International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
IJISAE
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Hybrid Filter Feature Selection Approach for Remaining Useful Life Prediction of Industrial Machinery

Ku Amirul Asyraf Ku Amir*[1], Shakirah Mohd Taib[2], Mohd Hilmi Hasan[3]

Submitted: 10/09/2022 **Accepted:** 20/12/2022

Abstract: Data-driven predictive maintenance commonly uses machine learning algorithms to conduct prognostics of an asset's condition over its life cycle. Asset information and domain expert knowledge are essential in data-driven predictive maintenance to support maintenance-related decisions. Using a general feature selection approach in data-driven prognostics can cause misinterpretation, removal, or loss of domain-specific information of assets. The high dimensionality characteristics of asset data due to a large number of features sourced from various sensor measurements can affect the performance and reliability of machine learning algorithms. This paper presents a feature selection approach to overcome the challenges of retaining domain-specific asset data information by utilising the Safe Operating Limit of an asset. The asset information is combined with the filter method to reduce the high dimensional aspects of asset data for application in equipment's remaining useful life prediction. The proposed feature selection approach is demonstrated on an oil and gas equipment dataset that contains multiple run-to-failure situations of a gas compressor.

Keywords: *feature selection, filter method, high dimensional data, prognostics, remaining useful life*

1. Introduction

Unplanned downtime is any unforeseen event that could be an unexpected shutdown or failure of equipment or process. Unplanned downtime can incur economic losses from $22,000 to $50,000 per minute [1] in the automotive industry. The same nature of the event would cost around $5 million per day in the mining industry [2] and from $38 to $88 million annually in the oil and gas industry [3].

In the oil and gas industry, according to [3], data-driven predictive maintenance can reduce downtime by 36% compared to reactive maintenance strategy. The results from the reduced downtime are, in economic terms, $34 million in cost savings annually and a 5-10% decrease in overall maintenance cost. Productivity values result in a 10-20% increase in uptime and availability.

There are several approaches to maintenance being practised in the industrial environment. The earliest being used are reactive maintenance, where repairs are done after breakdown occurs. Preventive maintenance is where replacement and inspection are done based on intervals and schedules. Condition-based maintenance is a practice where critical equipment is monitored closely for any changes and to perform just-in-time repairs. Predictive maintenance is where the future state of equipment is estimated using mathematical models or artificial intelligence techniques.

In predictive maintenance, failure prognostics consist of activities to anticipate failure time by predicting the future health state and Remaining Useful Life (RUL) of equipment [4]. RUL is the approximation of the time to failure based on the possibility of future and existing failure modes to occur. [5]

There are two fundamental types of prognostics: 1) Data-driven prognostics, where data from sensors are transformed into learning models [4]. Pattern recognition and machine learning techniques are used to discover changes in system behaviour. Historical condition monitoring data from sensors are used to predict RUL using interpolation, extrapolation, and machine learning techniques. 2) Model-based prognostics, where system-specific knowledge from experts and physical characteristics of the system are used to produce physical failure models for degradation, such as crack, wear, and corrosion. Statistical models are applied in model-based prognostics such as Particle filter, Kalman filter, and hidden Markov model [6]

From this perspective, this study aims to explore the approaches in the selection of features to predict the remaining useful life of industrial equipment. The key motivation for the selection of this prediction scope for the area of industrial equipment is associated with the fact that adequate information on possible failure allows the respective industrial personnel to assess and implement necessary actions or prevention steps within an ample period before an unplanned equipment failure happens. This study conducts a case study in which historical data from an oil and gas industrial equipment are used for feature selection and prediction of their remaining useful life.

The main contribution of this study is in the application,

[1] *Computer and Information Sciences Department, Universiti Teknologi PETRONAS, Perak – 32610, Malaysia*
ORCID ID : 0000-0002-5786-0007
[2] *Computer and Information Sciences Department, Universiti Teknologi PETRONAS, Perak – 32610, Malaysia*
ORCID ID : 0000-0001-6598-3536
[3] *Computer and Information Sciences Department, Universiti Teknologi PETRONAS, Perak – 32610, Malaysia*
ORCID ID : 0000-0002-4065-3968
** Corresponding Author Email: ku_20000142@utp.edu.my*

evaluation and addition of discussions not previously established concerning the application of domain knowledge for data-driven prognostics to predict the remaining useful life of industrial equipment with the use of historical data. While the papers identified in Section 2 focus on the study of feature selection methods, this study proposes the investigation of the performance of remaining useful life prediction using the feature selection method with the combination of domain knowledge from domain experts.

The organisation of the rest of the paper is arranged as follows. In Section 2, a brief review of feature selection methods that are emphasised in relation to the problem statement. Next, in Section 3 we explain the process flowchart of the study with the details of the proposed feature selection method and the details of the data analysis, whereas Section 4 discusses the experimental results and analysis. Section 5 is the conclusion of the paper.

2. Literature Survey

The earliest definition of big data consists of 3Vs: Volume, Velocity and Variety. [23] Volume refers to the quantity of data that is produced. Velocity relates to the speed at which data is being produced and processed. Variety is the number of data types. According to [7], two additional dimensions are introduced: Veracity and Value. Veracity refers to the uncertain nature of data caused by many factors such as incompleteness, irregularities, and abnormalities. Value refers to the need to enhance raw and unprocessed data by obtaining higher-level information for use in various scenarios. Data can be classified into three types: structured data such as sensor signals and controller data, semi-structured data such as information in Extensible Business Reporting Language (XBRL) format [23], and unstructured data such as sound, images, and videos.

