. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.

[22] Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. TimeCMA: Towards LLM-empowered multivariate time series forecasting via cross-modality alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18780–18788, 2025. doi: 10.1609/aaai.v39i18.34067.

[23] Hengbo Liu, Ziqing Ma, Linxiao Yang, Tian Zhou, Rui Xia, Yi Wang, Qingsong Wen, and Liang Sun. Sadi: A self-adaptive decomposed interpretable framework for electric load forecasting under extreme events. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. doi: 10.1109/icassp49357.2023.10096002.

[24] Peiyuan Liu, Beiliang Wu, Naiqi Li, Tao Dai, Fengmao Lei, Jigang Bao, Yong Jiang, and Shu-Tao Xia. WFTNet: Exploiting global and local periodicity in long-term time series forecasting. *arXiv preprint arXiv:2309.11319*, 2024.

[25] Xu Liu, Juncheng Liu, Gerald Woo, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Moirai-moe: Empowering time series foundation models with sparse mixture of experts. *arXiv preprint arXiv:2410.10469*, 2024.

[26] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024.

[27] Alysha M. De Livera, Rob J. Hyndman, and Ralph D. Snyder. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496):1513–1527, 2011.

[28] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[29] Yao Nie, Ziqing Zhang, Ruixiang Wang, Zhaocheng Liu, and Junchi Tang. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2023.

[30] Boris N. Oreshkin, Dimitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-

beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020.

[31] Xihao Piao, Zheng Chen, Taichi Murayama, Yasuko Matsubara, and Yasushi Sakurai. Fredformer: Frequency debiased transformer for time series forecasting. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 2400–2410, 2024. doi: 10.1145/3637528.3671928.

[32] Derck WE Prinzhorn, Thijmen Nijdam, Putri A van der Linden, and Alexander Timans. Conformal time series decomposition with component-wise exchangeability. *arXiv preprint arXiv:2406.16766*, 2024.

[33] Vikas C. Raykar, Arindam Jati, Sumanta Mukherjee, Nupur Aggarwal, Kanthi Sarpatwar, Giridhar Ganapavarapu, and Roman Vaculin. TsSHAP: Robust model agnostic feature-based explainability for time series forecasting. *arXiv preprint arXiv:2303.12316*, 2023.

[34] Lei Ren, Tingyu Mo, Xuejun Cheng, and Xi Li. Temporal–frequency attention focusing for time series extrinsic regression via auxiliary task. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):15494–15506, 2023.

[35] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[36] Abdul Khalique Shaikh, Amril Nazir, Imran Khan, and Abdul Salam Shah. Short term energy consumption forecasting using neural basis expansion analysis for interpretable time series. *Scientific Reports*, 12(1):22562, 2022.

[37] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post-hoc explanation methods. In

Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pages 180–186, 2020. doi: 10.1145/3375627.3375830.

[38] Slawek Smyl. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. In *International Workshop on Neural Forecasting*, 2020.

[39] Artyom Stitsyuk and Jaesik Choi. xpatch: Dual-stream time series forecasting with exponential seasonal-trend decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20601–20609, 2025. doi: 10.1609/aaai.v39i19.34270.

[40] Sean J. Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

[41] Transportation Security Administration. Foia electronic reading room. `https://www.tsa.gov/foia/readingroom`, 2022. Accessed: 2026-01-13.

[42] Dongjing Wang, Gangming Guo, Tianpei Ouyang, Dongjin Yu, Haiping Zhang, Bao Li, Rong Jiang, Guandong Xu, and Shuiguang Deng. A lightweight spatio-temporal neural network with sampling-based time series decomposition for traffic forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 26(6):8682–8693, 2025. doi: 10.1109/TITS.2025.3552010.

[43] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Yiming Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. *arXiv preprint arXiv:2405.14616*, 2024.

[44] Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. *Advances in Neural Information Processing Systems*, 37:469–498, 2024.

[45] Wei Wei, Zi'ang Wang, Bowen Pang, Jiannan Wang, and Xue Liu. Wavelet transformer: An effective method on multiple periodic decomposition for time series forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8): 14063–14077, 2025. doi: 10.1109/TNNLS.2025.3525502.

[46] Xiulan Wei, Yong Zhang, Yun Wei, Yongli Hu, Shuzhen Tong, Wei Huang, and Jinde Cao. Metro passenger-flow representation via dynamic mode decomposition and its application. *IEEE Transactions on Neural Networks and Learning Systems*, 34(1): 157–170, 2021.

[47] Qingsong Wen, Liang Sun, Fan Yang, Xian Song, Jianmin Gao, Xue Wang, and Hongbo Xu. Robuststl: A robust seasonal-trend decomposition algorithm for long time series. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1055–1063, 2019.

[48] Peter R. Winters. Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3):324–342, 1960.

[49] Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. Etsformer: Exponential smoothing transformers for time-series forecasting. *arXiv preprint arXiv:2202.01381*, 2022.

[50] Haixu Wu, Yuxuan Xu, Jin Wang, Mingsheng Long, Jianmin Wang, and Michael I. Jordan. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pages 22419–22430, 2021.

[51] Haixu Wu, Yuxuan Xu, Mingsheng Long, Jianmin Wang, and Michael I. Jordan. Timesnet: Temporal 2d-variation modelling for general time series analysis. *arXiv preprint arXiv:2305.08891*, 2023.

[52] Shi Xiaoming, Wang Shiyu, Nie Yuqi, Li Dianqi, Ye Zhou, Wen Qingsong, and Ming Jin. Time-moe: Billion-scale time series foundation models with mixture of experts.

In *ICLR 2025: The Thirteenth International Conference on Learning Representations*. International Conference on Learning Representations, 2025.

[53] Yang Xing, Zhongxu Hu, Peng Hang, and Chen Lv. Learning from the dark side: A parallel time series modelling framework for forecasting and fault detection on intelligent vehicles. *IEEE Transactions on Intelligent Vehicles*, 9(2):3205–3219, 2024. doi: 10.1109/TIV.2023.3342648.

[54] Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with 10k parameters. *arXiv preprint arXiv:2307.03756*, 2024.

[55] Shashank Yadav and Vignesh Subbian. Failure modes of time series interpretability algorithms for critical care applications and potential solutions. *arXiv preprint arXiv:2506.19035*, 2025.

[56] Kun Yi, Jingru Fei, Qi Zhang, Hui He, Shufeng Hao, Defu Lian, and Wei Fan. Filternet: Harnessing frequency filters for time series forecasting. *Advances in Neural Information Processing Systems*, 37:55115–55140, 2024.

[57] Guoqi Yu, Jing Zou, Xiaowei Hu, Angelica I. Aviles-Rivero, Jing Qin, and Shujun Wang. Revitalizing multivariate time series forecasting: Learnable decomposition with inter-series dependencies and intra-series variations modelling. In *International Conference on Machine Learning*, 2024.

[58] Wenzhen Yue, Yong Liu, Xianghua Ying, Bowei Xing, Ruohao Guo, and Ji Shi. Freeformer: Frequency enhanced transformer for multivariate time series forecasting. *arXiv preprint arXiv:2501.13989*, 2025.

[59] Ailing Zeng, Qi Lai, Tongtong Lin, Yang Li, Haoyuan Peng, Yiying Zhou, Dong Ni, Yichuan Xu, Zheng Li, Weinan Zhang, et al. Are transformers effective for time series forecasting? In *International Conference on Machine Learning*, pages 25977–25992. PMLR, 2023.

[60] Yuyi Zhang, Qiushi Sun, Dongfang Qi, Jing Liu, Ruimin Ma, and Ovanes Petrosian. ShapTime: A general XAI approach for explainable time series forecasting. In *Lecture Notes in Networks and Systems: Intelligent Systems and Applications*, pages 659–673. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-47721-8_45.

[61] Zhiqiang Zhang, Yuxuan Chen, Dandan Zhang, Yining Qian, and Hongbing Wang. CTFNet: Long-sequence time-series forecasting based on convolution and time–frequency analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):16368–16382, 2024. doi: 10.1109/TNNLS.2023.3294064.

[62] Kaiping Zheng, Shaofeng Cai, Horng Ruey Chua, Wei Wang, Kee Yuan Ngiam, and Beng Chin Ooi. Tracer: A framework for facilitating accurate and interpretable analytics for high stakes applications. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 1747–1763, 2020.

[63] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

[64] Tian Zhou, Ziqing Ma, Qingsong Wen, Ziqi Zhou, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Learning Representations*, 2023.

5

Variable-Horizon Forecasting

Variable-Horizon Patch-Autoregressive Forecasting with Semantic Anchoring and Scenario-Based Robustness Analysis

Abstract

Different horizon time-series forecasting faces fundamental challenges: horizon fragmentation from fixed-horizon training, error accumulation in single-step autoregressive models, and a lack of diagnostic tools for understanding model behaviour under uncertainty. In this chapter, a forecasting framework is proposed based on variable-horizon Patch-Autoregressive (Patch-AR) training, augmented with semantic anchoring for interpretability and scenario-based diagnostics for robustness analysis, where a single model learns to forecast across multiple horizons by unrolling patch-level latent dynamics. Semantic anchoring uses frozen language models to softly regularise patch embeddings toward interpretable concepts while maintaining forecast accuracy. Diagnostic analyses clarify the scope and limitations of concept-based explanations.

Empirical validation on heterogeneous airport datasets demonstrates that variable-horizon Patch-AR achieves competitive accuracy across multiple horizons (24–720 steps) with a single model, while patch-level rollouts significantly outperform single-step autoregressive baselines, particularly at longer horizons, where error accumulation becomes more pronounced.

In addition, robustness under exogenous shocks is studied via scenario-based sensitivity analysis. Rather than proposing a new forecasting architecture for shocks, the approach characterises how forecast outputs respond when future exogenous trajectories are systematically perturbed while historical inputs are held fixed.

5.1 Introduction

Real-world forecasting systems are rarely deployed for a single lead time. Operational decision-making typically requires a coherent set of predictions spanning short-term

horizons (e.g., the next 24 hours) to longer-term planning horizons (e.g., the next 30 days). Despite this, many modern neural forecasting pipelines are trained, tuned, or even architected separately for each target horizon, implicitly encouraging a fragmented one-model-per-horizon workflow [14, 18, 24, 29]. Such fragmentation increases training and maintenance costs, complicates deployment, and can result in inconsistent behaviour across horizons [7].

This issue becomes particularly acute in long-horizon forecasting. Iterative single-step autoregressive (AR) generation is attractive for its flexibility, but its training objective is misaligned with long rollouts: prediction errors are repeatedly fed back into the model, often leading to compounding errors and rapidly degrading forecasts at extended horizons [20, 14]. At the same time, as neural forecasters grow more complex and are increasingly used in high-stakes settings, there is a growing demand for models that can be meaningfully inspected and diagnosed, rather than treated as opaque black boxes [6, 5]. In this chapter, explainability takes two forms. First, semantic explainability: the proposed framework uses semantic anchoring (via a frozen language model) to associate patch-level latent states with human-readable concepts (e.g., “spike”, “flat”, “periodic”), so that explanations are concept-based descriptions of which temporal patterns are expressed where, rather than input-level attribution. Second, scenario-based explainability: robustness diagnostics are introduced under alternative future exogenous trajectories, so that practitioners can assess how forecasts respond to counterfactual or uncertain shock scenarios.

Beyond horizon variability, many applied forecasting deployments must contend with exogenous shocks (e.g., public health disruptions, policy interventions, or other external signals) that introduce uncertainty into future conditions. Even when auxiliary covariates are used to improve forecast accuracy [14, 22], practitioners often need to answer “what-if” questions: how would forecasts change under alternative future shock trajectories, and are predictions stable (robust) under plausible perturbations [26, 27].

Our key insight is that rather than training separate models or prediction heads for each horizon, a single forecasting model should learn internal dynamics that can be

unrolled for variable numbers of steps. This perspective shifts the focus from horizon-specific architectural design to learning robust latent dynamics that generalise across rollout lengths [28, 20].

This chapter makes the following contributions:

1. **Variable-horizon training protocol**: A training procedure is introduced that enables a single model to forecast across multiple horizons by sampling horizons during training and unrolling patch-level latent dynamics, eliminating the need for horizon-specific models. Crucially, performance remains stable across the evaluated horizon range, suggesting that the model learns horizon-agnostic temporal patterns rather than horizon-specific mappings.

2. **Patch-autoregressive latent dynamics**: It is demonstrated that patch-level autoregressive rollouts significantly outperform single-step autoregressive baselines, particularly at longer horizons where error accumulation is severe. On airport datasets, patch-AR reduces MSE by 32% (daily) and 67% (hourly) compared to single-step AR at horizon 96.

3. **Semantic anchoring and diagnostic interpretability**: Semantic anchoring is proposed using frozen language models as a lightweight regularizer that provides interpretable, concept-based descriptions of latent dynamics, and explicit diagnostic evaluations are conducted to clarify the scope and limitations of these explanations.

4. **Scenario-based robustness diagnostics under exogenous shocks**: A scenario-based sensitivity analysis protocol is introduced to quantify how forecasts respond to counterfactual future exogenous trajectories, enabling robustness assessment beyond point accuracy.

5.2 Related Work

5.2.1 Long-Horizon Time-Series Forecasting

Long-horizon forecasting has been addressed through two primary paradigms: direct forecasting and iterative (recursive) forecasting [17]. Direct forecasting methods predict multiple future steps simultaneously, often using encoder-decoder or sequence-to-sequence models to output the entire horizon in one pass [24]. While direct approaches avoid the recursive error accumulation inherent to iterated forecasting, they still face challenges at very long horizons due to limited context windows and the difficulty of capturing long-range dependencies, which tends to increase forecast error with horizon length [13].

Patch-based and representation-based models have emerged as a promising direction for long-horizon forecasting. PatchTST [18] segments time series into patches and applies transformers over patch sequences, demonstrating strong performance on long-horizon benchmarks. iTransformer [16] inverts the attention mechanism to attend over variates rather than time steps, improving scalability. TimesNet [25] models temporal patterns using two-dimensional convolutions over time-frequency representations. These approaches share the insight that operating on local segments rather than individual time steps can improve long-range modelling; however, they are typically trained and evaluated under fixed-horizon settings and do not explicitly address horizon fragmentation within a single unified model.

Decomposition-based methods, such as Autoformer [24] and FEDformer [29], explicitly separate trend and seasonal components to capture different temporal scales. While this decomposition can improve long-horizon forecasting, these methods are typically trained separately for each target horizon and do not explicitly aim to learn horizon-agnostic latent dynamics that generalise across rollout lengths.

5.2.2 Autoregressive Latent Dynamics

Autoregressive models have a long history in time-series forecasting, ranging from classical ARIMA models to modern neural architectures [2, 20]. A central challenge of autoregressive forecasting is error accumulation: prediction errors introduced at each step are recursively fed back into the model and can compound over the forecast horizon [1, 20]. This issue becomes particularly severe for long-horizon forecasting, where even small per-step errors can lead to substantial degradation in forecast quality [17, 21]. In time-series forecasting, models such as DeepAR [20] rely on observation-level autoregressive generation, making them susceptible to error accumulation under long forecasting horizons. Other architectures, such as the Temporal Fusion Transformer [14], mitigate this issue through direct multi-horizon decoding with known future covariates, rather than iterative autoregressive rollout.

Our approach differs by operating at the patch level rather than the observation level. By predicting local segments instead of individual time steps, the number of autoregressive rollout steps required for a given forecast horizon is substantially reduced, mitigating the effects of error accumulation. This perspective aligns with prior empirical findings that recursive single-step forecasting degrades rapidly at long horizons, motivating alternative formulations that improve rollout stability [20, 18].

5.2.3 Language Models in Time-Series Modelling

The integration of language models with time-series data has been explored along several distinct directions. Broadly, existing approaches can be grouped into those that reprogram time series into prompt- or token-like representations for direct use by large language models, and those that align time-series representations with language representations through cross-modal learning [11, 15, 9]. In a separate line of work, text-augmented forecasting incorporates external textual signals, such as news or social media, as auxiliary inputs to time-series models, typically requiring both modalities during training and

inference [8, 12].

More recently, language-model-centric approaches repurpose pre-trained LLMs for time-series prediction by mapping time series into token-like representations or textual prompts. Time-LLM [11], for instance, reprograms time series into textual prototypes and uses a frozen LLM backbone to produce forecasts. TimeCMA [15] further improves multivariate forecasting via cross-modality alignment. TimeXL [10] proposes an LLM-in-the-loop multi-modal framework where prototype-based encoders generate initial predictions and multiple LLM agents iteratively refine predictions and textual explanations. While these approaches leverage the representational capacity and reasoning ability of LLMs, their forecasting pipelines are tightly coupled to prompt design, tokenisation schemes, or LLM-centric orchestration, and are often not explicitly designed around temporal inductive biases or long-horizon stability objectives [18].

In contrast, this chapter uses a frozen language model solely as a semantic reference rather than as a forecasting engine. Patch embeddings are softly regularised toward semantic concept embeddings derived from the language model, providing interpretable semantic explanations while preserving the temporal modelling capacity and rollout stability of the forecasting model.

5.3 Problem Setup and Notation

5.3.1 Forecasting Task Definition

Airport passenger flow forecasting is considered as a multivariate time-series forecasting task. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ denote a multivariate time series of length T, where each $\mathbf{x}_t \in \mathbb{R}^d$ is a d-dimensional observation vector at time t. The forecasting task is to predict future values $\mathbf{Y} = \{\mathbf{x}_{T+1}, \mathbf{x}_{T+2}, \ldots, \mathbf{x}_{T+H}\}$ given a context window $\mathbf{X}_{context} = \{\mathbf{x}_{T-L+1}, \mathbf{x}_{T-L+2}, \ldots, \mathbf{x}_T\}$ of length L, where H is the prediction horizon.

In this setting, variable-horizon forecasting is of particular interest, where a single model should be able to predict multiple horizons $H \in \mathscr{H}_{\text{eval}}$ without retraining. In addition,

for robustness analysis, forecasts conditioned on a future exogenous trajectory $\mathbf{U}_{T+1:T+H}$ (e.g., shock-related signals) are considered, and responses are evaluated when $\mathbf{U}_{T+1:T+H}$ is replaced by a counterfactual trajectory $\tilde{\mathbf{U}}_{T+1:T+H}$ while $\mathbf{X}_{context}$ is held fixed.

5.3.2 Horizon Sets

Three sets of horizons used in the experimental protocol are distinguished:

Training horizons ($\mathscr{H}_{\text{train}}$): The set of horizons used during training. In variable-horizon training, horizons are sampled from $\mathscr{H}_{\text{train}}$ uniformly at random for each training batch. For example, $\mathscr{H}_{\text{train}} = \{24, 48, 96, 192\}$ for daily data or $\mathscr{H}_{\text{train}} = \{96, 192, 336, 720\}$ for hourly data.

Validation horizons ($\mathscr{H}_{\text{val}}$): The set of horizons used for model selection and early stopping. Typically, $\mathscr{H}_{\text{val}} = \mathscr{H}_{\text{train}}$ to ensure the model is evaluated on the same horizons it was trained on.

Evaluation horizons ($\mathscr{H}_{\text{eval}}$): The set of horizons used for final evaluation and reporting. In our experiments, $\mathscr{H}_{\text{eval}} = \mathscr{H}_{\text{train}}$, but in principle, $\mathscr{H}_{\text{eval}}$ could include horizons not seen during training to test generalisation.

The key advantage of variable-horizon training is that a single model can handle all horizons in $\mathscr{H}_{\text{eval}}$ without retraining, eliminating the need for horizon-specific models.

5.3.3 Evaluation Metrics

Forecasting performance is evaluated using standard regression metrics [25]:

Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{H \cdot d} \sum_{h=1}^{H} \sum_{v=1}^{d} (\hat{x}_{T+h,v} - x_{T+h,v})^2 \tag{5.1}$$

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{H \cdot d} \sum_{h=1}^{H} \sum_{v=1}^{d} |\hat{x}_{T+h,v} - x_{T+h,v}| \tag{5.2}$$

where $\hat{x}_{T+h,v}$ is the predicted value and $x_{T+h,v}$ is the ground truth for variate v at lead step h (v is used for variates in these metrics and i is reserved for patch indices in Sections 5.4.2 and 5.4.5).

Horizon-wise reporting: Metrics are reported separately for each horizon $h \in \mathscr{H}_{\text{eval}}$ to understand how performance degrades with increasing horizon.

For interpretability evaluation, stability of concept rankings and semantic anchoring to patch statistics are additionally reported (Section 5.6.5). For robustness under exogenous shocks, scenario-based forecast deltas are additionally reported (Section 5.6.6).

5.4 Methodology

This section specifies the variable-horizon Patch-AR forecaster and its optional semantic-anchoring stage. Figure 5.1 follows the implementation order: the context series is patchified and embedded; each patch token is mixed with a finite concept bank in reprogramming (below); the resulting vectors are prefixed by the embedding of a single fixed English prompt, "Predict next values:", with tokenizer whitespace, so frozen GPT-2 can process the sequence by feeding continuous patch embeddings in place of discrete token indices (standard causal attention order). Only the patch positions are retained, optionally augmented by an amplitude residual from raw patch embeddings (Section 5.4.2), mapped to an initial latent state, and the trainable patch autoregressor predicts future patches up to horizon H. That prefix is the same for every series, variable, and horizon and does not list concepts or task instructions; semantic alignment to named concepts is carried entirely by the reprogramming step, where concept token embeddings supply the attention keys and values. Scalar forecasts are produced only by the patch-level autoregressor in Section 5.4.2 (GPT-2 is not used to decode numeric futures as text and is not fine-tuned for forecasting). Patch-level concept scores reported later are the reprogramming attention weights (Section 5.4.5).

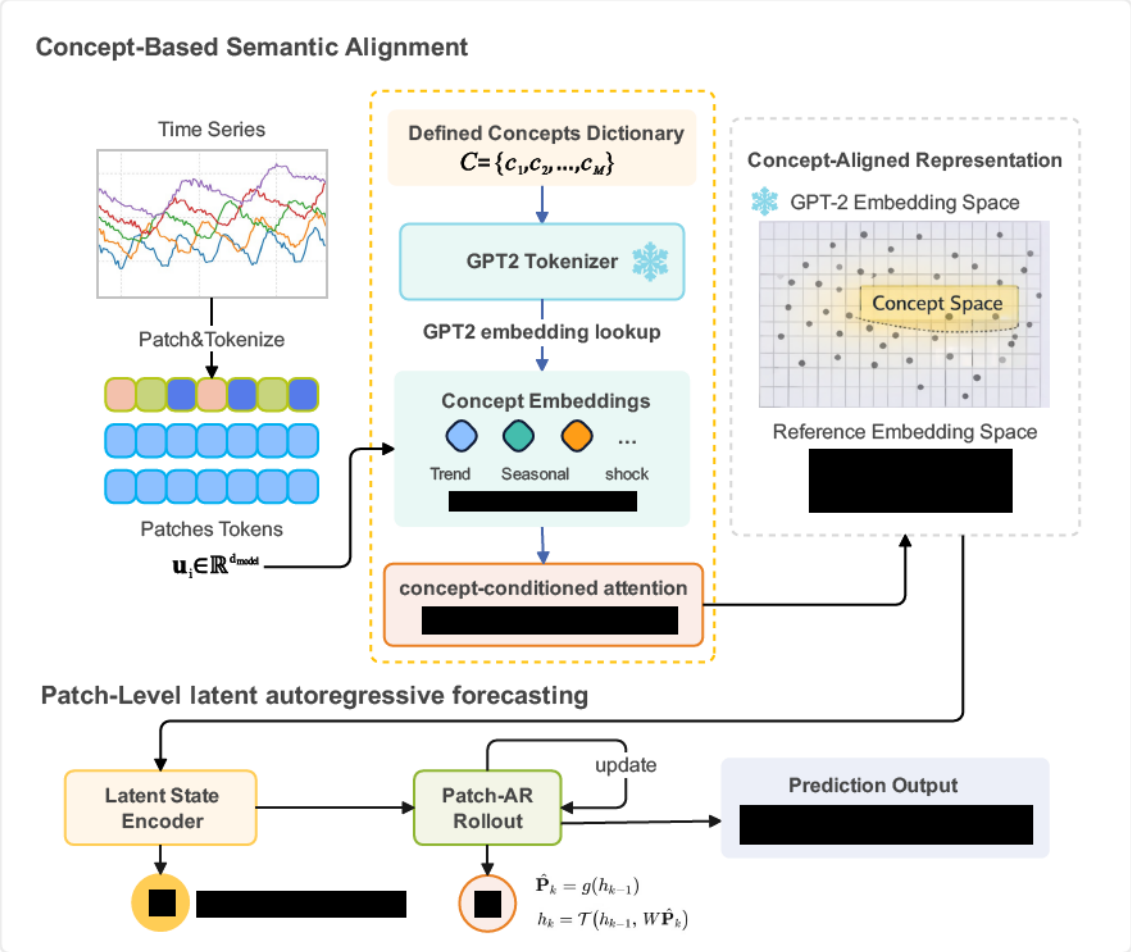


Figure 5.1: Overview of the implemented pipeline. Trainable forecasting path: patchification; reprogramming with a concept bank; a fixed "Predict next values:" prefix plus reprogrammed patches fed to frozen GPT-2 as continuous embeddings (patch positions only); optional amplitude residual from raw patch tokens (Section 5.4.2); encoding to an initial latent state; patch-level autoregressive rollout to horizon H. Interpretability: reprogramming attention yields per-patch concept scores (Section 5.4.5); GPT-2 does not decode future numeric values.

5.4.1 Patchification and Tokenisation

The input time series is first converted into a sequence of patch tokens. Given an input context window $\mathbf{X}_{context} \in \mathbb{R}^{L \times d}$ (length L, d variables), it is reshaped to (d, L) and patches are extracted with a patch length p and stride s using right-padding by replication. This yields a patch sequence

$$\mathbf{P}_i \in \mathbb{R}^p, \quad i = 1, \ldots, N, \quad N \approx \left\lfloor \frac{L+s-p}{s} \right\rfloor + 1, \tag{5.3}$$

for each variable, where each $\mathbf{P}_i$ is a length-p temporal segment.

To obtain tokens suitable for attention/LLM conditioning, each patch is projected into a d_{model}-dimensional token via a 1D convolutional token embedding, producing patch tokens

$$\mathbf{u}_i \in \mathbb{R}^{d_{\text{model}}}. \tag{5.4}$$

This reduces the effective sequence length from L to N and provides stable local summaries for long-horizon reasoning.

$L = 96$ is used as the default context length. For daily data, $p = 7$, $s = 2$ are used; for hourly data, $p = 16$, $s = 8$ are used. Time features are additionally injected by projecting a patch-level average of calendar features and adding it to each patch token prior to concept reprogramming.

5.4.2 Patch-Autoregressive Latent Dynamics

Our forecasting model operates autoregressively at the patch level. After encoding the N context patches $\mathbf{P}_1, \ldots, \mathbf{P}_N$, the encoder yields an initial latent state $\mathbf{h}_0 \in \mathbb{R}^{d_{\text{LLM}}}$ for each variable stream. For $k = 1, \ldots, K$ with $K = \lceil H/p \rceil$, the rollout predicts future patches $\hat{\mathbf{P}}_{N+k}$ and updates a recurrent latent state $\mathbf{h}_k$:

$$\hat{\mathbf{P}}_{N+k} = g(\mathbf{h}_{k-1}), \qquad \mathbf{h}_k = \text{GRUCell}\big(W\hat{\mathbf{P}}_{N+k}, \mathbf{h}_{k-1}\big), \tag{5.5}$$

where $\hat{\mathbf{P}}_{N+k} \in \mathbb{R}^p$ is the kth predicted future patch (consistent with Algorithm 1), $W : \mathbb{R}^p \to \mathbb{R}^{d_{\text{LLM}}}$ is a learned linear projection, and $g(\cdot)$ is a small MLP decoder with latent LayerNorm for stability. The model rolls out for K steps, the predicted patches are concatenated, and the output is truncated to the desired horizon H.

The context is patchified with stride s to obtain a dense token sequence, but the forecast is generated as consecutive length-p patches (no overlap) to minimise rollout steps and mitigate error accumulation.

The encoder and patch autoregressor operate on per-patch vectors in $\mathbb{R}^{d_{\text{LLM}}}$ built from

the convolutional patch embeddings by the reprogramming and frozen GPT-2 pathway in Section 5.4.4.

Amplitude residual channel. In the reported runs, the raw patch tokens $\mathbf{u}_i \in \mathbb{R}^{d_{\text{model}}}$ (before reprogramming) are additionally mapped through a learned linear projection into $\mathbb{R}^{d_{\text{LLM}}}$ and the result is added to the contextualised patch outputs with a learnable scalar gain. This skip path preserves level and local variability that concept mixing and the frozen language-model block can attenuate, which stabilises training and mitigates overly smoothed patch forecasts; it does not replace reprogramming, and concept attention weights in Section 5.4.5 still supply the interpretable concept scores.

5.4.3 Variable-Horizon Training Protocol

For every training batch, a horizon H is sampled uniformly at random from $\mathscr{H}_{\text{train}}$. The dataset produces a fixed maximum target window ($\max(\mathscr{H}_{\text{train}})$) and the target is sliced to the sampled horizon for each batch. The forward pass is evaluated with prediction horizon H, and the loss is computed only over the first H steps.

At validation and test time, the model is evaluated on all horizons in $\mathscr{H}_{\text{val}}$ or $\mathscr{H}_{\text{eval}}$, using the same model parameters.

Algorithm 1 Variable-Horizon Patch-AR Training

Input: Training data $\mathscr{D}$, horizon set $\mathscr{H}_{\text{train}}$, patch length p, stride s

Output: Trained model θ

Initialize model parameters θ
for epoch $= 1$ to max_epochs **do**
  for batch in $\mathscr{D}$ **do**
    Sample horizon $H \sim \text{Uniform}(\mathscr{H}_{\text{train}})$
    Sample context window $\mathbf{X}_{context}$ and target $\mathbf{Y}$ of length H
    // Patchification
    $\mathbf{P} = \text{Patchify}(\mathbf{X}_{context}, p, s)$ // N patches
    // Encode patches to latent space
    $\mathbf{h}_0 = \text{Encoder}(\mathbf{P}_1, \ldots, \mathbf{P}_N)$ // Initial latent state
    // Autoregressive rollout
    $K = \lceil H/p \rceil$
    for $k = 1$ to K **do**
      $\hat{\mathbf{P}}_{N+k} = g(\mathbf{h}_{k-1})$ // Predict patch from latent state
      $\mathbf{h}_k = \text{GRUCell}(W\hat{\mathbf{P}}_{N+k}, \mathbf{h}_{k-1})$ // Update latent state
    end for
    // Reconstruct forecast
    $\hat{\mathbf{Y}} = \text{Reconstruct}(\hat{\mathbf{P}}_{N+1}, \ldots, \hat{\mathbf{P}}_{N+K})$
    // Compute loss (masked for variable horizons)
    $\mathscr{L} = \text{MSE}(\hat{\mathbf{Y}}[:H], \mathbf{Y})$
    // Backpropagation
    $\theta \leftarrow \theta - \alpha \nabla_\theta \mathscr{L}$
  end for
end forreturn θ

5.4.4 Semantic Anchoring via Frozen Language Models

A frozen GPT-2 backbone is used as a fixed contextualiser for reprogrammed patches. The only natural-language prepended to those patches is the shared prefix "Predict next values:"; it supplies a short left context for a forward pass with continuous embeddings and is not used to spell out variables, horizons, or concept inventories. Semantic coupling to language instead happens in reprogramming, where each patch token attends over embeddings of the strings in $\mathscr{C} = \{c_1, \ldots, c_M\}$: those strings are the tags readers attach to columns of the attention matrix when interpreting $A_{i,j}$ (Section 5.4.5). **Defined concepts** are therefore the main design object for readable explanations; a **learned-concept** ablation keeps the same attention machinery but forms keys and values from trainable linear mixtures of the full GPT-2 vocabulary, so columns need not align with single dictionary words.

Backbone choice (GPT-2). Our implementation requires a frozen transformer that (i) exposes a token embedding table for mapping short concept strings to fixed vectors and (ii) accepts continuous input embeddings in place of token indices, so that patch tokens can be contextualised without updating the backbone weights. Independently of these interface constraints, GPT-2 has become a recurrent backbone and reporting baseline in recent LLM-oriented time-series forecasting, which motivates its use as a conventional reference point: TimeCMA [15] freezes GPT-2 to extract prompt embeddings for cross-modality alignment, while LLM-TPF [19] builds on GPT-2 with parameter-efficient adaptation (LoRA) and evaluates against a standalone GPT-2 baseline in the same LLM-based model family. GPT-2 is adopted here as a pragmatic default: publicly released checkpoints are stable and widely available, the model is small enough to keep memory and per-batch latency manageable when the encoder is invoked throughout training, and its moderate width matches the dimensionality used for latent patch states without necessitating a separate projection bottleneck beyond the reprogramming layer. No claim is made that GPT-2 is uniquely optimal for semantic anchoring; other frozen text encoders with the same interface could be substituted.

Curating the defined vocabulary $\mathscr{C}$. $\mathscr{C}$ is a finite list of short English strings (GPT-

2-tokenizable words or phrases). These are drawn from a hand-maintained vocabulary specification that is bundled with the training implementation. The vocabulary is organised into eight core families: trend, seasonality, volatility, events, statistics, forecasting, mathematical, and temporal. In addition, there are five optional domain groups: financial, weather, energy, traffic, and airport/passenger operations. The effective $\mathscr{C}$ is resolved as follows: if the experiment supplies an explicit word list or selects which domain groups to include, that choice overrides the defaults; otherwise, when the dataset name indicates airport or passenger traffic, core is merged with the airport/passenger group only; in all other cases only the core families are used. Lists are concatenated in a fixed order and deduplicated while preserving first occurrence, which fixes $M = |\mathscr{C}|$: core only yields M=144; core plus airport/passenger yields M=185; merging core with every optional domain group yields M=249, listed in full in Table 5.1. The passenger-flow and airport experiments reported in this chapter use the M=185 configuration unless stated otherwise. Each c_j is tokenised with the frozen GPT-2 tokenizer, embedded through the frozen input-embedding table, and mean-pooled across sub-tokens when a concept spans multiple byte-pair pieces, yielding $\mathbf{e}_j \in \mathbb{R}^{d_{\text{LLM}}}$. Tables 5.8 and 5.9 highlight "noisy", "spike", "periodic", and "flat" among the diagnostics. For the learned-concept ablation, keys and values are formed from trainable linear combinations of all GPT-2 token embeddings rather than from an explicit list $\mathscr{C}$.

Reprogramming (concept-conditioned projection into LLM space). Given patch tokens $\mathbf{u}_i \in \mathbb{R}^{d_{\text{model}}}$, a lightweight multi-head attention layer is applied that uses the patch tokens as queries and the concept embeddings as keys/values:

$$\tilde{\mathbf{u}}_i = \text{Attn}(\mathbf{u}_i; \{\mathbf{e}_j\}_{j=1}^{M}) \in \mathbb{R}^{d_{\text{LLM}}}. \tag{5.6}$$

The fixed prefix "Predict next values:" is embedded with the frozen tokenizer, concatenated ahead of the reprogrammed patch sequence $\{\tilde{\mathbf{u}}_i\}$, and frozen GPT-2 is run on the resulting continuous embedding sequence. Only the patch positions are retained from the output as contextualised patch representations.

5.4.5 Concept-Based Explanation Mechanism

In the defined-concept model, concept relevance is read directly from the reprogramming attention weights. Let $A_{i,j}$ denote the attention mass assigned from patch token i to concept j (aggregated across heads). $A_{i,j}$ is interpreted as a concept score for patch i with respect to concept c_j. This supports two explanation views used throughout our analysis: (i) top-K concepts per patch (semantic explanations of each local segment), and (ii) top-K patches per concept (where in the series a concept is expressed).

Table 5.1 enumerates the M=249 strings obtained by merging the core families with every optional domain group, after global deduplication in a fixed canonical order. Several entries are written with an internal underscore so that the tokenizer treats them as single literal strings (e.g. *rush_hour*, *check_in*). Runs that use core with the airport/passenger group only (M=185, the default when the dataset name indicates airport or passenger traffic) omit strings that enter solely through the financial, weather, energy, or traffic groups; Table 5.1 is therefore a strict superset in terms of membership.

Table 5.1: Flattened curated vocabulary after merging all core and optional domain groups with deduplication (M=249). Airport-focused runs in this chapter use M=185 (core plus airport/passenger only).

up, rise, increase, grow, ascend, climb, surge, down, fall, decrease, decline, drop, descend, plunge, flat, steady,
stable, constant, unchanged, plateau, drift, wander, fluctuate, oscillate, vary, seasonal, periodic, cyclic, recurring, repetitive, daily, weekly,
monthly, quarterly, yearly, annual, hourly, minute, second, business, weekend, holiday, cycle, period, frequency, rhythm, pattern, noisy,
volatile, unstable, erratic, chaotic, smooth, calm, consistent, spiky, jumpy, irregular, random, stochastic, variance, volatility, fluctuation, variation,
spike, jump, leap, burst, crash, collapse, dip, peak, maximum, summit, top, apex, valley, minimum, trough, bottom,
nadir, shift, change, transition, break, breakpoint, changepoint, regime, phase, state, mean, average, median, mode, std, deviation,

Continued on next page

Table 5.1: Flattened curated vocabulary (M=249, continued).

spread, correlation, autocorrelation, dependence, stationary, nonstationary, trend_stationary, outlier, anomaly, exception, forecast, prediction, forecasting, projection, extrapolation, interpolation,

horizon, lead, lag, backcast, nowcast, forecast_error, linear, exponential, logarithmic, polynomial, sinusoidal, cosine, sine, wave, oscillation, amplitude,

wavelength, harmonic, resonance, damping, decay, morning, afternoon, evening, night, midnight, rush, commute, off_peak, business_hours, weekday, workday,

currency, exchange, rate, return, yield, risk, market, trading, portfolio, bull, bear, inflation, deflation, temperature, humidity, pressure,

precipitation, wind, cloud, sunny, rainy, snowy, storm, climate, weather, season, winter, summer, load, demand, supply, capacity,

utilization, consumption, power, energy, grid, network, traffic, flow, congestion, jam, gridlock, bottleneck, speed, velocity, throughput, density,

volume, lane, highway, freeway, road, street, intersection, rush_hour, peak_hour, travel, heavy, light, moderate, dense, sparse, accident,

incident, construction, delay, queue, passenger, traveler, visitor, arrival, departure, terminal, gate, checkpoint, security, customs, boarding, disembark,

line, wait, baggage, luggage, check_in, checkout, processing, busy, crowded, empty, flight, airline, schedule, departure_time, arrival_time, seasonal_travel,

holiday_travel, business_travel, leisure, international, domestic, connection, transfer, occupancy, load_factor

5.4.6 Latent Conditioning on Exogenous Information

For robustness analysis under exogenous shocks, exogenous information is incorporated through a latent conditioning formulation that allows autoregressive predictions to depend jointly on previously predicted target values and future exogenous trajectories, while maintaining a fixed representation of historical observations.

Historical target values and historical exogenous variables are jointly encoded to

produce an initial latent state that summarises observed passenger flow dynamics and the corresponding shock context. This encoding is computed once from observed data and remains unchanged during autoregressive prediction.

During prediction rollout, future exogenous information is introduced through a dedicated projection that conditions latent state updates on both the predicted target output and the corresponding exogenous input. Formally, at patch prediction step k (where k indexes patches in the rollout sequence), the model produces a latent update of the form

$$\mathbf{h}_{k+1} = \mathscr{T}\Big(\mathbf{h}_k,\ \mathrm{Proj}_{\mathrm{exog}}\big([\hat{\mathbf{P}}_k,\ \alpha \cdot \mathbf{x}_k]\big)\Big), \tag{5.7}$$

where $\hat{\mathbf{P}}_k \in \mathbb{R}^p$ denotes the predicted patch at step k, $\mathbf{x}_k \in \mathbb{R}^{d_{\mathrm{exog}}}$ denotes the corresponding future exogenous input for patch k (aggregated or sampled to match the patch temporal extent), α is a learnable gating parameter, $\mathrm{Proj}_{\mathrm{exog}}(\cdot)$ is a learned linear projection, and $\mathscr{T}(\cdot)$ denotes the recurrent state transition.

This formulation decouples historical encoding from future exogenous assumptions. As a result, future exogenous trajectories can be modified post-training without altering the learned historical representation, enabling efficient scenario-based analysis while preserving forecast stability.

5.5 Datasets and Experimental Setup

Evaluation is conducted on airport passenger-flow time series spanning multiple temporal resolutions. The setting is a multivariate series over the top 20 busiest US airports, with each variable representing one airport, evaluated at daily and hourly frequencies. These datasets are derived from publicly released aviation statistics and Freedom of Information Act (FOIA) requests [23]. External variables are constructed using GDELT, as detailed in Appendix A. Unless stated otherwise, the main forecasting experiments in this chapter use the same top-twenty hub panels and UTC calendar span as the airport case study in Chapter 4.5.7 (Table 4.5). Scenario-based sensitivity analysis under exogenous shocks

(Section 5.6.6) uses a shorter temporal subset, 13 March 2019–31 December 2021, so that train, validation, and test partitions all contain shock-affected periods; this subset applies only to that robustness protocol.

5.5.1 Dataset Statistics

The primary datasets used in our standard experiments are the same top-twenty hub panels described in Chapter 4.5.7. The daily dataset spans 30 December 2018 to 31 May 2025 with $C = 20$ airports and 2,345 day-aligned timesteps; the hourly panel shares the same date span with 56,280 hourly timesteps. The assembled target channels contain no missing values in the released panels (Table 4.5). All time series exhibit strong seasonality (daily, weekly, and annual patterns) and moderate-to-high variance, with coefficient of variation (CV) ranging from 0.3 to 0.8 across airports. The datasets include both stable periods and disruption periods (e.g., COVID-19 impacts), providing a realistic testbed for long-horizon forecasting under varying conditions.

Figure 5.2 shows a Pearson correlation heatmap of daily hub-level throughput and two z-scored national exogenous indices (GDELT-based policy-related aviation restriction and public-health stress; Appendix A), with variables ordered alphabetically by IATA code (airports) followed by the two indices. Correlations between each hub and each index are read from the cells where the corresponding airport row/column meets each index row/column (the lower-right block of the matrix). Figure 5.3 presents STL (Seasonal and Trend decomposition using Loess) for Atlanta (ATL) daily throughput (original, trend, weekly seasonal, residual), illustrating weekly seasonality, longer-run trend movements (including COVID-19-era changes), and residual variation.

