International Journal for Multidisciplinary Research (IJFMR)

E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Transformer-Based Deep Learning for Optimized Vehicular Engine Health Monitoring

K. Thrilochana Devi[1], Durgaprasad Dondapati[2],
Harsha Vardhan Gottipati[3], Yesu Babu Chattu[4], Jaya Krishna Aripirala[5]

[1]Assistant Professor, Department of IT, Vasireddy Venkatadri Institute of Technology, Guntur
[2,3,4,5]B.Tech. Student, Department of IT, Vasireddy Venkatadri Institute of Technology, Guntur

ABSTRACT

The growing reliance on industrial machinery and critical infrastructure has increased the need for real-time, accurate engine health monitoring to prevent failures and optimize performance. Traditional fault detection methods often rely on rule-based systems or shallow machine learning models, which may fail to capture complex patterns in engine behavior. In this study, we suggest a combination of Long Short-Term Memory (LSTM) and Temporal Fusion Transformer (TFT) model for engine health classification. The LSTM captures sequential dependencies in sensor data, while the TFT enhances interpretability by focusing on multi-horizon forecasting and attention-based feature selection. The system is trained on a dataset of engine parameters, preprocessed using normalization techniques, and optimized with a real-time adaptive retraining mechanism to improve model robustness over time. Experimental results demonstrate that our hybrid model achieves high accuracy, outperforming conventional machine learning approaches. Additionally, we integrate our model into a Flask-based web application to enable real-time monitoring and user-friendly fault detection. This research contributes to the advancement of intelligent predictive maintenance, reducing operational costs and enhancing industrial safety.

Keywords: Engine Health Monitoring, Long Short-Term Memory (LSTM), Temporal Fusion Transformer (TFT), Time-Series Forecasting, Predictive Maintenance, Sequential Data Modeling.

1. INTRODUCTION

In modern industrial and automotive applications, ensuring the health and reliability of engines is crucial for operational efficiency and safety. Unexpected engine failures can result in costly downtime, increased maintenance expenses, and potential safety hazards. Traditional methods of engine health monitoring rely on scheduled maintenance and manual inspections, which may not accurately predict failures in advance. With the advent of deep learning, Predictive maintenance has gained substantial interest as a proactive approach to identifying engine faults before they lead to critical failures. In particular, time-series forecasting models, such as Long Short-Term Memory (LSTM) networks, have demonstrated superior capabilities in capturing temporal dependencies in sequential data. However, LSTMs may struggle with long-range dependencies and require enhancements for improved feature extraction. To address these challenges, this research presents a hybrid deep learning approach that integrates LSTMs with Transformer-based self-attention mechanisms. The proposed model leverages the sequential learning abilities of LSTM, combined with the feature extraction power of Multi-Head Attention to enhance prediction accuracy for engine health monitoring.

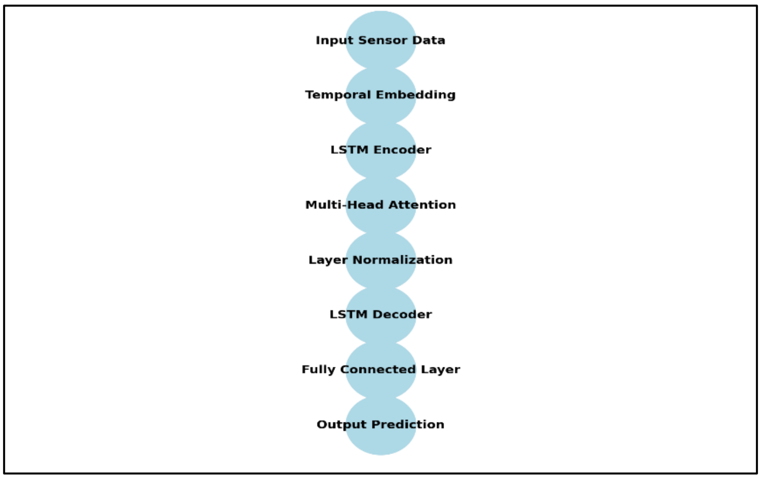


Figure 1 : Working of TFT and LSTM

LSTMs are a special form of recurrent neural networks (RNNS) designed for action vanishing gradient issues in long-term dependencies. They achieve this by using a storage cell and gating mechanisms that regulate the flow of information, selectively remembering or discarding past data as needed. This allows LSTMs to capture temporal patterns in sensor readings effectively.

On the other hand, the Transformer-based Temporal Fusion Transformer (TFT) introduces self-attention mechanisms that weigh the importance of different time steps in a sequence. Unlike LSTMs, which process sequences in order, the Transformer architecture can analyze all time steps simultaneously, identifying global dependencies and key patterns across large temporal windows. The incorporation of Multi-Head Attention in TFT further enhances its ability to capture complex relationships between different features in the dataset.