In industrial application, big data analysis requires specific knowledge of the domain [8] as industrial data can consist of multiple categories such as vibration, temperature, pressure, electric signal, rotating speed, and acoustic values [9]. The spatiotemporal property of industrial data needs to be put into consideration. A spatial attribute integrates spatially independent subsystems in a complex process [10], while a temporal attribute refers to different sensors having diverse sampling frequencies. In many big data applications, industrial data also suffers from the 'curse of dimensionality' problems [11], as illustrated in Figure 1, where the difficulty of getting an optimal number of features increases as the higher the number of features, the more data samples will be required.

Feature selection is one of the common practices to handle the curse of dimensionality problem, by reducing the number of features [30]. Feature selection is a pre-processing procedure used to detect significant attributes in a dataset, as a measure of dimensionality reduction [22]. Table 1 summarises the analysis of

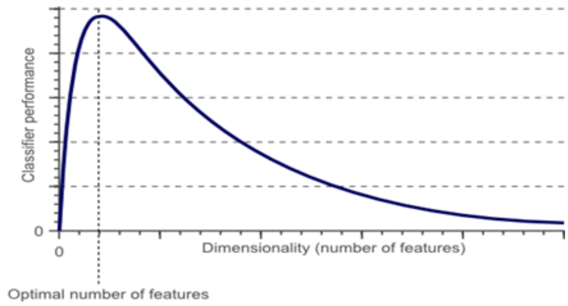

Fig. 1. Curse of dimensionality in feature selection [29]

four feature selection methods that are used to classify features in data which are the filter method, wrapper method, embedded method, and unsupervised learning method.

The filter method selects variables based on ranking methods [30] where a scoring function is used to measure the usefulness of a feature by using specified criteria or measure of statistical dependence to evaluate the relationship between features. The wrapper method uses machine learning algorithms to select features that can contribute to the measure of the algorithm's accuracy. Wrapper methods can perform better in selecting features as the feature space for training and testing can consider feature dependencies at the cost of higher computational complexity [31]. In the embedded method, feature selection is performed simultaneously during the modelling algorithm execution using an objective function such as goodness-of-fit term and penalty for a higher number of features [32]. The integration of modelling and feature selection in the embedded method results in a better computational complexity than the wrapper method. In unsupervised learning feature selection, algorithms such as neural networks and deep learning are used to

Table 1. Analysis on feature selection method

Feature Selection Method	***Filter***	***Wrapper***	***Embedded***	***Unsupervised Learning***
Use in high dimensional data	Yes [15][18]	Yes [16]	Yes [17]	Yes [15]
Incorporates domain knowledge	No [24]			
Computation cost	Low	High	Moderate	High
Advantages	1. Independent of classifier 2. Better computational complexity than wrapper methods 3. Good scalability	1. Models feature dependencies 2. Interacts with classifier	1. Better computational complexity than wrapper methods	1. Learning more complex structures from data due to the deep architecture. 2. Do not require feature extractor
Disadvantages	1. Ignores feature dependencies. 2. Ignores interaction with classifier [12]	1.Risk of overfitting [14] 2. Computationally intensive	1. Classifier dependent selection	1. Require large samples 2. No physical meaning 3. Long training time 4. Risk of overfitting

learn and select features where many layers of information processing stages are implemented to analyse and compute hierarchical features from the observed data [33].

There were several works of literature in studies on feature selection method to be used in a high dimensional dataset that has been conducted. These papers were studied, and the results are summed up in Table 2. Based on the findings, the four approaches can be used for high dimensional data according to [15][16][17][18]. The filter approach has been chosen as the feature selection method for the case study. It has good scalability for a high dimensional dataset [12] and a low computational cost compared to other approaches due to its independence from the classification algorithm. However, the drawback of the method is filter method does not interact with the classifier, and feature dependencies are ignored as features are considered separately. Although the wrapper method and embedded method can identify feature dependencies by interacting with the classifier, there is a risk of overfitting and high computational cost. The risk is due to the exponential growth of the feature subset as the number of features increases, and both methods do not have high generality [28]. Hence, the filter method is more suitable for our case as we will handle a large dataset.

There is room for improvement to the currently selected method as there are a few disadvantages to the filter method in selecting features. The filter method requires some enhancement to overcome the drawback of lacking feature dependencies.

In real-world data analytics applications, expert knowledge and data-driven approaches should be synergised to better understand data [19] in order to eliminate data inconsistencies and irrelevant data. In the medical domain, [25] proposed an expert-driven feature selection where domain experts prioritise and rank the features that could predict the Large Gestational Age of infants. A medical knowledge motivated feature selection (MFS) was proposed in [27] to evaluate the effectiveness of domain knowledge in feature selection for heart disease diagnostics. An expert-driven feature selection was proposed in [26] to predict accident-related causes of death from plaintext autopsy reports. These studies have shown acceptable results in the output of prediction by using expert-driven feature selection. However, the number of features used in the studies is considered relatively small compared with high dimensional data.