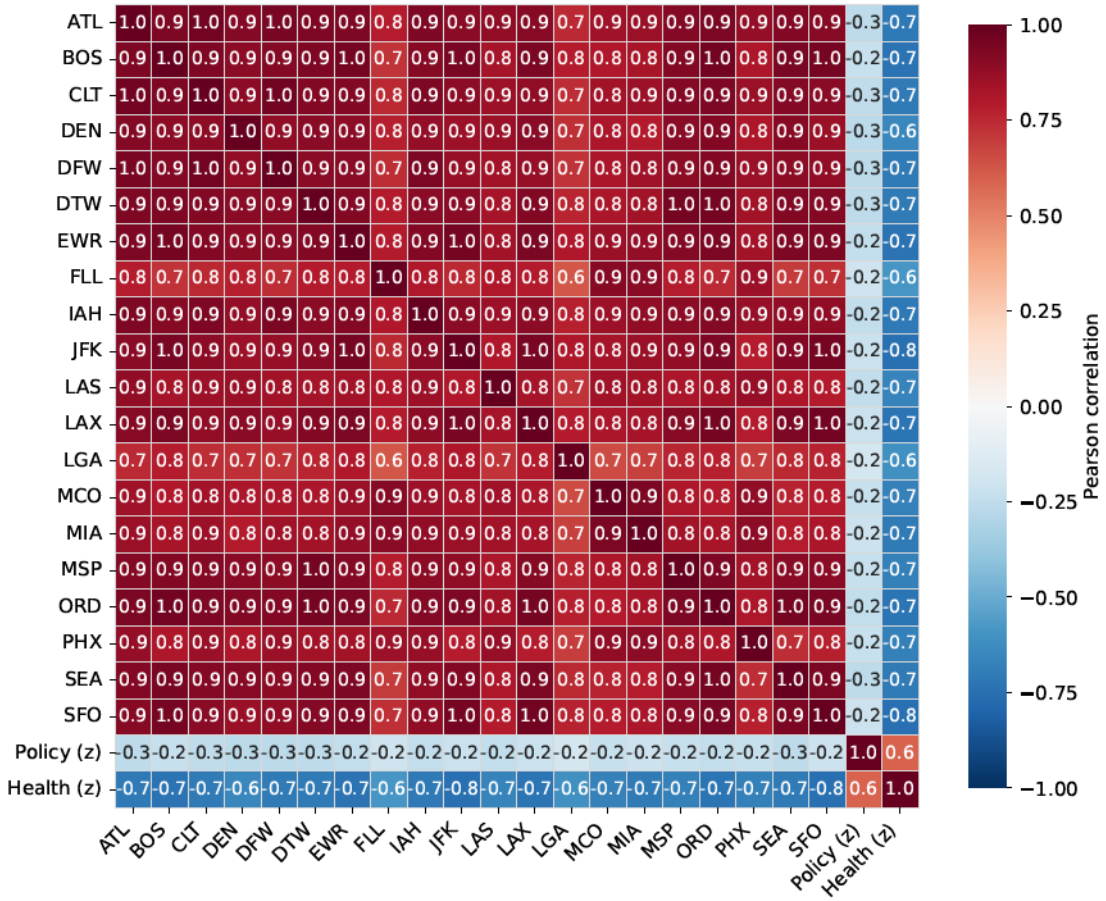


Figure 5.2: Pearson correlation heatmap for the top-twenty busiest US airports: daily TSA-reported throughput per airport and two z-scored national exogenous indices (policy-related aviation restriction and public-health stress; construction in Appendix A). Airport–index associations appear in the heatmap cells linking each IATA code with *Policy (z)* or *Health stress (z)*. The daily panel spans 30 December 2018 to 31 May 2025.

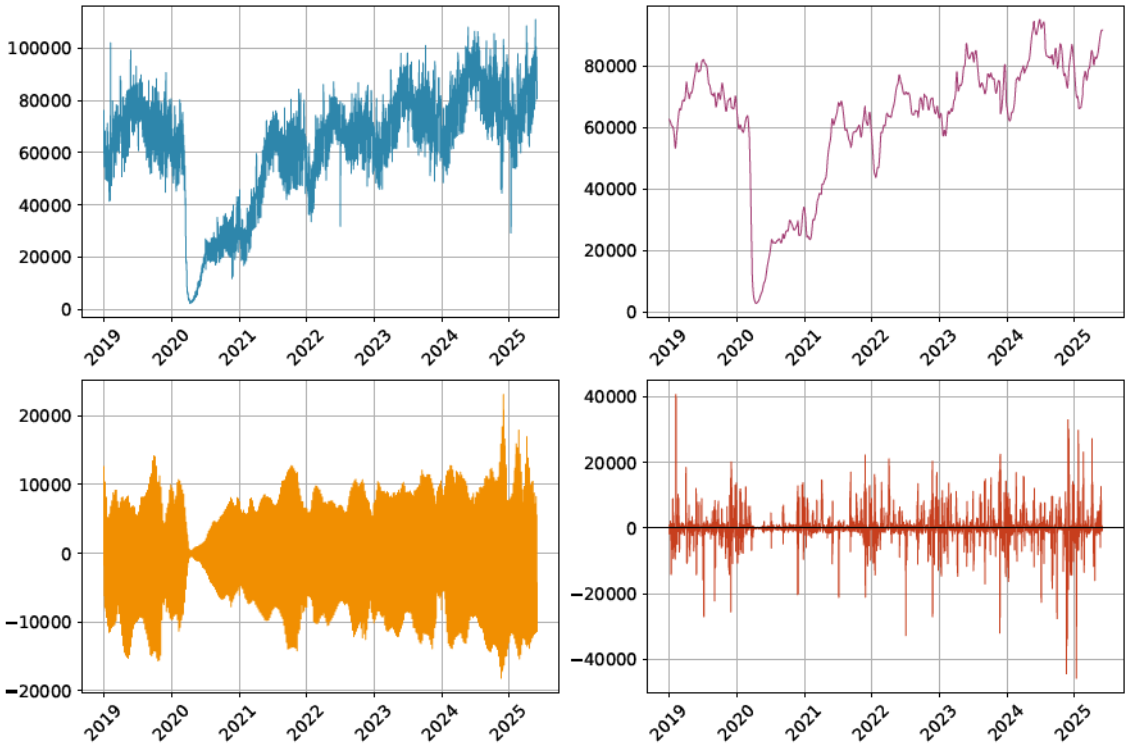


Figure 5.3: STL decomposition of Atlanta (ATL) daily throughput: original series, trend, weekly seasonal component, and residuals. Weekly seasonality, longer-run trend (including COVID-19-era shifts), and irregular residual variation are visible for this representative hub.

Datasets are split chronologically into train/val/test (70%/10%/20%): daily (1,641/235/469) and hourly (39,396/5,628/11,256) with C=20 airports. For daily data, horizons 24/48/96/192 are evaluated, and for hourly data, horizons 96/192/336/720 are evaluated. Unless otherwise stated, the context length is $L = 96$.

5.5.2 Implementation Details

Two-layer MLPs are used for the patch encoder and decoder, with patch embedding dimension 128. Latent dynamics are modelled by a Transformer with 4 layers, 8 attention heads, hidden dimension 256, and feedforward dimension 1024. Adaptive Moment Estimation (Adam) is used for optimisation (learning rate 10^{-4}, weight decay 10^{-5}), with batch size 32 and early stopping with patience 5 on validation loss.

Experiments are run on a single NVIDIA A100 GPU (24GB VRAM), with Python 3.8+, PyTorch 1.12+, and CUDA 11.0+.

5.6 Results

5.6.1 Multi-Horizon Forecasting Performance

Variable-horizon Patch-AR is evaluated on the Top-20 busiest U.S. airports datasets at daily and hourly granularities. Table 5.2 reports MSE/MAE across the evaluated horizon sets, comparing Patch-AR with iDecPred baseline.

Table 5.2: Top-20 busiest U.S. airports passenger flow forecasting performance at different granularities. Results are reported for a single variable-horizon Patch-AR model with defined concepts trained per granularity, compared with iDecPred baseline.

Granularity	Horizon	Patch-AR		iDecPred Baseline	
		MSE	MAE	MSE	MAE
Daily	24	0.244	0.319	–	–
	48	0.278	0.355	–	–
	96	0.325	0.406	0.329	0.422
	192	0.397	0.463	0.413	0.478
Hourly	96	0.103	0.204	0.106	0.207
	192	0.114	0.213	0.119	0.217
	336	0.126	0.225	0.131	0.229
	720	0.142	0.242	0.150	0.248

Performance degrades smoothly as the horizon increases. These results demonstrate that a single trained model can support multiple horizons within each granularity without retraining. Compared to iDecPred baseline, Patch-AR achieves competitive or better performance across all evaluated horizons.

5.6.2 Patch-AR vs Single-Step AR Ablation

Under matched backbone architectures, Patch-AR is compared against a single-step autoregressive baseline, differing in prediction granularity (patch-level vs step-level) and the autoregressive rollout mechanism. Tables 5.3 and 5.4 summarise the results.

Table 5.3: Patch-AR vs Single-Step AR (Daily)

Horizon	Patch-AR		Single-Step		Improvement
	MSE	MAE	MSE	MAE	
24	0.244	0.319	0.392	0.473	38% / 33%
48	0.278	0.355	0.413	0.492	33% / 28%
96	0.325	0.406	0.480	0.539	32% / 25%
192	0.397	0.463	0.533	0.573	26% / 19%

Table 5.4: Patch-AR vs Single-Step AR (Hourly)

Horizon	Patch-AR		Single-Step		Improvement
	MSE	MAE	MSE	MAE	
96	0.103	0.204	0.313	0.424	67% / 52%
192	0.114	0.213	0.436	0.496	74% / 57%
336	0.126	0.225	0.564	0.565	78% / 60%
720	0.142	0.242	0.824	0.701	83% / 66%

Patch-level rollouts substantially reduce error accumulation, with improvements increasing with horizon length.

The improvement pattern reveals the critical advantage of patch-level modelling at longer horizons. For daily data, the improvement grows from 38% at horizon 24 to 32% at horizon 96, then decreases slightly to 26% at horizon 192. This pattern suggests that patch-level rollouts provide the greatest benefit at intermediate horizons where error

accumulation becomes significant but not yet overwhelming. At very short horizons (24 steps), both methods perform relatively well, while at very long horizons (192 steps), even patch-level methods face inherent challenges.

For hourly data, the improvement pattern is more dramatic and consistent: improvements grow from 67% at horizon 96 to 83% at horizon 720, demonstrating that patch-level rollouts become increasingly advantageous as horizon length increases. The 83% improvement at horizon 720 (30 days ahead) is particularly notable, as it represents a substantial reduction in error accumulation that would be difficult to achieve with single-step methods. This pattern supports the hypothesis that patch-level modelling is especially effective for very long-horizon forecasting where error accumulation is most severe.

5.6.3 Concept Ablation Study

Three variants are evaluated: defined concepts (pre-specified interpretable concepts), learned concepts (data-driven concept embeddings), and no concepts (baseline without semantic anchoring). Defined concepts provide modest but consistent improvements over both learned and no-concept variants, particularly at longer horizons, suggesting a regularisation effect.

Figure 5.4 visualizes the relative performance change when using learned concepts compared to defined concepts, showing both MSE and MAE metrics across different horizons. For daily data (Figure 5.4, panel **a**), the relative change is near zero at short horizons (horizon 24: MSE change $< 0.1\%$, MAE change $\approx 3.1\%$), indicating minimal difference between defined and learned concepts at short forecast lengths. As horizon increases, the performance gap widens: at horizon 192, learned concepts show 4.8% higher MSE and 4.6% higher MAE compared to defined concepts. This pattern suggests that semantic anchoring provides a stabilising regularisation effect that becomes more critical as autoregressive rollouts extend, with predefined concepts helping to maintain forecast stability over long horizons.

For hourly data (Figure 5.4, panel **b**), the pattern is similar but with smaller absolute

differences. The relative changes are minimal at short horizons (horizon 96: MSE and MAE changes $< 0.5\%$) and grow to 3.9% (MSE) and 1.1% (MAE) at horizon 720. The smaller differences at hourly resolution may reflect the finer temporal granularity, providing more stable local patterns that are less dependent on semantic regularisation.

Notably, both panels show that MAE changes are generally smaller than MSE changes, suggesting that semantic anchoring primarily affects the variance of errors rather than their mean magnitude. The consistent pattern across both resolutions and both metrics supports the hypothesis that predefined semantic concepts act as a form of inductive bias that yields modest improvements to long-horizon forecasting stability.

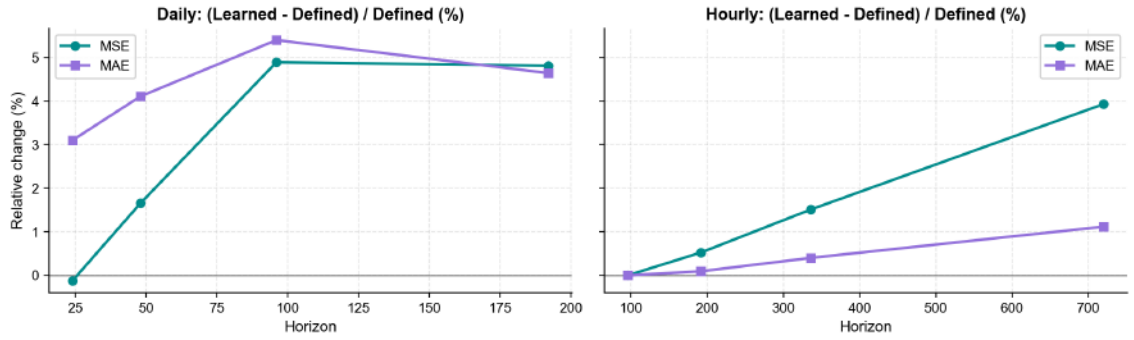


Figure 5.4: Concept ablation (relative MSE and MAE change vs defined concepts). Both MSE and MAE show the relative change when using learned concepts compared to defined concepts.

Tables 5.5 and 5.6 report complete concept-ablation results across all horizons, complementing the aggregated view in Figure 5.4.

Several patterns emerge from the full ablation results. First, defined concepts yield modest but consistent improvements over learned and no-concept variants across the evaluated horizons, with the gap widening at longer horizons. For daily data, the improvement grows from negligible at horizon 24 (MSE difference $< 0.1\%$) to 4.8% at horizon 192 (MSE 0.397 vs. 0.416). This pattern suggests that semantic anchoring provides a stabilising regularisation effect that becomes more critical as autoregressive rollouts expand. Second, learned and no-concept variants produce identical results. The learned-concept variant performs similarly to the no-concept baseline, suggesting that unconstrained concept embeddings do not form meaningful semantic structure without explicit inductive bias; this

highlights the role of predefined concepts as a structured semantic prior. Third, hourly models show smaller absolute differences but similar relative patterns, with defined concepts providing 3.9% improvement at horizon 720 (MSE 0.142 vs. 0.148). These results support the hypothesis that predefined semantic concepts act as a form of inductive bias that improves long-horizon forecasting stability.

Table 5.5: Full Concept Ablation (Daily, All Horizons)

Horizon	Defined		Learned		None	
	MSE	MAE	MSE	MAE	MSE	MAE
24	0.244	0.319	0.244	0.329	0.244	0.329
48	0.278	0.355	0.282	0.370	0.282	0.370
96	0.325	0.406	0.341	0.428	0.341	0.428
192	0.397	0.463	0.416	0.485	0.416	0.485

Table 5.6: Full Concept Ablation (Hourly, All Horizons)

Horizon	Defined		Learned		None	
	MSE	MAE	MSE	MAE	MSE	MAE
96	0.103	0.204	0.103	0.204	0.103	0.204
192	0.114	0.213	0.115	0.214	0.115	0.214
336	0.126	0.225	0.128	0.226	0.128	0.226
720	0.142	0.242	0.148	0.244	0.148	0.244

5.6.4 Concept-Based Explanations

Semantic concept scoring produces a score for each input patch and each concept in a fixed vocabulary (e.g., spike, periodic, flat, noisy). These scores are used to provide two complementary qualitative views: (i) a per-patch summary given by the top-ranked

concepts, and (ii) a per-concept localization that highlights which patches in the input window most strongly express a given concept.

Figure 5.5 illustrates concept-based explanations for a single input window, demonstrating how semantic concepts map to temporal patterns. The top panel shows the input time series with patch boundaries marked, providing context for the temporal structure. The bottom panel shows a heatmap of concept scores across patches, where each row represents a concept and each column represents a patch index. Concepts are selected based on highest activation in this window, and the heatmap reveals which patches express which concepts most strongly. For example, patches with high "spike" scores correspond to rapid increases in the time series, while patches with high "periodic" scores align with recurring patterns. This visualisation enables users to understand what temporal patterns the model is detecting and how these patterns relate to interpretable semantic concepts.

Figure 5.6 extends this analysis to multiple variables (airports), showing concept scores across variables for aligned patch indices. Each row represents a variable (airport), and each column represents a patch index over the same context window. The heatmap reveals which airports express which concepts at which temporal positions, enabling cross-variable comparisons. For instance, if multiple airports show high "spike" scores at the same patch index, this may indicate a system-wide event affecting multiple locations simultaneously. The z-scored visualisation ensures that concept scores are comparable across variables with different scales, making it easier to identify shared patterns and variable-specific characteristics.

These visualizations demonstrate that concept-based explanations provide interpretable, human-readable descriptions of model behaviour that align with intuitive temporal semantics, supporting their use as semantic explanations of patch-level dynamics.

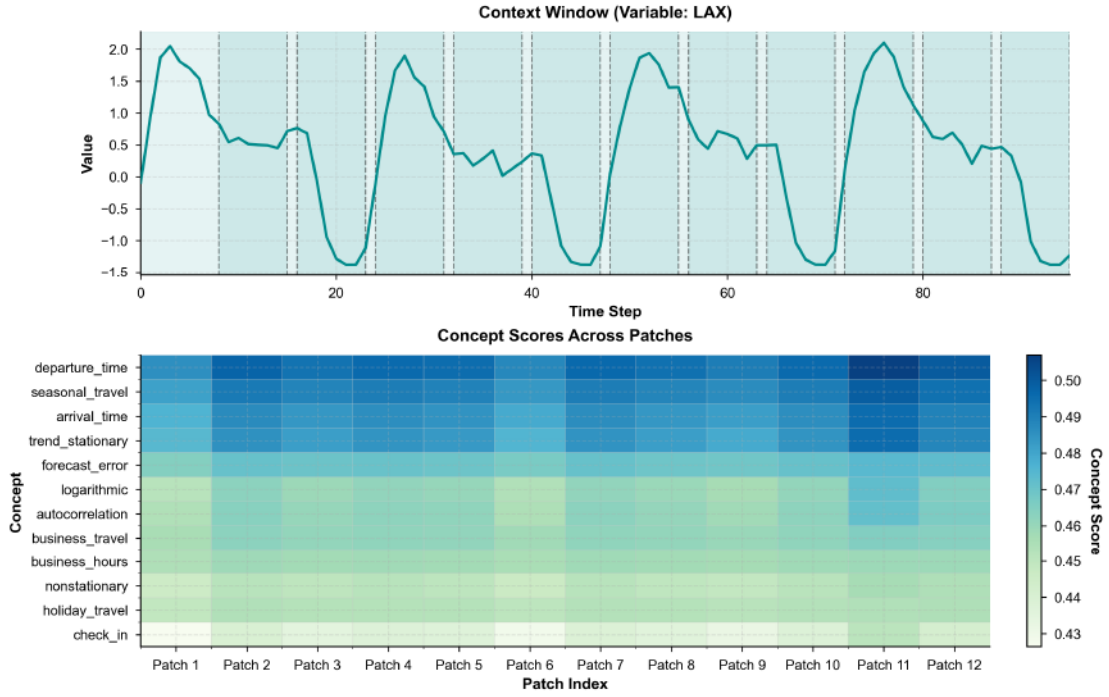


Figure 5.5: Concept-based explanations for one input window. Top: the selected context window (one variable) with patch boundaries. Bottom: heatmap of concept scores over patches; concepts are selected by highest activation in this window (max over patches).

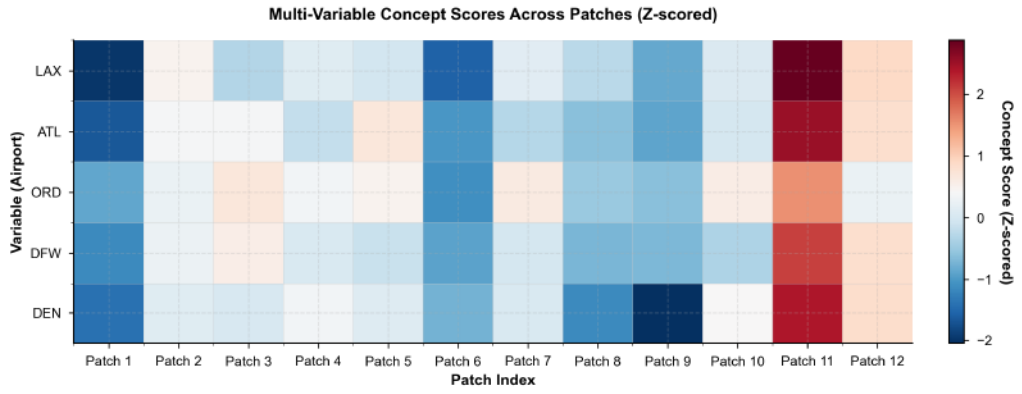


Figure 5.6: Multi-variable concept score heatmap at aligned patch indices. Rows are variables (airports) and columns are patch indices over the same context window; colors show the selected concept score across variables for each patch index (z-scored for visualisation).

5.6.5 Diagnostic Evaluation of Explanation Quality

Concept scores are intended as human-readable explanations of patch-level dynamics (e.g., "spike", "flat", "periodic"). This section evaluates semantic validity in two ways: stability of concept rankings across random seeds and under small input noise, and Spearman

alignment between concept scores and simple patch-level statistics.

Table 5.7 demonstrates exceptional stability of concept-based explanations. The Top-5 Jaccard similarity is 1.000 (perfect match) across different random seeds and under small input noise (std = 0.01), indicating that concept rankings are highly reproducible and robust to minor perturbations. This stability is crucial for interpretability applications, as it ensures that explanations remain consistent across different runs and are not artifacts of random initialization or minor data variations. The perfect stability under noise suggests that concept scores capture robust semantic patterns rather than noise-sensitive features.

Table 5.7: Stability Across Reruns and Noise

Setting	Seeds	Noise std	n_{win}	Top-5 Jaccard (mean / median)
Base repeatability	42, 43	0.00	200	1.000 / 1.000
Small-noise robustness	42, 43, 44	0.01	200	1.000 / 1.000

Table 5.8 quantifies the alignment between concept scores and intuitive patch-level statistics using Spearman rank correlations. The results show meaningful correlations that align with semantic expectations: the “noisy” concept shows positive correlations with variance (0.324) and range (0.334), while the “flat” concept shows negative correlations with variance (-0.151) and range (-0.171), as expected. The “spike” and “periodic” concepts also show positive correlations with variance (0.318 and 0.339, respectively), suggesting that these concepts capture patterns associated with high variability. These correlations provide evidence that concept scores reflect interpretable temporal characteristics, supporting their use as semantic explanations of patch-level behaviour.

Table 5.8: Concept-Score Anchoring to Patch Statistics (Selected Concepts)

Concept	ρ(slope)	ρ(var)	ρ(max_abs_diff)	ρ(range)
noisy	-0.080	0.324	0.263	0.334
spike	-0.051	0.318	0.252	0.330
periodic	-0.045	0.339	0.275	0.336
flat	-0.116	-0.151	-0.043	-0.171

Table 5.9 reports window-level significance tests for concept-statistic correlations, providing statistical evidence for the alignment patterns observed in Table 5.8. All reported concept-statistic pairs show highly significant correlations ($p < 10^{-12}$) with median Spearman ρ values ranging from 0.29 to 0.34 for positive correlations and -0.19 to -0.21 for negative correlations. The 95% confidence intervals are narrow and do not include zero, confirming that the observed correlations are statistically robust across windows. These significance tests strengthen the evidence that concept scores capture meaningful semantic patterns rather than random associations, supporting their validity as interpretable explanations.

Table 5.9: Window-Level Significance (Selected Concept-Statistic Pairs)

Concept	Stat	n_{win}	Median ρ	95% CI	p / q
noisy	var	200	0.294	[0.275, 0.326]	1.24×10^{-60} / 3.65×10^{-60}
noisy	range	200	0.313	[0.291, 0.342]	1.24×10^{-60} / 3.65×10^{-60}
spike	var	200	0.290	[0.271, 0.322]	1.24×10^{-60} / 3.65×10^{-60}
periodic	var	200	0.295	[0.268, 0.323]	1.24×10^{-60} / 3.65×10^{-60}
flat	var	200	-0.188	[-0.206, -0.150]	8.39×10^{-13} / 9.72×10^{-13}
flat	range	200	-0.208	[-0.231, -0.165]	2.21×10^{-21} / 2.78×10^{-21}

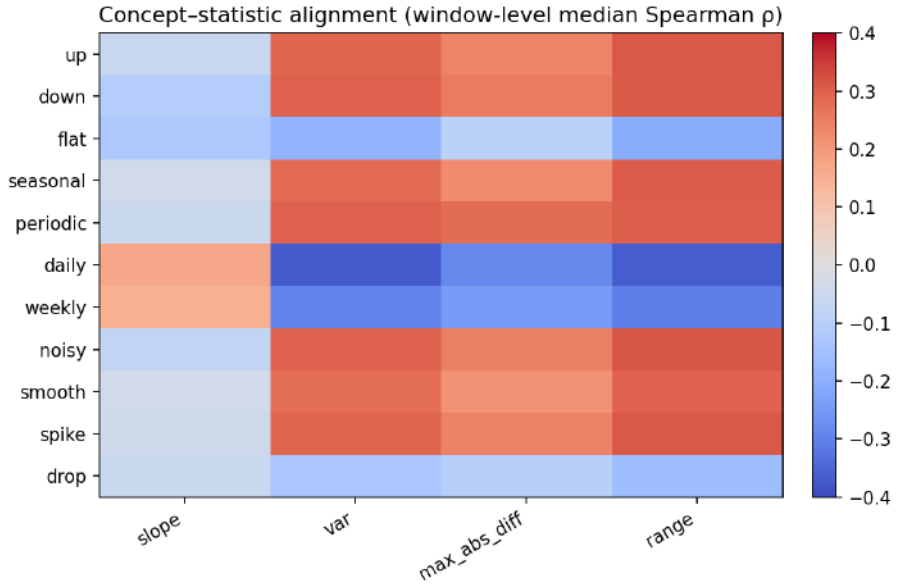


Figure 5.7: Concept-score anchoring to patch statistics (median Spearman ρ heatmap)

Figure 5.7 provides a comprehensive view of concept-statistic alignment across all concept-statistic pairs, complementing the selected results in Table 5.8. The heatmap visualizes median Spearman rank correlations, with color intensity indicating correlation strength and sign. The heatmap reveals clear patterns: concepts associated with high variability (noisy, spike, periodic) show positive correlations with variance and range statistics, while concepts associated with low variability (flat) show negative correlations. These patterns provide visual confirmation that concept scores capture interpretable temporal characteristics, supporting their validity as semantic explanations. The consistency of these patterns across multiple concept-statistic pairs strengthens the evidence that semantic anchoring successfully aligns patch embeddings with interpretable temporal semantics.

5.6.6 Robustness under Exogenous Shocks via Scenario-Based Sensitivity Analysis

Robustness under exogenous shocks is evaluated by comparing baseline forecasts (using the original future exogenous trajectory) to scenario forecasts (using a perturbed future exogenous trajectory), while keeping historical observations fixed. For a fixed window

ending at time t, the step-wise forecast difference is

$$\Delta(h) = \hat{x}_{t+h}^{\text{scenario}} - \hat{x}_{t+h}^{\text{baseline}}, \quad h = 1, \ldots, H, \tag{5.8}$$

where $\hat{x}_{t+h}$ denotes the predicted value at absolute time $t+h$. The window end t corresponds to the forecast origin T in Section 5.3; multivariate metrics there use $\hat{x}_{T+h,v}$ with variate index v.

Dataset and splits. For robustness analysis under exogenous shocks, a temporal subset of the system-level datasets spanning March 13, 2019 to December 31, 2021 is used to ensure that train, validation, and test splits all contain periods with exogenous shock patterns (e.g., COVID-19 impacts). This time period selection ensures that each data split includes both pre-shock and shock-affected periods, enabling meaningful evaluation of scenario-based sensitivity across all data splits. The temporal subset is then split chronologically into train/validation/test partitions using the same 70%/10%/20% procedure as described in Section 5.5. The key difference from the no-shock baseline setting is that models are trained with exogenous shock features included, enabling evaluation of how forecasts respond to counterfactual future shock trajectories. All scenario-based sensitivity analyses are conducted on the same validation and test splits to ensure comparability with baseline results.

Scenario design. Three interpretable scenario classes are considered: (i) *zero* (remove future shock effects by setting exogenous inputs to zero), (ii) *lastx amplification* (scale the most recent observed shock level by a factor, e.g. 10×), and (iii) *constant extreme-value* (fix future exogenous inputs at their historical maxima as a stress test). These scenarios are not intended as causal interventions; they provide controlled perturbations to probe counterfactual model behaviour.

Robustness diagnostics. Forecast deltas are reported under each scenario and directional consistency is summarised via the sign distribution of $\Delta(h)$ across the horizon.

5.6.6.1 Accuracy with and without exogenous shocks

Models trained with and without exogenous shock inputs are compared at daily and hourly resolutions, using a 96-step input and 96-step prediction window. When shocks are included, contemporaneous raw indicators only are used: a policy-related travel-restriction channel and a public-health stress channel, aligned with each time step, with no explicitly constructed lagged shock features in the runs reported here. The no-shock baseline omits both channels from the model input.

Table 5.10 reports MAE/MSE on validation and test splits for the with-shock configuration (both raw channels) against this no-shock baseline.

Table 5.10: Forecasting accuracy comparison across temporal resolutions (96-step horizon): no exogenous shock inputs vs. both contemporaneous raw shock channels.

Resolution	Split	MAE		MSE	
		Without Exog	With Exog	Without Exog	With Exog
Daily	Validation	0.544	0.493	0.521	0.435
	Test	0.573	**0.503**	0.544	**0.435**
Hourly	Validation	0.144	0.142	0.035	0.034
	Test	0.150	**0.144**	0.050	**0.046**

Figure 5.8 visualizes the accuracy comparison across temporal resolutions, showing both MAE and MSE for models with and without exogenous features. For daily resolution (panel a), both MAE and MSE show clear improvements when exogenous features are included: MAE decreases from 0.573 to 0.503 (12.2% improvement), while MSE decreases from 0.544 to 0.435 (20.0% improvement). The larger relative improvement in MSE suggests that exogenous features help reduce both the magnitude and variance of errors, with the variance reduction being particularly pronounced.

For hourly resolution (panel b), the improvements are smaller in relative terms but still meaningful: MAE decreases from 0.150 to 0.144 (3.7% improvement, calculated from source data), while MSE decreases from 0.050 to 0.046 (8.0% improvement). The smaller

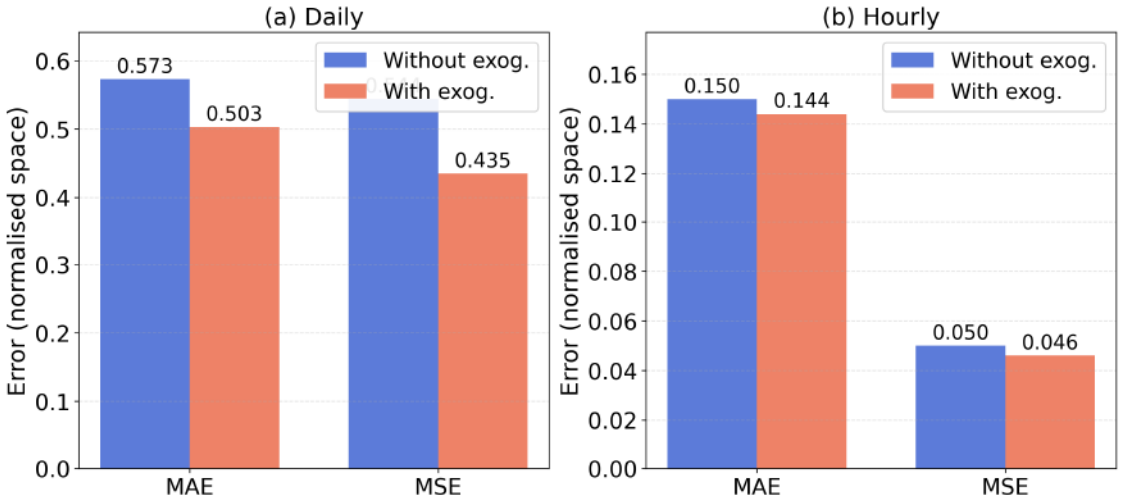


Figure 5.8: Forecasting performance comparison across temporal resolutions. Panel (a) shows Daily resolution with MAE and MSE for models with and without exogenous features. Panel (b) shows Hourly resolution with MAE and MSE. Hourly models achieve lower absolute error while daily models show larger relative improvements from exogenous features.

relative improvements at hourly resolution may reflect the already lower baseline error, making further improvements more challenging. However, the absolute error values are substantially lower at hourly resolution (MAE: 0.144 vs. 0.503 for daily), indicating that finer temporal granularity provides inherent advantages for forecasting accuracy.

Two observations merit attention. First, accuracy gains are stronger at daily resolution despite hourly models achieving lower absolute error (test MAE: 0.144 vs. 0.503), suggesting that exogenous information provides greater relative benefit when temporal aggregation smooths short-term dynamics. Second, the consistent improvement across validation and test splits indicates robust generalisation of the exogenous feature contribution.

5.6.6.2 Computational Complexity Analysis

The computational complexity of our method scales favourably with the horizon length due to patch-level operations. For a forecast horizon H and patch length p, the number of autoregressive steps is $K = \lceil H/p \rceil$, which is substantially smaller than H for typical patch sizes (e.g., $p = 7$ for daily data reduces 192 steps to 28 patch-level steps). This reduction in rollout steps directly translates to lower computational cost and reduced error accumulation.

Training complexity is $O(N \cdot |\mathscr{H}_{\text{train}}| \cdot K \cdot d_{\text{model}})$, where N is the number of training samples, $|\mathscr{H}_{\text{train}}|$ is the number of training horizons, K is the average number of patch steps, and d_{model} is the model embedding dimension (consistent with the notation in Section 5.4). The variable-horizon training adds a factor of $|\mathscr{H}_{\text{train}}|$ compared to fixed-horizon training, but this is typically small (e.g., 4 horizons) and is offset by the benefit of a single model handling all horizons.

Inference complexity is $O(K \cdot d_{\text{model}})$ per forecast, independent of the number of horizons the model was trained on, making it efficient for deployment scenarios requiring multiple horizon predictions.

Table 5.11 summarises computational costs for Patch-AR with and without exogenous features. Incorporating exogenous features increases training time by 11–22% with minimal parameter growth ($< 0.1\%$). Notably, hourly models exhibit a 13.5% inference speedup despite increased training time, suggesting that exogenous conditioning signals improve inference efficiency. The modest computational overhead is justified by accuracy improvements, particularly at hourly resolution where inference efficiency also improves.

The efficiency analysis reveals interesting resolution-dependent patterns. Daily models show a larger training time increase (22.0%) compared to hourly models (11.2%), likely due to the relative simplicity of daily patterns requiring more training iterations to learn exogenous dependencies. Conversely, hourly models benefit from an inference speedup (13.5% reduction in inference time), possibly because exogenous conditioning provides stronger signal-to-noise ratios that enable faster convergence during autoregressive rollout. The parameter count remains essentially unchanged ($< 0.1\%$ increase), confirming that exogenous integration is achieved through efficient latent conditioning rather than architectural expansion. These efficiency characteristics make the method practical for deployment scenarios where both accuracy and computational cost matter.

Table 5.11: Computational efficiency comparison across resolutions and models. Inference time is for a single 96-step prediction. Parameter counts are in millions (M).

Resolution	Model	Training Time (s)	Inference (s)	Params (M)	Epochs
Daily	w/o Exog	16.93	1.37	60.84	6
Daily	w/ Exog	20.66 (+22%)	1.60 (+17%)	60.87	7
Hourly	w/o Exog	640.92	28.98	60.86	14
Hourly	w/ Exog	712.84 (+11%)	25.06 (-14%)	60.91	16

5.6.6.3 Directional Response under Scenario Perturbations

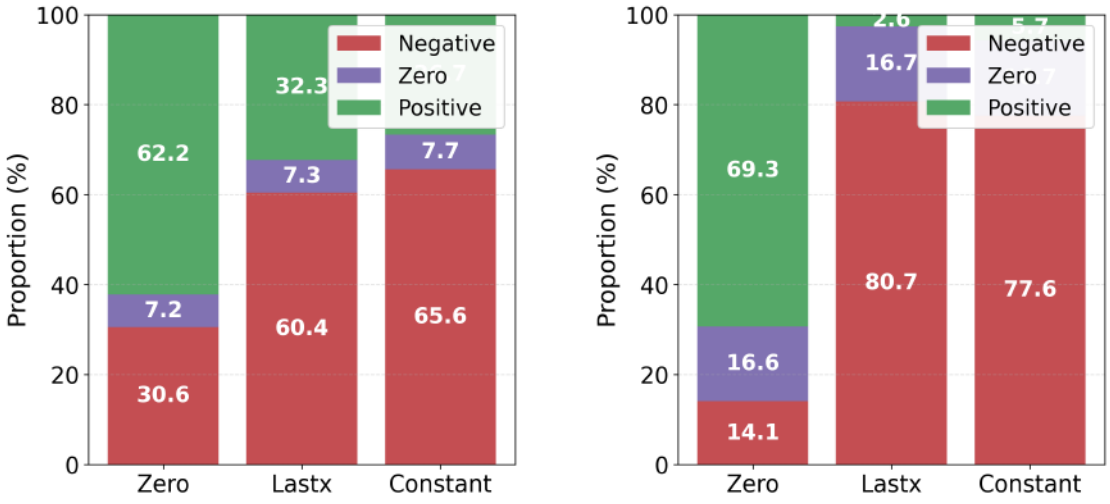
The sign distribution of forecast deltas is examined under each scenario to understand whether models respond coherently to scenario perturbations. Figure 5.9 visualizes the proportion of forecast steps with negative, zero, and positive deltas for each scenario type, revealing clear resolution-dependent patterns.

For daily models (Figure 5.9a), the sign distributions vary substantially across scenarios. Under the zero scenario (removal of exogenous effects), 62.2% of steps show positive deltas, indicating that removing exogenous information generally increases forecasts, while only 30.6% show negative deltas. This pattern suggests that exogenous information at daily resolution tends to suppress forecasts, so removing it leads to higher predictions. Under the lastx scenario (amplification), the pattern reverses: 60.4% of steps show negative deltas and 32.3% show positive deltas, indicating that amplifying exogenous shocks generally decreases forecasts. The constant scenario shows a similar pattern to lastx (65.6% negative, 26.7% positive), suggesting that extreme-value scenarios produce consistent directional responses.

For hourly models (Figure 5.9b), the sign distributions show much stronger directional consistency. Under the lastx scenario, 80.7% of steps show negative deltas and only 2.6% show positive deltas, demonstrating highly coherent responses to amplification. The constant scenario shows similar coherence (77.6% negative, 5.7% positive), while the zero scenario shows the opposite pattern (14.1% negative, 69.3% positive), consistent with the daily pattern but more pronounced.

These patterns indicate that hourly forecasts exhibit stronger directional consistency

than daily forecasts under amplification/stress-test scenarios, reflecting the finer temporal resolution providing more direct coupling between exogenous shocks and forecast trajectories. The high proportion of negative deltas under amplification scenarios (80.7% for hourly lastx) suggests that hourly models interpret amplified shocks as demand-reducing factors, producing coherent downward adjustments in forecasts.



(a) Daily **(b)** Hourly

Figure 5.9: Delta sign proportions for daily and hourly models on the test split. Each bar shows the proportion of forecast steps with negative, zero, and positive deltas under different scenarios (zero, lastx, constant).

The quantitative diagnostics above are computed in the offline evaluation pipeline described in this subsection. Figure 5.10 illustrates a prototype interactive view of the same scenario-based workflow for airport passenger throughput: users can configure shock scenarios (consistent with the zero, lastx, and constant stress-test classes), compare baseline versus scenario forecast traces, and inspect summary sensitivity indicators side by side. The screenshot is intended as a deployment-oriented mock-up, it does not introduce additional experiments beyond the protocol and metrics already reported.

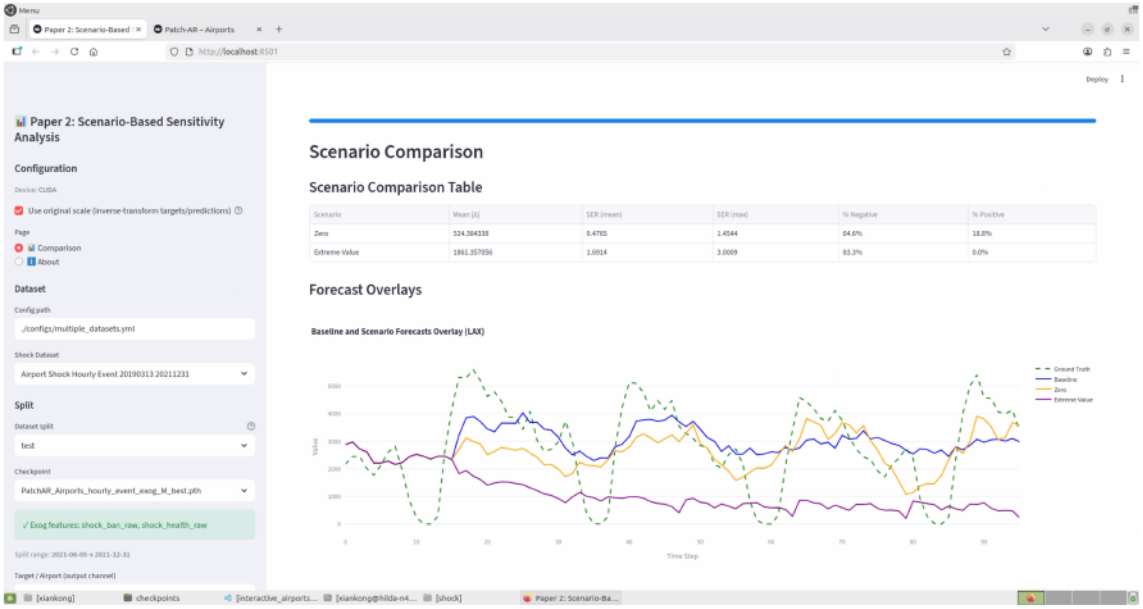


Figure 5.10: Interactive scenario-based sensitivity analysis interface for airport passenger flow forecasting. The interface allows users to configure scenarios, visualize forecast comparisons, and examine sensitivity metrics across different shock conditions.

5.6.7 Additional Accuracy-Related Results

5.6.7.1 Single-Horizon Sanity Runs

To check whether variable-horizon training compromises accuracy at matched horizons, single-horizon sanity runs are reported and compared with the corresponding variable-horizon model evaluated at the same horizons (Table 5.12).

The comparison reveals a small but consistent accuracy trade off. Single horizon models achieve marginally better performance (for example, daily horizon 96: MSE 0.310 versus 0.325, a 4.6% improvement; hourly horizon 96: MSE 0.100 versus 0.103, a 3.4% improvement). This trade off is expected because single horizon training allows the model to specialize for a specific forecast length. However, the performance gap is modest (typically less than 5%). The practical benefits of variable horizon training, namely allowing a single model to handle multiple horizons without retraining, outweigh this small accuracy cost. The consistency of this pattern across both daily and hourly resolutions suggests that the trade off is inherent to the variable horizon training objective rather than specific to a particular resolution.

