Exploring the Impact of Data Selection to Support Development of Predictive Twins

Ashit Harode, ashit02@vt.edu
Ph.D. Candidate, Myers-Lawson School of Construction, Virginia Tech, Blacksburg, VA, USA

Akhileswar Yanamala, akhileswar@vt.edu
Ph.D. Student, Myers-Lawson School of Construction, Virginia Tech, Blacksburg, VA, USA

Walid Thabet, Ph.D., CM-BIM, thabet@vt.edu
Professor, Myers-Lawson School of Construction, Virginia Tech, Blacksburg, VA, USA

Abstract

Current research often focusses on the use of Digital Twins for data collection and visualization and Machine Learning for data analysis to develop prediction models. However, these research lack discussion on how data to develop predictive models/twin needs to be selected and how they contribute to the models' accuracy and effectiveness. In this paper the authors focus their attention on how the data needs to be selected for the development of accurate and cost-effective prediction models. The paper developed two machine learning models, one containing redundant data and another with redundant data combined as single data points. During testing, both models achieved similar accuracy of 0.86, highlighting that redundant data did not add to the accuracy of the predictive model/twin. The results also show that collection of redundant variables can be eliminated to reduce the cost of data capture and storage.

Keywords: Machine Learning, Predictive Maintenance, Air Handling Unit, Fault Detection, Data Analysis

1 Introduction

With the advancements in digitization and data processing, the Architecture, Engineering, and Construction industry continues to improve the design, construction, and carbon footprint for modern buildings. As the maintenance cost of a building continues to account for more than 65% of its facility operation and management cost, building maintenance has become an integral part of facility management (Cheng *et al.* 2020). Facility managers often utilize one of three maintenance strategies or programs to perform maintenance of their buildings (Honeywell 2021): (1) Reactive maintenance, also known as corrective maintenance, refers to an equipment maintenance strategy where maintenance is only performed once an asset has broken down, , (2) Preventive maintenance (PvM) is a time-based maintenance program that is triggered at a predefined time interval based on usage patterns, the criticality of equipment to the building functions, and the historical performance of the asset, and (3) Predictive maintenance (PdM) relies on periodic or continuous real-time monitoring of an equipment's operational conditions to predict future trends in the equipment's performance. Using sensors to collect data in real-time that is then fed into AI-enabled applications, advanced data analysis tools and algorithms can identify potential problems and predict maintenance requirements before equipment breaks down. This continuous monitoring of the operational performance of the equipment makes this a real-time condition-based maintenance.

Condition-based Predictive Maintenance provides several benefits when compared to reactive or preventive maintenance. Predictive Maintenance allows for anticipating when an asset may need attention. Issues can then be addressed proactively, resulting in minimized downtime. This capability allows a facility manager to know when to dispatch the technician with the right tools. This provides for better asset performance, fewer repairs and less downtime. With an abundance of data coming in from multiple assets, using predictive maintenance, data can be analyzed across multiple channels throughout the facility to assist facility managers by drawing attention to the assets that are critical to building operations and are predicted to fail in the near future.

The U.S. Department of Energy reported that Predictive Maintenance can save 8% to 12% more energy when compared to preventive maintenance and up to 40% when compared to reactive maintenance. When comparing asset downtime, McKinsey & Co. reported that predictive maintenance can reduce asset downtime by 30% to 50% and increase life expectancy by 20% to 40% (Honeywell 2021).

The key to a predictive decision making process in Predictive Maintenance is data (Zonta *et al.* 2020). While developing prediction twins it is essential that useful knowledge is extracted from the data related to life-cycle of an asset or building system (Baptista *et al.* 2018). Integration of data from multiple sources like monitoring data, maintenance records, and work orders is required to support decision making for predictive maintenance (Cheng *et al.* 2020). However, processing data from a large number of data points does not always contribute to the accuracy of the prediction twins. Collected data can include redundant information that does not improve the accuracy of the model but increases the computational requirements or can contain noise data that can result in decreased accuracy.

This research work focuses on answering the question: "What are the effects of understanding available data on the development of Prediction model to support Predictive Twin?" This work will add to the body of knowledge by drawing attention to the data selection and analysis part of predictive maintenance. Appropriate understanding of data will not only improve the predictive twin's accuracy and goal, but will also make the predictive maintenance implementation more cost effective. When data requirements are identified, less money is spent on data collection and storage as only the required amount of data is collected. Another major benefit of understanding the data requirements is that it makes the predictive twins more explainable and understandable, increasing their use and adoption to support facility management.