From an industrial perspective, when handling high-dimensional industrial data, the interpretation of feature attributes varies between general data and equipment or asset data. For instance, in asset data, a feature of low variance can indicate a nominal state, while a feature with high variance in values could refer to an inaccurate sensor. [20] A feature's importance should be based on its capability to explain and relate to domain-specific information in asset data.

The application of standard feature selection can cause the dismissal of features with significant value of domain-specific information. The ability to quantify and retain domain knowledge to minimise loss of information during feature engineering is crucial to ensure the selected feature continues to be useful to data-driven prognostics. Also, improper use of asset data in developing machine learning models for prognostics can lead to poor model performance and high levels of imprecise predictions. The incorrect assessments could result in heightened loss of value during an asset's life cycle, operational inefficiencies, and unacceptable safety situations [21]. Therefore, there is a need to improve the selected method with domain knowledge and expert knowledge for the feature selection process.

Section 2 reviews the recent works related to feature selection methods used for high-dimensional data. In data-driven predictive maintenance, machine learning algorithms' performance and reliability depend on the quality of data used in model training and testing [37]. Factors that can cause poor algorithm performance include a large number of variables, commonly in a high dimensional dataset, features with low information and variance changes, and redundant features [38]. Difficulties in choosing an ideal sample from high dimensional data and disparities in understanding domain-specific knowledge also become a challenge to apply machine learning methods to asset data for predictive maintenance [39]

To ensure a conducive data representation for predictive maintenance analytics, domain knowledge that includes the details during asset failures, such as fault modes and their implications [40], needs to be identified and preserved during feature engineering. Considering the effects of failures and their occurrences on an asset life cycle can help produce good predictive maintenance analytics [41]. To avoid the 'curse of dimensionality problem, feature selection methods are typically utilised to identify the best features that contain the most distinctive information [11].

Traditional feature selection methods are suitable for general data. Still, they may not be adequate to capture domain-specific information in asset data due to the lack of information on an asset's process design and characteristics [42]. There is no generic approach to incorporating domain knowledge in high-dimensional equipment datasets feature selection. General feature selection tends to disregard prior knowledge of domain-specific information about which features are more relevant, which removes crucial features that contain such information. Features selected by standard feature selection methods may be useful for general machine learning applications but may not ensure the relevancy of

features when applied to data-driven prognostics. This paper proposes a hybrid feature selection method to avoid such limitations in the general feature selection method. The ideal feature subset is identified while retaining domain-specific asset data information that can represent sensitivity to changes for application in predictive maintenance.

3. Proposed Method

The overall process flow of the methodology in this study is broadly summarised in Fig. 2. The proposed feature selection method in this work is applied to time series data, where the values of features are measurements and readings of various sensors and parameters of an industrial machine.

3.1. Methodology

Data acquisition is the process of sampling signals acquired from physical sensors, converting analogue signals into digital numeric values that can be stored and analysed in a computer. Next, the data preparation stage will include data cleaning where irrelevant, incomplete, or incorrect values will be identified, replaced, modified, or removed to prepare data for analysis. Then, the feature selection process comprises a selection of relevant features that is useful and can contribute to the machine learning prediction model. This phase will propose an improved feature selection method that combines domain knowledge or expert-

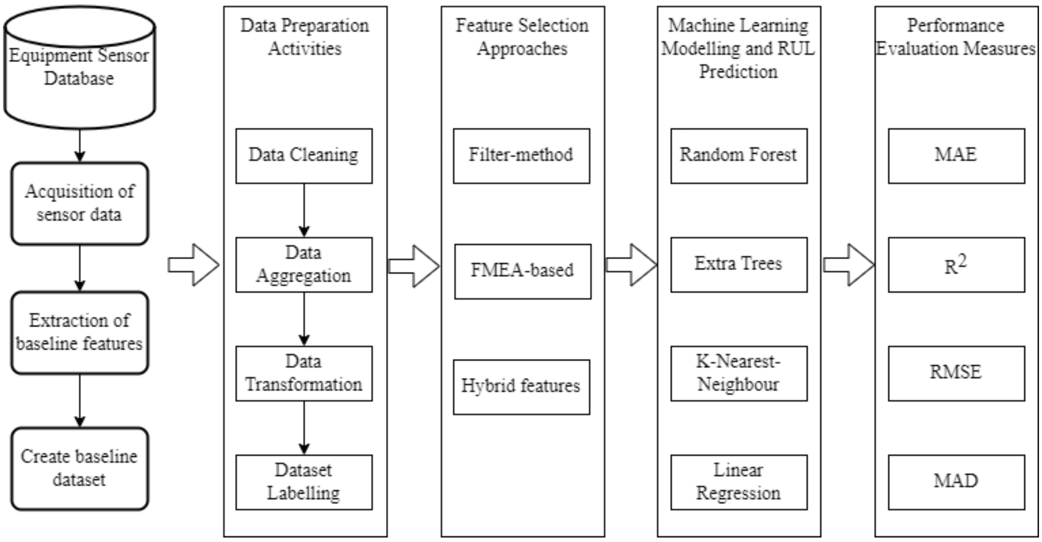

Fig. 2. Overall process flow

driven features with the existing feature selection method. The last step consists of generating a training dataset by identifying the target column for model training and formatting the dataset into an acceptable format for the machine learning module. The shutdown and maintenance log of the equipment is obtained and analysed to determine and classify between the planned shutdown and unplanned shutdown or failures that occurred. The dates and descriptions of the events are studied and utilised as the basis to generate the target column for the dataset for the ML model training. Based on discussion with equipment maintenance personnel, the duration for TTE prediction to be beneficial for intervention by personnel for the equipment is 120 hours before a failure. The TTE labelling method is performed based on the approach in [43] where the time to the next event is used as the label for each point of time in the dataset. The outcome of the TTE labelling process is visualised in Fig 3.

