Applied Computer Science, vol. 19, no. 2, pp. 112–124
doi: 10.35784/acs-2023-18

Submitted: 2023-05-07 | Revised: 2023-06-10 | Accepted: 2023-06-14

Keywords: Predictive Maintenance, Machine Learning, Features Selection, SMOTE-Tomek, Support Vector Machine

Mouna TARIK [0009-0008-1603-0067]*, *Ayoub MNIAI* *, *Khalid JEBARI* *

HYBRID FEATURE SELECTION AND SUPPORT VECTOR MACHINE FRAMEWORK FOR PREDICTING MAINTENANCE FAILURES

Abstract

The main aim of predictive maintenance is to minimize downtime, failure risks and maintenance costs in manufacturing systems. Over the past few years, machine learning methods gained ground with diverse and successful applications in the area of predictive maintenance. This study shows that performing preprocessing techniques such as oversampling and feature selection for failure prediction is promising. For instance, to handle imbalanced data, the SMOTE-Tomek method is used. For feature selection, three different methods can be applied: Recursive Feature Elimination, Random Forest and Variance Threshold. The data considered in this paper for simulation are used in literature. They are used to measure aircraft engine sensors to predict engine failures, while the prediction algorithm used is a Support Vector Machine. The results show that classification accuracy can be significantly boosted by using the preprocessing techniques.

1. INTRODUCTION

Maintenance costs are a major part of the total operating costs of all manufacturing or production plants (Mobley, 2002). U.S. industries spend more than 200$ billion each year on the maintenance of plant equipment which impacts their productivity and profit (Mobley, 2002). In fact, predictive maintenance (PdM) is a leading strategy aiming to improve the productivity, quality and the performance of overall equipment. It allows to schedule maintenance at the most convenient and most cost-efficient moment before that the failure occurs.

Predictive maintenance technologies measure and gather operations and equipment real-time data via sensor networks. It includes non-destructive testing methods such as acoustic, infrared, oil analysis, sound level measurements, vibration analysis, and thermal imaging.

Predictive maintenance uses data science and predictive analytics to estimate when a piece of equipment might fail. Many machine learning (ML) techniques are designed to analyze a large amount of data and can achieve outstanding performance (Wuest, 2016). Machine learning is a powerful tool for predictive analyses in different applications whose performance depends on the appropriate choice of ML techniques (Carvalho et al., 2019).

* LMA, FSTT, Abdelmalek Essaadi University, Tetouan, Morocco, tarik.mouna@gmail.com, ayoubm.m@gmail.com, khalid.jebari@gmail.com

In recent years, many machine learning techniques have been introduced to deal with failure prediction. In their work, Milena Nacchia et al. (Nacchia et al., 2021) presented analyses of the maturity level and the contribution of ML methods for predictive maintenance in smart manufacturing. Chia-Hung Yeh et al. (Yeh et al., 2019) proposed a method based on machine learning to predict the long cycle maintenance time of wind turbines in a power company. A hybrid network was used and reached good prediction results.

Furthermore, the main aim of Emiliano Traini's et al. (Traini et al., 2019) work is to give a general framework that is applicable to cases of predictive maintenance of generic manufacturing tools in order to improve the man-machine collaboration in production. The study is applied to a real milling data set as validation of the framework.

The aim of this paper is to introduce a framework that includes some preprocessing techniques combined with Support Vector Machine (SVM) algorithm.

Data preprocessing in machine learning is a crucial step that helps to enhance the quality of data to promote the extraction of meaningful feature subsets. It refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for building and training machine learning models. Once redundant or irrelevant features are eliminated, it will lead to significant impacts in terms of the performance of the ML methods used. Also, preprocessing techniques involves handling the class imbalance problem by oversampling or undersampling the data, real-world applications such as failures prediction frequently encounter this problem.

