Journal of Computer Science

Original Research Paper

Data Preparation in Machine Learning for Condition-based Maintenance

[1]Ons Masmoudi, [2]Mehdi Jaoua, [3]Amel Jaoua and [4]Soumaya Yacout

[1,2,4]Department of Mathematics and Industrial Engineering, École Polytechnique de Montréal, Montreal, Canada
[3]LR-OASIS, National Engineering School of Tunis, University of Tunis El Manar, Tunis, Tunisia

Article history
Received: 15-12-2020
Revised: 13-03-2021
Accepted: 16-03-2021

Corresponding Author:
Amel Jaoua
LR-OASIS, National Engineering School of Tunis, University of Tunis El Manar, Tunis, Tunisia
Email: amel.jaoua@polymtl.ca

Abstract: Using Machine Learning (ML) prediction to achieve a successful, cost-effective, Condition-Based Maintenance (CBM) strategy has become very attractive in the context of Industry 4.0. In other fields, it is well known that in order to benefit from the prediction capability of ML algorithms, the data preparation phase must be well conducted. Thus, the objective of this paper is to investigate the effect of data preparation on the ML prediction accuracy of Gas Turbines (GTs) performance decay. First a data cleaning technique for robust Linear Regression imputation is proposed based on the Mixed Integer Linear Programming. Then, experiments are conducted to compare the effect of commonly used data cleaning, normalization and reduction techniques on the ML prediction accuracy. Results revealed that the best prediction accuracy of GTs decay, found with the k-Nearest Neighbors ML algorithm, considerately deteriorate when changing the data preparation steps and/or techniques. This study has shown that, for effective CBM application in industry, there is a need to develop a systematic methodology for design and selection of adequate data preparation steps and techniques with the proposed ML algorithms.

Keywords: Data Preparation, Machine Learning, Condition-Based Maintenance, Performance Decay, Prediction

Introduction

Under the Industry 4.0 paradigm, reliability of industrial assets and production machines is very important. Towards smart factory, machine health monitoring and management aims to operate with near zero breakdown. In this context, (Aivaliotis *et al*., 2019) have proposed a methodology based on the Digital Twin concept in order to enable predictive maintenance for manufacturing systems using Prognostics and Health Management techniques. Also, as reported in, (Carvalho *et al*., 2019; Diez-Olivan *et al*., 2019), several recent Prognostics and Health Management projects were able to reach this level of profitable maintenance using Machine Learning (ML). Specifically, based on collected sensors data, ML intelligent predictive algorithms are implemented to reach successful Condition-Based Maintenance (CBM) strategy. This success is mainly related to the capability of such ML algorithms to handle high dimensional and multivariate data from various sensors and predict the degradation and future failure states (Accorsi *et al*., 2017). For example, (Márquez *et al*., 2020), have shown the success of Artificial Neural Network (ANN) in identifying deterioration of bearings. Also, in (Coraddu *et al*., 2016) authors have shown the potential of ML algorithms in predicting propulsive performance degradation of a naval vessel powered by Gas Turbines (GTs).

However, as pointed out in (Bennane and Yacout, 2010; Loukopoulos *et al*., 2017; Diez-Olivan *et al*., 2019), the majority of these works have centered in comparing performances of different ML algorithms in degradation prediction; however, they did not give enough insight on the data preparation phase. Only a few works, such as (Bukhsh *et al*., 2020) have exhibited the data preparation technique used before applying predictive ML models for bridges intelligent maintenance. The data preparation phase generally includes three steps: Data cleaning, i.e., handling missing data and outliers, data reduction, i.e., reducing the data size by aggregation, elimination redundant feature, etc. and data normalization, (Han *et al*., 2011). In other fields, such as biological and medical research, it is well known that this data preparation phase can greatly improve or deteriorate the ML prediction accuracy, (Perez-Rey *et al*., 2006). For example,

(Nawi *et al*., 2017) have shown that the prediction accuracy of the Artificial Neural Network (ANN) ML algorithm, considerably deteriorates when the data normalization step is conducted using the Min-Max technique rather than the Z-score one.

Thus, the present work aims to investigate the effect of the data preparation steps on ML prediction of GTs performance decay. Experiments are conducted on data generated from a simulator of a gas turbine and were formerly used in (Coraddu *et al*., 2016; Cipollini *et al*., 2018) to show the benefit of ML in predicting the decay of GTs performances installed on naval vessel. This work intends to go further, investigating the effect of the used technique during data preparation steps on the ML prediction performances. To address this issue, a new Mixed Integer Linear Programming (MILP) model is firstly proposed to implement a robust linear regression imputation as a data cleaning technique. This model is based on former works, in the biomedical field, conducted by (Omelchenko, 2014; Poos *et al*., 2016), which showed the benefit of MILP modelling in avoiding over-emphasizing outliers when building regression models. This MILP model is implemented and used as a cleaning technique in the data preparation step. The benefit of its use is shown through comparison with other more common techniques such as mean imputation. Then, in order to analyze the effect of the different data preparation steps and techniques, three ML algorithms are used: Linear Regression (LR), k-Nearest Neighbors (k-NN) and Neural Network (NN) to predict the GT degradation coefficients. The effect is measured by comparing the corresponding Mean Absolute Percentage Error (MAPE). The results show that this MAPE is not only sensitive to the steps but also to the technique used to prepare the data before applying ML algorithm for degradation state prediction.

This study is organized as follows. Section 2 reviews previous relevant studies related to the data preparation steps and techniques used in CBM context. Section 3 presents the MILP formulation of a data cleaning technique developed to handle missing data. Section 4 introduces the methodology followed in this study. It also includes a description of the considered dataset, followed by a visualization of the effect of data imputation techniques. Section 5 includes the computational experiments and comparison of ML algorithms after performing the data preparation techniques. Finally, conclusions are discussed.

Literature Review

The success of CBM over preventive strategies is mainly due to its capability to avoid unnecessary maintenance tasks by taking actions only when abnormal behaviors of a physical asset is detected. Diez-Olivan *et al.* (2019) gave an extensive review on machinery diagnostics and prognostics implementing CBM. Since the success of this CBM strategy is highly dependent on the prediction accuracy, several researchers have focused on using ML algorithm to enhance the prediction of the failure state, (Prajapati and Ganesan, 2013). According to (Coraddu *et al*., 2016), ML approaches have the capability to identify complex pattern from the received sensory data and provide better estimation of the degradation state. Review of recent works using ML algorithm to predict future degradation state and the remaining useful life of assets is given in (Carvalho *et al*., 2019). Although extensive work has been carried out under ML for CBM, yet little attention has been paid to the data preparation phase. According to (Bennane and Yacout, 2010; Loukopoulos *et al*., 2017; Diez-Olivan *et al*., 2019), the relevance of the data preparation phase has been widely recognized in the literature but still few research efforts have been carried out to address this issue in CBM context.