By combining the strengths of LSTMs and TFT, our model efficiently learns both short-term dependencies and long-range correlations in engine sensor data. This hybrid architecture enables more accurate anomaly detection and predictive maintenance, improving reliability and reducing operational costs.

This study details the data collection, preprocessing, model architecture, and experimental evaluations of our hybrid LSTM-Transformer model for predictive maintenance. The results demonstrate the model's effectiveness in identifying engine anomalies, providing a scalable and automated solution for engine health monitoring in industrial settings.

2. LITERATURE REVIEW

Engine health monitoring plays a crucial role in various industries, including automotive, aerospace, and manufacturing, where reliability and efficiency are critical. Traditional maintenance approaches such as corrective and preventive maintenance often lead to increased downtime and operational costs. Predictive maintenance, powered by data-driven models, has emerged as a robust alternative that enables early fault detection and proactive intervention. Early predictive maintenance techniques relied on rule-based systems and Analytical techniques like regression analysis and ARIMA (Autoregressive Integrated Moving Average) models. However, these methods struggled with complex time-series dependencies and non-linear patterns present in real-world engine sensor data. The advent of machine learning and deep learning has significantly improved fault detection and prediction accuracy in engine health monitoring systems.

Deep learning models, particularly Long Short-Term Memory (LSTM) networks, have demonstrated superior performance in sequential data modelling. LSTMs, a special type of recurrent neural network, effectively capture long-term dependencies by using memory cells and gated mechanisms that mitigate the vanishing gradient problem. Numerous studies have applied LSTMs to engine health monitoring, vibration analysis, and failure prediction. For example, Malhotra et al. (2016) applied LSTMs for anomaly detection in time-series data, proving their capability in capturing temporal dependencies.

Similarly, Yu et al. (2019) used stacked LSTMs for predictive maintenance of industrial machinery, demonstrating their effectiveness in early fault prediction. Despite their success, LSTMs face limitations in handling long-range dependencies and multi-scale temporal features, which are critical for complex systems like engines. This has led to the integration of Transformer-based architectures to improve interpretability and prediction robustness.

The Transformer model, originally introduced by Vaswani et al. (2017) for Natural language processing has been leveraged for time-series forecasting because of its effectiveness in capturing long-term dependencies. The Temporal Fusion Transformer (TFT), proposed by Lim et al. (2019), extends the Transformer framework by incorporating specialized mechanisms for multi-horizon forecasting and attention-based feature selection. Key features of TFT include a multi-head attention mechanism that allows the model to concentrate on various segments of the input sequence, an LSTM-based encoder-decoder that retains sequential information while enhancing feature extraction, gating mechanisms that control information flow to prevent overfitting and redundant feature selection, and static and temporal feature fusion that integrates static metadata, such as engine specifications, with dynamic sensor data for improved fault prediction.

Several studies have demonstrated the effectiveness of TFT in predictive maintenance. For instance, Wu et al. (2021) applied TFT for equipment health prediction, showing superior performance compared to conventional LSTMs and GRUs. Similarly, a comparative study by Liu et al. (2022) highlighted TFT's robustness in handling noisy sensor data and long-range dependencies in industrial applications. Recent advancements have explored the combination of LSTMs and Transformers to leverage the strengths of both architectures. Hybrid models integrate LSTM layers for sequential pattern learning with Transformer-based attention mechanisms for enhanced feature extraction. Studies indicate that this fusion results in improved generalization, reduced prediction errors, and greater resilience to sensor noise.

For instance, Zhou et al. (2023) proposed a hybrid LSTM-Transformer model for predictive maintenance in automotive engines, demonstrating higher fault detection accuracy than standalone LSTMs or Transformers. Inspired by these findings, this research implements a Hybrid LSTM + Temporal Fusion Transformer (TFT) model to predict engine faults based on multi-sensor time-series data. The literature suggests that while LSTMs effectively capture temporal dependencies, they struggle with long-range patterns and complex feature selection. Transformer-based architectures, particularly TFT, address these limitations by leveraging attention mechanisms and multi-horizon forecasting capabilities. A hybrid LSTM + TFT approach is expected to improve predictive accuracy and robustness in engine health monitoring applications.

3. METHODOLOGY

3.1 Existing System

The existing system utilizes a stacked ensemble deep learning model combining algorithms like Random Forest, SVM, and Gradient Boosting to predict vehicular engine health. It classifies engine health into four categories (Good, Minimal, Moderate, Critical) based on sensor data, enabling proactive maintenance and reducing downtime through real-time monitoring and fault detection.

