XAI for Machine Learning in PredictiveMaintenance

Harish Chandran Manohar, *MSC Data Science*
Faculty of Engineering Environment and Computing
Coventry University
Coventry, England
manoharh@coventy.ac.uk

***Abstract*—This paper uses various Machine Learning and Explainable Artificial Intelligence (XAI) methodologies to identify the binary classification of machine failures, which is critical in reducing downtime for industrial machines. This work implements and contrasts three classification methods: logistic regression, gradient tree classifier, and random forest. During the pre-processing stage, automatic outlier detection and oversampling analysis are performed, and the results are compared to determine the model's optimal performance. After hyperparameter tuning using the grid search method, the Random Forest model has the greatest average accuracy of 99% among the evaluated models. The best F1 score of 99.33% demonstrates a flawless combination of precision and recall measures. Moreover, adopting the Explainable AI tools like SHAP and LIME, improved the model interpretability to explain how the black box models predict the machine failures by providing the global and local explanations.**

Keywords—Explainable AI, Predictive Maintenance, Industry 4.0, imbalanced data, automatic outlier detection, classification, XAI

I. Introduction

Machine Learning (ML) is widely applied in Predictive Maintenance to identify any abnormality and deviation from standard performance and can even predict the failure of equipment [1]. Yet, the question is that these AI models are not transparent, and therefore, are called 'black box AI' which poses a question to these predictions [2]. In response, Explainable AI (XAI) techniques such as LIME, and SHAP are designed to deliver human-interpretable information concerning the AI model predictions [1].

[3] used Random Forest on AI4I 2020 Predictive Maintenance dataset for its Tree-based explainability and attained a baseline accuracy of 98.34%. [4] emphasizes that based on the characteristics of the data, both over-sampling and under-sampling, can significantly improve the classification success in predictive maintenance tasks with imbalanced data. However, there is a low Recall score for all the classification algorithms used in [4]. The Gradient Boosted Tree model used in [5] achieved an accuracy of 97.3%, outperforming other algorithms such as random forests (RF) and support vector machines (SVM). [6] proposed an ensemble model of several tree-based classifiers on AI4I 2020 dataset and achieved an accuracy of 98.93%.

II. DATASET

A. Dataset Description

The dataset chosen for experimentation is the AI4I 2020 Predictive Maintenance Dataset, an industrial dataset publicly released in 2020 by [3] during the International Conference on Artificial Intelligence. The dataset contains 10000 records and 14 attributes as shown in Table 1.

Table 1

Feature	Role	Type	Description
product type	Feature	Categorical	Identifier for product quality variant and serial number. Quality variants: L (50%), M (30%), H (20%).
air temperature [K]	Feature	Float	Air temperature during the process, normalized to a standard deviation of 2 K around 300 K.
process temperature [K]	Feature	Float	Process temperature, calculated by adding 10 K to the air temperature, normalized to a standard deviation of 1 K.
rotational speed [rpm]	Feature	Float	Rotational speed of the machine, calculated from a power of 2860 W with overlaid normally distributed noise.
torque [Nm]	Feature	Float	Torque values normally distributed around 40 Nm with a standard deviation of 10 Nm. Negative values are excluded.
tool wear [min]	Feature	Integer	Duration of tool wear. The quality variants add different durations: H adds 5 minutes, M adds 3 minutes, and L adds 2 minutes.
machine failure	Target	Binary (0 or 1)	Label indicating whether a machine failure occurred during the data point. Failure modes include Tool Wear Failure (TWF), Heat Dissipation Failure (HDF), Power Failure (PWF), Overstrain Failure (OSF), and Random Failure (RNF).

B. Problems in the Dataset

The main focus of this paper will be to perfect binary classification with oversampling and explain the model's behaviour with XAI. From Figure 1, the major outliers are from the minority class '1' for features such as Rotational Speed, Torque, and Process Temperature. The validity of

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

using those outliers for the model training and their elimination using IQR method will be challenged when such points are removed. A broader perspective of such typical outcasts is required.