In their work, Bekar et al. (Bekar et al., 2020) present an intelligent approach using unsupervised Machine Learning techniques for data preprocessing and analysis in predictive maintenance area. They demonstrate that, through this approach, it is possible to get useful information about component/machine behavior which serve as a foundation for decision support and the development of prognostic models. Fernandes et al. (Fernandes et al., 2019) performed data analysis and feature selection to build models in predictive maintenance within a metallurgical company. The results demonstrated that insights derived from the data will aid in developing adaptive learning models capable of handling complex information which can be effectively deployed across an entire product line of industrial equipment. In a paper by Lai & Leu (Lai & Leu, 2017), they explained that ensuring data preprocessing has become a significant concern issue of big data applications. Also, they proposed the Preprocessing Tasks Quality Measurement (PTQM) model to identify the quality defects of data preprocessing tasks in order to increase the big data applications efficiency and practicality. Abidi & Alkhalefah (Abidi & Alkhalefah, 2022) proposed a PdM planning model using five main phases: (a) data cleaning, (b) data normalization, (c) optimal feature selection, (d) prediction network decision-making, and (e) prediction. They demonstrated that the proposed model can efficiently predict the future condition of components for maintenance.

The highlighting points of this paper are listed below:

- Introduction of a predictive maintenance framework to prevent unexpected failures.
- Application of SMOTETomek method to overcome the problem of unbalanced data.
- Application of feature selection methods to select the most relevant features and optimize the performance of the model.

Evaluate the performance of the model using the SVM algorithm and demonstrate that the use of SVM combined with the above pre-processing techniques leads to better results.

The rest of this article is organized as follows: Section 2 introduces the proposed method and related works. In section 3 the experimental results are presented. Finally, our paper ends with a conclusion and future work in section 4.

2. THE PROPOSED METHOD AND RELATED WORKS

Preprocessing techniques represent a very important part of a data science project, it helps to reduce the dimensions of a dataset and remove the useless variables. In this work, the methodology followed is divided into three phases: (i) The oversampling technique is applied to the unbalanced aircraft engine dataset using SMOTE-Tomek method, (ii) then features selection techniques are used to select the most important features and drop the rest. (iii) The SVM is applied to the balanced dataset and measure the classification using the accuracy (Fig. 1).

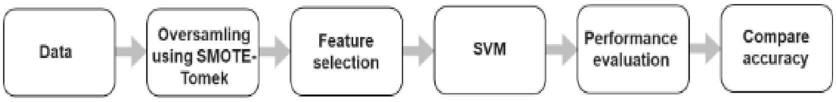

Fig. 1. Research flow

2.1. Imbalanced data and re-sampling

First, a dataset is imbalanced if the classification categories are not approximately equally represented (Nacchia et al., 2021). Preprocessing of data by resampling methods are commonly used to deal with the class-imbalance problem (Estabrooks & Japkowicz, 2004), it is used to upsample or downsample the minority or majority class. For an imbalanced dataset, we can oversample the minority class using replacement. This technique is called oversampling. Similarly, we can randomly delete rows from the majority class to match them with the minority class, and this technique is called undersampling.

A large number of approaches have been used to deal with the class imbalance problem. The most important and widely used methods for oversampling are Random Over-Sampling (ROS) (Rendon et al., 2020), Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) and Adaptive Synthetic Over-Sampling (ADASYN) (He, Garcia & Lee, 2008). Similarly to oversampling, we find different methods for undersampling, such as Random Under-sampling (RUS) (Kotsiantis & Pintelas, 2003), Tomek Links (TL) (Elhassan & Aljurf, 2016), Editing Nearest Neighbor (ENN) (Zhu et al., 2020) along with others.

One of these approaches is SMOTE-Tomek which combines SMOTE (Synthetic Minority Oversampling Technique), the famous oversampling methods and Tomek Links function for undersampling.

SMOTE-Tomek was introduced first by Batista et al. (Batista, Bazzan & Monard, 2003), the standard algorithm flow is as follows:

Step 1: For dataset D with unbalanced data distribution, we use the SMOTE method to obtain an extended dataset D’ by generating many new minority samples (Wang et al., 2019).
Step 2: Tomek Link pairs in dataset D’ are removed using the Tomek Link method (Wang et al., 2019).

