DOI: http://dx.doi.org/10.26483/ijarcs.v16i2.7224

ISSN No. 0976-5697

Volume 16, No. 2, March-April 2025

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

PREDICTIVE MAINTENANCE TO REDUCE MACHINE DOWNTIME IN FACTORIES USING MACHINE LEARNING ALGORITHMS

Manasa Koppula
Department of Computer and Data Science,
The University of Texas at Austin, Austin, TX 78712, United States

***Abstract*:** Accurate machine failure detection allows manufacturers to estimate potential machine deterioration and avoid machine downtime caused by unexpected performance issues. Predictive maintenance with the use of machine learning algorithms may anticipate machine faults and maximize maintenance efforts to solve machine downtime problems. To anticipate machine breakdowns and minimize downtime, this work applies a variety of machine learning methods, such as Random Forest, Decision Tree, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, and Logistic Regression. Based on the performance measurement values, Random Forest model has shown high levels of accuracy, precision, recall, and F-score. The sequence of order for accuracy of machine learning models follows as: Random Forest > Decision Tree> Gradient Booster Classifier and SVM > Logistic Regression and KVM. This work emphasizes that, through various machine learning models, machine manufacturers could optimize the machine maintenance and prolong the life of machines.

***Keywords*:** Predictive maintenance, machine downtime, machine learning, algorithms, optimization

1. INTRODUCTION

Maintaining the machine's performance over time is one of the biggest problems faced by machine manufacturers. To address machine failure and prevent needless maintenance expenses, predictive maintenance of the machinery is required [1]. Unplanned machine downtime in a manufacturing facility has the potential to increase product losses, raise maintenance costs, and interfere with production plans [2]. For any industry or manufacturing unit to remain safe and effective, machine maintenance is essential. An organization's operational performance and process safety are severely impacted by a poor maintenance system since it directly affects productivity levels [3].
The conventional way of scheduling maintenance depends on regular manufacturing shutdowns to carry out maintenance operations. By concentrating on when to schedule periodic maintenance, the conventional maintenance scheduling can be optimized to cut down the maintenance expenses. However, the conventional maintenance method ignores the connection between equipment deterioration and maintenance and is not the best strategy. Therefore, employing predictive maintenance offers improved industry practice for scheduling maintenance [4]. Predictive maintenance is required for manufacturing or production industries to plan numerous operational activities such as continuous production and long-term maintenance of machines [5]. Predictive maintenance detects the possibility of early detection maintenance failure and faults [6].
Machines frequently show symptoms and indications of poor performance prior to failure. Engineers employ predictive maintenance as a technique to foresee performance issues before failure occurs [7]. Predictive maintenance, a promising technique, has the potential to overcome the tradeoff issue associated with unplanned and preventative maintenance by optimizing uptime and a component's useful life at the same time. It is intended to track the state of out-of-service equipment and forecast when it may break down. It enables the approximation of future machine component behavior and condition, which will aid in the optimization of maintenance chores. As a result, while maintaining the lowest possible frequency of maintenance, machine downtime, and maintenance expenses can be greatly decreased [8]. While there are many advantages to implementing predictive maintenance at different design stages, there are drawbacks as well. Increased productivity, fewer system errors, less unscheduled downtime, and better resource efficiency are among the benefits. Additionally, predictive maintenance improves the optimization of maintenance intervention planning [9, 10].
Predictive maintenance utilizes machine learning and digital data to estimate when a machine needs a maintenance. The most common machine learning algorithms used are Random Forest, Decision Tree, Support Vector Machine, KNN etc. However, each study utilizes different data, and different machine learning methods will be applied to specific components of the machine. For this reason, it is necessary to employ and compare various machine learning algorithms [3].
Asti D et al. [11] developed a machine learning model and optimized the models based on the accuracy and prediction, F1 score for a prediction of machine failure. Additionally, they compared the outcomes between the models, such as Logistic Regression, Naive Bayes, K-Nearest Neighbors (KNN), Decision Tree, AdaBoost Classifier, Gradient Boosting, Random Forest, Extra Tree Classifier, and HistGradient Boosting to assess the machine’s downtime status. When compared to other models or algorithms, Random Forest demonstrated superior performance in terms of choosing criteria.
Mohammad Shahin et al., [2] created a machine learning and deep learning models to predict how the system

could malfunction and minimize downtime. According to performance assessment criteria including accuracy, precision, recall, and the F-score, the results suggest that deep forest and gradient boosting algorithms had very high levels of average accuracy (above 90%).

