# Predictive Maintenance Strategies for Fan Motors: An Examination of Sensor Technologies and Machine Learning Approaches

1st Truong Khang Thinh Nguyen
*Bachelor of Artificial Intelligence*
*Deakin University*
Melbourne, Australia
s223446545@deakin.edu.au

***Abstract*—Heating, Ventilation, and Air Conditioning (HVAC) systems are critical for maintaining comfort and air quality in buildings, but they are also prone to failures that can lead to costly downtime and inefficient energy usage. A key component of these systems is the fan, which plays a vital role in regulating airflow. Over time, wear and tear on fan components can cause system malfunctions. Predictive maintenance (PdM) aims to address these issues by detecting early-stage faults before they result in significant failures, helping to minimize unexpected breakdowns and reduce maintenance costs. This literature review aims to evaluate and discuss existing methods for fault detection in HVAC fan systems, addressing current challenges and limitations related to data collection and problem-solving approaches, in addition to proposing a novel solution that integrates hybrid methods with data fusion techniques and utilizes the Random Forest (RF) algorithm. By enhancing the accuracy and timeliness of fault detection, this review identifies gaps in existing research and lays the foundation for developing an innovative hybrid approach to advance predictive maintenance in HVAC systems.**

***Index Terms*—HVAC systems, Condition-Based Maintenance (CBM), fault detection, predictive maintenance, data fusion, Random Forest**

## Abbreviations and Acronyms

List of Abbreviations

- AUC : Area Under the Curve
- CBM : Condition-Based Maintenance
- EDA : Exploratory Data Analysis
- ECM : Electret Condenser Microphone
- ESN : Echo State Network
- FDD : Fault detection and diagnosis
- HVAC : Heating, Ventilation, and Air Conditioning
- ML : Machine Learning
- NRMSE : Normalized Root-Mean-Square Error
- PdM : Predictive Maintenance
- PHM: Prognostics and Health Management
- RF : Random Forest
- ROC : Receiver operating characteristic
- RMSE : Root Mean Square Error
- RNNs : Recurrent Neural Networks
- RUL : Remaining Useful Life
- RX : Receive Terminal
- SVM : Support Vector Machines
- TX : Transmit Terminal

## I. Introduction

With the rise of Industry 4.0, both physical and digital products have expanded more than ever before. This integration of physical and digital environments facilitates the collection of vast amounts of data from various pieces of equipment across different factory sectors [1]. Meanwhile, maintaining this equipment is crucial as it impacts operational uptime and efficiency. Thus, it is essential to detect and address equipment faults promptly to prevent interruptions in production processes, so-called predictive maintenance [2].

The building sector is respondible for 36% of the total global energy consumption. Of all the energy-consumer devices within a building, HVAC systems account for over 50% of the total energy consumed [6]. This makes HVAC systems a source of preventable and unexplored energy waste that can be tackled by incorporating intelligent operations. Meanwhile, malfunctioning of HVAC systems can endanger the lifespan of equipment, energy usage, and occupant thermal comfort. Therefore, it is crucial to find solutions that lower the energy consumption of HVAC systems while addressing their malfunctioning problem. FDD systems have the potential to address the issue by reducing equipment downtime, energy costs, and maintenance costs [3], [4].

There are three main categories of predictive maintenance: CBM, PHM and RUL. CBM focuses on monitoring the condition of equipment using real-time data to decide when maintenance should be performed. PHM, on the other hand, is a broader framework that includes predictive maintenance, diagnostics, and prognostics to manage the overall health of a system. While CBM is part of PHM, PHM also involves more sophisticated prognostic models that not only identify faults but also predict when a failure is likely to occur. RUL prediction goes a step further than CBM by predicting how much time remains before a failure occurs. RUL prediction takes this a step further by estimating the remaining operational time before a system failure happens. Regardless

Identify applicable funding agency here. If none, delete this.

of the specific approach employed, predictive maintenance typically falls under one of three types: physical model-based, knowledge-based, or data-driven methods [1], [4].

