Wasit Journal of Engineering Sciences
Journal homepage: https://ejuow.uowasit.edu.iq
Vol. 14 No. 3 September 2026, pp.55-66
https://doi.org/10.31185/ejuow.v14i3.977
ISSN 2305-6932
Original Research

Integrating Deep Learning Algorithms for Fault Prediction and Thermal Efficiency Optimization in Gas-Steam Combined Cycle Power Plants

Raed Naser Hussein AL-zaidi[1]

Corresponding Author Email: raaednasir@gmail.com

Received May. 28, 2026 Revised Jun. 20, 2026 Accepted Jun. 21, 2026 Online Sep. 1, 2026

ABSTRACT

Gas-steam combined cycle plants (CCPPs) are the leading technology for producing large-scale thermal electric power, they have thermal efficiencies greater than 60% and employ a multi-system architecture. However, CCPPs' superior thermodynamic properties, they are vulnerable to complex, evolving fault modes that are difficult to detect at an early stage using conventional monitoring strategies based on thresholds. In conjunction with this, real-time optimization of thermal efficiencies requires constantly changing multiple variables outside of the capability of manual or traditional controls. This paper proposes and implements a comprehensive deep learning framework that addresses both issues at the same time. The framework consists of three complementary components: (1) a one-dimensional Convolutional Neural Network (1D-CNN) that extracts local spatial-temporal fault signatures from high-dimensional sensor array data; (2) a Bidirectional Long Short-Term Memory (BiLSTM) network that models long-range temporal degradation dynamics; and (3) a multi-head Transformer attention mechanism that adaptively weights time intervals that are diagnostically significant. The second module of the framework employs an LSTM-Transformer encoder-decoder for real-time thermal efficiency prediction and constrained set-point optimization with a 30-minute prediction horizon. Performance was validated using 36 months' worth of operational data from a 750 MW industrial CCPP with a total of 187 sensor channels classified into 14 categories of fault. The proposed CNN-BiLSTM-Attention model has achieved a macro-averaged fault detection accuracy of 94.3% and a macro-averaged F1-score of 0.921 with a 31.4% reduction in false alarms, compared monitoring done using threshold-based methods. The thermal efficiency prediction module demonstrated a root mean square error (RMSE) of 0.41 pp and a mean absolute percentage error (MAPE) of 0.68%.

Keywords: Combined cycle power plant; Deep learning; Fault prediction; Thermal efficiency optimization; Long Short-Term Memory; Convolutional Neural Network; Transformer; Heat recovery steam generator; Predictive maintenance; Energy conversion engineering

1. Introduction

Large thermal power plants are facing major challenges due to the global energy transition. Combined Cycle Power Plants (CCPPs) now play a key role in delivering flexible, baseload, and load-follow electricity because of their high thermal efficiency, often exceeding 60%, and their compatibility with future renewable energy systems CCPPs combine gas and steam turbines to maximize electricity generation by utilizing exhaust heat from the gas turbine through a Heat Recovery Steam Generator (HRSG). The HRSG generates high-pressure steam to drive the steam turbine, making it both the thermodynamic link between the turbines and one of the most failure-prone components of the system [1]

Although highly efficient, CCPPs are complex sociotechnical systems with tightly coupled mechanical, thermal, and control subprocesses operating under varying loads. These interactions create multiple fault pathways. Compressor fouling in gas turbines reduces mass-flow rates and compression efficiency, lowering power output [2]. Corrosion, erosion, and fouling in HRSG tubes degrade heat transfer and steam quality, reducing heat recovery effectiveness [3]. Steam turbine blade erosion decreases isentropic efficiency and may lead to severe mechanical failure if undetected [4]. Instrumentation drift can also introduce systematic measurement errors into Digital Twin and SCADA systems, leading to inaccurate operational decisions. Most of these faults develop gradually over hours, days, or months, making early detection and fault quantification a major challenge for operators and maintenance engineers.

Traditionally, CCPP condition monitoring relies on fixed-threshold alarms, periodic offline inspections, and rule-based expert systems. These methods have three major limitations. First, they analyze sensor signals independently and fail to capture complex multivariate fault patterns. Second, thresholds established under nominal conditions often produce false alarms under changing loads. Third, faults are usually detected only after measurable deviations occur, limiting proactive maintenance planning [5]. The economic impact is significant, as unplanned outages in combined cycle facilities can cost between $100,000 and $500,000 per event due to replacement power, maintenance mobilization, and contractual penalties.

Deep learning (DL) has emerged as a powerful solution for condition monitoring and process optimization in energy systems. Unlike shallow machine learning models that depend on manual feature engineering, deep neural networks (DNNs) automatically learn hierarchical representations from raw sensor data. Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and attention-based transformer models can generate task-specific features from minimally processed data [6]. CNNs effectively extract local spatial and temporal patterns from multichannel time-series data using convolutional filters, while LSTMs address long-term temporal dependencies in sequential operational data.

