Research Article

Fusion of Neural Networks and Logistic Regression for Predictive Maintenance of Vehicle Engines

Jayden P. Chen

Diamond Bar High School, 21400 Pathfinder Road, Diamond Bar, CA, 91765, United States

ABSTRACT

The rise of predictive maintenance models has revolutionized vehicle maintenance, promising significant improvements in performance and lifespan. This research aims to develop a robust predictive maintenance model for automotive engines using a hybrid approach that combines neural networks and logistic regression models. By analyzing patterns within a publicly available dataset of 19,503 engine cases, which includes features such as engine rotations per minute, temperatures, and pressures, the study trains hybrid models to predict when a vehicle requires maintenance.

The methodology involves preprocessing the dataset, training individual models, and integrating them within a stacked ensemble framework. Neural networks are leveraged for their complex pattern recognition capabilities, and logistic regression models offer interpretability and simplicity. Metrics such as accuracy, precision, and recall evaluate the models' performance.

The ultimate goal is to enable vehicle owners and mechanics to address potential issues proactively, ensuring better vehicle performance and extending engine lifetimes. The hybrid models show enhanced success compared to traditional models, providing potential contributions to predictive analytics, and a new standard for various industries.

Keywords: Logistic regression, Neural Network, Predictive Analytics, Hybrid Models, Vehicle Engines

INTRODUCTION

As is well known, the vehicle has played an essential role in almost every aspect of our daily life for different transportation needs, including education, grocery shopping, work, entertainment, and so on. Unfortunately, similar to most other manufactured products, the vehicle has a certain lifespan with influential factors containing usage frequency, material quality, driving pattern, etc. Amongst all these factors, the maintenance may weigh in with a large extent for the extension of the vehicle's life. In the past, different vehicle maintenance practices have been employed, ranging from relatively simple manual ones to more sophisticated ones due to the increasing complexity of the associated system (1). Usually, the Vehicle manufacturers will provide detailed maintenance schedules, recommend regular check-ups, and predict the vehicle's performance and longevity.

The accuracy prediction of vehicle maintenance is essential in improving vehicle performance for several

Corresponding author: Jayden P. Chen, E-mail: jach14444@gmail.com.

Received July 25, 2024**; Accepted** August 10, 2024
https://doi.org/10.70251/HYJR2348.226472

benefits, such as early detection of issues, optimal component functioning, and minimized downtime, thus ensuring that the vehicle remains operational and performs reliably. Nonetheless, the traditional preventative methods have proven inconsistent and sometimes biased by personal judgment, mainly due to the need for more understandable data. The expected consequences include unexpected downtime, higher repair costs, extensive safety risks, and inefficient resource use (2). Therefore, the need for more advanced, data-driven maintenance strategies is urgent.

Many modeling-related methods have been proposed with the current advances in data collection and analysis technologies. Some of them use statistical models like linear regression (3), data mining (4, 5, 6), and machine learning-related technologies (7, 8, 9). However, different types of methods are associated with different strengths and weaknesses. For example, while statistical modeling has strengths in hypothesis testing, interpretability, and theoretical foundation, it often needs more flexibility, scalability, and performance than data mining and machine learning techniques. Also, while data mining excels in discovering patterns and relationships within large datasets, it needs to improve in theoretical grounding, interpretability, risk of overfitting, and integration complexity compared to statistical modeling and machine learning. Similarly, machine learning is highly effective for predictive modeling and handling complex, high-dimensional data. However, interpretability, theoretical grounding, risk overfitting, computational resource demands, and deployment complexity are weaknesses compared to statistical modeling and data mining (10).

To address the weaknesses associated with individual methods, a better alternative is to combine the strengths of the different techniques. To the author's best knowledge, there is a rare use of such hybrid approaches in vehicle maintenance prediction in the current literature. To fill the research gap, the author proposed fused methods that integrate machine learning and statistical modeling advantages. Specifically, neural networks and logistic regression are selected to represent machine learning and statistical modeling, respectively, due to their popularity and ease of implementation. Also, a set of performance metrics, including accuracy, precision, and recall, is chosen to compare individual methods' and hybrid approaches' modeling performance. Finally, the paper selects the engine as the testbed to demonstrate the performance of the proposed approach. This predictive model is anticipated to help vehicle owners and mechanics proactively address potential issues before they become severe, enhancing vehicle performance and extending engine lifetimes. Extending past the realm of vehicle engines, this paper's exploration of hybrid predictive models hopes to guide its consideration and implementation of other research projects.