The main objective of this work is to reduce machine downtime through the integration of machine learning models. Reducing waiting times and downtime brought on by equipment failures due to poor maintenance procedures is crucial for increasing output and lowering expenses. Together with incorporating a suitable machine learning model and proactive maintenance techniques, the lifespan of equipment could be increased greatly.

This work aims to reduce manufacturing downtime through the integration of machine learning methodologies. Reducing waiting times and downtime brought on by machine failures and poor maintenance procedures is important for increasing productivity, lowering expenses, and maximizing the lifespan of the equipment [2]. In this study, various machine learning algorithms have been used that may be able to detect faults early on before they cause future machine failure. Understanding how machine learning algorithms can be used to anticipate unscheduled machine downtime and lower maintenance costs is the main goal of this study.

2. METHODS

2.1. Data collection

The current study makes use of the "Optimization of Machine Downtime" dataset from Kaggle’s [12]. The dataset includes 2,500 records and 16 attributes (two categorical and 14 numerical) such as date, machine ID, load cells, hydraulic pressure, coolant pressure, air system pressure, coolant temperature, hydraulic oil temperature, proximity sensors, spindle vibration, tool vibration, spindle speed, voltage, torque, cutting force, and downtime. The dataset has around 50.6% of Machine Failure entries and 49.4% of operational entries. The dataset has missing values in most of the columns; hence, missing values in numerical columns are imputed with their mean, and categorical columns are imputed with the mode.

2.2. Data Split

While there are many ways to split the dataset, like Train-Test Split, Train-Validation-Test Split, k-fold Cross Validation, Stratified Sampling, and Time-primarily based split, a Train-Test Split have been used for dividing the dataset. 75% of the dataset is used for training purposes and 25% for testing purposes, with the dataset being divided into training and testing sections.

2.3. Data Preprocessing

The data preprocessing tasks, such as cleaning, transformation, and selection, are performed iteratively to prepare the dataset. This includes removing duplicates, handling missing values using imputation techniques (where the mean was applied for numerical columns and the mode for categorical columns), aligning data types, and standardizing numerical features to avoid biases. Outliers are identified and managed using winsorization technique, and feature engineering introduces new variables based on machine performance patterns. Quality assurance checks validate the data's reliability, and the dataset is split for training and testing to evaluate model performance.

2.4. Data Pipeline

The data pipeline is designed for flexibility and adaptability, facilitating the processing of raw sensor data, feature extraction, and application of machine learning models. Python was used for data acquisition, cleaning, and organizing, while tools like pandas and NumPy ensure the data is in a standard format. Descriptive analysis and visualization help uncover patterns related to machine downtime, and predictive models are built using scikit-learn.

2.5. Exploratory Data Analysis

Exploratory data analysis (EDA) was used to gain a deeper understanding of the dataset's complex characteristics. While there is no fixed methodology for EDA, typical techniques include summary statistics, correlation analysis, data visualization, and aggregation methods. The analysis specifically focuses on understanding the machine environment, examining machine dynamics and performance, analyzing statistical properties and variability, and identifying distribution patterns.

2.6. Distribution Characteristics

Metrics such as skewness, which measures the degree of asymmetry, and kurtosis, which describes the characteristics of the distribution tails, are essential in exploratory data analysis (EDA) [6]. When analyzing distribution patterns, it is observed that variables like applied force, hydraulic pressure, and air system pressure follow right-skewed distributions with prominent peaks and extended tails. This indicates that lower values are more common, with occasional higher values, suggesting a vulnerability to extreme values, particularly in the case of cutting force. This detailed, research-focused analysis forms the foundation for deeper exploration of machine dynamics and informs future efforts to optimize and improve industrial processes.

2.7. Statistical Analysis

Correlation matrices was developed to perform multivariate Exploratory Data Analysis (EDA). Correlation matrices involve evaluating the relationships between different variables through their correlation coefficients. This process offers valuable insights into how variables are interrelated within the dataset.

2.8. Algorithms

The output variable, ‘downtime’, was predicted using a variety of machine learning algorithms. For improved prediction, the algorithms were tuned using a variety of hyper parameters. Accuracy, precision, F1-score, MTBF (Mean time between failures), and AUROC are the primary metrics for all algorithms.