The research utilizes published experimental data for a roof top unit-variable air volume system (RTU-VAV) installed in a two-story light-commercial 3,200 sqft experimental facility designed to emulate a 1980s-era office building. The data was generated by the Oak Ridge National Laboratory (ORNL) in Tennessee, USA and published by the Lawrence Berkeley National Laboratory (Granderson *et al.* 2023). The data selected for analysis comprised 60 data points and included faulted and unfaulted scenarios of the damper positions of the RTU. Data for faulty damper scenarios included damper open position stuck at 5%, 10%, 50%, and 100%. Using the published data, the researcher first conducted a correlation analysis to determine the relevance of the different data points to the damper position fault scenario and identify any redundancy in the data. The data sets for faulty and fault free scenarios were then combined for a first run to train and test a predictive ML model using all the 60 datapoints. The training run time and accuracy of results were recorded and documented. Using the results of the correlation analysis, input data points (features) to the ML model were aggregated and other redundant data points were eliminated and a second run was conducted to train and test the ML model. The training run time and accuracy of results were again recorded and documented. The accuracy and training time of the model with and without the redundant data were compared to draw conclusions on the effects of data analysis on the prediction model to support the development of predictive twins.

Section 2 of this paper provides a literature review of the current state of the art for predictive maintenance and highlights the research gaps. Section 3 discusses in detail the methodology adopted to run the comparative analysis of the two ML predictive models to highlight the importance of robust data selection process towards developing predictive twins. Section 4 discusses the implementation of the methodology. The results of the implementation are discussed in section 5. Followed by discussion on the findings and future work to this research in section 6.

2 Literature Review

Industry 4.0, a term often used to describe the "fourth industrial revolution" has become synonymous with the use of enabling technologies related to connectivity, amount of data, new devices, inventory reduction, customization and controlled production (Zonta *et al.* 2020). Industry 4.0 has caught the attention of manufacturing industry enabling increased digitization and automation to develop digital value chain of product lifecycle from concept to use and maintenance (Hossain & Nadeem 2019). In the manufacturing industry, Industry 4.0 has aided in improving product quality while decreasing development cost and time (Hossain & Nadeem 2019). Even though the characteristics of construction projects differ from products developed by the manufacturing industry, the concepts of Industry 4.0 can be translated to match the needs of the construction industry (Hossain & Nadeem 2019). This translation leads to the coining of the term Construction 4.0.

Construction 4.0 is supported by use of transformative digital technologies such as Building Information Modelling (BIM), Common Data Environment (CDE), unmanned aerial systems, Augmented Reality, artificial intelligence, cybersecurity, big data and analytics, blockchain, and laser scanner (Forcael *et al.* 2020).

Benefiting from Construction 4.0, predictive maintenance is a facility maintenance and management strategy that relies on digital technologies like integration with internet of things, artificial intelligence, and integrated systems (Rockwell Automation). Condition-based Predictive Maintenance requires collection of asset or building system operational data transmitted by sensors, maintenance records, and work orders (Cheng *et al.* 2020). Condition-based Predictive Maintenance can improve facility maintenance and management by increasing equipment life, improving efficiency, and reducing labor cost (Hosamo *et al.* 2022). However these benefits are dependent on the accuracy and effectiveness of the developed predictive maintenance model (Florian *et al.* 2021). Using technologies such as machine learning, building information modelling, and internet of things, research has been conducted to develop and implement effective predictive maintenance strategies. Table 1 summarizes literature focused on the use of predictive maintenance and outlines the types of data and algorithms used, and results achieved.

Table 1. Literature focused on use of Predictive Maintenance Strategies.