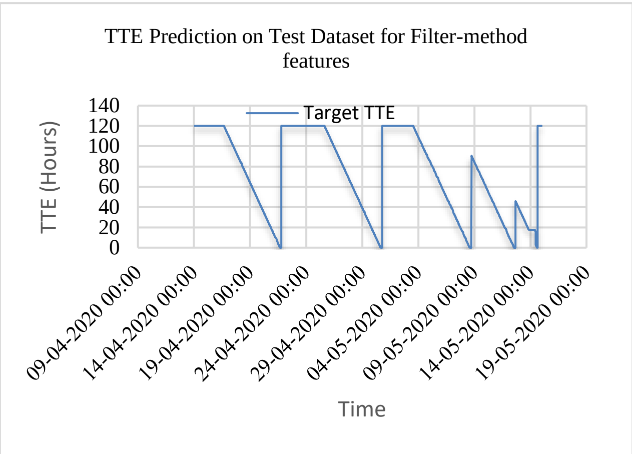

Fig. 3. TTE labelling process

In the model training stage, the machine learning model is used to predict the remaining useful life. Features from the feature selection process and training data generated from the data pre-processing phase will be used in this stage. Next, the validation stage is done, where the trained model is evaluated using test data. This phase is a critical step as it acts to test the reliability and generalisation capability of the model. This phase will focus on the machine learning model's performance in the prediction of test data. Mainly, the performance will be measured by the error measurement such as Mean Absolute Error (MAE), Coefficient of determination (R^2), Root Mean Squared Error (RMSE) and Median Absolute Deviation (MAD).

MAE is used to compute the projected value of absolute error loss, using the calculation method shown in (1). R^2 quantifies the amount of variance on dependent variables from the independent variables [45]. The formula is defined in equation (2). RMSE is a function to compute the risk metric related to the values of quadratic error or loss that is expected [44]. The general calculation method is as shown in (3). The general formula of MAD is shown in (4). MAD is the calculation of regression loss based on the absolute amount of dispersion around the median of the data.

$$MAE = \frac{\sum_{n=1}^{n}|y_i - x_i|}{n} \quad (1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (2)$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}} \quad (3)$$

$$MAD = median(|X_i - \tilde{X}|) \quad (4)$$

3.2. Feature Selection Process

Two main challenges when dealing with industrial high dimensional datasets are the "curse of dimensionality" and the missing of important features. One of the most direct methods to overcome the two challenges is to reduce the search space or features while retaining domain knowledge. Hence, this study proposes a hybrid feature selection algorithm that considers domain knowledge during the process, as illustrated in Fig. 3.

Given a collection of features, the filter method is used to find the mutual relationship or correlation between the features to remove duplicate and correlated features, which results in the deletion of irrelevant or weakly related features. Constant features are removed, with features that contain similar values for all the outputs in the dataset, which provide no information that can help in the machine learning process. Quasi-continuous features are removed, with features that give a continuous reading, but the values are almost constant; an example in asset data is where a sensor marks a certain condition by giving 0 and 1 values only.

The steps taken to for filter-method are explained in Algorithm 1.

Algorithm 1: **Filter-method feature selection**
Input: Dataset containing N features
Output: Dataset with j features excluding irrelevant features based on filter-method

1. Assign variance threshold object, vs_constant = VarianceThreshold(threshold=0)
2. Assign train dataset, x_train = df
3. Select numerical columns from x_train, assign to numerical_x_train
4. fit vs_constant to numerical_x_train
5. Get names of columns with constant values, assign to constant_columns
    a. For columns in numerical_x_train where
6. Select constant categorical columns from x_train, assign as constant_cat_columns
    a. Check if x_train[column].dtype == "0" and len(x_train[column]).unique == 1
7. Concatenate column list: constant_columns and constant_cat_columns. Assign as all_constant_columns
8. Drop the constant columns all_constant_columns from x_train
9. Set threshold = 0.95
10. Assign list for quasi_constant_column
11. For feature in x_train.columns:
    a. Calculate ratio of each row by x_train[feature].value_counts / x_train.sort_values.values
    b. Assign ratio to predominant
    c. If predominant >= threshold, append quasi_constant_column with feature
12. Create a set to hold correlated features, corr_features
13. Create correlation matrix using Pearson correlation, corr_matrix = x_train.corr
14. For each row in corr_matrix:
    a. If value > 0.9, add column name to corr_features
15. Exit with features remaining
16. End

Expert knowledge-driven features are acquired mainly from the experts' inputs on the physical boundaries of an asset derived from the combination of equipment design limits, process flow, and aspects of the operational processes [34]. In this study, Safe Operating Limit (SOL) [35] is used as a guideline to identify sensors with physical boundaries or limits in the range of parameters in which operations will result in safe and acceptable equipment performance. SOL can be defined as a range of values for a critical operating parameter such as pressure, temperature, pH, speed, and flow that defines a safe operating envelope for equipment or process unit where if the values exceed the envelope threshold, predetermined actions are to be taken in order to prevent the possibility of impending catastrophic events and failure or loss of containment. [36] The sensors in which SOL is considered critical by the experts during operational activities are taken as important features of the dataset.