The following paper is structured as follows: First, a data description is provided to show the different predictors and targets of the data. Then, the methodology is provided to show the details of preprocessing steps, modeling training procedures, and the hybrid approach used. Then, the detailed evaluation results of the model are presented, followed by the discussion and conclusion of the findings and potential use cases for hybrid models.

DATA DESCRIPTION

It is an essential foundation for implementing this research to have a reliable and understandable dataset. Through extensive search, the author found a compatible dataset from the esteemed data library Kaggle that perfectly satisfies the aforementioned research need. First, this dataset includes various features and measurements related to the engine health of vehicles, such as engine rotations per minute, pressure, temperature, and engine condition. Second, it consists of a large sample size (19,503) across different models, years, and mileage, providing the needed diversity of vehicles (11).

The detailed descriptive statistics of the data used are shown in Table 1.

The goal of this research extends beyond engine maintenance prediction due to its alternate focus on the potential advantages of hybrid data models. Instead of collecting and testing thousands of engines, the author chose an observational study, which can be completed much faster with a much smaller budget. One potential limitation of the dataset on engine maintenance is the relatively low number of variables. A lack of variables may lead to a less accurate model as some influential factors may not be in the dataset. However, this simplicity also makes it much more practical to analyze the effectiveness of hybrid data models on a variable-by-variable basis.

METHODOLOGY

Four different types of models will be created, and the data above will be used to predict engine conditions. In each model, 80% of the data will be used for training, and the remaining 20% will be used for testing. To ensure fairness across all models, the data will be split based on seeds so that each model will get the same training and

Table 1. Descriptive Statistics of the Variables Used for the Data

Variables	Description	Minimum	Maximum	Mean	S.D.
ERPM	Engine rotations per minute	61.00	2239	791.2	267.6
LOP	Lubricant oil pressure	0.0034	7.266	3.304	1.022
FP	Fuel pressure	0.0032	21.14	6.656	2.761
CP	Coolant pressure	0.0025	7.479	2.335	1.036
LOT	Lubricant oil temp	71.32	89.58	77.64	3.111
CT	Coolant temp	61.67	195.5	78.43	6.207
EC	Engine Condition	0.000	1.000	0.6305	0.4827

Notes: 1. S.D. represents the standard deviation; 2. Engine Condition is a binary value with 1 demonstrating a need for maintenance.

testing data (12). Additionally, each model will be run through eight random seeds to examine its consistency. Therefore, the results will reflect the applicability of the models to this dataset in a controlled manner.

The methods employed will be logistic regression, neural network, and two different combinations of the two previous methods (or the hybrid ones). As this paper aims to test the potential added effectiveness of merging models, the base logistic regression and neural network models will act as a control to compare with hybrid models. For the models ending in logistic regression, each feature will have its coefficient, standard error, z-value, p-value, and 95% confidence intervals displayed. These results will also be shown for the intercept. Graphs will represent the training and validation loss of the neural networks, and the absolute weights of each feature will be calculated through the weight matrices by finding the mean of the absolute values of the associated weights. The absolute value is used because it demonstrates how influential each feature is to the decision of the neural network.

Then, all models will be evaluated in comparison to each other. The accuracy will be estimated based on the proportion of how many of its guesses were correct in the testing data. Precision will be the proportion of true positives out of all positive predictions. Recall will be the proportion of true positives out of all positives.

Logistic Regression

Logistic regression was developed as an alternative to the probit model for prediction and classification in 1944 from the work of Joseph Berkson (13). A generalized linear model, logistic regression is a tried-and-tested, simple, fast, and efficient method. Its prediction works through an equation consisting of the sum of multiple features, each with its own coefficient, often including an intercept. (14). The binary nature of the dataset's "engine condition" output variable makes logistic regression a prime model, as the predictions will range between zero and one. This model will also output results that provide insight into how each variable impacts the outcome, giving an easy metric to compare to another model (15).