The pseudocode of SMOTE-Tomek Links is as follows:

1. (Start of SMOTE) Choose random data from the minority class.
2. Calculate the distance between the random data and its k nearest neighbors.
3. Multiply the difference with a random number between 0 and 1, then add the result to the minority class as a synthetic sample.
4. Repeat step number 2–3 until the desired proportion of minority class is met (end of SMOTE).
5. (Start of Tomek Links) Choose random data from the majority class.
6. If the random data nearest neighbor is the data from the minority class (i.e. create the Tomek Link), then remove the Tomek Link.

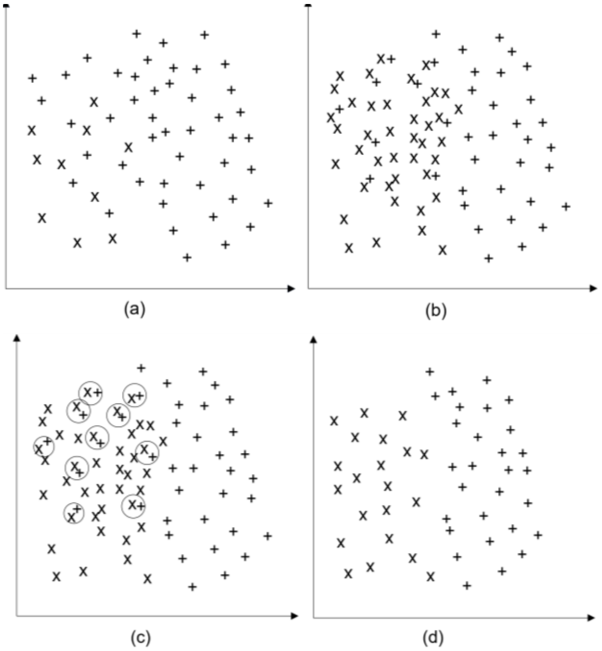

Fig. 2. Balancing a data set using SMOTE-Tomek

The process presented above is shown in Fig. 2. The dataset with new minority class examples created artificially is illustrated in (b). The identification of Tomek links is presented in (c) and the removal of those examples in (d) (Batista, Bazzan & Monard, 2003).

2.2. Feature selection methods

Feature selection (FS) consists on eliminating redudant or irrelevant features that might decrease the model performance (Huang, Li & Xie, 2015).

It is important in fault diagnosis in industrial applications, where numerous redundant sensors monitor the performance of a machine (Jović, Brkić & Bogunović, 2015). Feature selection methods can be classified based on two criteria: The search strategy and the evaluation criterion (Liu & Motoda, 1998). Both criteria typically belong to one of the three classes, determined by the evaluation metric of choice: filter, wrapper, embedded and hybrid methods (Chandrashekar & Sahin, 2014).

Most filter methods calculate a score for all features and then select the features with highest scores (Bommert et al., 2020).Tthey are independent of any learning algorithm. Wrapper methods look for features that are suitable for the machine learning algorithm used, they are evaluated based on the performance of the model (Huljanah et al., 2019). Embedded methods are methods that maintain each iteration of the model training process and extract features that contribute most to training for certain iterations (Huljanah et al., 2019). Hybrid methods are presented as a combination of the above methods.

In the literature, many studies use feature selection to improve the model's performance in the predictive maintenance field. Aremu et al. (Aremu et al., 2020) in their work, presented a feature selection framework, beneficial for predictive maintenance analytics. They proposed a correlation and relative entropy feature engineering framework specific to asset data. A novel and flexible parameterized PdM solution for event/log based equipment was proposed by Wang et al. (Wang et al., 2017). They explained how to optimize the model parameters by selecting the most effective features and tuning classifiers to build a high-performance prediction model.