2.8.1. Logistic Regression

Logistic Regression was applied to classification tasks where the objective was to forecast the likelihood that an instance will fall into a particular class or not. This algorithm was applied as it provides a clear explanation of how different factors affect downtime by explicitly modeling the likelihood of machine failures [13].

2.8.2. Naïve Bayes

Naïve Bayes algorithm is based on the Bayes theorem and the main idea behind this algorithm is each pair of features being classified is independent of the others [14].

2.8.3. K-Nearest Neighbors (KNN)

Considering the values of K nearest neighbors in the training dataset, the KNN algorithm predicts the label or value of a new data point based on the similarity principle [15].

2.8.4. Decision Trees

Key characteristics influencing machine reliability are identified by decision trees. AdaBoost combines several weak learners to increase prediction accuracy [16].

2.8.5. Gradient Boosting

Gradient Boosting combines several weak models to produce a single and more accurate predictive model [17].

2.8.6. Random Forest

Random forest enhances predictions by using multiple decision trees [18].

2.8.7. Support Vector Machine (SVM)

SVMs are frequently applied to classification issues. By identifying the best hyperplane that optimizes the margin between the nearest data points of opposing classes, they can differentiate between two classes [19].

3. RESULTS

To comprehend the multivariate analysis, a heatmap or correlation matrices is utilized **(Figure 1)**. This aids in our comprehension of how the variables relate to one another. There is a negative relationship between cooling component temperature and torque, suggesting that greater torque may facilitate cooling. There is a link between spindle speed, force, and vibration characteristics, suggesting that higher speeds lead to higher levels of force and vibration. Sensors, particularly those measuring physical loads and proximity, show strong correlations with vibrational factors (linked to the machine's spindle and tool) and cutting force. These correlations suggest that these measurements may share common underlying influences or exhibit similar patterns.

Within the hydraulic systems, significant correlations are observed between pressure readings (both hydraulic and coolant) and temperature measurements (coolant and oil), indicating a systemic connection among these variables. For dynamic interactions, spindle speed shows notable associations with vibration and force measures, particularly for tool and cutting forces. This implies that higher speeds tend to increase force and vibration levels. A subtle negative correlation is also observed between rotational force (torque) and the temperature of cooling components, implying that higher torque might slightly aid in cooling. To identify the most impactful features for modeling, the correlation strengths, variability, and predictive potential parameters were examined. Features that have strong correlations with the target variable and significant variability are typically key to the model development. Additionally, features that exhibit high predictive power, those that strongly influence the target outcome, are especially critical. Variables related to force application, hydraulic dynamics, and sensor measurements emerge as key influencers. The findings suggest that machining operations typically involve moderate applied forces, indicating a trend toward moderate force applications. At the same time, coolant pressure remains low, reflecting an ongoing need for temperature control. Additionally, air system pressure and coolant temperature both highlight moderate operational demands in these areas.

The insights gained from these correlations and dependencies laid a foundation for the next phase of model development. With a clearer understanding of the dataset's characteristics and influential features, the subsequent step will involve creating a model that leverages these insights for better prediction accuracy and operational optimization.