Literature	Data Used	Algorithms Used	Results
Hosamo et al. (2022)	BIM, Real-Time Sensor Data, and Facility Management Data	Artificial Neural Network, Support Vector Machine and, Decision Tree	Prediction of faults in AHUs using machine learning is both functional and beneficial towards facility maintenance.
Cheng et al. (2020)	Real-Time Sensor Data, BIM, and Facility Management Data	Artificial Neural Network, and Support Vector Machine	Prediction of MEP components' condition was possible.
Al-Aomar *et al.* (2024)	Data from building management system and computerized maintenance management system	Support Vector Machine, k-Nearest Neighbor, and Decision Tree	Prediction of asset condition based on scale of critical to excellent
Marzouk & Zaher (2020)	Images of different fire protection systems	Convolutional Neural Network	Validated the utility of artificial intelligence (AI) in identifying assets that require proactive maintenance
Assaf *et al.* (2020)	Past air conditioning complaints and weather-related data	Text Mining and Nonlinear Autoregressive Exogenous (NARX)	Predicted building occupants' complaints

Assaf & Srour (2021)	Unstructured occupant complaint logs	Multi-Layer Perceptron	Assist facility managers in better planning staffing based on predicted complaints.
Masdoua *et al.* (2022)	HVAC systems data generate by Pacific Northwest National Laboratory	Decision Tree, Random Forest, and SVM	Detect and diagnose AHU sensor faults using machine learning algorithm

The literature review showcases the diversity of research conducted to improve the effectiveness of predictive maintenance through the implementation of different machine learning algorithms. In parallel, the research also explores areas of facility maintenance that can benefit from predictive maintenance like predicting asset's conditions, and occupant complaints. However, the scope of these research does not focus on the importance of data selection to develop a predictive model. Hosamo *et al.* (2022) while developing a predictive model briefly introduces the importance of feature selection in machine learning approach to filter out redundant and noisy data but fails to expand on the topic in terms of effects of feature selection on the accuracy and training time of the prediction model.

This paper aims to fill this research gap by highlighting the importance of data analysis for developing a machine learning model for predictive maintenance. The authors in this paper explore how redundant data can be identified and eliminated and highlights how data analysis can affects the goal selection of the machine learning model towards predictive maintenance.

3 Methodology

This study explored the importance and need for defining data requirements to develop predictive models in support of predictive twins. The research utilized experimental hourly data for a roof top unit-variable air volume system (RTU-VAV) generated by the Oak Ridge National Laboratory (ORNL) in Tennessee, USA and published by the Lawrence Berkeley National Laboratory (Granderson *et al.* 2023). The data included 60 data points measured for the RTU-VAV system with the RTU Outside Air Damper (OAD) position in a fault-free operation as well as faulty operation. The data for the faulty damper position included data measured while the damper open position was manually forced to be mechanically stuck at 5%, 10%, 50%, and 100%. Data for a fault-free RTU damper position included data during normal damper modulation between 10% and 100% based on economizer instructions to open or close the outside air damper (OAD).

In an HVAC system an economizer is used to utilize the outside air to condition the facility. An economizer evaluates outside air temperature and humidity levels, and when appropriate, uses the outside air to cool buildings, reducing the load on the mechanical cooling system. A mixing box inside the RTU combines outside air with return air in calculated percentages using dampers to supply well-conditioned air to the building spaces. It should be noted that the outside air (OA) damper is synched with the return air (RA) damper in the RTU-VAV system to provide 100% air. For example, a 30% OA damper position requires a 70% RA damper position to maintain a balanced air pressure in the space. HVAC economizers use logic controllers and sensors to get an accurate read on outside air quality. As the economizer detects the right level of outside air to bring in, it utilizes the OA and RA dampers to control the amount of air that gets pulled in, recirculated and exhausted from a building space. For example, when the outside air temperature and humidity levels can effectively condition facility spaces, the outside air damper (OAD) is opened to 100% and the return air damper (RAD) is fully closed. During cold or hot weather, outside air will be mixed with return air to condition the air to the required set point used before being supplied to the spaces. This might require opening the outside air damper to 40% and the return air damper to 60%.

Using the data published, the research was divided into three steps presented in Figure 1. Step 1 focused on a search to acquire existing data related to HVAC system operation. Literature review was conducted to identify literature focused on providing data to the scientific community to support research and development of predictive twins and to support predictive maintenance. Once the needed data was identified, appropriate data was filtered and labeled to train and test two machine learning models. A Spearman correlation analysis was conducted in step-2 to calculate correlation between the data points (features) and the faulty/non-faulty status of the OA damper. Based on this correlation, redundant features would be identified and later combined or eliminated from the analysis. In step-3 two analysis runs are conducted for training and testing a neural network. In the first run, the neural network was trained and tested using 58 data points as input (features), and two data points (faulty/non-faulty status, and damper position) as output (labels). In the second run, results from the correlation analysis are applied to eliminate or combine redundant data points and reduce the number of input data (features) to 22 data points. Data points used as output is not changed. The new model was trained and tested.

