**AAI-530 IoT and Big Data Final Project**

# Maintenance Classification and Maintenance Prediction using Machine Learning and Deep Neural Networks

**Luis Lopez**

**University of San Diego**

**Department of Engineering**

**Instructor:** Roozbeh Sadeghian, Ph.D

## Background

According to the American Transportation Research Institute (ATRI), the tire cost per mile for fleets and independent truck owners averages $0.043. Trucks typically traveling 89,000 miles annually, this translates to an annual tire expense of approximately $3,827 per truck. Additionally, the cost index for tires has been on the rise since 2019, indicating an upward trend in operational expenses related to tire maintenance (American Transportation Research Institute [ATRI], 2021). Tires remain one of the primary maintenance concerns in transportation and logistics, directly tied to the cost of operations on a per-mile basis. Moreover, as truck drivers travel farther distances than they did in 2019, these costs continue to escalate.

Predictive maintenance for connected vehicle fleets relies on vast amounts of data from temperature, infrared, acoustic, vibration, battery-level, and sound sensors. However, IoT enabled maintenance systems face Big Data challenges that impact their efficiency and effectiveness depending on the network size and the number of devices per unit.

Thousands of vehicles transmitting real-time data create major storage and processing challenges. According to Chaudhuri (2018), only 10% of IoT generated data is analyzed, highlighting inefficiencies in data utilization. Additionally, high-velocity data streams require robust pipelines for near real-time processing to ensure timely maintenance and prevent failures. Sensor calibration plays a crucial role, as data collection frequency (per second, minute, or hour) affects processing speed and infrastructure demands.

IoT driven maintenance systems generate structured, semi-structured, and unstructured data, depending on sensor type. For example, live-cam feeds monitoring drivers and accidents require preprocessing before transmission, while temperature and GPS readings may arrive in different formats as well. Data quality is another challenge, faulty sensors, connectivity loss, and terrain variations can introduce missing values and outliers, distorting maintenance predictions. Raw sensor data requires cleaning and transformation before it becomes useful. Real-time vehicle location data in its raw form is useful for dispatch coordination, but other data needs preprocessing for insights.

## Approach

In this report, we analyze an IoT dataset that captures sensor-generated data from a logistics company, along with historical maintenance records and unit-specific information. The dataset provides key insights into the operational health of vehicles, detailing maintenance events, sensor readings, and other contextual information. However, one limitation is the lack of a unique identifier for each unit, which prevents direct tracking of individual maintenance histories. Despite this, the dataset contains a high volume of observations, allowing us to construct a detailed and data-driven analysis of the company’s maintenance practices.

For logistics companies, cost reduction, operational efficiency, and predictive asset management are fundamental to maintaining profitability and ensuring safety. Given this context, this project aims to address two critical questions that directly impact maintenance planning and fleet management:

1. Will a specific unit require maintenance?

- This represents a binary classification task, where the goal is to predict whether a unit will need maintenance in the near future based on available sensor and historical data.
- A reliable predictive model can help fleet managers proactively schedule maintenance, reducing unexpected breakdowns, avoiding costly emergency repairs, and ensuring continuous fleet operation.
- Additionally, preventive maintenance improves driver safety, minimizes hazardous situations on the road, and contributes to a more cost-effective supply chain by reducing downtime.

2. How many maintenance events will occur within a given time window?
    - This question requires predicting a continuous variable in a time-series context, making it a forecasting problem rather than classification.
    - By estimating the expected number of maintenance events over a given period, fleet managers can better allocate resources, optimize scheduling, and manage maintenance budgets more effectively.
    - Understanding maintenance frequency also helps companies implement predictive maintenance strategies, reducing unexpected failures and associated operational disruptions.

To address these questions, we propose a hybrid modeling approach:

- For classification (Question 1): We will build a binary classification model, utilizing machine learning techniques to determine the likelihood of a unit requiring maintenance.
- For time-series forecasting (Question 2): We will implement a Long Short-Term Memory (LSTM) model, a type of recurrent neural network (RNN) known for capturing sequential dependencies, to predict maintenance occurrences based on historical patterns.

By applying data-driven predictive modeling, this project aims to enhance fleet efficiency, reduce maintenance costs, and improve overall logistics operations. A robust maintenance forecasting system will enable the company to transition from reactive to proactive maintenance, ensuring higher reliability, lower operational risks, and improved long-term asset management.

