

Physics-Guided SCADA-Based Transformer Model for Robust Early Detection of Wind Turbine Gearbox Faults

Xiaomei Hu¹, Ravi Pandit¹, Yifan Zhao¹ and Jiarong Hong²

¹Faculty of Engineering and Applied Sciences, Cranfield University, Bedford, United Kingdom

²Mechanical Engineering & Saint Anthony Falls Laboratory, University of Minnesota, Rosemount, MN, USA

E-mail: xiaomei.hu.081@cranfield.ac.uk

Abstract. Among the various components of a wind turbine, the gearbox incurs the highest maintenance costs. Therefore, early fault prediction of wind turbine gearboxes has become an area of growing interest, with the aim of enabling predictive maintenance and reducing maintenance costs. This study has developed a Transformer-based deep learning model for early gearbox fault prediction using only SCADA data collected under healthy operating conditions. A reliable normal behaviour model is established through a systematic selection of SCADA variables that are correlated with the gear bearing temperature, combined with rigorous data cleaning and normalisation procedures. In selecting SCADA variables for training the normal behaviour model, the first principle of thermodynamics is applied to the bearing system to identify key variables that most strongly influence bearing temperature. In contrast, many existing studies rely primarily on Pearson correlation to select variables based on statistical relationships with the target signal. The proposed physics-based approach provides a more robust and physically interpretable representation of the relationships between input variables and the target output. The model is validated using data from a real wind turbine and demonstrates its ability to identify gearbox bearing temperature anomalies prior to fault occurrence, thereby providing valuable lead time for maintenance interventions before faults escalate into major gearbox failures. Furthermore, a comparative analysis with an LSTM-based model shows that while the Transformer does not consistently provide earlier fault alarms across all events, it demonstrates improved robustness in distinguishing genuine faults from non-critical high-temperature episodes, thereby reducing false alarms.

1. Introduction

Currently, wind farms primarily rely on regular inspections and reactive maintenance strategies, which are limited in their ability to detect early-stage faults. To address this, accurate prediction of wind turbine condition is crucial for operators. By forecasting the future condition of key components, maintenance can be scheduled proactively, resources can be allocated in advance, and informed maintenance decisions can be made. These capabilities are essential for reducing unplanned downtime, improving turbine availability for power generation, and



Content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](https://creativecommons.org/licenses/by/4.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

ultimately minimising the Levelised Cost of Energy (LCOE). Among the various components of a wind turbine, the gearbox is one of the most failure-prone, accounting for about 35% of total maintenance costs, with replacement averaging around US \$400,000 [1–3]. Temperature is widely recognised as one of the key indicators for gearbox fault diagnosis, as failing components typically generate excess heat. Consequently, early fault prediction for wind turbine gearboxes based on temperature prediction has become an area of growing interest in both academia and industry.

Studies have developed gearbox temperature prediction-based fault detection methods using SCADA data. Machine learning and deep learning techniques such as Artificial Neural Networks (ANN) [4], CNN–LSTM models [5], and PSO–BP algorithms [6] have been widely explored to model the nonlinear relationship between operational variables and gearbox temperatures. Recently, there has been growing interest in applying Transformer-based deep learning models for fault prediction in rotating machinery [7–9]. However, their application to wind turbine gearboxes, particularly through the modelling of bearing temperature, remains relatively under-explored. This study investigates the use of the Vanilla Transformer architecture [10] for early fault prediction using real-world SCADA data from a 2.5 MW wind turbine gearbox.

2. Fault Prediction Methodology

The proposed approach for the early prediction of gearbox faults is based on the creation of a normal behaviour model (NBM). SCADA data from selected variables under healthy conditions, which have strong correlations to and potential impact on the target variable, are used to train the model. Then, this model is employed to predict the target variable, namely the gearbox bearing temperature, which is the SCADA variable physically closest to the gearbox. The aim is to detect anomalies in the predictions, which may indicate potential gearbox faults. In this study, early fault prediction is defined as the issuance of an alarm prior to the gear bearing temperature exceeding the predefined over-temperature threshold of 90 °C, which corresponds to the SCADA-based fault declaration. The earliness of a prediction is quantified by the lead time, defined as the time difference between the first alarm generated by the model and the threshold-crossing time. Predictions with larger positive lead times are considered earlier, while alarms occurring close to or after the threshold crossing are regarded as late predictions.