In this context, (Bennane and Yacout, 2010; Ragab *et al*., 2016) mainly focused on data cleaning by identifying outliers using the Inter-Quartile Range (IQR) method and handling the missing data using different techniques, namely the complete case analysis, mean imputation and k-Nearest Neighbors (k-NN) imputation. Data were cleaned using the Logical Analysis of Data (LAD) model; then a supervised learning algorithm was used to predict the health state of an oil transformer system. Loukopoulos *et al*. (2017) have also presented different imputation techniques to handle the missing data, for the CBM application on centrifugal compressors. Among these techniques, autoregressive model, k-NN imputation, Self Organizing Map (SOM) and Bayesian Principal Components Analysis (BPCA) were used to fill the missing data. Tsang *et al*. (2006) proposed three data cleaning procedures to handle missing data, in the practice of CBM optimization. The first one is based on completely recorded observations, also known as the complete case analysis. The second is the imputation-based procedures such as mean imputation or regression imputation. The last proposed procedure is based on models in which the models' parameters are estimated using techniques such as the Expectation Maximization (EM) algorithm. Hu *et al*. (2012) implemented the z-score normalization technique to prepare data before the application of the Recurrent Neural Network (RNN) model used to predict the Remaining Useful Life (RUL). (Ghasemi *et al*., 2013) reduced the dimensionality of the condition monitoring data set using the Principal Component Analysis (PCA), before the application of the LAD model. Ragab *et al*. (2016) applied feature selection and extraction methods to reduce the data dimensionality. Feature extraction was performed using the Discrete Wavelet Transform (DWT) method, while the feature selection was performed using

the Distance Evaluation Technique (DET). These data preparation techniques improved the accuracy of LAD algorithm used to estimate the RUL of a turbofan engine. Bukhsh *et al*. (2020) have proposed a predictive model based on the Neural Network algorithm for efficient bridge maintenance planning. They have used the z-score data normalization.

As seen from the reviewed works, when data preparation phase is conducted in the CBM context, authors generally focused on a single aspect: Data cleaning, data reduction or normalization. However, in other fields, such as biomedical, it is well known that efficient data preparation may have a major impact on the ML algorithm performance (Nawi *et al*., 2017). For example, (Wu *et al*., 2019) have shown the benefit of conducting data cleaning and then data reduction to better predict fatty liver disease. Singh and Singh (2019) prepared clinical biomedical data set by performing three different steps. The data cleaning step consisted of imputing missing values either by their means or by their modes. The dataset was then normalized using the z-score normalization technique. The final step, data reduction, was conducted using three feature selection techniques, which are χ^2 statistic, symmetric uncertainty and gain ratio. Daoud and Mayo (2019) presented different techniques of data normalization and data reduction implemented before building ML models used for cancer prediction purpose. Data normalization was carried out using different methods. The PCA method was performed to reduce the data dimensionality. Kotsiantis *et al*. (2006) also conducted data cleaning and data normalization, in order to achieve better performance of supervised algorithms.

Obviously, according to these works done in the biomedical field, preparing data with different steps may considerably affect the ML model performance. More recently, some researchers have investigated the issue of using different techniques when conducting a data preparation step. Nawi *et al*. (2017) compared the effect on the performance of the Artificial Neural Networks (ANN), of three different normalization techniques; namely the Min-Max Normalization, Z-Score Normalization and Decimal Scaling Normalization. Results on all datasets revealed that when the z-score technique is applied, the ANN produces the best performance. Also (Lokman *et al*., 2019), have shown the effect of different normalization techniques on the accuracy performance of anomaly detection in cyberattacks.

This literature review reveals that researcher on CBM generally focused on proposing new ML models for better degradation and failure prediction. To the best of our knowledge and according to the recent review (Carvalho *et al*., 2019), comparing the effect of different data preparation steps and techniques on this prediction accuracy has not yet been addressed in the CBM context. Thus, in this study the effect of the widely used steps and techniques on the data preparation phase is analyzed. A new data cleaning technique named the robust Linear Regression is also introduced. The proposed model to implement this imputation technique is given in the next section.

Mathematical Model of the Robust Linear Regression

In this section, the MILP model proposed to implement a robust linear regression for imputation purpose is presented.

Problem Description

One of classical imputation methods to handle the missing data considers the simple LR model. However, in order to conduct regression a commonly known drawback of this simple LR model is its lack of robustness to some unusual observations, usually called *outliers.* To overcome this drawback, (Omelchenko, 2014) has proposed a more robust linear regression using MILP. This model allows the detection and the exclusion of potential outliers from the regression model. This approach was originally developed for predicting chemical compound's properties of peptides. This regression model was implemented and tested on real biological dataset examples, proving that it has better predictive performance compared to the regular LR technique. Poos *et al*. (2016) have also shown that MILP is a powerful approach which avoids over-emphasizing outliers when building LR models. Based on this study, a new MILP is proposed to adapt this robust Linear Regression approach for the imputation purpose.

Hence, supposing that n observations in a dataset $D = \left\{\left(x_i, y_i\right)\right\}_{i=1}^{n}$ are considered, the missing values are estimated using the following regression equation:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1,...n, \tag{1.1}$$

where, β_0 and β_1 are the regression coefficients and ε_i is the error term.

The simple LR model is not robust to *outliers*. In fact, when independent variables x_i or the dependent ones y_i include outliers, the estimate of the regression coefficients β_0 and β_1 will be biased according to Eq. (1.1). Thus, the simple LR imputation method might be not reliable in the presence of outliers to estimate missing data. Therefore, the least absolute deviations method is performed to estimate the regression coefficients β_0 and β_1. This approach consists of minimizing the sum of the absolute error ε_i, which is the distance between the actual dependent variable Y and the predicted dependent variable $\hat{Y}$, such as $\hat{y}_i = \beta_0 + \beta_1 x_i$ for $i = 1, \ldots n$, as follows:

$$Minimize \sum_{i=1}^{n} |\varepsilon_i| = \sum_{i=1}^{n} |y_i - \hat{y}_i| = \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|. \quad (1.2)$$

The linearization of this objective function results in having a MILP model. Constraints are added to this mathematical program so that only the "*good*" observations are taken into account when estimating the regression coefficient of the simple LR model. In other words, if the observation is considered as an outlier, the MILP model disregards it and does not take it into account to estimate the coefficients. Then, a maximum number of outliers could be detected and removed by the MILP model. For this imputation technique, the following MILP is proposed.

Mathematical Formulation

The sets, parameters and decision variables of the MILP model are defined as follows.

Sets

$N = \{1, \ldots, n\}$: Set of observations in a dataset (indexed by i).

Parameters

y_i: Value of the i^{th} observation of the dependent variable Y in the dataset
x_i: Value of the i^{th} observation of the independent variable X in the dataset
o: Maximum number of outliers that can be detected by the model, M: Maximum value of observations in the dataset

Decision Variables

β_0: A continuous variable that represents the Y-intercept of the regression line in the simple linear regression model
β_1: A continuous variable that represents the slope of the regression line in the simple linear regression model
θ_i: A binary variable equal to 1, if the observation i is an outlier, 0 otherwise
α_i: A continuous non-negative variable corresponding to each observation i in the dataset

In order to perform the robust simple linear regression, the MILP model is formulated as follows:

$$Minimize \sum_{i \in N} \alpha_i \quad (2.1)$$

Subject to:

$$y_i - \beta_0 - \beta_1 x_i \leq \alpha_i + M\theta_i, \ \forall i \in N, \quad (2.2)$$

$$y_i - \beta_0 - \beta_1 x_i \geq \alpha_i + M\theta_i, \forall i \in N, \quad (2.3)$$

$$\sum_{i}^{n} \theta_i \leq o, \forall i \in N, \quad (2.4)$$

$$\alpha_i \geq 0, \ \forall i \in N, \quad (2.5)$$

$$\theta_i \in \{0,1\}, \ \forall i \in N, \quad (2.6)$$

$$\beta_0 \in \mathbb{R}, \quad (2.7)$$

$$\beta_1 \in \mathbb{R}, \quad (2.8)$$

where, M is a parameter that has a large value.

In Model (2.1)-(2.8), the objective function (2.1) is to minimize the sum of the absolute distance between the actual dependent variable Y and the predicted variable $\hat{Y}$ in order to find the coefficient values β_0 and $\beta_1$which provide the best-fitting line through the data points. As mentioned in section 3.1, the regression model is constructed by solving the least absolute deviations problem. To linearize the non-linear objective function (1.2) in this problem and to remove the absolute value operator, the positive auxiliary variable α_i is introduced. Constraints (2.2) and (2.3) require excluding the outliers from the regression model.