## II. Literature Review

Several experiments have been conducted to predict failures in HVAC systems, in our interest research, focusing on fan systems. It is important to note that most tests applied in HVAC and predictive maintenance (PdM) rely on simulation environments as collecting real-world data presents significant challenges as well as just one source of sensor, which can be insufficient for analysis [1], [5], [7].One experiment [10] seeks to address the issue of frequent mill fan repairs in a power plant, where erosion leads to repairs every 2 to 2.5 months despite using high-resistance steels. Conducted at the Maritsa East 2 thermal power plant in Bulgaria, the study collected vibration data and key operational parameters (such as grinding productivity, fan productivity, and vibration state) through the Decentralized Control System (DCS) Historian system over an 8-month period in 2010. The researchers employed a hybrid method combining physics-based and knowledge-based models to analyze the vibration data for predictive maintenance purposes. They specifically focused on non-linear trends in the vibration data, utilizing MATLAB for regression analysis. The results showed that analyzing non-linear trends effectively predicts changes in the condition of the mill fan motor. To maximize the potential of the proposed diagnostic procedures, additional information on maintenance schedules and rotor replacement shutdowns is essential for interpreting vibration trends accurately. The study highlights the value of vibrosignals in diagnosing mill fan conditions, as they provide significant diagnostic insights that can isolate and clarify issues. Incorporating operational efficiency data into the analysis would enhance the overall understanding of the system's maintenance needs, ultimately optimizing predictive maintenance strategies by aligning vibration data with operational performance.

Another study [9] explored predictive maintenance for mill fan systems by using RNNs to predict potential failures based on performance degradation using data-driven method. Vibration data from bearings near the mill rotor were used to train two RNN types: the Elman Network and the ESN. The researchers aimed to assess these models' performance in terms of prediction accuracy and training time. Due to the presence of multiple vibration sources, wavelet de-noising was applied to filter the data. Results showed that the ESN significantly outperformed the Elman network, achieving a much lower NRMSE in both training and testing phases. Additionally, the ESN trained much faster, taking 20 minutes compared to over 10 hours for the Elman network. In a second experiment, the dataset was reduced by a factor of 10, with similar results favoring the ESN in both accuracy and speed. This study concluded that the ESN is particularly suitable for real-time predictive maintenance in scenarios requiring large datasets and quick response times.

In a recent study [8], the authors set out to predict faults or abnormalities in an industrial fan using a data-driven approach, specifically employing the RF algorithm. The data was collected through the MPU-6050 sensor, which measured both 3-axis accelerometer (vibration data) and temperature. The results showed that the RF algorithm performed exceptionally well, achieving an accuracy of 99%. Additionally, the regressor approach yielded a RMSE of 80, demonstrating the model's effectiveness. The authors also confirmed that overfitting was not an issue in the model.

While current solutions for detecting faults in fan systems within HVAC systems are effective, there are still areas where improvements can be made.

## III. Approach Proposal

The enhanced strategy will involve integrating various data sensors through data fusion and employing a hybrid method. This approach will combine a data-driven methodology utilizing multiple machine learning algorithms, for our report, Logistic Regression, SVM, and RF, with a physical model-based simulation using Simulink to validate hypotheses and reveal new insights and findings. The general workflow of this study is represented as shown in 1.