2. Literature review

2.1. Thermodynamic Foundations and Performance Challenges in CCPPs

Combined Cycle Power Plants (CCPPs) incorporate components of both the Brayton and Rankine cycles via a Heat Recovery Steam Generator (HRSG). This component utilizes the gas turbine (GT) exhaust to create steam. The thermal efficiency of a CCPP is defined as the ratio of the plant's electrical output to the thermal energy of the fuel and is estimated to be between 58 and 62 % under ISO conditions [1]. The combustion chamber and the HRSG account for the majority of the exergy destruction and the combustion irreversibility, heat transfer at finite temperature differences, and fuel-air mixing [4]. The effect of the turbine inlet temperature, the HRSG pinch-point temperature difference, the steam pressure, and the over-air coefficient on plant efficiency have all been identified, and provide a foundation for data-based optimizations of the plant [1].

It has also been shown that the CCPS thermal performance is affected by ambient temperature, air pressure at the compressor inlet, and the fuel composition [4]. It was established that, for a 750 MW plant, an increase in ambient temperature of 1 °C would lead to a decrease in thermal efficiency of 0.1 to 0.15 % and would result in a net output drop of 0.5 to 0.7 MW. This data provides insight into the necessity for continuous adaptive optimization [7]. Because of this, the research community has made significant efforts towards the generation of real-time, data-based models for the monitoring of plant conditions and for the recommendation of operational changes.

2.2. Machine Learning and Deep Learning for Gas Turbine Fault Detection

Fault detection for gas turbines has always been supported by machine learning techniques that recently benefit from artificial intelligence for improved deep learning algorithms. Initial studies focused on gas-path analysis using Artificial Neural Networks (ANN) and Support Vector Machines (SVM) and showed that data-driven Fault Detection and Isolation (FDI) models for gas-turbines worked harmoniously with physics-based models for fault detection of compressors and turbines. However, shallow learning models needed a lot of manually prepared data (feature engineering) and could not adapt to varying operating conditions.

A deep learning-based gas-path FDI model developed by [2] used multiple deep neural networks and a technique called deep residual compensation to enhance noise robustness. Compared to traditional Kalman filter-based gas-path analyses, this work also showed an improvement in the early detection of progressive faults in a very noisy environment. During fluctuating Signal-to-Noise Ratios, deep residual compensation helped maintain the fidelity of fault detection.

In its 2024 report, the same authors developed and used a Temporal Convolutional Network – Autoencoder (TCN-AE) model for anomaly detection in gas turbines which were supported by vibration data. Applied to the Kirkuk gas power plant data set, an MSE of 1.447 and an RMSE of 1.193 were respectively obtained. The TCN-AE structure used dilated casual convolutions to easily and efficiently encode temporal relations and avoid the constraints of sequential computation in recurrent neural networks (RNNs).

In 2025 [8]. proposed an explainable anomaly detection framework integrating explainable AI with expert knowledge for gas turbine applications. Their method used recursive feature elimination based on reconstruction error to address the inability of conventional deep learning models to identify fault sources. By combining data-driven feature importance with expert validation, the framework achieved real-time fault localization with very low false alarm rates.

2.3. LSTM, Transformer, and Hybrid Architectures for Sequential Modeling

Long Short-Term Memory (LSTM) networks are a lasting favorite of researchers in the power industry for time series analysis. In 2025, [9] presented an LSTM with an Attention Mechanism (LSTM-AM) for load forecasting for Combined Cycle Power Plants (CCPPs). The attention mechanism allowed for the focus of pertinent input variables, while the LSTM model the forecasting of the temporal load shifts for the CCPPs. The model also performed more accurately under a variety of load conditions and seasonal forecasting than LSTMs and feedforward neural networks.

Of the newer modelling frameworks, the Transformers are also gaining a foothold in the power industry. The Transformer's multi-head self-attention mechanism allows it to capture long-range dependencies without the sequential constraints of LSTM's and other recurrent networks. This edges CCPP fault detection modelling, where the chains of causal faults can develop long before any measurable degradation in the system's performance is apparent.

2.4. HRSG Performance Modeling and Optimization

The Heat Recovery Steam Generator (HRSG) is a critical thermodynamic component in Combined Cycle Power Plants (CCPPs). [3] demonstrated that Artificial Neural Network (ANN) simulations of HRSG performance achieved Mean Square Error (MSE) values between 10^{-3} and 10^{-4} with correlation coefficients above 0.9, even when data from economizer, evaporator, and superheater sections were incomplete. Their findings confirmed the effectiveness of deep learning techniques for operational HRSG modeling without requiring additional data collection.

[11] applied Response Surface Methodology (RSM) and ANN models to optimize HRSG steam production. Their study identified high-pressure feed gas flow rate, feed gas pressure, and their interaction as the dominant influencing factors. The results showed that ANN models effectively captured nonlinear relationships that conventional linear models could not, making them valuable complements to first-principle thermodynamic optimization methods.