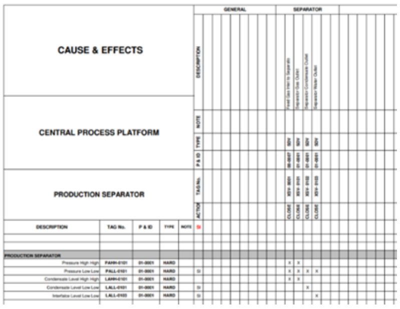

Fig. 4. Cause and Effects matrix

Finally, a subset of features containing a dimensionally reduced form of the original data while retaining adequate information on the asset is generated. The selected feature subset is expected to have limited correlation after removing highly correlated and duplicate features during the filter method process. Limiting feature correlation and redundancy reduces the risk of overfitting during machine learning model application. The selected feature subset should have unbiased properties as the features selected with experts' guidance are obtained through operation and maintenance strategies using risk assessment techniques like Failure Modes, and Effects Analysis (FMEA). The features gathered from FMEA should cover sensors that can show significant changes in an abnormal event of an asset. Based on the FMEA, domain experts will develop a cause-and-effects matrix that maps failure events to their related sensors from these sensor-based indicators. The cause-and-effects matrix as shown in Fig. 4 becomes a reference to identify the sensors that are most likely relevant to identified failures from the FMEA analysis. The analysis of the matrix allows us to understand the relationship between failure causes and their effects. By understanding the logical implementation in system safety, relevant sensors can be identified for the failure events. The sensors will be considered as features and combined with the features obtained using the filter method. The feature selection evaluation process will be done to select relevant features among the result of the previous steps. Features that do not show a trend or deviation during a confirmed failure event can be inferred that they do not show detectable changes or contain minimal discriminatory information about an asset during changes of state in a machine failure event. The chosen features are expected to be adequately sensitive and represent the condition changes experienced by an asset, which will be useful for machine learning model training and analytics. Based on the following properties, a conducive feature subset selection is prepared to fulfil the requirements of machine learning and predictive maintenance.

The machine learning model training will be using an ensemble of Random Forest, Extra Trees, KNearestNeighbour, and Linear Regression algorithms. These algorithms are commonly used for regression models and numeric values prediction [43]. The dataset obtained from the feature selection process is used in the model training, and the performance in predicting the time-to-failure for each model and feature subset is compared. K-fold cross-validation is utilised on all the experiments to ensure every case from the original dataset is efficiently incorporated into the training and test splits.

4. Experiment and Findings

For our case study, a sample dataset from one oil and gas equipment is analysed for this research. The dataset consists of data from 17000 sensors. In the equipment instrumentation system, the records in the dataset are updated at a rate of 3500 rows per minute. The sensor values are sourced from multiple components with varying categories of measurements. The data contains missing values and inconsistencies, which require further understanding. There is domain-specific information that needs to be considered during feature selection, such as failure modes, operational limits and the relationship of components that is identified by the operational personnel. Table 2 summarises the details of dataset.

Table 2. Dataset details and number of features

Dataset	Description	Number of features
Baseline	Target equipment and related systems	7543
Filter-based features	Constant filter, Quasi-constant filter, Pearson correlation, low variance filter, and high correlation filter	778
FMEA-based features	Analysis from FMEA, Cause & Effect matrix	857
Hybrid	Combine filter-based and FMEA-based features, remove duplicate features	1509

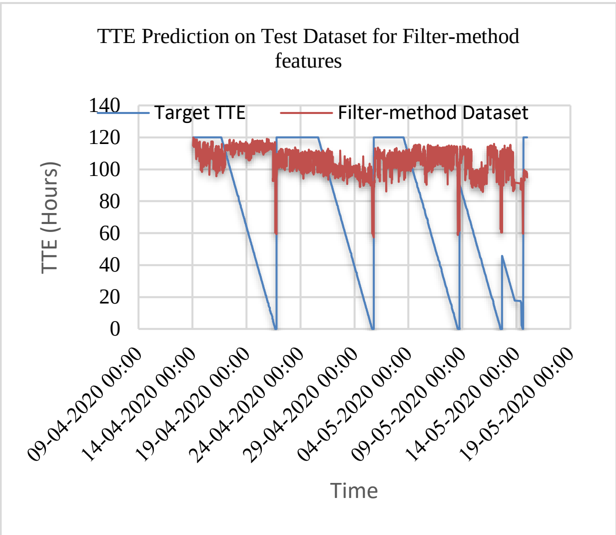

Fig. 5. TTE Prediction on Filter-method dataset

The baseline features in this case study of gas compressor equipment are determined by grouping sensors based on asset structure documentation, and identification of sensors belonging to the equipment and its nearest related systems. The baseline features consist of 7543 features. The filter-based features are obtained after the removal of sensors with constant, quasi-constant, low variation and highly correlated sensors from the baseline using the filter method. The number of features retained after the implementation of the filter method is 778 features.