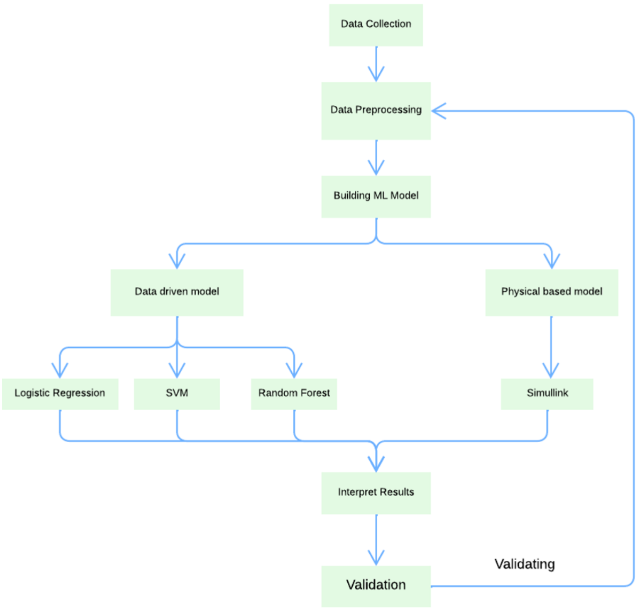

Fig. 1. Implementation Workflow

### A. Data Collection

During the data acquisition phase, this study aims to collect temperature data using the DHT22 sensor, vibration data through the built-in LSM6DS3 module, and audio signal data from the fan motor's ECM. Initially, data will be recorded from the fan operating under normal conditions, followed by the simulation of fault scenarios. Essentially, the author utilizes two Arduino Nanos for data collection: one is for capturing accelerometer data, while the other processes the data from the sensors. These two Arduinos are connected through Serial Communication using the TX and RX terminals on each device. The outputs are transmitted from the Arduino sketch through `Serial` communication used in Python for every 1 second, and subsequently stored in a CSV file. The data includes the `Timestamp`, `Temperature` measured in degrees Celsius, `Sound Frequency`, three axis accelerometer `Accel_X`, `Accel_Y`, and `Accel_Z`.

To simulate faulty conditions, the study uses a hair dryer to create a scenario in which the fan motor experiences excessive load. Fans depend on airflow to regulate the motor; therefore, inadequate cooling hinders effective heat dissipation, resulting in overheating. For vibration data, the author attaches tape to the fan blades to induce imbalances, misalignments, and bearing wear, reflecting real-world scenarios. This approach also introduces faulty data in the sound measurements, as the imbalances cause the fan to operate with vibrations, leading to increased noise levels.

The data collection spans a total of 6 hours. Initially, the study records data from the fan operating under normal conditions for 3 hours. The remaining time is allocated for collecting faulty data: 1 hour is dedicated to simulating overheating, the following hour addresses imbalances and bearing wear, and the final hour combines all these faulty conditions. Essentially, if any of these indicators are detected, the model will classify it as faulty. This approach provides a robust approach for measuring both faulty and normal condition data.

### B. Data Preprocessing

After completing the data collection, the study advances to the preprocessing stage, where the dataset consists of exactly 6,820 samples. Before proceeding to the EDA step, it is essential to convert the recorded frequency sound data into a suitable format using the Fast Fourier Transform (`fft`) function from the `scipy` package. This transformation allows for the analysis of the frequency components within the data, enabling a clearer understanding of the underlying patterns and characteristics before conducting further analysis. Following the successful collection and verification of sensor data, the author proceeds to label the `Target` variable according to the interval transitions from normal conditions to faulty conditions. In the next step, after handling `Null` values, feature scaling is performed, followed by encoding the `Target` output variable. Subsequently, the distribution of the `Target` variable is checked, and if imbalances are found, oversampling or undersampling is applied to balance the dataset in terms of the `Target` variables. To proceed with building the machine learning algorithm, the data is firstly split into two sets: a training set and a testing set. This approach ensures robust results in predicting whether the fan has malfunctioned or not.

### C. ML Models

Our objective is to predict faulty symptoms of the machine, focusing on binary classification with `Normal` and `Faulty` labels. To address this problem, the author intends to utilize Logistic Regression, SVM, and Random Forest, as these algorithms are particularly well-suited for binary classification tasks. Additionally, this study combines all three models to leverage the strengths of each, aiming to achieve a more comprehensive estimate of the output variable.

*1) Logistic Regression:* models the relationship between the input features—Temperature, Vibration, and Sound data—and the probability of the `Target` variable indicating whether the machine is faulty. This algorithm estimates the likelihood that a given set of input features corresponds to a faulty condition by fitting a logistic function to the data. It transforms the input features into a probability value between 0 and 1, where values closer to 1 indicate a higher likelihood of the machine being faulty, while values closer to 0 suggest normal operation.

*2) SVM:* finds the optimal hyperplane that effectively separates the data into two classes: Normal and Faulty. This hyperplane is determined by maximizing the margin between the closest data points of each class, known as support vectors. By doing so, SVM not only makes sure that the classes are distinctly separated but also enhances the model's generalization capabilities when predicting new, unseen data.