Actually, supposing that a good observation, θ_i will be equal to 0. If the observation is an outlier, θ_i will be equal to 1 and α_i will be restricted to be equal to 0 while $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$ will always be in the range of $[-M, M]$. Constraints (2.4) express the restricted number of outliers that could be detected by the model. Finally, Constraints (2.5)-(2.8) guarantee the variables nature. This MILP model is implemented using CPLEX (version 12.7). The following section will show how the imputation is conducted using this robust Linear regression approach.

Methodology

This section presents the used methodology in this study. Section 4.1 describes the ML process and the considered Gas Turbine (GT) dataset. Section 4.2 lists the selected techniques for each data preparation step. In section 4.3, the data is visualized after the cleaning step using the robust LR imputation technique.

ML Process and Data Description

The main goal of using ML algorithm is to characterize the behavioral pattern of the assets based on the sensors data herein provided in the GT dataset. To conduct this ML process, the phases presented by (Diez-Olivan *et al*., 2019), summarized in the following schematic diagram, Fig. 1, are used. The considered data is split into two sets: Training set and test set: The training set representing 70% of the dataset and the test set which contains the remaining 30%. Data preparation phase is conducted using different steps with the corresponding different techniques. These techniques are presented later in section 4.2. The ML model is built using the training data. Then the model makes predictions of the

output variable which indicates the asset's health using the test data. Finally, an evaluation of the model performance is carried out by calculating the error, which is resulted from a comparison between the actual performance decay and the one predicted by the ML model.

This methodology is applied to predict the performance decay of GTs used for a naval propulsion plant. The corresponding dataset was generated by (Coraddu *et al*., 2016), using a simulator of a naval vessel with GT propulsion plant. This dataset is given in UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/). Researchers in (Coraddu *et al*., 2016; Cipollini *et al*., 2018) worked on designing a CBM approach and applying it to GTs data used for naval propulsion plants, enabling the diagnosis and the prognosis of the naval assets. The dataset was used to derive a ML predictive model in order to monitor performance decay over time of the propulsion system and to identify in advance potential failures.

In the present work, the system's behavior is described by two parameters, the turbine and the compressor degradation coefficients. Both parameters represent the output variables in the dataset. Each of these two outputs' ranges has been sampled with a uniform grid of precision 10^{-3}, in order to get a good granularity representation. Given that the dataset is labeled with degradation coefficients which represent a continuous range, regression models are used to investigate these two parameters in order to perform the CBM approach. The dataset also includes the following 16 features: Lever position, Ship speed, Gas turbine shaft torque, Gas generator rate of revolutions, Gas turbine rate of revolutions, Starboard propeller torque, Port propeller torque, HP turbine exit temperature, GT compressor inlet air temperature, GT compressor outlet air temperature, HP turbine exit pressure, GT compressor inlet air pressure, GT compressor outlet air pressure, GT exhaust gas pressure, Turbine injection control and Fuel flow.

In order to predict the performance decay of GTs, the following three ML algorithms: Linear Regression (LR), k-Nearest Neighbors (k-NN) and Neural Network (NN) are used. The k-NN is a supervised learning algorithm. k-NN algorithm first selects the k target variables whose associated feature (input) is the closest to the new input, according to a distance and then the algorithm determines the output value to be predicted based on the k selected target variables. A Neural Network (NN) algorithm; specifically, a multilayer perceptron, which is an ANN, is implemented. It is a supervised learning model that can learn a non-linear function by training a set of inputs and an output in a dataset. For more insight on these ML algorithms, interested reader can consult (Basheer and Hajmeer, 2000; Kramer, 2013). These algorithms were selected based on their popularity among the practitioners and researchers (Shukla and Kumar, 2019; Ray, 2019). Each of these algorithms has its own characteristics. The LR models are known for being easily interpretable, as the regression coefficients indicate the most important features. The k-NN algorithm, on the opposite of the LR algorithm, does not require a linear relationship between the inputs and the target variable, providing a more flexible approach. On the other hand, the NN algorithm is known by its flexibility provided by the hyper parameters, making it able to learn and model non-linear and complex relationships. Nevertheless, NN considers a greater number of hyper parameters compared to k-NN. Also, the k-NN and NN models are more complex to interpret compared to LR models. The next paragraph exhibits the implemented steps and techniques at the data preparation phase.

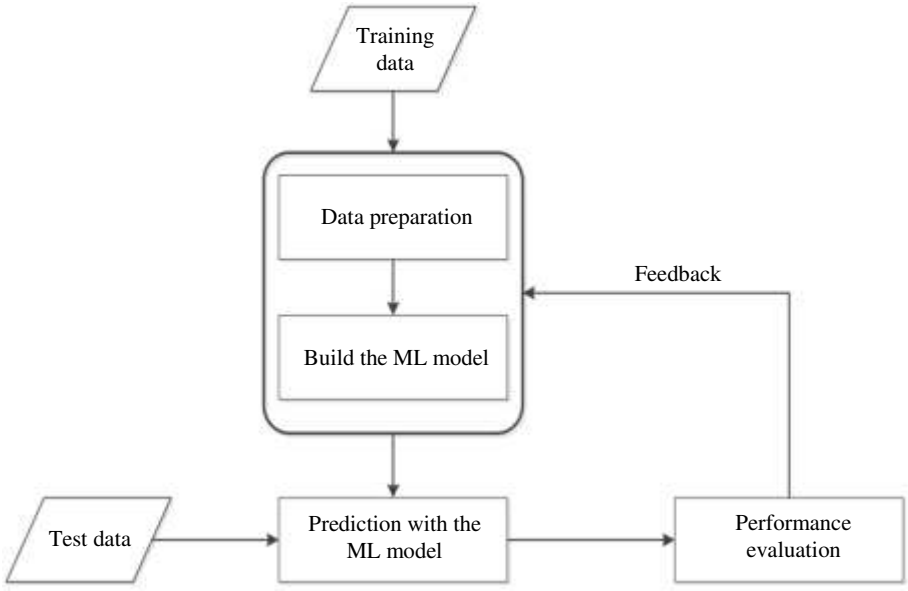

Fig. 1: Schematic diagram of a machine learning process (Diez-Olivan *et al*., 2019)

Data Preparation Techniques

In this study, the effect of the three commonly used steps in data preparation, namely data cleaning, data normalization and data reduction, are investigated.

For data cleaning the following techniques were investigated:

- **Mean Imputation** consists of replacing missing values with the average of the non-missing values in a variable, (Malarvizhi and Thanamani, 2012)
- **Linear Regression (LR) imputation** aims to model a linear relationship between a dependent variable and one independent variable, if a simple LR model is considered, (Yan and Su, 2009). The idea is to set the feature that has missing values as the dependent variable and set another feature as the independent variable, assuming the existence of a linear relationship between these two features. In order to construct the LR model, the observations which contain missing values in the two features are removed. In fact, regression coefficients of the model can only be estimated by using a complete set of observations. Once the simple LR model is built, it is possible to estimate the missing values of the feature involved. In this study, the coefficients are estimated using the least-squares method
- **EM Algorithm** is an iterative algorithm that finds the maximum likelihood estimates of a model parameters defined by an incomplete dataset. The number of iterations of this algorithm was set at 50
- **Robust Linear Regression (LR) imputation** uses the same methodology as the regular LR imputation technique; however, it relies on a different approach to estimate the regression coefficients. This approach is based on the MILP presented in section 3. For the following experimentation, the number of maximum outliers o that the MILP can detect was set to be equal to 10% of the total observations in the dataset. This choice is based on preliminary experiments conducted with a range of o values corresponding to five different percentages of the total observations in the dataset (10 to 50%) and selecting the o value providing the lowest MAPE when implementing the regression algorithms. Bartoli and Olsen (2006; Filzmoser, 2005) used a similar method to select the maximum percentage of outliers
- **k-NN imputation** determines the average value of the k closest neighbors that are the most similar to the feature with the missing value. The calculated average is then imputed in the missing value. The closest or most similar k neighbors are selected by using the similarity metric, also called the distance metric (Troyanskaya *et al*., 2001). In this study, the distance metric selected is the Euclidean distance and the number of neighbors k was set to 5. The neighborhood size k selection plays an important role in resulting in a good performance of k-NN. However, as pointed out by (Loukopoulos *et al*., 2017), no global rule is set for determining this optimal k. In the present paper, preliminary experiments as the one done by (Thanh Noi and Kappas, 2018), with different values of k between 1 and 20, are conducted. Then the k value which gave the lowest value of MAPE is selected