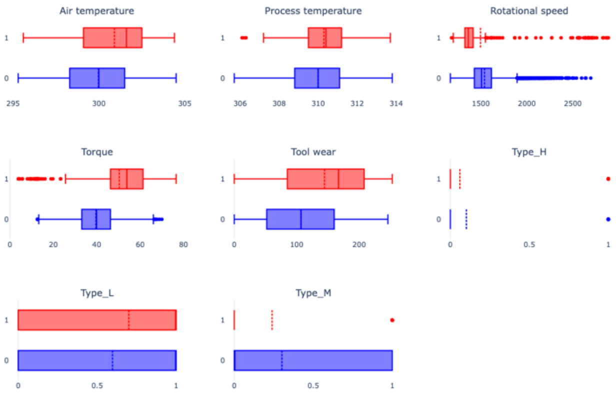
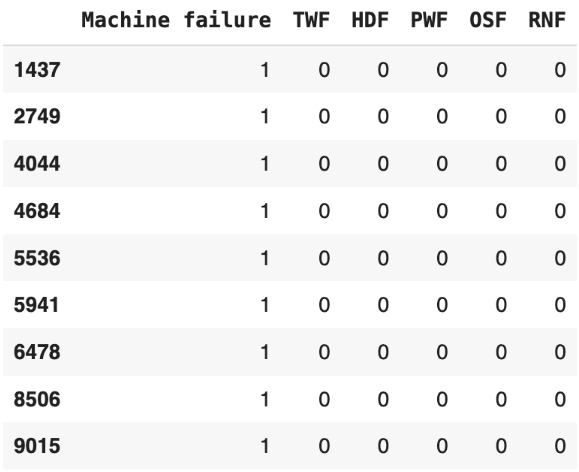


Figure 1

Machine failures with no identified failure mode imply that the failure is a true anomaly or an undiscovered random failure mod. In a classification task, these anomalies can be seen as outliers if not handled appropriately with additional prior knowledge. These machine failure records, as illustrated in Figure 2, do not contribute to meaningful dataset preparation for machine learning.

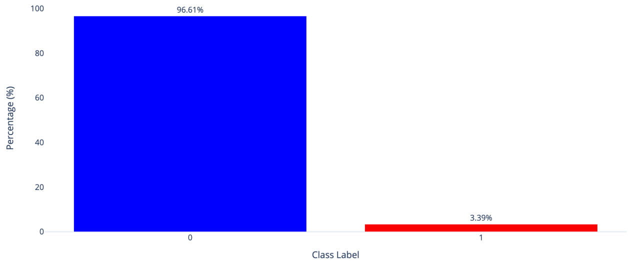


	Machine failure	TWF	HDF	PWF	OSF	RNF
1437	1	0	0	0	0	0
2749	1	0	0	0	0	0
4044	1	0	0	0	0	0
4684	1	0	0	0	0	0
5536	1	0	0	0	0	0
5941	1	0	0	0	0	0
6478	1	0	0	0	0	0
8506	1	0	0	0	0	0
9015	1	0	0	0	0	0

Figure 2

The class distribution demonstrated in Figure 3 indicates that there are 96. 61% of machines without failure and 3. 39% with failure. This imbalance induces the model to learn more from the majority class than the minority; raising major concerns when training the model to work effectively without being influenced by the heuristic nature of the model.

Figure 3



From the data distribution shown in Figure 4, features are normally distributed, but a few features such as Rotational Speed, Process Temperature, and Air Temperature have high variance. The maximum values of those features affect the model, making it assign more importance to them.

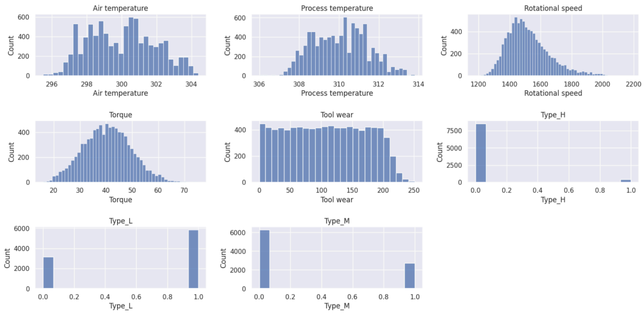


Figure 4

III. Methods

These are the approaches or algorithms employed in this paper, which include the methods for data pre-processing, classification, and model interpretability for developing effective and reliable PM solutions.

A. Automatic Outlier Detection

These are the four outlier detection algorithms used:

- Local Outlier Factor finds outliers by calculating the density difference between a data point and the points in the locality. The method is appropriate for identifying local shifts in high-density data areas.
- Isolation Forest isolates observations by picking a feature randomly and the number of splits needed to separate a data point from the rest is used to calculate the anomaly score of the data point.
- One-class SVM is an unsupervised learning algorithm that aims to find the decision boundary for separating new data from the training data.
- Elliptic Envelope first brings the dataset into a Gaussian distribution and then uses the Mahalanobis distance to find the outliers. This method is most appropriate when dealing with datasets that have a normal distribution because it does an excellent job of capturing the underlying structure and identifying outliers.