*3) RF:* an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (Normal or Faulty). Each tree is created using a random subset of the training data and a random subset of features (Temperature, Vibration and Sound data), promoting diversity among the trees. This randomness enhances the model's performance and robustness. While individual decision trees may be susceptible to overfitting, the aggregation of numerous trees in a Random Forest typically results in better generalization to unseen data. The final prediction is determined by the class that receives the majority of votes, representing the most common class in the `Target` variable.

*4) Voting Classifier:* An ensemble learning method that enhances classification performance by combining predictions from multiple base models, capturing various data patterns and reducing overfitting risk. It utilizes two voting methods: hard voting, where each classifier votes for a class and the majority wins, and soft voting, where classifiers provide probability distributions over classes, and the class with the highest average probability is selected. Soft voting is particularly effective for imbalanced classes or well-calibrated probabilities.

*5) Hyperparameters Tuning:* Once the models are built, the next step involves identifying the optimal parameters and hyperparameters that will yield the best performance for the specific dataset. This process is crucial for achieving a good fit and enhancing the overall effectiveness of the model. To systematically explore various combinations of parameters and hyperparameters, one can utilize the `GridSearchCV` function from the `sklearn` library.

`GridSearchCV` performs an exhaustive search over specified parameter values for an estimator. It evaluates each combination of parameters using cross-validation, ensuring that the model's performance is assessed on different subsets of the data. This helps in selecting the combination that maximizes model accuracy and minimizes overfitting, ultimately leading to a more reliable and generalizable model.

### D. Evaluation

After the models have been built and the parameters and hyperparameters have been optimized, the next step is to evaluate their performance using various evaluation metrics. For this study, the metrics employed will include accuracy_score, recall_score, f1_score , confusion_matrix, and classification_report. These metrics provide a comprehensive overview of the models' performance on unseen data, allowing for an in-depth assessment of their effectiveness in predicting outcomes. Additionally, the ROC curve and AUC are calculated using the roc and auc functions to provide further insights into the models' performance. These metrics help assess the trade-off between true positive rates and false positive rates, offering a visual representation of how effectively the models distinguish between classes. Analyzing the ROC curve and AUC reinforces the decision-making process when selecting the most suitable model for this type of dataset. Finally, at the conclusion of the evaluation stage, if any key components fail—such as if the evaluation metrics are too low, there is a risk of underfitting or overfitting, or the dataset presents challenges for the models to learn effectively—it may be necessary to revisit the data processing step. In some cases, it might even require returning to the data collection stage to gather additional or improved data. This iterative process ensures that the models are built on a solid foundation, ultimately enhancing their performance and reliability.

## IV. Results and Discussion

In the earlier sections, the study focused on outlining the theoretical key factors related to predicting the failure of the fan motor. This section will shift its focus to interpreting the recorded results obtained by the author, assessing whether the model performs effectively on the dataset, and evaluating the quality of the data as a reliable source for addressing this specific problem.
Before delving into the results, it is essential to make an assumption to develop an intuition regarding the problem. The data collection process is divided into two distinct parts: one for recording normal operational data and another for capturing faulty data. It is reasonable to expect that the normal data will differ significantly from the faulty data. For instance, when the fan motor operates normally at a temperature of 30 degrees Celsius, any excessive overload will likely cause the motor to overheat, resulting in a sustained increase in temperature over time. Thus, it becomes intuitively clearer to analyze and interpret the data and patterns within the dataset.

### A. EDA

Following the essential steps of data transformation and encoding, the study now focuses on examining the distribution of the classes Normal and Faulty. It is important to determine whether one class significantly dominates the other. If a class imbalance is observed, it may be necessary to perform oversampling or undersampling to ensure a more balanced dataset. This adjustment jusst helps improve the model's performance and reliability when predicting outcomes. The figure 2 shows that there is little difference in the number of samples classified as Normal versus Faulty, with 1 representing Normal and 0 representing Faulty. Therefore, there is no need to apply any sampling techniques.