Based on domain expert knowledge from operational engineers, the sensors that function as first-up alarms and flags are removed as they are not useful in giving information prior to abnormal equipment events. The analysis of the cause-and-effects matrix that is developed during the FMEA process is used to determine relevant sensors to past failure events and system logical design. This step includes the SOL as a guideline where sensors containing operating envelopes of Higher High, High, Lower Low, and Low are grouped as critical sensors that can impact the equipment performance. The number of features retained after the FMEA-based selection is 857 features.

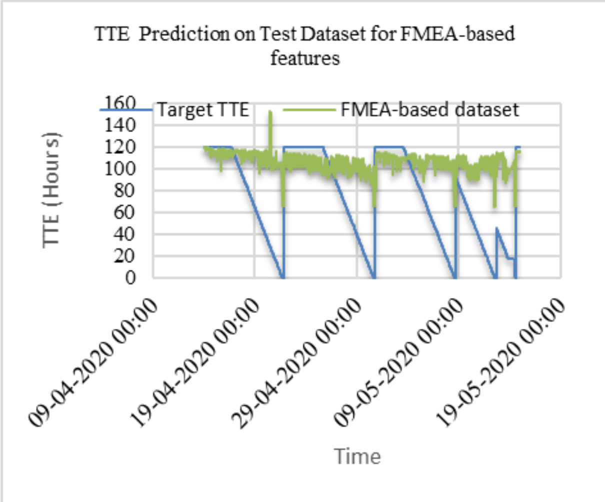

Fig. 6. TTE Prediction on FMEA-based dataset

The hybrid set of features is a combination of the FMEA-based features and the filter-method features. Ideally, this selection of features should incorporate the domain-knowledge information while keeping only statistically relevant features. After the removal of any duplicate features from the combined selection, the number of features in the hybrid selection is 1509 features. This group of sensors will be the features used for the model training process.

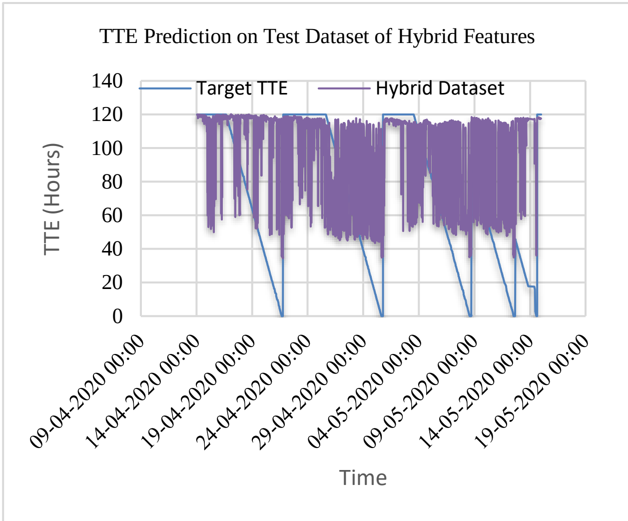

Fig. 7. TTE Prediction on hybrid dataset

The machine learning models' predictions are plotted as line charts to visualise the difference in prediction against target values. Fig. 5 and Fig 6 show the TTE prediction values for ML models trained using the filter-method dataset and FMEA-based dataset respectively. Based on the observed TTE prediction values generated from the ML model trained using the hybrid dataset as shown in Fig. 7, the hybrid dataset model displayed more sensitivity in the predictions based on the significant range of fluctuations in predictions as the time towards actual event decreases.

Therefore, in this study, predictions of the models using Filter-based features and FMEA-based features are generally unable to display a significant change in prediction values for upcoming events compared to the prediction using Hybrid Features. The

Hybrid Features predictions also drop slightly closer to the target TTE than the other predictions.

Table 3. Performance metrics based on testing dataset prediction

Metrics/ Features	MAE	R2	RMSE	MAD
Filter-based	2906.8094	0.0918	3885.9657	1934.9761
FMEA-based	2717.5796	0.1578	3873.0734	1932.9175
Hybrid	2438.8284	0.2836	3147.9908	1919.9727

The experimental results of this study show that machine learning models trained with hybrid features incorporating domain-specific information and filter-based selection can predict the time-to-failure of equipment with a better performance. Table 3 shows the calculation of performance measures of up to 16% lower MAE compared to the filter method and up to 10% lower MAE compared to domain-specific features only. In terms of correlation, the hybrid features improved the correlation score (R2) to 67% compared to the filter method and up to 44% against FMEA-based features. From the performance metrics evaluation of the predictions, a significant reduction in error rate can be deducted for the predictions using the hybrid dataset.

5. Conclusion

The challenges concerning feature selection of big data in the industrial dataset have been studied. The comparison study of existing feature selection methods and their implementation in a high-dimensional dataset is documented in this study.

Firstly, the objective of this research was to investigate the feasibility of the feature selection method for high-dimensional data and the research was summarised in the literature review. The second objective was to develop an improved feature selection method with domain knowledge for remaining useful life prediction of an industrial equipment dataset to improve prediction accuracy. The final objective was to evaluate and validate the proposed feature selection method in terms of accuracy in predicting the remaining useful life.