Table 5.12: Variable-horizon versus single-horizon training at matched horizons. Variable-horizon results are taken from Table 5.2.

Granularity	Horizon	Variable-horizon		Single-horizon	
		MSE	MAE	MSE	MAE
Daily	96	0.325	0.406	0.310	0.399
Hourly	96	0.103	0.204	0.100	0.202
	192	0.114	0.213	0.113	0.212

5.7 Analysis and Discussion

5.7.1 Why Variable-Horizon Training Works

Variable-horizon training mitigates horizon specialisation that commonly arises when models are optimised for a single fixed forecast length. By sampling horizons during training, a single model is encouraged to learn latent dynamics that transfer across rollout lengths rather than relying on horizon-specific mappings. Patch-level representations plausibly support this transferability because each patch summarises a local window, so the same latent dynamics can be unrolled for different numbers of patch steps without relabelling the patch content by horizon. Empirically, one variable-horizon model remains competitive across the horizon grids reported for Patch-AR (Table 5.2). At the matched horizons in Table 5.12, dedicated single-horizon training is only marginally stronger, indicating a modest accuracy cost for horizon-agnostic training in this comparison (as detailed in the Single-Horizon Sanity Runs discussion).

5.7.2 Role of Semantic Anchoring

Semantic anchoring contributes to both predictive performance and interpretability. In the concept ablation study (Figure 5.4; Tables 5.5 and 5.6), predefined concepts match or

improve on the learned-concept and no-concept variants at every reported horizon, with the hourly horizon 96 case showing no numerical separation between variants and larger gaps emerging at longer horizons. This pattern is consistent with semantic anchoring acting as a stabilising inductive bias during long autoregressive rollouts where explicit concepts matter most. For interpretability, concept scores are stable across reruns and small perturbations (Table 5.7) and show statistically supported association with simple patch-level statistics (Tables 5.8 and 5.9; Figure 5.7), supporting their use as human-readable explanations of patch-level dynamics.

5.7.3 Robustness under Exogenous Shocks

Exogenous features improve test-set point accuracy at both resolutions (Table 5.10), while scenario perturbations reveal resolution-dependent response patterns. Using delta sign proportions as a diagnostic (Figure 5.9), hourly forecasts exhibit markedly stronger directional consistency under amplification and stress-test scenarios, whereas daily forecasts show more mixed responses. Taken together, this is consistent with temporal resolution shaping how exogenous information is reflected in rollout behaviour under the scenarios used here, and supports reporting scenario-based robustness diagnostics alongside point accuracy when shocks are a concern.

5.8 Conclusion

This chapter studies long-horizon forecasting by learning patch-level latent dynamics under a variable-horizon training objective, so one model can be rolled out to multiple forecast lengths within the horizon grids used for training and evaluation without retraining. On Top-20 U.S. airports, Patch-AR matches or improves on an iDecPred baseline across the reported daily and hourly horizons. In addition, a matched-architecture ablation that differs only in the autoregressive mechanism shows materially lower MSE for patch-level rollouts than for single-step autoregressive rollouts, with the gap increasing toward longer horizons.

Semantic anchoring regularises patch embeddings against concept embeddings from a frozen language model used as a semantic reference, not as a forecaster. In the reported concept ablations, predefined concepts match or improve on learned-concept and no-concept variants at each evaluated horizon, with negligible separation at short horizons (including tied metrics for all variants at hourly horizon 96) and larger differences emerging at longer horizons. Separate diagnostics indicate stable concept rankings under rerun and small input noise and statistically supported rank association with simple patch-level statistics.

For settings with exogenous shock covariates, test-set accuracy is reported with and without exogenous inputs, and responses of step-wise forecast deltas to counterfactual future exogenous paths are summarised. Under the scenarios implemented here, hourly forecasts display more directional consistency in those deltas than daily forecasts, whereas claims about operational robustness would require additional criteria and external validation beyond this case study.

The present implementation adopts a fixed patch length, which may limit its ability to capture temporal heterogeneity in real-world time series. Adaptive or multi-scale patchification offers a promising extension for modelling such variability [4, 3]. Evaluating and integrating these mechanisms within the Patch-AR framework remains an important direction for future work.

References

[1] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[2] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[3] Yizhi Cao, Zijian Tian, Wenjie Guo, and Xinggao Liu. Mspatch: A multi-scale patch mixing framework for multivariate time series forecasting. *Expert Systems with Applications*, 273:126849, 2025. doi: 10.1016/j.eswa.2025.126849.

[4] Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. *arXiv preprint arXiv:2402.05956*, 2024.

[5] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 (artificial intelligence act), 2024. URL `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ%3AL_202401689`. Official Journal of the European Union, L series, 12 July 2024. Article 13 addresses transparency obligations for certain AI systems.

[6] European Union Aviation Safety Agency (EASA). Artificial intelligence roadmap 2.0: Human-centric approach to ai in aviation. Technical report, EASA, May 2023. URL `https://www.easa.europa.eu/sites/default/files/dfu/easa_ai_roadmap_2_0_2023.pdf`. Version 2.0.

[7] Shanghua Gao, Teddy Koker, Owen Queen, Tom Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. Units: A unified multi-task time series model. *Advances in Neural Information Processing Systems*, 37:140589–140631, 2024.

[8] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, pages 261–269, 2018. doi: 10.1145/3159652.3159667.

[9] Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. GPT4MTS: Prompt-based large language model for multimodal time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23343–23351, 2024. doi: 10.1609/aaai.v38i21.30383.

[10] Yushan Jiang, Wenchao Yu, Geon Lee, Dongjin Song, Kijung Shin, Wei Cheng, Yanchi Liu, and Haifeng Chen. Timexl: Explainable multi-modal time series prediction with llm-in-the-loop. *arXiv preprint arXiv:2503.01013*, 2025.

[11] Ming Jin, Shiyu Wang, Lijun Ma, Zhixuan Chu, J. Y. Zhang, Xiaoming Shi, et al. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2024.

[12] Kai Kim, Howard Tsai, Rajat Sen, Abhimanyu Das, Zihao Zhou, Abhishek Tanpure, Mathew Luo, and Rose Yu. Multi-modal forecaster: Jointly predicting time series and textual data. *arXiv preprint arXiv:2411.06735*, 2024.

[13] Hans Krupakar and V. A. Kandappan. A review of the long horizon forecasting problem in time series analysis. *arXiv preprint arXiv:2506.12809*, 2025.

[14] Bryan Lim, Sercan Ö. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012.

[15] Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. TimeCMA: Towards LLM-empowered multivariate time series forecasting via cross-modality alignment. *arXiv preprint arXiv:2406.01638*, 2024.

[16] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lijun Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2024.

[17] Massimiliano Marcellino, James H Stock, and Mark W Watson. A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135(1-2):499–526, 2006.

[18] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2023.

[19] Qihong Pan, Haofei Tan, Guojiang Shen, Xiangjie Kong, Mengmeng Wang, and Chenyang Xu. LLM-TPF: Multiscale temporal periodicity-semantic fusion LLMs for time series forecasting. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 25*, 2025. doi: 10.24963/ijcai.2025/671.

[20] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020. doi: 10.1016/j.ijforecast.2019.07.001.

[21] Souhaib Ben Taieb, Gianluca Bontempi, Amir F Atiya, and Antti Sorjamaa. A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition. *Expert Systems with Applications*, 39(8):7067–7083, 2012.

[22] H. Tang, J. Yu, B. Lin, Y. Geng, Z. Wang, X. Chen, L. Yang, T. Lin, and F. Xiao. Airport terminal passenger forecast under the impact of COVID-19 outbreaks: A case study from China. *Journal of Building Engineering*, 65:105740, 2023. doi: 10.1016/j.jobe.2022.105740.

[23] U.S. Department of Transportation. Freedom of information act (foia) program. `https://www.transportation.gov/foia`, 2022. Accessed 2026-01-05.

[24] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pages 22419–22430, 2021.

[25] Haixu Wu, Tengge Hu, Yong Liu, Hangcheng Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modelling for general time series analysis. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=ju_Uqw384Oq`.

[26] Zhijian Xu, Hao Wang, and Qiang Xu. Intervention-aware forecasting: Breaking historical limits from a system perspective. *arXiv preprint arXiv:2405.13522*, 2024.

[27] Chengyang Zhang, Yong Zhang, Qitan Shao, Bo Li, Yisheng Lv, Xinglin Piao, and Baocai Yin. Chattraffic: Text-to-traffic generation via diffusion model. *IEEE Transactions on Intelligent Transportation Systems*, 26(2):2656–2668, 2025. doi: 10.1109/tits.2024.3510402.

[28] Linqi Zhou, Michael Poli, Winnie Xu, Stefano Massaroli, and Stefano Ermon. Deep latent state space models for time-series generation. In *International Conference on Machine Learning*, pages 42625–42643. PMLR, 2023.

[29] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, pages 27268–27286, 2022. URL `https://proceedings.mlr.press/v162/zhou22a.html`.

6

Conclusions and Future Work

This thesis addresses explainable multivariate time-series forecasting with an emphasis on airport passenger flow and related high-dimensional forecasting problems. Across the thesis, four complementary contributions are made: (i) post-hoc explainability methods for forecasting models, (ii) saliency post-processing strategies to improve human interpretability and practical usefulness, (iii) intrinsically interpretable forecasting via explicit component decomposition, and (iv) forecasting with variable-horizon patch-level dynamics, augmented with concept-based interpretability and scenario-based robustness diagnostics under exogenous-shock trajectories.

Rather than advocating a single explanation technique, evidence in this thesis suggests that explanation quality is context-dependent within the evaluated settings: it varies with temporal pattern characteristics, model architecture, explainer choice, and the operational question being asked (diagnosis, justification, or scenario sensitivity).

6.1 Comparison of Explainability Paradigms

Throughout this thesis, the term explainability is used in a technical sense and takes different forms depending on the chapter; the following comparison summarises the kind of explanation offered by each contribution.

The thesis considered three complementary paradigms for explaining forecasts:

- **Post-hoc feature attribution (saliency).** Gradient-based and model-agnostic meth-

ods can highlight input features (and time steps) tied to forecasts. Chapter 2 found architecture- and regime-dependent behaviour without a uniformly best explainer. Deletion AUC yielded a comparative explainer ranking that showed empirical association with predictive error under the fixed masking conditions used in that chapter, but is sensitive to the masking strategy. Common baselines (e.g., zero replacement) can introduce distribution shift and artefacts, so the metric reflects responses to perturbations rather than true feature removal.

- **Post-processed saliency for human readability and reuse.** Chapter 3 demonstrated that raw saliency outputs can be dense and noisy in multivariate forecasting (especially with sliding-window inputs). By enforcing sparsity and normalisation and optionally aligning saliency back to the original time axis, explanations can become more interpretable and can also be leveraged for downstream gains (e.g., saliency-guided augmentation and transferring informative structure to simpler models).

- **Intrinsic interpretability through structured forecasting models.** Chapter 4 introduced iDecPred, which produces explicit trend, seasonal, and residual components as part of the forecasting process. This shifts interpretability from post-hoc attribution to a structured representation that supports transparent diagnostics and component-wise inspection.

In addition, Chapter 5 explored **concept-based interpretability** via semantic anchoring of latent patch dynamics. The key takeaway is that semantic descriptions can offer stable, high-level characterisations of model behaviour, but do not constitute causal or feature-importance explanations without additional validation.

6.2 Key Lessons for Explainable Time-Series Forecasting

Synthesising evidence across chapters reveals three recurring but misaligned objectives: predictive accuracy, attribution faithfulness, and operational utility. Chapter 2 shows weak coupling between test error and oracle explainability in synthetic settings, indicating that accuracy is a poor proxy for explanation quality. This necessitates combining oracle-grounded evaluation with behavioural probes on real data, as plausible saliency does not imply correctness. Under distributional uncertainty, Chapter 5 further augments error-based evaluation with scenario-based sensitivity (e.g., forecast deltas and directional consistency), capturing aspects of robustness not reflected in point metrics. Consequently, the relative emphasis on these objectives should be task-dependent (e.g., debugging, insight, communication, auditing, or scenario planning).

- **No single explainer is universally reliable.** Explanation behaviour depends on temporal structure, model class, and the operational question. Explainers should therefore be selected and validated under conditions that match the intended deployment (data distribution, horizon, and decision context).

- **Interpretability should be matched to the decision abstraction.** When localised attribution is required, sparsified and temporally aligned saliency can improve readability. When stakeholders require structured reasoning about dynamics, intrinsically interpretable formulations (e.g., decomposition into trend/seasonal/residual components) provide a more appropriate level of abstraction.

- **Interpretability and accuracy are not inherently a zero-sum trade-off.** If interpretability-oriented constraints are aligned with the forecasting task (e.g., enforcing readable saliency structure or using explicit component decompositions), they can preserve, and in some settings modestly improve, forecast accuracy; however, such benefits are not guaranteed and should be established empirically for each deployment context.

- **Patch-level autoregressive strategies show improved long-horizon behaviour.** Empirical results indicate that patch-level rollouts produce more coherent predictions across extended horizons than single-step autoregressive schemes, particularly where error accumulation becomes more pronounced.

- **Robustness under uncertainty is not well captured by single-point error metrics.** Evaluation based on fixed test targets can obscure how forecasts respond to plausible variations in future conditions. Scenario-based sensitivity analysis (e.g., directional consistency) provides a more informative way to assess model behaviour under distributional uncertainty.

6.2.1 Implications for Real-World Deployment

For airport passenger flow forecasting, these insights lead to the following practical recommendations for deployment:

- **Tie explanations to operational questions.** Explanations should be elicited and reported in service of specific tasks, such as peak diagnosis, detection of spurious seasonal proxies, and sensitivity of forecasts under plausible disruption scenarios.

- **Prioritise auditability in high-stakes workflows.** Structured outputs (e.g., component decompositions) and scenario-based sensitivity diagnostics are often easier to review, communicate, and operationalise than dense attribution heatmaps.

- **Evaluate robustness with explicit futures.** When future exogenous trajectories are uncertain, robustness should be characterised by comparing forecasts across a set of plausible scenario trajectories rather than assuming a single "correct" future path.

More broadly, these implications support a pragmatic approach for transportation decision-making: pair strong predictive models with explanations and diagnostics that are verifiable, decision-relevant and compatible with the risk and accountability requirements of the deployment setting.

6.3 Limitations

This thesis has several limitations that motivate follow-up research:

- **Scope of explainers and metrics.** The empirical studies prioritise widely used post-hoc explainers and standard evaluation metrics. Consequently, forecasting-specific or time-series-native explainers were not included in the evaluation, and the behavioural metrics used for real-world datasets are generally not specifically designed for forecasting settings [7, 1, 5].
- **Temporal-structure preservation in evaluation.** Perturbation-based evaluations and masking strategies can inadvertently distort temporal dependence, including autocorrelation, seasonality, and cross-variable coupling. This may introduce distribution shifts that confound both faithfulness tests and robustness assessments, motivating evaluation protocols that better preserve plausible temporal structure under interventions, as well as explanation methods that explicitly constrain perturbations to remain locally and temporally plausible [6, 3].
- **Human-centred validation.** The thesis focuses primarily on technical evaluation and diagnostic analyses. Additional user studies with domain experts would strengthen claims about decision utility, usability, and trust calibration in operational forecasting settings [4, 2].
- **Limits of concept-based explanations.** Concept scores provide stable semantic descriptions but do not constitute causal attributions or reliable importance measures. Additional validation is required to establish causal interpretability.

6.4 Future Research Directions

Building on the findings of this thesis, several directions emerge that extend current explainability approaches in time-series forecasting while remaining grounded in the empirical observations of Chapters 2–5.

- **Forecasting-aware explanation methods grounded in temporal structure.** Chapters 2–3 show that existing explainers and post-hoc pipelines do not explicitly account for temporal dependence, feature redundancy, or sliding-window representations. Future work should develop methods that operate on forecasting-specific structures (e.g., rollout trajectories and patch-level dynamics) and explicitly model temporal dependencies, rather than treating inputs as independent features.

- **Faithfulness metrics robust to perturbation artefacts and temporal distortion.** Results in Chapter 2 indicate that deletion-based metrics are sensitive to masking strategies that may not represent a true absence of information. Common baselines (e.g., zero replacement) can induce distribution shift and artefacts, motivating evaluation protocols that preserve plausible temporal structure under intervention and better reflect failure modes in multi-step forecasting.

- **Bridging post-hoc and intrinsic interpretability.** Chapters 3 and 4 highlight complementary strengths between post-hoc saliency and intrinsically interpretable models. Hybrid approaches that combine structured decomposition with feature-level attribution could support both faithful explanations and interpretable component-level reasoning within integrated or hybrid frameworks.

- **From semantic descriptions to validated concept-level explanations.** Chapter 5 shows that semantic anchoring provides stable, high-level characterisations of latent dynamics, but remains descriptive rather than causal. Future work should develop validation protocols and intervention-based methods to connect concept-level representations with causal or counterfactual interpretations.

- **Robustness evaluation under structured scenario models.** Scenario-based sensitivity analysis offers a more informative view of robustness under uncertain exogenous inputs (Chapter 5). Extending this approach to richer scenario generators (e.g., policy interventions, demand shocks, or capacity constraints), combined with

uncertainty quantification, can enable more realistic and actionable robustness assessments.

- **Human-centred evaluation and decision integration.** While this thesis focuses on technical evaluation, the ultimate value of explanations lies in their role in decision-making. Future work should investigate how different explanation forms (e.g., saliency, decomposition, and scenario-based diagnostics) influence human understanding, trust calibration, and operational decisions in real-world settings.

References

[1] Yueshan Chen and Sihai Zhang. Wae: An evaluation metric for attribution-based xai on time series forecasting. *Neurocomputing*, 622:129379, 2025. doi: 10.1016/j.neucom.2025.129379.

[2] Brigt Håvardstun, Cèsar Ferri, Kristian Flikka, and Jan Arne Telle. XAI for time series classification: Evaluating the benefits of model inspection for end-users. In *World Conference on Explainable Artificial Intelligence*, pages 439–453. Springer, 2024. doi: 10.1007/978-3-031-63800-8_22.

[3] Zichuan Liu, Yingying Zhang, Tianchun Wang, Zefan Wang, Dongsheng Luo, Mengnan Du, Min Wu, Yi Wang, Chunlin Chen, Lunting Fan, et al. Explaining time series via contrastive and locally sparse perturbations. *arXiv preprint arXiv:2401.08552*, 2024.

[4] Hongnan Ma, Kevin McAreavey, and Weiru Liu. Tsfeatlime: An online user study in enhancing explainability in univariate time series forecasting. In *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 578–585. IEEE, 2024. doi: 10.1109/ictai62512.2024.00087.

[5] Udo Schlegel and Daniel A. Keim. A deep dive into perturbations as evaluation technique for time series XAI. In *Explainable Artificial Intelligence*, Communications in Computer and Information Science, pages 165–180. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-44070-0_9.

[6] Hugues Turbé, Mina Bjelogrlic, Christian Lovis, and Gianmarco Mengaldo. Evaluation of post-hoc interpretability methods in time-series classification. *Nature Machine Intelligence*, 5(3):250–260, 2023. doi: 10.1038/s42256-023-00620-w.

[7] Xu Zheng, Farhad Shirani, Zhuomin Chen, Chaohao Lin, Wei Cheng, Wenbo Guo, and Dongsheng Luo. F-fidelity: A robust framework for faithfulness evaluation of explainable ai. *arXiv preprint arXiv:2410.02970*, 2024.

A

Airport Passenger Dataset Construction and Analysis

PAPER V – Airport Passenger Dataset Construction and Analysis

International Exposure and Heterogeneous Recovery of U.S. Airports after COVID-19

Abstract

The COVID-19 pandemic generated a large and sudden information shock that coincided with sharp declines in airport passenger flow, often preceding formal policy interventions. This study examines heterogeneous passenger flow responses and recovery dynamics using daily passenger throughput data for the 20 busiest U.S. airports and GDELT-based indices capturing health-related and control-related aviation news. Anticipatory passenger flow declines are documented, with passenger volumes beginning to fall in 2020 early March, prior to major policy escalations in mid-March. An exposure-interaction framework with airport and date fixed effects shows that health-related information is associated with larger passenger flow reductions at airports with higher international exposure, while control-related information is associated with relatively higher passenger flow in later periods. Recovery analysis reveals persistent heterogeneity: airports with higher international passenger shares display lower recovery levels across early, mid, and late stages, with the strongest correlation observed during the mid-recovery period (90-120 days). Early recovery speed is negatively associated with international exposure, contributing to recovery gaps that persist through day 180. Airport size shows positive but imprecise associations with recovery, whereas international exposure remains a robust correlate across multiple robustness checks. Overall, the findings indicate that recovery patterns are consistent with a trade-off between exposure and capacity, and suggest that uniform policy responses may be less well aligned when passenger flow adjustments precede formal interventions.

A.1 Introduction

The COVID-19 pandemic triggered the most severe disruption of global air travel since the Second World War, leading to an abrupt collapse in passenger flow and exposing

structural vulnerabilities in air transportation systems [1, 9]. Existing research has documented widespread connectivity losses and highly uneven recovery patterns across airline networks and airports [16]. While this literature has substantially advanced understanding of infrastructure performance and operational resilience [17, 22], less is known about how information environments, shaped by public discourse on health risks and control measures, influence heterogeneous passenger flow responses and recovery trajectories across major airports.

A growing body of work has examined aviation passenger flow reductions during the COVID-19 pandemic and the role of policy interventions, epidemiological indicators, and mobility restrictions. However, two limitations remain. First, evidence from mobility and consumption behaviour suggests that individuals often respond to perceived risk and information before formal policies are implemented [4]. If airport passenger flow reacts anticipatorily to information shocks, policy responses based solely on observed passenger flow declines may be systematically mistimed. Second, despite facing broadly similar national conditions, airports experienced markedly heterogeneous passenger flow shocks and recovery paths during the pandemic. The mechanisms underlying this heterogeneity remain insufficiently understood.

This study argues that these divergent outcomes reflect the interaction between national information shocks and airport structural characteristics. The COVID-19 pandemic is conceptualised as a large-scale information shock transmitted through mass media coverage and public discourse, altering perceived health risk and uncertainty associated with air travel [18]. While all airports are exposed to a common national information environment, their passenger flow responses may differ systematically depending on structural exposure and recovery capacity. In particular, airports with greater international passenger shares may be more sensitive to global health-related information and cross-border uncertainty, whereas airport size may proxy for recovery capacity through network redundancy and operational flexibility [5].

To examine these mechanisms, daily passenger throughput data for the 20 largest U.S.

airports from 2019 to 2020 are combined with national aviation-related information indices constructed from the Global Database of Events, Language, and Tone (GDELT) [7]. Health-related information is distinguished, capturing perceived epidemiological risk, from control-related information, reflecting coverage of travel restrictions and operational controls [18]. Because these indices are constructed at the national level, they characterise a common information environment while allowing identification of heterogeneous airport responses driven by differences in exposure and capacity, consistent with exposure interaction designs used in related empirical work [8].

The empirical analysis proceeds in three steps. First, using an event study framework, the temporal relationship between information shocks, policy interventions, and airport passenger flow is documented, showing that passenger declines precede major policy escalations. Second, an exposure interaction model with airport and date fixed effects is estimated to identify heterogeneous passenger flow responses to information shocks under a common national environment. Third, recovery dynamics are analysed using multiple complementary metrics to assess how structural characteristics shape both the depth of the initial shock and subsequent recovery trajectories.

The results reveal four main findings. First, passenger volumes decline sharply in advance of major policy interventions, consistent with information-driven anticipatory behaviour. Second, airports with higher international exposure experience significantly larger passenger flow reductions in response to health-related information and persistently slower recovery across multiple stages. Third, control-related information is positively associated with passenger flow during later recovery phases, particularly for internationally exposed hubs. Fourth, airport size exhibits weaker and less precisely estimated effects, suggesting that exposure-related vulnerabilities dominate early recovery dynamics.

Taken together, these findings highlight the importance of information-driven behavioral responses prior to policy implementation and their implications for recovery dynamics. While existing work has largely focused on the effects of policy interventions, it is documented that demand adjustments begin earlier and are systematically shaped by the

interaction between information shocks and airports' international exposure. Airports more exposed to international traffic respond more strongly to health-related information, leading to sharper initial declines. Importantly, these early responses have persistent consequences: they translate into heterogeneous recovery paths, suggesting that initial information-driven shocks play a central role in generating medium-run divergence across locations.

The remainder of the paper is organised as follows. Section A.2 reviews the related literature on aviation passenger flow shocks, information effects, and structural vulnerability. Section A.3 describes the data sources and empirical strategy. Section A.4 presents the empirical results. Section A.5 discusses implications for resilience and policy design, and Section A.6 concludes.

A.2 Related Work

This study builds on and contributes to three related strands of literature: aviation passenger flow shocks and recovery dynamics, information effects and anticipatory behaviour in transportation passenger flow, and international exposure, airport structure, and vulnerability. Together, these literatures provide important context for understanding how large-scale disruptions affect airport passenger flow and recovery [12, 17], while also highlighting gaps that motivate the present analysis.

A.2.1 Aviation Demand Shocks and Recovery Dynamics

A substantial body of research has examined the impacts of major shocks on aviation passenger flow, including terrorist attacks, financial crises, natural disasters, and public health emergencies [12, 11, 3, 24, 16]. Early studies of events such as the September 11 attacks documented sharp passenger flow contractions followed by gradual recovery, emphasising the role of security concerns and policy responses [11, 10]. Subsequent work extended this perspective to economic shocks, showing that macroeconomic downturns and fuel price volatility can generate prolonged effects on airport passenger flow [3, 2].

More recent research has focused on the COVID-19 pandemic, documenting unprecedented declines in passenger volumes across global aviation networks. Studies have linked passenger flow reductions to mobility restrictions, border closures, and epidemiological indicators such as infection rates and fatalities. At the airport and airline levels, empirical analyses have highlighted substantial heterogeneity in the depth of passenger flow decline and the pace of recovery, often attributing these differences to market structure, route composition, and network connectivity [16]. This work provides important insights into structural recovery patterns but places less emphasis on the role of information environments and behavioural passenger flow responses.

Despite these advances, much of the aviation shock literature treats recovery as an aggregate outcome, commonly measured as the time required to return to pre-shock passenger flow levels [17]. Such measures abstract from the dynamic and stage-dependent nature of recovery processes. Emerging work in transportation resilience emphasises that recovery is path dependent, with early recovery dynamics playing a disproportionate role in shaping longer-term outcomes [14, 19]. However, empirical applications of this perspective to aviation, particularly using high-frequency data, remain limited.

A.2.2 Information Effects and Anticipatory Behaviour in Transportation

A parallel literature in transportation economics and behavioural research highlights the role of information in shaping travel decisions [13, 23]. Travellers respond not only to prices and constraints but also to perceptions of risk, uncertainty, and service reliability, which are often mediated by information flows. Evidence from various transportation contexts shows that information availability and framing can significantly influence travel behaviour, especially during disruptive events.

Several studies emphasise anticipatory behaviour, whereby individuals adjust travel decisions in advance of realised constraints based on expectations formed through media coverage and public discourse [6]. During the COVID-19 pandemic, health-related news

coverage and risk communication intensified rapidly, and a growing number of studies have linked media attention to changes in mobility patterns and consumer behaviour [18]. These findings suggest that information can influence behaviour independently of, and sometimes prior to, formal policy interventions.

However, much of this literature is conducted at aggregate spatial scales and focuses on general mobility rather than aviation-specific outcomes. Moreover, information effects are often inferred indirectly and rarely measured using systematic, time-varying indicators. As a result, it remains difficult to isolate the contribution of information environments to airport passenger flow separately from policy measures and epidemiological conditions.

A.2.3 International Exposure, Airport Structure, and Vulnerability

A third strand of literature examines how structural characteristics shape vulnerability and resilience in transportation systems. In aviation, international connectivity has long been recognised as a source of both opportunity and risk. Airports with extensive international networks benefit from higher passenger flow potential but may also be more exposed to global shocks and cross-border uncertainty.

Empirical studies during the COVID-19 pandemic show that internationally connected airports and airlines were disproportionately affected by border closures and international travel restrictions. At the same time, international hubs are often large airports with diversified networks and greater operational resources, which may support recovery once constraints are lifted [17, 5]. The net effect of international exposure and airport size on recovery therefore remains theoretically ambiguous.

Existing studies often examine international exposure and airport size separately, without explicitly modelling their joint effects [15]. This limits understanding of how exposure and capacity interact to shape recovery trajectories. Moreover, few studies consider how these structural characteristics interact with information environments, despite the likelihood that internationally exposed airports are more sensitive to global information flows and risk perceptions.

A.3 Data and Empirical Strategy

This section describes the data sources, construction of key variables, and empirical strategy used to examine heterogeneous passenger flow responses and recovery dynamics under a common information shock. Our approach combines high-frequency airport-level passenger flow data with national measures of aviation-related information exposure, allowing us to study both the timing of passenger flow responses and cross-airport heterogeneity in impact and recovery.

Figure A.1 summarises the construction of key variables and the empirical design. Aviation-related information indices are constructed from national GDELT news coverage and are common across airports [7]. Airport passenger throughput and structural characteristics are measured at the airport level [21]. Identification of heterogeneous passenger flow responses relies on interactions between national information indices and pre-pandemic airport characteristics under airport and date fixed effects.

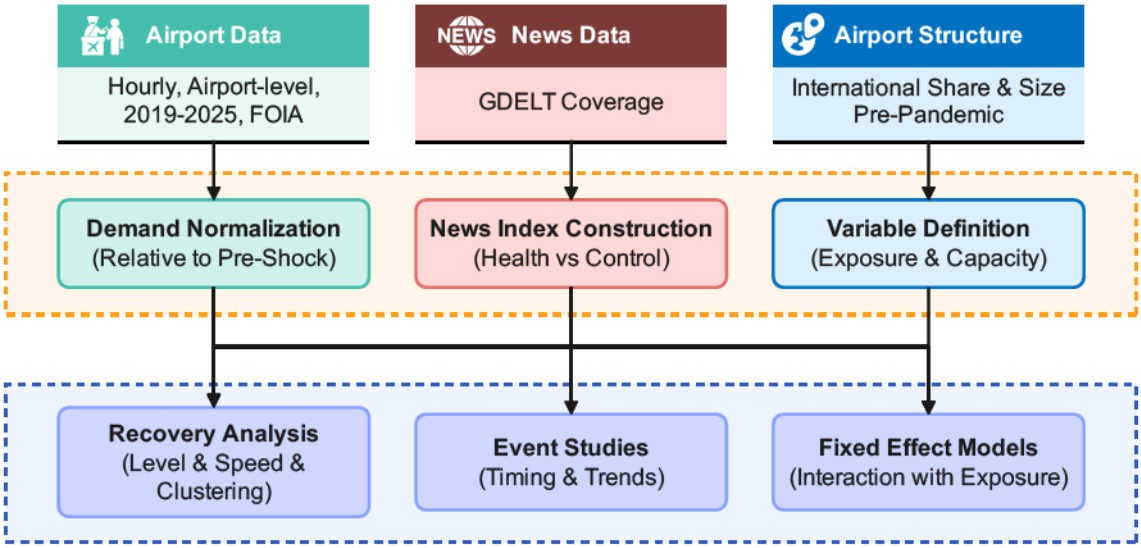


Figure A.1: Data construction and empirical framework. This figure illustrates the data sources (airport throughput, GDELT news articles, policy timeline), data processing steps (normalisation, filtering), construction of information indices (health and control), airport structural attributes (international exposure and size), and the empirical framework (event study, exposure-interaction fixed effects models, and recovery analysis).

A.3.1 Dataset Construction

A.3.1.1 Airport Passenger Throughput

Daily passenger throughput data are used for the 20 largest commercial airports in the United States [20]. The sample period spans January 1, 2019 to December 31, 2020, covering a full pre-pandemic baseline year and the initial pandemic shock and recovery period. Focusing on the largest airports ensures consistent data availability and captures the majority of U.S. passenger traffic, while also limiting heterogeneity driven by very small or seasonal facilities.

Figure A.2 shows the geographic distribution of the study airports across the United States, with airports coloured by their international passenger share in 2019 and sized by their recovery capacity (measured as average normalised passenger volume over days 30–180 post-shock). The map illustrates the spatial coverage of our sample, which includes major hubs across all regions of the country. This geographic diversity helps ensure that our findings are not driven by region-specific factors. The visualisation also suggests a negative association between international exposure and recovery capacity, consistent with the main results.

The primary outcome variable is the daily passenger volume at airport i on day t. To facilitate comparisons across airports of different sizes and to focus on relative changes, passenger volumes are normalised using a pre-shock baseline mean:

$$\bar{V}_i^{\text{pre}} = \frac{1}{|\mathcal{T}_{\text{pre},i}|} \sum_{k \in \mathcal{T}_{\text{pre},i}} V_{i,k}, \quad \tilde{V}_{it} = \frac{V_{it}}{\bar{V}_i^{\text{pre}}}. \tag{A.1}$$

Here k indexes relative event time, and $\mathcal{T}_{\text{pre},i}$ denotes the set of pre-shock event days $k \in [-60, -30]$ for airport i.

The normalised throughput $\tilde{V}_{it}$ is used for descriptive comparisons and recovery metrics (e.g., recovery levels, recovery speed, and dynamic correlations across recovery stages).

For the fixed-effects regressions and shock–demand timing analyses (including lead–lag correlations), the dependent variable is $\log(1 + V_{it})$, where V_{it} denotes raw daily passenger

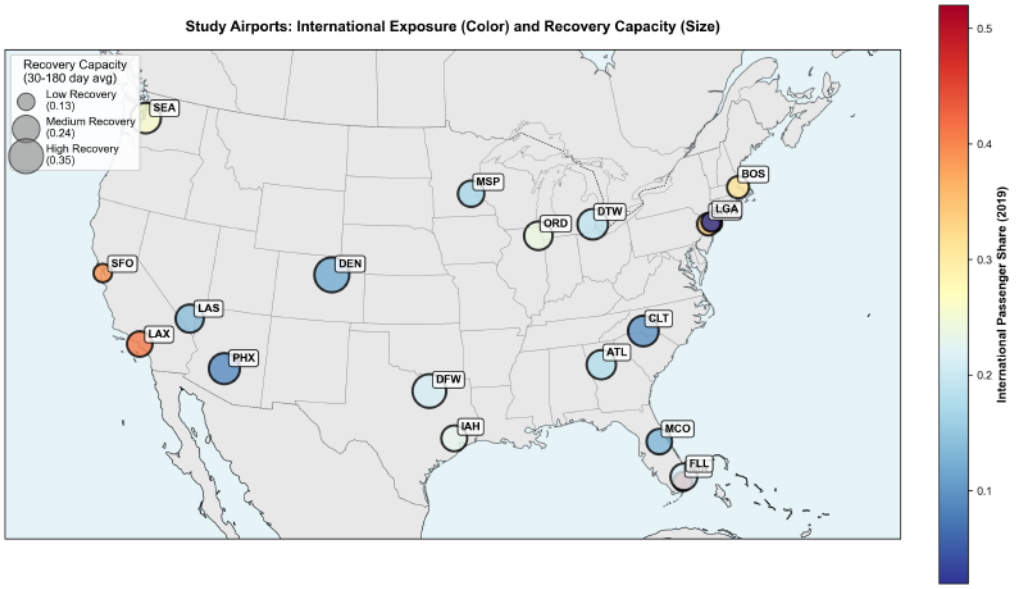


Figure A.2: Geographic distribution of study airports: international exposure and recovery capacity. Colour indicates international passenger share in 2019 (redder = higher exposure), and marker size indicates recovery capacity (measured as average normalised passenger volume over days 30–180 post-shock).

throughput.

A.3.1.2 Information Shock Measures from GDELT

To characterise the information environment faced by travellers, daily aviation-related information indices are constructed using the Global Database of Events, Language, and Tone (GDELT) [7].

Two types of information are distinguished:

- Health-related information, capturing coverage of epidemiological risk.
- Control-related information, capturing coverage of travel restrictions and operational controls.

Health-related articles are identified using keyword combinations that include pandemic-related terms (e.g., “COVID-19”, “coronavirus”, “pandemic”, “outbreak”) together with aviation-related terms (e.g., “airport”, “airline”, “flight”). Control-related articles are identified using keywords related to restrictions and interventions (e.g., “travel ban”, “lockdown”, “quarantine”, “flight suspension”), again restricted to aviation contexts.

For each day t, the share of health-related (or control-related) aviation articles is computed relative to the total number of aviation-related articles, and each index is standardised to have mean zero and unit variance. The resulting standardised series are denoted by H_t (health-related) and C_t (control-related). In addition, an aggregate aviation information index, denoted by S_t, is constructed, which is used to summarise overall information pressure and to define the national shock peak.

Figure A.3 illustrates the evolution of these indices. Health-related information rises earlier and more gradually, while control-related information increases more visibly in early March. Although both indices peak around mid-March, their temporal evolution differs in terms of onset and growth dynamics. These indices are constructed at the national level and are common to all airports. Cross-airport differences in outcomes are therefore attributed to differences in exposure and structural characteristics.

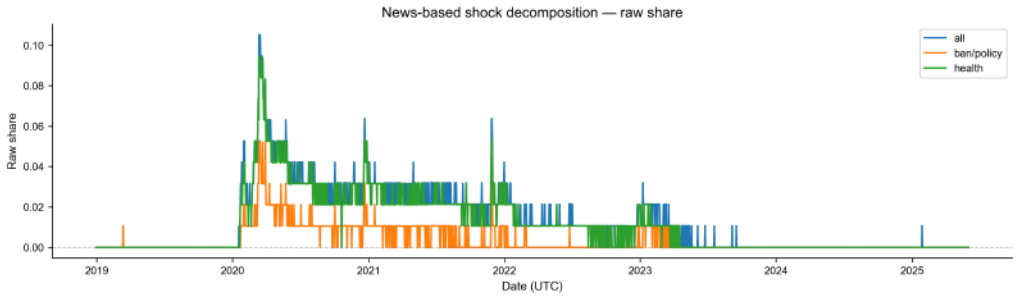


Figure A.3: News-based shock decomposition. Health-related information rises earlier and more gradually, while control-related information increases more visibly in early March.

A.3.2 Empirical Design and Identification Strategy

Our empirical design combines descriptive analysis and regression-based estimation to examine the timing of passenger flow responses, cross-airport heterogeneity, and recovery dynamics. **International exposure** is measured as the share of international passengers in 2019 [21], and **airport size** is measured using total passenger volume in 2019 as a proxy for recovery capacity.

A.3.2.1 Timing and Sensitivity to Information Shocks

Event time is defined relative to the peak date of the aggregate information index S_t, and normalised passenger volumes are used to examine timing patterns.

To characterise airport-level sensitivity to information shocks, the following is estimated:

$$\log(1+V_{it}) = \alpha_i + \beta_i S_t + \varepsilon_{it}, \tag{A.2}$$

where S_t denotes the standardised aggregate aviation information index, and β_i captures airport-specific sensitivity. These estimates are interpreted as associations.

A.3.2.2 Exposure-Interaction Model and Recovery Analysis

To examine heterogeneous responses under a common information environment, the following is estimated:

$$\log(1+V_{it}) = \alpha_i + \delta_t + \phi_H(H_t \times Intl_i) + \phi_C(C_t \times Intl_i) + \varepsilon_{it}. \tag{A.3}$$

Date fixed effects absorb common shocks, so identification comes from cross-airport differences in exposure. This specification is extended by interacting information with airport size.

Finally, additional exploratory analyses are conducted to examine recovery patterns, including clustering airports based on sensitivity and recovery performance and analysing associations between exposure and recovery metrics computed from $\tilde{V}_{it}$. These analyses are descriptive and not used for causal interpretation.

A.4 Results

This section presents the empirical results in four parts. First, the timing of information shocks and passenger flow responses is documented, showing that demand adjustments begin during the buildup of information shocks and prior to major policy interventions.

Then, it is examined whether these responses vary systematically across airports with different levels of international exposure. Next, exposure–information interaction effects are estimated using fixed effects models to identify differential sensitivities under a common national information environment. Finally, it is analysed how these initial differences translate into heterogeneous recovery dynamics across airports, and the roles of international exposure and airport size are assessed across different recovery stages.

A.4.1 Timing of Information Shocks and Demand Responses

The temporal alignment between information shocks, policy interventions, and passenger flow dynamics is first examined. Figure A.4 plots the standardized information indices constructed from GDELT data, including health-related and control-related components, as well as their aggregate measure. The aggregate index rises throughout early March and reaches its peak on March 15, 2020.

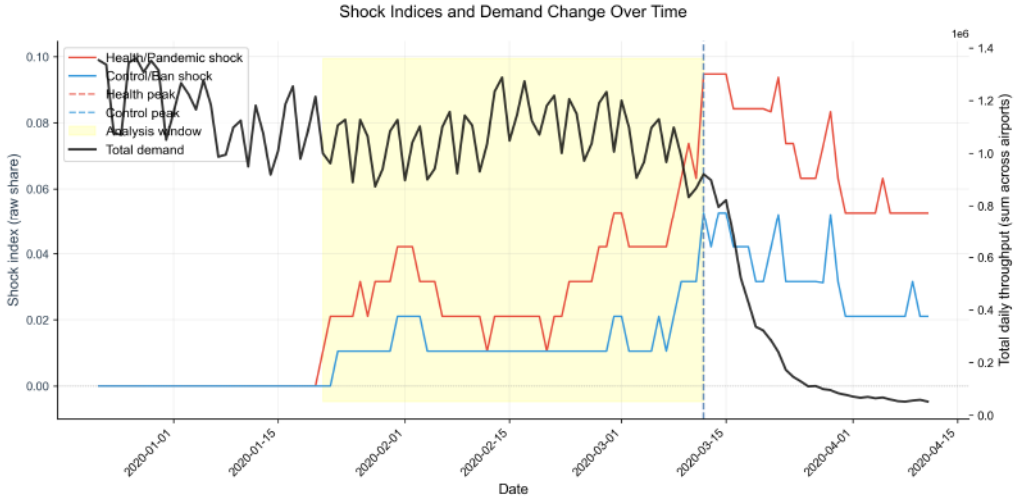


Figure A.4: Standardized health-related and control-related aviation information indices over time. Health-related information rises earlier and exhibits a pronounced increase in early March 2020, while control-related information remains relatively low through February and increases more visibly from early March. Both indices reach their peak around mid-March, though their temporal evolution differs in terms of onset and growth dynamics.

Health-related information exhibits an earlier and more gradual evolution. The index is already elevated relative to its baseline from late January onward, and displays a distinct upward shift around March 5, indicating a sharp increase in perceived epidemiological

risk during this period. In contrast, control-related information remains relatively low through most of February and increases more visibly only from early March, with a clear rise around March 8, closer to the timing of policy announcements.

Passenger flow dynamics show a corresponding early adjustment. National aggregate passenger volumes, measured as the sum across the 20 largest U.S. airports and smoothed using a seven-day moving average, begin to decline in early March. Relative to late-February baseline levels and year-over-year comparisons, passenger volumes fall below key thresholds (e.g., -5%) around March 2 and -10% around March 6-7.