Data normalization consists of scaling the features so they can fall within a smaller range, improving the efficiency and the accuracy of ML algorithms, (Han *et al*., 2011). Herein follows the two data normalization techniques used in this study:

- **Min-Max normalization** is performed using the following formula, from (Kotsiantis *et al*., 2006):

$$v' = \frac{v - \min_A}{\max_A - \min_A}\left(new\_\max_A - new\_\min_A\right) + new\_\min_A$$

v is the value of the old feature and v' is the value of the normalized feature. In this study, the range chosen is [0, 1], in other words, $new\_\min_A$ is equal to 0 and $new\_\max_A$ is equal to 1. This means that, for every feature, its minimum value is converted to 0 and its maximum value is transformed into 1 and all other values get transformed into a decimal between 0 and 1:

- **Z-score normalization** uses the mean of all the values of the feature and its standard deviation, (Kotsiantis *et al*., 2006), as stated below:

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

Data reduction step creates a reduced representation of a dataset based on the original one making its volume much smaller while maintaining its integrity. Data dimensionality reduction is chosen among the data reduction approaches. Dimensionality reduction consists of reducing the number of features in the original dataset, aiming to train ML algorithms with fewer features (Han *et al*., 2011). The following techniques were selected to reduce the data dimensionality:

- **Principal Component Analysis (PCA)** reduces the number of features by converting the correlated ones into linearly new uncorrelated features called principal components. PCA thus creates new orthogonal linear combinations based on the initial features with the largest variance. The selection of the number of principal components is related to the percentage of the total variance which needs to be explained in (Abdi and Williams, 2010). In this

study, the number of components is selected such as the percentage of the total variance explained by the PCs is equal or greater to 95%

- **Factor Analysis (FA)** is also a linear dimensionality reduction method. FA assumes that the observed variables depend on a lower number of some unknown latent variables. The purpose of FA is to uncover such linear relations and to explain the covariance among the observed variables. FA can therefore reduce the datasets by using the new independent latent variables, called factors (Fodor, 2002)

Visualization of Data Cleaning

The purpose of this section is to visualize data before and after the data cleaning step, allowing us to observe the results of some techniques proposed in the previous sections, particularly the robust LR imputation technique, presented in section 3. As mentioned by (Kohavi, 2001) data visualization tools provide an easier way to visualize trends, understand patterns and the relationship between the features in the dataset. They make it easier to identify areas which need attention and that can affect the performance of the ML algorithms. It is worth to mention that numerical comparison will be conducted in the next section. The objective is to visualize the effect of the imputation technique in data cleaning step. Knowing that there are no missing values in the dataset, 10% from each feature, are randomly deleted, in order to study the effect of handling the missing data with the proposed techniques. The trend of the feature "HP turbine exit temperature" in the original dataset is shown in Fig. 2. The *x* axis represents the first 200 observations or measures of the dataset and the *y* axis represents the feature value corresponding to each observation.

The line plot of the same feature is represented, in Fig. 3, after artificially creating the 10% of missing values. The randomly and discontinuously dispersed missing data are shown in Fig. 3 with gray circle.

Then the proposed MILP model is applied, in section 3, to perform the data cleaning step using the robust Linear Regression imputation technique. The corresponding imputed data are plotted in Fig. 4. The comparison of the Fig. 2 with the Fig. 4 shows that the difference between the feature trend in the original dataset and the one after handling the missing data using the robust LR imputation technique is minimal.

In order to visually compare the effect of different imputation techniques in filling the missing data, the mean imputation technique is applied and the corresponding trend is plotted in Fig. 6. This plot shows that the feature trend in the imputed dataset using the robust LR imputation technique (Fig. 5) is more similar to the actual feature trend in the original dataset than when using the mean imputation technique (Fig. 6).

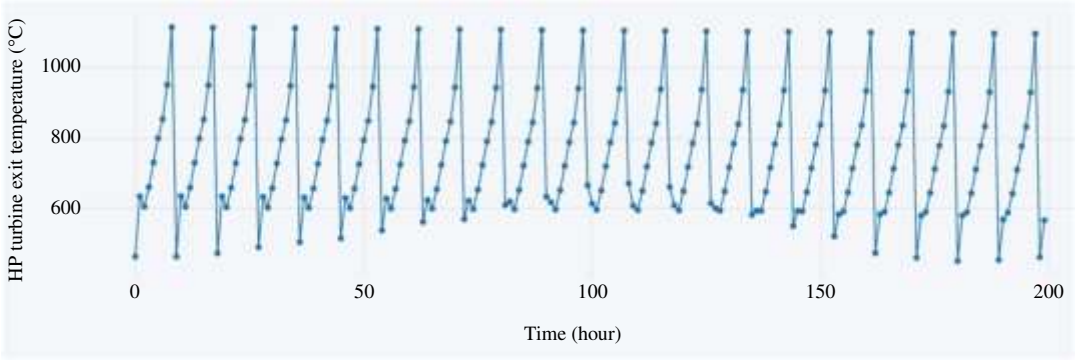

Fig. 2: Line plot of the feature "HP turbine exit temperature" in the original dataset

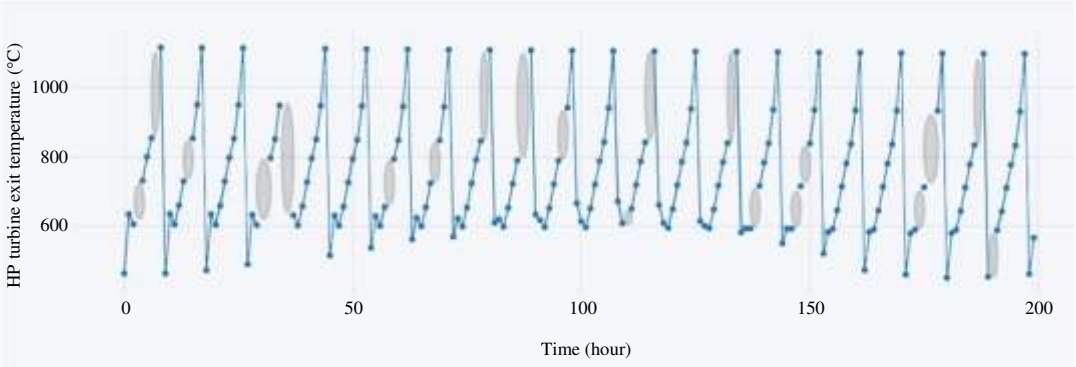

Fig. 3: Plot of feature "HP turbine exit temperature" with missing values

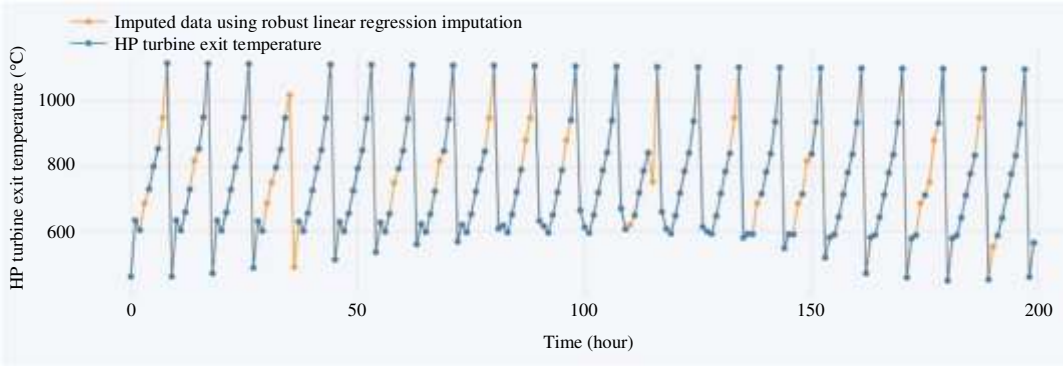

Fig. 4: Plot of feature "HP turbine exit temperature" using the robust LR imputation technique