In this paper, a hybrid feature selection method for remaining useful life prediction based on filter method and domain expert knowledge has been proposed. The filter method has been used to optimise the high dimensional data and a large number of parameters in industrial asset data. Domain expert knowledge is utilised to ensure critical and valuable features based on system logic are not missing from the dataset used in machine learning applications. As asset data does not always follow general data characteristics, a feature selection method that retains critical domain expert knowledge is required. This work is to show that feature selection that is done with consideration of domain-specific knowledge and information can produce a selection of features useful for application in predictive maintenance and able to obtain higher accuracy of machine learning model prediction.

Acknowledgements

This research was endorsed by the Centre for Research in Data Science (CeRDaS), Universiti Teknologi PETRONAS, Malaysia. The work presented in this paper is part of the ongoing research funded by the Yayasan Universiti Teknologi PETRONAS fund (015LC0-311).

Author contributions

Ku Amirul Asyraf Ku Amir: Conceptualization, Methodology, Software, Field study, Visualization, Writing, Editing, Investigation. **Shakirah Mohd Taib:** Validation, Reviewing. **Mohd Hilmi Hasan:** Writing-Reviewing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Advanced Technology Services, I. (2006, March 29). Downtime Costs Auto Industry $22k/Minute - Survey. Retrieved December 19, 2020, from https://news.thomasnet.com/companystory/downtime-costs-auto-industry-22k-minute-survey-481017

[2] The modern mine: How digitisation is transforming industry from pit to port. (2020, July 9). Schneider Electric. https://blog.se.com/mining-metals-minerals/2018/11/06/the-modern-mine-how-digitization-is-transforming-industry-from-pit-to-port/

[3] Boyapati, B. ., and J. . Kumar. "Parasitic Element Based Frequency Reconfigurable Antenna With Dual Wideband Characteristics for Wireless Applications". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 6, June 2022, pp. 10-23, doi:10.17762/ijritcc.v10i6.5619.

[4] Baker Hughes, "The Impact of Digital on Unplanned Downtime: An Offshore Oil and Gas Perspective", 2016

[5] Tobon-Mejia, D. A., Medjaher, K., Zerhouni, N., & Tripot, G. (2012). A data-driven failure prognostics method based on mixture of Gaussians hidden Markov models. IEEE Transactions on reliability, 61(2), 491-503.

[6] ISO. (2004). Condition Monitoring and Diagnostics of Machines Prognostics Part1: General Guidelines.

[7] Kozlowski, J. D., Watson, M. J., Byington, C. S., Garga, A. K., & Hay, T. A. (2001, July). Electrochemical cell diagnostics using online impedance measurement, state estimation and data fusion techniques. In Intersociety energy conversion engineering conference (Vol. 2, pp. 981-986). SAE; 1999.

[8] Agarwal, D. A. . (2022). Advancing Privacy and Security of Internet of Things to Find Integrated Solutions. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(2), 05–08. https://doi.org/10.17762/ijfrcsce.v8i2.2067

[9] Debattista, J., Lange, C., Scerri, S., & Auer, S. (2015, December). Linked 'Big' Data: towards a manifold increase in big data value and veracity. In 2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC) (pp. 92-98). IEEE.

[10] Fernandes, M., Canito, A., Bolón-Canedo, V., Conceição, L., Praça, I., & Marreiros, G. (2019). Data analysis and feature selection for predictive maintenance: A case-study in the metallurgic industry. International journal of information management, 46, 252-262.

[11] Zhang, W., Yang, D., & Wang, H. (2019). Data-driven methods for predictive maintenance of industrial equipment: A survey. IEEE Systems Journal, 13(3), 2213-2227.

[12] Yan, J., Meng, Y., Lu, L., & Li, L. (2017). Industrial big data in an industry 4.0 environment: Challenges, schemes, and applications for predictive maintenance. IEEE Access, 5, 23484-23491.

[13] Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70-79.

[14] Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. bioinformatics, 23(19),