Importantly, these demand adjustments occur prior to major policy escalations, including the U.S. travel restrictions on Europe and the World Health Organization's pandemic declaration on March 11, followed by the U.S. national emergency declaration on March 13. At the same time, passenger volumes begin to weaken during the buildup of information shocks, preceding the peak of the aggregate index on March 15.

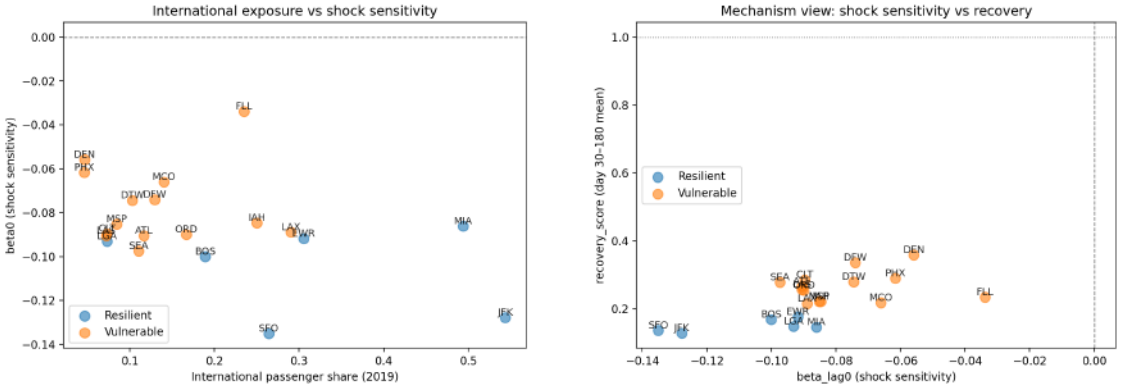
Taken together, these timing patterns suggest that passenger flow responds in advance of both peak information intensity and formal policy interventions. Moreover, the alignment between early demand declines and the rise in health-related information—rather than the later increase in control-related information—is consistent with anticipatory behavioural responses to perceived health risks and uncertainty.

A.4.2 Differential Sensitivity to Information by International Exposure

It is next examined whether the early demand adjustments documented in Section A.4.1 vary systematically across airports with different levels of international exposure. Under a common national information environment, airports may respond differently to information shocks depending on their structural exposure to global travel risks.

Airport-specific sensitivities to the aggregate information shock are first estimated. Figure A.5 summarises the distribution of these sensitivities and relates them to airport characteristics. The results reveal substantial heterogeneity across airports, with estimated

coefficients ranging from strongly negative to near zero. Notably, airports with higher international passenger shares exhibit systematically more negative sensitivity coefficients, indicating a stronger decline in passenger flow in response to increases in information intensity. The correlation between international exposure and sensitivity is negative and statistically significant, suggesting that internationally exposed airports are more responsive to information shocks.



(a) Sensitivity vs. international passenger share. **(b)** Sensitivity vs. recovery performance.

Figure A.5: Airport-level sensitivity to information shocks. Panel (a) relates sensitivity estimates to international passenger share, and panel (b) relates sensitivity to recovery performance. Across airports, sensitivity coefficients exhibit substantial heterogeneity, ranging from strongly negative to near zero.

To more formally assess this relationship, fixed effects models are estimated that exploit cross-airport variation in exposure under a common information environment. Table A.1 reports estimates from specifications that interact information indices with international passenger share while controlling for airport and date fixed effects. The results confirm that health-related information interacts strongly with international exposure: airports with higher international shares experience significantly larger declines in passenger volumes in response to increases in health-related information. By contrast, control-related information exhibits a positive interaction with international exposure. This pattern is consistent with recovery-window evidence, where control-related information becomes more prominent during later stages of the period.

These patterns highlight that different types of information play distinct roles in shaping passenger flow responses. Health-related information, which captures perceived epidemi-

Table A.1: Exposure–information interaction fixed effects estimates

	Health × Intl	Control × Intl
Coefficient	−0.464	+0.279
	(0.014)	(0.012)
p-value	< 0.001	< 0.001
R^2	0.560	
Observations	37,520	

ological risk, is associated with sharper demand contractions, especially at internationally exposed airports. In contrast, control-related information appears to be associated with stabilisation or partial recovery, potentially reflecting a role in reducing uncertainty during later stages of the period.

Taken together, these findings provide evidence that passenger flow responses to information shocks are not uniform, but are systematically shaped by airport-level exposure to international travel. This heterogeneity is central to understanding how a common information environment can generate divergent demand responses across locations, and motivates the subsequent analysis of how these differences translate into heterogeneous recovery dynamics.

A.4.3 Recovery Dynamics: Persistent but Not Permanent Heterogeneity

It is next examined how the heterogeneous responses to information shocks documented in Section A.4.2 translate into recovery dynamics across airports. In particular, it is assessed whether early differences in sensitivity to information shocks are associated with persistent divergence in passenger flow during the recovery period.

Figure A.6 shows average recovery trajectories for airports grouped by international exposure. While all airports experience sharp initial declines, recovery paths diverge substantially. Airports with lower international exposure recover more quickly and reach higher fractions of baseline passenger flow, whereas highly exposed airports exhibit

persistently weaker recovery during the early and middle stages.

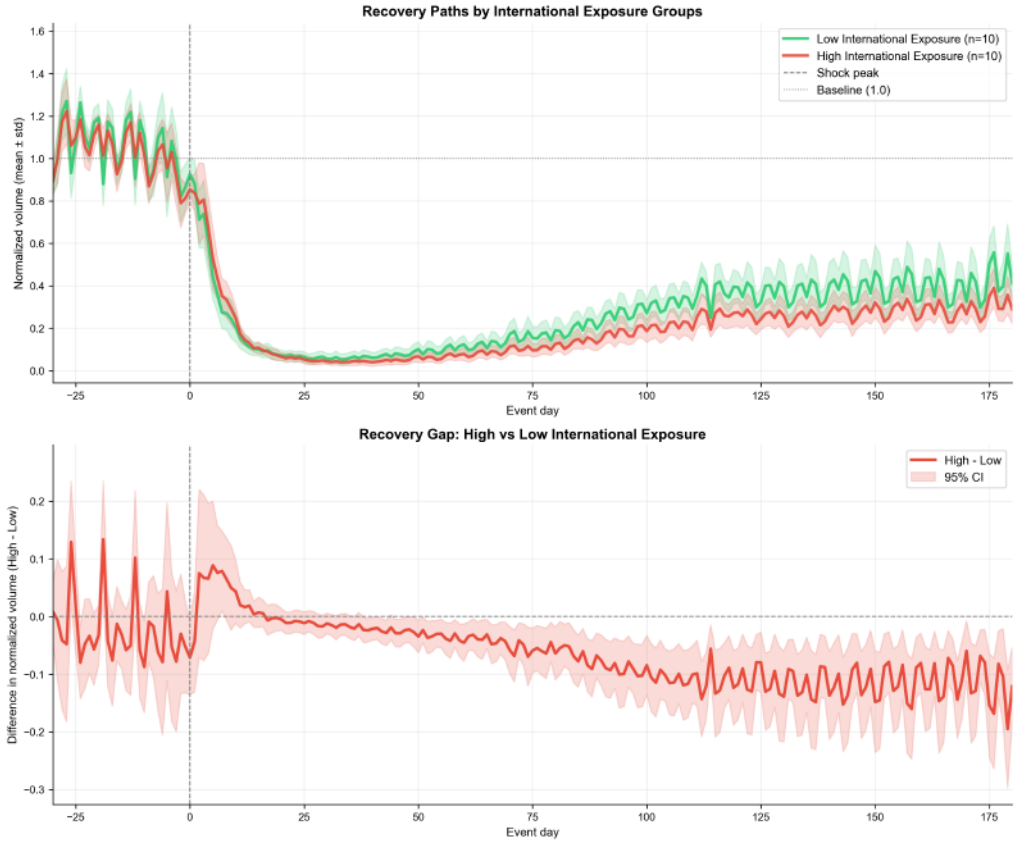


Figure A.6: Recovery trajectories by international exposure group. While both groups experience sharp initial declines, recovery paths diverge substantially, with lower-exposure airports recovering faster and reaching higher fractions of baseline passenger flow. High international exposure airports show persistently lower recovery throughout the recovery period.

This pattern is reflected across multiple recovery metrics. International exposure is negatively associated with recovery levels throughout the post-shock period, with the strongest relationship observed during the mid-recovery stage (approximately 90–120 days, $r = -0.739$, $p = 0.0002$). Similarly, early recovery speed is significantly lower for airports with higher international exposure ($r = -0.701$, $p = 0.0006$), indicating that recovery gaps emerge early and widen during the initial recovery phase.

However, this heterogeneity is not permanent. Extending the analysis to longer horizons shows that the negative association between international exposure and recovery gradually weakens over time. While differences remain pronounced within the first year following the shock, they become smaller and statistically insignificant over longer horizons, as

passenger volumes across airports converge toward baseline levels. By contrast, airport size exhibits positive but imprecisely estimated associations with recovery, suggesting that exposure-related frictions dominate any scale-related buffering capacity within this sample.

Taken together, these findings suggest that early differences in responses to information shocks generate persistent but ultimately diminishing divergence in recovery trajectories. Recovery disadvantages associated with international exposure emerge early, remain pronounced through the first year, and then gradually attenuate over longer horizons. This pattern is consistent with path-dependent recovery dynamics in the short to medium run, combined with convergence as broader economic conditions, policy adjustments, and evolving travel environments reshape demand over time.

A.4.4 Robustness Checks

A series of robustness checks is conducted to assess the stability of the main findings. Across all specifications, the results consistently show a negative interaction between health-related information and international exposure, and a positive interaction between control-related information and international exposure.

First, the role of time fixed effects is examined. Omitting date fixed effects leads to larger coefficient magnitudes, reflecting the presence of common time-varying factors, but does not alter the direction or statistical significance of the estimated interactions. This indicates that the main results are not driven by aggregate time trends.

Second, robustness to alternative definitions and scaling of international exposure is tested. Standardising the exposure variable, discretising it into terciles, or excluding airports at the extremes of the exposure distribution yields consistent estimates, suggesting that the results are not sensitive to functional form choices or outliers.

Third, the sensitivity of the results to sample period selection is examined. Estimating the model over alternative time windows—including the full pandemic period, peak impact period, and extended windows—produces stable and statistically significant interaction

effects, indicating that the findings are not dependent on a particular sample definition.

Fourth, it is assessed whether the information indices are capturing epidemiological conditions rather than information effects. Controlling for state-level COVID-19 case counts interacted with international exposure leads to only modest attenuation of the estimated coefficients, while the main interaction effects remain statistically significant. This suggests that the information measures capture distinct signals beyond epidemiological dynamics.

Finally, placebo tests are conducted by randomly reassigning international exposure across airports. The resulting placebo coefficients are centred near zero, while the baseline estimates lie far from this distribution, indicating that the observed effects are unlikely to arise by chance.

Taken together, these robustness checks confirm that the main findings are stable across alternative specifications, exposure measures, and sample definitions, and are not driven by outliers, omitted epidemiological factors, or random variation (Table A.2).

A.5 Discussion

This study provides a unified perspective on how information environments shape passenger flow dynamics during large-scale disruptions. Taken together, the results suggest a coherent mechanism linking information shocks, behavioural responses, and recovery heterogeneity. Passenger flow begins to decline as health-related information intensifies, prior to both the peak of the information shock and major policy interventions, indicating anticipatory demand adjustment. Under a common national information environment, airports with greater international exposure exhibit systematically stronger responses to health-related information, leading to sharper initial declines. These early differences translate into persistent divergence in recovery trajectories, with highly exposed airports recovering more slowly in the short to medium run before gradually converging over longer horizons.

Table A.2: Robustness checks: Stability of exposure–information interaction effects across alternative specifications

Specification	Health × Intl	Control × Intl
Baseline		
With Date FE	-0.464^{***}	$+0.279^{***}$
Alternative Specifications		
Without Date FE	-1.941^{***}	$+0.495^{***}$
Standardized Exposure	-0.059^{***}	$+0.038^{***}$
Time Windows		
Jan–Jun 2020	-0.664^{***}	$+0.388^{***}$
Feb–May 2020	-0.567^{***}	$+0.332^{***}$
Extended Window	-0.627^{***}	$+0.382^{***}$
Alternative Exposure Measures		
Terciles	-0.052^{***}	$+0.037^{***}$
Exclude Extremes	-0.370^{***}	$+0.324^{***}$
Information vs. Epidemiology		
With State Case Controls	-0.342^{***}	$+0.171^{***}$
Placebo Tests		
Placebo p-value	< 0.05	< 0.001

Notes: This table reports estimates from exposure–information interaction models under alternative specifications. All models include airport fixed effects. "Standardized Exposure" refers to using a standardized international passenger share. "Terciles" refers to discretizing exposure into three groups. "Exclude Extremes" removes the highest and lowest exposure airports. "State Case Controls" includes state-level COVID-19 case counts interacted with international exposure. Placebo tests are based on random reassignment of exposure across airports. Statistical significance is denoted by $^{***}p < 0.001$.

A.5.1 Information and Anticipatory Demand Behaviour

The observed timing patterns are consistent with anticipatory behavioural responses to information. Rather than reacting solely to binding policy constraints, passengers appear to adjust travel decisions in response to evolving perceptions of risk and uncertainty. The early rise in health-related information, combined with the decline in passenger flow prior to major policy escalations, suggests that information plays an independent role in shaping demand.

This interpretation is consistent with the observed timing patterns, where passenger behaviour adjusts in response to widely reported information before formal policy measures are implemented. In the context of aviation, health-related information may increase perceived risk and uncertainty associated with travel, leading passengers to cancel or postpone trips even in the absence of formal restrictions. Importantly, this mechanism does not require precise information or fully rational expectations; it is consistent with heuristic responses to rapidly changing and highly visible information environments.

A.5.2 International Exposure and Recovery Heterogeneity

A central finding of this study is that international exposure systematically shapes both initial responses and recovery dynamics. Airports with higher international passenger shares exhibit stronger sensitivity to health-related information and experience more pronounced and persistent recovery disadvantages. This pattern suggests that exposure to global travel networks amplifies vulnerability to information shocks and cross-border uncertainty.

The persistence of recovery gaps in the short to medium run indicates that early differences in demand adjustment can have lasting consequences. Even as recovery speeds converge in later stages, airports that experience deeper initial declines remain below baseline levels for longer periods. This pattern is consistent with path-dependent recovery dynamics, in which early disadvantages become embedded in subsequent trajectories.

A.5.3 The Role of Control-Related Information

The positive interaction between control-related information and international exposure indicates that, for more internationally exposed airports, control-related information is associated with relatively higher passenger flow, consistent with later stages of the period, when control-related information becomes more prominent. While health-related information is closely tied to perceived risk and drives early demand reductions, control-related information becomes more relevant during later phases.

One possible interpretation is that information about policies and operational measures may reduce uncertainty for travellers who remain active in the market, particularly in more complex international travel contexts. This interpretation is consistent with the observed interaction effects, but does not imply that restrictions increase demand.

A.5.4 Airport Size and Structural Trade-offs

The role of airport size appears more limited in this setting. While larger airports may possess greater operational capacity and network flexibility, the empirical results show positive but imprecisely estimated associations with recovery outcomes. This suggests that scale alone does not offset the vulnerabilities associated with international exposure.

Taken together, the findings point to a structural trade-off between exposure and capacity. Internationalisation increases sensitivity to information shocks and slows recovery, while size may provide only partial buffering capacity within this sample. This trade-off helps explain why highly international, large airports can still experience persistent recovery disadvantages during global disruptions.

A.5.5 Policy Implications

The findings have several implications for policy design. First, if passenger flow responds to information prior to policy implementation, then reactive interventions may be poorly aligned with actual demand dynamics. Policies introduced after substantial demand ad-

justment has already occurred may have limited immediate impact. Second, the persistent recovery disadvantages associated with international exposure suggest that uniform recovery strategies may be insufficient. Airports with greater exposure to international traffic may benefit from targeted support during early recovery stages, when divergence in recovery trajectories emerges. Third, the distinct roles of health-related and control-related information underscore the importance of clear and consistent communication, particularly during early recovery stages when uncertainty is highest. Providing timely and transparent information about risks and travel conditions may help reduce uncertainty and support more stable recovery patterns. These implications are suggestive rather than prescriptive, as the analysis does not identify causal policy effects.

A.5.6 Limitations and Future Research

This study has several limitations. First, the analysis focuses on a relatively small sample of large U.S. airports, which may limit external validity. Second, the study examines a single, unprecedented shock, and the generality of the findings to other contexts remains an open question. Third, while measures of information exposure are constructed, information environments are inherently complex and may not be fully captured by the indices used here. Future research could extend this framework to other settings, incorporate alternative information sources such as search or social media data, and explore causal identification strategies to better isolate the mechanisms linking information and passenger behaviour.

A.6 Conclusion

This chapter examines how information shocks, international exposure, and airport structural characteristics jointly shape airport passenger flow and recovery during the COVID-19 pandemic. Using daily passenger throughput data for the 20 largest U.S. airports and aviation-specific information indices constructed from GDELT, both the timing of passenger flow responses and the heterogeneity of recovery trajectories under a common national

information environment are analysed.

Several conclusions emerge from the analysis. First, air travel passenger flow declines precede major policy escalations and closely follow increases in health-related aviation information. This timing pattern provides evidence consistent with information-driven anticipatory behaviour, suggesting that travellers adjust decisions based on evolving perceptions of risk and uncertainty rather than solely in response to formal policy constraints.

Second, airports exhibit substantial heterogeneity in their responses to the same information shock. International exposure emerges as a key dimension of this heterogeneity. Airports with higher international passenger shares experience stronger passenger flow reductions in response to health-related information, with the exposure–interaction coefficient estimated at $\phi_H = -0.464$ ($p < 0.001$), even after controlling for aggregate time effects. In contrast, control-related information shows a positive association ($\phi_C = +0.279$, $p < 0.001$) in the exposure–interaction specification, suggesting that different information types play distinct roles. This finding indicates that structural exposure to international travel amplifies sensitivity to global information shocks.

Third, recovery dynamics are highly uneven across airports and persist well beyond the initial shock. International exposure is associated with lower recovery levels across early, mid, and late recovery stages, with correlations of $r = -0.582$ (early, $p = 0.0071$), $r = -0.739$ (mid, $p = 0.0002$), and $r = -0.669$ (late, $p = 0.0013$), with the strongest associations observed during the mid-recovery period (90–120 days). Differences in recovery speed are concentrated in the early recovery phase ($r = -0.701$, $p = 0.0006$), after which recovery trajectories tend to evolve in parallel. This pattern implies that early recovery dynamics play a critical role in shaping longer-term outcomes.

Fourth, airport size exhibits positive but imprecise associations with recovery outcomes, while interactions between size and international exposure are not statistically significant within this sample. These results suggest that scale may provide some recovery capacity, but it does not fully offset the recovery frictions associated with international exposure during the period studied. Together, the findings point to a structural trade-off between

exposure and capacity: internationalisation increases vulnerability to information shocks and slows recovery, while capacity-related advantages are more difficult to identify in the short to medium term.

The study contributes to the literature in several ways. It provides systematic airport-level evidence consistent with information-driven passenger flow responses during a large-scale crisis, complementing policy-centric analyses of aviation passenger flow. It documents persistent heterogeneity in recovery dynamics under a common national shock, emphasising that recovery is not solely determined by the magnitude of the initial disruption. Finally, it frames airport recovery as a stage-dependent process shaped by structural characteristics and information environments, offering a nuanced perspective on resilience in aviation systems.

The findings also have important policy implications. If passenger flow responds to information before policy implementation, then uniform, reactive interventions may be misaligned with observed behavioural dynamics. Highly internationalised airports may benefit from early, targeted support aimed at accelerating initial recovery, rather than relying exclusively on broad measures introduced after passenger flow has already collapsed. More broadly, the results underscore the importance of considering information environments and structural heterogeneity when designing responses to future disruptions in air transportation.

Several limitations should be noted. The analysis focusses on a relatively small sample of large U.S. airports and examines a single, unprecedented shock. While this allows for detailed high-frequency analysis, the findings may not generalise to smaller airports, other countries, or different types of disruptions. In addition, although the information indices capture salient aspects of the information environment, they do not encompass all channels through which information influences behaviour. Finally, the empirical approach documents robust associations and timing patterns but does not identify causal treatment effects of information or policy interventions.

Future research could extend this framework in several directions. Expanding the

analysis to a broader set of airports or to international contexts would enhance external validity. Applying similar methods to other shocks could help assess the generality of information-driven passenger flow responses. Incorporating alternative information sources, such as social media or search data, may further refine measurement of information environments. Finally, combining observational data with quasi-experimental or structural approaches could provide deeper insight into the causal mechanisms linking information, structure, and recovery.

In conclusion, this study shows that information shocks play a central role in shaping airport passenger flow and that recovery outcomes reflect systematic structural heterogeneity rather than random variation. By highlighting the interaction between information environments and airport characteristics, this chapter contributes to a more comprehensive understanding of resilience in air transportation systems and offers relevant insights to manage future large-scale disruptions.

References

[1] Air Transport Bureau. Effects of novel coronavirus (COVID-19) on civil aviation: economic impact analysis. Technical Report 178, International Civil Aviation Organization, Montréal, Canada, 2020. URL `https://www.icao.int/economic-impacts-covid-19-civil-aviation`.

[2] Jungho Baek and Soojoong Nam. Oil shocks and air travel demand: Insights from korea and japan. *Australian Economic Papers*, 64(4):457–472, 2025. doi: 10.1111/1467-8454.70001.

[3] Martijn Brons, Eric Pels, Peter Nijkamp, and Piet Rietveld. Price elasticities of demand for passenger air travel: a meta-analysis. *Journal of Air Transport Management*, 8(3):165–175, 2002. doi: 10.1016/S0969-6997(01)00050-3.

[4] Ho Fai Chan, Jordan W Moon, David A Savage, Ahmed Skali, Benno Torgler, and Stephen Whyte. Can psychological traits explain mobility behavior during the covid-19 pandemic? *Social Psychological and Personality Science*, 12(6):1018–1029, 2021. doi: 10.1177/1948550620952572.

[5] Kevin L Clark, Udit Bhatia, Evan A Kodra, and Auroop R Ganguly. Resilience of the us national airspace system airport network. *IEEE Transactions on Intelligent Transportation Systems*, 19(12):3785–3794, 2018. doi: 10.1109/tits.2017.2784391.

[6] Koen Deforche, Jurgen Vercauteren, Viktor Müller, and Anne-Mieke Vandamme. Behavioral changes before lockdown and decreased retail and recreation mobility during lockdown contributed most to controlling COVID-19 in western countries. *BMC Public Health*, 21:654, 2021. doi: 10.1186/s12889-021-10676-1.

[7] GDELT Project. A global database of society. https://www.gdeltproject.org/, 2022. Accessed: 2026-01-13.

[8] Paul Goldsmith-Pinkham, Isaac Sorkin, and Henry Swift. Bartik instruments: What, when, why, and how. *American Economic Review*, 110(8):2586–2624, 2020. doi: 10.1257/aer.20181047.

[9] International Civil Aviation Organization. Annual report 2020. https://www.icao.int/annual-report-2020, 2020. Accessed: 2026-01-13.

[10] Harumi Ito and Darin Lee. Comparing the impact of the september 11th terrorist attacks on international airline demand. *International Journal of the Economics of Business*, 12(2):225–249, 2005. doi: 10.1080/13571510500127931.

[11] Harumi Ito and Darin Lee. Assessing the impact of the september 11 terrorist attacks on U.S. airline demand. *Journal of Economics and Business*, 57(1):75–95, 2005. doi: 10.1016/j.jeconbus.2004.06.003.

[12] Milan Janić. Reprint of "modelling the resilience, friability and costs of an air transport network affected by a large-scale disruptive event". *Transportation Research Part A: Policy and Practice*, 81:77–92, 2015. doi: 10.1016/j.tra.2015.07.012.

[13] David Levinson. The value of advanced traveler information systems for route choice. *Transportation Research Part C: Emerging Technologies*, 11(1):75–87, 2003. doi: 10.1016/S0968-090X(02)00023-2.

[14] YuanChi Liu, Sue McNeil, Jürgen Hackl, and Bryan T. Adey. Prioritizing transportation network recovery using a resilience measure. *Sustainable and Resilient Infrastructure*, 7(1):70–81, 2020. doi: 10.1080/23789689.2019.1708180.

[15] Oriol Lordan, Jose M Sallan, Pep Simo, and David Gonzalez-Prieto. Robustness of the air transport network. *Transportation Research Part E: Logistics and Transportation Review*, 68:155–163, 2014. doi: 10.1016/j.tre.2014.05.011.

[16] Pere Suau-Sanchez, Augusto Voltes-Dorta, and Natàlia Cugueró-Escofet. An early assessment of the impact of covid-19 on air transport: Just another crisis or the end of aviation as we know it? *Journal of Transport Geography*, 86:102749, 2020. doi: 10.1016/j.jtrangeo.2020.102749.

[17] Xiaoqian Sun, Sebastian Wandelt, and Anming Zhang. Resilience of cities towards airport disruptions at global scale. *Research in Transportation Business & Management*, 34:100452, 2020. doi: 10.1016/j.rtbm.2020.100452.

[18] Takahisa Suzuki, Hitoshi Yamamoto, Yuki Ogawa, and Ryohei Umetani. Effects of media on preventive behaviour during the covid-19 pandemic. *Humanities and Social Sciences Communications*, 10(1):1–8, 2023. doi: 10.1057/s41599-023-01554-9.

[19] Fang Tang, Xiangyong Luo, and Xuesong Zhou. Mobility resilience and recovery dynamics: Parsimonious framework beyond V-shapes. *Transportation Research Part C: Emerging Technologies*, 175:105122, 2025. doi: 10.1016/j.trc.2025.105122.

[20] Transportation Security Administration. Foia electronic reading room. `https://www.tsa.gov/foia/readingroom`, 2022. Accessed: 2026-01-13.

[21] United States Department of Transportation Bureau of Transportation Statistics. 2019 traffic data for u.s. airlines and foreign airlines u.s. flights - final, full-year. `https://www.bts.gov/newsroom/final-full-year-2019-traffic-data-us-airlines-and-foreign-airlines`, 2019. Accessed: 2026-01-13.

[22] Sebastian Wandelt, Anming Zhang, and Xiaoqian Sun. Global airport resilience index: Towards a comprehensive understanding of air transportation resilience. *Transportation Research Part D: Transport and Environment*, 138:104522, 2025. doi: 10.1016/j.trd.2024.104522.

[23] Lei Zhang and David Levinson. Determinants of route choice and value of traveler in-

formation. *Transportation Research Record: Journal of the Transportation Research Board*, 2086(1):81–92, 2008. doi: 10.3141/2086-10.

[24] Lei Zhou and Zhenhua Chen. Measuring the performance of airport resilience to severe weather events. *Transportation Research Part D: Transport and Environment*, 83:102362, 2020. doi: 10.1016/j.trd.2020.102362.

B

Explainable Interface Survey

Explainable Interface for Human-Autonomy Teaming: A Survey

Abstract

Nowadays, large-scale foundation models are being increasingly integrated into numerous safety-critical applications, including Human-Autonomy Teaming (HAT) within transportation, medical, and defence domains. Consequently, the inherent 'black-box' nature of these sophisticated deep neural networks heightens the significance of fostering mutual understanding and trust between humans and autonomous systems. To tackle the transparency challenges in HAT, this chapter conducts a thoughtful study on the underexplored domain of Explainable Interface (EI) in HAT systems from a human-centric perspective, thereby enriching the existing body of research in Explainable Artificial Intelligence (XAI). The design, development, and evaluation of EI within XAI-enhanced HAT systems are explored. To do so, the distinctions between these concepts are first clarified: EI, explanations and model explainability, aiming to provide researchers and practitioners with a structured understanding. Second, a novel framework for EI is contributed, addressing the unique challenges in HAT. Last, the summarised evaluation framework for ongoing EI offers a holistic perspective, encompassing model performance, human-centered factors, and group task objectives. Based on extensive surveys across XAI, HAT, psychology, and Human-Computer Interaction (HCI), this review offers multiple novel insights into incorporating XAI into HAT systems and outlines future directions.

B.1 Introduction

B.1.1 Background

Over the past decade, advancements in machine learning, computing, and communication, have significantly increased the machine autonomy [285]. This evolution from automated

to autonomous systems signifies their transition from mere tools to teammates in Human-Autonomy Teaming (HAT). However, the increase in machine autonomy brings challenges in ethics and value alignment, authority transfer explanation, and supervision complexities [142, 163].

Essential to addressing these challenges is the need for mutual explanation, understanding, and trust, particularly regarding the machine's capabilities in perception, processing, decision-making, and planning in uncertain situations [128, 19, 181]. Klein et al. [128] emphasize that effective HAT requires four key commitments: team agreement, predictable mutual actions, directability, and maintaining common ground. They also identify ten crucial challenges for creating intelligent and efficient HAT, ranging from adherence to agreements and predictability to shared goal setting, collaborative planning, and optimizing energy in joint activities. Joseph et al. [153] identify three key challenges in HAT: creating team affordances for shared awareness and motivation, understanding beneficial tasks and interactions for social cueing, and developing techniques to use these cues to improve HAT performance.

Mutual understanding within HAT systems relies on transparency, accountability, and predictability. Accountability in HAT, as defined [66] is "process of account to some agent for a specific task". While related to responsibility and liability, Mulgan [172] differentiates them by highlighting that accountability demands justifications for actions and stands indivisible, whereas liability pertains to legal or financial obligations.

Mutual trust in HAT systems is closely tied to clarity in accountability. Trust grows when users witness autonomous systems performing tasks with reliability and transparency, particularly in elucidating decisions in uncertain scenarios. Minimising uncertainty and improving transparency in learning are crucial for trust development. This must be coupled with an understanding of human capabilities to optimise task collaboration [275, 229].

To build efficient HAT for time-critical and safe-critical tasks leveraging AI, explainability is central for fostering mutual understanding and trust [10, 155]. Without two-way explanations between humans and autonomy, the latter becomes a "black box", severing

a vital communication link and threatening the collective objective [296]. Humans need to clearly convey their motivations and intentions to machines for goal alignment, while machines should provide clear explanations of their actions.

In conclusion, Explainable Interface (EI) is crucial in HAT for effective collaboration. The core benefits of integrating EI with Explainable Artificial Intelligence (XAI) in HAT encompass:

- **Enhancing understanding:** EI clarifies how autonomy decisions are made, making complex AI processes transparent and accessible.
- **Enhancing effective communication and trust:** EI facilitates clear, two-way communication between humans and machines, essential for building or restoring trust.
- **Facilitating seamless coordination:** By fostering interaction among HAT agents, EI strengthens the overall system, enabling smoother and more intuitive teamwork.

B.1.2 Contribution

Table B.1 presents an examination of 30 XAI surveys. Their coverage of HAT, model explainability, Large Language Model (LLM) for explanation generation, human perspectives, system design paradigm and evaluation framework was reviewed. Based on the summarised results, the frequently reported lack of structured frameworks and reviews for Explainable Interface enhanced Human-Autonomy Teaming (EI-HAT) systems is addressed, and the call for deeper integration of human-centric considerations in XAI system is answered.

These previous reviews collectively serve as a robust underpinning for the literature synthesis presented in this chapter. In contrast to previous surveys, our study takes a step further by considering the design, development, and evaluation of EI, exploring its critical role in HAT and discussing future trends and challenges in creating more intuitive and user-friendly explainable systems. Key contributions of this study include:

- **Clarifying 'explainable interface':** Differentiating between 'explainability', 'explanation', and 'explainable interface', clarifying roles of AI researchers, developers, and designers.
- **Exploring conceptual and design framework for EI in HAT:** Introducing a framework for developing EI for HAT. Distinguish key points of task-oriented and experience-oriented ones.
- **Summarising model explainability enhancement:** Organising model explainability enhancement methods in practical way, specially summarising the ongoing research on LLMs for explanation generation.
- **Proposing evaluation framework for EI-HAT:** Advocating for a comprehensive assessment of EI in HAT by embracing model performance, human-centred, and group performance evaluation methods.
- **Highlighting challenges and directions:** Enumerating ongoing challenges in the integration of EI for HAT.
- **Providing application perspective insights:** Analysing the evolution and adaptability of EI across healthcare, transportation and digital twin.

B.1.3 Project Organisation

As outlines in Figure B.1, the design approach for EI-HAT system begins with stakeholder requirements analysis. This is followed by defining the system's objectives, which can be task-oriented, experience-oriented, or both. Next is the explanation generation, laying the groundwork for a design approach that encompasses interactivity, adaptivity, personality, and multi-modality. The final phase is system evaluation, which can be model-based, human-centred, or teaming task-oriented. This completes a full cycle of key elements essential for the functionality of the EI-HAT system.

Table B.1: Overview of Relevant Surveys on XAI and HAT, with Main Topic for Contribution

Ref	Year	Topic	HAT	Model	LLMs	Human	Design	Evaluation
Ours	2024	Explainable Interface, HAT	■	■	■	■	■	■
[1]	2018	XAI Systems, HCI	-	■	-	■	■	-
[60]	2018	User Interface, Interactive ML	■	-	-	■	■	■
[212]	2019	Explainability, Human-Agent Systems	■	■	-	■	■	■
[84]	2019	XAI	-	■	-	-	-	-
[22]	2020	XAI, Responsible AI	-	■	-	■	-	■
[168]	2021	XAI Systems	-	■	-	■	■	■
[256]	2021	XAI	-	■	-	-	-	■
[272]	2021	Explainable Reinforcement Learning	■	■	-	■	■	■
[46]	2022	Counterfactual and Causality	-	■	-	■	■	■
[117]	2022	Contrastive Explanations	-	■	-	-	-	■
[162]	2022	XAI, Medical Images	-	■	-	-	-	■
[259]	2022	Explainable Reinforcement Learning	-	■	-	-	-	■
[305]	2022	NLP, Explainability	-	■	-	-	-	■
[181]	2022	XAI, Autonomous Driving	■	■	-	■	■	■
[119]	2022	Trustworthy AI	-	■	-	■	-	■
[61]	2023	XAI	-	■	-	-	■	■
[89]	2023	Interpreting Adversarial Examples	-	■	-	-	-	-
[106]	2023	Explainable CNNs	-	■	-	-	-	■
[142]	2023	Trustworthy AI	-	■	-	■	■	■
[165]	2023	Explainable Reinforcement Learning	-	■	-	-	-	■
[175]	2023	XAI Evaluation	-	■	-	-	-	■
[193]	2023	XAI, Medical Images	-	■	-	-	-	■
[200]	2023	Graph Counterfactual Explanations	-	■	-	-	-	■
[215]	2023	Explainable Goal-Driven Agents	■	■	-	■	■	■
[7]	2023	XAI, Healthcare	-	■	-	■	■	■
[8]	2023	XAI, Concepts, Challenges	-	■	-	■	■	■
[268]	2023	XAI Model Improvement, Challenges	-	■	-	-	-	■
[72]	2024	Explanation-Guided Learning	-	■	-	■	-	■
[163]	2024	Trustworthy AI, Variable Autonomy	■	■	-	■	■	-
[277]	2024	XAI in LLM Era	-	■	■	■	■	■

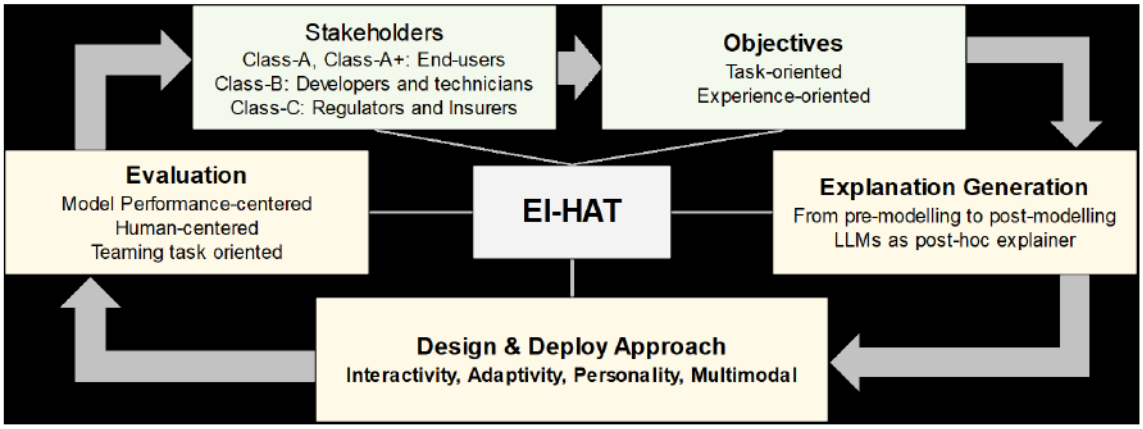


Figure B.1: Design Approach for Explainable Interface enhanced Human-Autonomy Teaming.

This chapter is structured as follows. In Section B.2, the high-level architecture of EI-HAT is presented, laying the foundational understanding of what the EI is, why it is needed, and how it functions. Building on this, Section B.3 delves into the specific frameworks and techniques for model explainability across various stages, from pre-modelling to post-modelling stage, specially LLMs for explanation generation is discussed. Section B.4 explores the human-centred EI-HAT. To assess the effectiveness of EIs, Section B.5 provides a comprehensive summary of evaluation framework. In Section B.6, challenges and directions are discussed. The study is concluded in Section 4.7.

B.2 High-Level Architectures of Explainable Interface in HAT

B.2.1 Clarification of Explainable Interface

In analysis of the surveys presented in Table B.1, frequent discussions around various XAI-related concepts were observed, but there is still lack of consistency in terminology usage.

BJ Fogg's Prominence-Interpretation Theory, proposed in 2003, can be viewed as an early exploration of the concept of an "Explainable Interface" [69]. It consists of two main components: Prominence and Interpretation. Prominence pertains to the noticeable and distinct elements on a website that capture user attention. Interpretation involves how these elements are judged by users. For example, a colorful, animated banner ad is prominent and may lead users to judge the site's credibility . The term "Explainable Interface" first emerged in the ACM archives in 2010 [140] in research on the role of trust in self-adaptive ubiquitous computing systems, focusing on the pervasive display environment. It revealed that elements like transparency, controllability, security, privacy, and seriousness positively impact user trust development, while user discomfort, uncertainty, and annoyance negatively correlate with it. The Defense Advanced Research Projects Agency (DARPA)'s

XAI program presents it as a two-stage process, separating the XAI model from the user interface for explanations [87]. Study [1] calls for in-depth research and application from professionals in HCI and cognitive sciences. The studies [168, 263, 57] discuss EI as a layer presenting reorganized explanatory information, facilitating interaction between humans and autonomy, though they don't provide explicit definitions.

In this chapter, the concept of an "Explainable Interface" is defined within the context of AI-enhanced HAT systems, distinguishing it from conventional software systems where explanations primarily serve as user guidance. The "Explainable Interface" is designed to enhance two-way communication between humans and autonomy by reasoning, communicating and adapting to the key working states from both sides.

As shown in Figure B.2, from bottom to top, **Model Explainability** concerns itself with two primary facets: the inherent interpretability of AI models and the enhancement of understanding through post-hoc explainability. Interpretability and transparency can help form explanations but are only part of the process [212]. The metrics of success here include the complexity of the model and the accuracy, completeness, robustness and efficiency of the generated explanations. Such endeavors necessitate the dedicated efforts of AI algorithm researchers. **Explanations** serve as the information communication with users by different forms of explanations considering their influence of human reasoning[239]. This necessitates collaborative efforts from experts in data scientists to translate computational outcomes into multimodal information such as graph, image, text, speech [139]. Building upon the foundation of Explainability and Explanations, **the Explainable Interface** addresses specific domain tasks by customising the presentation of complex AI processes and outcomes in user-understandable and actionable formats, promoting effective HCI through organised multimodal presentations. Achieving this requires expertise from designers[304], developers, and cognitive scientists versed in AI and specific domains. Starting from the bottom, tasks become increasingly specific and evaluations shift from focusing on the model to being more human-centered. Clarifying these definitions assists researchers in deconstructing the complexity of XAI-related issues

arising from its interdisciplinary nature.

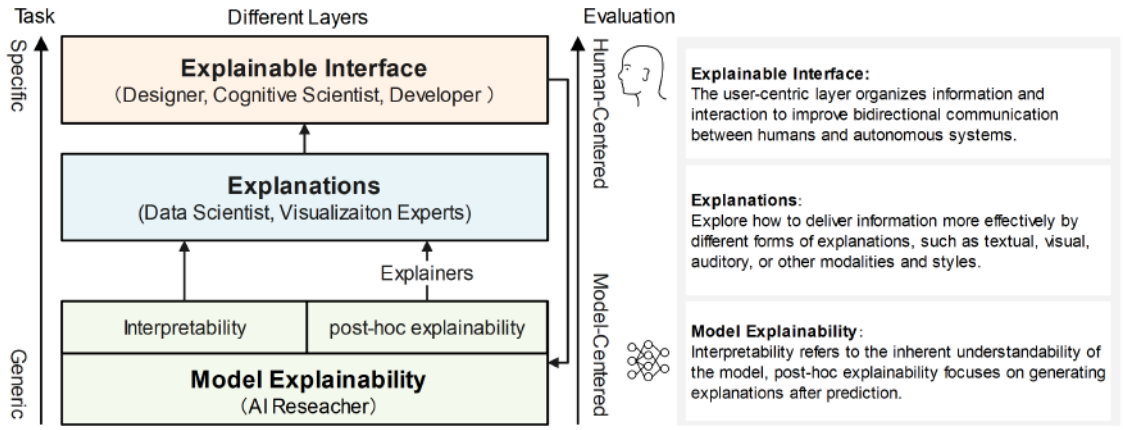


Figure B.2: Clarification of Model Explainability, Explanations, Explainable Interfaces.

B.2.2 The Need of Explainable Interface

Shahin et al. identified three key reasons for employing EI in HAT, analysed through psychological, sociotechnical, and philosophical perspectives [17]. Psychologically, EI enhances safety and reduces accidents by leveraging concepts like the Theory of Mind for user-centric explanations [6]. Sociotechnically, EI is essential for the human-focused design of systems, improving collaboration. Philosophically, it argues that AI systems should proactively provide explanations of their decisions and future strategies, aiding in deeper human understanding.

B.2.2.1 Regulation and Guidelines

DARPA formulated the XAI program in 2015 with the goal to enable end users to better understand, trust, and effectively manage AI systems [87]. General Data Protection Regulation (GDPR) emphasize on clarity in decisions stemming from human data, underscoring the imperative of user comprehension and their right to question system decisions [257]. In 2023, the US National Institute of Standards and Technology (NIST) introduced the AI Risk Management Framework 1.0 (RMF). This framework provides valuable guidance for tech entities engaged in AI, emphasizing the importance of trust and responsibility in

implementation. One common misconception is the idea that AI systems are inherently objective and error-free. This misconception can have detrimental effects on individuals and broader ecosystems. To address this, the RMF outlines key pillars of AI trustworthiness, including safety, security, resilience, explainability, privacy, fairness, accountability, transparency, validity, and reliability [5].

In healthcare, clinical stakeholders must comprehend AI recommendations to build trust and make informed decisions[42]. This becomes even more significant when certain AI models are classified as medical devices by regulatory bodies like the US Food and Drug Administration (FDA), subjecting them to strict regulations [196, 94]. In this situation, effective EIs are essential in areas such as diagnostic support tools, treatment recommendation systems, risk prediction models, and medical image analysis tools.