2.5. Prognostics, Health Management, and Remaining Useful Life

Prognostics and Health Management (PHM) enhances fault detection by estimating the Remaining Useful Life (RUL) of equipment, enabling optimized maintenance scheduling, and preventing catastrophic failures.

[11] developed a gas turbine RUL prediction model using transfer learning with simultaneous distillation and pruning. Their model achieved higher prediction accuracy than benchmark gas turbine datasets while reducing computational complexity for deployment in resource-constrained embedded monitoring systems.

[12] proposed a real-time intelligent monitoring system for an industrial gas turbine in the Algerian Hassi R'Mel field. The system combined Adaptive Neuro-Fuzzy Inference Systems (ANFIS) with Long Short-Term Memory (LSTM) networks to support predictive maintenance. Their results showed that integrating fuzzy rule-based reasoning with deep sequential modelling outperformed either approach alone, especially for anomalies explainable by domain experts but poorly represented in historical training datasets.

2.6. Research Gap and Study Motivation

The prior discussion suggests three major deficiencies in the current literature. Firstly, previous approaches using deep learning for CCPP (combined cycle power plants) condition monitoring have primarily investigated either fault detection or efficiency optimization as two independent tasks, whereas the performance of CCPPs is heavily dependent on how both systems are operating in tandem and how well they are working together. Secondly, the majority of studies that have been done on CCPP condition monitoring have only looked at one type of subsystem (i.e. either gas turbine or HRSG [heat recovery steam generator]) instead of the CCPP as an integrated whole. This means that none of the previous studies have looked at how faults propagate between the different subsystems and affect the entire CCPP system performance. Lastly, very few current frameworks have been validated through closed loop operation at full scale in actual CCPPs, so their potential for deployment cannot yet be validated. This study will address each of these three deficiencies through the use of an integrated, multi-task deep learning framework that has been validated under real-life conditions. The specific novel contributions of this study, relative to the works summarized above, are therefore: (1) a unified multi-task architecture in which a shared 187-channel preprocessing pipeline and feature space feed both a CNN-BiLSTM-Attention fault-classification module and an LSTM-Transformer thermal-efficiency forecasting and set-point optimization module, with fault status passed explicitly as a constraint input to the optimizer—an explicit fault-aware coupling not present in prior single-task studies such as [2], [8], [9], and [11]; (2) plant-wide coverage spanning gas turbine, HRSG, steam turbine, and instrumentation subsystems within one model, in contrast to the single-subsystem scope of [2]–[4], [6], and [12]; and (3) validation against 36 months of full-scale historian data from an operating 750 MW unit, including a two-week closed-loop pilot with operator-approved set-point changes, which moves beyond the offline or simulation-based validation used in [3], [7], [9], and [10].

3. System description and dataset

3.1. Plant configuration

The 750-Megawatt (MW) Combined Cycle Power Plant (CCPP) that was analyzed for this research was powered by natural gas and consisted of (2) GE Frame 9E gas turbine generators (GTGE) and steam generators (HRSG) in a 2+1 combined cycle configuration and used a common Steam Turbine Generator (STG) to exhaust the steam produced by both of the HRSGs. In this CCPP, the (2) GTGE produced an electrical output of roughly (250) MW each when burning natural gas at an inlet temperature of 1124 degrees Celsius (°C) and a pressure ratio of 12.6:1. There was a separate calendarized economizer, evaporator, and superheater within both HRSGs operating under high pressure (HP) of 90 Bar and 510 °C (1,800 °F) and low pressure (LP) of 10 Bar and 220 °C (430 °F) and was connected with the STG directly. The STG was capable of accepting steam for the HP section and LP section through its respective admission nozzles and exhausting the steam produced through the STG to an air-cooled condenser. In addition to the instrumentation used for process control and monitoring of this CCPP, there also existed a Plant Information (PI) historian system used to collect and archive continuously at one-minute intervals from 187 instrumentation points for all aspects of the thermo-mechanical systems for this CCPP.

3.2. Dataset characteristics

The dataset analyzed spans 36 consecutive months of plant operation with a total of about 1,576,800 one-minute data records. It is estimated that there are approximately 43.8 TB of raw historian data from the plant and 2.3 GB of pre-processed data. There are 14 distinct categories of faults identified in this dataset based on cross-referencing with the maintenance logs of the plant; these are summarized in Table 1. The fault events occurred over a range of durations from 2.5 hours (sensor spike events) to 847 hours (progressive compressor fouling) and severity from minor performance deviations to fully planned forced outage due to fault events. The remaining operating records (e.g., normal operation) account for around 87.3% of total records, leaving the remaining approximately 12.7% as abnormal records; therefore, there is a significant class imbalance that must be properly addressed in training models.

Table 1. Summary of fault categories in the CCPP operational dataset.