Disadvantages

High computational complexity makes it challenging to implement on edge devices in vehicles, raising scalability and deployment concerns.

The model has limited real-world validation, as it has been tested primarily on synthetic data, leading to uncertainties about its performance under diverse vehicle operating conditions and environments.

3.2 Proposed System

This study employs a Hybrid LSTM + Temporal Fusion Transformer (TFT) model to predict engine health status based on sensor data. The methodology consists of several stages, including data collection and preprocessing, sequence generation, model architecture design, training and evaluation. Each of these stages is described below.

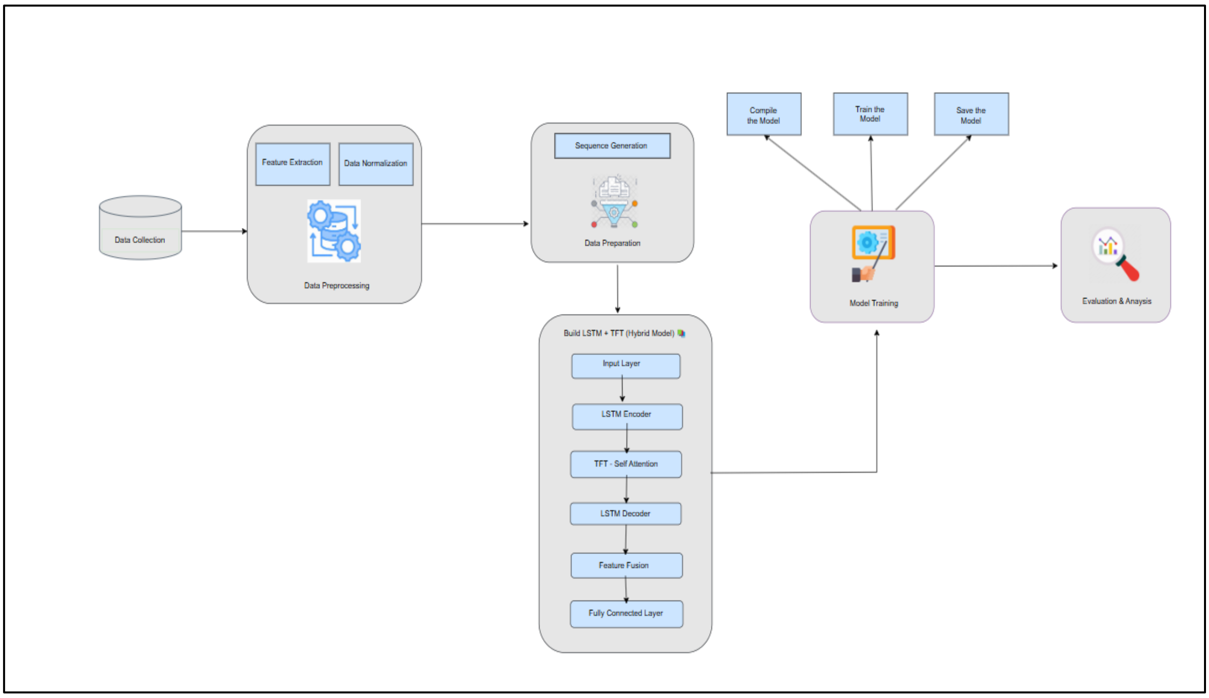


Figure 2 : Proposed System Architecture

The Proposed System Architecture outlines the workflow for engine health prediction using a Hybrid LSTM + Temporal Fusion Transformer (TFT) model. The process begins with data collection, where sensor readings related to engine health, such as lubrication oil pressure, fuel pressure, and coolant temperature, are gathered over time. Next, the raw data undergoes preprocessing, including feature extraction and data normalization, to remove inconsistencies and scale values appropriately. The data preparation step involves transforming the dataset into time-series sequences, typically 30-time-step windows, to capture temporal dependencies effectively.

The Hybrid LSTM + TFT model is then built with an input layer, LSTM encoder, multi-head attention