B.2.2.2 Expectation of Different Stakeholders

When designing EI-HAT system, it is imperative to consider stakeholders which can be categorized by roles and expertise from the start because of their different requirements.

Arrieta et al. [22] identify stakeholder categories including domain experts, data scientists, managers, and regulatory entities. Langer et al. expand this to five classes: users, developers, affected parties, deployers, and regulators, noting differing expectations like acceptance and accountability [136]. Building upon the categories outlined in [181], an additional key category is introduced: Class-A+, which includes individuals with high technical proficiency, such as pilots or operators of advanced military autonomous systems. Class-A covers end-users without specialized knowledge, who require user-friendly, simplified, understandable AI explanations and privacy considerations [115, 76]. Class-A+, for technically adept end-users, demands in-depth, sector-specific AI information [76, 226]. Class-B, encompassing developers and technicians, concentrates on detailed AI model insights and ethical development approaches, using explanations for model improvement, debugging, and modification [136, 107, 72, 246]. Class-C includes regulators and insurers, focusing on adherence to standards like GDPR and dependability

in critical applications [62, 257]. Importantly, developers' dual role is recognized: as users needing comprehensive model understanding, and as creators who must consider EI characteristics for their intended audiences.

Moreover, the range of stakeholders in XAI systems differs according to the design domain. In the healthcare sector, for example, key stakeholders include data scientists, clinical researchers, and clinicians [55, 126]. Additionally, due to the similarity in target groups, for interfaces primarily aimed at Class-A, some design experiences from business-to-customer can be referenced. For interfaces mainly targeting Class-A+ and Class B, some design experiences from business-to-business products can be useful.

B.2.3 Architecture for Explainable Interface in HAT

The design of EI-HAT systems does not adhere to a one-size-fits-all solution due to the vast diversity in applications, users, scenarios, and environments. It necessitates a case-by-case examination. Despite this, common design principles emphasise human-centred automation [42, 186, 168, 161]. This approach highlights the importance of human control, requiring human engagement and informed collaboration through effective communication. Automation should only be introduced when necessary and should be transparent, understandable, predictable, and subject to human direction. It is essential for the system to be able to understand its human counterpart, promoting a shared understanding of intentions, situational awareness, actions, and emotions between humans and automated teammates. Furthermore, the design of these systems should prioritise ease of learning and operation, ensuring simplicity and user-friendliness in human interaction.

As depicted in Figure B.3, The architecture of EI-HAT system consists of user interface, functional module, support systems, and human-autonomy loop. The interface includes functional and explainable components. The functional module comprises perception, localisation, prediction, decision, and action [215, 282]. Support systems include situation awareness, human physical behaviour recognition, mental/cognitive behaviour inference, and multimodal sensor systems. The human-autonomy loop ensures continuous feedback

and promotes mutual understanding and trust.

EI in the architecture includes explanations of models, behaviours, environments, decision-making, and sometimes external explanation [1]. It also gathers feedback from users, while the functional interface continues to focus on direct system interaction and control. EI and functional interface can either be integrated into a single, cohesive user interface or they can be separate, depending on the design and the requirements of the system. There is also a hybrid approach where the core functional interface is supplemented by on-demand explainable features. The design choice should be driven by considerations such as user experience, system complexity, and the cognitive load on the users. Moreover, to ensure the human operator can be effectively explained for that to build proper mutual understanding and mutual trust, human behaviour/intention recognition is also required.

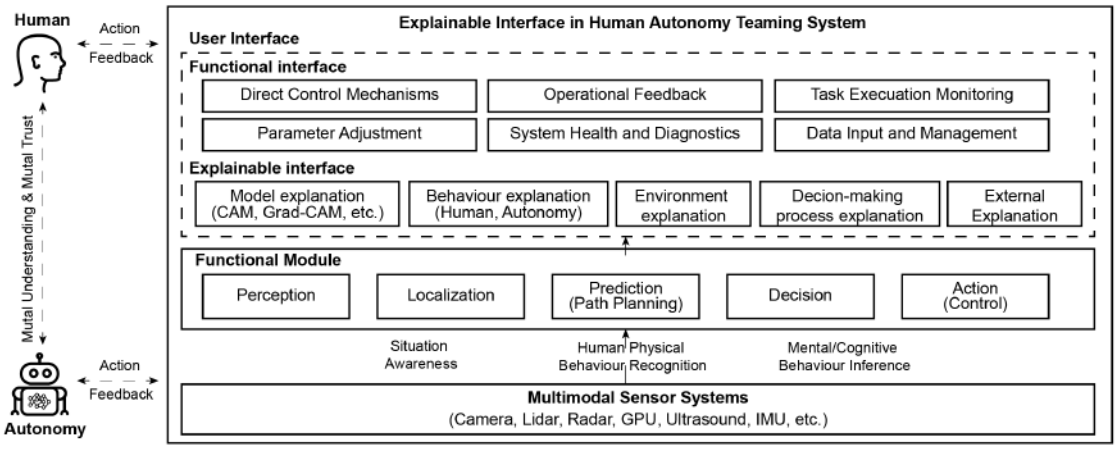


Figure B.3: Explainable Interface in Human-Autonomy Teaming System.

Within such a collaboration system, several negative consequences to the human who work together with the automation can be estimated such as out-of-loop performance, loss of situation awareness, complacency, or lack of vigilance, over-trust, and automation surprises. Accordingly, two different types of assistance from automation can be developed, which are soft protection and hard protection [110]. Soft protection in automation helps humans avoid collisions by providing warnings and hazard explanations, and addresses unsafe human behaviours within safety parameters. However, it may not always prevent negative events, especially if humans with inadequate situational awareness override

it. In contrast, hard protection actively prevents humans from taking dangerous actions through machine-initiated decisions and controls, which can sometimes lead to unexpected responses from operators unfamiliar with the system. Therefore, there is a need for EIs that allows humans to share their situational awareness with automation, comprehend and interpret its decisions and judgments, and recognize its functional limitations.

B.3 Frameworks and Techniques for Model Explainability

Explanations generated across varying phases can serve as input data for EIs. Based on the requirements of different stakeholders, appropriate information can be selected for integration. Explainability methodologies can be categorized into three stages: pre-modelling Explainability, inherently Interpretable model by Design, and post-modelling explainability [166]. This categorization facilitates the consideration and selection of explainability methods at various stages of engineering practice. The classification of model explainability methods is diverse, partly due to the multidisciplinary nature of the field. Timo Speith suggests that confusion over taxonomies in the field may delay the potential social benefits of XAI [233].

As shown in Figure B.4, a stages-based approach has been adopted to minimise overlap and provide practical guidance. In the pre-modelling stage, data analysis is conducted to set clear expectations and understand the data, addressing biases and inconsistencies early on datasets. The model design stage focuses on building inherent interpretability through transparent structures, interpretability constraints, and modular components. Post-modelling methods involve explaining the trained functional models after they made predictions.

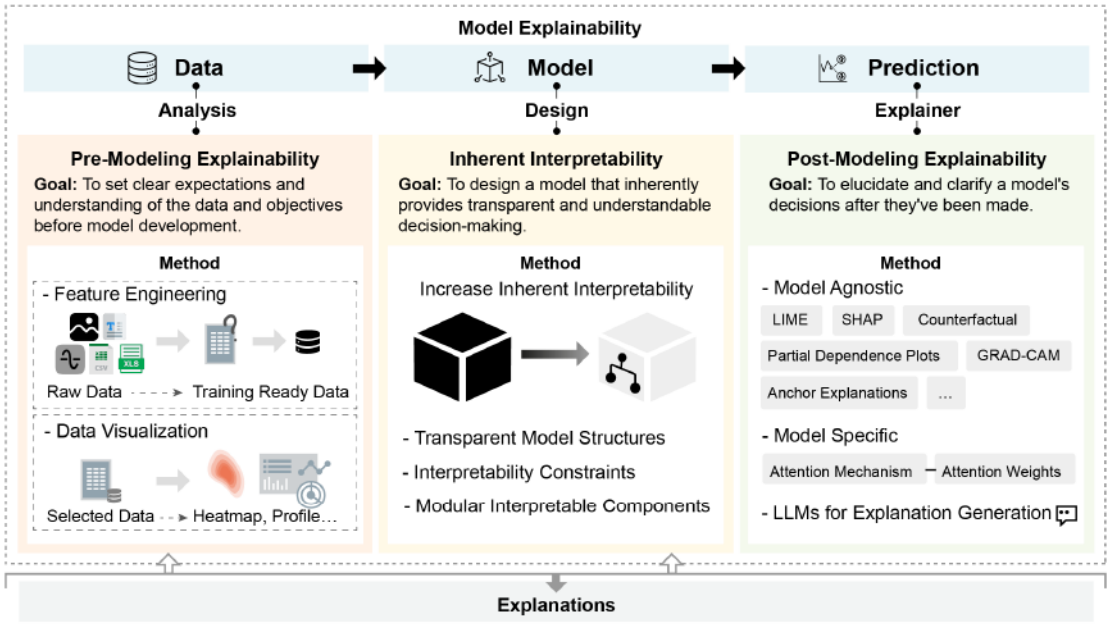


Figure B.4: A Comprehensive Framework for Model Explainability in Machine Learning.

B.3.1 Pre-modelling Explainability

Pre-modelling explainability involves data pre-processing and feature exploration for better understanding and preparation of datasets prior to model training, which is crucial for model effectiveness and interpretability [166]. It encompasses two main areas: data visualisation and feature engineering. Data visualisation utilises graphical representations, like graphs and heatmap, to reveal data patterns, with advanced methods incorporating ML to enhance design and explanation (e.g., ML4VIS studies) [265]. This technique not only helps in identifying data trends but also improves stakeholder understanding and memory retention [261]. In sectors such as healthcare, finance, and environmental monitoring, data visualisation effectively communicates complex information to non-experts, impacting attitudes and decisions [190]. Feature engineering involves creating interpretable features through domain knowledge, thus aiding in model decision-making processes [101]. This includes selecting and transforming data into relevant features to simplify the model and boost its performance [280]. The created features should be easily understandable and applicable to the specific domain, like tumor characteristics in medical imaging [42] or word frequency in natural language processing [305]. By incorporating expert input

in the feature selection process, the human-in-the-loop approach greatly enhances the accuracy and interpretability of models, particularly in fields such as image analysis and risk assessment [48]. This refinement in selecting relevant features leads to more precise and insightful results.

B.3.2 Inherent Interpretability of Model by Design

In regulated domains like healthcare and safety, inherent interpretable models are essential due to their clarity and ease of justification during regulatory reviews [236]. Post-hoc explanations for black box models may not accurately reflect the original model's computations, or they might be too vague, leading to overly complex decision-making processes prone to errors. Therefore, in high-risk scenarios, inherently interpretable models are more recommended [214]. These models are designed with a focus on embedding interpretability within their fundamental architecture and optimisation processes from the beginning, as opposed to relying on retrospective explanations applied to black box models. There are several design methods to consider when creating inherently interpretable models.

B.3.2.1 Using Transparent Model Structures

Early AI systems, using logic and symbolic reasoning, could perform deductions and provide explanations for their actions. Some commonly used methods include linear regression, decision trees, and rule-based systems.

- **Linear regression:** Statistical method for analysing the relationship between variables.
- **Decision trees:** Visual decision-making structures that use if-else rules to classify data or make predictions [148]. Their interpretability can be quantified through 'explanation size', referring to the attributes assessed from the root to a leaf [232].
- **Rule-based systems:** Transparent systems that deliver decisions based on predefined rules. Highly interpretable, commonly used in expert systems for guidance and

recommendations.

One example of an interpretable system is Full Spectrum Command, designed to facilitate training in light infantry command. It allows users to review computer-controlled soldiers' actions and decisions post-mission by clicking on them for detailed explanations[252]. However, despite its potential, this kind of systems encounter challenges such as inefficiency, high construction costs, and limitations in handling real-world complexity.

B.3.2.2 Imposing Interpretability Constraints

Beyond the architecture of models, imposing specific constraints during model development plays a pivotal role in enhancing inherent interpretability. The following types of constraints are discussed:

- **Enforced sparsity:** This constraint focuses on reducing the complexity of models by limiting the number of active features or parameters. The Sparse Neural Additive Models (SNAM) apply group sparsity regularization to achieve feature selection and improved interpretability [288]. Yu et al. [295] propose that representation learning's objective is to compress and transform data distribution into a mixture of low-dimensional Gaussian distributions on incoherent subspaces. They introduce 'sparse rate reduction' as a unified measure for the quality of representations.

- **Directed knowledge-integrated training:** This constraint aligns machine learning models with established principles and causality in a specific field. This approach involves integrating domain-specific knowledge into the training process. Various methods, such as physics-informed neural networks, saliency-guided training, causal knowledge integration, merging knowledge graphs with neural networks, and explanation-guided learning, are used to achieve this alignment [225, 109, 276, 213, 72]. In healthcare, Domain Knowledge Guided Recurrent Neural Networks (DG-RNN) demonstrate state-of-the-art performance in various

healthcare prediction tasks using real-world electronic health record (EHR) data [294].

- **Model streamlining constraints:** This approach encompasses Progressive Mimic Learning (PML) which aims to shape lightweight CNN models by emulating the learning trajectories of more complex models [154]. The use of knowledge distillation, which involves training a smaller student model to mimic the behaviourof a larger teacher model achieves a better balance of interpretability and performance [88]. These strategies effectively address both the complexity and efficiency aspects in model development.

B.3.2.3 Combining different interpretable modules

Combining different interpretable modules can have a profound impact on both the performance and interpretability of models. Several approaches have been explored that leverage the strengths of interpretable modules to achieve these goals.

- **Generalized Additive Models (GAMs):** GAMs are statistical models that enable flexible, non-linear modelling of relationships between variables. They combine multiple smooth functions to capture the complex patterns in data, with each function representing a different predictor's influence. These models are additive, with the overall output being the sum of these individual functions, and are widely used for their interpretability in fields requiring analysis of complex, non-linear relationships. such as identification of key factors in the success rate of cardiac paralysis surgery [93].

- **Neural Additive Models (NAMs):** NAMs blend the principles of interpretable regression components with the power of neural networks. This combination maintains an interpretable structure while benefiting from the expressivity of neural networks [4]. Recent advancements in NAMs include approaches from a bayesian perspective to enhance transparency, provide calibrated uncertainties, select relevant features,

and identify interactions. The development of Laplace-approximated NAMs (LA-NAMs) shows improved performance on tabular datasets and real-world medical tasks[38].

- **Attention Mechanisms:** Attention Mechanisms enable models to focus selectively on relevant parts of input data, assigning varying weights to different input components. This process aids in tracing the model's decision-making back to specific input features. It began with Bahdanau-style attention [20], which introduced the additive attention, enabling a model to automatically learn to focus on various parts of the input sequence during decoding, thereby enhancing translation performance. This concept was refined by Luong et al. with more effective and computationally efficient multiplicative attention techniques [151]. The application of attention mechanisms expanded to image captioning in [287] where hard attention was applied to selectively concentrate on specific image parts for generating descriptive text. The introduction of transformer architecture, which relies solely on self-attention mechanisms without recurrent or convolutional layers, has drastically improved the quality of machine translations and setting a new standard for sequence-to-sequence models [253]. The advancements continued with the development of Sparse Transformers, Longformer, and Big Bird, addressing the limitations of standard attention mechanisms in processing long sequences. By employing sparse attention patterns they reduce computational complexity and enable the handling of longer sequences more efficiently [45, 30, 298].

In conclusion, it's important to recognize that the methods discussed are not mutually exclusive. For example, [295] illustrates a synergistic method that combines different techniques, such as unrolling iterative optimisation schemes, reinterpreting self-attention layers, and implementing an idealised model for token distribution, all within a cohesive framework aimed at developing transformer-like network architectures. This approach not only enhances the model's mathematical interpretability but also boosts its practical performance.

B.3.3 Post-modelling Explainability

Post-modelling explainability methods are classified by approach (model agnostic or specific), scope (local or global) [233], and the type of output they generate (textual, visual, or quantitative) [255]. The process from pre- to post-modelling explainability is iterative and involves sequential optimisation. Three methods of post-hoc explanation are outlined for image and time-series datasets: factual, counterfactual, and semi-factual approaches [121].

B.3.3.1 Feature-based Explanation

Feature-based explanation is a widely used approach in XAI methods for helping users understand the importance of features. The concept of feature importance is the most prevalent explanation type [175]. Gradient-based methods, such as Grad-CAM [224] and Integrated gradients [238], compute the gradient of the model's output with respect to its inputs to generate saliency maps. DeepLIFT [227] is a method for decomposing the output prediction of a neural network on a specific input by backpropagating the contributions of all neurons in the network to every feature of the input. Perturbation-based methods alter or remove portions of the input to measure their impact on the model's output. Examples include RISE [197] and LIME [207], which are computationally intensive but applicable to various models. SHAP is a well-known feature-based explanation method that applies Shapley values derived from cooperative game theory to the context of ML. Furthermore, TsSHAP [206] targets robust, model-agnostic explainability in univariate time series forecasting.

In addition to analysing individual feature importance, understanding the interactions between features and their impact is crucial. However, generating feature interactions is more complex compared to assessing individual feature importance. Current research has explored measuring interactions through tree-based models like CART and random forests [149] and visualising partial dependence plots [108]. Nonetheless, further investigation is needed to fully understand the contribution of feature interactions to prediction results.

B.3.3.2 Example-based Explanation

Example-based methods explain models by selecting instances from the dataset rather than creating feature summaries. Example-based explanations are meaningful only when data instances can be represented in a way that is understandable to humans. It is particularly effective for images, as they can be directly viewed.

Jacob Bien et al. [34] introduce prototype methods for classification, which aim to select minimal subsets of representative samples to condense the datasets. In [123], Kim et al. highlight the limitations of prototypes in complex distributions. They propose the "MMD-critic" approach, which combines prototypes with a criticism mechanism to enhance human understanding. Jonathan Crabbé et al. [50] address the challenge of interpreting unsupervised black-box models. They introduce label-free feature importance and label-free feature example importance.

B.3.3.3 Concept-based Explanation

Concept-based explanation methods aim to interpret the internal representations or neuron activations of models against high-level concepts. These concepts are characterised by a set of samples that share specific features. These explanations can assist in scientific research by aligning model predictions with known scientific concepts, potentially uncovering new insights.

Anchors technique presents a rule-based, model-agnostic approach for generating precise, user-friendly explanations of predictions [208]. Kim et al. introduced Concept Activation Vectors (CAVs) to interpret neural networks in relation to human-friendly concepts [124]. CAVs represent the internal states of a neural network as vectors indicating a concept's direction in the model's latent space. Furthermore, Zhang et al. [300] introduced an Invertible Concept-based Explanation (ICE) framework to enhance concept-based explanations. The framework utilizes non-negative matrix factorization to generate Non-negative Concept Activation Vectors (NCAVs). Human subject experiments demonstrated that NCAVs outperformed previous methods such as PCA. In 2022, CARs is proposed as

an extension. CARs represent concepts as regions rather than fixed directions in the Deep Neural Network (DNN)'s latent space, which enable it capture more complex relationships between concepts and the DNN's internal representations [49].

In summary, feature-based methods prioritize individual data attributes, example-based methods focus on specific instances, and concept-based methods connect decisions to human-understandable concepts. These methods are often model-agnostic, allowing different explainers to be used with a functional model. The selection of an explainer should consider the domain, AI task, and stakeholder requirements. Additionally, interactive tools can enhance the selection process and improve the XAI process.

B.3.4 Causal Inference and Counterfactual Explanation

Causality and counterfactual reasoning are key in improving mental models by establishing causal connections or directing focus in HAT dynamics [12]. The D-BIAS system exemplifies this with a causality-based human-in-the-loop framework to mitigate algorithmic bias in HAT [78]. Similarly, the trust-POMDP model highlights causality's role in HCI, enabling robots to gauge human trust, understand their actions' causal effects on trust, and choose trust-enhancing actions for team success. Importantly, this model reveals that prioritizing team trust, sometimes a counterfactual objective, may not always coincide with maximizing team performance [44].

Traditional Causal inference is a field origin in 20th century in the research of statistics and medicine that dedicated to understanding and quantify the causal relationships between variables [194]. Causal inference involves discerning causal relationships between features and variables via causal learning. This entails the extraction of high-level causal variables from low-level observations, thereby facilitating a comprehensive understanding of the underlying causal mechanisms [222]. Causal learning is important in ethical AI for actionable healthcare [201].

By modelling causal relationships, machines can better support counterfactual inference therefore be more effective at modifying their behaviours to engage in a truly adaptive team

[234]. By incorporating causal relationships, which encompass valuable human knowledge, into the structural design, users can gain deeper insights into the decision-making processes of complex deep learning models, thereby enhancing their interpretability [276, 222].

Judea Pearl has proposed a theoretical framework known as the ”ladder of causality,” which elucidates three distinct levels of causal inference: association, intervention, and counterfactual [195]. Using counterfactual to create feature attribution explanations from hypothetical scenarios allows for the identification of key features influencing model classifications. This approach offers a deeper analysis than attribution-only methods by determining which features are necessary or sufficient for altering outcomes [130]. Psychological field also have demonstrated that counterfactual can provide psychologically intuitive and plausible explanations, further affirming their utility in fostering a comprehensive understanding [120]. Sandra Wachter et al. recommend utilizing the approach of counterfactual explanations without opening the black box to achieve explainability. This approach is based on three explanatory objectives: (1) informing and aiding individuals in understanding why specific decisions were made, (2) providing reasons for questioning those decisions by presenting alternative outcomes, and (3) understanding the necessary changes that need to be made to achieve desired results based on the current decision model [260].

The Recent studies show a merging of neuro-symbolic methods in causal inference and counterfactual explanation [278, 157], emphasizing the importance of domain knowledge. However, experiments on Amazon Mechanical Turk reveal that the influence of causality on decision-making varies with individual domain knowledge and doesn’t always enhance decision support [302]. Counterfactual explanations face limitations including lack of user studies, defining plausibility, sparsity issues, coverage assessment, and comparative testing [120]. Research [46] indicates that existing model-agnostic counterfactual algorithms for XAI lack a causal theoretical foundation, thus limiting their ability to convey causality to decision-makers.

B.3.5 Explainable Reinforcement Learning

Explainable Reinforcement Learning (XRL) presents unique challenges in creating and presenting explanations. These arise from its sequential, temporal, and stochastic nature, as well as the complexity of its policies and state spaces [105]. This distinctiveness positions XRL as a particularly challenging branch of XAI.

XRL seeks to improve human collaboration, visualisation, policy summarisation, query-based explanations, and verification, but faces issues like limited detail in human experiments, scalability, and the complexity of explanations for non-experts [272]. Current XRL methods often derive from supervised learning, but there is a need for RL-specific approaches that include feature importance, learning process, Markov Decision Process, and policy-level explanations [165]. XRL-Bench benchmark suite provides a standard and unified platform for researches and practitioners to develop, test, and compare state importance-based XRL algorithms [286].

Herman Yau et al. argue that traditional post-hoc explanations are impossible in traditional reinforcement learning. They propose collecting explanations during the agent's training process and derive approaches to extract local explanations based on intention. Their method ensures consistency with learned Q-values and focuses on the intended outcomes of the agent. The researchers evaluate their approach on standard environments using OpenAI Gym and show its superiority over vanilla Q-Learning by providing contrastive explanations. Even for deep Q-networks (DQN) without mathematical guarantees, they are able to approximate the agent's intention [293].

Unlike existing methods that require extensive access to an agent's inner workings, Huber et al. [105] redefine the generation of counterfactual states as a domain transfer problem, leveraging the StarGAN architecture to transform features relevant to an agent's decision while maintaining other features. This method aims to create counterfactual states that prompt the agent to choose a desired action, thereby offering a more straightforward and effective way to understand and explain RL agent strategies.

Study [122] introduces the Prototype-Wrapper Network (PW-Net), which integrates

human-friendly prototypes into the decision-making process of agents. Additionally, the Human-in-the-loop Explainability via Deep Reinforcement Learning (HEX) method aims to generate personalized explanations for individual decision makers [137]. Study [59] creates an EI for Amidar video game with six major components: agent selection, agent description, live game stream, agent performance characteristics, narrative explanation, ”what if” scenarios. Results indicate that EI increases user trust and acceptance, though not necessarily perception of prediction accuracy. Building on this, future research could explore the design and evaluation of EIs in scenarios where AI functions as a trusted collaborator in HAT, rather than operating autonomously.

In RL-enhanced systems, by maintaining human-in-the-loop through EIs, additional context information can be provided to improve the learning process. Human interaction and guidance during the learning process aligns AI decisions with human understanding and values, resulting in more explainable and reasonable behaviour from AI agents. And In combining XRL in HAT system, future research could explore immersive visualisation and symbolic representation integration [272].

B.3.6 LLMs for Explanation Generation

The development of LLMs such as GPT-4 [3] and LLaMA [249] represents a significant progression in natural language processing (NLP). The integration of multi-modality into LLMs has expanded traditional NLP, encompassing diverse data formats such as images [277], audio, and video.

LLMs assist in analysing specific behaviours of smaller models. The Transformer Debugger [171], utilising automated interpretability methods and sparse autoencoders. Study [301] proposes an approach to explain agent behaviour by distilling the agent’s policy into a decision tree, forming a behaviour representation from it, and then querying an LLM for explanations based on this representation.

Nicholas Kroeger et al. [133] explore advanced LLMs’ utility for post-hoc explanations through in-context learning (ICL) [58]. They developed various strategies like

perturbation and explanation-based methods, comparing them against standard explainers such as LIME and SHAP. The results highlighted LLMs' potential as effective post-hoc explainers. Additionally, Feldhus et al. [65] examines how converting saliency maps into natural language can make AI explanations more understandable, comparing two novel methods (model-free and instruction-based verbalizations) and evaluating their effectiveness through human and automated tests. The study [41] focused on human prediction of Visual Question Answering (VQA) model responses after receiving explanations and revealed no significant improvement in team collaboration tasks such as Failure Prediction (FP) and Knowledge Prediction (KP). Another study [213] explored integrating LLMs with Knowledge Graphs (KGs) to enhance model transparency and interpretability in automotive demand forecasting, aiming to mitigate hallucination challenges and improve understanding of LLMs with real-world data and exogenous factors. While enhanced capabilities of LLMs facilitate their application in various domains such as healthcare, finance, and transportation, for tasks like forecasts, anomaly detection, and classification, it is also important to discuss the interpretability of LLMs themselves and to improve their ability to discern cause-and-effect relationships [114].

In summary, LLMs are predominantly employed to produce explanations akin to traditional formats, enhanced with natural language integration. LLMs not only revolutionize the method of generating explanations but also how these processes and results are integrated into systems. More comprehensive strategies towards exploiting explainability in the LLM Era can refer to [277]. Further exploration is suggested into integrating LLMs as agents and connects LLMs to private data sources and APIs with frameworks such as LangChain and MetaGPT [100].

B.4 Human-Centered Explainable Interface in HAT

HAT has been defined as "the dynamic, interdependent coupling between one or more human operators and one or more automated systems requiring collaboration and coordin-

ation to achieve successful task completion" [51]. Teamwork among human involves three dimensions: cognition (team's mission and objectives understanding), skills (adaptability, communication, leadership), and attitudes (feelings about the task, mutual trust, and team cohesion) [40]. Successful HAT requires mutual understanding of team members' attitudes, abilities, intentions, and states, along with a commitment to task completion and support [43]. Additionally, cultural, experiential, and individual differences influence HAT performance [220]. Astrid Bertrand et al. [32] also discuss the impact of both normal and problematic cognitive biases on XAI design and evaluation, stressing the need for adaptive interfaces and the mitigation of biases.

In Figure B.5, the diagram depicts human-autonomy interaction mediated by an EI. The interface acts as a mediator, facilitating two-way feedback and action between humans and autonomy. The autonomy also perceives the human's reaction states, influencing the adaptability, personalisation, and optimisation of the EI. This continuous feedback loop is specifically designed to modelling improve team performance and collaboration.

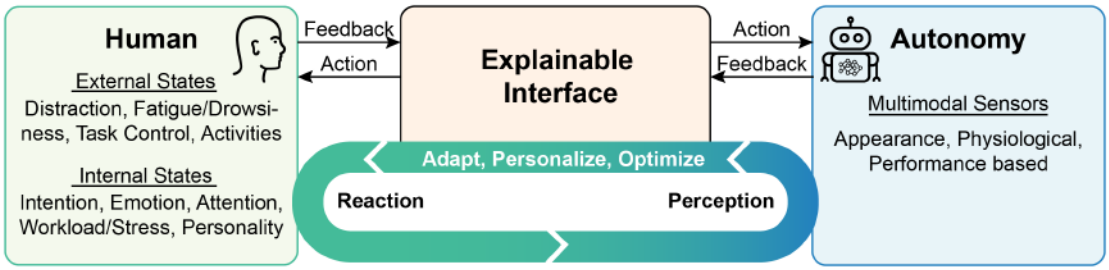


Figure B.5: Interactive Dynamics between Humans and Autonomous Systems through an Explainable Interface.

B.4.1 Human Behaviours modelling in HAT

To build a complete human behaviourunderstanding model for an EI in HAT, a variety of human states should be obtained, modelled, and incorporated. The current research for human modelling has been widely used in intelligent vehicles, autonomous piloting, and air traffic management [284]. On the human side, it lists external states such as distraction, fatigue/drowsiness, task control, and activities, along with internal states like intention,

emotion, attention, workload/stress, and personality. The autonomy side is characterized by multimodal sensors that may include data on appearance, physiological states, and performance-based metrics [102]. Appearance-based sensors include the RGB camera, depth sensor, infrared sensor, etc. They are utilized to obtain facial and bodily movements. The physiological-based sensors monitor the various physiological factors, including the heart rate, brain waves, neuromuscular electromyography signals [283] and blood pressure. The performance-based sensors collect autonomy and task-related data during the operating process, such as the control and interaction quality, etc. Existing studies focus on a specific task and ignore the collaboration between various functions, thereby limiting recognition performance. Therefore, a unified human behaviourmodel that considers both external and internal conditions is indispensable to endow the system with robust, precise, and time-variant sensing capabilities to assist the understanding of human behaviors, and hence contribute to better explanation qualities.

Human distraction, often a result of secondary activities, is a key cause of errors and misinterpretations in team settings. Research in this area primarily focuses on using appearance-based detection and physiological data such as electrodermal activity, heart rate, breathing rate, and electroencephalogram (EEG) for early identification [64, 118]. In critical and emergency situations, it is essential to communicate important signals clearly to heavily distracted operators. Incorporating audio or directional tactile warning signals can facilitate operators'rapid responses to potential collision events [95, 160].

Fatigue and drowsiness are another indication of human physical behaviors. Fatigue can be caused by either long-term repeated physical tasks or long-term monitoring and interpretation of autonomy's decisions and actions, which cause a heavy mental workload and mental fatigue [56]. Detect human fatigue and drowsiness with eye-tracking technologies to monitor the variation and dynamics of the pupil is a common solution [143]. Also, EEG and electromyography-based data can be used to analyze the brain and muscle fatigues levels [258]. To keep the human engaged and motivated, heavy fatigue must be avoided for the HAT, as heavy fatigue usually can be viewed as a key index of the failure

of HAT [204]. Thus, a key design requirement for EI is explaining the machine's thought by adapting to human's mental workload [267].

Regarding the impact of human psychological behaviors on HAT and EI designs, human intention is one of the critical aspects. Intention recognition is the process of understanding whether the ongoing activities of the agent are goal-oriented or not, and what is the goal behind these specific actions. Michon [164] stated that the cognitive structure of human user-level intention can be divided into three sub-levels: strategical, tactical, and operational level. Understanding such an intention hierarchical plays an essential role in EI design for safer and more seamless HAT [25]. Intention communication between humans and autonomy can helps building human's trust in the autonomy and identifying potential risks in advance. Similarly, understanding human intention can help autonomy better support the human and contribute to a higher team performance [185, 262, 241].

Human attention estimation is another critical aspects for EI to support conditional teamwork and help the machine provide intervention or assistance in critical scenarios [103, 113]. A typical approach for attention estimation is to build a mapping model between the context of the external scenario and the human field of attention based on deep learning [209, 79]. When designing the human-attention-oriented EI, three basic benefits can be identified. First, estimate human attention to provide corrected cues to explanation contents. For example, if the human operator is highly focused on the task, abstracted key information can be provided to avoid a heavy workload. Second, it provides a cognitive correction and self-adjustment of autonomy. In this case, the autonomy can identify if the explanation satisfies the human's focus points and assess if the most expected explanation is provided. Last, estimating human attention enables the machine to learn human behaviors and the way human's percept current situations so that a human-like situation awareness process for autonomy can be learned.

Last, the emotional states of the human for HAT has been investigated in several studies [169, 37]. These studies indicate that the human emotional state critically impacts team

performance and that the negative state is associated with aggressive collaboration that leads to poor team performance and even serious accidents. Adverse emotional states like frustration also prevent humans from receiving a correct understanding of autonomy, thereby reducing their sense of mastery, control, and task performance [270]. There are two common types of approaches can be adopted to analyze human emotional states: discrete and continuous [141, 85]. The discrete approach categorizes emotional states into specific states such as happiness, surprise, sadness, and anger. The continuous approach uses a continuous space to indicate the emotional state. Emotion recognition has been studied from various perspectives with a variety of sensors using ML technology [170].

B.4.2 Explainable Interface Design Principles

B.4.2.1 Considerations in Explainable Interface Design

Survey [163] underscores the potential consequences of inadequate adjustments in the system's level of autonomy, such as errors, control conflicts, user frustration, and ultimately, the disuse of the system . Based on our previous discussion about human behavior, to ensure that EI-HAT can intuitively respond to human behaviourand meet functional needs, it is essential to consider human-centered design principles by following approaches:

Presets, Context and Situation Awareness for Common Ground: In XAI, the concepts of presets, context, and situation awareness are pivotal for establishing common ground between human operators and AI systems. The IMPACT program illustrates this through its use of high-level 'plays' as presets, simplifying complex multi-UxV control into manageable commands, thereby enhancing situational comprehension [28]. In Transparent Value Alignment (TVA), the focus shifts to robots discerning human context and values, highlighting the significance of perceiving, and comprehending human actions to align robot behavioureffectively [218]. Meanwhile, the SAFE-AI framework delves deeper into situation awareness by categorizing it into perception, comprehension, and projection levels, each integral for AI systems to understand and explain their operational context

to human users [219]. Collectively, these approaches underscore the importance of AI systems not only perceiving their environment but also interpreting and communicating this understanding back to human collaborators, thus establishing a shared common ground essential for effective and harmonious human-AI interaction.

Personality and Adaptivity: Considering the vast cognitive diversity among human users, influenced by cultural contexts, personality traits, and past experiences, a universal approach to interface design may not suffice [161, 202]. In a study investigating the effects of adaptive versus nonadaptive robots on information exchange and social relations, it was found that when the robot's dialogue was adapted to expert knowledge, expert participants perceived the robot as more effective, authoritative, and less patronizing [248]. Furthermore, to ensure safe and efficient HAT, it's important to adjust trust in AI agents appropriately. Over-reliance on autonomous systems can lead to serious safety issues. Developing adaptive trust calibration methods is essential. These methods monitor users' reliance behaviourand use cognitive cues, called "trust calibration clues," to detect and prompt recalibration of trust [179]. Personalized interfaces cater to individual preferences and behaviors, rendering interactions more intuitive and relatable. For example, in the healthcare domain, clinical doctors and patients have significantly different needs for explanations, requiring different designs for EIs [126]. Different patients have different preferences regarding the emphasis on negative emotions in explanations and their preference for detailed or holistic explanations. Exploring interactive features such as click or hover interactions or real-time personalized explanations through eye-tracking data can be beneficial to increase system understanding and acceptance [240].

Interactivity: Interactivity is most mentioned among these considerations of EI design. It allows users to question, explore, and comprehend AI decisions, which promotes trust and alignment with human values. User feedback, visualization, and interactive tools are effective methods for enhancing transparency and facilitating user understanding of AI processes. For example, the Interactive Attention Alignment (IAA) framework improves user trust and understanding by fine-tuning DNNs to synchronize their behaviourwith

human cognitive models, optimizing both prediction accuracy and attention mechanisms. [73]. Additionally, methods such as incorporating user feedback for continuous AI learning, visualizing AI processes, and developing interactive tools that allow for the manipulation and exploration of AI models' decisions are important components. The introduction of Argumentative Exchanges (AXs) as a dynamic framework for interactive conflict resolution in XAI demonstrates the effectiveness of these methods [203]. Furthermore, the integration of domain knowledge into AI development seeks to enhance both the design of EI and the accuracy of AI algorithms by implementing dialogical guidelines and knowledge acquisition [75]. Moreover, research in Interactive Machine Learning (IML) consolidates existing knowledge on interface design. The proposed structural and behavioral model for IML systems, along with the identified principles for effective interface building, emphasizes the ongoing necessity for research in this area, further underscoring the vital role of interactivity in creating AI systems that are transparent, understandable, and aligned with human-centric values [60, 246].

Multi-modality: The presentation and interaction of information in multiple modalities are also important for EIs. Currently, explanations are primarily conveyed through visual and auditory means, such as charts, heatmap, knowledge graphs, structured text, natural language, sound cues, and speech [9, 139]. Although touch, smell, and even taste are important components of the human sensory system, they have received less exploration compared to the aforementioned modalities [290].

Sharon Oviatt et al. [184] have discussed the transition from traditional Graphical User Interfaces (GUIs) to multimodal User Interfaces (MUIs), highlighting the advantages of MUIs in natural human-computer interaction and emphasizing their alignment with natural human communication processes. Dong Huk Park et al. introduced a model designed for simultaneous textual rationale generation and attention visualization based on two multimodal explanation modules, the Visual Question Answering Explanation (VQA-X) and the Activity Explanation (ACT-X) [188]. MIRIAM, a prototype multimodal interface, provides voice assistance for operators. This system integrates natural language under-

standing, aiding in the development of clear and accurate mental models for autonomous systems. It demonstrates the potential of multimodal interfaces in improving understanding and trust for remote autonomous system operators in hazardous environments [92]. A study by Schneider et al. examines how different feedback modalities can improve the user experience in autonomous driving by testing feedback modalities like light, sound, visualization, text, and vibration. The results show that light and visualization feedback significantly enhanced the user experience in both practical and hedonic qualities compared to no feedback [221]. Holzinger et al. propose a holistic approach to automated medical decision-making that integrates ML research with human expertise. The authors introduce counterfactual graphs and emphasize the importance of incorporating multiple modalities throughout the decision pipeline by aligning local geometries across different input feature spaces [98].

Last, research on information olfaction plays a key role in human responses, despite not being a primary information channel [192]. The study reveals that olfactory channels, including aspects like direction, saturation, airflow, and air quality, effectively augment information visualization. These channels are especially beneficial in enhancing memory, emotion, and immersion, and in providing clear signals when visual information is unclear or distracted. Additionally, olfaction complements other senses by delivering crucial information, such as specific scents for gas leak warnings or driver safety alerts [82]. Although olfactory cues are limited in providing complex explanations, they can be a potential vital element for more intuitive explanations for future multimodal EIs.

In conclusion, effective design principles for EIs hinge on four core considerations: establishing common ground through context awareness, personalizing adaptivity to accommodate user diversity, fostering interactivity for deeper understanding and trust, and integrating multi-modality for richer, more intuitive interactions. These principles underscore the necessity for EI designs that are not only technically advanced but also deeply attuned to human cognitive and sensory processes. By embracing these guidelines, HAT systems can be developed that are more accessible, transparent, and aligned with human

needs, ultimately leading to more effective and harmonious human-AI collaborations.

B.4.2.2 Extended User Experience Framework for Explainable Interface

Susanne Bødker emphasizes the role of HCI from the perspective of human activity theory in 1989 [36], this approach serves as the foundation of user centred design in HCI, advocating for interfaces that are intuitive, accessible, and tailored to the specific needs of users. For example, in the field of medical AI, the incorporation of Human-centered design approaches holds particular significance as it ensures that AI not only assists medical professionals in decision-making but also aligns with their workflow and cognitive processes [99, 281].

The development of EIs has led to the formulation of guidelines and frameworks that bridge the gap between algorithm-centric and user-centred explanations [146, 271]. These guidelines emphasis on involving UX and design practitioners in the creation of AI products that not only possess technical proficiency but also cater to user needs and contexts. Additionally, there is a need for tools that make it easier for designers and developers to design, develop, and evaluate EIs. For instance, ProtoAI introduces a new workflow called Model-Informed Prototyping (MIP) for designing AI-enhanced user interfaces, enabling designers to incorporate real-time AI model outputs into UI prototypes. This approach simplifies data-driven design, aligns with human-AI interaction principles, enables detailed prototyping with AI model insights, and supports interface simulation with various data and model scenarios [235].

Garrett introduced ”The Elements of User Experience.” The diagram delineates various levels of considerations that are essential for crafting a high-quality user experience (UX) [74]. In the context of HAT, where automation takes different forms and operations are influenced by both human input and AI. As shown in Figure B.6, the following extensions to this framework are proposed based on earlier discussions, aimed at enhancing EIs within HAT.

- **Task-oriented explainable interface:** Ideal for environments where the primary

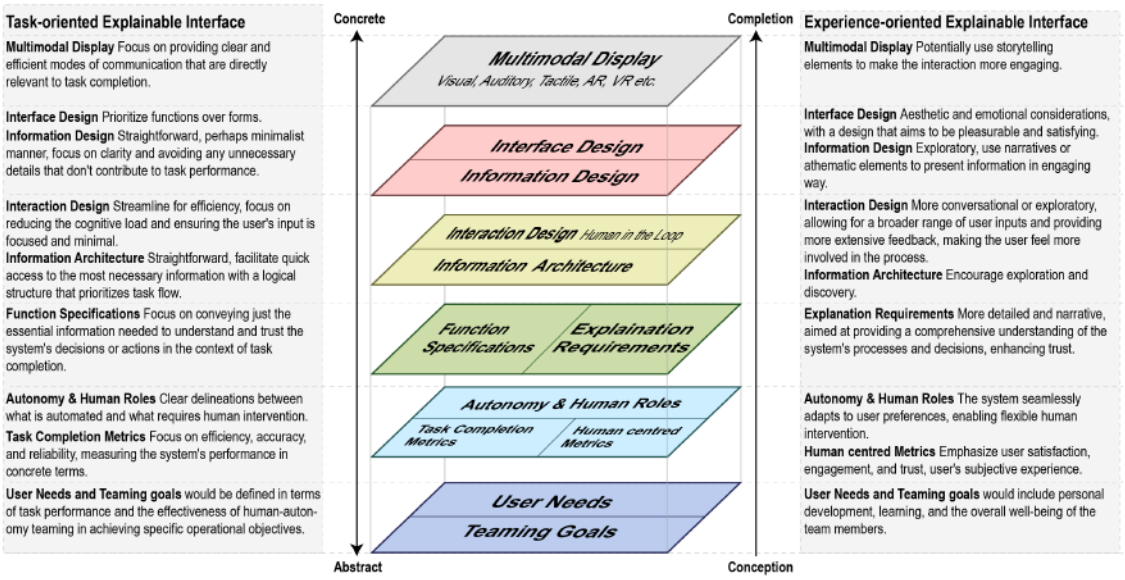


Figure B.6: Extended User Experience Framework for Explainable Interface: Contrasting Task-oriented and Experience-oriented Approaches in Human-Autonomy Teaming.

objective is to efficiently and accurately complete tasks, such as in industrial, medical, or military settings.