The development of the proposed methodology follows a structured three-step procedure. First, SCADA data are analysed to select relevant variables for constructing the normal behaviour model, followed by data cleaning to address missing values, faulty measurements, and outliers, as well as data normalisation and division into training and testing sets. Second, Transformer- and LSTM-based architectures are designed and implemented for the gearbox bearing temperature prediction task. Third, a fault indicator for early gearbox fault detection is developed and described.

3. SCADA Data Set

This study uses the Eolos 2.5 MW Liberty C96 wind turbine located at the University of Minnesota, Rosemount, MN, USA, whose gearbox consists of two parallel–helical stages. The datasets comprise SCADA data and met-tower data recorded at 1-minute intervals during 2018, as well as weather data obtained from OpenWeather[11]. Three gearbox bearing over-temperature fault events (exceeding 90 °C) were identified in 2018, occurring on March 5 (21:57–22:13), April 14 (04:57–13:54), and December 1 (22:22–23:06).

3.1. Data Analysis

For the selection of SCADA variables for the normality model, an analysis of the variables is conducted based on the first principle of thermodynamics in the control volume surrounding the gearbox bearing (Eq. 1).

$$C_p \frac{dT_{\text{bear}}}{dt} \cong R^{-1}(T_{\text{bear}} - T_{\text{nacelle}}) + \mu \omega_{\text{gear}}^2 + \alpha P \quad (1)$$

where C_p is the heat capacity of the nacelle, T_{bear} is the bearing temperature, T_{nacelle} is the nacelle temperature, ω_{gear} is the rotational gearbox shaft speed, P is the active power, R is the thermal resistance of the system, μ is the friction coefficient, and α is the heating performance coefficient that describes the energy transformed into heat per unit of electricity produced. The first term in the right-hand side is the expression of heat exchange between bearing and nacelle via conduction. The second term is the expression of heat produced by rotation friction. The third term is the Joule effect produced by the generator.

We also know that the relationship between rotational gearbox shaft speed and the rotor speed is given by Eq. 2, and the finite difference approximation is given by Eq. 3.

$$\omega_{\text{gear}} = \frac{\omega_{\text{rotor}}}{\text{gear ratio}} \quad (2)$$

$$\frac{dT_{\text{bear}}}{dt} \approx T_{\text{bear}}(t + \Delta t) - T_{\text{bear}}(t) \text{ if } \Delta t \rightarrow 0 \quad (3)$$

Substituting Eq. 2 and 3 into Eq. 1 yields Eq. 4.

$$T_{\text{bear}}(t + \Delta t) \cong \beta_1 T_{\text{bear}}(t) + \beta_2 T_{\text{nacelle}}(t) + \beta_3 \omega_{\text{rotor}}^2(t) + \beta_4 P(t) \quad (4)$$

Eq. 4 indicates that, in addition to bearing temperature, nacelle temperature, rotor speed, and active power strongly influence bearing temperature behaviour. These four variables are therefore selected as inputs for the prediction model. In cases where nacelle temperature measurements are unavailable, one could consider using the ambient temperature as a suitable substitute, as adopted in this study.

3.2. Data Cleaning

Data cleaning is a crucial preliminary step in the analysis, as it removes faulty and noisy data that may otherwise distort the results. Missing values and outliers are excluded. Operating conditions in which the active power is zero or negative while the wind speed exceeds the cut-in speed are considered faulty and are therefore removed. In addition, to ensure data quality and relevance, the selected SCADA variables are also filtered to exclude values outside the manufacturer-specified normal operating ranges, as defined in Table 1.

3.3. Data Normalization

The four selected input variables have different units and scales. They need to be scaled before training to ensure all features contribute fairly to the model. In this study, the standard scaler (z-score) method is adopted, which uses mean value and standard deviation to perform scaling. This method has an advantage over the conventional minimum and maximum normalisation since it is less dependent on extreme values.