The two models were compared to analyze the impact of reducing the total number of input data points (features) from 58 to 22 on model accuracy and training time and to draw conclusions on the need to specify the data requirements to use as input while developing predictive twins.

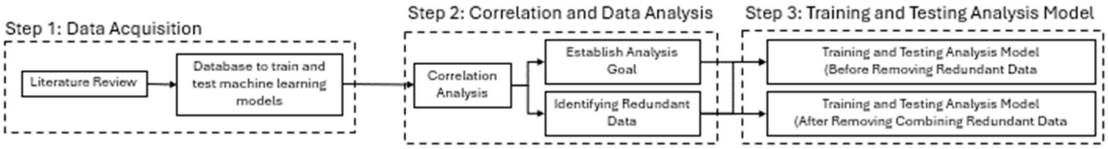
Step 1: Data Acquisition
Literature Review
Database to train and test machine learning models
Step 2: Correlation and Data Analysis
Correlation Analysis
Establish Analysis Goal
Identifying Redundant Data
Step 3: Training and Testing Analysis Model
Training and Testing Analysis Model (Before Removing Redundant Data
Training and Testing Analysis Model (After Removing Combining Redundant Data

Figure 1. Methodology adopted to highlight the importance feature selection.

4 Implementation

4.1 Step 1: Data Acquisition

To explore the impact of data selection on the implementation of Predictive Twins, the research explored sources for available fault detection and diagnostics data collected by others. Literature review was conducted to identify referenced open-source experimental fault detection and diagnostics data collected for mechanical systems. Ahern *et al.* (2023) provided a dataset collected with the objective of progressing the research work for fault detection and diagnostics. However, the datasets provided were unlabeled and included data issues like missing features, missing time interval, and inaccurate data. Lawrence Berkeley National Laboratory in support of Fault Detection and Diagnostics of Roof Top Units collected and made available various data for fault free and faulty runs of mechanical systems (Granderson *et al.* 2023). The data selected for this research was for a Roof Top Unit – Variable Air Volume (RTU-VAV) system collected by Oak Ridge National Laboratory in their light-commercial flexible research platform (FRP). The FRP depicts a 1980s-era two story 3,200 sq-ft facility with 10 spaces which are reserved for experiments and remains unoccupied with internal loads emulated. Due to the labeled nature and completeness of the data, it was found suitable and used in this research.

The data contained operational variables (data points) collected from 56 different sensors located in the HVAC system and the facility space. An additional 4 variables were also included in the dataset to depict the different seasons (Fall, Summer, Spring, and Winter) during which the data was collected. A description for some examples of the 60 variables is shown in Table 2.

Table 2. Example of variables collected by Granderson *et al.* (2023).

No.	Data Point Name	Data Point Abbreviation	Description	Unit
1	RTU: Outdoor Air Damper Control Signal	RTU_OA_DMPR_DM	Outside Air Damper Position	0-1
2	RTU: Outside Air Temperature	RTU_OA_TEMP	Outside Air Temperature	°F
3	RTU: Electricity	RTU_TOT_WATT	RTU Electricity Consumption Rate	W
4	Terminal: Room 102 Air Temperature	TERM_RM_TEMP_102	Room 102’s Ambient Temperature	°F
5	VAV Box: Room 102 Power Consumption	VAV_RM_WATT_102	Room 102’s VAV Box Power Consumption rate	W

These 60 data points define the feature space for the training of the predictive model. The features represented data describing the damper control signals, air temperatures, volumetric flow rate, electric and gas consumption rate for VAVs, fans, and RTU, heating and cooling temperature setpoint, room temperature and humidity, and, occupancy mode (Granderson *et al.* 2023). The data contained feature values belonging to fault-free runs and faulty runs for issues with dampers. Data for fault free cases was collected for a day immediately prior to the faulty-run. Faults were introduced at 12 am following the fault free run and data was collected for a day before restoring the normal run.

The research used data associated with the operation of the OA damper in fault-free and faulty modes. This included data for the damper operating in normal conditions (fault-free), and data collected while the damper was operating in a user-induced faulty condition. This included data collected while the damper is forced to be stuck at 5%, 10%, 50%, and 100% open position scenarios.