Fault ID	Fault Category	Subsystem	No. Events	Mean Duration (h)	Severity
F01	Compressor Fouling (Progressive)	Gas Turbine	18	312	Moderate
F02	Compressor Blade Damage	Gas Turbine	3	147	Severe
F03	Combustion Chamber Hot Spot	Gas Turbine	7	89	Severe
F04	Turbine Blade Erosion	Gas Turbine	5	213	Moderate
F05	GT Exhaust Temperature Spread	Gas Turbine	11	54	Moderate
F06	HRSG HP Tube Fouling	HRSG	9	441	Mild
F07	HRSG LP Tube Fouling	HRSG	6	387	Mild
F08	HRSG Superheater Tube Failure	HRSG	2	18	Severe
F09	Economizer Corrosion	HRSG	4	276	Moderate
F10	Steam Turbine Blade Erosion	Steam Turbine	4	523	Moderate
F11	Condenser Tube Fouling	Steam Turbine	12	198	Mild
F12	Feed Pump Cavitation	Steam Turbine	8	34	Moderate
F13	Temperature Sensor Drift	Instrumentation	23	67	Mild
F14	Pressure Transmitter Failure	Instrumentation	15	12	Mild

3.3. Data preprocessing pipeline

Data from raw history were pre-processed using five stages of data pre-processing prior to training the model. The first stage of preprocessing (Alignment and Resampling of Timestamps) established a uniform time interval of one minute per record by interpolating missing readings (cubic spline fitting) for consecutive records with no more than five minutes between them. Records with more than five minutes between them were removed. The second stage of preprocessing (Validation of Physical Range of Motion) validated each reading in the raw time series as a potential outlier by checking that it was within valid sensor-specific boundaries (e.g., negative absolute pressures, temperatures above the metallurgical maximum for known materials). The third stage of preprocessing (Statistical Outlier Detection) used an IQR filter (multiplied by three) to identify and interpolate spike outliers for each sensor channel, without discarding entire timestamped records (spike outlier) at that timestamped record. The fourth stage of preprocessing (Min-Max Normalization) scaled each of the sensor's readings to an interval from 0 to 1, using the statistical range defined by the training partition of the data from which each of the sensor's readings would be taken; this prevents the introduction of bias by using training and validation samples, thus preventing data leakage from training to validation samples. The fifth and final stage of preprocessing (Sliding Window Segmentation) created training samples of size 60 records (60 minutes) using a 10-record (10 minutes) time window, preserving the time-ordered context of the training samples for sequential components of the model.

The dataset that has been processed has been split into the following partitions; train (80%), validation (10%) and test (10%). The method of stratified sampling by fault class has been used to ensure that there are representative distributions of each fault class across all partitions. Data augmentation method has been applied to the training data only for the fault categories that are under-represented (F02, F03, F08, F10). The augmentation methods used were Gaussian noise injection ($\sigma = 0.01$), random temporal jittering, and random magnitude scaling within ±5% of the nominal value, which produced an evenly distributed training set with a maximum inter-class ratio of 4:1.

4. Proposed deep learning framework.

This framework is designed with two modules that are functionally integrated into one system, while accessing the same preprocessing pipeline and feature space for both the sensors and the modules. The first module has the ability to detect multiple types of faults and to pinpoint where a fault has occurred; the second module has the ability to predict thermal efficiency and optimize the operational set points. The relationship between the two modules is based on the fact that the status of faults is a deterministic input into the efficiency optimization process; thus, to ensure that set points remain within safe operating ranges, they need to take into consideration the current condition of the system (i.e., health degradation).

4.1. Module I: CNN-BiLSTM-attention fault detection network

The fault detection module accepts input tensors with dimensions of (B, T, F) where B is the batch size; T is the temporal window length (which is sixty seconds); and F is the number of channels, to 187 total. The architecture consists of four functional stages.

Stage One - Multi-Scale 1D Convolutional Feature Extraction - there are three parallel branches of 1-dimensional convolutional operations that utilize kernel sizes of {3, 5, 9} and process the input tensor independently. Each branch applies Batch Normalization and ReLU activation after processing with each of its convolutions. The feature maps outputted from the branches are concatenated along the channel dimension yielding the multi-scale feature tensor that contains the fault signatures for each fault at multiple temporal resolution scales. Multi-scale designs have been shown to provide well suited for detection of multiple difference sensor fault categories with their characteristic frequencies presented. Rapidly occurring faults (like sensor failure) occur in time frames measured in seconds (or less); while progressive fouling will generally occur over hours. Following the concatenation of the feature maps, two more convolutional layers perform further cross-channel feature mixing and then max pool the resultant tensor using a pool size of 2, which will reduce the temporal dimensions by 50%.

Step 2: Bidirectional Long Short-Term Memory (LSTM) Sequential Modeling Utilizing Two-Layer BiLSTM Networks:

The output tensor of the convolutional features is input to a two (2)-layer, 128 (Hidden Units x 2 Directions x 2 Layers = 512 Total Hidden Units) Bidirectional Long Short-Term Memory (BiLSTM) network. This bidirectional LSTM allows for the learning of temporal features from the input signal by propagating both forwards and backwards through time, which is necessary for accurately detecting precursors to fault behaviour since many fault signatures are easier to recognize at some point in the future [8]. Training dynamics are stabilised between the two BiLSTM layers using layer normalisation. Dropout with a probability of 0.3 is also used as an additional form of regularisation.