Table 1. Selected SCADA variables used to develop the normality model

SCADA variables	Range	Units
Bearing Temperature	(-90, 90)	°C
Ambient Temperature	(-20, 40)	°C
Rotor Speed	[0, 17.6]	rpm
Active Power	[0, 2500]	kW

3.4. Data Split

In the proposed methodology, only data corresponding to healthy operating conditions are used for model training, as the objective is to construct a normal behaviour model. After training, the model performance is evaluated using testing datasets that contain gearbox fault events, in order to assess whether the model can predict faults prior to their actual occurrence times. Accordingly, following the 80/20 principle, the dataset is divided into training and testing periods defined as follows: February 1–28 for training and March 1–7 for testing, March 12–April 8 for training and April 9–16 for testing, and October 29–November 25 for training and November 26–December 3 for testing.

4. Transformer vs LSTM

Both models use the four SCADA variables identified in Section 3.1 as input features and predict the gearbox bearing temperature as the output. To ensure a fair comparison between the two architectures, both models are trained using identical training configurations. The Adam optimizer is used with an initial learning rate α_0 of 0.0001, exponential decay coefficient $\beta_1 = 0.9$ and $\beta_2 = 0.992$, and $\epsilon = 10^{-7}$. The input sequence length is set to 60 time steps, the batch size is 32 samples, and training is conducted for 100 epochs.

The Transformer model consists of a single encoder block. The four-dimensional input at each time step is first projected into a 32-dimensional embedding space using a dense layer. Fixed sinusoidal positional encoding is then added to incorporate temporal order information. The encoder block employs multi-head self-attention with two attention heads and an embedding dimension of 32. The feed-forward network contains a hidden layer of 64 neurons with ReLU activation. Residual connections and layer normalization (with $\epsilon = 10^{-6}$) are applied after both the attention and feed-forward sublayers. Dropout with a rate of 0.1 is applied after the multi-head attention output, after the feed-forward network, and before the penultimate dense layer to improve generalization. The resulting Transformer model contains approximately 13,577 trainable parameters.

For the LSTM baseline, a two-layer architecture is adopted consisting of an LSTM layer with 49 units followed by a second LSTM layer with 12 units. A final dense layer is used to produce the temperature prediction. The LSTM configuration results in approximately 13,573 trainable parameters, which closely matches the parameter count of the Transformer model to ensure a fair comparison between the two architectures. The architectural details and parameter counts of both models are summarized in Table 2.

Table 2. Model architecture comparison

Model	Architecture	Key Settings	Trainable Parameters
Transformer	Input projection $\rightarrow$ positional encoding $\rightarrow$ Transformer block $\rightarrow$ global average pooling $\rightarrow$ dense layers	Embedding dimension = 32 Attention heads = 2 Feed-forward size = 64 Dropout = 0.1	$\approx 13,577$
LSTM	LSTM(49) $\rightarrow$ LSTM(12) $\rightarrow$ Dense(1)	Two-layer recurrent architecture	$\approx 13,573$

5. Fault Indicator

The fault indicator (FI) relies on the generation of a nominal residual by comparing the measurements of the physical variable of the system y_t with their estimate $\hat{y}_t$. In this work, the residual value is calculated as $|y_t - \hat{y}_t|$. To mitigate potential false positives, the FI is smoothed using an Exponentially Weighted Moving Average (EWMA), given by equation (5).

$$Z_t = \lambda e_t + (1 - \lambda)Z_{t-1} \quad (5)$$

Here, e_t is the residual, Z_t the EWMA variable at time t , and Z_{t-1} is the EWMA variable at $t - 1$. The smoothing parameter λ is set to 0.6 [12]. The upper control limit of EWMA is defined by equation (6) and (7). An alarm will be raised if the EWMA point is outside this upper control limit.

$$UCL = Z_0 + ks_{EWMA}^2 \quad (6) \quad s_{EWMA}^2 = \frac{\lambda}{2 - \lambda} s^2 \quad (7)$$

Here, Z_0 and s are the mean and variance of the residuals during normal (fault-free) operation, respectively. The constant k , typically set to 3 [12], controls the sensitivity of the alarm.