4.2 Step 2: Correlation and Data Analysis

The published damper fault detection and diagnostics data consisted of 60 data points. If all of these variables are used to develop a ML model to predict the operational status of the damper (fault-free or faulty), 60 variables would be used as input (features) with one output (label) indicating whether the system OA damper is running in fault-free or faulty mode. In this step, the researchers conducted a correlation analysis in order to determine if all 60 variables are relevant inputs and to identify the variables which had the most impact on the accuracy and the run time of the model. It is hypnotized that some of these 60 variables may be redundant and need to be combined or eliminated.

A Spearman Correlation, also known as Spearman's rank correlation coefficient, was used to assess the direction and strength of the monotonic relationships between the 60 variables. Specifically, we were interested to identify which out of the 60 variables were the highest indicators of whether the OA Damper is operating in a fault-free or faulty mode. In other words, we wanted to know what are the variables that had the highest correlation with the damper's operational status.

To facilitate the correlation analysis, a preliminary predictive goal of binary classification was defined where for a given data row the model would predict a faulty or fault-free run. As per this goal, the predictive model was provided with data that included all 60 data variables as features and each row of the data was labelled as faulty or fault-free run.

A spearman coefficient was calculated. Spearman coefficient is a "non-parametric rank statistical measure of the strength and the direction of the arbitrary monotonic association between two ranked variables or one ranked variable and one measured variable" (Xiao *et al.* 2016). Spearman coefficient is a convenient correlation matrix to use as it does not need to make assumption about the distribution frequency and linear relationship between the variables (Xiao *et al.* 2016). Since the distribution of data selected in this paper was not known, the Spearman coefficient is selected to perform correlation analysis.

Spearman Correlation Coefficient is calculated between features and features, and features and labels. The value of Spearman Coefficient ranges from -1 to 1. A value of 1 shows a high positive correlation i.e. increase in the value of one result in increase in the value of another. -1 indicates a high negative correlation i.e. increase in the value of one result in decrease in the value of another. If no correlation exists a value of 0 is assigned. Figure 2 shows the results of correlation analysis for the dataset.