- **Experience-oriented explainable interface:** Most suitable for scenarios where user engagement, long-term trust, and an intuitive understanding, such as in educational, customer service, or domestic environments.

The task-oriented perspective emphasizes efficiency, clarity, and practicality in achieving specific tasks, with a focus on the concrete aspects of the interaction. The Experience-oriented perspective, however, focuses on the quality of the interaction, aiming to enrich the user's experience and foster long-term benefits such as learning and personal growth. Both perspectives are valuable, and the optimal approach may involve balancing the two to create an interface that is both efficient and engaging for human-autonomy collaboration.

B.5 Evaluation of Explainable Interface in HAT

In the HAT context, evaluating EI extends beyond traditional XAI methods, necessitating a human-centered approach that aligns evaluation methods with various roles' goals [110]. A

paradigm shifts in XAI evaluation, as suggested in [211], prioritizes metrics that measure the explanation's relevance to the XAI's objectives, such as the discrepancy between the explanation's logic and the agent's actual performance, the number of generated rules, the complexity of the explanation, and its stability. Bhatt et al. [33] suggest three criteria for ideal feature-based explanations: low sensitivity to input changes, high faithfulness in identifying relevant features, and simplicity, all of which can be automated without human involvement. DARPA's framework [86] introduces measures including user satisfaction, mental model accuracy, task performance, trust assessment, and correctability. Different experimental conditions are used for evaluation, ranging from no explanation to full explanations for each decision and action, along with a control condition using a non-explainable system. Research [256] categorizes explanation evaluation methods into objective and human-centered evaluations. Objective evaluations apply automated approaches and metrics to assess explainability, while human-centered evaluations focus on aspects like explanation quality, user satisfaction, comprehension, and performance. A distinction is made between the plausibility and convincingness of explanations versus their correctness, highlighting the importance of not conflating these criteria [175]. Enyan Dai et al. [52] outline essential qualities for explanations, such as correctness, completeness, consistency, contrastivity, user-friendliness, and causality, and suggests quantitative metrics for evaluating Graph Neural Network (GNN) explanations, including accuracy, fidelity, sparsity, and stability.

Building on existing research, as shown in Figure B.7, dividing evaluation metrics into three categories is proposed: model performance-centered, human-centered, and those aligned with HAT goals. This comprehensive evaluation framework is designed to ensure both technical efficiency and user-centric effectiveness in HAT systems, thereby promoting optimal integration and collaboration between humans and autonomous systems.

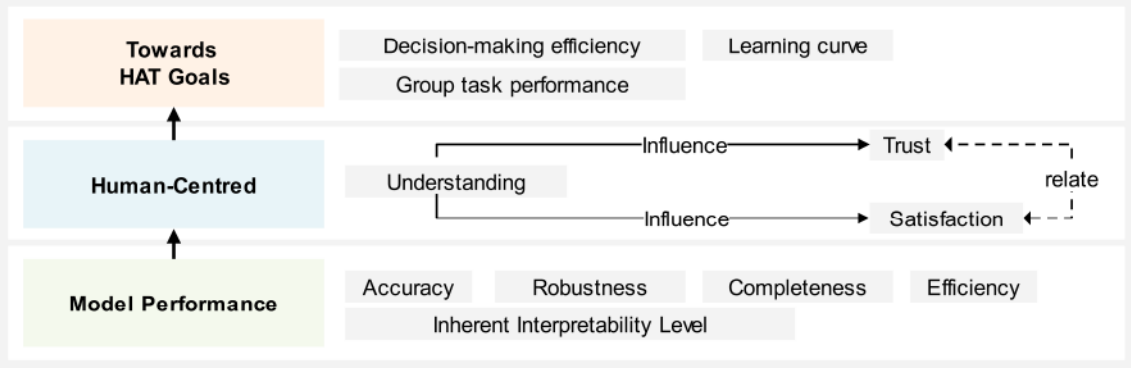


Figure B.7: Evaluation Framework of Explainable Interface for Human-Autonomy Teaming.

B.5.1 Model Performance Centred Evaluation Methods

B.5.1.1 Inherent Interpretability Level Evaluation for Functional Model

The inherent interpretability of a model largely hinges on its architecture. Simple models, such as linear regressions, small decision trees, or rule-based systems, are more interpretable due to their straightforward decision-making processes [22]. Within the same architecture, interpretability also varies with the model's complexity. In decision trees, fewer nodes or leaves indicate simpler decision rules, making the model more interpretable. Conversely, larger trees with more nodes capture complex patterns but are harder to understand and visualize. This complexity can be measured by the tree's number of nodes or depth [232]. Saliency or activation maps help interpret the focus of each layer in CNN or RNN [49, 191]. However, as the network depth increases, generating explanations becomes more challenging, especially for non-experts.In addition, when addressing complex tasks like vehicle state recognition and fault diagnosis, researchers often propose frameworks that integrate multiple basic models [282] to enhance performance. However, quantitatively comparing the interpretability of different architectures within such frameworks remains a more challenging endeavor.

B.5.1.2 Quantitative Evaluation for Post-hoc Explainers

Accuracy involves correctly representing the reasons for generating a specific output, particularly when the ground-truth rationale is known. It is closely linked to fidelity,

which measures the faithfulness of identified substructures using heuristic metrics [52]. CLEVR-XAI [14] is a sizable visual question answering dataset featuring questions paired with pixel-level ground truth masks, serving as a benchmark for assessing visual explanations. In practical applications, the ground-truth, such as real feature importance, is often unknown. Simulating models and treating the system's output as ground truth is a common approach to assess explanation fidelity. This method utilizes ML metrics, including accuracy, precision, recall, F1-score, AUC, and others, to evaluate post-hoc explanations [168, 207]. Researchers also evaluated explanation accuracy by removing higher-order features and introducing noise to the inputs [159]. Hugues et al. [250] created a synthetic dataset with adjustable complexity and known discriminative features. They evaluated six popular explainers on both synthetic and real datasets like FordA and ECG. They introduced two new metrics, AUC$\widetilde{S}$top and F1$\widetilde{S}$, for assessing interpretability in time-series classification. AUC$\widetilde{S}$top measures the area under the curve as the top-k time steps are corrupted, reflecting explainers' accuracy in identifying important time steps.

Robustness is crucial for ensuring the stability and precision of AI explanations across varying conditions. Although a model's predictions might be consistent, its explanations can vary, highlighting the importance of robust explanations for reliable decision-making [135]. Section B.3 outlines the heterogeneity in XAI explanation methods, such as feature importance and heatmap, complicating the identification of the most reliable method without a standard ground truth. To tackle this, Huang et al. introduce two novel black-box evaluation techniques for assessing interpretation discrepancy and probabilistic robustness. These include using a genetic algorithm for worst-case scenario analysis and subset simulation for assessing overall system robustness [104]. However, current post-hoc explainers exhibit limitations in robustness, especially in identifying biases in classifiers for sensitive applications. Common methods like LIME and SHAP often fail to detect biases related to protected attributes such as race [231]. Furthermore, the robustness of GNN post-hoc explainers against label noise is another critical evaluation aspect, underscoring the need for accurate explanations in noisy data [303].

Completeness in post-hoc explanations involves a balance between detailed insights into a model's internal mechanisms (reasoning-completeness) and the clarity and relevance of its outputs (output-completeness). Reasoning-completeness focuses on the depth of understanding a model's internal workings, where white-box models are essential for complete transparency in system operations. In contrast, output-completeness concerns how comprehensively the explanation represents the model's output [175]. To assess output-completeness, the F1$\widetilde{S}$ score is used as a metric, providing a balanced evaluation of an explainer's ability to identify important time steps [250]. However, while aiming for high completeness in explanations, it's crucial to strike a balance between sparsity and completeness for user comprehension. Explanations that are too sparse may omit critical information, whereas overly detailed ones can lead to complexity or redundancy, impeding user understanding.

Efficiency is particularly important in real-time or interactive systems, where prompt explanations are essential for maintaining user engagement and trust. In critical domains like medical diagnostics, financial decision-making, and autonomous vehicles, the ability to provide fast and clear explanations is crucial for timely decision-making, system trust, and avoiding potential harm or economic loss. Efficiency metrics include the speed of generating explanations and the computational resources needed. A proposed benchmark in [107] aims to address limitations in current explainers and foster the development of more effective, utility-driven methods. Study [237] indicates that gradient-based methods like GradCAM outperform perturbation-based methods in efficiency in medical image analysis. This efficiency stems from gradient-based methods requiring only a single input pass for explanation, as opposed to multiple perturbed inputs used by perturbation-based methods, leading to quicker processing times.

Above all, the model performance evaluation discussed can function independently of end-users, and its performance significantly impact the generation of explanations, which is a fundamental step in developing EIs.

B.5.2 Human-Centered Evaluation Methods

Different individuals perceive EIs distinctively. In coronary heart disease interface evaluations, non-experts showed higher satisfaction with various explanations compared to experts [186]. Mohammad et al. found that in clinical decision support systems, example-based and counterfactual explanations were more comprehensible and less likely to foster over-trust in AI than local and global ones [77]. The study [223] examines the effects of self-driving car explanations on user trust and understanding. It found that causal explanations were least satisfying, intentional explanations increased trust the most, and mixed explanations yielded the best overall understanding and consistent trust. These findings illustrate that EI-HAT differs from general model explainability, with no universal best explanation applicable across scenarios and audiences. These highlight the necessity for human-centered evaluation such as understanding, trust, and satisfaction.

Human understanding makes a basic of HAT in complex, dynamic environments because it is essential to build a shared mental model for human and autonomy. Situation Awareness (SA) playing a key role in this process [173]. For a lower level of autonomy, it's important to maintain human engagement and responsibility for the task, with the autonomous system providing warnings and support via sophisticated assistance technologies [47]. Paul Salmon et al. [216] propose that the achievement and maintenance of SA is influenced by both individual and task factors, such as experience, training, workload and also interface design. The method for measuring SA must be consistent and reliable, producing comparable assessments under similar circumstances. Evaluating SA should satisfy three core criteria: it should enable simultaneous measurement across different situations, assess awareness on both an individual and team level, and provide insights in real-time.

Human trust is shaped by the user's AI knowledge, task understanding, and objectives. EIs are designed to foster appropriate trust, aligning with AI's capabilities to avoid both under utilization and excessive reliance. For instance, excessive trust in autonomous driving could lead to neglecting necessary manual interventions, risking accidents [138].

In autonomous vehicles, establishing trust requires understanding the system's abilities and limitations [91]. Trust is sensitive to system errors and tends to fluctuate over time [23]. Measuring human trust in EIs can be a key indicator of their effectiveness, looking at how trust is built and adjusted through various types of explanations [96]. Trust is connected to team interaction and failure recovery, suggesting that it is a dynamic process best measured through real-time, sensor-based, and behavior-based methods, beyond traditional surveys [54].

Human satisfaction with an EI doesn't necessarily imply correct system understanding or trust [97]. Satisfaction with explanations is determined by the extent to which users perceive the ease of use and utility of an AI system or process. Key characteristics of satisfactory explanations include completeness, actionability/persuasiveness, usefulness, relevance, personalization, and timeliness. These attributes can be measured using various methods like surveys, questionnaires, user interviews, feedback after interaction, user experience metrics, metrics from recommendation systems, and assessments of contextual relevance [245]. Ashraf Abdul et al. [2] introduced the Cognitive-GAM (COGAM) framework, which formalizes cognitive load for XAI by using visual chunks and aims for a balance between cognitive load and accuracy in generating explanations.

Measuring these aspects effectively in practice necessitate active participation from individuals. Comparative experiments are key in evaluating EIs, by contrasting interfaces with and without explanations or examining different types of EIs. Hoffman's research [97] delves into the explanation process, highlighting curiosity's role, and introduces a detailed set of scales to measure aspects like explanation quality, satisfaction, mental models, and trust, useful in question-and-answer assessments. Myounghoon Jeo's research [112] suggests that AI systems evoking positive emotions can boost trust. Mishra et al. [167] developed a hierarchical LSTM model using eye-tracking data for sentiment analysis in text. Mengyao Li et al. [144] proposed using conversational data, focusing on lexical and acoustic features, to assess trust dynamically in NASA missions. Advanced affective computing techniques involve emotion recognition through text, audio, and

visual cues, as well as physiological data like EEG or ECG, often analyzed using ML [266]. The DEVCOM Army Research Laboratory's HAT3 toolkit [177] integrates various indicators, including behavioural and physiological cues, for trust analysis in human-autonomy interactions.

While human-centered evaluation methods are vital, focusing exclusively on one aspect, like high trust levels, is not always indicative of superior results. It's important to value both objective and subjective measurements. Affective computing frequently utilizes emotional models, including Ekman's six basic emotions [63], Plutchik's wheel [199], or the Pleasure-Arousal-Dominance (PAD) model [21], and is supported by well-developed multimodal databases [266]. However, the relationship between these emotions and aspects like understanding, trust, and satisfaction need further exploration. What's more, in HAT, Dynamic evaluation of human-centered metrics such as trust calibration can allow for continuous improvement and adaptation of the system to better support the team's needs.

B.5.3 Evaluation Methods towards Teaming Goals

Human-centered evaluation are not always align with the performance of HAT [97]. Though they are important, the ultimate goal is to successfully complete tasks. Systems that fail to enhance collective goal achievement, despite advantages in interpretability and satisfaction, are deemed suboptimal. This underscores the importance of evaluating EIs in the context of HAT objectives. Factors such as learning curves, decision-making efficiency, and group task performance play a critical role in the dynamics of HAT.

The learning curve can be evaluated by measuring the performance of the HAT on specific tasks. Over time, the team should become more proficient, completing tasks more quickly and with fewer errors [67]. The learning curve can be evaluated through various measures such as task completion time, error rates, physiological indicators, psychological factors (such as self-reported affective state and stress), and perceived workload and team efficiency. Characteristics of autonomous agents, team composition, task attributes, human individual difference variables and training significantly influence the HAT performance

[182].

Group task performance represents the most direct measure of their effectiveness or usability. The Perceptual-Cognitive Explanation (PeCoX) framework, which aims to improve HAT performance, underscores the importance of generating, conveying, and receiving explanations [176]. When considering EIs within this framework, they act as a bridge facilitating team interaction. Thus, assessing EIs must focus on key aspects of team performance, including the efficiency, completeness, and quality of task execution. These factors offer tangible metrics to evaluate how EIs enhance the overall effectiveness and collaborative success of the HAT system.

B.6 Challenges and Future Directions

B.6.1 Challenges

Incorporating XAI into HAT systems and facilitating interaction through EIs poses a variety of critical challenges:

- **Timely explanation generation:** In high-stakes, real-time HAT settings, it's essential for EIs to offer timely and understandable explanations in proper time to foster appropriate trust and assist in decision-making.
- **Explanation generation methods selection:** The right method for generating explanations is complex and depends on factors like data complexity, target audience, and context of use. The range of methods varies from simple rule-based systems to complex ML models, and the use of different post-hoc explainers can lead to inconsistent results [132].
- **Explanation presentation methods selection:** How explanations are presented in EIs is as important as the explanations themselves. The presentation method can greatly affect HAT performance. The same information can be perceived differently depending on whether it is presented via text, visuals, audio, tactile, or olfactory

means. The choice of presentation method varies with the task type and environment, requiring careful selection to ensure effectiveness and appropriateness for the specific context and user needs.

- **Complexity in designing explainable interface frameworks:** Creating these frameworks is an interdisciplinary effort, requiring knowledge from AI, HCI, psychology, and relevant domains. This necessitates a collaborative approach that extends beyond traditional disciplinary limits, Personalizing explanations to accommodate individual differences is essential [273].

- **Personalized and adaptive design:** Interfaces should be tailored to users' varying cognitive abilities and expertise levels. This involves modelling human behaviourdynamics and adjust EIs accordingly, thereby fostering accurate understanding, reasonable trust, and increased satisfaction.

- **Hybrid dynamic evaluation:** As discussed in Section B.5, evaluating these interfaces requires a balanced approach. This includes assessing the model's performance, the human-centered evaluation, and the overall performance of the HAT system. Methods like questionnaires and affective computing should be utilized, and more empiral research is needed for specific tasks and domains.

- **Preparing users for HAT through explainable interfaces:** It is vital to equip users with the necessary knowledge about the system for effective teaming. The design of EIs that can achieve this educational goal is an area that requires further exploration.

- **Multi-disciplinary collaboration:** Incorporating EIs into HAT systems goes beyond technical considerations. It requires viewing autonomy as a participant in human society and designing responsible AI ecosystems. Bridging the gap between responsible AI research, industry, policy, and regulatory efforts necessitates a comprehensive integration of art, humanities, and other disciplines [39].

B.6.2 Directions

To effectively enhance HAT systems through EIs, a comprehensive and structured human participated approach is essential. The approach should commence with a re-evaluation of traditional prototype design and software development processes, prioritizing the seamless integration of XAI either as a distinct module or within the functional interface. A pivotal element of this strategy involves establishing a decision framework. This framework should facilitate the selection of explanation methods that align with the intricacies of the data, the user's comprehension level, and the contextual application of the system.

The debate between the use of interpretable models and post-hoc explanations for black-box models has gained attention in recent research [214]. Efforts have been made to explore neuro-symbolic AI, which aims to integrate the continuous and discrete aspects of neuro and symbolic AI [29]. Additionally, causal inference and IML approaches have been examined [150, 246]. However, despite the emergence of more interpretable model architectures, there are challenges in designing systems that effectively explain these models to humans. Moreover, it is crucial to focus on developing user-friendly tools for interactive post-hoc explainer selection, allowing non-AI experts to participate in system design [90].As discussed in section B.3.6, the utilization of advanced multimodal LLMs to explain simpler black box models is another valuable direction, providing human-comprehensible explanations and simplifying the EI development process. However, the application of LLMs in design, the mitigation of potential issues arising from LLM hallucinations to produce reliable explanations, and the interpretability LLM itself remain areas for exploration.

For explanation presentation, it is good to draw on insights from visualization and conduct empirical researches. These researches should evaluate the impact of multimodal content presentation methods, like visual, textual, or auditory formats, on user understanding and interaction. When establishing bridges between autonomy and humans through EIs, considering presets, along with incorporating interactivity, adaptability, and multimodality, has the potential to improve the performance of HAT. To ensure comprehensive evaluation,

a hybrid dynamic multimodal approach is recommended. This combines quantitative methods like model performance metrics with qualitative methods such as user feedback and surveys. Additionally, integrating affective computing to assess human behaviourprovides valuable insights for interface refinement.

B.7 Conclusion

Drawing on the survey findings, this chapter emphasizes the necessity for ongoing research to improve aspects like understandability, trustworthiness, user satisfaction, and the overall efficiency and effectiveness of XAI-enhanced HAT systems. Crucial concepts such as explainability, explanations, and EIs within HAT systems are clarified, helping researchers taking a first step into this field from various domains to discern the current state, challenges, and potential focal points. This chapter serves as a valuable resource for stakeholders involved in XAI systems, proposing a design framework for EIs. It also introduces an elaborate evaluation framework that containing model performance, human-centric factors, and HAT task objectives. This framework deconstructs the intricacies involved in assessing EIs and XAI systems. Furthermore, the challenges and potential avenues for future research in this area are specifically detailed, both from explanation generation and from the design and integration of these explanations within the system by EIs.

B.8 Explainable Interface Application Scenarios in HAT: Case Studies

Though EIs play a pivotal role in enhancing the efficiency of HAT in high-risk sectors like healthcare, transportation, and manufacturing, there is a lack of empirical studies examining end-users' explainability needs and behaviors around XAI explanations in real-world contexts [127]. Research [26] outlines a framework to evaluate robot autonomy levels in human-robot interaction (HRI). It redefines autonomy as a robot's ability to perceive,

plan, and act in its environment for specific tasks. A 10-point taxonomy categorizes levels of robot autonomy (LORA), assessing how these levels influence key HRI factors like acceptance and reliability. The study also underscores autonomy's interplay with constructs like safety and trust in HRI, essential for comprehending autonomy's role and impacts in these interactions. Related research is summarised in Table B.2, which illustrates how, across varying levels of autonomy, the roles of humans and autonomous systems differ, thereby influencing the functionality of EIs in these contexts. Notably, the concept of variable autonomy in HAT for firefighting, introduced in [254], allows for the dynamic adjustment of robot autonomy based on situational demands. This approach leads to efficient task sharing and role adaptation between humans and robots, implying a need for similarly dynamic EIs in these contexts.

B.8.1 Healthcare

B.8.1.1 Clinical Decision Support Systems

EIs are vital in Clinical Decision Support Systems (CDSS) for clinician acceptance and transparency of AI'decision-making is more effective than merely presenting AI confidence levels [187]. Factors influencing the effectiveness of explainability in CDSS include technical integration, algorithm validation, deployment context, decision-making roles, and user demographics [11]. Computer-based CDSS differ in operation, knowledge sources, and workflow impact. A substantial portion focuses on patient-centered decision support, while a smaller percentage provides online support for clinicians. Most require additional personnel for data management [31]. Implementing CDSS involves considering delivery methods, governance, incentives, workflow integration, content updates, and data quality [274]. AI-based CDSS are evolving with complex algorithms, analyzing large datasets for precise, personalized support, improving healthcare quality and efficiency. However, issues such as data privacy, transparency of algorithms, and integration into existing workflows present ongoing challenges [205]. Key elements in the design and

Table B.2: Design of Explainable Interface for Different Autonomy Levels

Level	Human Role	Autonomy Role	Functions of EIs
Low Autonomy	Direct Controller: Executes immediate control, frequent commands, continuous monitoring	Active Executor: Acts upon human input nearly real-time, no independent decision-making.	Clarify operational mechanisms and respond to direct human commands. [26, 31]
Medium Autonomy	Supervisor/Guide: Involves setting parameters, providing intermittent input, and intervening in complex decisions beyond the agent's programming.	Assistant: Offers insights into decision-making under specific conditions	Provide insights into decision-making, especially under specific conditions or parameters. [26, 147]
High Autonomy	Overseer/Monitor: Engages in high-level monitoring and intervenes in unexpected situations.	Self-regulated Actor/Adaptive Manager: Performs tasks independently, seeks input when uncertain	Offer detailed explanations of complex decisions, highlighting the influencing algorithms, behaviors, or data. Justify actions retrospectively [152].
Full Autonomy	Initializer/Auditor: Establishes initial parameters and high-level goal setting. Beyond the setup, humans are mostly hands-off, only intervening in exceptional cases or for maintenance.	Sovereign Operator: Manages all tasks autonomously, constantly improving or adapting its strategy based on goals and feedback from its environment. Human input is rare.	Explain overall strategies and objectives, clarify adjustments and rationale in autonomous strategies, and ensure consistency between human intentions and autonomous actions. [152]

implementation of EIs for CDSS include:

- Real-time visualization: EIs should display system's decision-making processes and actions in real-time, allowing users to observe and understand how the system arrives at its conclusions. This feature enables clinicians, operators, or users to monitor and intervene when necessary [11].
- Alerts and notifications: EIs generate alerts and notifications to inform users about potential complications, anomalies, or critical events. This feature enables timely intervention by support staff, who can take appropriate actions based on the provided information.
- Context-aware assistance: EIs should deliver situation-specific information, like patient condition, current procedures, and relevant literature, aiding support staff, nurses, and anesthesiologists in recognizing potential issues and making informed decisions [292]. External Explanations also matters, such as biomedical literature or prediction process allow clinicians to validate system reasoning against their own knowledge and determine if recommendations are clinically appropriate [198, 292].
- Natural language explanation: Providing natural language explanations and visualizations can help clinicians understand the robot's control logic and reasoning, improving trust and confidence in the technology [158]. Study [251] shows that LLMs in CDSS can clarify medical diagnoses but must be used cautiously.
- User-Centered Design: Emphasizing the importance of aligning with clinical workflows and user needs, human-centered design is crucial in CDSS development [264]. This approach involves iterative processes, including prototyping, testing, and redesigning [186].

B.8.1.2 Surgical Robotics

The development of autonomy in surgical robots can be categorized based on a combination of human involvement and the robot's level of proficiency. At Level 0, the robot is

completely under the control of the surgeon, serving merely as a tool. In contrast, Level 5 envisions a robot performing an entire surgery independently. Intermediate levels involve the robot transitioning from providing mechanical assistance to autonomously executing specified tasks, suggesting surgical methods awaiting human confirmation, and even making medical judgments under the supervision of a doctor [291]. This range from Level 0 to Level 5 encapsulates a progression from total human oversight to complete self-sufficiency of the robot. Currently, most surgical robots operate at the level 1 and level 2, with the surgeon playing a significant role in the procedure. However, advancements in AI and sensors are improving the precision, grip, feedback, and autonomy of robotic surgery systems [68].

Figure B.8 illustrates Level 1, 'Robot Assistant', where a continuous interaction loop exists between the developer and the doctor. In this level, EI provides tool status updates and enhanced visualizations, like 3D renderings of the surgical area, for clearer insights [53]. Level 2, 'Task Autonomy', positions the doctor in a supervisory role. Here, control over the robot is discrete rather than continuous. An application of this is in spinal surgery, where robotic guidance is integrated with preoperative planning software [147]. EIs for doctor include feedback on decisions, reasons, confidence levels, alternative suggestions, and real-time updates. In the future, it is possible that surgical robots will achieve higher levels of autonomy, allowing them to perform more complex tasks independently while still being supervised by human surgeons. However, there also are researches arguing for a shift in focus from complete autonomy to enhancement of human capabilities in the context of autonomous surgery [24]. As AI becomes more integrated into healthcare, challenges arise, especially concerning explainability, utility, trust, legal, regulatory, and ethical [291, 183, 279].

The surgical robotics field is characterized by varied interaction methods to enhance surgical precision or improve experience. Developments include robots capable of environment reconstruction and force feedback [145]. Minimally invasive surgery primarily consists of a series of sub-tasks, which can be broken down into basic gestures or contexts.

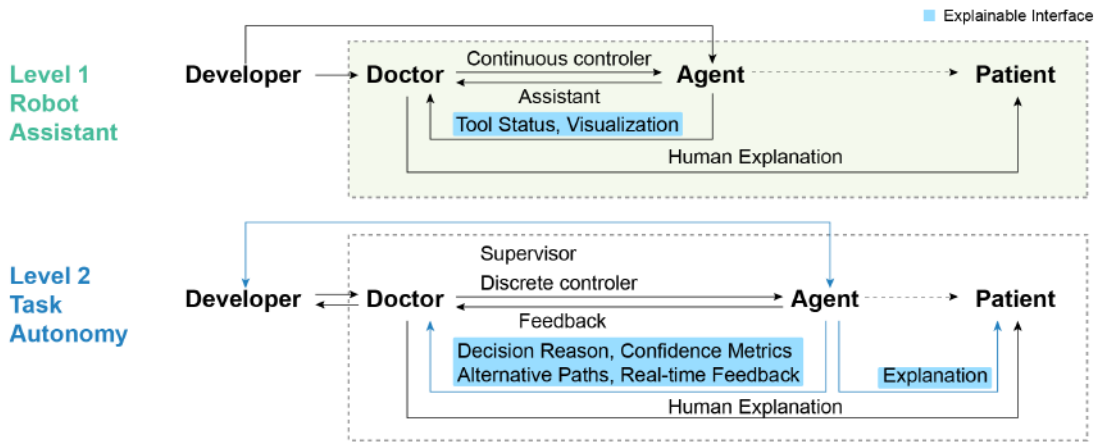


Figure B.8: Explainable Interface in level 1 and level 2 surgical robotics

Surgical gesture recognition can aid in motion planning and decision-making, as well as establish context-aware knowledge to improve the control quality of surgical robots [299]. Furthermore, as robots become more autonomous, the importance of strategic task planning becomes increasingly clear [80]. There is a trend towards optimizing shared autonomy, advocating for a more collaborative approach between humans and robots. This human-in-the-loop collaborative method enhances human capabilities alongside robotic autonomy [81]. A study reviews current surgical robotic interfaces, highlighting advancements and limitations in feedback and visualization, and proposes a novel design featuring bedside access, sterile procedure, 3D head-mounted visualization, and intuitive control for enhanced surgical outcomes [230].

B.8.1.3 Challenges fo Explainable Interface in Healthcare

Based Advancements in healthcare suggest that their future depends not only on technological progression but also on creating interfaces that integrate seamlessly with AI, prioritizing clarity and comprehensibility. Design and implementing EIs in healthcare faces unique challenges:

- Complexity of Medical Data: The inherently intricate and varied nature of medical data, including electronic health records, medical imaging, genomics, and sensor data, presents a significant challenge. XAI systems must be adept at interpreting and

explaining decisions based on this diverse, high-dimensional data [13]. Furthermore, the lack of standardized datasets exacerbates the difficulty of model training.

- Real-time Clinical Decision Making: Given the high stakes involved in healthcare decisions, the critical nature of surgical procedures, the necessity for real-time decision-making is paramount.[11].

- Varied User Expertise: The range of technical and medical knowledge among XAI system users in healthcare, such as doctors, nurses, and patients, varies widely [126]. Crafting explanations that are understandable to this diverse user base is a considerable challenge, necessitating customization to suit different comprehension levels.

- Regulatory Compliance and Ethical Considerations: Healthcare's strict regulatory environment demands compliance, especially as the FDA starts to view certain AI models as medical devices, necessitating stringent regulatory frameworks [94].

- Integration with Clinical Workflows: The effectiveness of EIs in healthcare hinges on their seamless integration with existing clinical workflows. This necessitates interdisciplinary collaboration to ensure that the implementation of these interfaces does not introduce additional complexity or time-consuming burdens for healthcare professionals [186].

- Evaluating Explanation Effectiveness: The difficulty in evaluating healthcare explanations lies in their subjective and objective combined nature, as effectiveness may vary based on their role in clinical decision-making, patient understanding, and improving outcomes.

- Data Privacy and Security: Ensuring the privacy and security of sensitive health data used in XAI systems is paramount. This includes addressing concerns related to data sharing, consent, and anonymization [7].

B.8.2 Transportation

B.8.2.1 Autonomous Vehicles

HAT in transportation and autonomous vehicles (AVs) can be categorized into six levels of driving automation, as defined by the Society of Automotive Engineers (SAE) [47]. As the level of automation increases, the role of the human driver shifts from active control to supervision. This transition can lead to challenges in maintaining driver engagement, situation awareness, and trust in the automation system [27]. Level 3 and higher AVs are equipped to handle full control, allowing drivers to focus on non-driving activities. However, human intervention is still required in certain situations. Providing explanations within AVs enhances the driver's sense of control and understanding of the vehicle's status and capabilities, reducing potential anxiety. These intervention needs differ depending on non-driving tasks, driving conditions, and takeover methods. Personalized AVs, tailored to individual driver behaviors, enhance reliability and foster greater trust in the system [217, 134]. Shahin et al.'s survey about XAI in autonomous driving, proposes an framework for a safer, transparent, publicly approved, and environmentally friendly intelligent vehicles [17]. As self-driving cars become more prevalent, regulatory bodies might see the value in having these types of interfaces [289]. Implementing such interfaces not only be beneficial for users but can also help manufacturers meet demands and gain approvals from regulatory agencies [35]. To present EIs inside and outside vehicles, visual, auditory, and haptic/tactile aspects should be considered:

- Visual interface: Visual interfaces in AVs have evolved from traditional physical controls to advanced touchscreens and augmented reality head-up displays (AR-HUD). They enable passengers to set destinations, monitor progress, environment, vehicle status, and manage settings like climate and entertainment, thereby improving efficiency in control takeover, enhancing user satisfaction, and trust [180, 116].

- Auditory interface: Speech output is highly relevant to improve driver attitudes that affect acceptance of automated systems is brought forth [70]. In AVs, auditory

user interfaces are designed to ensure safe control transitions, especially during the conditional self-driving phase. These interfaces use a combination of voice and sound feedback to communicate with the user. Compared to generic auditory output, communicating upcoming automated manoeuvres additionally by speech led to a decrease in self-reported visual workload and decreased monitoring of the visual HMI [174].

- Haptic/tactile interface: AVs can enhance user experience and control through the integration of haptic feedback interfaces. A study compared two types of driver-vehicle interfaces for controlling vehicles: one using a haptic interface with both kinesthetic and tactile feedback, and the other using a hand-gesture interface with augmented reality feedback. Compared to manual driving, semi-autonomous control significantly reduced the perceived workload. The haptic interface, in particular, proved to be especially reliable and effective in comparison to the gesture interface, as it provides direct and immediate feedback to the driver [156].

Multimodal explanations utilizing two types of displays are more effective than unimodal or three-display multimodal explanations, with speech-based displays producing the best takeover performance. The combination of visual, auditory, and tactile cues, especially visual-auditory warnings, improves response efficiency in scenarios like vehicle malfunctions and highway exits [139, 297]. The MIRIAM study demonstrates a multimodal interface combining visual indicators with a conversational agent, maintaining situation awareness across various cognitive styles and enhancing user engagement through natural language interaction [210]. The timing of information delivery is as crucial as its content, with traffic risk probability guiding when explanations are provided, particularly to avoid overwhelming passengers with excess information [125]. Figure B.9 delineates the requirements for EIs in AVs across three operational conditions: 'good condition' for standard autonomous operation, 'specific condition' for complex scenarios like adverse weather requiring increased notifications, and 'takeover condition' triggered by system errors or unpredictable circumstances necessitating control transfer. EIs should avoid

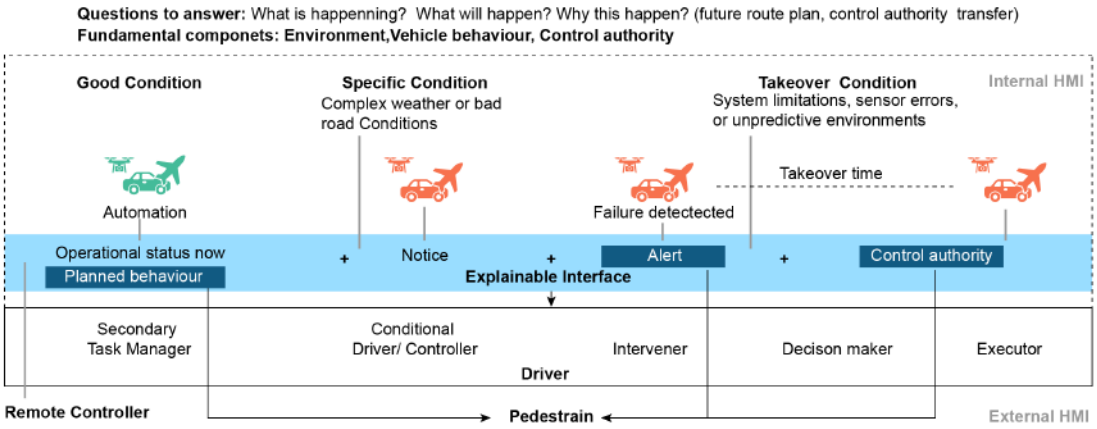


Figure B.9: Explainable Interface Requirements for Human Vehicle Interaction in Conditional Driving Automation

overloading users with complex technical details, instead clearly presenting current and anticipated environmental conditions, vehicle behaviors, and control information. The interface should also allow users to retrospectively track significant environmental changes, behaviors, and control shifts during the journey. Incorporating predictive maintenance data, such as remaining useful life and time-to-failure, is important [228]. As for remote control, more comprehensive dynamic team situation awareness is needed [173].

In evaluating autonomous systems, it is imperative to assess understanding, trust, and satisfaction in real-time interactions during the operation of AVs. This approach ensures more accurate and useful results compared to traditional questionnaires, which may not be effective in this context. Measurements should integrate the use of existing vehicle devices, such as cameras, and wearable devices capable of monitoring human psychological activities for dynamic modelling of human behaviourin HAT.

B.8.2.2 Air Transportation Systems

The integration of EIs is pivotal in enhancing the functionality and safety of transportation systems. The European Union Aviation Safety Agency (EASA) is actively working towards developing guidelines for AI applications intended for safety-related applications in aviation. The stakeholders, including air traffic controllers, airport operators, and Uncrewed Traffic Management (UTM) service providers, seek assurance that the AI

solutions are trustworthy and safe [247]. Developing robust solutions to manage, measure, and maintain trust between humans and autonomous systems is important, particularly in maintaining existing safety standards in aviation.

In Air Traffic Management (ATM), XAI is instrumental in managing air traffic flow and capacity, ensuring efficient resource utilization, and minimizing delays. leading to the funding of various projects by SESAR (Single European Sky ATM Research). Among these, the TAPAS specializes in air traffic flow, capacity management, and air traffic control services within ATM. The project's scope includes establishing transparency requirements, developing a prototype, implementing an explainable layer, and formulating principles for transparency in ATM automation processes, as referenced in [244]. ARTIMATION improves runway delay predictions and optimizes departures and inspections by enhancing explainability with heatmaps and decision-based storytelling. Additionally, the Mindtooth tool evaluates air traffic controllers' reactions, including Approach-Withdrawal, stress, and workload using neurophysiological metrics. This is particularly relevant considering that time pressure is a critical factor in high-stakes environments like air traffic control, where rapid and accurate decision-making is imperative [15].

Concerning Demand-Capacity Balancing (DCB), challenges such as congestion or hotspots occur when airspace demand exceeds capacity. Research in this area includes a multi-agent RL approach that uses a rule-based environment and a modified proximal policy optimization (MAPPO) framework. This approach integrates a supervisor for enhanced learning, effectively balancing demand, and capacity in air traffic flow management [242]. In [131], EI enhances the explainability of deep multi-agent reinforcement learning solutions for demand-capacity balance (DCB) problems, employing visual analytics for transparent decision-making. Additionally, XAI can drive secure data chains in conjunction with blockchain networks for autonomous ATM, ensuring data integrity in ATM operations [18].

B.8.2.3 Challenges for Explainable Interface in Transportation

Summarizing key findings, integrating XAI into transportation systems is crucial for high safety and operational standards. Major challenges include:

- Integration with Existing Infrastructure: Merging XAI with existing transportation infrastructure, such as traffic signals and GPS systems.
- Adapting to Dynamic Environments such as unpredictable traffic, weather, and pedestrian behavior.
- Real-Time Data Processing and Interaction: In high-stake real-time HAT, swiftly processing dynamic data, provide understandable explanations in proper time and adjust EIs according to feedback.
- Real-time Evaluation: Collecting evaluation data during HAT process, without distracting the user.
- Challenges in Advanced Urban Transportation: The emergence of intricate urban air traffic, featuring drones and eVTOLs (electric vertical take-off and landing vehicles), presents unique challenges and prospects for incorporating EIs in AVs and traffic control systems. Addressing these issues calls for creative solutions to preserve both efficiency and safety within complex urban transportation networks.

B.8.3 Explainable Interface and Digital Twin

B.8.3.1 Relationship of Physical System, Digital Twin, and XAI

The relationship between a physical system and its digital twin (DT) is dynamic and symbiotic. The DT mirrors the life of its corresponding physical entity, adapting its real-time behaviourin response to inputs from the physical system. In the context of Industry 4.0, Cyber-Physical Systems (CPS) provide real-time data and control capabilities [83]. Utilizing DT for decision-making, they enable the simulation and modelling of various

aspects of a physical system, offering valuable insights and predictions. These simulations can serve multiple purposes, such as optimizing performance, predicting maintenance needs, and simulating potential scenarios in smart cities, manufacturing, healthcare [71]. EIs play a crucial role in this relationship by ensuring transparency and understandability in the decision-making processes within the DT.

B.8.3.2 XAI Methods for Digital Twin

XAI methods play a pivotal role in DT. For example, LSTM neural networks, employed for lane change prediction in autonomous driving DT, use techniques such as layer-wise relevance propagation for better clarity in their predictions [269]. Furthermore, architectures like TwinExplainer integrate XAI into deep learning-powered DT systems, prioritizing the interpretability of their predictive outcomes [178]. The study [129] investigated the remaining useful life prediction in DT systems through interpretable machine learning, employing the PiML library for transparent and reliable AI decision-making.

B.8.3.3 Digital Twin as an Explainable Interface

DTs surpass traditional explainable interfaces (EIs) by offering interactive features and counterfactual analysis. This improves understanding of AI-driven decisions across various industries, like healthcare and manufacturing. For instance, In drug discovery, DTs explain AI's reasoning behind diagnoses and treatments, boosting trust and confident decision-making by healthcare professionals. [16]. In the manufacturing industry, DTs can provide real-time explanations for quality control decisions, helping operators identify the root causes of defects and optimize production processes [189].

B.8.4 Challenges for XAI and Digital Twin Integration

The integration of XAI methodologies into DT systems presents interface design and optimization challenges [111]. First, incorporating XAI not only as a post-hoc approach to explain the outputs, but also integrating internal interpretable AI from the initial stages

of design. Moreover, the progression of DT technology is hindered by a general lack of understanding of its advantages and a deficiency in technical expertise. Contributing technologies such as 3D simulations, IoT, AI, big data, and cloud computing are also in developmental stages, which adds to the complexity [243]. Furthermore, advancements in high-performance computing and real-time interaction technologies are crucial for the enhancement of XAI-enabled DTs. This necessitates a high level of technical understanding from users. To address these issues, there is a need for the development of new methods and frameworks. These should enable the seamless integration of XAI into DT systems, thus lowering the learning curve for users and making it easier for them to understand and trust these systems.

References

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3173574.3174156

[2] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y. Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376615

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report.

[4] Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey Hinton. 2021. Neural Additive Models: Interpretable Machine Learning with Neural Nets. arXiv:2004.13912 [cs, stat]

[5] NIST AI. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0).

[6] Arjun R. Akula, Keze Wang, Changsong Liu, Sari Saba-Sadiya, Hongjing Lu, Sinisa Todorovic, Joyce Chai, and Song-Chun Zhu. 2022. CX-ToM: Counterfactual Explanations with Theory-of-Mind for Enhancing Human Trust in Image Recognition

Models. *iScience* 25, 1 (2022), 103581. https://doi.org/10.1016/j.isci.2021.103581

[7] A.S. Albahri, Ali M. Duhaim, Mohammed A. Fadhel, Alhamzah Alnoor, Noor S. Baqer, Laith Alzubaidi, O.S. Albahri, A.H. Alamoodi, Jinshuai Bai, Asma Salhi, Jose Santamaría, Chun Ouyang, Ashish Gupta, Yuantong Gu, and Muhammet Deveci. 2023. A Systematic Review of Trustworthy and Explainable Artificial Intelligence in Healthcare: Assessment of Quality, Bias Risk, and Data Fusion. *Information Fusion* 96 (2023), 156–191. https://doi.org/10.1016/j.inffus.2023.03.008

[8] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy Artificial Intelligence. *Information Fusion* 99 (2023), 101805. https://doi.org/10.1016/j.inffus.2023.101805

[9] Kamran Alipour, Jurgen P. Schulze, Yi Yao, Avi Ziskind, and Giedrius Burachas. 2020. A Study on Multimodal and Interactive Explanations for Visual Question Answering. https://doi.org/10.48550/arXiv.2003.00431 arXiv:2003.00431 [cs]

[10] Francesco Alzetta, Paolo Giorgini, Amro Najjar, Michael I. Schumacher, and Davide Calvaresi. 2020. In-Time Explainability in Multi-Agent Systems: Challenges, Opportunities, and Roadmap. *Explainable, Transparent Autonomous Agents and Multi-Agent Systems* 12175 (2020), 39–53. https://doi.org/10.1007/978-3-030-51924-7_3

[11] Julia Amann, Dennis Vetter, Stig Nikolaj Blomberg, Helle Collatz Christensen, Megan Coffee, Sara Gerke, Thomas K. Gilbert, Thilo Hagendorff, Sune Holm,

Michelle Livne, Andy Spezzatti, Inga Strümke, Roberto V. Zicari, and Vince Istvan Madai. 2022. To Explain or Not to Explain?—Artificial Intelligence Explainability in Clinical Decision Support Systems. *PLOS Digital Health* 1, 2 (2022), e0000016. https://doi.org/10.1371/journal.pdig.0000016

[12] Robert W Andrews, J Mason Lilly, Divya Srivastava, and Karen M Feigh. 2023. The Role of Shared Mental Models in Human-AI Teams: A Theoretical Review. *Theoretical Issues in Ergonomics Science* 24, 2 (2023), 129–175.