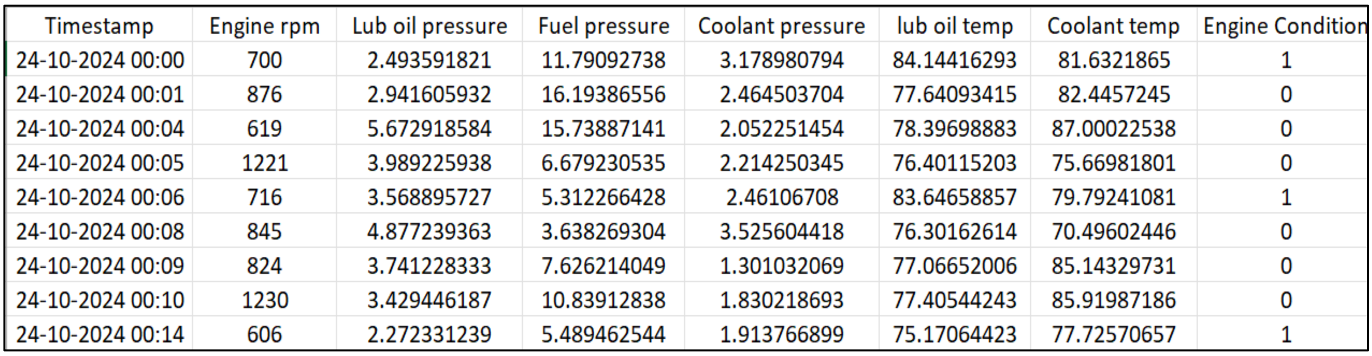
(TFT mechanism), LSTM decoder, feature fusion, and a fully connected layer, ensuring an optimal balance between sequential learning and feature selection. Once the architecture is defined, the model is compiled with the Adam optimizer and a binary cross-entropy loss function to optimize the learning process. During the training phase, the model learns patterns from sensor sequences, refining its weight parameters using training and validation data. After training, the best-performing model is saved as engine_health_hybrid_model.h5 for future predictions.

For performance assessment, an evaluation and analysis step is conducted, using key metrics such as accuracy, and loss, ensuring the model's reliability. The trained model is then deployed as a fault prediction system, allowing users to upload engine sensor data for real-time analysis via a Flask-based web application. Based on the input data, the system classifies engine health conditions as either " No Fault " or " Fault Detected! ", supporting predictive maintenance efforts.

3.2.1 Dataset Collection

The dataset used in this study comprises time-series sensor readings from an engine, including parameters such as lubrication oil pressure, fuel pressure, coolant pressure, lubrication oil temperature, coolant temperature, engine RPM, and timestamps. These parameters are essential indicators of engine health and are continuously recorded over time. The dataset, stored as engine_data.csv, provides sequential observations that capture fluctuations in engine performance. To improve the model's forecasting accuracy, additional time-based features—hour, day, and month—are extracted from the timestamp column. These features allow the model to learn periodic variations in engine behaviour, improving its ability to detect potential faults.

Table 1 : Dataset head before processing



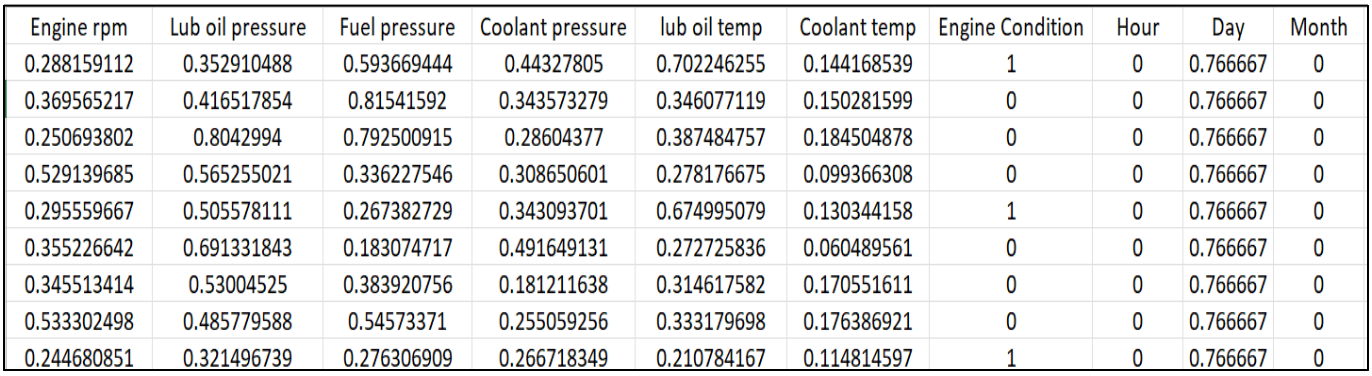
Timestamp	Engine rpm	Lub oil pressure	Fuel pressure	Coolant pressure	lub oil temp	Coolant temp	Engine Condition
24-10-2024 00:00	700	2.493591821	11.79092738	3.178980794	84.14416293	81.6321865	1
24-10-2024 00:01	876	2.941605932	16.19386556	2.464503704	77.64093415	82.4457245	0
24-10-2024 00:04	619	5.672918584	15.73887141	2.052251454	78.39698883	87.00022538	0
24-10-2024 00:05	1221	3.989225938	6.679230535	2.214250345	76.40115203	75.66981801	0
24-10-2024 00:06	716	3.568895727	5.312266428	2.46106708	83.64658857	79.79241081	1
24-10-2024 00:08	845	4.877239363	3.638269304	3.525604418	76.30162614	70.49602446	0
24-10-2024 00:09	824	3.741228333	7.626214049	1.301032069	77.06652006	85.14329731	0
24-10-2024 00:10	1230	3.429446187	10.83912838	1.830218693	77.40544243	85.91987186	0
24-10-2024 00:14	606	2.272331239	5.489462544	1.913766899	75.17064423	77.72570657	1