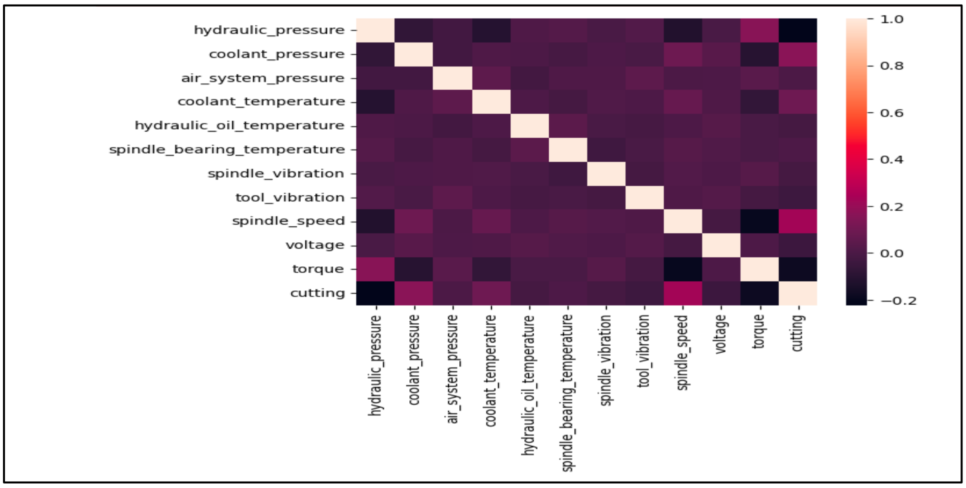

Figure 1. Correlation Matrix / Heatmap

We further explore the machine's dynamics and performance, focusing on key factors that influence its operation. Vibration measurements from the spindle and tool offer essential insights into the stability of the machining process. The spindle's higher rotational speed plays a significant role in determining the efficiency and speed of machining operations **[Figure 2].** The machine's high electrical voltage demands highlight its considerable power consumption. Furthermore, torque values represent the rotational force applied to the spindle, indicating the power used in machining. The cutting force levels demonstrate the moderate force typically applied during the machining process.

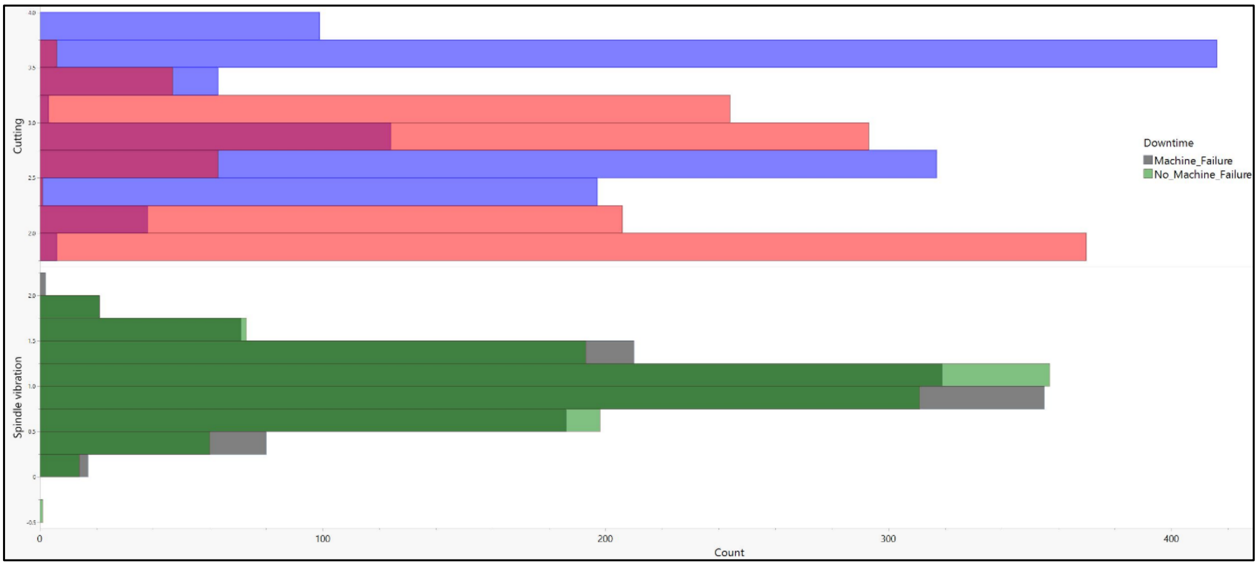

Figure 2: Performance analysis measurement for spindle vibration and cutting force

The analysis also includes a detailed examination of the statistical properties and variability within the dataset. This involves carefully investigating fluctuations, trends, and patterns **[Figure 3],** offering a thorough understanding of the data's inherent variability and revealing potential factors that contribute to its diverse behaviors. Hydraulic pressure shows significant variability, suggesting changes in power requirements. Temperature readings display varying levels of consistency, influencing the machine's thermal conditions. Vibration levels also exhibit fluctuating patterns, which directly affect the stability of the machining process. Additionally, spindle speed, electrical voltage, torque, and cutting force all demonstrate unique levels of variability.