Various FS methods are used in literature, and some of these methods are applied to our case study in section 3:

- Random Forest (RF): is an embedded method. It is presented as an ensemble of unpruned classification or regression trees (Breiman, 2001). Each individual tree in the random forest spits out a class prediction, and the forest chooses the class with the most votes. RF performs feature selection while a classification rule is built. The two commonly used variable importance measures in RF are Gini importance index and permutation importance index (Hasan, 2016).
- Recursive Feature Elimination (RFE): is a wrapper method, it selects features by iteratively training a set of data with the current set of features and eliminating the least significant feature indicated (Themistocleous, Papadaki & Kamal, 2020). These features are repeatedly eliminated until a certain threshold is met. The RFE ranks features according to some measure of their importance (Granitto, 2006). At each iteration features importance are measured and the less relevant one is removed (Granitto, 2006).
- Variance Threshold: is a filter method. It removes all features whose variance doesn't meet a specific threshold (Themistocleous, Papadaki & Kamal, 2020). The variance value can be calculated using equation (1):

$$\textit{Variance score (f)= p (1-p)} \tag{1}$$

in which p represents the percentage of instances taking the feature value 1. Features with the variance score below the threshold can be deleted immediately. The purpose of this filter is to remove features that have a very little variation or that consist only of noise (Ambarwati & Uyun, 2020).

2.3. Support Vector Machine

A Support Vector Machine (SVM) is a supervised machine learning algorithm introduced by Vapnick (Vapnik, 1999) and can be used for both classification and regression tasks. The objective of SVM is to find an optimal separating hyperplane in an N-dimensional space that correctly classifies the data points (Fig. 3). SVM has been used in a wide variety of applications such as financial fraud detection (Ravisankar et al., 2011), credit rating analysis (Huang et al., 2004), predictive maintenance (Gohel et al., 2020), among others.

In a SVM, the aim is maximizing the margin between the data points and the hyperplane. The data points that lie closest to the separating line between two classes are called: 'Support Vectors'. An SVM can be linear or nonlinear, the linear formulation is the simplest one.

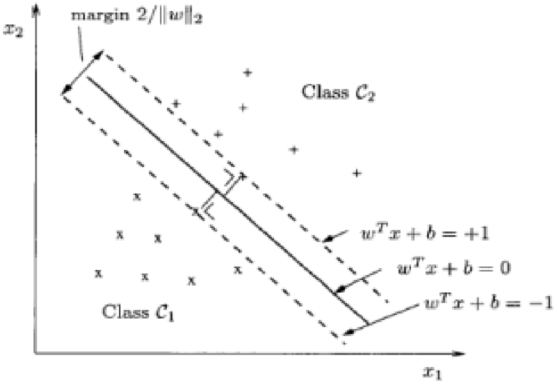

Fig. 3. Definition of a separating hyperplane illustrated in a two-dimensional input space: Linear classification

The SVM problem for the training data set of N points $\{x_i, y_i\}$, *i = 1, . . ., M* is given by:

Minimize:

$$\frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{m} \xi_i^2 \tag{2}$$

Subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi, i = 1, \ldots\ldots, M$$

$$\xi_i \geq 0, \text{ i=1,…,M} \tag{3}$$

Where w is the weight vector, C is the regularization parameter and b is the bias term corresponding to the hyperplane (Singla & Shukla, 2020).

$$f(x_i) = w^T x_i + b \tag{4}$$

ξi is measuring the distance between the margin and the examples x_i that exist on the wrong side of the margin (Tarik & Jebari, 2020).

To solve no linear problems, various SVM kernel functions are used, the most popular ones are: The Linear, Polynomial, Sigmoid and Gaussian Radial Basis function.

The different kernels are mentioned in Table 1.

Tab. 1. SVM Kernel functions

Kernel function	Formula
Linear	$K(x_i, x) = x_i^T x$

Polynomial	$K(x_i, x) = \left(x_i^T x + \theta\right)^d$
RBF	$K(x_i, x) = e^{\frac{-1}{2\sigma^2}\|x - x_i\|^2}$
Sigmoid	$K(x_i, x) = tanh(\eta x x_i + \theta)$

Where *d* is the degree of the kernel function. *θ, σ* and *η* are kernel parameters.

3. EXPERIMENTAL STUDY

3.1. Data

It is crucial that Aircraft Engines should undergo proper maintenance, but it is very expensive as a routine. Hence, airlines are interested to predict engine failures of in-service equipment in order to reduce flight delays and ensure cost savings.