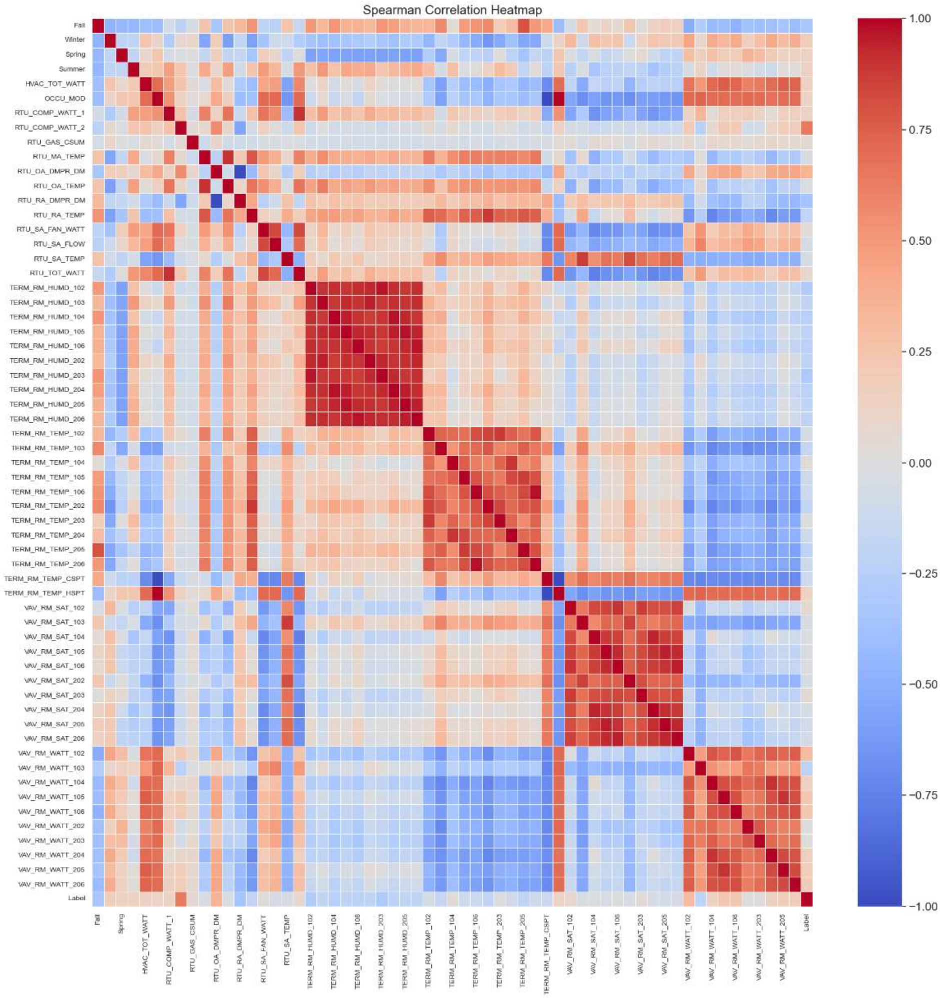

Figure 2. Result of Correlation Analysis

As shown in Figure 2, the x axis and y axis are identical with the first 60 rows/columns representing the 60 features provided in the dataset and the last row/column representing the labels i.e., faulty or fault-free run.

Figure 2 shows a high feature-label correlation between the OAD position and the label, and the RAD position and the label. The correlation results indicated that while predicting fault-free and faulty run accurately, the value of the OAD and the RAD position contributed the most. However, a deeper analysis of the data revealed that this high correlation was due to the fact that in the dataset, large quantity of data rows that had the OAD open position value as 5%, 10%, 50%, or 100% belonged to the faulty run class. This created a class imbalance in the data. If the dataset gets used to predict fault free runs from faulty runs, the trained prediction model will most likely label all data with OAD values of 5%, 10%, 50%, or 100% as faulty run irrespective of other features.

Based on this analysis, the selected dataset cannot be used for the binary classification of fault-free and faulty run with OAD and RAD position present as features in the dataset, as this would result in a class imbalance. Therefore, modifications were made to the training of the prediction model. The OAD and RAD position variables were removed as features, and the OAD position was added as a label. Based on this initial modification, the goal of the predictive model was not only to predict the status of the OAD position (i.e. faulty or fault free), but also to predict the damper position.

Figure 2 also shows a high feature-feature correlation between the humidity of different rooms, temperature of different rooms, VAV box supply air temperature for different room, and VAV boxes power consumption for different rooms. To reduce redundancy, variables that showed high feature-feature correlation were combined using a calculated average value..

Based on the results of the correlation analysis, the OAD and RAD position variables were removed from the features to eliminate class imbalance. The label space of the dataset was also modified to allow for the prediction of faulty and fault free run and damper position. This modification created dataset 1 of this research which included 58 features and 2 labels. The dataset was further modified by combining redundant features resulting in a further reduction of the features from 58 to 22 creating dataset 2. The features that were combined included variables that provided information on the humidity levels in different rooms, temperature in different rooms, VAV supply air temperature for different room, and VAV power consumption for different rooms.

4.3 Step 3: Training and Testing Prediction Models

The two datasets, one using the 58 data points and the other with 22 data points were used to develop two ML models used for the analysis. These datasets were trained and tested using an Artificial Neural Network (ANN). The first training was conducted with the first dataset containing the 58 variables to set up a baseline accuracy and training time. The second training was conducted using the second aggregated dataset to determine the impact of combining redundant data on the accuracy and training time of the model.

The two datasets were first split into two sub-sets, with 80% being used for training and 20% being used for testing. The training dataset was used to train-validate-train the ANN. Using GridSearchCV, two hyperparameters were tuned during the train-validate cycle to find their optimum value. The options for hyperparameters value are defined in Table 3.

Table 3. Possible Hyperparameter values.

Hyperparameters	Possible values
Layers	[20], [40, 20], [45, 30, 15]
Activation Function	Sigmoid, Relu

Using the two optimized hyperparameter values, the two ANNs are retrained during the train cycle and the training time is measured. The trained ANNs are then tested using testing data. The training time and accuracy of each model were measured and compared.