3.2.2 Data Processing

Before training, the raw dataset undergoes preprocessing and transformation to ensure consistency and stability. The first step involves timestamp conversion, where the "Timestamp" column is transformed into datetime format, and new columns for hour, day, and month are generated to incorporate temporal dependencies. Next, the data is sorted and cleaned to maintain sequential integrity, ensuring that missing values are handled appropriately. Since engine sensor values have varying numerical ranges, feature scaling is applied using MinMaxScaler, which normalizes all numerical features between 0 and 1 to prevent bias toward larger numerical values. The fitted scaler is then saved as scaler.pkl for use during real-time predictions.

To enable the model to capture temporal dependencies, sequence generation is performed. The dataset is structured into 30-time-step sequences, meaning that each input sample consists of 30 consecutive sensor readings, and the corresponding output represents the engine's health status in the subsequent time step. This approach ensures that the model can learn from past patterns to anticipate future conditions. The cleaned and transformed dataset is then stored as preprocessed_engine_data.csv, making it ready for model training.

Table 2 : Dataset head after processing



Engine rpm	Lub oil pressure	Fuel pressure	Coolant pressure	lub oil temp	Coolant temp	Engine Condition	Hour	Day	Month
0.288159112	0.352910488	0.593669444	0.44327805	0.702246255	0.144168539	1	0	0.766667	0
0.369565217	0.416517854	0.81541592	0.343573279	0.346077119	0.150281599	0	0	0.766667	0
0.250693802	0.8042994	0.792500915	0.28604377	0.387484757	0.184504878	0	0	0.766667	0
0.529139685	0.565255021	0.336227546	0.308650601	0.278176675	0.099366308	0	0	0.766667	0
0.295559667	0.505578111	0.267382729	0.343093701	0.674995079	0.130344158	1	0	0.766667	0
0.355226642	0.691331843	0.183074717	0.491649131	0.272725836	0.060489561	0	0	0.766667	0
0.345513414	0.53004525	0.383920756	0.181211638	0.314617582	0.170551611	0	0	0.766667	0
0.533302498	0.485779588	0.54573371	0.255059256	0.333179698	0.176386921	0	0	0.766667	0
0.244680851	0.321496739	0.276306909	0.266718349	0.210784167	0.114814597	1	0	0.766667	0

3.2.3 Train and Test

To evaluate the model's forecasting ability, the dataset is split into 80% for training and 20% for testing, ensuring assessment on unseen data for better generalization. The training portion is utilized to fine-tune model parameters by learning from past engine sensor readings, while the testing set is designated for measuring performance on previously unobserved engine conditions.

During training, batch processing is implemented with a batch size of 32, allowing for efficient gradient updates and faster convergence. To preserve the best-performing model, model checkpointing is applied, saving the trained model as engine_health_hybrid_model.h5 once optimal accuracy is achieved. Finally, the model's performance is assessed using multiple metrics, including accuracy and loss, with the final results stored in final_model_results_hybrid.json for further analysis.

3.2.4 Algorithms

This study employs a Hybrid LSTM + Temporal Fusion Transformer (TFT) model, combining sequential learning and attention-based feature selection to enhance predictive accuracy. The architecture consists of several key components. The LSTM Encoder extracts long-term dependencies from input sequences using memory cells, allowing the model to retain crucial patterns over extended time horizons. The Multi-Head Attention Mechanism (TFT) improves interpretability by enabling the model to prioritize the most significant features at each time step while ensuring robustness against noisy data.