6. Results and Discussion

Table 3 compares the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) obtained by the Transformer and LSTM models during the March, April and December testing periods. The LSTM consistently achieves lower MAE and RMSE than the Transformer across all three periods. This behaviour is also evident in Fig. 1, which shows that the LSTM predictions fit more closely to the training distribution (healthy SCADA data) than the Transformer, yielding lower MAE/RMSE. However, this higher pointwise accuracy often indicates overfitting to normal behaviour and reduced sensitivity to early anomalies. In contrast, the Transformer's predictions deviate more (higher MAE/RMSE), and these deviations become more pronounced near abnormal conditions (e.g., the fault events in March, April, and December, marked by the vertical green dotted line). This larger deviation from actual values during incipient failure acts as a useful signal: the model is effectively saying “this pattern no longer looks healthy”. Hence, the Transformer may issue earlier alarms, even if it sacrifices pointwise accuracy under normal conditions.

Fig. 2 demonstrates that the Transformer model outperformed the LSTM in predicting the March over-temperature fault. The Transformer issued its first alarm on March 5 at 15:30—6 hours and 27 minutes prior to the fault—whereas the LSTM alarm was triggered at 20:00, only 1 hour and 57 minutes earlier. At the time of the Transformer's alarm, the bearing temperature

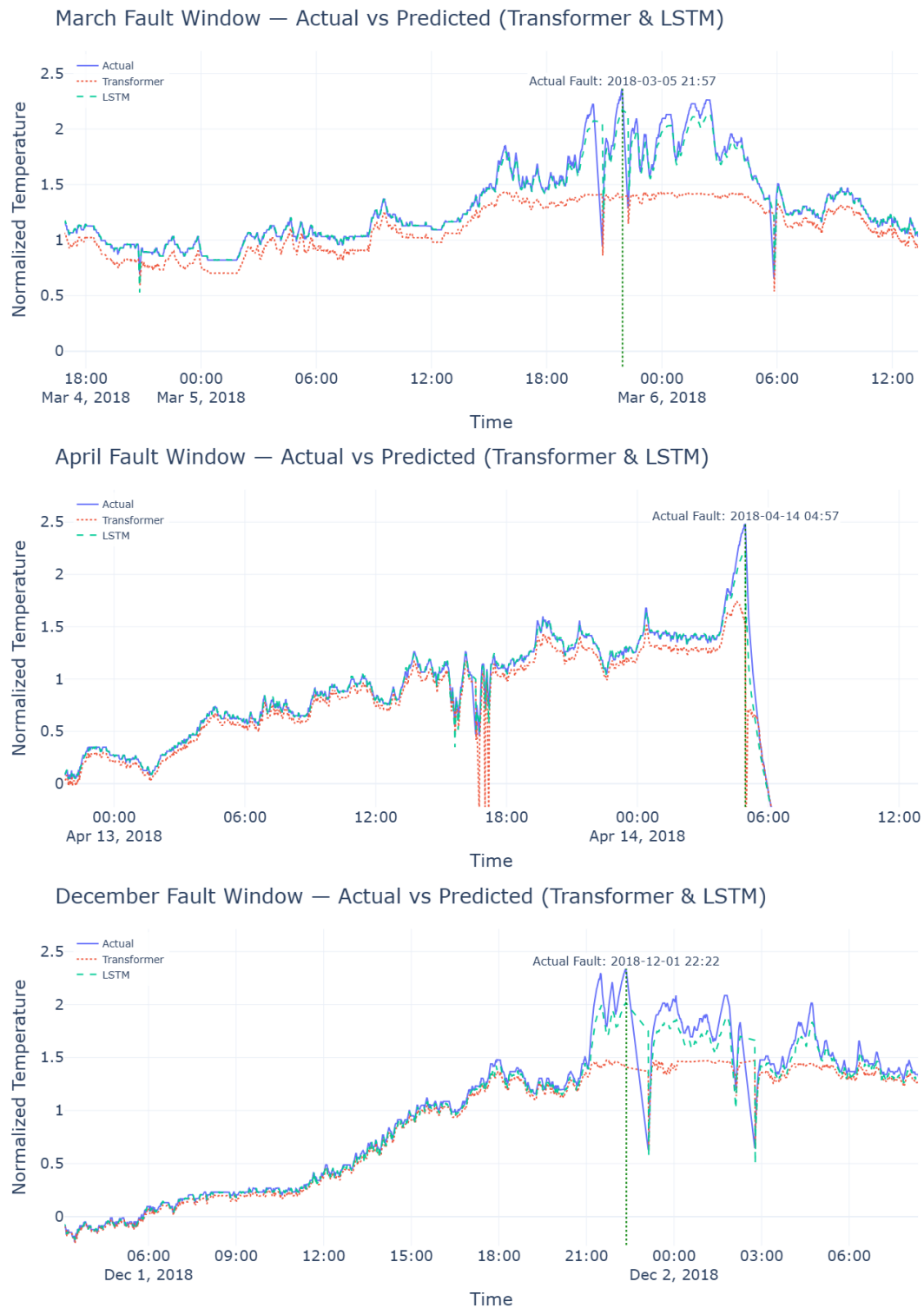


Figure 1. Gearbox Bearing Temperature: Actual vs. Transformer Prediction vs. LSTM Prediction (March, April & December)

Table 3. MAE and RMSE Transformer vs. LSTM (March, April & December)

March	Transformer	LSTM
MAE	0.138668	0.015615
RMSE	0.183133	0.041980

April	Transformer	LSTM
MAE	0.067728	0.016709
RMSE	0.103163	0.040774

December	Transformer	LSTM
MAE	0.065103	0.025430
RMSE	0.115427	0.062438