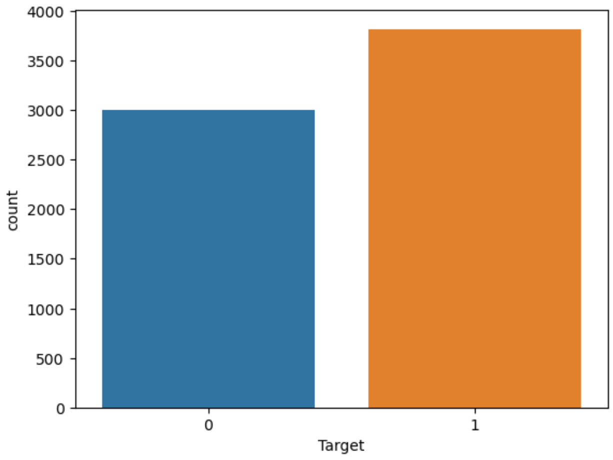

Fig. 2. Classes distribution of the Target variable

### B. Fine-Tuned Models

This section will discuss the key hyperparameters that require fine-tuning for each model using GridSearchCV. The parameters will be configured with cv set to 5 for five-fold cross-validation, verbose set to 2 for detailed output, and n_jobs set to $-1$ to utilize all available CPU cores to speed up the fine tuning time.

TABLE I
Best Hyperparameters and Cross-Validation Scores of Each Model

| Model | Best Hyperparameters |
|---|---|
| Logistic Regression | {C: 1, multi_class: ’auto’, solver: ’liblinear’} |
| SVM | {C: 100, decision_function_shape: ’ovo’, gamma: ’scale’} |
| Random Forest | {max_depth: None, max_features: None, min_samples_leaf: 1, min_samples_split: 5, n_estimators: 100} |

### C. Computational Metrics

TABLE II
Evaluation Metrics Overview for Each Model

| Metric | Logistic Regression | SVM | Random Forest | Voting Classifier |
|---|---|---|---|---|
| Accuracy | 0.9208 | 0.9917 | 0.9917 | 0.9892 |
| Precision | 0.99 | 1.00 | 1.00 | 1.00 |
| Recall | 0.82 | 0.98 | 0.98 | 0.98 |
| F1-Score | 0.90 | 0.99 | 0.99 | 0.99 |

At first glance, all the models presented demonstrate high performance, not only in terms of accuracy but also in precision and recall, with all metrics considerably exceeding 90%, which is considered excellent. Focusing on the Logistic Regression model, it achieved an accuracy of approximately 92%, indicating that the model effectively predicts the fan motor's state (whether it fails or not) in most instances. However, the recall of 0.82 suggests that the model overlooks about 18% of actual failures. This oversight can be critical in situations where failing to detect a motor failure could lead to significant operational challenges. Interestingly, both the SVM and Random Forest models exhibit identical evaluation metrics, with an accuracy of approximately 99%. A precision of 1.00 signifies that these models make no false positive predictions; whenever they predict a failure, it is accurate. The recall of 0.98 indicates that they correctly identify 98% of actual failures, resulting in only 2% of missed predictions. This makes both models particularly well-suited for scenarios where minimizing both false positives and false negatives is crucial, such as in fan motor monitoring. The Voting Classifier also demonstrates strong performance, achieving an accuracy of approximately 98.92%. Similar to the other models, it attains perfect precision, indicating that every predicted failure is correct. However, with a recall of 0.98, it shares the 2% rate of missed failures observed in the other models. This result suggests that, although the Voting Classifier combines predictions from all three models, its performance is slightly diminished due to the inclusion of Logistic Regression, which does not perform as well as the SVM and RF models.

To further support our findings, it needs to perform ROC and AUC analysis as well as the confusion matrix to provide a comprehensive overview of each model's performance.