B. Over-sampling

To handle the problem of class imbalance that hinders the model from focusing on the minority class, oversampling methods were applied to balance the proportions of the classes in the training set.

- SMOTE creates synthetic samples for the minority class by computing the distance between the samples of the minority class and selecting the nearest neighbours to create new samples. This is useful in ensuring that the numbers of classes are balanced and prevents the model from being inclined to the majority class

- ADASYN is an approach that aims at creating synthetic samples that belong to the minority class, particularly in the areas of the feature space that lacks sufficient samples of the minority class.

- SMOTE combined with Tomek links, not only creates synthetic samples but also eliminates the cases that are difficult to classify. This combination enhances the decision boundary by applying the technique of oversampling on the minority class and removing noisy samples from the dataset.

- SMOTE is used together with ENN, by which the minority class is oversampled, while the data set is cleaned from instances that were misclassified by their neighbours. This leads to generation of a cleaner and more balanced data set which helps in increasing the performance of the model.

C. *Machine Learning Algorithms*

- Logistic Regression: A statistical model that maps one or more predictor variables to a binary response variable using a linear equation and a logistic function. Logistic Regression is easy to implement, easy to interpret and works best for binary classification problems.

- Gradient Booster: It is an ensemble learning technique in which models are created one at a time and each is designed to address mistakes of the former.

- Random Forest: It is an ensemble method for decision trees that combines the results from a set of trees to ensure accuracy and reliability. This technique helps to minimize over-fitting and brings better generalization to the model.

D. *Explainable AI Methods*

- Shapely additive explanation (SHAP) is a model-agnostic feature importance interpretation technique that assigns values to each feature based on Shapley values by delivering insights into the feature impact on the trained model. SHAP values show how relevant each feature is for the outcome and provide global feature importance and local explanation for each record.

- Local Interpretable Model-agnostic Explanations (LIME) aims to fit the local region of the black box model with an interpretable model. LIME uses random selective perturbation by generating perturbed samples around the instance data point and overlooking the generalized feature impact of different samples. In contrast to other models, LIME gives immediate and simple interpretations of individual predictions, making it appropriate for real-time use

IV. Experimental Setup

The model development process for this paper is depicted in the Figure 5 below:

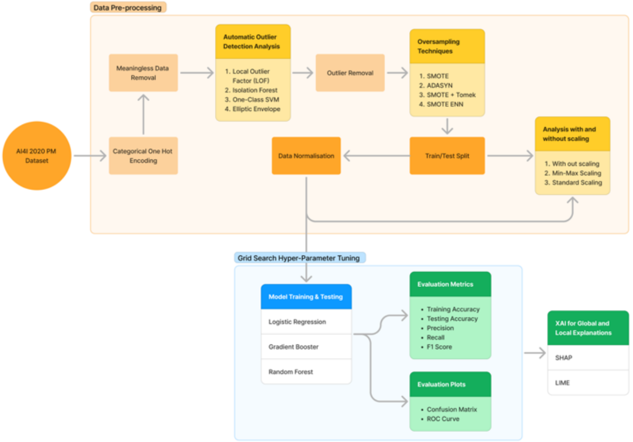


Figure 5

A. *Data Pre-processing*

For categorical data, one hot encoding is preferred over ordinal/integer encoding as integer encoding introduces a natural order relationship between categorical variables which is not true. Since categorical encoding, there will be 8 features for the machine learning algorithm to train with. In addition, the meaningless data mentioned in the problem section are also removed. Next, outlier analysis is performed on 4 different automatic outlier detections such as Local outlier factor, isolation forest, one class SVM, and Elliptic envelope, because the LQR approach can neglect the outliers from the minority class which might be very important in the model training phase. To eliminate the problem of class imbalance, four different oversampling techniques such as ADASYN, SMOTE, SMOTE + Tomek and SMOTE ENN are used to train the three classification algorithms mentioned in this paper.

B. *Model Development and Tuning*

After the data cleaning, the dataset is split into 80% for the training set and 20% for the testing set, respectively. To explore the different scenarios, an initial model is trained without applying any normalisation. This initial model is compared with the other data scaling methods, such as min-max and standard scaler. The best scaler will be selected for training all three classifier models, and grid search is used in the hyper-parameter tuning step to find the optimal parameters of each model to maximise the model's performance in all aspects. Explainable approaches are applied to the best-performing model to obtain global and local explanations.

V. Results

Out of the four automatic outlier detection methods, the Elliptic Envelope performed well on this dataset shown in Figure 6 with a line indicator, scoring an accuracy of 99.09%. After outlier removal, the data records were reduced from 10000 to 8990 records.