2507-2517.
[15] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.
[16] Georges, N., Mhiri, I., Rekik, I., & Alzheimer's Disease Neuroimaging Initiative. (2020). Identifying the best data-driven feature selection method for boosting reproducibility in classification tasks. Pattern Recognition, 101, 107183.
[17] Khumprom, P., Yodo, N., & Grewell, D. (2020, January). Neural Networks Based Feature Selection Approaches for Prognostics of Aircraft Engines. In 2020 Annual Reliability and Maintainability Symposium (RAMS) (pp. 1-7). IEEE.
[18] Chen, G., & Chen, J. (2015). A novel wrapper method for feature selection and its applications. Neurocomputing, 159, 219-226.
[19] Zhao, J., Chen, L., Pedrycz, W., & Wang, W. (2018). Variational inference-based automatic relevance determination kernel for embedded feature selection of noisy industrial data. IEEE Transactions on Industrial Electronics, 66(1), 416-428.
[20] Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2015). Feature selection for high-dimensional data. Cham: Springer International Publishing.
[21] Komas, T., Daub, R., Karamat, M. Z., Thiede, S., & Herrmann, C. (2019, August). Data-and expert-driven analysis of cause-effect relationships in the production of lithium-ion batteries. In 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE) (pp. 380-385). IEEE.
[22] Aremu, O. O., Palau, A. S., Parlikad, A. K., Hyland-Wood, D., & McAree, P. R. (2018). Structuring data for intelligent predictive maintenance in asset management. IFAC-PapersOnLine, 51(11), 514-519
[23] Aremu, O. O., Cody, R. A., Hyland-Wood, D., & McAree, P. R. (2020). A relative entropy-based feature selection framework for asset data in predictive maintenance. Computers & Industrial Engineering, 106536.
[24] Mlambo, N., Cheruiyot, W. K., & Kimwele, M. W. (2016). A survey and comparative study of filter and wrapper feature selection techniques. International Journal of Engineering and Science (IJES), 5(8), 57-67.
[25] Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. Business Horizons, 60(3), 293-303.
[26] Liu, Y., Wu, J. M., Avdeev, M., & Shi, S. Q. (2020). Multi-Layer Feature Selection Incorporating Weighted Score-Based Expert Knowledge toward Modeling Materials with Targeted Properties. Advanced Theory and Simulations, 3(2), 1900215.
[27] Akhtar, F., Li, J., Pei, Y., Imran, A., Rajput, A., Azeem, M., & Liu, B. (2020). Diagnosis of large-for-gestational-age infants using a semi-supervised feature learned from expert and data. Multimedia Tools and Applications, 1-31.
[28] Paithane, P. M., & Kakarwal, D. (2022). Automatic Pancreas Segmentation using A Novel Modified Semantic Deep Learning Bottom-Up Approach. International Journal of Intelligent Systems and Applications in Engineering, 10(1), 98–104. https://doi.org/10.18201/ijisae.2022.272
[29] Mujtaba, G., Shuib, L., Raj, R. G., Rajandram, R., Shaikh, K., & Al-Garadi, M. A. (2017). Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. PloS one, 12(2), e0170242.
[30] Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Systems with Applications, 40(1), 96-104.
[31] Gnana, D. A. A., Balamurugan, S. A. A., & Leavline, E. J. (2016). Literature review on feature selection methods for high-dimensional data. International Journal of Computer Applications, 975, 8887.
[32] Open Data Science, ODSC. "Confronting the Curse of Dimensionality." Medium, Medium, 4 Apr. 2019, medium.com/@ODSC/confronting-the-curse-of-dimensionality-5bcf2998b30d.
[33] Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70-79.
[34] Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. Advances in bioinformatics, 2015.
[35] Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. Journal of biomedical informatics, 85, 189-203.
[36] Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E., & Greenspan, H. (2018). Chest pathology identification using deep feature selection with non-medical training. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualisation, 6(3), 259-263.
[37] Stauffer, T., & Chastain-Knight, D. (2019, March). Don't Let Your Safe Operating Limits Leave You SOL (Out of Luck). In 2019 Spring Meeting and 15th Global Congress on Process Safety. AIChE.
[38] American Petroleum Institute. (2016). Recommended Practice for Completion/Workover Risers (2nd ed.). American Petroleum Institute.
[39] Center for Chemical Process Safety (CCPS). (2010). Guidelines for Risk Based Process Safety. John Wiley & Sons.
[40] Ellefsen, A. L., Bjørlykhaug, E., Æsøy, V., Ushakov, S., & Zhang, H. (2019). Remaining useful life predictions for turbofan engine degradation using semi-supervised deep architecture. Reliability Engineering & System Safety, 183, 240-251.
[41] Stuart, R., & Peter, N. (2016). Artificial intelligence-a modern approach 3rd ed.
[42] Lei, Y., Li, N., Guo, L., Li, N., Yan, T., & Lin, J. (2018). Machinery health prognostics: A systematic review from data acquisition to RUL prediction. Mechanical systems and signal processing, 104, 799-834.
[43] Downey, A., Sadoughi, M., Laflamme, S., & Hu, C. (2018). Incipient damage detection for large area structures monitored with a network of soft elastomeric capacitors using relative entropy. IEEE Sensors Journal, 18(21), 8827-8834.
[44] Maleki, E., Belkadi, F., Boli, N., van der Zwaag, B. J., Alexopoulos, K., Koukas, S., ... & Mourtzis, D. (2018). Ontology-based framework enabling smart product-service systems: application of sensing systems for machine health monitoring. IEEE internet of things journal, 5(6), 4496-4505.
[45] Zhang, B., Zhang, L., & Xu, J. (2016). Degradation feature selection for remaining useful life prediction of rolling element bearings. Quality and Reliability Engineering International, 32(2), 547-554.
[46] Mashhadi, P. S., Nowaczyk, S., & Pashami, S. (2019). Stacked Ensemble of Recurrent Neural Networks for Predicting Turbocharger Remaining Useful Life. Applied Sciences, 10(1), 69. MDPI AG. Retrieved from http://dx.doi.org/10.3390/app10010069
[47] Pedregosa, F., et al., (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.
[48] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ. Computer science, 7, e623. https://doi.org/10.7717/peerj-cs.623
[49] Ahmed Cherif Megri, Sameer Hamoush, Ismail Zayd Megri, Yao Yu. (2021). Advanced Manufacturing Online STEM Education Pipeline for Early-College and High School Students. Journal of Online Engineering Education, 12(2), 01–06. Retrieved from http://onlineengineeringeducation.com/index.php/joee/article/view/47