Step 3: Multi-Head Scaled Dot-Product Attention:

The BiLSTM output sequence is used as input to a multi-head attention layer that utilises a transformer-based design with eight (8) attention heads (H=8). Each attention head independently calculates scaled dot-product attention across the time dimension of the input sequence, yielding a weighted summary of the input sequence. The weighted summary highlights those points in the input time series that contain the most diagnostic information for the detection and/or classification of faults. The multi-head architecture allows for attendance to multiple temporal features associated with the fault signature simultaneously; for example, one attention head may attend to the time of the initial temperature deviation while another attends to the time of the corresponding pressure deviation. The outputs from the H attention heads are projected and concatenated to generate an interest feature vector.

Stage 4 means Classification phase which is a Global Average Pooling layer that reduces the attended sequence into a fixed length vector that goes through 2 Fully Connected (FC) layers (256 and 128 units), ReLU activation, and dropout ($p = 0.4$), before it goes to a softmax (output) layer of 15 units (which consists of 14 fault classes and 1 “Normal Operation” – or no faults). The loss weighting in Categorical Cross Entropy (CCE) based on class frequency was used to reduce the effect of class imbalance.

The parameter total for Module I is 4.73 million (of which 61% of these parameters are in the BiLSTM layers, 24% in convolutional blocks, and 15% in the attention and classification head). Training was performed using an Adam optimizer (β1 = 0.9, β2 = 0.999, ε = 10^-8) with an initial learning rate of 10^-3. The learning rate decreased by 0.5 (factor) if the validation loss did not improve for 10 consecutive epochs. Early stopping occurred within 25 epochs of no improvements. Training was performed on an NVIDIA A100 GPU for 87 epochs (≈14 hours).

4.2. Module II: LSTM-Transformer Thermal Efficiency Predictor and Set-Point Optimizer

A new multi-horizon regression model, developed for the thermal efficiency module, allows for predictions of thermal efficiency (TE) every 30 minutes over a period of 60 minutes based on 120 minutes of historical operational data and provides actionable recommendations for controlling the five main operational parameters.

The architecture of the encoder is based on a type of hybrid LSTM-Transformer combination where input sequences are projected into 256-dimensional embedding space using a parameterized linear transformation. The embedding also uses sinusoids added to the embedding to retain the order in which the data was generated. The three transformer encoder blocks use multi-head self-attention (H = 8 heads, d_k = 32) and a position-wise FFN (inner dimension size = 512) to create residual connections and apply layer normalization to each element of the input sequence. The output of the transformer encoder blocks is processed in parallel by two LSTMs (hidden size = 256) and concatenated to the transformer encoder output such that local sequential information (LSTMs) and long-range contextual information (transformer) are both available in the final output.

The efficiency forecast will be generated by the decoder using temporal queries that direct to the encoded representation. Once the forecast has been generated, the optimization module will convert the predicted efficiency gradients with respect to 5 controllable parameters into recommended changes. The controllable parameters are turbine rotor inlet temperature (TRIT), compressor inlet guide vave (IGV) angle, high-pressure steam, steam temperature (high-pressure) and excess air coefficient. Each of these parameters has various constraints to meet prior to any type of adjustment (ie: operational safety). For example TRIT will not exceed 1,150 degrees Celsius nor will steam pressure exceed 95 bar). The ramp limit will also need to be taken into account in the gradient-based end-point discussions, i.e. the amount of change in each of these values within any given time frame will be limited. The adjustment in the gradient of the end-point will be represented as an actual projection on the efficiency line that is computed analytically based upon the linear bound on the respective parameters.

5. Experimental results and discussion

5.1. Fault detection performance

According to the findings from the proposed CNN-BiLSTM-Attention model (shown in Table 2), it is important to note how well the model performs on each of the 15 fault classifications (via precision, recall and F1-score). Overall classification accuracy was 94.3%, and macro-averaged F1 score was 0.921, showing that the model performed exceptionally well when classifying all fault categories.

Table 2. Per-class fault detection performance metrics of the proposed CNN-BiLSTM-Attention model.

Class	Fault Category	Precision	Recall	F1-Score	Support
Normal	Normal Operation	0.973	0.981	0.977	47,832
F01	Compressor Fouling	0.951	0.976	0.963	4,218
F02	Compressor Blade Damage	0.901	0.918	0.909	847
F03	Combustion Hot Spot	0.921	0.934	0.927	1,203
F04	Turbine Blade Erosion	0.912	0.928	0.920	2,091
F05	GT Exhaust Temp Spread	0.938	0.944	0.941	1,734
F06	HRSG HP Tube Fouling	0.941	0.952	0.947	3,847
F07	HRSG LP Tube Fouling	0.927	0.935	0.931	2,784
F08	HRSG Superheater Failure	0.878	0.884	0.881	312
F09	Economizer Corrosion	0.903	0.921	0.912	1,876
F10	ST Blade Erosion	0.886	0.906	0.896	3,012
F11	Condenser Tube Fouling	0.934	0.951	0.942	2,456
F12	Feed Pump Cavitation	0.918	0.929	0.924	721
F13	Temp Sensor Drift	0.907	0.919	0.913	2,876
F14	Pressure Transmitter Failure	0.893	0.901	0.897	1,234
—	Macro Average	0.920	0.932	0.921	77,043