was 82.4 °C, well below the 90 °C sensor threshold, underscoring its capability for early fault detection. However, the Transformer did not show a clear advantage in the April and December cases: in April, its alarm preceded the fault by 57 minutes and only led the LSTM by 30 minutes, while in December both models alarmed simultaneously, 1 hour and 22 minutes before the fault.

In addition to these three faults, three high-temperature episodes (80-90 °C) occurred on May 24, 21:30 (~80 °C), June 29, 19:29 (~80 °C), and October 20, 13:37 (~83 °C), none of which developed into actual faults. Both models were evaluated on these episodes to test their ability to distinguish between critical and non-critical events, as shown in Fig. 3. Results show that the Transformer produced no false alarms, whereas the LSTM generated false alarms on October 20 by incorrectly interpreting the high temperature as a fault.

Overall, while the LSTM consistently achieved lower MAE and RMSE across all testing periods, this advantage primarily reflects its strong fit to normal operating conditions rather than superior fault prediction capability. The Transformer, despite exhibiting higher pointwise prediction errors, demonstrated greater sensitivity to incipient abnormal behaviour; however, based on the three fault events evaluated in this study, it does not show a definitive advantage over the LSTM in terms of earlier fault prediction. Nevertheless, the Transformer exhibited stronger robustness against false alarms, correctly distinguishing non-critical high-temperature episodes from genuine fault conditions, whereas the LSTM misclassified at least one such event. Consequently, the Transformer offers a slight but practically meaningful advantage for predictive maintenance applications, where reliable fault discrimination and false-alarm avoidance are often more critical than minimising pointwise prediction error under normal operating conditions.