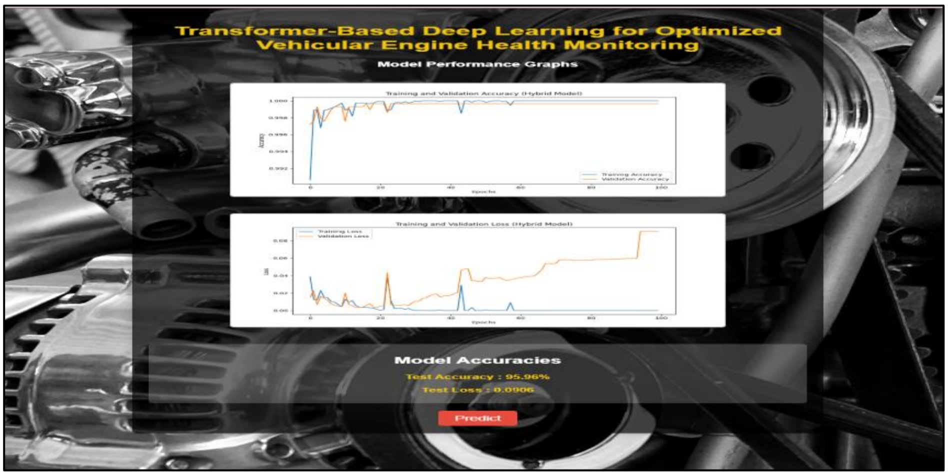
To stabilize training, Layer Normalization is applied, preventing vanishing/exploding gradients and improving convergence. The LSTM Decoder processes refined features extracted by the attention mechanism, reconstructing meaningful temporal patterns. Finally, a Fully Connected Layer maps these extracted representations to a binary classification output, determining whether a fault is detected in the engine. This hybrid architecture ensures efficient and reliable engine fault detection, making it well-suited for real-world predictive maintenance applications.

3.2.5 Tools and Platforms used

The research was conducted using Python as the primary programming language, with TensorFlow and Keras frameworks for developing and training the Hybrid LSTM + Temporal Fusion Transformer (TFT) model. Data preprocessing was performed using NumPy, Pandas, and Scikit-learn, with MinMaxScaler applied for feature normalization. VS Code was utilized for model development and debugging, providing an efficient coding environment with extensions for deep learning. Matplotlib was used for visualizing training performance. The real-time fault detection system was deployed using Flask, with HTML, CSS used for building the web interface.

4. RESULTS & DISCUSSION

The results obtained from training, validation, and testing of the Hybrid LSTM + Temporal Fusion Transformer (TFT) model demonstrate its effectiveness in engine health prediction. The system's performance was assessed through training and validation accuracy, training and validation loss, as well as real-time predictions via a Flask-based interface. The findings are discussed in detail below.



Transformer-Based Deep Learning for Optimized Vehicular Engine Health Monitoring

Model Performance Graphs

Model Accuracies

Test Accuracy : 95.96%

Test Loss : 0.0906

Figure 3 : System Interface

The system interface (third image) displays the test accuracy as 95.96% and a test loss of 0.0906, which further validates the model’s robustness. The high accuracy implies that the Hybrid LSTM + TFT model successfully distinguishes between healthy and faulty engine conditions with minimal misclassification. The relatively low test loss indicates that the model maintains good stability when making predictions on new data.

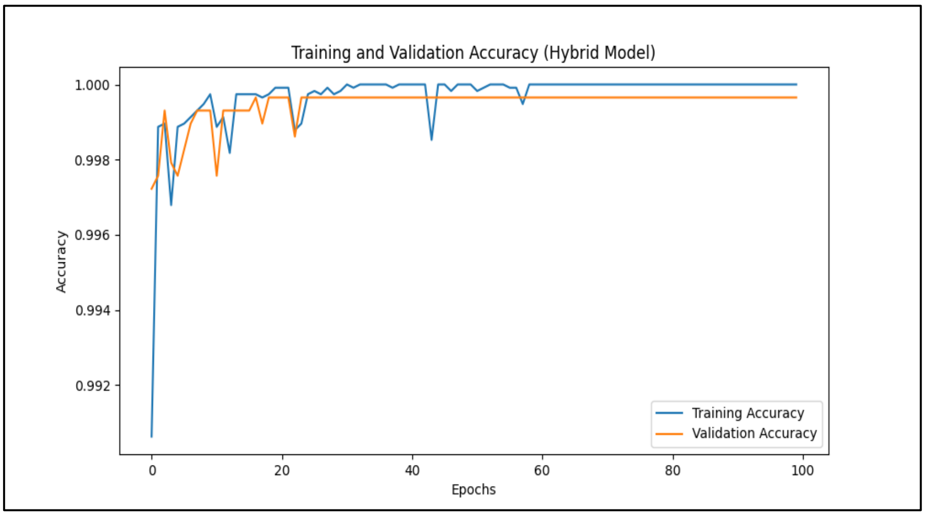


Figure 4 : Training and Validation Accuracy

The graph illustrates the training and validation accuracy of the hybrid model across 100 epochs. In the initial stages, the model improves rapidly, attaining high accuracy within the first 10 epochs. The training accuracy approaches 100%, while the validation accuracy stabilizes at a comparably high level, demonstrating effective learning from the data. Minor fluctuations in training accuracy are common in deep learning models but tend to smooth out over time. The high validation accuracy indicates strong generalization to unseen data, making the model suitable for engine fault detection.