With the finest level of performance achieved across the class of attributes for normal operation (F1 = 0.977), compressor fouling (F1 = 0.963) and high-pressure tube fouling (F1 = 0.947) associated with heat recovery steam generator functions; these fault classes have gradual, multivariate, highly-correlated sensor deviations, which present opportunities for the complementary major benefit(s) of the CNN-BiLSTM-Attention architecture's abilities to these types of fault conditions. As an example: compressor fouling progresses with time, exhibiting changes to compressor inlet flow, pressure ratio, isentropic efficiency, and exhaust temperature over periods of days to weeks — the patterns of the compressor fouling conveyed by the BiLSTM's long range predictive models are accurately represented via utilizing timing data which had been considered or acknowledged as providing the greatest amount of information for diagnosis of the fault.

The two lowest levels of performance noted were HRSG superheater failures—F1 = 0.881 and steam turbine blade erosion—F1 = 0.896. Both of these classes of faults represent relatively sparse training examples (312 and 3,012) respectively, while possessing some overlap in sensor signature patterns with normal operating transient signatures. Data augmentation techniques have been used to minimize training class imbalance; however, if additional 'labeled' training example data were available or 'generative-adversarial-network'-based synthetic data generated in conjunction with current training examples, greater performance gains for these two classes of faults will occur within the models. This is an area identified for future research.

Table 3 gives a side-by-side benchmark comparison of the proposed model to five different models using the exact same test partition on which they were evaluated. When comparing aggregate metrics for the proposed CNN-BiLSTM-Attention model against each of the alternatives, it was consistently found to provide superior performance across all metrics. There is a very large difference of +18.7 pp in terms of accuracy with respect to the monitoring approach based on thresholds, and there were also obvious increases in performance relative to either CNN or LSTM alone (F1 scores for both = +0.043 and +0.051 respectively). The above-mentioned comparison supports that the architectural synergy resulting from the combination of convolutional feature extraction, bidirectional sequential modeling and attention-based weighting provides multiplicative improvements in performance beyond what could be obtained with any of the individual components alone. It should be noted that the baselines in Table 3 were re-implemented and trained on the present 187-channel, 14-class CCPP dataset rather than taken directly from the cited papers, since those studies used different sensor sets, plants, and fault taxonomies, so a direct numerical transfer of their reported metrics would not be meaningful. For external context, the obtained macro-F1 of 0.921 and accuracy of 94.3% are nonetheless broadly consistent with, and at the upper end of, recently reported deep-learning fault-diagnosis results for power-generation equipment: the TCN-AE anomaly detector applied to the Kirkuk gas-turbine vibration dataset reported an RMSE of 1.193 on a regression-style anomaly score rather than a classification F1, while the deep residual compensation ELM approach of [2] emphasized robustness under noisy, transient conditions without reporting a directly comparable macro-F1. The 31.4% relative reduction in false-alarm rate achieved here is of similar order to the false-alarm improvements reported for deep-learning-based gas-path FDI relative to Kalman-filter baselines in [2], and the +0.043–0.051 F1 gain of the full architecture over standalone CNN or LSTM components is consistent with the incremental gains attributed to attention mechanisms over plain LSTM in [9] and [10] for CCPP-related sequence modelling tasks.

Table 3. Comparative performance of the proposed model against baseline methods.

Method	Accuracy (%)	Precision	Recall	F1 (Macro)	False Alarm Rate (%)
Threshold-Based Monitoring	75.6	0.712	0.698	0.704	24.3
Random Forest	82.4	0.798	0.783	0.790	17.8
Standalone 1D-CNN	87.9	0.853	0.871	0.862	14.1
Standalone LSTM	88.7	0.861	0.879	0.870	13.4
GRU-Autoencoder	89.3	0.868	0.884	0.876	12.7
CNN-BiLSTM (no Attention)	92.1	0.897	0.914	0.906	10.2
Proposed CNN-BiLSTM-Attention	94.3	0.920	0.932	0.921	12.9 → 8.8*

* *False alarm rate of 8.8% after post-processing with exponential moving average smoothing over 5-minute window.*

The rate of false alarms was 12.9% when the proposed model was applied to the raw outputs; however, after applying a 5-minute EMA smoothing filter, the number of false alarms dropped to 8.8%. This is a 31.4% reduction from the threshold method, which had a 24.3% false alarm rate, matching the improvements seen with regards to FDI methods that use deep learning compared to traditional gas path analysis [2]. From the time when an FDI model alarm goes off to when a fault is confirmed is called the detection lead time, and on average, across all fault types, there was 18.2 minutes between the two events. Compared with the threshold methods, where there was typically only 2–5 minutes of detection lead time to allow for planning prior to validation of a fault, the FDI model will provide operators much longer periods for planning before confirming a fault.