The data set (D) is provided by NASA .The train set consists of run-to-failure data from 100 aircraft engines. A brief description of the data set is presented in table 2:

Tab. 2. Data description

Feature name	Data type	Description
Id	Integer	Values from 1 to 100 that represent each aircraft engine
Cycle	Integer	Number of cycles for each engine Id
Setting 1 to setting 3	Double	Values of operational settings
s1 to s21	Double	Sensor's data values for 21 sensors

The 'Cycle' column in the train set has increasing values for every Id. The last value of 'Cycle' for a particular Engine Id represents the failure of that engine.

In the training set, the engine with id=69 took a maximum number of cycles to fail (i.e., 362 cycles) and the engine with id=39 took minimum number of cycles to fail (i.e., 128 cycles).

The engine is assumed to operate normally at the start of each time series and it starts to degrade at some point during the series of the operating cycles. When a predefined threshold is reached, the engine is considered unsafe for further operation (Tarik & Jebari, 2020).

If the remaining cycles are less than the specified number of cycles (e.g.. period=30), the engine will fail in this period otherwise the engine is fine.

The training data file contains 20630 cycle records with 3100 positive targets and 17530 negative targets as presented in Fig. 4.

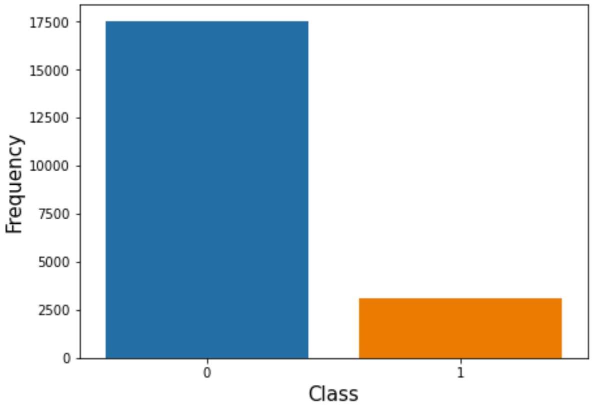

Fig. 4. Target Vs Records

The dataset is imbalanced, that means, the percentage of the normal working records of the engine (healty) is higher than the faulty ones. Thus, the oversampling technique was applied using the SMOTETomek technique. Fig. 5. presents the new values of the two classes.

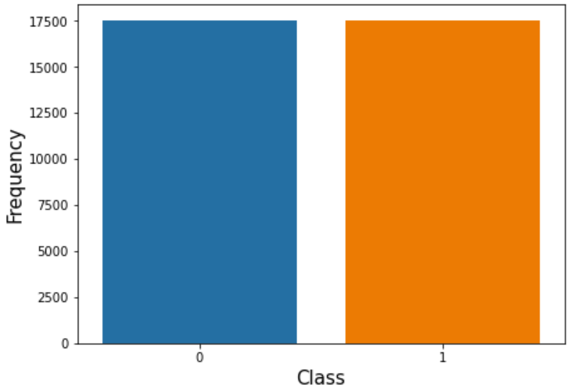

Fig. 5. The records after oversampling technique

All the values in the dataset are numeric, there are no missing values and the data is balanced and not noisy. The dataset is divided into training and testing set (respectively 80% and 20%).

The code was performed using jupyter notebook running on python 3.8 language environment and executed on a core I5-73000 CPU processor.

3.2. Experimental results

For feature selection methods, three algorithms are used: Random Forest, RFE and CFS. Table 3 shows the relevant features selected after performing the three methods.

Tab. 3. Features selected after applying the FS methods

FS method	Features selected
RF	's4', 's7', 's11', 's12', 's15', 's20', 's21'
RFE	's2', 's6', 's7', 's8', 's11', 's12', 's13', 's15', 's20', 's21'
Variance Threshold	‘s2', 's3', 's4', 's7', 's9', 's11', 's12', 's14', 's17', 's20', 's21'

Different metrics are used for performance evaluation and the generalization ability of the trained classifier. In our case, the evaluation metric used is the accuracy known as the ratio of the number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number of correct prediction}{Total number of predictions made} \qquad (5)$$

In addition, the athors also use the SVM under default parameters (C and sigma of the Radial basis function) to classify and predict the data. Table 4 shows the accuracy results before applying FS methods.