5 Results

As a result of hyperparameter tunning, a 3 hidden layer AAN with 45 nodes in the first hidden layer, 30 nodes in the second hidden layer, and 15 nodes in the third hidden layer was the most preferred architecture. This AAN used ‘relu’ function as the activation function. The accuracy of both the trained ANN models was found to be 0.86. These accuracy measurements showed that combining redundant data had no adverse effect on the predictive capability of the model. This may be an indicator that, for future implementation of predictive models and predictive twin for facility maintenance and management, the cost and time of capturing real-time data can be reduced by identifying critical variables and eliminating redundant data points that has low impact. In this research, variables that had low impact on the fault detection and diagnostics of OAD included humidity, temperature, VAV box supply air temperature, and VAV boxes power consumption in the different building spaces.

Training time for the first predictive model with 58 variables was measured to be 65.26 seconds. Training time for the aggregated second model with 22 variables was 65.37 seconds. Further study needs to be conducted to draw conclusions on the increase in the training time when redundant data is removed. Figure 3 shows the comparison of training time and accuracy respectively for the ANN models trained using dataset 1 and dataset 2.

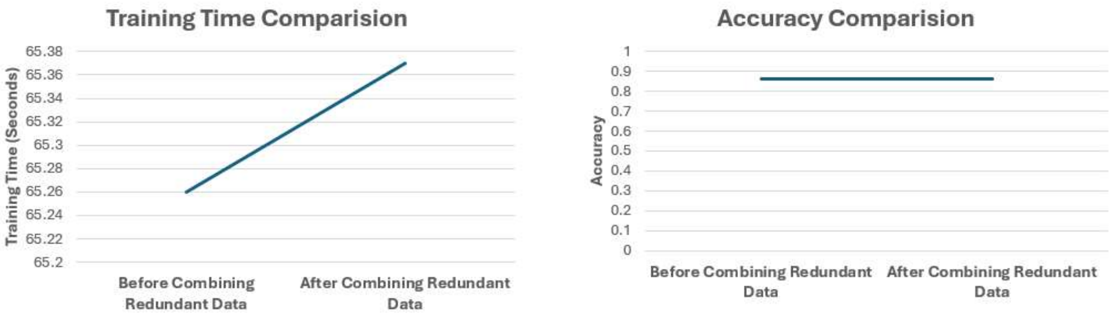

Figure 3. Training Time and Accuracy Comparison

6 Discussion and Conclusion

The research showed that combing and eliminating redundant data in the feature space to develop a predictive model capable of identifying fault free runs from faulty runs along with damper position has no adverse effect on the accuracy of the model. However, when thinking about the cost of implementation of predictive maintenance and predictive twins, collection and storage of redundant data points can increase the cost of implementation. Therefore, identifying clear data requirements plays an important role in reducing the cost of implementation.

The research work also utilized correlation analysis as a data analysis technique to measure the quality of data being used for predictive maintenance. Class imbalance in the training data can impact the effectiveness of the predictive model and needs to be identified early on. Data points highly correlated with the expected output need to be studied further to ensure that they provide necessary information to enable an accurate prediction. The correlation analysis in this research work identified OAD and RAD position as the data points providing the most information on predicting a fault free run from a faulty run. However, a deeper analysis showed that this correlation was a result of class imbalance which could significantly decrease the effectiveness of the predictive model. To mitigate this issue, the researchers eliminated the OAD and RAD position variables as features and changed the predictive goal to predict fault free runs from faulty runs and the current OAD position.

Through this research the authors set out to convey the importance of proper data selection, when training prediction models to support predictive maintenance and developing predictive twins. As predictive maintenance is a complex process, the model developed to support it will also be involved in complex decision making. This research shows the importance of proper data selection for the development of cost-effective prediction models. This research adds to the body of knowledge by providing an initial investigation into how important it is to select appropriate features when developing effective prediction models to reduce implementation costs. The authors hope that as research in the field of predictive twins matures, more attention is given to data selection.

As part of the continued research agenda, the authors plan to improve on the results of this research by conducting more detailed analysis on the impact of data selection on the training time and accuracy. The authors would also like to test different methods to identify, combine and/or eliminate redundant data and to address data challenges like class imbalance.

In this research Spearman Correlation Coefficient was used to understand the feature-feature and feature-label relationships. As part of future research, the authors will explore different correlation matrix and feature selection methods to provide comparative analysis on how different methods can be selected based on predictive maintenance requirements.