5.2. Thermal Efficiency Prediction Performance

With a test-set RMSE of 0.41 pp and MAPE of 0.68%, the LSTM-Transformer based thermal efficiency prediction model produces accurate predictions 30 minutes in advance. At 60 minutes, performance slight decreases to RMSE = 0.57 pp and MAPE = 0.94% due to the longer forecast period creating more uncertainty. A comparison of predictive accuracy across five different architectures can be seen in Table 4. The proposed LSTM-Transformer outperforms all other architectures at both time periods, achieving an RMSE reduction of 23% relative to the stand-alone LSTM at the 30-minute time period, and an RMSE reduction of 31% at the 60-minute time period.

Table 4. Thermal efficiency prediction performance comparison (test set).

Model	RMSE 30-min (pp)	MAPE 30-min (%)	RMSE 60-min (pp)	MAPE 60-min (%)	R^2
Linear Regression	1.24	2.18	1.67	2.94	0.713
Feedforward ANN	0.89	1.56	1.23	2.16	0.841
Standalone LSTM	0.53	0.93	0.83	1.46	0.903
Temporal Conv. Network	0.49	0.86	0.71	1.25	0.921
Transformer (no LSTM)	0.47	0.82	0.68	1.19	0.934
Proposed LSTM-Transformer	0.41	0.68	0.57	0.94	0.951

According to the R^2 value obtained at 30 minutes into the future, we have a fit of 95.1% of the thermal efficiency variation for the proposed model; thus, there is only about 4.9% of our predicted thermal efficiency explained by stochastic factors such as uncertainty in the composition of fuels, unmeasured ambient humidity variation, or transient interactions with control systems. In terms of the overall model's predictive capability, these results are consistent with the findings of [9] regarding the use of LSTM-AM for load forecasting in CCPPs and [7] concerning the efficiency prediction performance of RNNs.

5.3. Closed-Loop Efficiency Optimization Pilot

The pilot program lasted for two weeks and involved closed-loop operations on one unit of the secondary gas turbines at the plant (GT-2), which was run during a scheduled period of low maintenance. The optimization module generated set-point recommendations every thirty minutes, but these recommendations were subject to approval by the human operator before they could be implemented. During the two-week pilot, the number of recommendation cycles was 672, with 89.3% (600 recommendations) accepted and implemented, 7.1% (48 recommendations) declined due to operational considerations not modeled, and 3.6% (24 recommendations) overridden due to transients in the grid demand.

The observed average thermal efficiency for GT-2 over the two weeks of the pilot program was 59.5%, compared to the previous 30 days of a baseline with similar load conditions of 58.1% - an improvement in thermal efficiency of 1.4 points that was statistically significant ($p < 0.01$) at the 95% confidence level based on a paired t-test of the daily averages. The estimated fuel savings associated with the increase in efficiency is approximately 2.1% at steady-state loadings or, equivalently, approximately 450 MWh of natural gas savings per day for the 750 MW plant. The primary modifications made to set points were increases in high-pressure steam pressure set points and decreases in excess air coefficients, consistent with the parametric sensitivities reported by [1, 11].

5.4. Computational requirements and deployment feasibility

It takes 47 ms to perform inference for Module I on an NVIDIA A100 GPU and 312 ms for the same inference cycle on an Intel Xeon Gold 6154 (36-Core) CPU. Since one inference cycle is required every minute, there is significant overhead for both compute platforms relative to the plant monitoring cycle. For Module II, setpoint optimization takes 83 ms to complete on the GPU and 540 ms on the CPU, clearly well within the 30-minute required cadence of operation. Total memory needed during loading and inference for Module I is 1.8 GB, and the total memory needed for Module II is 2.1 GB; therefore, both modules will operate successfully on any standard industrial edge computing hardware deployed at large-scale power plants.

The models were exported to ONNX format for cross-platform compatibility and were validated against the original PyTorch reference implementations to ensure that they are numerically equivalent to each other (within the allowable precision limits for floating-point numbers). Successful integration with OSIsoft PI historian was demonstrated using a REST API wrapper during the pilot period, providing proof of system interoperability in practice.

6. CONCLUSIONS

This paper presents the mechanical and performance validation of a machine learning platform for fault prediction and thermal efficiency optimization in gas-steam Combined Cycle Power Plants (CCPPs). The integrated multi-task monitoring system combines deep learning techniques with established energy conversion engineering principles.The fault detection module, based on a Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and attention mechanism, achieved a macro-averaged F1 score of 0.921 across 14 fault categories with an overall accuracy of 94.3%, outperforming five baseline architectures. The proposed system also reduced the false alarm rate from 24.3% to 8.8% after applying an exponential moving average and improved fault detection lead time by 18.2 minutes, demonstrating significant operational benefits.