For this model, the formula is:

$$\text{logit}(P(EC=1)) = \beta_0 + \beta_1 \times ERPM + \beta_2 \times LOP + \beta_3 \times FP + \beta_4 \times CP + \beta_5 \times LOT + \beta_6 \times CT \quad (1)$$

Where β_0 is the intercept and β_1 to β_6 are the coefficients of each variable. The variable descriptions are listed above in Table 1.

Neural Network

Both logistic and neural networks are great models for binary classification (16). While neural networks require more computation, they can find more complex relationships between features and the output. In contrast, logistic regression models can only find linear relations between features and the target. However, it is much harder to interpret how a neural network gets to its output from its input. Neural networks consist of multiple layers, each with nodes connected to every other node in the preceding and following layer. Between the input and output, there is a significant quantity of connections that are difficult to comprehend fully due to a vast amount of weighted parameters and paths that are difficult to trace. These are the hidden layers; a black box or a simple diagram often represents them. (17)

For the model to capture a wide array of relations, it will be trained with two hidden layers connecting the inputs to the output. Using two layers allows for multiple stages of data abstraction while not requiring an immense amount of computing. This neural network will have two hidden layers of 64 and 32 sizes respectively. The large amount of neurons in the first hidden layer will find an extensive range of patterns, while the second smaller layer will refine the results. (18). While there are many hidden layer nodes compared to the input layer size, there is a vast amount of training data, and the model uses early stopping and cross-validation to prevent overfitting (19).

A neural network predicts the "engine condition," with the six other variables as the inputs. The output node will conclude with a value between zero and one as a probability of the engine requiring maintenance. When calculating the model results, a probability value of 0.5 or greater will be considered a prediction of requiring maintenance. Figure 1 represents the structure of the neural network.

Hybrid Model 1: Logistic Regression into Neural Network

As shown before, the output of the logistic regression model demonstrated a multitude of linear relations. It relates the features to the engine condition, but there may be more connections that are not linear. In this method, the logistic regression model can serve as a strong base as an input to the neural network mode, while the neural network may continue to refine the results through its training (20). Once the logistic regression model is finished, its outputs for each set of inputs are saved. This data is then used as a seventh input node in a new neural network, replicating the structure of the previous neural network aside from the additional input. The implementation of logistic regression into a neural network model is shown in Figure 2.

Hybrid Model 2: Neural Network into Logistic Regression

In contrast to the first hybrid model, this one starts with the first neural network and then uses the result as a feature for a logistic regression model. The logistic regression model could provide more in-depth results of each variable concerning the neural network and tune the results (21). After the neural network model is completed, the output is used in a slightly altered version of the original logistic regression model represented by the following equation.

$$\text{logit}(P(EC=1)) = \beta_0 + \beta_1 \times ERPM + \beta_2 \times LOP + \beta_3 \times FP + \beta_4 \times CP + \beta_5 \times LOT + \beta_6 \times CT + \beta_7 \times NN \quad (2)$$

Compared with Equation 1, Equation 2 is slightly different as it has $\beta_7 \times$ NN added, where NN is the Figure 1 below output and β_7 is its coefficient. Below, Figure 3 demonstrates the inputs of the logistic regression model, which includes the six features of the dataset and the results of the previous neural network model.

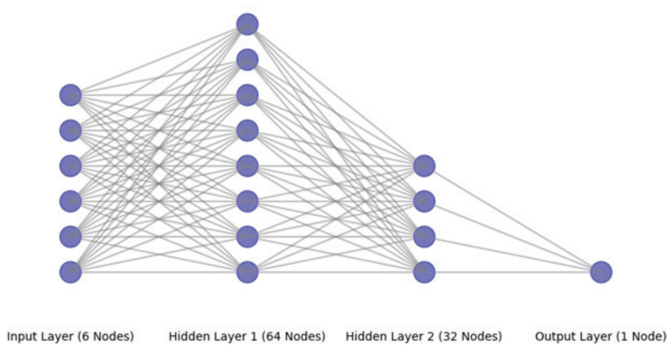

Figure 1. Neural Network Diagram.
Note: For illustration convenience, each node in the hidden layers represents eight nodes.