Acknowledgements

The authors would like to thank the Construction Infrastructure Research Affiliate Program at Myers-Lawson School of Construction for providing necessary funding to undertake this research work and contribute to the body of knowledge. The view and findings expressed in this paper are those of the authors and do not reflect those of the funding agencies.

References

Ahern, M., O'Sullivan, D.T.J. & Bruton, K. (2023) A dataset for fault detection and diagnosis of an air handling unit from a real industrial facility. *Data in Brief.* Vol. 48. pp. 109208. https://doi.org/https://doi.org/10.1016/j.dib.2023.109208.

Al-Aomar, R., AlTal, M. & Abel, J. (2024) A data-driven predictive maintenance model for hospital HVAC system with machine learning. *Building Research & Information.* Vol. 52. No. 1-2. pp. 207-224.

Assaf, S., Awada, M. & Srour, I. (2020) 'Data driven approach to forecast building occupant complaints'. *Construction Research Congress 2020,* 2020. American Society of Civil Engineers Reston, VA, pp.172-180.

Assaf, S. & Srour, I. (2021) Using a data driven neural network approach to forecast building occupant complaints. *Building and Environment.* Vol. 200. pp. 107972.

Baptista, M., Sankararaman, S., de Medeiros, I.P., Nascimento Jr, C., Prendinger, H. & Henriques, E.M. (2018) Forecasting fault events for predictive maintenance using data-driven techniques and ARMA modeling. *Computers & Industrial Engineering.* Vol. 115. pp. 41-53.

Cheng, J.C.P., Chen, W., Chen, K. & Wang, Q. (2020) Data-driven predictive maintenance planning framework for MEP components based on BIM and IoT using machine learning algorithms. *Automation in Construction.* Vol. 112. pp. 103087. https://doi.org/https://doi.org/10.1016/j.autcon.2020.103087.

Florian, E., Sgarbossa, F. & Zennaro, I. (2021) Machine learning-based predictive maintenance: A cost-oriented model for implementation. *International Journal of Production Economics.* Vol. 236. pp. 108114.

Forcael, E., Ferrari, I., Opazo-Vega, A. & Pulido-Arcas, J.A. (2020) Construction 4.0: A literature review. *Sustainability.* Vol. 12. No. 22. pp. 9755.

Granderson, J., Lin, G., Chen, Y., Casillas, A., Wen, J., Chen, Z., Im, P., Huang, S. & Ling, J. (2023) A labeled dataset for building HVAC systems operating in faulted and fault-free states. *Scientific data.* Vol. 10. No. 1. pp. 342.

Honeywell (2021) Evolution of Predictive Maintenance. https://www.honeywellforge.ai/us/en/whitepaper/the-evolution-of-predictive-maintenance.

Hosamo, H.H., Svennevig, P.R., Svidt, K., Han, D. & Nielsen, H.K. (2022) A Digital Twin predictive maintenance framework of air handling units based on automatic fault detection and diagnostics. *Energy and Buildings.* Vol. 261. pp. 111988. https://doi.org/https://doi.org/10.1016/j.enbuild.2022.111988.

Hossain, M.A. & Nadeem, A. (2019) 'Towards digitizing the construction industry: State of the art of construction 4.0'. *Proceedings of the ISEC,* 2019. pp.1-6.

Marzouk, M. & Zaher, M. (2020) Artificial intelligence exploitation in facility management using deep learning. *Construction Innovation.* Vol. 20. No. 4. pp. 609-624.

Masdoua, Y., Boukhnifer, M. & Adjallah, K.H. (2022) 'Fault detection and diagnosis in AHU system with data driven approaches'. *2022 8th International Conference on Control, Decision and Information Technologies (CoDIT),* 2022. IEEE, pp.1375-1380.

Rockwell Automation *Predictive maintenance (PdM).* Available at: https://fiixsoftware.com/maintenance-strategies/predictive-maintenance (Accessed: April 15).

Xiao, C., Ye, J., Esteves, R.M. & Rong, C. (2016) Using Spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation: Practice and Experience.* Vol. 28. No. 14. pp. 3866-3878.

Zonta, T., Da Costa, C.A., da Rosa Righi, R., de Lima, M.J., da Trindade, E.S. & Li, G.P. (2020) Predictive maintenance in the Industry 4.0: A systematic literature review. *Computers & Industrial Engineering.* Vol. 150. pp. 106889.