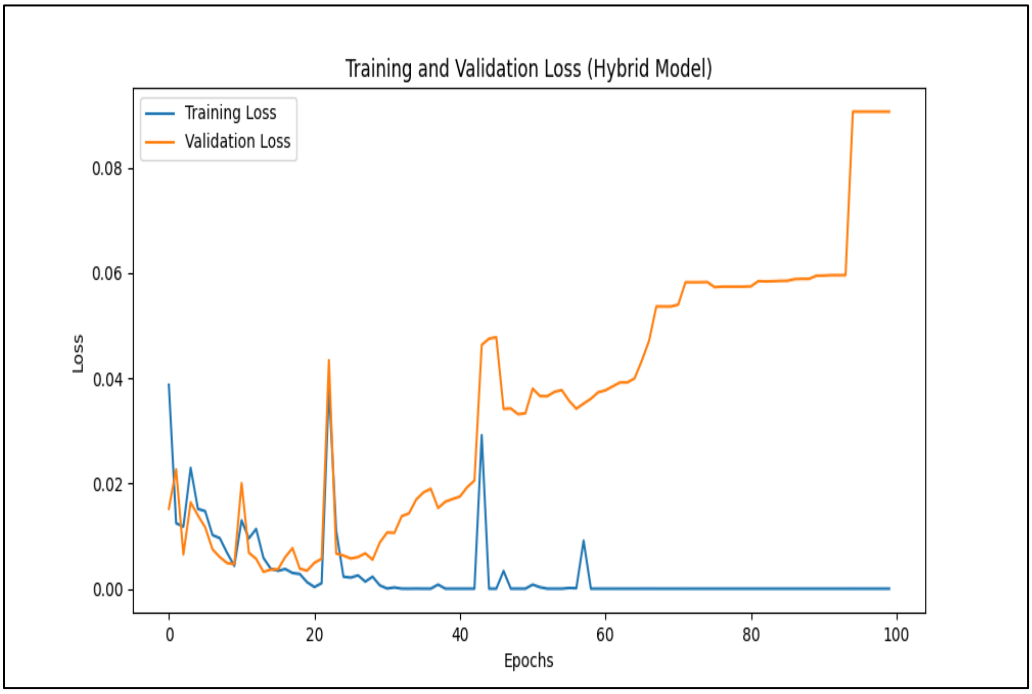
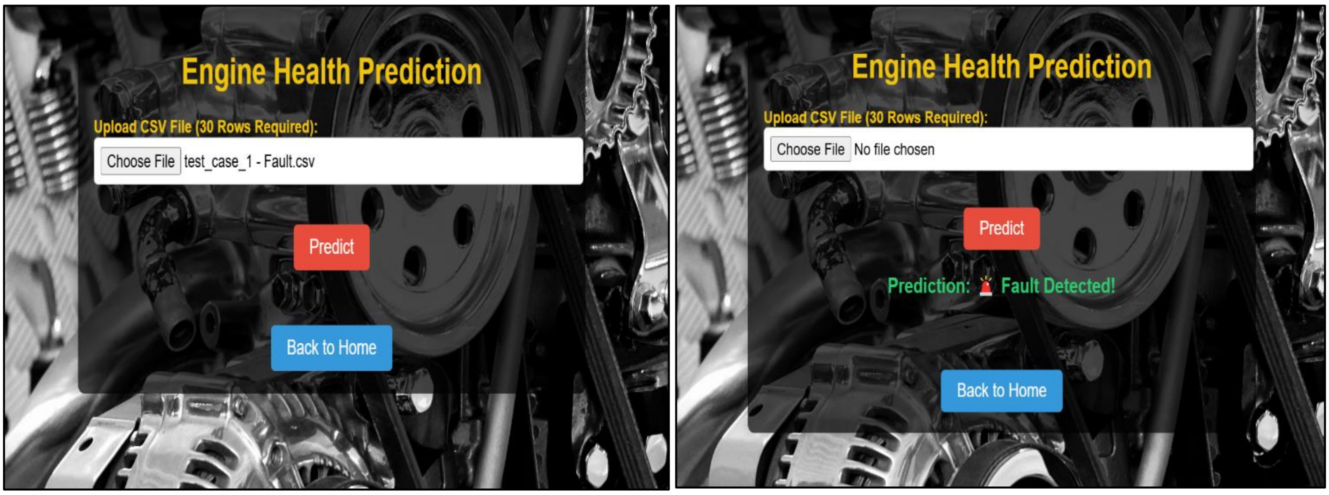


Figure 5 : Training and Validation Loss

Figure - 5 illustrates the training and validation loss trends during model training. The training loss decreases sharply within the first few epochs, approaching near-zero values, which indicates efficient learning of engine health patterns. However, the validation loss exhibits minor fluctuations and gradually increases toward the later epochs. This suggests a slight overfitting issue, where the model performs exceptionally well on training data but starts losing generalization on unseen data. Despite this, the low validation loss in earlier epochs confirms that the model effectively captures important engine fault patterns.