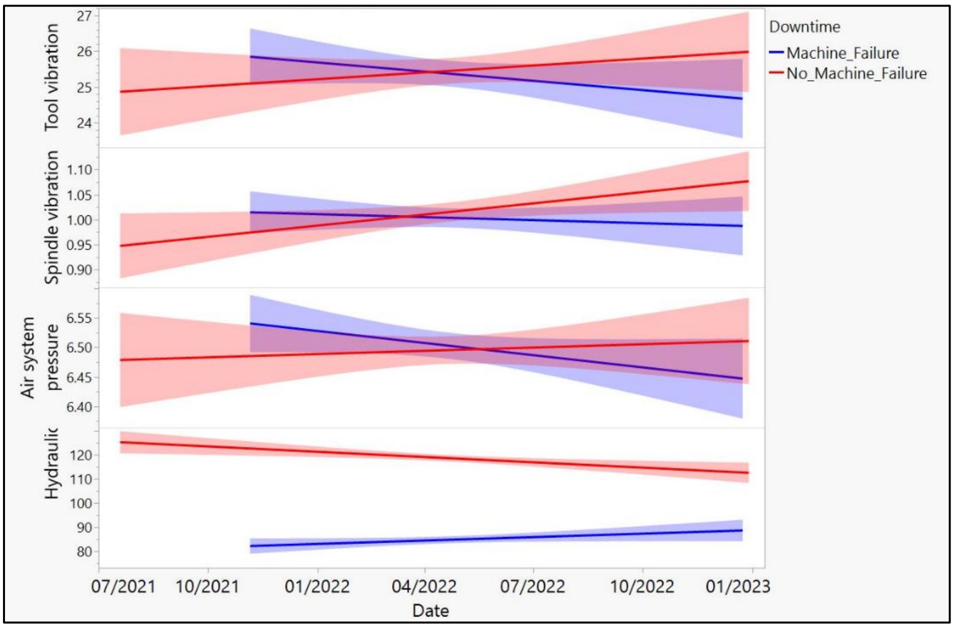

Figure 3: Performance trend for machine tool over the period

Table 1 depicts the model performance metrics for all the 7 algorithms. A crucial statistic that ranges from 0 to 1 is the F1-score, which is computed as F1 = (2 × Precision × Recall) / (Precision + Recall) [14]. Higher values signify a better balance between precision and recall. It is especially helpful in dealing with false negatives, which anticipate downtime when none exists, and false positives, which miss real downtime. A high F1-score guarantees that the model minimizes false alarms while accurately detecting and forecasting machine downtime. Another crucial statistic that assesses the trade-off between false positive and true positive rates is the Area Under the Receiver Operating Characteristic (AUROC). It has a range of 0 to 1, where a model that performs better is indicated by values nearer 1. A thorough grasp of the model's capacity to distinguish between classes across a range of probability thresholds is offered by the AUROC curve.

The reliability of the machine is calculated using MTBF (mean time between failure). MTBF is calculated by dividing the total operating time with the number of failures.

Numerous models, including Random Forest, Logistic Regression, Naive Bayes, KNN, Decision Trees, SVM, and Gradient Boosting, were utilized for model optimization. Five assessment metrics—accuracy, precision, recall, F1-score, AUROC, and MTBF were used to evaluate their performance **(Table 1).**

Table 1. Model performance details

Model Name	Accuracy	Precision	Recall	F1-Score	AUROC	MTBF
Logistic Regression	0.85	0.85	0.85	0.85	0.99	2.0
Gradient Booster Classifier	0.91	0.91	0.91	0.91	0.99	2.0
KNN	0.85	0.85	0.85	0.85	0.99	2.0
Naïve Bayes	0.85	0.85	0.85	0.85	0.99	2.1
SVM	0.91	0.91	0.91	0.91	0.99	2.0
Decision Tree	0.95	0.95	0.95	0.95	0.99	2.0
Random Forest Classifier	0.99	0.99	0.99	0.99	0.99	2.1

Despite having lower values on most metrics and an accuracy of 0.85, Logistic Regression, KNN, and Naïve Bayes are nonetheless reliable models, with Naïve Bayes having a little higher MTBF than the others. Gradient Boost Classifier, Support Vector Machine (SVM), and Decision Tree come in second and third, respectively, with strong accuracy, precision, recall, and F1-score. The Random Forest Classifier has the best performance on all parameters, with accuracy, precision, recall, F1-score, and AUROC all at 0.99.