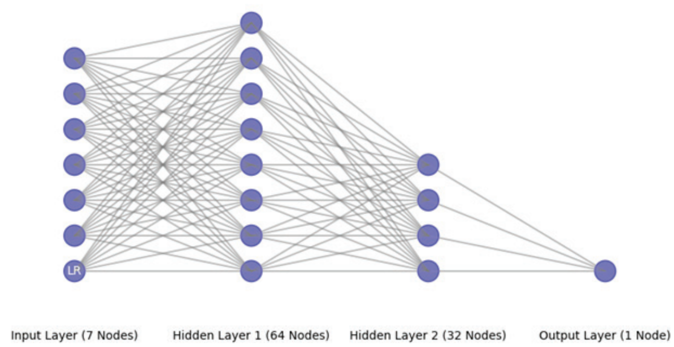

Figure 2. Hybrid Model 1 Diagram.
Notes: 1. For illustration convince, each node in the hidden layers represents eight nodes; 2. The node labeled LR represents the output of the logistic regression model being used as an input for the neural network.

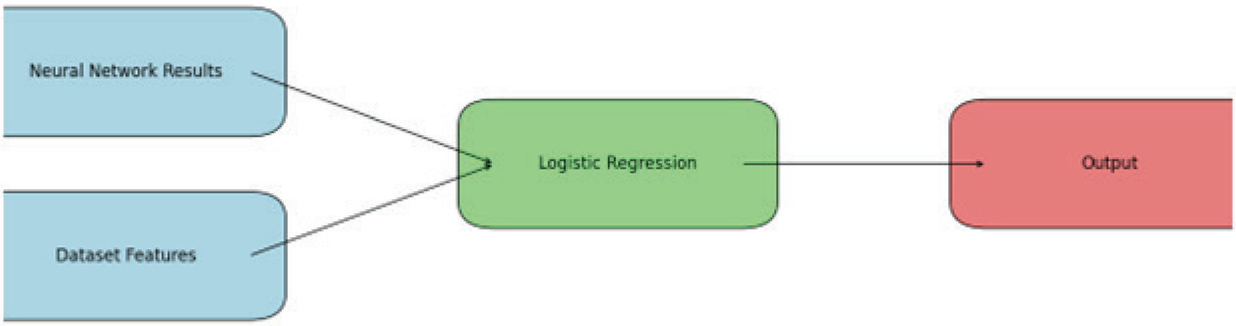

Figure 3. Hybrid Model 2 Diagram.

RESULTS

First, the results of the logistic regression model and Hybrid Model 2 will be compared. Since Hybrid Model 2 is essentially the logistic regression model with an additional input from the neural network model's output, both models yield similar types of results, such as p-values and coefficients. These results will be evaluated and compared.

Next, the results of the neural network model amd Hybrid Model 1 will be compared. Hybrid Model 1 is a neural network model that incorporates the output of the logistic regression model as an additional input. Metrics such as weights, training loss, and validation loss will be evaluated and compared.

Finally, the paper will discuss metrics common to all models, including accuracy, precision, and recall. Descriptive statistics, such as mean, maximum, and minimum values for each metric will be evaluated and compared. The analysis will determine the most and least successful models.

Logistic Regression Results

In review of Table 2, the coefficient represents the change in engine condition for each unit of change for the feature. ERPM has a negative coefficient, which indicates that the higher the ERPM value, the lower the engine condition value. Alongside its statistical significance, an engine with a higher ERPM value is significantly less likely to require maintenance. CP and LOT also have

Table 2. Logistic Regression Results

Feature	Coef	Std. Error	Z-value	P-value	95% CL L	95% CL U
Intercept	2.3390	0.4834	4.8671	0.6144	1.3914	3.2865
ERPM	-0.0021	0.0001	-31.5028	<0.0001	-0.0023	-0.0020
LOP	0.1599	0.0172	9.2824	<0.0001	0.1261	0.1936
FP	0.1038	0.0068	15.2963	<0.0001	0.0905	0.1171
CP	-0.0607	0.0167	-3.6331	0.0021	-0.0934	-0.0280
LOT	-0.0153	0.0056	-2.7899	0.1128	-0.0262	-0.0044
CT	0.0005	0.0028	0.1613	0.0237	-0.0051	0.0060

Notes: 1. Refer to Table 1 for variable descriptions; 2. The bolded p-values represent a feature of statistical significance at the level of 0.05; 3. Coef. Represents coefficient; 4. CL L and CL U represents the lower and upper ends of the confidence level.

negative coefficients while LOP, FP, and CT have positive coefficients. Therefore, an engine with higher CP and LOT will be less likely to need maintenance while an engine with higher LOP, FP, and CT will be more likely to need maintenance. The intercept starts at 2.339 and the prediction value is brought down to a value between zero and one by the other features.