**Dataset**

The dataset consists of 92,000 entries with 26 features, categorized into:

- Vehicle Information: Make and Model, Year of Manufacture, Vehicle Type
- Usage Metrics: Usage Hours, Load Capacity, Actual Load
- Maintenance History: Last Maintenance Date, Maintenance Type, Maintenance Cost
- Sensor Data: Engine Temperature, Tire Pressure, Fuel Consumption, Battery Status, Vibration Levels, Oil Quality, Brake Condition
- Anomaly Detection & Predictive Scores: Failure History, Anomalies Detected, Predictive Score
- Environmental & Operational Conditions: Weather Conditions, Road Conditions, Delivery Times

- Operational Efficiency Metrics: Downtime Maintenance, Impact on Efficiency
- Target Variable: Maintenance Required
- Engineered Target: Maintenance Count

The Exploratory Data Analysis conducted on this logistics fleet maintenance dataset provided valuable insights into vehicle maintenance trends, cost distributions, and operational inefficiencies. The dataset contained various features, including vehicle information, usage metrics, maintenance history, sensor data, and operational conditions. A critical limitation identified during the initial review was the lack of a unique identifier for each unit, which prevented direct tracking of individual maintenance histories. However, by leveraging the dataset's high volume of observations, we were able to extract meaningful patterns in maintenance operations.

## Data Cleaning and Preprocessing

The dataset was well-structured, with no missing values (until some features were engineered) requiring imputation. However, adjustments were made to improve data usability, such as converting date fields into a datetime format and encoding categorical variables. Redundant features, including vehicle-specific IDs that did not contribute to predictive analysis, were removed. Additionally, Failure History, Anomalies Detected, and Maintenance Required were found to be highly correlated, leading to their consolidation into a single feature for more streamlined analysis.

## Key Findings from EDA

Mostly the vehicles in the dataset were manufactured between 2005 and 2022, with most falling within the 2015-2022 range with an Average usage per vehicle of 2,989 hours, and with some vehicles accumulating more than 36,000 hours of operational time. The Older vehicles tended to have higher maintenance costs and downtime, confirming that aging fleets require more frequent servicing.

The Maintenance Type and Cost Analysis showed that Oil changes and tire rotations were the most frequent maintenance types, while engine overhauls had the highest average cost. Seasonal spikes in maintenance costs were observed, particularly in the last quarter of each year, suggesting pre-winter servicing trends. The average maintenance cost per event was $1,043, with some maintenance events exceeding $5,999, highlighting the need for better budgeting and cost control.

When analyzing the Operational Efficiency and Downtime, we see that Overloaded vehicle (carrying more than their rated capacity) had slightly higher downtime but no significant impact on maintenance costs which is an interesting observation when it comes to cost, it is hard to see why but further analysis can promote other findings. A hypothesis could be that regular maintenance helps to avoid higher cost but since the units are overloaded, it requires more time to load or onload without affecting maintenance.

The Battery status was inversely correlated with maintenance downtime, indicating that vehicles with well-maintained electrical systems experienced fewer disruptions. Lastly Road conditions played a significant role in maintenance needs, with vehicles operating in rural or mixed terrain experiencing more wear and tear compared to those primarily on highways.

## Importance of These Insights

Understanding these patterns is essential for optimizing fleet maintenance strategies. By identifying high-cost maintenance events, companies can prioritize preventive maintenance and reduce unexpected breakdowns. Seasonal spikes in maintenance costs indicate the importance of preemptive service scheduling, while correlations between road conditions and downtime help optimize vehicle routing decisions. Additionally, the relationship between battery health and maintenance downtime suggests that monitoring electrical systems proactively could significantly enhance fleet efficiency.

In conclusion, this EDA has provided critical operational insights that can be used to improve maintenance scheduling, reduce costs, and enhance vehicle longevity. These findings emphasize the importance of data-driven decision-making in fleet management, allowing logistics companies to move from reactive maintenance to a more predictive and preventive approach.

## Machine and Deep Learning

When building the machine learning models, it was necessary to prepare the data to ensure maximim results when doing the binary classification. The models utilized int his project was Logistics regression, XGBoost, Random Forest, HGBoost classifier and lastly a Deep Learning model for classification. Each of this models we carefully chose due to their unique mathematical structure for classification. Through using SMOTE to balance the class and normalizing the data, the results of each algorithm is as follows:

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Logistics Regression | 86% | 78% | 79% | 78% |
| XGBoost | 80% | 78% | 79% | 78% |
| Random Forest | 87% | 83% | 80% | 81% |
| HGBoost | 78% | 90% | 78% | 81% |
| Deep Learing Model | 78% | 89% | 78% | 81% |

**All averages are weighted**

A Deep Neural Network classifier was constructed to further enhance classification performance. The model consisted of seven fully connected Dense layers, totaling 4,036,251 parameters. The architecture was designed to optimize learning stability while preventing overfitting.