7. Conclusion

This study demonstrates the effectiveness of a Transformer-based deep learning model for early gearbox fault prediction in wind turbines using only SCADA data from healthy operations. A reliable normal behaviour model has been developed through a systematic selection of SCADA variables correlated with the gearbox bearing temperature, combined with rigorous data cleaning and normalisation procedures. In contrast to many existing studies that rely primarily on Pearson correlation to identify statistically correlated variables, this work adopts a physics-based approach grounded in the first law of thermodynamics to identify the key factors influencing bearing temperature. This methodology provides a more robust and physically interpretable

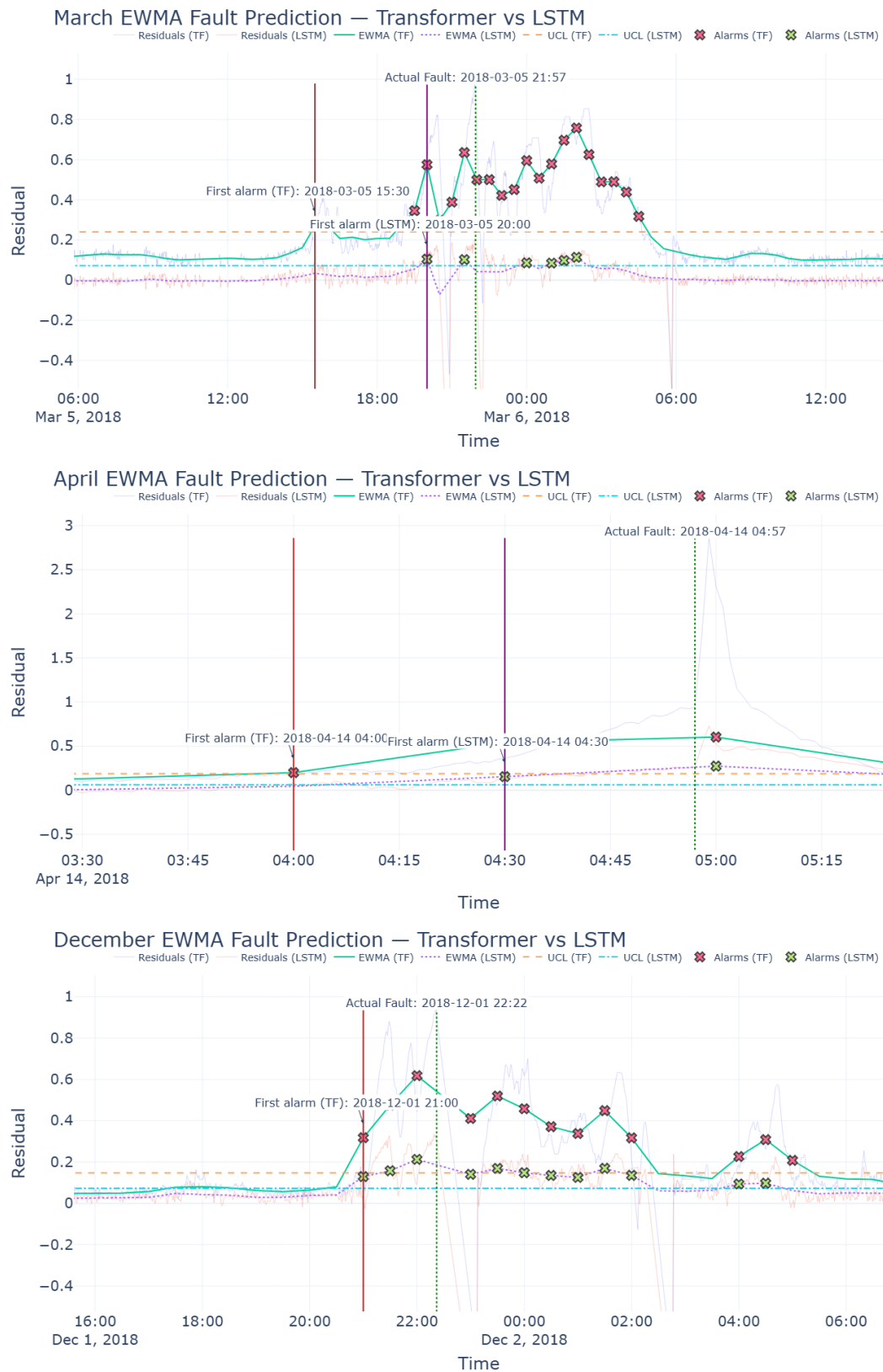


Figure 2. EWMA Over-Temperature Fault Prediction with Residual Layers - Transformer vs LSTM (March, April & December)