In the review of Table 3, the most influential feature in this model is the neural network due to its high coefficient and low p-value. The neural network has a very high and positive coefficient. All the other features have negative coefficients, suggesting that the neural network overestimates the engine condition as requiring maintenance more often than it doesn't. The coefficients and p-values strongly differ from the logistic regression model as their purpose in this logistic regression model is to refine the results of the neural network.

The logistic regression model has multiple statistically significant features. On the other hand, Hybrid Model 2 only has one statistically significant feature not including the neural network or the intercept. While the logistic model uses the features for its predictions, Hybrid Model two only slightly considers the other features, mainly determining its output based on the neural network's output.

Neural Network Results

As the neural networks were going through their validation and training phases, the losses were recorded. The results are displayed in Figures 4 and 5 below.

In Figures 4 and 5, the training loss in the neural network and Hybrid Model 1 quickly drops off before ending at around 0.59. The validation loss oscillates around 0.60. While the neural networks were allowed to run for up to 1000 epochs, they stopped at around 100 to prevent overfitting, so more epochs would unlikely make the

Figure 4. FNeural Network Training and Validation Loss.

Figure 5. Hybrid Model 1 Training and Validation Loss.

Table 3. Hybrid Model 2 Logistic Regression Results

Feature	Coef	Std. Error	Z-value	P-value	95% CL L	95% CL U
NN	4.7675	0.4187	12.2896	<0.0001	3.9468	5.5882
Intercept	2.5444	0.5568	4.5126	0.0038	1.4531	3.6357
ERPM	-0.0003	0.0002	-4.3468	0.1346	-0.0006	<0.0001
LOP	-0.0149	0.0221	-0.4800	0.1397	-0.0583	0.0284
FP	-0.0006	0.0107	0.2045	0.2870	-0.0215	0.0203
CP	-0.0144	0.0175	-0.8359	0.5126	-0.0488	0.0198
LOT	-0.0514	0.0056	-9.1933	<0.0001	-0.0625	-0.0405
CT	-0.0098	0.0029	-3.4134	0.0501	-0.0154	-0.0042

Notes: 1. NN stands for neural network; 2. Refer to Table 1 for variable descriptions; 3. The bolded p-values represent a feature of statistical significance at the level of 0.05. 4. Coef. Represents coefficient; 5. CL L and CL U represent the lower and upper ends of the confidence level.

neural network more accurate. The training loss getting lower represents the neural network model adapting to the training data, but the consistency in validation loss shows that more epochs are not necessarily helping the model be accurate.

Based on the absolute weights in Table 4, the engine rpm is the most influential feature in the neural network, Similar to the logistic regression model. However, the differences between the absolute weights of the features in the neural network are minor. Each feature noticeably contributes to the final decision of the neural network.

The logistic regression output has a comparable absolute weight to the other features. It is considered in the final output but not the major deciding factor, unlike the role the neural network plays in Hybrid Model 2.

Compared to the neural network model, the feature weights in Hybrid Model 1 generally have slightly lower values due to the added logistic regression model's weight. Engine rpm has the highest feature weight in both of the neural network models and coolant pressure has the lowest weight. However, coolant temp has a relatively high weight for Hybrid Model 1 while it has a lower weight for the base neural network model. The weight values are likely altered based on how much the logistic regression model overestimated or underestimated the importance of a feature.

All Models Evaluation

Each model was run through 8 different permutations of the data randomly being split into the training and testing groups. These splits of data are controlled by being associated with a seed, which allows each model to use the same 8 sets of data. In this section, descriptive statistics for each model on each seed are taken and evaluated.

As shown in Table 5, the accuracies across the methods range between 0.6532 and 0.6734. While the accuracy differences are minimal, the hybrid techniques show a noticeable improvement compared to the logistic regression and neural network models. Specifically, HM1 has the most excellent performance in terms of minimum and mean accuracy, while it is tied with HM2 for the highest maximum accuracy. On the other hand, HM2 has the lowest standard deviation. Therefore, using hybrid models accomplishes improvements in model performance with this dataset.