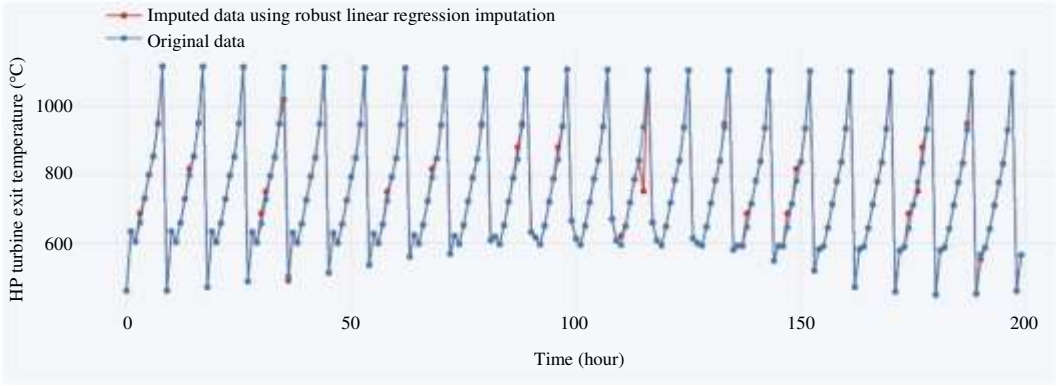

Fig. 5: Plot of the feature "HP turbine exit temperature" using the robust LR imputation technique Vs. original data

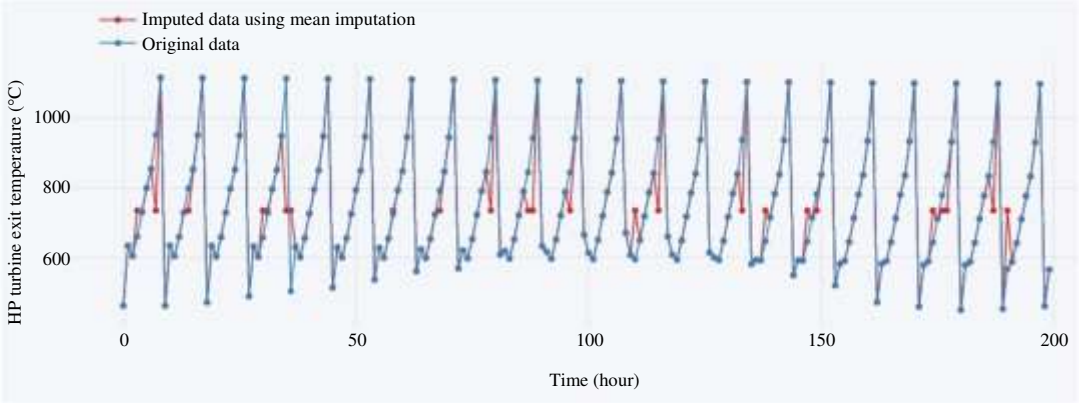

Fig. 6: Plot of the variable "HP turbine exit temperature" using the mean imputation technique Vs. original data

In this section, the approach of imputing missing data in the feature "HP turbine exit temperature" with different cleaning techniques is shown. In the following section, the effect of these different data preparation steps and techniques on the ML accuracy prediction of GTs performance decay will be explored in more depth.

Computational Experiments and Results

In order to measure the effect of the data preparation steps and techniques, the metric used to evaluate the performance of the ML algorithms has to be first defined. Then, the experimental results of testing the data preparation steps with the different techniques listed in section 4.2 are presented.

Experiments were conducted using Python 3.6 programming language.

Metric

To evaluate the effect of different steps and techniques on the performance of the prediction accuracy of the ML algorithms, the Mean Absolute Percentage Error (MAPE) was used, it is defined as follows:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{\left|y_i - \hat{y}_i\right|}{y_i}$$

y_i is the value of the i^{th} observation of the output in the original dataset, $\hat{y}_i$ is the value of the i^{th} observation of the predicted output and n is the number of observations in the test dataset. The MAPE, as a relative measure, is a frequently used ML evaluation metric and is known for its advantages of scale-independency and interpretability. Coraddu *et al*. (2016; Cipollini *et al*., 2018), the authors chose the same metric on the same GT dataset example. Also, this metric has been successfully adopted in similar studies in recent years (Wisyaldin *et al*., 2020; Velasco-Gallego and Lazakis, 2020). Given that there are two outputs, the "GT compressor decay coefficient" and the "GT decay coefficient", this multi-target problem is tackled by decomposing it into two single target sub-problems. This same calculation approach was used in (Coraddu *et al*., 2016; Cipollini *et al*., 2018). This means that the ML algorithms predict each output separately, by considering one decay coefficient at a time and then the average of the two MAPEs is computed. A better prediction accuracy is given with the lowest MAPE.

Effect of Data Preparation Steps on ML Performances

In order to study the effect of data preparation steps on the performance of ML algorithms, different combinations of these steps on the GT dataset were applied. The following four combinations are considered:

1. The application of only one step which is the data cleaning step
2. The application of both steps: Data cleaning and data normalization
3. The application of both steps: Data cleaning and data reduction
4. The application of all 3 steps: Data cleaning, data reduction and data normalization

Following the ML process presented in section 4.1, these four combinations are applied considering the different techniques presented in section 4.2. Results for the three ML algorithms: The LR, k-NN and the NN are computed. For each of these algorithms, if only the data cleaning step is performed, five different MAPEs are obtained, each one corresponding to a data cleaning technique. If both steps of data cleaning and data normalization are performed, the results show ten different MAPE, considering the five data cleaning techniques and the two data normalization techniques used. For each combination and for each algorithm, the lowest MAPE is selected. The following graph, Fig. 7 summarizes the outcome of this experimentation.

The Fig. 7 shows that the NN model's MAPE, after only cleaning the data, is significantly high, indicating that this algorithm is considerably sensitive to the data normalization and/or the data reduction steps. Besides, results point out that the k-NN and NN algorithms perform better than the LR algorithm, regardless of the data preparation steps. This figure also shows that the combination of steps chosen to prepare the data has an impact on the performance of these algorithms. In fact, the best performance obtained by the LR (1.35%) is resulted from implementing only the data cleaning step. Whereas, when the k-NN or the NN models are used, better performance is found when the 3 data preparation steps are performed. As shown in the Fig. 7, the *K*-NN model's MAPE trained using the prepared dataset by performing all the 3 steps (0.01%) is remarkably lower than the one resulted from only performing the data cleaning step, which is equal to 0.93%. Likewise, the MAPE of the NN model built after performing all the data preparation steps, is the lowest (0.88%). This illustrates that MAPE of the ML model varies according to the used data preparation steps. In other words, the prediction accuracy of GTs performance decay is related to the choice of data preparation steps. The next section further investigates the effect of the chosen technique when conducting a data preparation step. In order to elucidate this issue, the effect of using different data cleaning techniques on the performance of ML algorithms is analyzed.