**DNN Architecture and Regularization**

- **Leaky ReLU Activation Function:**
  - Applied in all layers with alpha = 0.1, mitigating the dying neuron problem common in standard ReLU.
  - Improved model convergence and training stability.
- **L2 Regularization (λ = 0.001)**:
  - Applied to all layers to reduce overfitting and improve generalization.
- **Batch Normalization:**
  - Implemented after each layer to accelerate training and stabilize learning curves.

- **Dropout (0.50):**
  - A 50% dropout rate was applied across all layers to prevent overfitting and enhance robustness.

**Optimization and Training Strategy**

- Dynamic Learning Rate Adjustment (ReduceLROnPlateau):
- Adjusted learning rate when validation loss plateaued.
  - Patience = 3, with a reduction factor of 0.9, ensuring gradual optimization.
- **Early Stopping Callback:**
  - Monitored validation loss, stopping training after 10 epochs of no improvement.
  - Restored best weights to prevent overfitting.
- **Adam Optimizer (lr = 0.001)**:
  - Chosen for its adaptive learning rate properties, improving convergence speed.
- **Binary Crossentropy Loss Function:**
  - Standard for binary classification tasks, ensuring a stable loss landscape.

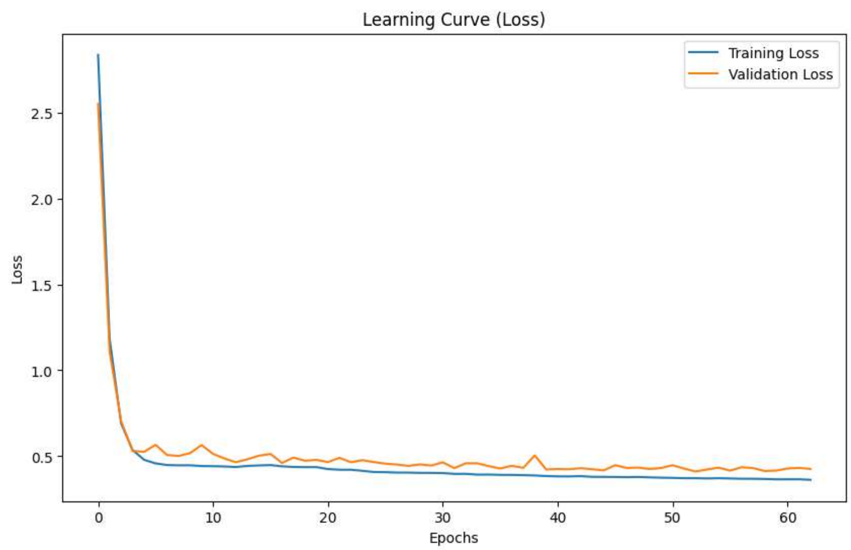
**DNN Training Performance**

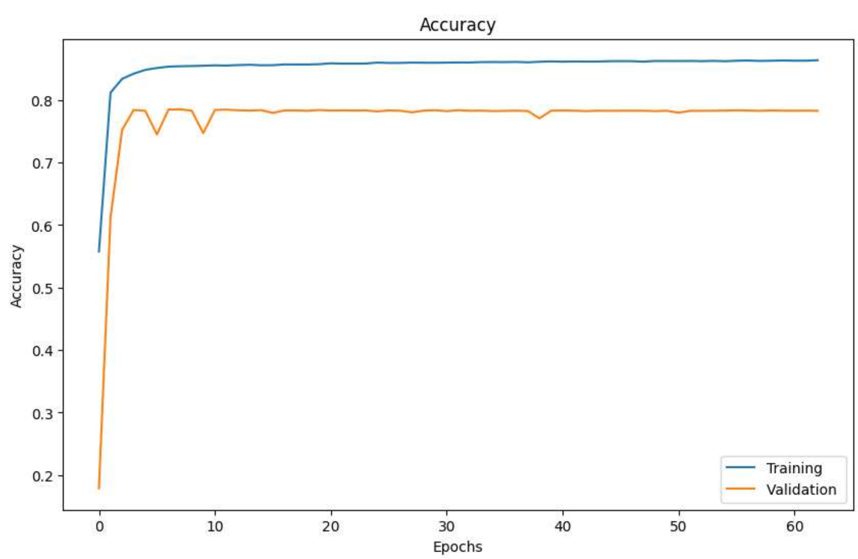
The model was trained using:

- 100 epochs
- Batch size = 300

- Validation data from unseen test samples

The results demonstrated superior performance compared to traditional ML models, showing an increase in recall and overall predictive accuracy:



## LSTM

Before constructing the Long Short-Term Memory model for time-series forecasting, several crucial preprocessing steps were performed to ensure the data was structured correctly for sequential learning. These steps included data preparation, feature engineering, scaling, and sequence generation, all of which were essential for optimizing the model's predictive capabilities.