In Table 7, the model with the best statistic in precision has the worst value of that statistic for recall. Often, having higher precision results in a lower recall and vice versa. There is no model with a clear advantage in both the precision and recall measurements. However, it

Table 4. Neural Network Feature Absolute Weights

Feature	Absolute Weight NN	Absolute Weight HM1
Engine rpm	0.1647	0.1595
Fuel pressure	0.1624	0.1427
lub oil temp	0.1521	0.1579
Lub oil pressure	0.1477	0.1365
Coolant temp	0.1458	0.1573
Coolant pressure	0.1431	0.1292
log_reg_probs	DNE	0.1446

Note: log_reg_probs represents the output of the logistic regression model; 2. NN stands for neural network; 3. HM1 stands for Hybrid Model 1.

Table 5. Descriptive Statistics of Each Method's Accuracy Measurements

	LR	NN	HM1	HM2
Maximum	0.6685	0.6732	0.6734	0.6734
Minimum	0.6532	0.6573	0.6596	0.6586
Mean	0.66054	0.66724	0.66756	0.66468
S.D.	0.00466	0.00512	0.00479	0.00432

Notes: 1. S.D. stands for standard deviation, LR stands for logistical regression, NN stands for neural network, HM1 stands for Hybrid Model 1, and HM2 stands for Hybrid Model 2; 2. The best value of each statistic is bolded and the worst value is highlighted.

Table 6. Descriptive Statistics of Each Method's Precision Measurements

	LR	NN	HM1	HM2
Maximum	0.6863	0.7097	0.7067	0.6979
Minimum	0.6689	0.6915	0.6893	0.6779
Mean	0.67726	0.69886	0.69670	0.68917
S.D.	0.00587	0.00553	0.00576	0.00610

Notes: 1. S.D. stands for standard deviation, LR stands for logistical regression, NN stands for neural network, HM1 stands for Hybrid Model 1, and HM2 stands for Hybrid Model 2; 2. The best value of each statistic is bolded and the worst value is highlighted.

Table 7. Descriptive Statistics of Each Method's Recall Measurements

	LR	NN	HM1	HM2
Maximum	0.8884	0.8523	0.8575	0.8640
Minimum	0.8701	0.7989	0.8176	0.8369
Mean	0.87997	0.82807	0.83551	0.85198
S.D.	0.00563	0.01497	0.01363	0.01005

Notes: 1. S.D. stands for standard deviation, LR stands for logistical regression, NN stands for neural network, HM1 stands for Hybrid Model 1, and HM2 stands for Hybrid Model 2; 2. The best value of each statistic is bolded and the worst value is highlighted.

is clear that all models have higher recall than precision measurements, meaning that they are better at predicting a model that requires maintenance than a model that does not require maintenance.

Table 1 for the data description shows that around 63% of the engines in the dataset require maintenance, which could explain why these models are better at correctly detecting when an engine needs maintenance than when it does not as more engines require maintenance than those that do not. It is likely that as the training data had more engines requiring maintenance, the final models are better tuned to finding engines that require maintenance.

CONCLUSIONS AND RECOMMENDATIONS

All four models had similar accuracies, around 66%, which could result from a need for more data. There may be more significant outside factors not included in the dataset that impact whether or not an engine requires maintenance. However, this percentage of average accuracy among many seeds and the thousands of data points strongly indicates an influential impact of the six given variables on the engine condition. Due to the questionable accuracy rate, the outputted model could see much improvement in practically predicting engine maintenance. This model could likely see an improvement in its accuracy if the dataset had more variables.

Though minimal, results such as accuracy display that hybrid models of logistic regression and neural networks show a clear improvement compared to their individual parts. This research goes beyond engines as it provides a basis for a deeper exploration of hybrid data models due to their slight edge over traditional models. The implementation of another model as a feature in an alternate model is a technique that can be used for many other models, and therefore potentially utilized in many forms of datasets. Specific models can be put into a hybrid model based on how they complement each other's strengths and weaknesses.

While proving to have an edge in accuracy, hybrid models face the limitation of requiring more training as they contrive of extra parts. This could be a dealbreaker in situations where massive amounts of data has to be processed due to the extra computing or time requirements. Additionally, the results of a hybrid model is more difficult to comprehend as the results of each model affects the other. If the goal of a data model is to see how impactful individual features are, it would be better to only use a singular simplistic model such as a logistic regression model.