Effect of Data Cleaning Techniques on ML Performances

This section focuses solely on the effect of changing the data cleaning technique. The aim is to study the effect of these techniques on the performance of ML algorithms, specifically the k-NN and the NN algorithms, since they gave the best accuracies, in section 5.2. For that purpose, the effect of using the EM algorithm as a data cleaning technique is analyzed compared to the robust Linear Regression technique. The present section will not focus on the case where only data cleaning is conducted before NN implementation, because as shown in Fig. 7, the

corresponding MAPE is extremely high. Results are provided in the following graph, Fig. 8.

As seen in Fig. 8, performance of the k-NN algorithm considerabely deteriorates when using the EM cleaning technique rather than Robust LR. In fact, the NN algorithm produces now the best performance rather than k-NN. Actually, the choice of data cleaning technique has also a significant impact on the performance of ML algorithms. In fact, the chosen data preparation steps and the used techniques to implement these steps have an impact on the ML prediction accuracy of the GTs performance decay. This result is in concordance with the one provided in (Nawi *et al*., 2017) where authors have found that ANN prediction accuracy considerably deteriorates when data normalization step is conducted using the Min-Max technique rather than Z-score.

Experiments to investigate the effect of data normalization and reduction techniques on the prediction accuracy of the GTs performance degradation are also conducted. Results for the k-NN algorithm are plotted in the following graph, Fig. 9.

As shown in Fig. 9, performance of k-NN algorithm varies according to the used data normalization and reduction techniques. Specifically, this figure reveals that the best accuracy of the k-NN algorithm is reached when data cleaning is conducted with the robust Linear Regression, data normalization is done with the z-score technique and finally the data reduction with PCA. Given that these techniques gave the best prediction accuracy, next section aims to investigate whether this performance is affected by the combination of data preparation steps.

Effect of Data Cleaning Techniques in Data Preparation Steps

This study is conducted for the k-NN algorithm. For each of the 4 combinations of data preparation steps and for each data cleaning technique implemented, the results for z-score normalization and PCA reduction which gave the best prediction of the k-NN algorithm are plotted. Results are given in Fig. 10, to visualize the MAPE of the k-NN algorithm after preparing the data.

The Fig. 10 indicates that the best performance of the k-NN algorithm is obtained by selecting the robust Linear Regression imputation as a data cleaning technique. However, this algorithm gives the best performance, only if its application is carried out on normalized and reduced data. In fact, results reveal that the MAPE of the k-NN algorithm is equal to 0.01% when applying all the three steps of data preparation while using the robust linear regression imputation technique. However, when only data cleaning and data reduction are conducted, the MAPE deteriorates and reaches 0.94%. For this case, the k-NN imputation technique provides the best accuracy. That means that the choice of the steps is also sensitive to the technique used to implement these steps. Thus, to benefit from the good prediction accuracy of the GTs performance decay using the k-NN, it is mandatory to impute data using the robust Linear Regression technique and then conduct data normalization using the z-score technique and finally reduce data using the PCA technique. The importance of normalizing the data using the z-score technique before applying PCA was also found in (Gao *et al*., 2019; Obaid *et al*., 2019).

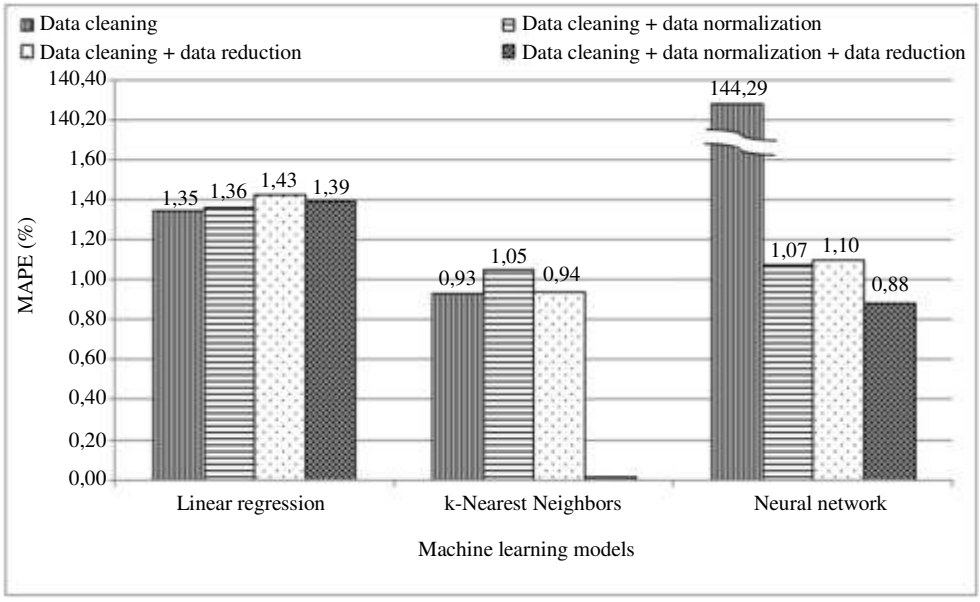

Fig. 7: Performance of ML algorithms with lowest MAPE

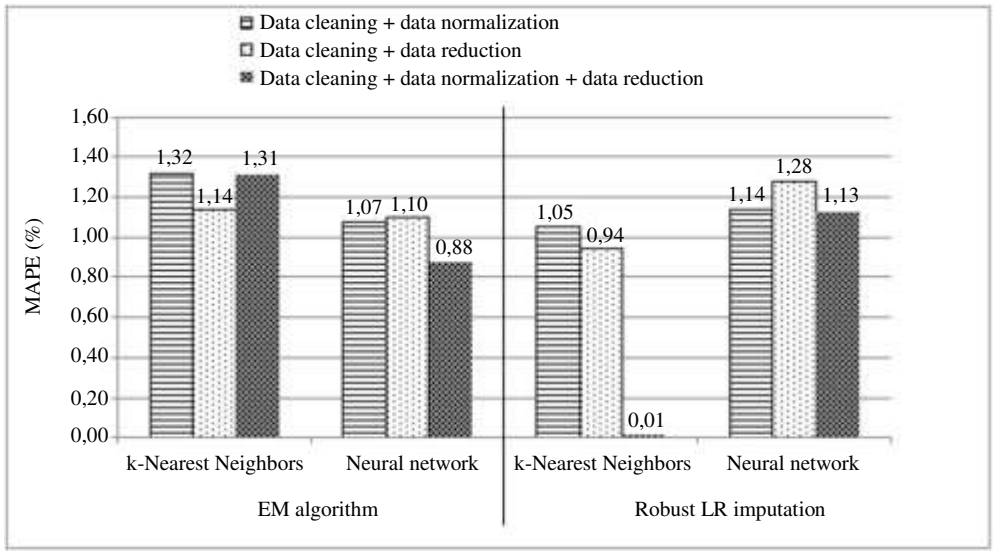

Fig. 8: Performance of k-NN and NN models when cleaning the data using the EM technique and the Robust LR imputation technique