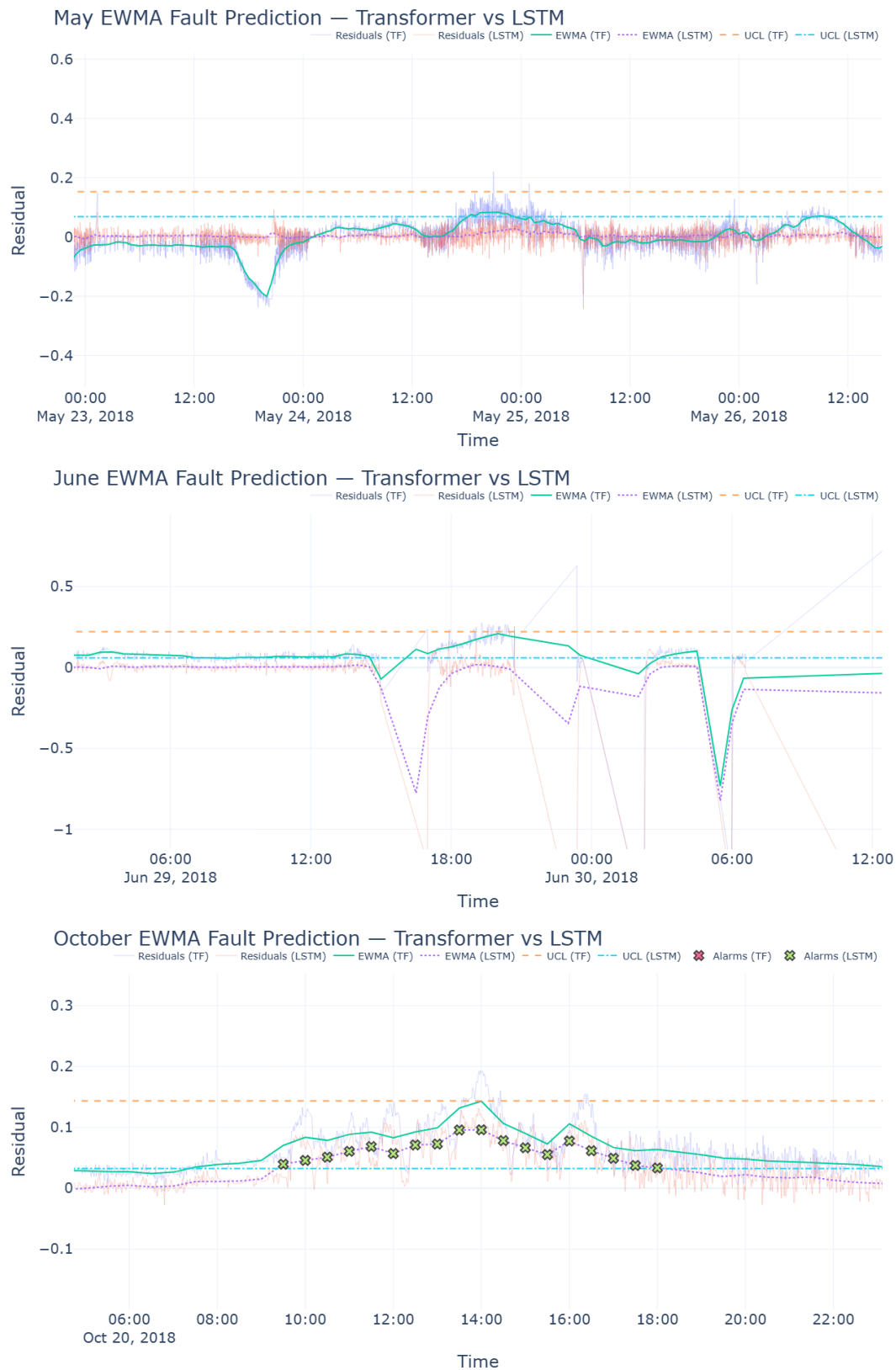


Figure 3. EWMA Over-Temperature Fault Prediction with Residual Layers - Transformer vs LSTM (May, June & October)

representation of the relationships between input variables and the target output. The proposed model is validated on a real wind turbine, showing its ability to identify gearbox bearing temperature anomalies prior to fault occurrence, thereby providing valuable lead time for maintenance interventions before faults escalate into major gearbox failures.