[13] Anmol Arora, Joseph E. Alderman, Joanne Palmer, Shaswath Ganapathi, Elinor Laws, Melissa D. McCradden, Lauren Oakden-Rayner, Stephen R. Pfohl, Marzyeh Ghassemi, Francis McKay, Darren Treanor, Negar Rostamzadeh, Bilal Mateen, Jacqui Gath, Adewole O. Adebajo, Stephanie Kuku, Rubeta Matin, Katherine Heller, Elizabeth Sapey, Neil J. Sebire, Heather Cole-Lewis, Melanie Calvert, Alastair Denniston, and Xiaoxuan Liu. 2023. The Value of Standards for Health Datasets in Artificial Intelligence-Based Applications. *Nature Medicine* 29, 11 (2023), 2929–2938. https://doi.org/10.1038/s41591-023-02608-w

[14] Leila Arras, Ahmed Osman, and Wojciech Samek. 2022. CLEVR-XAI: A Benchmark Dataset for the Ground Truth Evaluation of Neural Network Explanations. *Information Fusion* 81 (2022), 14–40. https://doi.org/10.1016/j.inffus.2021.11.008

[15] ARTIMATION. 2022. ARTIMATION: Measuring Acceptance and Human Performance in ARTIMATION. https://www.artimation.eu/measuring-acceptance-and-human-performance-in-artimation/. Accessed on 2024-03-31.

[16] Heba Askr, Enas Elgeldawi, Heba Aboul Ella, Yaseen A. M. M. Elshaier, Mamdouh M. Gomaa, and Aboul Ella Hassanien. 2023. Deep Learning in Drug Discov-

ery: An Integrative Review and Future Challenges. *Artificial Intelligence Review* 56, 7 (2023), 5975–6037. https://doi.org/10.1007/s10462-022-10306-1

[17] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. 2023. Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions. https://doi.org/10.48550/arXiv.2112.11561 arXiv:2112.11561 [cs]

[18] Louise Axon, Dimitrios Panagiotakopoulos, Samuel Ayo, Carolina Sanchez-Hernandez, Yan Zong, Simon Brown, Lei Zhang, Michael Goldsmith, Sadie Creese, and Weisi Guo. 2023. Securing Autonomous Air Traffic Management: Blockchain Networks Driven by Explainable AI. arXiv:2304.14095 [cs]

[19] Carlos RB Azevedo, Klaus Raizer, and Ricardo Souza. 2017. A vision for human-machine mutual understanding, trust establishment, and collaboration. In *2017 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE, 1–3.

[20] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural Machine Translation by Jointly Learning to Align and Translate. https://doi.org/10.48550/arXiv.1409.0473 arXiv:1409.0473 [cs, stat]

[21] Iris Bakker, Theo van der Voordt, Peter Vink, and Jan de Boon. 2014. Pleasure, Arousal, Dominance: Mehrabian and Russell Revisited. *Current Psychology* 33, 3 (2014), 405–421. https://doi.org/10.1007/s12144-014-9219-4

[22] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[23] Kenny Basso and Cristiane Pizzutti. 2016. Trust Recovery Following a Double Deviation. *Journal of Service Research* 19, 2 (2016), 209–223. https://doi.org/10.1177/1094670515625455

[24] Edoardo Battaglia, Jacob Boehm, Yi Zheng, Andrew R. Jamieson, Jeffrey Gahan, and Ann Majewicz Fey. 2021. Rethinking Autonomous Surgery: Focusing on Enhancement over Autonomy. *European Urology Focus* 7, 4 (2021), 696–705. https://doi.org/10.1016/j.euf.2021.06.009

[25] Mahmut Bayazit and Elizabeth A. Mannix. 2003. Should I Stay Or Should I Go? Predicting Team Members' Intent to Remain in the Team. *Small Group Research* 34, 3 (2003), 290–321. https://doi.org/10.1177/1046496403034003002

[26] Jenay M. Beer, Arthur D. Fisk, and Wendy A. Rogers. 2014. Toward a framework for levels of robot autonomy in human-robot interaction. *J. Hum.-Robot Interact.* 3, 2 (jul 2014), 74–99. https://doi.org/10.5898/JHRI.3.2.Beer

[27] Matthias Beggiato, Franziska Hartwich, Katja Schleinitz, Josef Krems, Ina Othersen, and Ina Petermann-Stock. 2015. What would drivers like to know during automated driving? Information needs at different levels of automation.. In *7. Tagung Fahrerassistenzsysteme*.

[28] Kyle Behymer, Clayton Rothwell, Heath Ruff, Michael Patzek, Gloria Calhoun, Mark Draper, Scott Douglass, Derek Kingston, and Doug Lange. 2017. Initial evaluation of the intelligent multi-uxv planner with adaptive collaborative/control technologies (IMPACT). *Beavercreek: Infoscitex Corp* (2017).

[29] Vaishak Belle, Michael Fisher, Alessandra Russo, Ekaterina Komendantskaya, and Alistair Nottle. 2024. Neuro-Symbolic AI + Agent Systems: A First Reflection on Trends, Opportunities and Challenges. In *Autonomous Agents and Multiagent Systems. Best and Visionary Papers*, Francesco Amigoni and Arunesh Sinha (Eds.). Springer Nature Switzerland, Cham, 180–200.

[30] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. arXiv:2004.05150 [cs]

[31] Amy Berlin, Marco Sorani, and Ida Sim. 2006. A Taxonomic Description of Computer-Based Clinical Decision Support Systems. *Journal of Biomedical Informatics* 39, 6 (2006), 656–667. https://doi.org/10.1016/j.jbi.2005.12.003

[32] Astrid Bertrand, Rafik Belloum, James R. Eagan, and Winston Maxwell. 2022. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 78–91. https://doi.org/10.1145/3514094.3534164

[33] Umang Bhatt, Adrian Weller, and José M. F. Moura. 2020. Evaluating and Aggregating Feature-based Model Explanations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan, 3016–3022. https://doi.org/10.24963/ijcai.2020/417

[34] Jacob Bien and Robert Tibshirani. 2011. Prototype Selection for Interpretable Classification. *The Annals of Applied Statistics* 5, 4 (2011), 2403–2424. https://doi.org/10.1214/11-AOAS495

[35] Francesco Biondi, Ignacio Alvarez, and Kyeong-Ah Jeong. 2019. Human - System Cooperation in Automated Driving. *International Journal of Human–Computer Interaction* 35, 11 (2019), 917–918. https://doi.org/10.1080/10447318.2018.1561793

[36] Susanne Bødker. 1989. A Human Activity Approach to User Interfaces. *Human-Computer Interaction* 4, 3 (1989), 171–195. https://doi.org/10.1207/s15327051hci0403_1

[37] Patrícia J. Bota, Chen Wang, Ana L. N. Fred, and Hugo Plácido Da Silva. 2019. A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals. *IEEE access : practical innovations, open solutions* 7 (2019), 140990–141020. https://doi.org/10.1109/ACCESS.2019.2944001

[38] Kouroche Bouchiat, Alexander Immer, Hugo Yèche, Gunnar Rätsch, and Vincent Fortuin. 2024. Improving Neural Additive Models with Bayesian Principles. https://doi.org/10.48550/arXiv.2305.16905 arXiv:2305.16905 [cs, stat]

[39] BRAID UK. 2024. https://braiduk.org/. Accessed: 2024-04-08.

[40] Janet A. Cannon-Bowers, Scott I. Tannenbaum, Eduardo Salas, and Catherine E. Volpe. 1995. Defining Team Competencies: Implications for Training Requirements and Strategies. In *Team Effectiveness and Decision Making in Organizations*, Richard A. Guzzo and Eduardo Salas (Eds.). Jossey-Bass, San Francisco, CA, 333–380.

[41] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do Explanations Make VQA Models More Predictable to a Human? arXiv:1810.12366 [cs]

[42] Haomin Chen, Catalina Gomez, Chien-Ming Huang, and Mathias Unberath. 2022. Explainable Medical Imaging AI Needs Human-Centered Design: Guidelines and Evidence from a Systematic Review. *npj Digital Medicine* 5, 1 (2022), 1–15. https://doi.org/10.1038/s41746-022-00699-2

[43] Jessie YC Chen and Michael J Barnes. 2014. Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues. *IEEE Transactions on Human-Machine Systems* 44, 1 (2014), 13–29.

[44] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa.

2018. Trust-Aware Decision Making for Human-Robot Collaboration: Model Learning and Planning. arXiv:1801.04099 [cs]

[45] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating Long Sequences with Sparse Transformers. arXiv:1904.10509 [cs, stat]

[46] Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. 2022. Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications. *Information Fusion* 81 (2022), 59–83. `https://doi.org/10.1016/j.inffus.2021.11.003`

[47] On-Road Automated Driving (ORAD) Committee. 2021. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles.

[48] Alvaro H. C. Correia and Freddy Lecue. 2019. Human-in-the-Loop Feature Selection. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (2019), 2438–2445. `https://doi.org/10.1609/aaai.v33i01.33012438`

[49] Jonathan Crabbé and Mihaela van der Schaar. 2022. Concept Activation Regions: A Generalized Framework For Concept-Based Explanations. arXiv:2209.11222 [cs]

[50] Jonathan Crabbé and Mihaela van der Schaar. 2022. Label-Free Explainability for Unsupervised Models. arXiv:2203.01928 [cs]

[51] Haydee M Cuevas, Stephen M Fiore, Barrett S Caldwell, and Laura Strater. 2007. Augmenting Team Cognition in Human-Automation Teams Performing in Complex Operational Environments. *Aviation, space, and environmental medicine* 78, 5 (2007), B63–B70.

[52] Enyan Dai, Tianxiang Zhao, Huaisheng Zhu, Junjie Xu, Zhimeng Guo, Hui Liu, Jiliang Tang, and Suhang Wang. 2023. A Comprehensive Survey on Trustworthy Graph Neural Networks: Privacy, Robustness, Fairness, and Explainability. arXiv:2204.08570 [cs]

[53] Evelyn Dappa, Kai Higashigaito, Jürgen Fornaro, Sebastian Leschka, Simon Wildermuth, and Hatem Alkadhi. 2016. Cinematic Rendering – an Alternative to Volume Rendering for 3D Computed Tomography Imaging. *Insights into Imaging* 7, 6 (2016), 849–856. https://doi.org/10.1007/s13244-016-0518-1

[54] Mustafa Demir, Nathan J. McNeese, Jaime C. Gorman, Nancy J. Cooke, Christopher W. Myers, and David A. Grimm. 2021. Exploration of Teammate Trust and Interaction Dynamics in Human-Autonomy Teaming. *IEEE Transactions on Human-Machine Systems* 51, 6 (2021), 696–705. https://doi.org/10.1109/THMS.2021.3115058

[55] Sanjoy Dey, Prithwish Chakraborty, Bum Chul Kwon, Amit Dhurandhar, Mohamed Ghalwash, Fernando J. Suarez Saiz, Kenney Ng, Daby Sow, Kush R. Varshney, and Pablo Meyer. 2022. Human-Centered Explainability for Life Sciences, Healthcare, and Medical Informatics. *Patterns* 3, 5 (2022), 100493. https://doi.org/10.1016/j.patter.2022.100493

[56] Jesús Díaz-García, Inmaculada González-Ponce, José Carlos Ponce-Bordón, Miguel Ángel López-Gajardo, Iván Ramírez-Bravo, Ana Rubio-Morales, and Tomás García-Calvo. 2021. Mental Load and Fatigue Assessment Instruments: A Systematic Review. *International Journal of Environmental Research and Public Health* 19, 1 (2021), 419. https://doi.org/10.3390/ijerph19010419

[57] Vicente Dominguez, Pablo Messina, Ivania Donoso-Guzmán, and Denis Parra. 2019. The Effect of Explanations and Algorithmic Accuracy on Visual Recommender Systems of Artistic Images. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 408–416. https://doi.org/10.1145/3301275.3302274

[58] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu

Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A Survey on In-context Learning. arXiv:2301.00234 [cs]

[59] Jeff Druce, Michael Harradon, and James Tittle. 2021. Explainable Artificial Intelligence (XAI) for Increasing User Trust in Deep Reinforcement Learning Driven Autonomous Systems. arXiv:2106.03775 [cs]

[60] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (2018), 1–37. https://doi.org/10.1145/3185517

[61] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, and Rajiv Ranjan. 2023. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *Comput. Surveys* 55, 9 (2023), 194:1–194:33. https://doi.org/10.1145/3561048

[62] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3411764.3445188

[63] Paul Ekman. 1971. Universals and Cultural Differences in Facial Expressions of Emotion. *Nebraska Symposium on Motivation* 19 (1971), 207–283.

[64] Hesham M. Eraqi, Yehya Abouelnaga, Mohamed H. Saad, and Mohamed N. Moustafa. 2019. Driver Distraction Identification with an Ensemble of Convolutional Neural Networks. https://doi.org/10.48550/arXiv.1901.09097 arXiv:1901.09097 [cs, stat]

[65] Nils Feldhus, Leonhard Hennig, Maximilian Dustin Nasert, Christopher Ebert, Robert Schwarzenberg, and Sebastian Möller. 2023. Saliency Map Verbalization:

Comparing Feature Importance Representations from Model-free and Instruction-based Methods. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, Bhavana Dalvi Mishra, Greg Durrett, Peter Jansen, Danilo Neves Ribeiro, and Jason Wei (Eds.). Association for Computational Linguistics, Toronto, Canada, 30–46. https://doi.org/10.18653/v1/2023.nlrse-1.4

[66] Jacques Ferber. 1999. *Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence* (1 ed.). Addison-Wesley Longman Publishing Co., Inc., USA.

[67] Eléonore Ferrier-Barbut, Philippe Gauthier, Vanda Luengo, Geoffroy Canlorbe, and Marie-Aude Vitrani. 2022. Measuring the Quality of Learning in a Human–Robot Collaboration: A Study of Laparoscopic Surgery. *ACM Transactions on Human-Robot Interaction (THRI)* 11, 3 (2022), 1–20.

[68] Paolo Fiorini, Ken Y. Goldberg, Yunhui Liu, and Russell H. Taylor. 2022. Concepts and Trends n Autonomy for Robot-Assisted Surgery. *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers* 110, 7 (2022), 993–1011. https://doi.org/10.1109/JPROC.2022.3176828

[69] B. J. Fogg. 2003. Prominence-Interpretation Theory: Explaining How People Assess Credibility Online. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems (CHI EA '03)*. Association for Computing Machinery, New York, NY, USA, 722–723. https://doi.org/10.1145/765891.765951

[70] Yannick Forster, Frederik Naujoks, and Alexandra Neukum. 2017. Increasing Anthropomorphism and Trust in Automated Driving Functions by Adding Speech Output. In *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 365–372. https://doi.org/10.1109/IVS.2017.7995746

[71] Aidan Fuller, Zhong Fan, Charles Day, and Chris Barlow. 2020. Digital Twin:

Enabling Technologies, Challenges and Open Research. *IEEE Access* 8 (2020), 108952–108971. https://doi.org/10.1109/ACCESS.2020.2998358

[72] Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. 2024. Going Beyond XAI: A Systematic Survey for Explanation-Guided Learning. *Comput. Surveys* (2024), 37. https://doi.org/10.1145/3644073

[73] Yuyang Gao, Tong Steven Sun, Liang Zhao, and Sungsoo Ray Hong. 2022. Aligning Eyes between Humans and Deep Neural Network through Interactive Attention Alignment. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–28. https://doi.org/10.1145/3555590

[74] Jesse James Garrett. 2011. *The Elements of User Experience: User-Centered Design for the Web and Beyond* (2nd ed ed.). New Riders, Berkeley, Calif.

[75] Anne Gerdes. 2021. Dialogical Guidelines Aided by Knowledge Acquisition: FTC 2020 - Future Technologies Conference 2020. *Proceedings of the future technologies conference (FTC) 2020* 1 (2021), 243–257. https://doi.org/10.1007/978-3-030-63128-4_19

[76] Julie Gerlings, Millie Søndergaard Jensen, and Arisa Shollo. 2022. Explainable AI, But Explainable to Whom? An Exploratory Case Study of xAI in Healthcare. In *Handbook of Artificial Intelligence in Healthcare: Vol 2: Practicalities and Prospects*, Chee-Peng Lim, Yen-Wei Chen, Ashlesha Vaidya, Charu Mahorkar, and Lakhmi C. Jain (Eds.). Springer International Publishing, Cham, 169–198. https://doi.org/10.1007/978-3-030-83620-7_7

[77] Bhavya Ghai. 2023. Towards Fair and Explainable AI Using a Human-Centered AI Approach. https://doi.org/10.48550/arXiv.2306.07427 arXiv:2306.07427 [cs]

[78] Bhavya Ghai and Klaus Mueller. 2022. D-BIAS: A Causality-Based Human-in-

the-Loop System for Tackling Algorithmic Bias. https://doi.org/10.48550/arXiv.2208.05126 arXiv:2208.05126 [cs, stat]

[79] Shreya Ghosh, Abhinav Dhall, Garima Sharma, Sarthak Gupta, and Nicu Sebe. 2021. Speak2Label: Using Domain Knowledge for Creating a Large Scale Driver Gaze Zone Estimation Dataset. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE, Montreal, BC, Canada, 2896–2905. https://doi.org/10.1109/ICCVW54120.2021.00324

[80] Michele Ginesi, Daniele Meli, Andrea Roberti, Nicola Sansonetto, and Paolo Fiorini. 2020. Autonomous Task Planning and Situation Awareness in Robotic Surgery. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 3144–3150. https://doi.org/10.1109/IROS45743.2020.9341382 arXiv:2004.08911 [cs]

[81] Deepak Gopinath, Siddarth Jain, and Brenna D. Argall. 2017. Human-in-the-Loop Optimization of Shared Autonomy in Assistive Robotics. *IEEE Robotics and Automation Letters* 2, 1 (2017), 247–254. https://doi.org/10.1109/LRA.2016.2593928

[82] Richard Grace and Sonya Steward. 2001. Drowsy Driver Monitor and Warning System. In *Driving Assessment Conference*, Vol. 1. University of Iowa.

[83] Milan Groshev, Carlos Guimarães, Jorge Martín-Pérez, and Antonio de la Oliva. 2021. Toward Intelligent Cyber-Physical Systems: Digital Twin Meets Artificial Intelligence. *IEEE Communications Magazine* 59, 8 (2021), 14–20. https://doi.org/10.1109/MCOM.001.2001237

[84] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5 (2019), 1–42. https://doi.org/10.1145/3236009

[85] Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie. 2011. Emotion Representation, Analysis and Synthesis in Continuous Space: A Survey. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. 827–834. https://doi.org/10.1109/FG.2011.5771357

[86] David Gunning and David Aha. 2019. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine* 40, 2 (2019), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

[87] David Gunning, Eric Vorm, Yunyan Wang, and Matt Turek. 2021. DARPA's Explainable AI (XAI) Program: A Retrospective. *Authorea Preprints* (2021). https://doi.org/10.22541/au.163699841.19031727/v1

[88] Hyeongrok Han, Siwon Kim, Hyun-Soo Choi, and Sungroh Yoon. 2023. On the Impact of Knowledge Distillation for Model Interpretability. arXiv:2305.15734 [cs]

[89] Sicong Han, Chenhao Lin, Chao Shen, Qian Wang, and Xiaohong Guan. 2023. Interpreting Adversarial Examples in Deep Learning: A Review. *Comput. Surveys* 55, 14s (2023), 1–38. https://doi.org/10.1145/3594869

[90] Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. 2022. Which Explanation Should i Choose? A Function Approximation Perspective to Characterizing Post Hoc Explanations. *Advances in neural information processing systems* 35 (2022), 5256–5268.

[91] Jacob Haspiel, Na Du, Jill Meyerson, Lionel P. Robert Jr., Dawn Tilbury, X. Jessie Yang, and Anuj K. Pradhan. 2018. Explanations and Expectations: Trust Building in Automated Vehicles. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Chicago IL USA, 119–120. https://doi.org/10.1145/3173386.3177057

[92] Helen Hastie, Francisco J. Chiyah Garcia, David A. Robb, Atanas Laskov, and Pedro Patron. 2018. MIRIAM: A Multimodal Interface for Explaining the Reasoning

Behind Actions of Remote Autonomous Systems. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM, Boulder CO USA, 557–558. https://doi.org/10.1145/3242969.3266297

[93] Trevor Hastie and R. Tibshirani. 2014. Generalized Additive Models. In *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat03141

[94] Center for Devices and Radiological Health. Fri, 10/20/2023 - 12:46. Artificial Intelligence and Machine Learning in Software as a Medical Device. *FDA* (Fri, 10/20/2023 - 12:46).

[95] Cristy Ho, Nick Reed, and Charles Spence. 2007. Multisensory In-Car Warning Signals for Collision Avoidance. *Human Factors* 49, 6 (2007), 1107–1114. https://doi.org/10.1518/001872007X249965

[96] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. *arXiv preprint arXiv:1812.04608* (2018). arXiv:1812.04608

[97] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for Explainable AI: Explanation Goodness, User Satisfaction, Mental Models, Curiosity, Trust, and Human-AI Performance. *Frontiers in Computer Science* 5 (2023), 1096257. https://doi.org/10.3389/fcomp.2023.1096257

[98] Andreas Holzinger, Bernd Malle, Anna Saranti, and Bastian Pfeifer. 2021. Towards Multi-Modal Causability with Graph Neural Networks Enabling Information Fusion for Explainable AI. *Information Fusion* 71 (2021), 28–37. https://doi.org/10.1016/j.inffus.2021.01.008

[99] Andreas Holzinger and Heimo Müller. 2021. Toward Human–AI Interfaces to Support Explainability and Causability in Medical AI. *Computer* 54, 10 (2021), 78–86. https://doi.org/10.1109/MC.2021.3092610

[100] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2023. MetaGPT: Meta Programming for a Multi-Agent Collaborative Framework. arXiv:2308.00352 [cs.AI]

[101] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. https://doi.org/10.1145/3392878 arXiv:2004.11440 [cs]

[102] Zhongxu Hu, Shanhe Lou, Yang Xing, Xiao Wang, Dongpu Cao, and Chen Lv. 2022. Review and Perspectives on Driver Digital Twin and Its Enabling Technologies for Intelligent Vehicles. *IEEE Transactions on Intelligent Vehicles* 7, 3 (2022), 417–440. https://doi.org/10.1109/TIV.2022.3195635

[103] Zhongxu Hu, Chen Lv, Peng Hang, Chao Huang, and Yang Xing. 2022. Data-Driven Estimation of Driver Attention Using Calibration-Free Eye Gaze and Scene Features. *IEEE Transactions on Industrial Electronics* 69, 2 (2022), 1800–1808. https://doi.org/10.1109/TIE.2021.3057033

[104] Wei Huang, Xingyu Zhao, Gaojie Jin, and Xiaowei Huang. 2023. SAFARI: Versatile and Efficient Evaluations for Robustness of Interpretability. https://doi.org/10.48550/arXiv.2208.09418 arXiv:2208.09418 [cs]

[105] Tobias Huber, Maximilian Demmler, Silvan Mertes, Matthew L. Olson, and Elisabeth André. 2023. GANterfactual-RL: Understanding Reinforcement Learning Agents' Strategies through Visual Counterfactual Explanations. https://doi.org/10.48550/arXiv.2302.12689 arXiv:2302.12689 [cs]

[106] Rami Ibrahim and M. Omair Shafiq. 2023. Explainable Convolutional Neural Networks: A Taxonomy, Review, and Future Directions. *Comput. Surveys* 55, 10 (2023), 1–37. https://doi.org/10.1145/3563691

[107] Maximilian Idahl, Lijun Lyu, Ujwal Gadiraju, and Avishek Anand. 2021. Towards Benchmarking the Utility of Explanations for Model Debugging. arXiv:2105.04505 [cs]

[108] Alan Inglis, Andrew Parnell, and Catherine Hurley. 2021. Visualizing Variable Importance and Variable Interaction Effects in Machine Learning Models. arXiv:2108.04310 [stat]

[109] Aya Abdelsalam Ismail, Héctor Corrada Bravo, and Soheil Feizi. 2021. Improving Deep Learning Interpretability by Saliency Guided Training. https://doi.org/10.48550/arXiv.2111.14338 arXiv:2111.14338 [cs]

[110] Makoto Itoh and Toshiyuki Inagaki. 2014. Design and Evaluation of Steering Protection for Avoiding Collisions during a Lane Change. *Ergonomics* 57, 3 (2014), 361–373. https://doi.org/10.1080/00140139.2013.848474

[111] Abdul Rehman Javed, Waqas Ahmed, Sharnil Pandya, Praveen Kumar Reddy Maddikunta, Mamoun Alazab, and Thippa Reddy Gadekallu. 2023. A Survey of Explainable Artificial Intelligence for Smart Cities. *Electronics* 12, 4 (2023), 1020. https://doi.org/10.3390/electronics12041020

[112] Myounghoon Jeon. 2023. The Effects of Emotions on Trust in Human-Computer Interaction: A Survey and Prospect. *International Journal of Human–Computer Interaction* 0, 0 (2023), 1–19. https://doi.org/10.1080/10447318.2023.2261727

[113] Sumit Jha and Carlos Busso. 2023. Estimation of Driver's Gaze Region from Head Position and Orientation Using Probabilistic Confidence Regions. *IEEE Transactions on Intelligent Vehicles* 8, 1 (2023), 59–72. https://doi.org/10.1109/TIV.2022.3141071

[114] Ming Jin, Qingsong Wen, Yuxuan Liang, Chaoli Zhang, Siqiao Xue, Xue Wang, James Zhang, Yi Wang, Haifeng Chen, Xiaoli Li, Shirui Pan, Vincent S. Tseng,

Yu Zheng, Lei Chen, and Hui Xiong. 2023. Large Models for Time Series and Spatio-Temporal Data: A Survey and Outlook. https://doi.org/10.48550/arXiv.2310.10196 arXiv:2310.10196 [cs]

[115] Weina Jin, Jianyu Fan, Diane Gromala, Philippe Pasquier, and Ghassan Hamarneh. 2023. Invisible Users: Uncovering End-Users' Requirements for Explainable AI via Explanation Forms and Goals. https://doi.org/10.48550/arXiv.2302.06609 arXiv:2302.06609 [cs]

[116] Chunhui Jing, Chenguang Shang, Dongyu Yu, Yaodong Chen, and Jinyi Zhi. 2022. The Impact of Different AR-HUD Virtual Warning Interfaces on the Takeover Performance and Visual Characteristics of Autonomous Vehicles. *Traffic Injury Prevention* 23, 5 (2022), 277–282. https://doi.org/10.1080/15389588.2022.2055752

[117] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *Comput. Surveys* 55, 5 (2022), 95:1–95:29. https://doi.org/10.1145/3527848

[118] Alexey Kashevnik, Roman Shchedrin, Christian Kaiser, and Alexander Stocker. 2021. Driver Distraction Detection Methods: A Literature Review and Framework. *IEEE access : practical innovations, open solutions* 9 (2021), 60063–60076. https://doi.org/10.1109/ACCESS.2021.3073599

[119] Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. 2022. Trustworthy Artificial Intelligence: A Review. *ACM Computing Surveys (CSUR)* 55, 2 (2022), 1–38.

[120] Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. 2021. If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the

Evaluation of Counterfactual XAI Techniques. https://doi.org/10.48550/arXiv.2103.01035 arXiv:2103.01035 [cs]

[121] Eoin M. Kenny, Eoin D. Delaney, Derek Greene, and Mark T. Keane. 2021. Post-Hoc Explanation Options for XAI in Deep Learning: The Insight Centre for Data Analytics Perspective. In *Pattern Recognition. ICPR International Workshops and Challenges*, Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani (Eds.). Springer International Publishing, Cham, 20–34. https://doi.org/10.1007/978-3-030-68796-0_2

[122] Eoin M Kenny, Mycal Tucker, and Julie Shah. 2022. Towards interpretable deep reinforcement learning with human-friendly prototypes. In *The Eleventh International Conference on Learning Representations*.

[123] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability. In *Advances in Neural Information Processing Systems*, Vol. 29. Curran Associates, Inc.

[124] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). https://doi.org/10.48550/arXiv.1711.11279 arXiv:1711.11279 [stat]

[125] Gwangbin Kim, Dohyeon Yeo, Taewoo Jo, Daniela Rus, and SeungJun Kim. 2023. What and When to Explain? On-road Evaluation of Explanations in Highly Automated Vehicles. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 104:1–104:26. https://doi.org/10.1145/3610886

[126] Minjung Kim, Saebyeol Kim, Jinwoo Kim, Tae-Jin Song, and Yuyoung Kim. 2024. Do Stakeholder Needs Differ? - Designing Stakeholder-Tailored Explainable Artifi-

cial Intelligence (XAI) Interfaces. *International Journal of Human-Computer Studies* 181 (2024), 103160. https://doi.org/10.1016/j.ijhcs.2023.103160

[127] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. https://doi.org/10.1145/3544548.3581001 arXiv:2210.03735 [cs]

[128] G. Klein, D.D. Woods, J.M. Bradshaw, R.R. Hoffman, and P.J. Feltovich. 2004. Ten Challenges for Making Automation a "Team Player" in Joint Human-Agent Activity. *IEEE Intelligent Systems* 19, 6 (2004), 91–95. https://doi.org/10.1109/MIS.2004.74

[129] Kazuma Kobayashi, Bader Almutairi, Md Nazmus Sakib, Souvik Chakraborty, and Syed B. Alam. 2023. Explainable, Interpretable & Trustworthy AI for Intelligent Digital Twin: Case Study on Remaining Useful Life. https://doi.org/10.48550/arXiv.2301.06676 arXiv:2301.06676 [cs, stat]

[130] Ramaravind Kommiya Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. 2021. Towards Unifying Feature Attribution and Counterfactual Explanations: Different Means to the Same End. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Virtual Event USA, 652–663. https://doi.org/10.1145/3461702.3462597

[131] Theocharis Kravaris, Konstantinos Lentzos, Georgios Santipantakis, George A Vouros, Gennady Andrienko, Natalia Andrienko, Ian Crook, Jose Manuel Cordero Garcia, and Enrique Iglesias Martinez. 2023. Explaining deep reinforcement learning decisions in complex multiagent settings: towards enabling automation in air traffic flow management. *Applied Intelligence* 53, 4 (2023), 4063–4098.

[132] Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. 2022. The Disagreement Problem in Explainable

Machine Learning: A Practitioner's Perspective. https://doi.org/10.48550/arXiv.2202.01602 arXiv:2202.01602 [cs]

[133] Nicholas Kroeger, Dan Ley, Satyapriya Krishna, Chirag Agarwal, and Himabindu Lakkaraju. 2023. Are Large Language Models Post Hoc Explainers? arXiv:2310.05797 [cs]

[134] M. Kyriakidis, J. C. F. de Winter, N. Stanton, T. Bellet, B. van Arem, K. Brookhuis, M. H. Martens, K. Bengler, J. Andersson, N. Merat, N. Reed, M. Flament, M. Hagenzieker, and R. Happee. 2019. A Human Factors Perspective on Automated Driving. *Theoretical Issues in Ergonomics Science* 20, 3 (2019), 223–249. https://doi.org/10.1080/1463922X.2017.1293187

[135] Vivian Lai, Zheng Cai, and Chenhao Tan. 2019. Many Faces of Feature Importance: Comparing Built-in and Post-hoc Feature Importance in Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 486–495. https://doi.org/10.18653/v1/D19-1046

[136] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What Do We Want from Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research. *Artificial Intelligence* 296 (2021), 103473. https://doi.org/10.1016/j.artint.2021.103473

[137] Michael T. Lash. 2022. HEX: Human-in-the-loop Explainability via Deep Reinforcement Learning. arXiv:2206.01343 [cs, stat]

[138] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[139] Sangwon Lee, Jeonguk Hong, Gyewon Jeon, Jeongmin Jo, Sanghyeok Boo, Hwiseong Kim, Seoyoon Jung, Jieun Park, Inheon Choi, and Sangyeon Kim. 2023. Investigating Effects of Multimodal Explanations Using Multiple In-vehicle Displays for Takeover Request in Conditionally Automated Driving. *Transportation Research Part F: Traffic Psychology and Behaviour* 96 (2023), 1–22. https://doi.org/10.1016/j.trf.2023.05.014

[140] Karin Leichtenstern, Elisabeth André, and Ekatarina Kurdyukova. 2010. Managing User Trust for Self-Adaptive Ubiquitous Computing Systems. In *Proceedings of the 8th International Conference on Advances in Mobile Computing and Multimedia (MoMM '10)*. Association for Computing Machinery, New York, NY, USA, 409–414. https://doi.org/10.1145/1971519.1971589

[141] Jennifer S. Lerner, Ye Li, Piercarlo Valdesolo, and Karim S. Kassam. 2015. Emotion and Decision Making. *Annual Review of Psychology* 66, 1 (2015), 799–823. https://doi.org/10.1146/annurev-psych-010213-115043

[142] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2023. Trustworthy AI: From Principles to Practices. *Comput. Surveys* 55, 9 (2023), 177:1–177:46. https://doi.org/10.1145/3555803

[143] Fan Li, Chun-Hsien Chen, Gangyan Xu, Li Pheng Khoo, and Yisi Liu. 2019. Proactive Mental Fatigue Detection of Traffic Control Operators Using Bagged Trees and Gaze-Bin Analysis. *Advanced Engineering Informatics* 42 (2019), 100987. https://doi.org/10.1016/j.aei.2019.100987

[144] Mengyao Li, Isabel M Erickson, Ernest V Cross, and John D Lee. 2022. Estimating Trust in Conversational Agent with Lexical and Acoustic Features. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 66. SAGE Publications Sage CA: Los Angeles, CA, 544–548.

[145] Xiao Li and Thenkurussi Kesavadas. 2018. Surgical Robot with Environment

Reconstruction and Force Feedback. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 1861–1866. https://doi.org/10.1109/EMBC.2018.8512695

[146] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–15. https://doi.org/10.1145/3313831.3376590

[147] Isador H. Lieberman, Stanley Kisinde, and Shea Hesselbacher. 2020. Robotic-Assisted Pedicle Screw Placement During Spine Surgery. *JBJS Essential Surgical Techniques* 10, 2 (2020), e0020. https://doi.org/10.2106/JBJS.ST.19.00020

[148] Wei-Yin Loh. 2011. Classification and Regression Trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery* 1, 1 (2011), 14–23.

[149] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence* 2, 1 (2020), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[150] Yunan Luo, Jian Peng, and Jianzhu Ma. 2020. When Causal Inference Meets Deep Learning. *Nature Machine Intelligence* 2, 8 (2020), 426–427. https://doi.org/10.1038/s42256-020-0218-x

[151] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. arXiv:1508.04025 [cs]

[152] Joseph B Lyons, Nhut T Ho, Anna Lee Van Abel, Lauren C Hoffmann, W Eric Fergueson, Garrett G Sadler, Michelle A Grigsby, and Amy C Burns. 2018. Exploring trust barriers to future autonomy: a qualitative look. In *Advances in Human*

Factors in Simulation and modelling: Proceedings of the AHFE 2017 International Conference on Human Factors in Simulation and modelling, July 17–21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8. Springer, 3–11.

[153] Joseph B. Lyons, Katia Sycara, Michael Lewis, and August Capiola. 2021. Human–Autonomy Teaming: Definitions, Debates, and Directions. *Frontiers in Psychology* 12 (2021).

[154] Hongbin Ma, Shuyuan Yang, Dongzhu Feng, Licheng Jiao, and Luping Zhang. 2021. Progressive Mimic Learning: A New Perspective to Train Lightweight CNN Models. *Neurocomputing* 456 (2021), 220–231. `https://doi.org/10.1016/j.neucom.2021.04.086`

[155] A. V. Shreyas Madhav and Amit Kumar Tyagi. 2023. Explainable Artificial Intelligence (XAI): Connecting Artificial Decision-Making and Human Trust in Autonomous Vehicles. In *Proceedings of Third International Conference on Computing, Communications, and Cyber-Security (Lecture Notes in Networks and Systems)*, Pradeep Kumar Singh, Sławomir T. Wierzchoń, Sudeep Tanwar, Joel J. P. C. Rodrigues, and Maria Ganzha (Eds.). Springer Nature, Singapore, 123–136. `https://doi.org/10.1007/978-981-19-1142-2_10`

[156] Udara E. Manawadu, Mitsuhiro Kamezaki, Masaaki Ishikawa, Takahiro Kawano, and Shigeki Sugano. 2016. A Haptic Feedback Driver-Vehicle Interface for Controlling Lateral and Longitudinal Motions of Autonomous Vehicles. In *2016 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*. 119–124. `https://doi.org/10.1109/AIM.2016.7576753`

[157] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. arXiv:1904.12584 [cs]

[158] Aniek F. Markus, Jan A. Kors, and Peter R. Rijnbeek. 2021. The Role of Ex-

plainability in Creating Trustworthy Artificial Intelligence for Health Care: A Comprehensive Survey of the Terminology, Design Choices, and Evaluation Strategies. *Journal of Biomedical Informatics* 113 (2021), 103655. https://doi.org/10.1016/j.jbi.2020.103655

[159] David Alvarez Melis and Tommi Jaakkola. 2018. Towards Robust Interpretability with Self-Explaining Neural Networks. In *Advances in Neural Information Processing Systems*, Vol. 31. Curran Associates, Inc.

[160] Fanxing Meng and Charles Spence. 2015. Tactile Warning Signals for In-Vehicle Systems. *Accident; Analysis and Prevention* 75 (2015), 333–346. https://doi.org/10.1016/j.aap.2014.12.013

[161] Christian Meske and Enrico Bunde. 2022. Design Principles for User Interfaces in AI-Based Decision Support Systems: The Case of Explainable Hate Speech Detection. *Information Systems Frontiers* (2022). https://doi.org/10.1007/s10796-021-10234-5

[162] Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. 2022. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Computing Surveys (CSUR)* 54, 10s (2022), 1–40.

[163] Leila Methnani, Manolis Chiou, Virginia Dignum, and Andreas Theodorou. 2024. Who's in Charge Here? A Survey on Trustworthy AI in Variable Autonomy Robotic Systems. *Comput. Surveys* (2024), 32. https://doi.org/10.1145/3645090

[164] John A. Michon. 1985. A Critical View of Driver behaviourModels: What Do We Know, What Should We Do? In *Human Behavior and Traffic Safety*, Leonard Evans and Richard C. Schwing (Eds.). Springer US, Boston, MA, 485–524. https://doi.org/10.1007/978-1-4613-2173-6_19

[165] Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. 2023. Explainable Reinforcement Learning: A Survey and Comparative Review. *Comput. Surveys* (2023), 1–35. https://doi.org/10.1145/3616864

[166] Dang Minh, H. Xiang Wang, Y. Fen Li, and Tan N. Nguyen. 2022. Explainable Artificial Intelligence: A Comprehensive Review. *Artificial Intelligence Review* 55, 5 (2022), 3503–3568. https://doi.org/10.1007/s10462-021-10088-y

[167] Abhijit Mishra, Srikanth Tamilselvam, Riddhiman Dasgupta, Seema Nagar, and Kuntal Dey. 2018. Cognition-Cognizant Sentiment Analysis With Multitask Subjectivity Summarization Based on Annotators' Gaze Behavior. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (2018). https://doi.org/10.1609/aaai.v32i1.12068

[168] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems* 11, 3–4 (2021), 1–45. https://doi.org/10.1145/3387166

[169] Chris D. Monaco and Sean N. Brennan. 2020. RADARODO: Ego-Motion Estimation From Doppler and Spatial Data in RADAR Images. *IEEE Transactions on Intelligent Vehicles* 5, 3 (2020), 475–484. https://doi.org/10.1109/TIV.2020.2973536

[170] Arthur G. Money and Harry Agius. 2009. Analysing User Physiological Responses for Affective Video Summarisation. *Displays* 30, 2 (2009), 59–70. https://doi.org/10.1016/j.displa.2008.12.003

[171] Dan Mossing, Steven Bills, Henk Tillman, Tom Dupré la Tour, Nick Cammarata, Leo Gao, Joshua Achiam, Catherine Yeh, Jan Leike, Jeff Wu, and William Saunders. 2024. Transformer Debugger.

[172] Richard Mulgan. 2000. 'Accountability': An Ever-Expanding Concept? *Public Administration* 78, 3 (2000), 555–573. https://doi.org/10.1111/1467-9299.00218

[173] Clare Mutzenich, Szonya Durant, Shaun Helman, and Polly Dalton. 2021. Updating Our Understanding of Situation Awareness in Relation to Remote Operators of Autonomous Vehicles. *Cognitive Research: Principles and Implications* 6, 1 (2021), 9. https://doi.org/10.1186/s41235-021-00271-8

[174] Frederik Naujoks, Yannick Forster, Katharina Wiedemann, and Alexandra Neukum. 2017. Improving Usefulness of Automated Driving by Lowering Primary Task Interference through HMI Design. *Journal of Advanced Transportation* 2017 (2017), e6105087. https://doi.org/10.1155/2017/6105087

[175] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *Comput. Surveys* 55, 13s (2023), 1–42. https://doi.org/10.1145/3583558

[176] Mark A. Neerincx, Jasper van der Waa, Frank Kaptein, and Jurriaan van Diggelen. 2018. Using Perceptual and Cognitive Explanations for Enhanced Human-Agent Team Performance. In *Engineering Psychology and Cognitive Ergonomics (Lecture Notes in Computer Science)*, Don Harris (Ed.). Springer International Publishing, Cham, 204–214. https://doi.org/10.1007/978-3-319-91122-9_18

[177] Catherine Neubauer. 2023. HAT3: The Human Autonomy Team Trust Toolkit. In *Companion Publication of the 25th International Conference on Multimodal Interaction* (¡conf-loc¿, ¡city¿Paris¡/city¿, ¡country¿France¡/country¿, ¡/conf-loc¿) *(ICMI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 115–118. https://doi.org/10.1145/3610661.3620660

[178] Subash Neupane, Ivan A. Fernandez, Wilson Patterson, Sudip Mittal, Milan Parmar, and Shahram Rahimi. 2023. TwinExplainer: Explaining Predictions of an Automotive Digital Twin. arXiv:2302.00152 [cs]

[179] Kazuo Okamura and Seiji Yamada. 2020. Adaptive Trust Calibration for Human-AI Collaboration. *Plos one* 15, 2 (2020), e0229132.