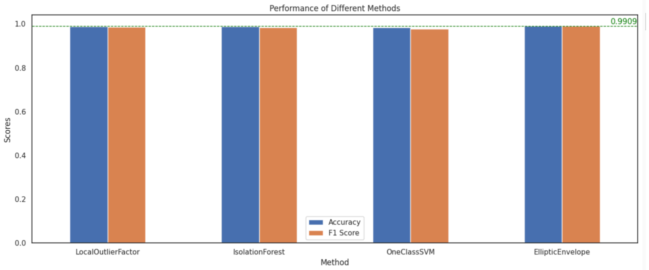


Figure 6

While looking at the Over-sampling analysis graph from Figure 7, there is no notable comparison. However, after training the 4 different oversampled data, it is noted that SMOTE-ENN performs well.

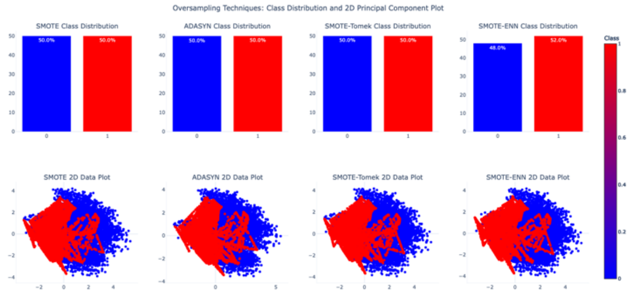


Figure 7

After applying Grid Search for the three machine-learning classification models chosen in this paper, the chosen best hyperparameters are given below for those respective models:

A. Logistic Regression (LR):

{'C': 0.01, 'max_iter': 100, 'penalty': 'l2', 'solver': 'lbfgs'}

B. Gradient Boosting (GB):

{'learning_rate': 0.2, 'max_depth': 5, 'n_estimators': 200}

C. Random Forest (RF):

{'bootstrap': False, 'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}

After the tuning process the final performance of the three classification models is as follows:

Table 2

Metrics	LR	GB	RF
Training Accuracy	0.8824	0.9972	0.9988
Testing Accuracy	0.8763	0.9924	0.9933
Precision	0.8700	0.9874	0.9902
Recall	0.8981	0.9982	0.9971
F1-Score	0.8838	0.9928	0.9936

From the confusion matrix depicted in Figure 8, Random Forest shows an improvement by having fewer False Negatives than the other two algorithms.

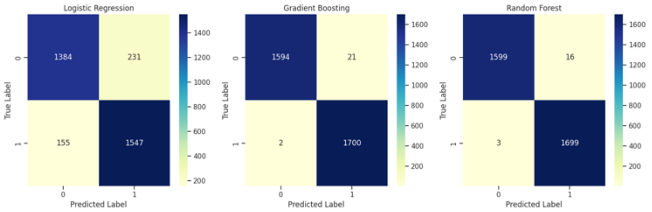


Figure 8

Also, ROC Curve from Figure 9 shows that Random Forest and Gradient Boosting performs well on this dataset with an AUC score of 1.

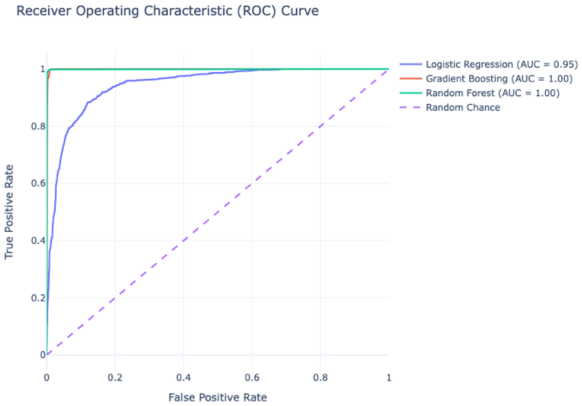


Figure 9

The feature importance of the model produced by SHAP is shown in Figure 10. Model developers such as data scientists or machine learning engineers can utilize SHAP to understand and retrain the model with the most appropriate feature impact. Using SHAP is a bit memory-intensive, and it took 7 minutes in approx. in MacBook Air M2.