The comparative analysis between the Transformer and LSTM models indicates that the Transformer does not consistently outperform the LSTM in terms of earliest fault prediction across all evaluated events. However, the Transformer demonstrates improved robustness in distinguishing genuine fault conditions from non-critical high-temperature episodes, thereby reducing the occurrence of false alarms. Future work will focus on extending the analysis to additional datasets to further assess the robustness and general applicability of the proposed approach, as well as benchmarking against a broader range of baseline methods. Such efforts represent an important step toward optimising maintenance strategies.

Acknowledgements

The authors acknowledge that this work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) and Cranfield University.

References

- [1] Dao C, Kazemtabrizi B and Crabtree C 2019 Wind turbine reliability data review and impacts on levelised cost of energy
- [2] Pulikollu R, Erdman W, McLaughlin J, Alewine K, Sheng S and Bezner J 2023 Wind turbine generator reliability analysis to reduce operations and maintenance (o & m) costs Tech. rep. National Renewable Energy Laboratory (NREL)
- [3] Moros D, Berrabah N and Ashton I 2024 Neural networks for offshore wind turbine converter failure prognosis *IFAC-PapersOnLine* vol 58 (Elsevier B.V.) pp 592–597 ISSN 24058963
- [4] Puruncajas B, Castellani F, Vidal Y and Tutivén C 2024 *Early Detection of Gearbox Failures in Wind Turbines Using Artificial Neural Networks and SCADA Data* (Springer) pp 337–350
- [5] Zhang F, Zhu Y, Zhang C, Yu P and Li Q 2023 Abnormality detection method for wind turbine bearings based on cnn-lstm *Energies* **16**(7) 3291 ISSN 1996-1073
- [6] Zhang J, Liu L, Wang L and Xi W 2024 Fault detection of key parts of wind turbine based on bp neural network combination prediction model *Energy Informatics* **7**(1) 124 ISSN 2520-8942
- [7] Zhu S, Liao B, Hua Y, Zhang C, Wan F and Qing X 2023 A transformer model with enhanced feature learning and its application in rotating machinery diagnosis *ISA Transactions* **133** 1–12 ISSN 00190578
- [8] Fan Z, He C, Zhi Y, Hu Y and Liu Y 2024 *Research on Remaining Useful Life Prediction Method of Rotating Machinery Based on the Fusion of GRU and Transformer* pp 219–228
- [9] He J, Ma Z, Liu Y and Yang Z 2025 Remaining useful life prediction of rotating machinery based on transformer-kan *Measurement Science and Technology* **36**(5) 056136 ISSN 0957-0233
- [10] Vaswani A, Brain G, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Łukasz Kaiser and Polosukhin I 2023 Attention is all you need Tech. rep. Google Research
- [11] OpenWeather 2025 Openweather history bulk URL https://home.openweathermap.org/history_bulks/new
- [12] Cambron P, Tahan A, Masson C and Pelletier F 2017 Bearing temperature monitoring of a wind turbine using physics-based model *Journal of Quality in Maintenance Engineering* **23**(4) 479–488 ISSN 1355-2511



DSPACE

<https://dspace.org/>

Physics-guided SCADA-based transformer model for robust early detection of wind turbine gearbox faults

Hu, Xiaomei; Pandit, Ravi; Zhao, Yifan; Hong, Jiarong

2026-05-01

IOP Publishing

<https://doi.org/10.1088/1742-6596/3224/6/062025>

Hu X, Pandit R, Zhao Y, Hong J. (2026) Physics-guided SCADA-based transformer model for robust early detection of wind turbine gearbox faults. Journal of Physics: Conference Series, Volume 3224, Issue 6, The Science of Making Torque from Wind 2026 (TORQUE 2026), 03-05 June 2026, Bruges, Belgium, Article number 062025