The dataset contained maintenance history, operational conditions, and vehicle sensor readings, all of which were organized chronologically to maintain temporal integrity. Since LSTM models require a properly formatted time-series, the unix timestamp feature was converted into a datetime format, and the dataset was reindexed by date to ensure continuity.

To improve the model's ability to capture maintenance patterns, feature selection and engineering were applied. The target variable, maintenance count, represented the number of maintenance events per day. To smooth short-term fluctuations, a 7-day moving average was applied to maintenance count.

LSTM models perform best when input features are within a consistent range, so MinMax Scaling was applied to normalize values between 0 and 1. This ensured that large variations in maintenance counts did not cause instability in training. Additionally, since LSTMs require sequential data, a sliding window approach was used to generate training sequences. Each input sequence consisted of the past 30 days of maintenance data (X), while the target output (y) represented the maintenance count on day 31. This allowed the model to learn from historical trends and make informed predictions about future maintenance needs.

To evaluate the model's performance, the dataset was split into training and testing sets, preserving the chronological order of events to prevent data leakage. Approximately 70% of the data was used for training, allowing the model to learn maintenance patterns, while the remaining 30% was reserved for testing, ensuring robust evaluation on unseen data. This approach enabled the LSTM model to generalize well and accurately predict future maintenance requirements.

These preprocessing steps were essential in preparing the dataset for time-series forecasting. By implementing time-based feature engineering, sequence generation, and data normalization, the model was equipped to capture long-term maintenance trends, making it a valuable tool for predicting servicing needs and optimizing fleet maintenance strategies.

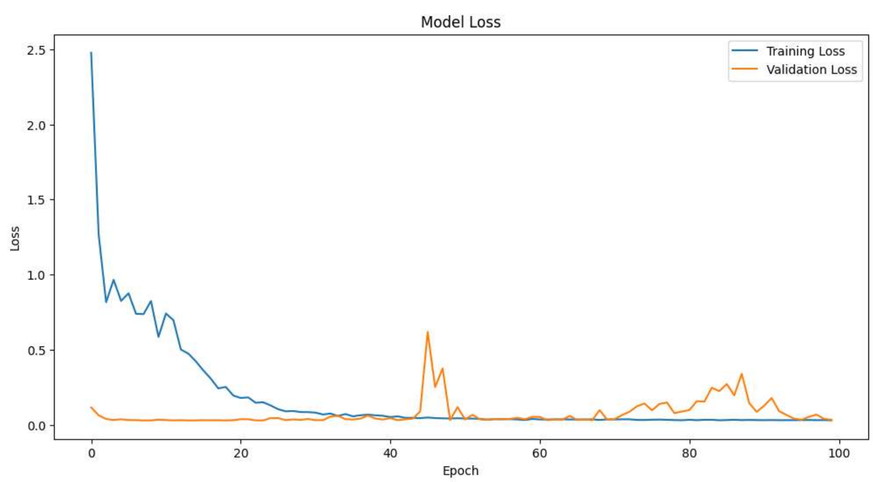
The architecture of the LSTM was fairly simple due to the size of the aggregated data of the number of services in a given day. Although it is a simple model there were severe issues with overfitting, therefore regularization and tuning was fundamental. The model's summary is as follows:

| Layer (type) | Output Shape | Param # |
|---|---|---|
| lstm_2 (LSTM) | (None, 30, 560) | 1,258,880 |
| batch_normalization_8 (BatchNormalization) | (None, 30, 560) | 2,240 |
| dropout_8 (Dropout) | (None, 30, 560) | 0 |
| lstm_3 (LSTM) | (None, 230) | 727,720 |
| batch_normalization_9 (BatchNormalization) | (None, 230) | 920 |
| dropout_9 (Dropout) | (None, 230) | 0 |
| dense_8 (Dense) | (None, 1) | 231 |