Hybrid models hold much potential for widespread applications and should be a subject of interest for future research. While more complicated, the slight increase in accuracy from hybrid models can be substantial in fields such as medical diagnosis, where each percentage of accuracy is exceedingly important. It is possible that hybrid models may have much higher accuracies with specific datasets, but the opposite could also be true, where hybrid models are less accurate with other datasets. Therefore, also using the models that the hybrid models are contrived of is important as a point of comparison for the hybrid model's improvement.

ACKNOWLEDGEMENTS

The author gives great gratitude to the user Parv Modi, the one who uploaded the dataset to kaggle.com. Additionally, the author is deeply thankful to the reviewers of the paper. This research is a testament to the author's pursuit of knowledge and advancement in data science.

REFERENCE

1. You S, Krage M & Jalics L. Overview of remote diagnosis and maintenance for automotive systems (No. 2005-01-1428). SAE Technical Paper. 2005.
2. Chaudhuri A. Predictive maintenance for industrial iot of vehicle fleets using hierarchical modified fuzzy support vector machine. arXiv preprint arXiv:1806.09612. 2018.
3. Arena F, Collotta M, Luca L, Ruggieri M & Termine FG. Predictive maintenance in the automotive sector: A literature review. Mathematical and Computational Applications. 2021; 27 (1): 2.

4. Giobergia F. Data mining in the automotive industry for predictive maintenance (Doctoral dissertation, Politecnico di Torino). 2017.
5. Giobergia F, Baralis E, Camuglia M, Cerquitelli T, Mellia M, Neri A, ... & Tuninetti A. Mining sensor data for predictive maintenance in the automotive industry. In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). 2018, October; pp. 351-360. IEEE.
6. Esteban A, Zafra A & Ventura S. Data mining in predictive maintenance systems: A taxonomy and systematic review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2022; 12 (5): e1471.
7. Prytz R. Machine learning methods for vehicle predictive maintenance using off-board and on-board data (Doctoral dissertation, Halmstad University Press). 2014.
8. Theissler A, Pérez-Velázquez J, Kettelgerdes M & Elger G. Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry. Reliability engineering & system safety. 2021; 215: 107864.
9. Jain M, Vasdev D, Pal K & Sharma V. Systematic literature review on predictive maintenance of vehicles and diagnosis of vehicle's health using machine learning techniques. Computational Intelligence. 2022; 38 (6): 1990-2008.
10. Ratner B. Statistical and machine-learning data mining: Techniques for better predictive modeling and analysis of big data. Chapman and Hall/CRC., ISBN: 9781315156316. 2017.
11. Automotive Vehicles Engine Health Dataset. Kaggle. https://www.kaggle.com/datasets/parvmodi/automotive-vehicles-engine-health-dataset/data (accessed on 2024-4-5).
12. Qian S, Pham VH, Lutellier T, Hu Z, Kim J, Tan L, ... & Shah S. Are my deep learning systems fair? An empirical study of fixed-seed training. Advances in Neural Information Processing Systems. 2021; 34: 30211-30227.
13. Cramer JS. The early origins of the logit model. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences. 2004; 35 (4): 613-626.
14. Dobson AJ & Barnett AG. An introduction to generalized linear models. Chapman and Hall/CRC., ISBN: 9781315182780. 2018.
15. Hosmer Jr, DW, Lemeshow S & Sturdivant RX. Applied logistic regression. John Wiley & Sons., ISBN: 9781118548387. 2013.
16. Terrin N, Schmid CH, Griffith JL, D'Agostino Sr, RB & Selker HP. External validity of predictive models: a comparison of logistic regression, classification trees, and neural networks. Journal of clinical epidemiology. 2003; 56 (8): 721-729.
17. Yosinski J, Clune J, Nguyen A, Fuchs T & Lipson H. Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579. 2015.
18. Rauber PE, Fadel SG, Falcao AX & Telea AC. Visualizing the hidden activity of artificial neural networks. IEEE transactions on visualization and computer graphics. 2016; 23 (1): 101-110.
19. Ghojogh B & Crowley M. The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. arXiv preprint arXiv:1905.12787. 2019.
20. Liang M & Hu X. Recurrent convolutional neural network for object recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2015; pp. 3367-3375.
21. Friedman J, Hastie T & Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics. 2000; 28 (2): 337-407.