The choice of the best model ultimately depends on specific optimization goals. Random Forest, with its ensemble approach, stands out as a strong contender, offering high accuracy and precision. However, the final decision should be based on the priorities of the downtime optimization task, whether minimizing false positives or maximizing overall accuracy. Further fine-tuning or the use of ensemble techniques could further improve the model's performance, offering a tailored solution for machine downtime prediction.

After a thorough model selection process, the Random Forest model underwent detailed hyper parameter tuning to enhance its accuracy and reduce misclassification rates. This optimized model was then integrated into Streamlit, providing an interactive and user-friendly platform for predicting machine downtime events. The fine-tuning process not only improved prediction accuracy but also minimized errors, delivering reliable real-time insights within the Streamlit application. This seamless integration demonstrates the model's readiness for real-world deployment, supporting the goal of reducing false positives and maximizing accuracy in downtime predictions. Ongoing improvements will focus on further refining the Random Forest model through continuous hyperparameter adjustments. Additionally, incorporating advanced ensemble methods and user feedback will be key to adapting the model to evolving operational conditions. Future updates will aim to enhance the model's robustness and maintain its effectiveness in dynamic machine downtime prediction scenarios.

Optimizing the hyper parameters of each model is essential for accurately predicting machine downtime. The model's ability to generalize and make reliable predictions is heavily influenced by the careful selection and evaluation of critical parameters such as learning rates, tree depths, and regularization strengths [20, 21]. A systematic approach, involving in-depth analysis and testing of these hyper parameters, is necessary to achieve optimal results in machine downtime prediction.

Table 2 shows details of model specific hyper parameters.

Table2: Details of model specific hyper parameters.

Model	Hyperparameters
Logistic Regression	- Penalty: ('l1', 'l2', 'elastic net', 'none') - C: Inverse of regularisation strength
Gradient Boosting	- n_estimators: Number of boosting stages

	- learning_rate: Weight applied to each tree - max_depth: Maximum depth of individual trees
KNN	- n_neighbors: Number of neighbors - weights: Weight function ('uniform' or 'distance') - p: Power parameter for Minkowski distance
Naive Bayes	Type: Gaussian, Multinomial, or Bernoulli
SVM	C- regularization parameter degree Degree of polynomial Gamma-kernal coefficient Kernel-kernel type
Decision Tree	- Criterion: Measure of split quality ('gini' or 'entropy') - max_depth: Maximum depth of the tree - min_samples_split: Minimum samples to split an internal node
Random Forest classifier	n_estimators: Number of trees in the forest - criterion: Measure of split quality ('gini' or 'entropy') - max_depth: Maximum depth of the trees

CONCLUSION

The study shows how machine learning models can be used to forecast machine downtime and enhance maintenance procedures by implementing predictive maintenance. Several machine learning algorithms were used to predict machine failures and optimize maintenance plans by utilizing a dataset that included a variety of parameters, such as machine health indicators and downtime status. Based on the evaluation criteria (accuracy, precision, recall, F1-score, and AUROC), the Random Forest Classifier was the best-performing model and was demonstrated to be the most dependable model for machine downtime prediction.

With excellent accuracy and other metric scores, gradient boosting and support vector machines (SVM) also demonstrated strong performance, making them viable options for predictive maintenance. KNN, Naïve Bayes, and logistic regression all performed consistently across metrics but were less accurate and precise compared to the random forest model. For predictive maintenance, these models remain reliable despite their relatively low accuracy (0.85), with Naïve Bayes showing a little improved MTBF (Mean Time Between Failures). All models' performance was optimized by hyperparameter tuning, but Random Forest and SVM benefited from this, producing better outcomes.

This study emphasizes the use of machine learning models in the industrial setup and the significance of predictive maintenance in lowering maintenance expenses and machine downtime.

Manufacturers may guarantee increased productivity, longer equipment life, and lower operating costs by selecting the best model based on performance parameters. The study also confirms that, depending on the situation and requirements of the application, simpler models can still be successful even when more complex models, such as Random Forest, yield the best results. Additionally, maximizing hyperparameter tweaking in machine learning models will be the focus of future research. Furthermore, examining a variety of deep learning techniques outside of LSTM may enhance modeling and yield more accurate forecasting results.

ACKNOWLEDGEMENTS

Manasa Koppula would like to thank Professor Jiao and Segun Jun Choi at the University of Texas at Austin.