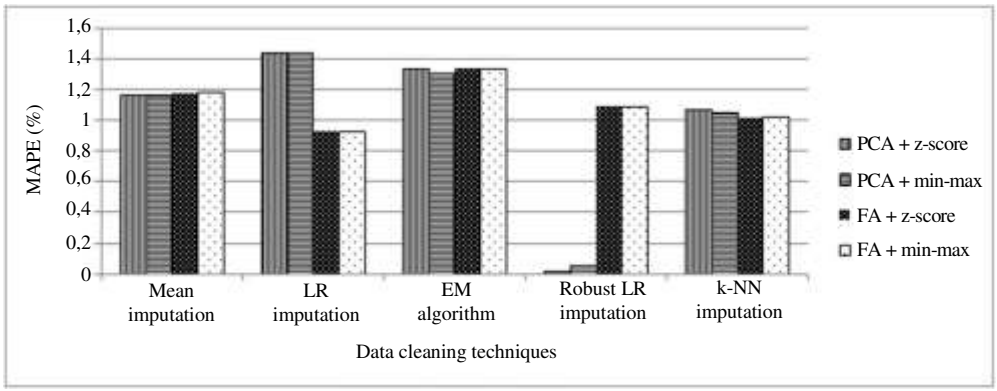

Fig. 9: Performance of k-NN with different data normalization and reduction techniques

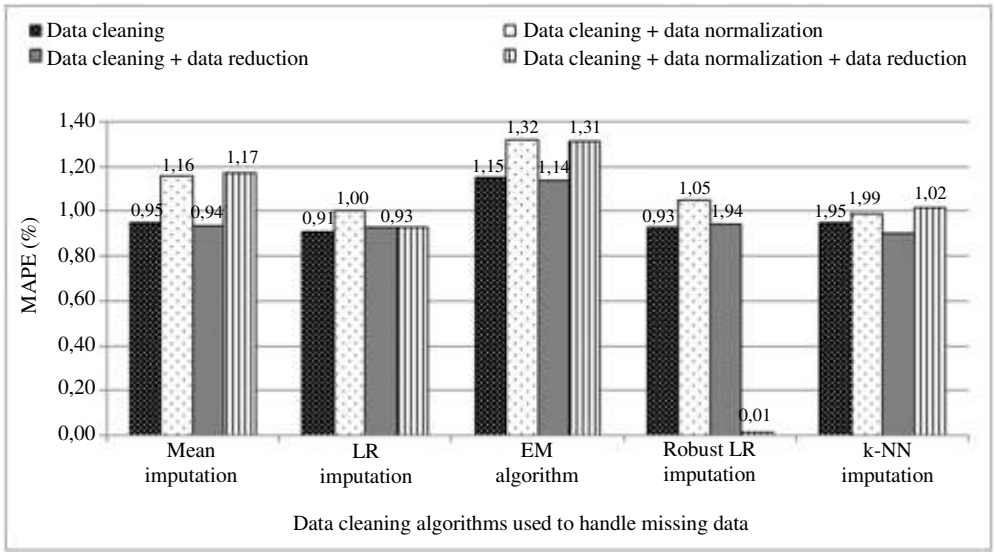

Fig. 10: Performance of k-NN algorithm after data preparation

The figure also shows that the MAPE is the lowest when performing both steps of data cleaning and data reduction and using one of the following techniques to handle missing data: Mean imputation, EM algorithm, k-NN imputation. On the other hand, the performance of the k-NN algorithm is the best when applying all the three steps of data preparation in case the Robust Linear Regression imputation technique is implemented to handle the missing data. Hence, this confirms that the data preparation steps should be considered depending on the data cleaning technique to get the best performance of the ML algorithm. This also proves that the choice of the data preparation steps is sensitive to the technique used to clean the data.

These experimentations lead to the conclusion that changing the steps and/or the techniques during the data preparation may have considerable effect on the prediction accuracy of the ML algorithms. In fact, even though cleaning the data with the proposed MILP for robust Linear Regression imputation have led to the best prediction of the GTs performance decay; but to reach this prediction accuracy it is imperative to apply z-score normalization and PCA data reduction.

Conclusion

This paper investigates the effect of different data preparation steps and techniques on the ML prediction accuracy of GTs performance decay. An accurate degradation prediction model of GTs performance is highly desired to reach an effective CBM strategy which predicts the degradation of the propulsion plant over time and schedule maintenance in advance.

First, a literature review was conducted to distinguish the used techniques and steps for data preparation in the CBM context. Then, based on former works in the biomedical field, a new MILP model was proposed to implement a robust Linear Regression imputation technique. The effect of this imputation technique in the data cleaning step is shown with trend visualization. Computational experiments are conducted using three different Machine Learning algorithms to predict the performance decay of GTs. Results have revealed that the k-NN prediction model may provide the best prediction accuracy when data are cleaned using the robust linear regression technique, normalized using the z-score and reduced with the PCA technique. Otherwise, if data are prepared using different steps or with different techniques, the prediction accuracy deteriorates. In fact, the results show that when only data cleaning step is conducted, prediction accuracy of the NN surpasses k-NN. That means that the ML algorithm prediction accuracy of the degradation and failure state is affected by changes in the used data preparation steps and techniques.

The main finding is that, in order to benefit from the high prediction capability of the proposed Machine Learning algorithm in CBM, researchers should clarify how data have been prepared. Specifically, with which steps and techniques they have conducted the data preparation phase before applying ML algorithm for prediction of the degradation and failure state.

Future research is intended to explore in more depth this complex interaction between the proposed ML algorithms for CBM and the data preparation steps and techniques. It is evident that this interaction may considerably deteriorate prediction accuracy of degradation and failure, which in turn can have major impact on the applicability of these ML algorithms in practice. In fact, for effective CBM application there is a need to develop a systematic methodology for design and selection of the adequate data preparation steps and techniques with the proposed ML algorithms.

Acknowledgement

We want to express our thanks and gratitude to all those who helped us throughout this work.

Author's Contributions

Ons Masmoudi and Mehdi Jaoua: Conducted experiments, data-analysis and contributed to the writing of the manuscript.

Amel Jaoua and Soumaya Yacout: Coordinated the data-analysis and contributed to the writing of the manuscript.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433-459. https://doi.org/10.1002/wics.101

Accorsi, R., Manzini, R., Pascarella, P., Patella, M., & Sassi, S. (2017). Data mining and machine learning for condition-based maintenance. Procedia Manufacturing, 11, 1153-1161. https://doi.org/10.1016/j.promfg.2017.07.239

Aivaliotis, P., Georgoulias, K., & Chryssolouris, G. (2019). The use of Digital Twin for predictive maintenance in manufacturing. International Journal of Computer Integrated Manufacturing, 32(11), 1067-1080. https://doi.org/10.1080/0951192X.2019.1686173

Bartoli, A., & Olsen, S. I. (2006). A batch algorithm for implicit non-rigid shape and motion recovery. In Dynamical Vision (pp. 257-269). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-70932-9_20

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design and application. Journal of Microbiological Methods, 43(1), 3-31. https://doi.org/10.1016/S0167-7012(00)00201-3

Bennane, A., & Yacout, S. (2010, July). Processing missing and inaccurate data in a condition based maintenance database. In The 40th International Conference on Computers & Indutrial Engineering (pp. 1-5). IEEE. https://doi.org/10.1109/ICCIE.2010.5668354

Bukhsh, Z. A., Stipanovic, I., Saeed, A., & Doree, A. G. (2020). Maintenance intervention predictions using entity-embedding neural networks. Automation in Construction, 116, 103202. https://doi.org/10.1016/j.autcon.2020.103202

Carvalho, T. P., Soares, F. A., Vita, R., Francisco, R. D. P., Basto, J. P., & Alcalá, S. G. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. Computers & Industrial Engineering, 137, 106024. https://doi.org/10.1016/j.cie.2019.106024

Cipollini, F., Oneto, L., Coraddu, A., Murphy, A. J., & Anguita, D. (2018). Condition-based maintenance of naval propulsion systems with supervised data analysis. Ocean Engineering, 149, 268-278. https://doi.org/10.1016/j.oceaneng.2017.12.002

Coraddu, A., Oneto, L., Ghio, A., Savio, S., Anguita, D., & Figari, M. (2016). Machine learning approaches for improving condition-based maintenance of naval propulsion plants. Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment, 230(1), 136-153. https://doi.org/10.1177/1475090214540874