Tab. 4. Accuracy results with/without oversampling technique

Model	Accuracy before oversampling	Accuracy with SMOTE-Tomek
SVM	84,49 %	86,81 %

The SVM applied to the data oversampled (D’) gave better accuracy results compared with the initial data set (D). or the next experiment, FS methods were applied to the balanced data set according to the flowchart in Figure 6.:

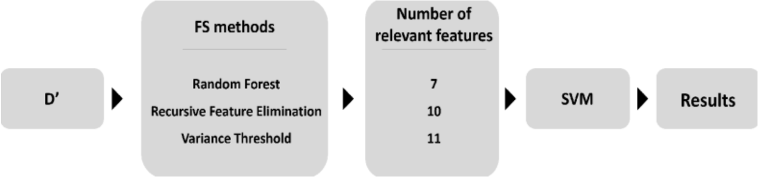

Fig. 6. Research workflow

The performance accuracy was calculated for SVM with three kernel functions: the Radial Basis Function (RBF), the Sigmoid and the Polynomial function. The SVM classification has the best accuracy using the FS methods. The RF-SVM model reaches the highest accuracy 97,77% using the polynomial function. The results are shown in the table 5.

Tab. 5. Accuracy results vs kernels

SVM Kernel	Test Accuracy		
	RF-SVM	RFE-SVM	VT-SVM

RBF	0.9653451226468	0.9396748431260	0.8783513976041
Polynomial	0.9777524244152	0.9556474614945	0.9060182544209
Sigmoid	0.9255561893896	0.8849115801483	0.7954934398174

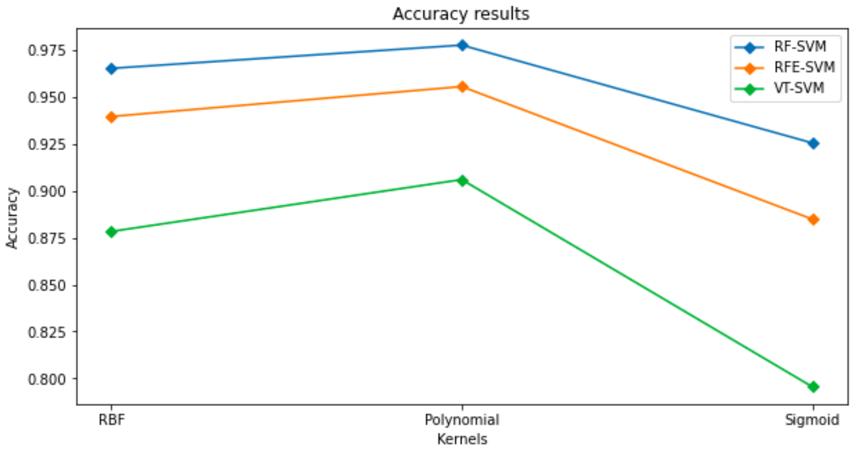

Fig. 7. Accuracy plots for the three models using different kernel functions

The SVM performance for balanced dataset using feature selection methods proves to be better than that with SVM applied for the initial data set. Also, the results of SVM employing polynomial kernel are better than other kernel functions.

Fig. 7. shows that the Random Forest combined with SVM using polynomial function reaches the highest accuracy and the lowest accuracy value is 79% for SVM with Sigmoid kernel.

4. CONCLUSION

Anticipating equipment failure or maintenance needs presents several challenges. Some of these challenges are the following: the selection of the most relevant features from the available data and identification of which sensor data, variables, or features are most indicative of impending failures or maintenance needs.

Moreover, the issue of imbalanced classes can result in biased models, leading to a poor performance in identifying failures.

Overcoming these challenges can lead to more efficient and cost-effective maintenance practices, reduced downtime, and improved equipment reliability. In this context, this study is more relevant to develop solutions for real-world industrial problems, consequently promoting the adoption of machine learning in various manufacturing industries.

Through this work, the main purpose of this article is to perform a systematic assessment of the effectiveness of data preprocessing techniques on ML methods in the area of predictive maintenance. It demonstrates that the data preprocessing techniques may significantly influence the final prediction results.