REFERENCES

1. Adebayo, R. A., Ogundipe, O. B., & Bolarinwa, O. G. (2021). Development of a Motorcycle Trailer Hitch for Commercial Purposes.
2. Shahin, M., Chen, F. F., Hosseinzadeh, A., & Zand, N. (2023). Using machine learning and deep learning algorithms for downtime minimization in manufacturing systems: An early failure detection diagnostic service. The International Journal of Advanced Manufacturing Technology, 128(9-10), 3857-3883.
3. Yeardley, A. S., Ejeh, J. O., Allen, L., Brown, S. F., & Cordiner, J. (2022). Integrating machine learning techniques into optimal maintenance scheduling. Computers & Chemical Engineering, 166, 107958.
4. Zonta, T., Da Costa, C. A., Zeiser, F. A., de Oliveira Ramos, G., Kunst, R., & da Rosa Righi, R. (2022). A predictive maintenance model for optimizing production schedule using deep neural networks. Journal of Manufacturing Systems, 62, 450-462.
5. Pech, M., Vrchota, J., & Bednář, J. (2021). Predictive maintenance and intelligent sensors in smart factory. Sensors, 21(4), 1470.
6. Li, D., Landström, A., Fast-Berglund, Å., & Almström, P. (2019). Human-centred dissemination of data, information and knowledge in industry 4.0. Procedia CIRP, 84, 380-386.
7. Liu, Z., Jin, C., Jin, W., Lee, J., Zhang, Z., Peng, C., & Xu, G. (2018, June). Industrial AI enabled prognostics for high-speed railway systems. In 2018 IEEE international conference on prognostics and health management (ICPHM) (pp. 1-8). IEEE.

8. Lee, W. J., Wu, H., Yun, H., Kim, H., Jun, M. B., & Sutherland, J. W. (2019). Predictive maintenance of machine tool systems using artificial intelligence techniques applied to machine condition data. Procedia Cirp, 80, 506-511.
9. Yadav, D. K., Kaushik, A., & Yadav, N. (2024). Predicting machine failures using machine learning and deep learning algorithms. Sustainable Manufacturing and Service Economics, 3, 100029.
10. Wang, K., & Wang, Y. (2018). How AI affects the future predictive maintenance: a primer of deep learning. In Advanced Manufacturing and Automation VII 7 (pp. 1-9). Springer Singapore.
11. Asit D, Karanam S, Susmitha P, Indira K, Bharani K D. (2023). Maximising Operational Uptime: A Strategic Approach to Mitigate Unplanned Machine Downtime and Boost Productivity using Machine Learning Techniques. International Journal of Innovative Science and Research Technology, 8, 12.
12. https://www.kaggle.com/datasets/srinivasanusuri/optimization-of-machine-downtime
13. Miyazaki, Y., Kawakami, M., Kondo, K., Hirabe, A., Kamimoto, T., Akimoto, T., ... & Tsuji, T. (2024). Logistic regression analysis and machine learning for predicting post-stroke gait independence: a retrospective study. Scientific Reports, 14(1), 21273.
14. Wolfson, J., Bandyopadhyay, S., Elidrisi, M., Vazquez-Benitez, G., Vock, D. M., Musgrove, D., ... & O'Connor, P. J. (2015). A Naive Bayes machine learning approach to risk prediction using censored, time-to-event data. Statistics in medicine, 34(21), 2941-2957.
15. Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. Annals of translational medicine, 4(11).
16. Matzavela, V., & Alepis, E. (2021). Decision tree learning through a predictive model for student academic performance in intelligent m-learning environments. Computers and Education: Artificial Intelligence, 2, 100035.
17. Sarker, I. H. (2021). Machine learning: algorithms, real-world applications and research directions. SN Comput Sci 2: 160.
18. Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. The Stata Journal, 20(1), 3-29.
19. Rodríguez-Pérez, R., & Bajorath, J. (2022). Evolution of support vector machine and regression modeling in chemoinformatics and drug discovery. Journal of Computer-Aided Molecular Design, 36(5), 355-362
20. Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2018). Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data. arXiv preprint arXiv:1803.11266.
21. Kee, T., & Ho, W. K. (2024). Optimizing Machine Learning Models for Urban Studies: A Comparative Analysis of Hyperparameter Tuning Methods.