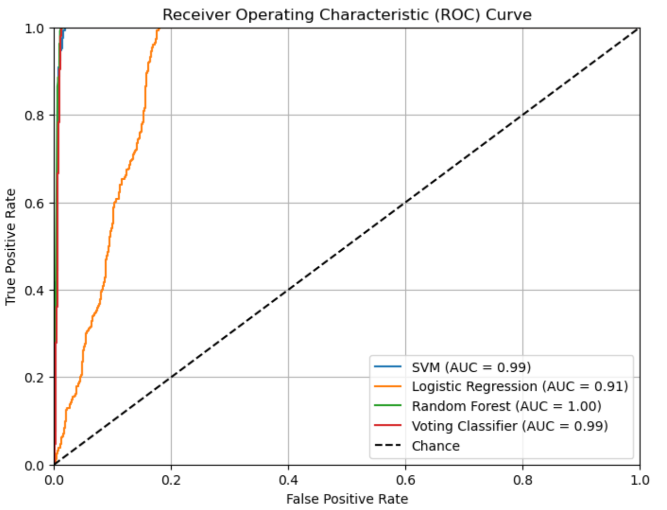

Fig. 3. ROC Cruve and Calculated AUC of the 4 models

Based on the ROC curve and the calculated AUC 3, along with the confusion matrices for SVM and Random Forest 4, and the evaluation metrics discussed earlier, the most suitable model for this problem is either SVM, Random Forest, or a combination of both using a Voting Classifier. The exceptionally high performance of these models, which might be mistaken for overfitting, is actually due to the intrinsic nature of the data. As previously assumed, when the motor exhibits symptoms of failure, the sensor readings differ significantly from those during normal operation, enabling the machine learning models to learn and generalize effectively.

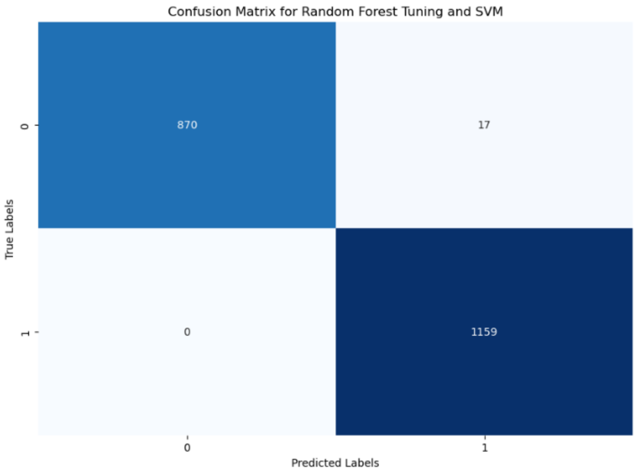

Fig. 4. Confusion Matrix of RF and SVM

## D. Limitation

The most significant and challenging aspect of predicting failures, RUL, and PHM does not solely lie in model building, data analysis, or data capture. Instead, the difficulty arises from the accessibility of the equipment. Predictive maintenance can be applied not only to small-scale equipment but also to an entire system of interrelated devices. However, gaining access to such equipment is often difficult, and these machines are typically expensive to acquire, making them not suitable for most individuals who wish to conduct research using real-world data. Consequently, many researchers rely on simulation data. Another limitation of this study is that it only captures a few days of data recording, which may not totally reflect real-world scenarios where most equipment needs to be monitored continuously over months or years.

## E. Future Works

This implementation can be expanded beyond the current report, which primarily focuses on a physical-based model using Simulink and incorporates fundamental and basic blocks representing key components of a fan motor. The model can be enhanced into a more complex system that integrates multiple sensors and equipment working together.Besides, a current sensore can be integrated, which monitors the electrical current flowing to machines, motors, or other equipment, can detect when a machine or motor is drawing more current than usual, potentially indicating mechanical issues such as increased friction or failing bearings. Furthermore, this study's approach can be applied to various types of equipment and combined with other methodologies to achieve a more robust estimate of the model's behavior regarding predictive maintenance.

Lastly, it has the potential to monitor specific equipment over extended periods, such as months or years.

## V. Conclusion

This study advances existing approaches by integrating a data-driven methodology that employs sensor fusion alongside machine learning algorithms with a physical-based model using Simulink to achieve a comprehensive understanding. The findings indicate that SVM and RF models are the most effective for predicting failures in the fan motor by introducing the predictors Temperature, Vibration and Sound Data.

## References

[1] T. P. Carvalho, F. A. A. M. N. Soares, R. Vita, R. da P. Francisco, J. P. Basto, and S. G. S. Alcalá, "A systematic literature review of machine learning methods applied to predictive maintenance," Computers & Industrial Engineering, vol. 137, no. 1, p. 106024, Nov. 2019, doi: https://doi.org/10.1016/j.cie.2019.106024.