Daoud, M., & Mayo, M. (2019). A survey of neural network-based cancer prediction models from microarray data. Artificial Intelligence in Medicine, 97, 204-214. https://doi.org/10.1016/j.artmed.2019.01.006

Diez-Olivan, A., Del Ser, J., Galar, D., & Sierra, B. (2019). Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0. Information Fusion, 50, 92-111. https://doi.org/10.1016/j.inffus.2018.10.005

Filzmoser, P. (2005). Identification of multivariate outliers: a performance study. Austrian Journal of Statistics, 34(2), 127-138. https://ajs.or.at/index.php/ajs/article/view/vol34%2C%20no2%20-%207

Fodor, I. K. (2002). A survey of dimension reduction techniques (No. UCRL-ID-148494). Lawrence Livermore National Lab., CA (US). https://doi.org/10.2172/15002155

Gao, Z., Ding, L., Xiong, Q., Gong, Z., & Xiong, C. (2019). Image compressive sensing reconstruction based on z-score standardized group sparse representation. IEEE Access, 7, 90640-90651. https://doi.org/10.1109/ACCESS.2019.2927009

Ghasemi, A., Esmaeili, S., & Yacout, S. (2013). Development of Equipment Failure Prognostics Model Based on Logical Analysis of Data (LAD). Engineering Letters, 21(4). http://www.engineeringletters.com/issues_v21/issue_4/EL_21_4_12.pdf

Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems, 5(4), 83-124. https://doi.org/10.1016/B978-0-12-381479-1.00003-4

Hu, C., Youn, B. D., Wang, P., & Yoon, J. T. (2012). Ensemble of data-driven prognostic algorithms for robust prediction of remaining useful life. Reliability Engineering & System Safety, 103, 120-135. https://doi.org/10.1016/j.ress.2012.03.008

Kohavi, R. (2001, March). Data mining and visualization. In Sixth Annual Symposium on Frontiers of Engineering (pp. 30-40). National Academy Press.

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. International Journal of Computer Science, 1(2), 111-117. https://doi.org/10.1007/978-3-642-38652-7

Kramer, O. (2013). Dimensionality reduction with unsupervised nearest neighbors (p. 18). Berlin: Springer. https://link.springer.com/book/10.1007%2F978-3-642-38652-7

Lokman, S. F., Othman, A. T., Bakar, M. H. A., & Musa, S. (2019, July). The Impact of Different Feature Scaling Methods on Intrusion Detection for in-Vehicle Controller Area Network (CAN). In International Conference on Advances in Cyber Security (pp. 195-205). Springer, Singapore. https://doi.org/10.1007/978-981-15-2693-0_14

Loukopoulos, P., Zolkiewski, G., Bennett, I., Pilidis, P., Duan, F., & Mba, D. (2017). Dealing with missing data as it pertains of e-maintenance. Journal of Quality in Maintenance Engineering. https://doi.org/10.1108/JQME-08-2016-0032

Malarvizhi, M. R., & Thanamani, A. S. (2012). K-nearest neighbor in missing data imputation. International Journal of Engineering Research and Development, 5(1), 5-7. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.375.925&rep=rep1&type=pdf

Márquez, A. C., de la Fuente Carmona, A., Marcos, J. A., & Navarro, J. (2020). Designing CBM Plans, Based on Predictive Analytics and Big Data Tools, for Train Wheel Bearings. Computers in Industry, 122, 103292. https://doi.org/10.1016/j.compind.2020.103292

Nawi, N. M., Hussein, A. S., Samsudin, N. A., Hamid, N. A., Yunus, M. A. M., & Ab Aziz, M. F. (2017). The effect of pre-processing techniques and optimal parameters selection on back propagation neural networks. International Journal on Advanced Science, Engineering and Information Technology, 7(3), 770-777. https://doi.org/10.18517/ijaseit.7.3.2074

Obaid, H. S., Dheyab, S. A., & Sabry, S. S. (2019, March). The impact of data pre-processing techniques and dimensionality reduction on the accuracy of machine learning. In 2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON) (pp. 279-283). IEEE. https://doi.org/10.1109/IEMECONX.2019.8877011

Omelchenko, O. (2014). Mixed-integer linear programming robust regression with feature selection (Doctoral dissertation, Applied Sciences:). http://summit.sfu.ca/item/14901

Perez-Rey, D., Anguita, A., & Crespo, J. (2006, December). Ontodataclean: Ontology-based integration and preprocessing of distributed data. In International Symposium on Biological and Medical Data Analysis (pp. 262-272). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11946465_24

Poos, A. M., Maicher, A., Dieckmann, A. K., Oswald, M., Eils, R., Kupiec, M., ... & König, R. (2016). Mixed Integer Linear Programming based machine learning approach identifies regulators of telomerase in yeast. Nucleic Acids Research, 44(10), e93-e93. https://doi.org/10.1093/nar/gkw111

Prajapati, A., & Ganesan, S. (2013). Application of statistical techniques and neural networks in condition-based maintenance. Quality and Reliability Engineering International, 29(3), 439-461. https://doi.org/10.1002/qre.1392

Ragab, A., Ouali, M. S., Yacout, S., & Osman, H. (2016). Remaining useful life prediction using prognostic methodology based on logical analysis of data and Kaplan–Meier estimation. Journal of Intelligent Manufacturing, 27(5), 943-958. https://doi.org/10.1007/s10845-014-0926-3

Ray, S. (2019, February). A quick review of machine learning algorithms. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon) (pp. 35-39). IEEE. https://doi.org/10.1109/COMITCon.2019.8862451

Shukla, S., & Kumar, S. (2019, July). Applicability of neural network based models for software effort estimation. In 2019 IEEE World Congress on Services (SERVICES) (Vol. 2642, pp. 339-342). IEEE. https://doi.org/10.1109/SERVICES.2019.00094

Singh, N., & Singh, P. (2019). Cardiac arrhythmia classification using machine learning techniques. In Engineering Vibration, Communication and Information Processing (pp. 469-480). Springer, Singapore. https://doi.org/10.1007/978-981-13-1642-5_42

Thanh Noi, P., & Kappas, M. (2018). Comparison of random forest, k-nearest neighbor and support vector machine classifiers for land cover classification using Sentinel-2 imagery. Sensors, 18(1), 18. https://doi.org/10.3390/s18010018

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ... & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6), 520-525. https://doi.org/10.1093/bioinformatics/17.6.520

Tsang, A.H., Yeung, W.K., Jardine, A.K. and Leung, B.P. (2006) 'Data management for CBM optimization', Journal of quality in maintenance engineering, Vol. 12 No.1, pp. 37-51. https://doi.org/10.1108/13552510610654529

Velasco-Gallego, C., & Lazakis, I. (2020). Real-time data-driven missing data imputation for short-term sensor data of marine systems. A comparative study. Ocean Engineering, 218, 108261. https://doi.org/10.1016/j.oceaneng.2020.108261

Wisyaldin, M. K., Luciana, G. M., & Pariaman, H. (2020, September). Using LSTM Network to Predict Circulating Water Pump Bearing Condition on Coal Fired Power Plant. In 2020 International Conference on Technology and Policy in Energy and Electric Power (ICT-PEP) (pp. 54-59). IEEE. https://doi.org/10.1109/ICT-PEP50916.2020.9249905

Wu, C. C., Yeh, W. C., Hsu, W. D., Islam, M. M., Nguyen, P. A. A., Poly, T. N., ... & Li, Y. C. J. (2019). Prediction of fatty liver disease using machine learning algorithms. Computer Methods and Programs in Biomedicine, 170, 23-29. https://doi.org/10.1016/j.cmpb.2018.12.032