The paper presents the study of various feature selection algorithms combined with SVM and analyzes their performance for aircraft engine dataset.

Experiments demonstrate that there is a significant performance difference in accuracy using the aircraft engine data and its balanced version using SMOTE-Tomek method. Indeed, the SVM classification accuracy achieved better results using the data oversampled with features selection methods.

In the future, the study can be enhanced by applying hybrid feature selection algorithms combined with metaheuristics to improve the performance of ML algorithms.

Conflicts of Interest

The authors have no conflicts of interest to declare.

REFERENCES

Abidi, M. H., Mohammed, M. K., & Alkhalefah, H. (2022). Predictive maintenance planning for industry 4.0 using machine learning for sustainable manufacturing. *Sustainability*, *14*(6), 3387. https://doi.org/10.3390/su14063387

Ambarwati, Y. S., & Uyun, S. (2020). Feature selection on magelang duck egg candling image using variance threshold method. *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* (pp. 694-699). IEEE. http://doi.org/10.1109/isriti51436.2020.9315486

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*, *145*, 106536. http://doi.org/10.1016/j.cie.2020.106536

Batista, G. E., Bazzan, A. L., & Monard, M. C. (2003). *Balancing training data for automated annotation of keywords: a case study* (pp. 10-18). WOB.

Bekar, E. T., Nyqvist, P., & Skoogh, A. (2020). An intelligent approach for data pre-processing and analysis in predictive maintenance with an industrial case study. *Advances in Mechanical Engineering*, *12*(5), 1687814020919207. https://doi.org/10.1177/1687814020919207

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, *143*, 106839. http://doi.org/10.1016/j.csda.2019.106839

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32. http://doi.org/10.1023/A:1010933404324

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. http://doi.org/10.1016/j.cie.2019.106024

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357. http://doi.org/10.1613/jair.953

Elhassan, T., & Aljurf, M. (2016). Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method. *Global Journal of Technology & Optimization*, *1*, 2016. http://doi.org/10.4172/2229-8711.S1111

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, *20*(1), 18-36. https://doi.org/10.1111/j.0824-7935.2004.t01-1-00228.x

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. https://doi.org/10.1016/j.ijinfomgt.2018.10.006

Gohel, H. A., Upadhyay, H., Lagos, L., Cooper, K., & Sanzetenea, A. (2020). Predictive maintenance architecture development for nuclear infrastructure using machine learning. *Nuclear Engineering and*

Technology, *52*(7), 1436-1442. http://doi.org/10.1016/j.net.2019.12.029

Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and intelligent laboratory systems*, *83*(2), 83-90. http://doi.org/10.1016/j.chemolab.2006.01.007

Hasan, M. A. M., Nasser, M., Ahmad, S., & Molla, K. I. (2016). Feature selection for intrusion detection using random forest. *Journal of information security*, *7*(3), 129-140. http://doi.org/10.4236/jis.2016.73009

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). IEEE. http://doi.org/10.1109/IJCNN.2008.4633969

Huang, J., Li, Y. F., & Xie, M. (2015). An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Information and software Technology*, *67,* 108-127. https://doi.org/10.1016/j.infsof.2015.07.004

Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision support systems*, *37*(4), 543-558. https://doi.org/10.1016/S0167-9236(03)00086-1

Huljanah, M., Rustam, Z., Utama, S., & Siswantining, T. (2019, June). Feature selection using random forest classifier for predicting prostate cancer. *IOP Conference Series: Materials Science and Engineering*, *546*(5), 052031. http://doi.org/10.1088/1757-899X/546/5/052031

Jović, A., Brkić, K., & Bogunović, N. (2015). A review of feature selection methods with applications. In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 1200-1205). IEEE. http://doi.org/10.1109/MIPRO.2015.7160458

Kotsiantis, S. B., & Pintelas, P. E. (2003). Mixture of expert agents for handling imbalanced data sets. *Annals of Mathematics, Computing & Teleinformatics*, *1*(1), 46-55.