[2] J. Wan et al., "A Manufacturing Big Data Solution for Active Preventive Maintenance," IEEE Transactions on Industrial Informatics, vol. 13, no. 4, pp. 2039–2047, Aug. 2017, doi: https://doi.org/10.1109/tii.2017.2670505.

[3] P. Movahed, S. Taheri, and A. Razban, "A bi-level data-driven framework for fault-detection and diagnosis of HVAC systems," Applied Energy, vol. 339, p. 120948, Jun. 2023, doi: https://doi.org/10.1016/j.apenergy.2023.120948.

[4] N. Es-sakali, M. Cherkaoui, M. O. Mghazli, and Z. Naimi, "Review of predictive maintenance algorithms applied to HVAC systems," Energy Reports, vol. 8, pp. 1003–1012, Nov. 2022, doi: https://doi.org/10.1016/j.egyr.2022.07.130.

[5] Z. Chen et al., "A review of data-driven fault detection and diagnostics for building HVAC systems," Applied Energy, vol. 339, pp. 121030–121030, Jun. 2023, doi: https://doi.org/10.1016/j.apenergy.2023.121030.

[6] M. S. Mirnaghi and F. Haghighat, "Fault detection and diagnosis of large-scale HVAC systems in buildings using data-driven methods: A comprehensive review," Energy and Buildings, vol. 229, p. 110492, Dec. 2020, doi: https://doi.org/10.1016/j.enbuild.2020.110492.

[7] T. Zonta, C. A. da Costa, R. da Rosa Righi, M. J. de Lima, E. S. da Trindade, and G. P. Li, "Predictive maintenance in the Industry 4.0: A systematic literature review," Computers & Industrial Engineering, vol. 150, p. 106889, Dec. 2020.

[8] P. Koprinkova-Hristova, Mincho Hadjiski, L. Doukovska, and S. Beloreshki, "Recurrent Neural Networks for Predictive Maintenance of Mill Fan Systems," International Journal of Electronics and Telecommunications, vol. 57, no. 3, 2024, doi: https://journals.pan.pl/publication/100978.

[9] F. Alamr and F. Alamr, "Industrial Duct Fan Maintenance Predictive Approach Based on Random Forest," CS & IT Conference Proceedings, vol. 10, no. 5, 2024, Accessed: Sep. 17, 2024. [Online]. Available: https://csitcp.org/abstract/10/105csit16

[10] M. B. Hadjiski, L. A. Doukovska, and S. L. Kojnov, "Nonlinear Trend Analysis of Mill Fan System Vibrations for Predictive Maintenance and Diagnostics," International Journal of Electronics and Telecommunications, vol. 58, no. 4, pp. 351–356, Dec. 2012, doi: https://doi.org/10.2478/v10177-012-0048-9.

[11] C. Murphy, "Choosing the Most Suitable Predictive Maintenance Sensor," 2020. Accessed: Sep. 26, 2024. [Online]. Available: https://www.allelcoelec.my/datasheets.b5/ADXL1004BCPZ.pdf

[12] M. Trčka and J. L. M. Hensen, "Overview of HVAC system simulation," Automation in Construction, vol. 19, no. 2, pp. 93–99, Mar. 2010, doi: https://doi.org/10.1016/j.autcon.2009.11.019.

[13] E. Gilabert, S. Fernandez, A. Arnaiz, and E. Konde, "Simulation of predictive maintenance strategies for cost-effectiveness analysis," Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, vol. 231, no. 13, pp. 2242–2250, Apr. 2015, doi: https://doi.org/10.1177/0954405415578594.

[14] P. Riederer, "MATLAB/SIMULINK for building and HVAC simulation : state of the art," Osti.gov, Jul. 2005. https://www.osti.gov/etdeweb/biblio/20685518 (accessed Sep. 28, 2024).

[15] M. Achouch et al., "On Predictive Maintenance in Industry 4.0: Overview, Models, and Challenges," Applied Sciences, vol. 12, no. 16, p. 8081, Aug. 2022, doi: https://doi.org/10.3390/app12168081