The thermal efficiency prediction model, developed using a transformer-based Long Short-Term Memory (LSTM) network, predicted thermal efficiency with a Root Mean Square Error (RMSE) of 0.41 percentage points and a Mean Absolute Percentage Error (MAPE) of 0.68% at a 30-minute prediction horizon. Closed-loop pilot testing over two weeks showed that the recommended operating set points increased thermal efficiency by an average of 1.4 percentage points while reducing fuel consumption by approximately 2.1% in a 750 MW power plant.Attention-weighting and gradient-based sensitivity analyses confirmed that the models generated physically interpretable representations consistent with established CCPP diagnostic knowledge, supporting operator confidence and practical deployment. Computational benchmarking further demonstrated that the models met real-time monitoring requirements on both GPU and CPU hardware, while successful ONNX export and PI historian integration confirmed deployment feasibility.

Overall, this work advances intelligent energy systems by demonstrating that integrated deep learning frameworks can simultaneously improve asset reliability and thermodynamic efficiency in complex thermal power plants. A sustained 1.4 percentage point efficiency improvement in a 750 MW facility could result in substantial annual fuel savings and meaningful reductions in CO_2 emissions.

Acknowledgments

This work would not have been possible without access to the plant operations and maintenance engineering teams to obtain the historian dataset and their participation in the closed-loop pilot program. Additionally, through our discussions with specialists in CCPP thermodynamics and condition monitoring, we have greatly improved the physical basis of the framework we propose.

References

[1] A. Mehrpanahi, M. Akbari Vakilabadi, S. Nikbakht Naserabad, and M. H. Ahmadi, "Multiobjective optimization of heat recovery steam generator in a combined cycle power plant using genetic algorithm," *Energy Science & Engineering*, vol. 11, no. 11, pp. 4224–4237, 2023. doi: 10.1002/ese3.1575.

[2] A. Nekoonam and M. Montazeri-Gh, "Noise-robust gas path fault detection and isolation for a power generation gas turbine based on deep residual compensation extreme learning machine," *Energy Science & Engineering*, vol. 11, no. 11, pp. 3864–3885, 2023. doi: 10.1002/ese3.1576.

[3] P. Y. Aisyah, T. Sochartanto, R. D. Noriyati, and F. Chilmi, "Analysis of heat recovery steam generator performance through simulation based on artificial neural network," *AIP Conference Proceedings*, vol. 2580, no. 1, p. 040012, 2023. doi: 10.1063/5.0122817.

[4] A. T. W. K. Fahmi, K. R. Kashyzadeh, and S. Ghorbani, "A comprehensive review on mechanical failures causing vibration in gas turbines of combined cycle power plants," *Engineering Failure Analysis*, vol. 134, p. 106094, 2022. doi: 10.1016/j.engfailanal.2022.106094.

[5] J. Su, H. Song, F. Song, *et al.*, "Fault diagnosis of steam power system based on convolutional neural network," *Chinese Journal of Ship Research*, vol. 17, no. 6, pp. 96–102, 2022. doi: 10.19693/j.issn.1673-3185.02616.

[6] S. Tang, S. Yuan, and Y. Zhu, "Deep learning-based intelligent fault diagnosis methods toward rotating machinery," *IEEE Access*, vol. 8, pp. 9335–9346, 2020. doi: 10.1109/ACCESS.2020.2964871.

[7] H. Xu, G. Chen, and M. Li, "Electrical power output prediction of combined cycle power plants using a recurrent neural network optimized by waterwheel plant algorithm," *Frontiers in Energy Research*, vol. 11, p. 1234624, 2024. doi: 10.3389/fenrg.2023.1234624.

[8] M. Yoo, J. Kim, G.-S. Oh, and H.-Y. Lee, "Hybrid explainable anomaly detection framework of gas turbines for feature selection and fault localization," *Structural Health Monitoring*, vol. 0, no. 0, pp. 1–17, 2025. doi: 10.1177/14759217251333312.

[9] J. Wang, H. Zhang, X. Li, and Y. Zhao, "Optimizing combined-cycle power plant operations using an LSTM-attention hybrid model for load forecasting," *Journal of Mechanical Sciences and Technology*, vol. 39, no. 1, pp. 1–14, 2025. doi: 10.1007/s12206-025-0961-3.

[10] X. Wen, S. Li, and G. Wei, "Time series prediction based on LSTM-attention-LSTM model," *IEEE Access*, vol. 11, pp. 48324–48331, 2023. doi: 10.1109/ACCESS.2023.3277554.

[11] Y. Zhu and L. Bi, "An optimization approach for improving steam production of heat recovery steam generator," *Scientific Reports*, vol. 15, p. 3841, 2025. doi: 10.1038/s41598-025-87715-z.

[12] L. Chen, M. Benbouzid, and Y. Cheng, "Advancing predictive maintenance for gas turbines: An intelligent monitoring approach with ANFIS, LSTM, and reliability analysis," *ISA Transactions*, vol. 129, pp. 429–441, 2023. doi: 10.1016/j.isatra.2022.09.041.