[180] Luis Oliveira, Jacob Luton, Sumeet Iyer, Chris Burns, Alexandros Mouzakitis, Paul Jennings, and Stewart Birrell. 2018. Evaluating How Interfaces Influence the User Interaction with Fully Autonomous Vehicles. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM Association for Computing Machinery. `https://doi.org/10.1145/3239060.3239065`

[181] Daniel Omeiza, Helena Webb, Marina Jirotka, and Lars Kunze. 2022. Explanations in Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems* 23, 8 (2022), 10142–10162. `https://doi.org/10.1109/tits.2021.3122865`

[182] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. 2022. Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors* 64, 5 (2022), 904–938. `https://doi.org/10.1177/0018720820960865`

[183] Shane O'Sullivan, Nathalie Nevejans, Colin Allen, Andrew Blyth, Simon Leonard, Ugo Pagallo, Katharina Holzinger, Andreas Holzinger, Mohammed Imran Sajid, and Hutan Ashrafian. 2019. Legal, Regulatory, and Ethical Frameworks for Development of Standards in Artificial Intelligence (AI) and Autonomous Robotic Surgery. *The international journal of medical robotics + computer assisted surgery: MRCAS* 15, 1 (2019), e1968. `https://doi.org/10.1002/rcs.1968`

[184] Sharon Oviatt and Philip Cohen. 2015. *The Paradigm Shift to Multimodality in*

Contemporary Computer Interfaces. Morgan & Claypool Publishers. https://doi.org/10.2200/S00636ED1V01Y201503HCI030

[185] April Rose Panganiban, Gerald Matthews, and Michael D. Long. 2020. Transparency in Autonomous Teammates: Intention to Support as Teaming Information. *Journal of Cognitive Engineering and Decision Making* 14, 2 (2020), 174–190. https://doi.org/10.1177/1555343419881563

[186] Cecilia Panigutti, Andrea Beretta, Daniele Fadda, Fosca Giannotti, Dino Pedreschi, Alan Perotti, and Salvatore Rinzivillo. 2023. Co-Design of Human-Centered, Explainable AI for Clinical Decision Support. *ACM Transactions on Interactive Intelligent Systems* (2023), 3587271. https://doi.org/10.1145/3587271

[187] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the Impact of Explanations on Advice-Taking: A User Study for AI-based Clinical Decision Support Systems. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–9. https://doi.org/10.1145/3491102.3502104

[188] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, 8779–8788. https://doi.org/10.1109/CVPR.2018.00915

[189] Heejong Park, Arvind Easwaran, and Sidharta Andalam. 2019. Challenges in Digital Twin Development for Cyber-Physical Production Systems. , 28–48 pages. https://doi.org/10.1007/978-3-030-23703-5_2 arXiv:2102.03341 [cs]

[190] Seungeun Park, Betty Bekemeier, Abraham Flaxman, and Melinda Schultz. 2022. Impact of Data Visualization on Decision-Making and Its Implications for Public

Health Practice: A Systematic Literature Review. *Informatics for Health & Social Care* 47, 2 (2022), 175–193. https://doi.org/10.1080/17538157.2021.1982949

[191] Prathyush S. Parvatharaju, Ramesh Doddaiah, Thomas Hartvigsen, and Elke A. Rundensteiner. 2021. Learning Saliency Maps to Explain Deep Time Series Classifiers. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 1406–1415. https://doi.org/10.1145/3459637.3482446

[192] Biswaksen Patnaik, Andrea Batch, and Niklas Elmqvist. 2019. Information Olfactation: Harnessing Scent to Convey Data. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 726–736. https://doi.org/10.1109/TVCG.2018.2865237

[193] Cristiano Patrício, João C. Neves, and Luís F. Teixeira. 2023. Explainable Deep Learning Methods in Medical Image Classification: A Survey. *Comput. Surveys* 56, 4 (2023), 1–41. https://doi.org/10.1145/3625287

[194] Judea Pearl. 2009. Causal Inference in Statistics: An Overview. *Statistics Surveys* 3 (2009). https://doi.org/10.1214/09-SS057

[195] Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic books.

[196] Filippo Pesapane, Caterina Volonté, Marina Codari, and Francesco Sardanelli. 2018. Artificial Intelligence as a Medical Device in Radiology: Ethical and Regulatory Issues in Europe and the United States. *Insights into Imaging* 9, 5 (2018), 745–753. https://doi.org/10.1007/s13244-018-0645-y

[197] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. arXiv:1806.07421 [cs]

[198] Robin L. Pierce, Wim Van Biesen, Daan Van Cauwenberge, Johan Decruyenaere, and Sigrid Sterckx. 2022. Explainability in Medicine in an Era of AI-based Clinical Decision Support Systems. *Frontiers in Genetics* 13 (2022), 903600. https://doi.org/10.3389/fgene.2022.903600

[199] Robert Plutchik. 2003. *Emotions and Life: Perspectives from Psychology, Biology, and Evolution*. American Psychological Association, Washington, DC, US. xix, 381 pages.

[200] Mario Alfonso Prado-Romero, Bardh Prenkaj, Giovanni Stilo, and Fosca Giannotti. 2023. A Survey on Graph Counterfactual Explanations: Definitions, Methods, Evaluation, and Research Challenges. *Comput. Surveys* 56, 7 (2023), 1–35. https://doi.org/10.1145/3618105

[201] Mattia Prosperi, Yi Guo, Matt Sperrin, James S. Koopman, Jae S. Min, Xing He, Shannan Rich, Mo Wang, Iain E. Buchan, and Jiang Bian. 2020. Causal Inference and Counterfactual Prediction in Machine Learning for Actionable Healthcare. *Nature Machine Intelligence* 2, 7 (2020), 369–375. https://doi.org/10.1038/s42256-020-0197-y

[202] Erasmo Purificato, Cataldo Musto, Pasquale Lops, and Ernesto William De Luca. 2022. First Workshop on Adaptive and Personalized Explainable User Interfaces (Apex-Ui 2022). In *27th International Conference on Intelligent User Interfaces*. 1–3.

[203] Antonio Rago, Hengzhi Li, and Francesca Toni. 2023. Interactive Explanations by Conflict Resolution via Argumentative Exchanges. https://doi.org/10.48550/arXiv.2303.15022 arXiv:2303.15022 [cs]

[204] Akilesh Rajavenkatanarayanan, Varun Kanal, Konstantinos Tsiakas, James Brady, Diane Calderon, Glenn Wylie, and Fillia Makedon. 2019. Towards a Robot-Based Multimodal Framework to Assess the Impact of Fatigue on User beha-

viourand Performance: A Pilot Study. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '19)*. Association for Computing Machinery, New York, NY, USA, 493–498. https://doi.org/10.1145/3316782.3322776

[205] Sriram Ramgopal, L. Nelson Sanchez-Pinto, Christopher M. Horvat, Michael S. Carroll, Yuan Luo, and Todd A. Florin. 2023. Artificial Intelligence-Based Clinical Decision Support in Pediatrics. *Pediatric Research* 93, 2 (2023), 334–341. https://doi.org/10.1038/s41390-022-02226-1

[206] Vikas C. Raykar, Arindam Jati, Sumanta Mukherjee, Nupur Aggarwal, Kanthi Sarpatwar, Giridhar Ganapavarapu, and Roman Vaculin. 2023. TsSHAP: Robust Model Agnostic Feature-Based Explainability for Time Series Forecasting. https://doi.org/10.48550/ARXIV.2303.12316

[207] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. https://doi.org/10.48550/arXiv.1602.04938 arXiv:1602.04938 [cs, stat]

[208] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (2018). https://doi.org/10.1609/aaai.v32i1.11491

[209] Rafael F. Ribeiro and Paula D. P. Costa. 2019. Driver Gaze Zone Dataset With Depth Data. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. 1–5. https://doi.org/10.1109/FG.2019.8756592

[210] David A. Robb, Francisco J. Chiyah Garcia, Atanas Laskov, Xingkun Liu, Pedro Patron, and Helen Hastie. 2018. Keep Me in the Loop: Increasing Operator Situation Awareness through a Conversational Multimodal Interface. In *Proceedings of the*

20th ACM International Conference on Multimodal Interaction. ACM, Boulder CO USA, 384–392. https://doi.org/10.1145/3242969.3242974

[211] Avi Rosenfeld. 2021. Better Metrics for Evaluating Explainable Artificial Intelligence. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*. 45–50.

[212] Avi Rosenfeld and Ariella Richardson. 2019. Explainability in Human–Agent Systems. *Autonomous Agents and Multi-Agent Systems* 33, 6 (2019), 673–705. https://doi.org/10.1007/s10458-019-09408-y

[213] Jože M. Rožanec, Blaž Fortuna, and Dunja Mladenić. 2022. Knowledge Graph-Based Rich and Confidentiality Preserving Explainable Artificial Intelligence (XAI). *Information Fusion* 81 (2022), 91–102. https://doi.org/10.1016/j.inffus.2021.11.015

[214] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[215] Fatai Sado, Chu Kiong Loo, Wei Shiung Liew, Matthias Kerzel, and Stefan Wermter. 2023. Explainable Goal-driven Agents and Robots - A Comprehensive Review. *Comput. Surveys* 55, 10 (2023), 211:1–211:41. https://doi.org/10.1145/3564240

[216] Paul Salmon, Neville Stanton, Guy Walker, and Damian Green. 2006. Situation Awareness Measurement: A Review of Applicability for C4i Environments. *Applied Ergonomics* 37, 2 (2006), 225–238. https://doi.org/10.1016/j.apergo.2005.02.001

[217] Kevin Joel Salubre and Dan Nathan-Roberts. 2021. Takeover Request Design in Automated Driving: A Systematic Review. *Proceedings of the Human Factors and*

Ergonomics Society Annual Meeting 65, 1 (2021), 868–872. https://doi.org/10.1177/1071181321651296

[218] Lindsay Sanneman. 2023. *Transparent Value Alignment: Foundations for Human-Centered Explainable AI in Alignment*. Thesis. Massachusetts Institute of Technology.

[219] Lindsay Sanneman and Julie A. Shah. 2022. The Situation Awareness Framework for Explainable AI (SAFE-AI) and Human Factors Considerations for XAI Systems. *International Journal of Human–Computer Interaction* 38, 18-20 (2022), 1772–1788. https://doi.org/10.1080/10447318.2022.2081282

[220] Beau G. Schelble, Christopher Flathmann, and Nathan McNeese. 2020. Towards Meaningfully Integrating Human-Autonomy Teaming in Applied Settings. In *Proceedings of the 8th International Conference on Human-Agent Interaction*. ACM, Virtual Event USA, 149–156. https://doi.org/10.1145/3406499.3415077

[221] Tobias Schneider, Sabiha Ghellal, Steve Love, and Ansgar R.S. Gerlicher. 2021. Increasing the User Experience in Autonomous Driving through Different Feedback Modalities. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 7–10. https://doi.org/10.1145/3397481.3450687

[222] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Towards Causal Representation Learning. https://arxiv.org/abs/2102.11107v1.

[223] Jan Maarten Schraagen, Pia Elsasser, Hanna Fricke, Marleen Hof, and Fabyen Ragalmuto. 2020. Trusting the X in XAI: Effects of Different Types of Explanations by a Self-Driving Car on Trust, Explanation Satisfaction and Mental Models. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 64, 1 (2020), 339–343. https://doi.org/10.1177/1071181320641077

[224] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 618–626. https://doi.org/10.1109/ICCV.2017.74

[225] Azra Seyyedi, Mahdi Bohlouli, and Seyedehsan Nedaaee Oskoee. 2023. Machine Learning and Physics: A Survey of Integrated Models. *Comput. Surveys* 56, 5 (2023), 115:1–115:33. https://doi.org/10.1145/3611383

[226] Amit Sheth, Manas Gaur, Kaushik Roy, and Keyur Faldu. 2021. Knowledge-Intensive Language Understanding for Explainable AI. https://doi.org/10.48550/arXiv.2108.01174 arXiv:2108.01174 [cs]

[227] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2019. Learning Important Features Through Propagating Activation Differences. arXiv:1704.02685 [cs]

[228] Bibhudhendu Shukla, Ip-Shing Fan, and Ian Jennions. 2020. Opportunities for explainable artificial intelligence in aerospace predictive maintenance. In *PHM Society European Conference*, Vol. 5. 11–11.

[229] Keng Siau and Weiyu Wang. 2018. Building Trust in Artificial Intelligence, Machine Learning, and Robotics. *Cutter Business Technology Journal* 31 (2018), 47–53.

[230] Anton Simorov, R. Stephen Otte, Courtni M. Kopietz, and Dmitry Oleynikov. 2012. Review of Surgical Robotics User Interface: What Is the Best Way to Control Robotic Surgery? *Surgical Endoscopy* 26, 8 (2012), 2117–2125. https://doi.org/10.1007/s00464-012-2182-y

[231] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*

(AIES '20). Association for Computing Machinery, New York, NY, USA, 180–186. https://doi.org/10.1145/3375627.3375830

[232] Victor Feitosa Souza, Ferdinando Cicalese, Eduardo Sany Laber, and Marco Molinaro. 2022. Decision Trees with Short Explainable Rules. In *Advances in Neural Information Processing Systems.*

[233] Timo Speith. 2022. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 2239–2250. https://doi.org/10.1145/3531146.3534639

[234] Kimberly Stowers, Lisa L. Brady, Christopher MacLellan, Ryan Wohleber, and Eduardo Salas. 2021. Improving Teamwork Competencies in Human-Machine Teams: Perspectives From Team Science. *Frontiers in Psychology* 12 (2021), 590290. https://doi.org/10.3389/fpsyg.2021.590290

[235] Hariharan Subramonyam, Colleen Seifert, and Eytan Adar. 2021. ProtoAI: Model-Informed Prototyping for AI-Powered Interfaces. In *26th International Conference on Intelligent User Interfaces (IUI '21)*. Association for Computing Machinery, New York, NY, USA, 48–58. https://doi.org/10.1145/3397481.3450640

[236] Agus Sudjianto and Aijun Zhang. 2021. Designing Inherently Interpretable Machine Learning Models. https://doi.org/10.48550/arXiv.2111.01743 arXiv:2111.01743 [cs, stat]

[237] Jimin Sun, Wenqi Shi, Felipe O. Giuste, Yog S. Vaghani, Lingzi Tang, and May D. Wang. 2023. Improving Explainable AI with Patch Perturbation-Based Evaluation Pipeline: A COVID-19 X-ray Image Analysis Case Study. *Scientific Reports* 13, 1 (2023), 19488. https://doi.org/10.1038/s41598-023-46493-2

[238] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. arXiv:1703.01365 [cs]

[239] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, Textual or Hybrid: The Effect of User Expertise on Different Explanations. In *26th International Conference on Intelligent User Interfaces (IUI '21)*. Association for Computing Machinery, New York, NY, USA, 109–119. https://doi.org/10.1145/3397481.3450662

[240] Maxwell Szymanski, Vero Vanden Abeele, and Katrien Verbert. 2022. Explaining Health Recommendations to Lay Users: The Dos and Dont's. In *Joint Proceedings of the IUI 2022 Workshops: APEx-UI, HAI-GEN, HEALTHI, HUMANIZE, TExSS, SOCIALIZE Co-Located with the ACM International Conference on Intelligent User Interfaces (IUI 2022)*. CEUR Workshop Proceedings, 1–10.

[241] Karim A. Tahboub. 2006. Intelligent Human-Machine Interaction Based on Dynamic Bayesian Networks Probabilistic Intention Recognition. *Journal of Intelligent and Robotic Systems* 45, 1 (2006), 31–52. https://doi.org/10.1007/s10846-005-9018-0

[242] Yifan Tang and Yan Xu. 2021. Multi-Agent Deep Reinforcement Learning for Solving Large-scale Air Traffic Flow Management Problem: A Time-Step Sequential Decision Approach. In *2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)*. IEEE, San Antonio, TX, USA, 1–10. https://doi.org/10.1109/DASC52595.2021.9594329

[243] Fei Tao, Jiangfeng Cheng, Qinglin Qi, Meng Zhang, He Zhang, and Fangyuan Sui. 2018. Digital Twin-Driven Product Design, Manufacturing and Service with Big Data. *The International Journal of Advanced Manufacturing Technology* 94, 9 (2018), 3563–3576. https://doi.org/10.1007/s00170-017-0233-1

[244] TAPAS. 2022. https://tapas-atm.eu/. Accessed on 2024-03-31.

[245] Sule Tekkesinoglu. 2024. Exploring Evaluation Methodologies for Explainable

AI: Guidelines for Objective and Subjective Assessment. *SSRN Electronic Journal* (2024). https://doi.org/10.2139/ssrn.4667052

[246] Stefano Teso, Öznur Alkan, Wolfgang Stammer, and Elizabeth Daly. 2023. Leveraging Explanations in Interactive Machine Learning: An Overview. *Frontiers in Artificial Intelligence* 6 (2023). https://doi.org/10.3389/frai.2023.1066049

[247] Christoph Torens, Umut Durak, and Johann C. Dauer. 2022. Guidelines and Regulatory Framework for Machine Learning in Aviation. In *AIAA SCITECH 2022 Forum*. American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2022-1132

[248] Cristen Torrey, Aaron Powers, Matthew Marge, Susan R Fussell, and Sara Kiesler. 2006. Effects of Adaptive Robot Dialogue on Information Exchange and Social Relations. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*. 126–133.

[249] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288* (2023). arXiv:2307.09288

[250] Hugues Turbé, Mina Bjelogrlic, Christian Lovis, and Gianmarco Mengaldo. 2023. Evaluation of Post-Hoc Interpretability Methods in Time-Series Classification. *Nature Machine Intelligence* 5, 3 (2023), 250–260. https://doi.org/10.1038/s42256-023-00620-w

[251] D. Umerenkov, G. Zubkova, and A. Nesterov. 2023. Deciphering Diagnoses: How Large Language Models Explanations Influence Clinical Decision Making. arXiv:2310.01708 [cs]

[252] Michael van Lent, William Fisher, and Michael Mancuso. 2004. An Explainable Artificial Intelligence System for Small-unit Tactical Behavior.

[253] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

[254] Ruben S Verhagen, Mark A Neerincx, and Myrthe L Tielman. 2024. Meaningful human control and variable autonomy in human-robot teams for firefighting. *Frontiers in Robotics and AI* 11 (2024), 1323980.

[255] Giulia Vilone and Luca Longo. 2021. Classification of Explainable Artificial Intelligence Methods through Their Output Formats. *Machine Learning and Knowledge Extraction* 3, 3 (2021), 615–661. https://doi.org/10.3390/make3030032

[256] Giulia Vilone and Luca Longo. 2021. Notions of Explainability and Evaluation Approaches for Explainable Artificial Intelligence. *Information Fusion* 76 (2021), 89–106. https://doi.org/10.1016/j.inffus.2021.05.009

[257] Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10, 3152676 (2017), 10–5555.

[258] N. K. Vøllestad. 1997. Measurement of Human Muscle Fatigue. *Journal of Neuroscience Methods* 74, 2 (1997), 219–227. https://doi.org/10.1016/s0165-0270(97)02251-6

[259] George A. Vouros. 2022. Explainable Deep Reinforcement Learning: State of the Art and Challenges. *Comput. Surveys* 55, 5 (2022), 92:1–92:39. https://doi.org/10.1145/3527448

[260] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal* (2017). https://doi.org/10.2139/ssrn.3063289

[261] Kassie Waller, Nellie Hill, Courtney Meyers, Amber McCord, and Courtney D. Gibson. 2020. The Effect of Infographics on Recall of Information about Genetically Modified Foods. *Journal of Agricultural Education* 61, 3 (2020), 22–37. https://doi.org/10.5032/jae.2020.03022

[262] James C. Walliser, Ewart J. de Visser, Eva Wiese, and Tyler H. Shaw. 2019. Team Structure and Team Building Improve Human–Machine Teaming With Autonomous Agents. *Journal of Cognitive Engineering and Decision Making* 13, 4 (2019), 258–278. https://doi.org/10.1177/1555343419867563

[263] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

[264] Liuping Wang, Zhan Zhang, Dakuo Wang, Weidan Cao, Xiaomu Zhou, Ping Zhang, Jianxing Liu, Xiangmin Fan, and Feng Tian. 2023. Human-Centered Design and Evaluation of AI-empowered Clinical Decision Support Systems: A Systematic Review. *Frontiers in Computer Science* 5 (2023).

[265] Qianwen Wang, Zhutian Chen, Yong Wang, and Huamin Qu. 2022. A Survey on ML4VIS: Applying Machine Learning Advances to Data Visualization. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (2022), 5134–5153. https://doi.org/10.1109/TVCG.2021.3106142 arXiv:2012.00467 [cs]

[266] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, and Wenqiang Zhang. 2022. A Systematic Review on Affective Computing: Emotion Models, Databases, and Recent Advances. arXiv:2203.06935 [cs]

[267] Aiden Warren and Alek Hillas. 2020. Friend or Frenemy? The Role of Trust in

Human-Machine Teaming and Lethal Autonomous Weapons Systems. *Small Wars & Insurgencies* 31, 4 (2020), 822–850. https://doi.org/10.1080/09592318.2020.1743485

[268] Leander Weber, Sebastian Lapuschkin, Alexander Binder, and Wojciech Samek. 2023. Beyond Explaining: Opportunities and Challenges of XAI-based Model Improvement. *Information Fusion* 92 (2023), 154–176. https://doi.org/10.1016/j.inffus.2022.11.013

[269] Christoph Wehner, Francis Powlesland, Bashar Altakrouri, and Ute Schmid. 2022. Explainable Online Lane Change Predictions on a Digital Twin with a Layer Normalized LSTM and Layer-wise Relevance Propagation. https://doi.org/10.48550/arXiv.2204.01292 arXiv:2204.01292 [cs]

[270] Alexandra Weidemann and Nele Rußwinkel. 2021. The Role of Frustration in Human–Robot Interaction – What Is Needed for a Successful Collaboration? *Frontiers in Psychology* 12 (2021), 640186. https://doi.org/10.3389/fpsyg.2021.640186

[271] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2021. "Let Me Explain!": Exploring the Potential of Virtual Agents in Explainable AI Interaction Design. *Journal on Multimodal User Interfaces* 15, 2 (2021), 87–98. https://doi.org/10.1007/s12193-020-00332-0

[272] Lindsay Wells and Tomasz Bednarz. 2021. Explainable AI and Reinforcement Learning—A Systematic Review of Current Approaches and Trends. *Frontiers in Artificial Intelligence* 4 (2021). https://doi.org/10.3389/frai.2021.550030

[273] Philipp Wintersberger, Hannah Nicklas, Thomas Martlbauer, Stephan Hammer, and Andreas Riener. 2020. Explainable Automation: Personalized and Adaptive UIs to Foster Trust and Understanding of Driving Automation Systems. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular*

Applications. ACM, Virtual Event DC USA, 252–261. https://doi.org/10.1145/3409120.3410659

[274] Adam Wright, Shobha Phansalkar, Meryl Bloomrosen, Robert A. Jenders, Anne M. Bobb, John D. Halamka, Gilad Kuperman, Thomas H. Payne, S. Teasdale, A. J. Vaida, and D. W. Bates. 2010. Best Practices in Clinical Decision Support. *Applied Clinical Informatics* 1, 3 (2010), 331–345. https://doi.org/10.4338/ACI-2010-05-RA-0031

[275] Julia L. Wright, Jessie Y. C. Chen, and Shan G. Lakhmani. 2020. Agent Transparency and Reliability in Human–Robot Interaction: The Influence on User Confidence and Perceived Reliability. *IEEE Transactions on Human-Machine Systems* 50, 3 (2020), 254–263. https://doi.org/10.1109/THMS.2019.2925717

[276] Chenwang Wu, Xiting Wang, Defu Lian, Xing Xie, and Enhong Chen. 2023. A Causality Inspired Framework for Model Interpretation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*. Association for Computing Machinery, New York, NY, USA, 2731–2741. https://doi.org/10.1145/3580305.3599240

[277] Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, and Ninghao Liu. 2024. Usable XAI: 10 Strategies Towards Exploiting Explainability in the LLM Era. arXiv:2403.08946 [cs]

[278] Zijun Wu, Zi Xuan Zhang, Atharva Naik, Zhijian Mei, Mauajama Firdaus, and Lili Mou. 2023. Weakly Supervised Explainable Phrasal Reasoning with Neural Fuzzy Logic. arXiv:2109.08927 [cs]

[279] Oskar Wysocki, Jessica Katharine Davies, Markel Vigo, Anne Caroline Armstrong, Dónal Landers, Rebecca Lee, and André Freitas. 2023. Assessing the Communication Gap between AI Models and Healthcare Professionals: Explainability, Utility

and Trust in AI-driven Clinical Decision-Making. *Artificial Intelligence* 316 (2023), 103839. https://doi.org/10.1016/j.artint.2022.103839

[280] Ziyu Xiang, Mingzhou Fan, Guillermo Vázquez Tovar, William Trehern, Byung-Jun Yoon, Xiaofeng Qian, Raymundo Arroyave, and Xiaoning Qian. 2021. Physics-Constrained Automatic Feature Engineering for Predictive modelling in Materials Science. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 12 (2021), 10414–10421. https://doi.org/10.1609/aaai.v35i12.17247

[281] Yao Xie, Ge Gao, and Xiang'Anthony' Chen. 2019. Outlining the Design Space of Explainable Intelligent Systems for Medical Diagnosis. *arXiv preprint arXiv:1902.06019* (2019). arXiv:1902.06019

[282] Yang Xing, Zhongxu Hu, Peng Hang, and Chen Lv. 2023. Learning from the Dark Side: A Parallel Time Series Modelling Framework for Forecasting and Fault Detection on Intelligent Vehicles. *IEEE Transactions on Intelligent Vehicles* (2023), 1–15. https://doi.org/10.1109/TIV.2023.3342648

[283] Yang Xing, Zhongxu Hu, Xiaoyu Mo, Peng Hang, Shujing Li, Yahui Liu, Yifan Zhao, and Chen Lv. 2024. Driver Steering Behaviour Modelling Based on Neuromuscular Dynamics and Multi-Task Time-Series Transformer. *Automotive Innovation* 7, 1 (2024), 45–58. https://doi.org/10.1007/s42154-023-00272-x

[284] Yang Xing, Chen Lv, and Dongpu Cao. 2020. *Advanced Driver Intention Inference: Theory and Design*. Elsevier.

[285] Yang Xing, Chen Lv, Dongpu Cao, and Peng Hang. 2021. Toward Human-Vehicle Collaboration: Review and Perspectives on Human-Centered Collaborative Automated Driving. *Transportation Research Part C: Emerging Technologies* 128 (2021), 103199. https://doi.org/10.1016/j.trc.2021.103199

[286] Yu Xiong, Zhipeng Hu, Ye Huang, Runze Wu, Kai Guan, Xingchen Fang, Ji Jiang, Tianze Zhou, Yujing Hu, Haoyu Liu, Tangjie Lyu, and Changjie Fan. 2024. XRL-

Bench: A Benchmark for Evaluating and Comparing Explainable Reinforcement Learning Techniques. arXiv:2402.12685 [cs]

[287] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2016. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. arXiv:1502.03044 [cs]

[288] Shiyun Xu, Zhiqi Bu, Pratik Chaudhari, and Ian J. Barnett. 2022. Sparse Neural Additive Model: Interpretable Deep Learning with Feature Selection via Group Sparsity. https://doi.org/10.48550/arXiv.2202.12482 arXiv:2202.12482 [cs, math, stat]

[289] Wei Xu and Marvin Dainoff. 2023. Enabling Human-Centered AI: A New Junction and Shared Journey between AI and HCI Communities. https://doi.org/10.48550/arXiv.2111.08460 arXiv:2111.08460 [cs]

[290] Xuhai Xu, Anna Yu, Tanya R. Jonker, Kashyap Todi, Feiyu Lu, Xun Qian, João Marcelo Evangelista Belo, Tianyi Wang, Michelle Li, Aran Mun, Te-Yen Wu, Junxiao Shen, Ting Zhang, Narine Kokhlikyan, Fulton Wang, Paul Sorenson, Sophie Kim, and Hrvoje Benko. 2023. XAIR: A Framework of Explainable AI in Augmented Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–30. https://doi.org/10.1145/3544548.3581500

[291] Guang-Zhong Yang, James Cambias, Kevin Cleary, Eric Daimler, James Drake, Pierre E. Dupont, Nobuhiko Hata, Peter Kazanzides, Sylvain Martel, Rajni V. Patel, Veronica J. Santos, and Russell H. Taylor. 2017. Medical Robotics—Regulatory, Ethical, and Legal Considerations for Increasing Levels of Autonomy. *Science Robotics* 2, 4 (2017), eaam8638. https://doi.org/10.1126/scirobotics.aam8638

[292] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr

Kuleshov, and Fei Wang. 2023. Harnessing Biomedical Literature to Calibrate Clinicians' Trust in AI Decision Support Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3544548.3581393

[293] Herman Yau, Chris Russell, and Simon Hadfield. 2020. What Did You Think Would Happen? Explaining Agent Behaviour Through Intended Outcomes. https://doi.org/10.48550/arXiv.2011.05064 arXiv:2011.05064 [cs, stat]

[294] Changchang Yin, Rongjian Zhao, Buyue Qian, Xin Lv, and Ping Zhang. 2019. Domain Knowledge Guided Deep Learning with Electronic Health Records. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, Beijing, China, 738–747. https://doi.org/10.1109/ICDM.2019.00084

[295] Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin D. Haeffele, and Yi Ma. 2023. White-Box Transformers via Sparse Rate Reduction. https://doi.org/10.48550/arXiv.2306.01129 arXiv:2306.01129 [cs]

[296] Luyao Yuan, Xiaofeng Gao, Zilong Zheng, Mark Edmonds, Ying Nian Wu, Federico Rossano, Hongjing Lu, Yixin Zhu, and Song-Chun Zhu. 2022. In Situ Bidirectional Human-Robot Value Alignment. *Science Robotics* 7, 68 (2022), eabm4183. https://doi.org/10.1126/scirobotics.abm4183 arXiv:https://www.science.org/doi/pdf/10.1126/scirobotics.abm4183

[297] Hanna Yun and Ji Hyun Yang. 2020. Multimodal Warning Design for Take-over Request in Conditionally Automated Driving. *European Transport Research Review* 12, 1 (2020), 34. https://doi.org/10.1186/s12544-020-00427-5

[298] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr

Ahmed. 2021. Big Bird: Transformers for Longer Sequences. arXiv:2007.14062 [cs, stat]

[299] Dandan Zhang, Ruoxi Wang, and Benny Lo. 2021. Surgical Gesture Recognition Based on Bidirectional Multi-Layer Independently RNN with Explainable Spatial Feature Extraction. https://doi.org/10.48550/arXiv.2105.00460 arXiv:2105.00460 [cs]

[300] Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A. Ehinger, and Benjamin I. P. Rubinstein. 2021. Invertible Concept-based Explanations for CNN Models with Non-negative Concept Activation Vectors. https://doi.org/10.48550/arXiv.2006.15417 arXiv:2006.15417 [cs]

[301] Xijia Zhang, Yue Guo, Simon Stepputtis, Katia Sycara, and Joseph Campbell. 2023. Explaining Agent Behavior with Large Language Models. arXiv:2309.10346 [cs]

[302] Min Zheng, Jessecae K. Marsh, Jeffrey V. Nickerson, and Samantha Kleinberg. 2020. How Causal Information Affects Decisions. *Cognitive Research: Principles and Implications* 5, 1 (2020), 6. https://doi.org/10.1186/s41235-020-0206-z

[303] Zhiqiang Zhong, Yangqianzi Jiang, and Davide Mottin. 2023. On the Robustness of Post-hoc GNN Explainers to Label Noise. https://doi.org/10.48550/arXiv.2309.01706 arXiv:2309.01706 [cs]

[304] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G Michael Youngblood. 2018. Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE conference on computational intelligence and games (CIG)*. IEEE, 1–8.

[305] Julia El Zini and Mariette Awad. 2022. On the Explainability of Natural Language Processing Deep Models. *Comput. Surveys* 55, 5 (2022), 1–31. https://doi.org/10.1145/3529755 arXiv:2210.06929 [cs]

C

Ethical Approval

15 January 2026

Dear Ms Kong ,

Reference: CURES/26975/2026
Project ID: 20975
Title: Explainable AI for Time-Series Forecasting of Airport Passenger Flow

Thank you for your application to the Cranfield University Research Ethics System (CURES).

We are pleased to inform you your project title change request has been actioned and has been amended as shown in the table below.

Submitted Form Reference	Status Date	Short_Title	full_title
Cranfield University Research Ethics System (CURES)	05/04/2023	Flexible Airport Passenger Flow Estimation	Explainable Deep Sequential Network for Flexible Airport Passenger Flow Estimation
Project title change	14/01/2026	Explainable AI for Time-Series Forecasting of Airport Passenger Flow	Explainable AI for Time-Series Forecasting of Airport Passenger Flow

If you have any queries, please contact CURES Support.

Regards,

CURES Team

Data Protection

We take your data privacy very seriously and will only use your information to provide the products and services you have requested from us. We will not sell, license or trade your information without your consent. For more information about how your data will be processed, please see our privacy policy. Where required, we may share your information across the University and with our commercial subsidiaries.

External ethics committees approval

Some research projects may require approval from research ethics committees external to Cranfield. These committees include, but are not limited to, the Ministry of Defence Research Ethics Committee (MODREC) and the NHS Research Ethics Committee (NHS REC).

Before completing this application, please check whether you need to apply for approval from external committees by visiting the intranet.
If external approval is required, this should be completed prior to submitting a CURES application. Completed external applications should be uploaded to this application in Part 5 with the accompanying decision from the committee.

☑ I confirm I have read and understood the above statement

Part 1: Summary Details

Lead Researcher (Applicant)

Title	First Name	Surname
Ms	Xiangqi	Kong

School: School of Aerospace, Transport and Manufacturing

Course code: DPAROFRC

Email: Xiangqi.Kong@cranfield.ac.uk

Lead Researcher status

◉ Student ○ Staff

Staff only - please provide the details of any internal budget codes set up for this project

Please confirm mode of study

☑ Full-time
☐ Part-time

Please confirm method of study

- ☐ Taught (MSc, MBA)
- ☑ Research (PhD, EngD, DBA, MSc by Research)

Please confirm which project type your application is supporting

- ◉ Thesis
- ○ Group Project
- ○ Other

Lead Supervisor

Title	First Name	Surname
Dr	Yang	Xing

School: School of Aerospace, Transport and Manufacturing

Email: Yang.X@cranfield.ac.uk

If the supervisor details above are missing or incorrect, please contact cures-support@cranfield.ac.uk and we will add or amend the information.

Do not sign the form in Part 6 until we have replied to you that the supervisor information has been updated.

Does this research project involve any co-researchers?

- ○ Yes
- ◉ No

Short title of project

Flexible Airport Passenger Flow Estimation

Full title of project

Explainable Deep Sequential Network for Flexible Airport Passenger Flow Estimation

In 150 - 300 words, describe your proposed project including research methods, identification of any potential risks and measures to reduce these risks (if necessary). You may wish to use your abstract if you have one.

Accurately determining Airport passenger Flow (APF) is critical to efficiently manage airport processes and allocate resources and is fundamental to the provision of a seamless passenger journey experience.

Currently, the estimation for APF have several limitations. One major limitation is the lack of explainable ability for the APF estimation model. An explainable network will allow the design of a transparency network that helps to clearly identify reasons behind the bottlenecks in airport operations and resources that relate to passenger flow, by providing a traceable inference process that enables effective decision-making to optimise
operations.

The PhD project will demonstrate explainable time-series data modelling with the particular focus on flexible airport passenger flow prediction. Futher more, explainable interface will be considered. The research outcomes will have significant opportunity to accelerate the development of smart airport.

Intended start date of data collection: 12/04/2023

Please note: you must not begin your research until approval has been given by CURES.

Will the research be sponsored or funded by an external organisation?

- (•) Yes
- ( ) No

Please enter the name of the external organisation

Saab

Please click on 'Next' in left hand actions bar to move forward to Part 2

Part 2: Ethical Risk

The following questions will help determine the level of ethical risk that your project entails.

Please select any that apply:

- ☐ Living animals are involved in my research
- ☐ The research will use tissues from deceased animals that have been killed as part of the research project
- ☐ The research involves living organisms/biological agents that could potentially impact other entities outside the laboratory
- ☐ Participants may be seriously harmed in the course of the research
- ☐ The proposed research could adversely affect the reputation of the university
- ☐ The research involves the collection of data from those engaged in illegal activities, or involves interaction with illegal content/material

- ☑ None of the above

Please select any that apply:

- ☐ The research will involve the collection and use of 'relevant material' under the Human Tissue Act 2004
- ☐ There are significant power differences present, or dual or other complicating relationships exist
- ☐ The research will deal with material that is obscene or violent in content
- ☐ The risk is greater to the participant(s) than they would normally experience in their daily life (physical, emotional, psychological, career, financial or spiritual)
- ☐ The research interacts with members of potentially vulnerable groups (e.g. children, the elderly, those with learning difficulties, prisoners etc.)
- ☐ The research could negatively impact national infrastructure (e.g. power grid)

- ☑ None of the above

Please select any that apply:

- ☐ The research involves human participants where sensitive personal or special category data will be gathered
- ☐ The research will use tissues from deceased animals that have come from the food supply chain, roadkill or an authorised source
- ☐ The research will use tissues from living animals that have come from normal husbandry, a therapeutic routine, or operation
- ☐ The research involves the collection, use or production of security sensitive material or data
- ☐ The research involves data obtained via social media
- ☐ The research is about illegal activities
- ☐ The research is based on unpublished material

- ☑ None of the above

Please select any that apply:

- ☐ The research involves human participants
- ☐ The research involves human participants who are fully able to provide their informed and valid consent
- ☐ The research involves human participants where personal data or identifying information will be gathered
- ☐ The research consists of interviews where any published or recorded material does not contain personal data, unless permission has been obtained for its use (for example, name, job title or years in post)
- ☐ The research involves human participants where commercially sensitive data will be gathered

- ☑ None of the above

With reference to your answers above, please confirm whether your research requires obtaining informed consent

- ○ Yes
- ○ No
- ◉ not applicable

Please explain why consent is not required for your research

The acquisition of research data does not originate from the participation of human individuals.

Will participants be able, during the data gathering phase, to freely withdraw or modify their consent and to ask for the destruction of all or part of the data that they have contributed, up to a pre-defined date?

- ○ Yes
- ○ No
- ◉ Not applicable

Will any participants be assured of anonymity or confidentiality?

- ○ Yes
- ○ No
- ◉ Not applicable

Are there any conflicts of interest?

- ○ Yes
- ○ No
- ◉ Not applicable

Are you collecting commercially sensitive data?

- ◉ Yes
- ○ No

Is there a Non-Disclosure Agreement in force?

- (•) Yes
- ( ) No

Please comment on any other ethical issues that may arise from your project

Please click on 'Next' in left hand actions bar to move forward to Part 3

Part 3: Methodology and Expertise

This section will help you to indicate what a reviewer might expect when you are using one the data collection methods below. You are strongly advised to upload supporting documents (such as consent forms, experimental protocols, etc.) in Part 5 to avoid unnecessary delays.

Please assign your research project to one of the University Themes

Aerospace

Please select any data collection methods that apply to your research. Refer to the help buttons for more information.

Experiment

- [x] I will use experiments in my research

Please provide further details. Supporting documents can be uploaded in Part 5.

Questions reviewers may have:

- What are the conditions?
- How will informed consent be obtained if humans are participating?
- What are human participants exposed to and for how long?
- Are the levels of risk greater than participants or experimenters would be exposed to in their everyday lives?

The experiment will be conducted on computers without human individuals as the experimental objects.

Direct observation

☐ I will use direct observation in my research

Interviews

☐ I will use interviews in my research

Questionnaires and surveys

☐ I will use questionnaires and/or surveys in my research

Video or audio recording

☐ I will use video or audio recording in my research

Electrophysiological and optical measures from humans

☐ I will use electrophysiological and/or optical measures from humans in my research

Measures from a biological sample

☐ I will use measures from a biological sample in my research

Secondary data

☐ I will use secondary data in my research

Other

☑ No category above applies to my research

Please provide further details. Supporting documents can be uploaded in Part 5.

The initial data is expected to be obtained from open dataset.

Please click on 'Next' in left hand actions bar to move forward to Part 4

Part 4: Specialist Studies

This section relates to research around health and defence topics. For most applications the questions will not be applicable.

Have you ensured that your study complies with The Human Tissue Act 2004 and its nine codes of practice?

- ○ Yes
- ○ No
- ◉ Not applicable

Are you conducting research involving human participants which is funded by the Ministry of Defence (MOD), or involves MOD-employed staff or participants?

- ○ Yes
- ◉ No

Is the research taking place in or through the National Health Service (NHS)?

- ○ Yes
- ◉ No

Please click on 'Next' in left hand actions bar to move forward to Part 5

Part 5: Supporting Documents

Please upload any supporting documents for your research project you may have that will help the reviewers understand your research design such as research protocol, questionnaires, interview schedules, participant information sheets, consent forms, case for support, etc. (dependent on study)

☑ I confirm I have uploaded all supporting documents relevant to my application

Please click on 'Next' in left hand actions bar to move forward to Part 6

Part 6: Declarations and Signatures

Researcher
Declaration

- The completed form is accurate to the best of my knowledge and belief.
- I undertake to abide by Cranfield University's Ethics Policy in undertaking this project.
- I confirm that I have completed the Cranfield University Ethics Training successfully.
- I understand that I am responsible for obtaining ethical approval and any other necessary approvals for my project.
- I understand that I must not begin the data collection until I have received approval from Cranfield University Research Ethics System (CURES).
- I understand that if any significant changes are subsequently made to my project then I will have to submit a new ethics application and until such time as new ethics approval is granted I will have to stop data collection on my project.
- I understand that if any of the above conditions are not met then the University reserves the right to refuse to assess work for any intended degree qualification.

Researcher Signature

Signed: This form was signed by Ms Xiangqi Kong (Xiangqi.Kong@cranfield.ac.uk) on 05/04/2023 9:05 AM

Students must sign their application **before** obtaining a signature from their supervisor.

Once both you and your supervisor have signed the form, your application will be queued for automatic submission.

Supervisor
Declaration

- I confirm that I have read and fully support this application and will be acting as the supervisor of the lead researcher (student) for this project.
- I have checked that the application has been completed correctly and is of good quality.
- In my opinion, the research proposal is viable.
- I confirm that my student has been instructed to complete the Cranfield University Ethics Training.
- I understand that the lead researcher I am supervising must not begin the data collection until they have received approval from the Cranfield University Research Ethics System (CURES).
- I understand that if any significant changes are subsequently made to this project then the lead researcher will be required to submit a new ethics application and until such time as new ethics approval is granted they must stop data collection on the project.
- I understand that if any of the above conditions are not met then the University reserves the right to refuse to assess the student's work for any intended degree qualification.

Lead Supervisor Signature

Signed: This form was signed by Dr Yang Xing (Yang.X@cranfield.ac.uk) on 05/04/2023 9:49 AM