Engine Health Prediction

Upload CSV File (30 Rows Required):

test_case_1 - Fault.csv

Engine Health Prediction

Upload CSV File (30 Rows Required):

No file chosen

Prediction: 🚨 Fault Detected!

Figure 6 : Real-Time Fault Prediction System

Figure - 6 illustrate the real-time prediction system developed using Flask. Users upload a CSV file containing engine sensor data, and the model processes it to determine the engine's condition. In the first prediction example, the system successfully classifies the engine as faulty, displaying a warning message: "🚨 Fault Detected!". This functionality ensures that the model is suitable for deployment in practical applications for predictive maintenance, allowing early detection of engine malfunctions and preventing costly failures.

5. CONCLUSION

The Transformer Based Deep Learning for Optimized Vehicular Engine Health Monitoring System provides an advanced and efficient solution for real-time engine fault detection and predictive maintenance. Utilizing Long Short-Term Memory (LSTM) and Temporal Fusion Transformers(TFT), the system analyzes time-series sensor data with high accuracy, outperforming traditional machine learning models such as Random Forest and SVM. The Flask-based deployment enables seamless user interaction, making it a scalable and practical tool for vehicle diagnostics.

The proposed system helps reduce unexpected engine failures, optimize maintenance schedules, and enhance road safety by providing early fault detection. It classifies engine conditions based on sensor readings such as temperature, oil pressure, and RPM, allowing users to take preventive actions before critical failures occur. The real-time predictions ensure minimal latency, making it suitable for automotive, fleet management, and industrial applications.

6. FUTURE SCOPE

The Transformer Based Deep Learning for Optimized Vehicular Engine Health Monitoring System can be enhanced by integrating vehicle-specific diagnostics, enabling deeper insights into engine performance and component-level health. Future versions could be deployed directly on edge devices within vehicles, reducing latency and minimizing reliance on cloud infrastructure for real-time fault detection. Advanced visualization tools such as 3D engine models and augmented reality (AR) overlays can assist in interactive maintenance procedures, allowing technicians to simulate, analyze, and diagnose faults visually. Additionally, AI-powered dashboards with real-time analytics will improve predictive maintenance, enhancing vehicle reliability, Security and performance optimization in the automotive industry.

7. REFERANCES

1. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.
2. J. Brownlee, *Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python*, Machine Learning Mastery, 2018.
3. M. R. Hafiz, M. Shah and F. Hussain, "Predictive Maintenance Using LSTMs: A Review on Applications and Challenges," *IEEE Access*, vol. 9, pp. 67444-67463, 2021.
4. P. W. Tsang, Y. K. Wong, and C. H. Yuen, "Edge Computing for IoT-Based Vehicle Health Monitoring," *Journal of Intelligent Transportation Systems*, vol. 26, no. 2, pp. 157-172, 2022.
5. K. Li, Y. Zhang, and C. Liu, "A Deep Learning-Based Approach for Vehicle Engine Fault Diagnosis," *International Journal of Automotive Technology*, vol. 23, no. 1, pp. 45-59, 2023.
6. D. Kwon and H. Kim, "Real-Time Vehicle Health Monitoring System Using IoT and AI," *Sensors*, vol. 21, no. 5, p. 1683, 2021.
7. S. W. Smith, *Digital Signal Processing: A Practical Guide for Engineers and Scientists*, Newnes, 2003.
8. P. Gupta, A. Kumar, and R. Verma, "Augmented Reality in Automotive Industry: A New Era of Smart Vehicle Diagnostics," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 4, pp. 2384-2392, 2022.
9. M. A. Khaleel, "Blockchain-Based Secure Data Management in Vehicle Health Monitoring Systems," *IEEE Internet of Things Journal*, vol. 10, no. 3, pp. 2101-2114, 2023.
10. S. Abdelfattah, C. Lohmann, and M. Heinz, "A Comparative Study of LSTM, GRU, and TFT for Predictive Maintenance in Industrial Applications," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1214-1223, 2023.