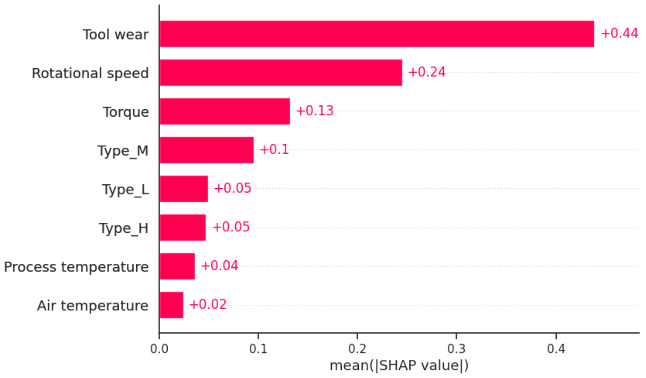
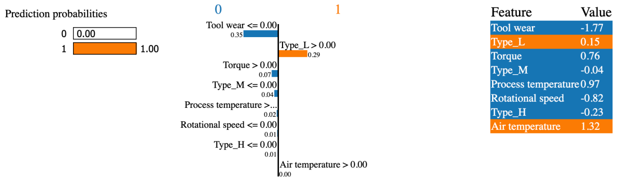


Figure 10

On the contrary, LIME is suitable for real-time use cases, and Figure 11 represents the quick local interpretation of the prediction made on new instances.



Feature	Value
Tool wear	-1.77
Type_L	0.15
Torque	0.76
Type_M	-0.04
Process temperature	0.97
Rotational speed	-0.82
Type_H	-0.23
Air temperature	1.32

Figure 11

VI. Ethical Considerations

When applying machine learning to predictive maintenance, ethical considerations are critical. Making sure the models are fair is critical for avoiding predicted bias, especially when used in a variety of industrial contexts. The training data should be representative and neutral so that no particular equipment or situation is unfairly disadvantaged. From a privacy perspective, protecting sensitive operational data is critical, necessitating compliance with relevant legislation such as the CCPA and GDPR. It's also important that the model's operation is transparent. Explainable AI (XAI) solutions such as SHAP and LIME guarantee that stakeholders understand how predictions are made, hence increasing the credibility and responsibility of predictive maintenance systems.

VII. Discussion and Conclusion

Oversampling is useful in balancing the model heuristics on an imbalanced dataset, as suggested by [4]. Furthermore, grid search proved to derive the best hyper-parameters. In addition to traditional data science steps, XAI techniques add more transparency to the developed model, which should be followed as a necessary step after model training to understand and fine-tune the model further. XAI techniques explored in this paper are good for model interpretability and promote trust for a healthy human-AI future. However, SHAP is computationally expensive, and LIME's selective perturbation is unstable for the same instance when triggered multiple times.

Thus, this paper showcases how ML and XAI methods can be used to enhance the prediction of machine failures. As a result of the high performance and the possibility of using SHAP and LIME for explainability, the best explainable model for Predictive Maintenance from the chosen dataset is the Random Forest Classifier.

VIII. Appendix

- AI4I 2020 Predictive Maintenance Dataset Link: https://archive.ics.uci.edu/dataset/601/ai4i+2020+predictive+maintenance+dataset
- Code Link: https://github.com/being-invincible/BinaryClassificationPM

References

[1] J. Gama, S. Nowaczyk, S. Pashami, R. P. Ribeiro, G. J. Nalepa, and B. Veloso, "XAI for Predictive Maintenance," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2023, pp. 5798–5799. doi: 10.1145/3580305.3599578.

[2] S. Gawde, S. Patil, S. Kumar, P. Kamat, and K. Kotecha, "An explainable predictive maintenance strategy for multi-fault diagnosis of rotating machines using multi-sensor data fusion," *Decision Analytics Journal*, vol. 10, Mar. 2024, doi: 10.1016/j.dajour.2024.100425.

[3] S. Matzka, "Explainable Artificial Intelligence for Predictive Maintenance Applications," in *Proceedings - 2020 3rd International Conference on Artificial Intelligence for Industries, AI4I 2020*, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 69–74. doi: 10.1109/AI4I49448.2020.00023.

[4] S. Cicak and U. Avci, "Handling Imbalanced Data in Predictive Maintenance: A Resampling-Based Approach," in *HORA 2023 - 2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/HORA58378.2023.10156799.

[5] M. Marzuki, S. Prayogi, and M. Abdillah, "Data-Driven Based Model For Predictive Maintenance Applications In Industrial System," European Alliance for Innovation n.o., Dec. 2023. doi: 10.4108/eai.23-11-2022.2341596.

[6] P. Sengupta, A. Mehta, and P. S. Rana, "Predictive Maintenance of Armoured Vehicles using Machine Learning Approaches," Jul. 2023, [Online]. Available: http://arxiv.org/abs/2307.14453