**Total params:** 1,989,991 (7.59 MB)

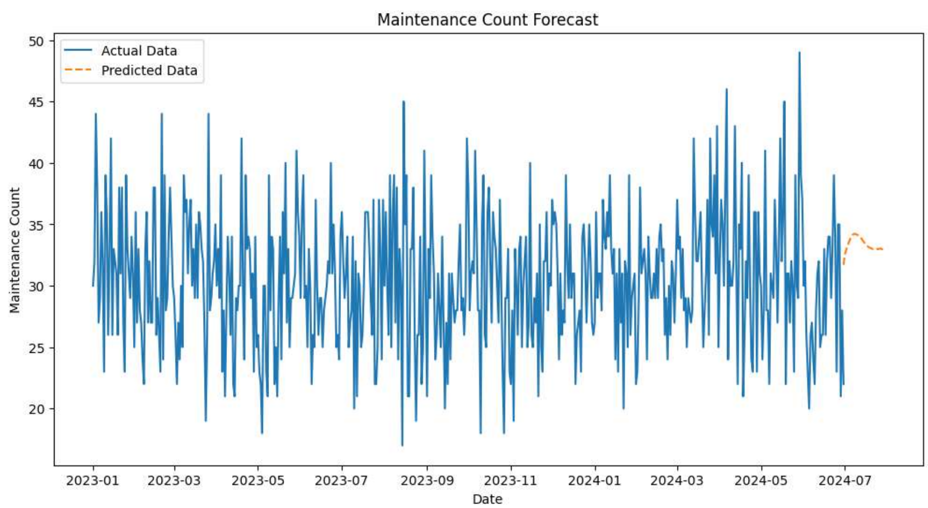
**Trainable params:** 1,988,411 (7.59 MB)

The Model was trained on a learning rate of .0005 at 100 epochs with a batch size of 20 using an Adam optimizer and using the Mean Square Error(MSE) function.



## Results

After training we tested the model to predict the next 30 days of maintenance. From the plot we can see that the model attempts to follow the fluctuation of the plot and stays near the average. This could indicate that the model is attempting to use the average as a benchmark for prediction instead of using the temporal dependence of the data to predict accurately. Having said that the prediction looks within a range that is not abnormal.




## Future Work

The widespread adoption of Internet of Things technology has fundamentally transformed the way we manage society, businesses, and even households. In the context of maintenance, various strategies can be employed depending on the application, including run-to-failure (R2F), preventive maintenance, and predictive maintenance (Dash et al., 2021). The framework utilized in this project presents multiple alternatives for enhancing the predictive capability of the Long Short-Term Memory (LSTM) model. One such approach involves generating synthetic data from the original dataset using Variational Autoencoders (VAEs) and Time Series Generative Adversarial Networks (TimeGANs).

TimeGANs, in particular, offer significant advantages due to their ability to capture and replicate temporal dependencies within time-series data, thereby enhancing the variability and richness of the training dataset (Yoon et al., 2019). By generating realistic synthetic sequences that reflect the statistical properties of the original dataset, these models can help mitigate issues related to limited sample sizes or data imbalance. Once the enriched dataset is created, it can be fed into a more robust LSTM architecture, which has been demonstrated to outperform other models such as Support Vector Machines (SVMs) and Deep Neural Networks (DNNs) in predictive maintenance tasks (De Vita & Bruneo, 2019).

By incorporating these advanced data augmentation techniques, the predictive model can achieve higher accuracy and reliability in forecasting the number of units requiring maintenance within a given time window. This improvement enables logistics and fleet management companies to optimize resource allocation, reduce unexpected failures, and improve overall operational efficiency.

References

American Transportation Research Institute. (2021). *An analysis of the operational costs of trucking: 2021 update.* Retrieved from https://www.truckingresearch.org/

Chaudhuri, A. (2018). *Predictive Maintenance for Industrial IoT of Vehicle Fleets using Hierarchical Modified Fuzzy Support Vector Machine.* Samsung R&D Institute Delhi.

D. Bruneo and F. De Vita, "On the Use of LSTM Networks for Predictive Maintenance in Smart Industries," 2019 IEEE International Conference on Smart Computing (SMARTCOMP), Washington, DC, USA, 2019, pp. 241-248, doi: 10.1109/SMARTCOMP.2019.00059

Dash, S. K., Raj, S., Agarwal, R., & Mishra, J. (2021). *Automobile predictive maintenance using deep learning*. *International Journal of Artificial Intelligence and Machine Learning, 11*(2). https://doi.org/10.4018/IJAIML.20210701.oa7

Yoon, J., Jarrett, D., & van der Schaar, M. (2019). *Time-series generative adversarial networks*. Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada. Retrieved from