Lai, S. T., & Leu, F. Y. (2017). Data preprocessing quality management procedure for improving big data applications efficiency and practicality. In Barolli, L., Xhafa, F., Yim, K. (Eds.) *Advances on Broad-Band Wireless Computing, Communication and Applications: Proceedings of the 11th International Conference On Broad-Band Wireless Computing, Communication and Applications (BWCCA–2016)* (pp. 731-738). Springer. https://doi.org/10.1007/978-3-319-49106-6_73

Liu, H., & Motoda, H. (Eds.). (1998). Feature extraction, construction and selection: A data mining perspective, *Springer Science & Business Media*, *453*. Springer. https://doi.org/10.1007/978-1-4615-5725-8

Mobley, R. K. (2002). *An introduction to predictive maintenance*. Elsevier.

Nacchia, M., Fruggiero, F., Lambiase, A., & Bruton, K. (2021). A systematic mapping of the advancing use of machine learning techniques for predictive maintenance in the manufacturing sector. *Applied Sciences*, *11*(6), 2546. http://doi.org/10.3390/app11062546

Ravisankar, P., Ravi, V., Rao, G. R., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision support systems*, *50*(2), 491-500. http://doi.org/10.1016/j.dss.2010.11.006

Rendon, E., Alejo, R., Castorena, C., Isidro-Ortega, F. J., & Granda-Gutierrez, E. E. (2020). Data sampling methods to deal with the big data multi-class imbalance problem. *Applied Sciences*, *10*(4), 1276. http://doi.org/10.3390/app10041276

Singla, M., & Shukla, K. K. (2020). Robust statistics-based support vector machine and its variants: a survey. *Neural Computing and Applications*, *32*(15), 11173-11194. http://doi.org/10.1007/s00521-019-04627-6

Tarik, M., & Jebari, K. (2020). Maintenance Prediction by Machine Learning: Study Review of Some Supervised Learning Algorithms. *Proceedings of the 2nd African International Conference on Industrial Engineering and Operations Management. Harare* (pp. 2678-2686). Zimbabwe: IEOM Society International.

Themistocleous, M., Papadaki, M., & Kamal, M. M. (Eds.). (2020). Information Systems: *17th European, Mediterranean, and Middle Eastern Conference, EMCIS 2020, Dubai, United Arab Emirates, November 25–26, 2020, Proceedings* , *402*. Springer. http://doi.org/10.1007/978-3-030-63396-7

Traini, E., Bruno, G., D'antonio, G., & Lombardi, F. (2019). Machine learning framework for predictive maintenance in milling. *IFAC-PapersOnLine*, *52*(13), 177-182. http://doi.org/10.1016/j.ifacol.2019.11.172

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, *10*(5), 988-999. http://doi.org/10.1109/72.788640

Wang, J., Li, C., Han, S., Sarkar, S., & Zhou, X. (2017). Predictive maintenance based on event-log analysis: A case study. *IBM Journal of Research and Development*, *61*(1), 11-121.

http://doi.org/10.1147/jrd.2017.2648298

Wang, Z. H. E., Wu, C., Zheng, K., Niu, X., & Wang, X. (2019). SMOTETomek-based resampling for personality recognition. *IEEE Access*, *7*, 129678-129689. http://doi.org/10.1109/ACCESS.2019.2940061

Wuest, T., Weimer, D., Irgens, C., & Thoben, K. D. (2016). Machine learning in manufacturing: advantages, challenges, and applications. *Production & Manufacturing Research*,*4*(1), 23-45. doi.org/10.1080/21693277.2016.1192517

Yeh, C. H., Lin, M. H., Lin, C. H., Yu, C. E., & Chen, M. J. (2019). Machine learning for long cycle maintenance prediction of wind turbine. *Sensors*, *19*(7), 1671. http://doi.org/10.3390/s19071671

Zhu, Y., Jia, C., Li, F., & Song, J. (2020). Inspector: a lysine succinylation predictor based on edited nearest-neighbor undersampling and adaptive synthetic oversampling. *Analytical biochemistry*, *593*, 113592. http://doi.org/10.1016/j.ab.2020.113592