IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTIVE MAINTENANCE USING ENSEMBLE LEARNING AND RARE EVENT CLASSIFICATION ON INDUSTRIAL IOT DATA

Rajasekhar Kalla

HCL Technologies Ltd.

Bangalore, India

ABSTRACT

The rapid advancement of Industrial Internet of Things (IIoT) technologies and sensor-enabled smart manufacturing systems has significantly transformed modern industrial operations. Continuous telemetry data generated from industrial sensors captures critical operational parameters at regular time intervals, representing both normal machine behavior and abnormal failure conditions. Machine failure events are inherently rare in industrial environments, resulting in highly imbalanced datasets that present substantial challenges for conventional classification algorithms.

This study proposes a predictive maintenance framework for early machine failure detection using historical multivariate sensor data. The primary objective is to develop a robust classification model capable of predicting machine failures approximately 24 hours prior to occurrence. In addition to predictive modeling, the dataset also facilitates comprehensive multivariate time-series analysis for industrial process monitoring and anomaly detection.

To address the challenges associated with extreme class imbalance, several imbalance-aware machine learning approaches are investigated, including ensemble-based classification techniques and resampling strategies. Furthermore, a hybrid classification methodology is proposed to enhance rare-event prediction performance and improve failure detection capability in industrial IoT environments.

Keywords: Predictive Maintenance, Industrial Internet of Things (IIoT), Rare Event Classification, Extreme Class Imbalance, Ensemble Learning, Machine Failure Prediction, Multivariate Time Series Analysis, XGBoost, Balanced Random Forest, Easy Ensemble Classifier, Industrial Analytics, Fault Detection, Condition Monitoring, Machine Learning, Sensor Data Analytics.

1. INTRODUCTION

Unexpected machine failures in industrial manufacturing environments can result in substantial production losses, increased operational expenditure, unplanned downtime, and inefficient utilization of manpower and maintenance resources. To minimize such disruptions and improve operational reliability, modern industries are increasingly adopting Industrial Internet of Things (IIoT) and Artificial Intelligence (AI)-driven predictive maintenance solutions. These intelligent monitoring systems continuously analyze real-time machine health conditions and generate early warning alerts for potential equipment failures before actual breakdown occurs.

In industrial IoT ecosystems, large volumes of operational data are continuously generated through embedded sensors installed on manufacturing equipment and critical industrial assets. These sensors monitor various machine health indicators such as pressure, voltage, vibration, and rotational speed at regular time intervals. The accumulated sensor observations form multivariate time-series telemetry data that can be utilized for machine condition monitoring, anomaly detection, and predictive maintenance applications.

In this study, a similar industrial machine dataset is utilized in which sensor measurements corresponding to pressure, voltage, vibration, and rotational speed are recorded across multiple timestamps for different machine identifiers. The complete dataset consists of five relational data tables containing telemetry measurements, machine information, maintenance records, machine failures, and operational error logs. Among these, three primary datasets are selected for data preprocessing, exploratory analysis, feature engineering, and predictive modeling. The dataset used in this work is publicly available through the Kaggle predictive maintenance repository [9]. A detailed description of the utilized data tables is presented in the subsequent sections.

Table 1:TELEMETRY DATA TABLE

Column name	*Column Information*
Datetime	Date and time information
Machine ID’s	Number from 1 to 100 (machine ID)
Pressure, Vibration, Voltage, Rotation	Pressure, rotation, voltage, and vibration measured by sensors

Table 2: Failure table

Column name	Column Information
Datetime	Date and time at which failure occurred
Machine ID’s	Number from 1 to 100 (machine ID)
Failure	Comp1, comp2, Comp3, comp4 kind of failures.

Table 3: Error Table

Column name	Column Information
Datetime	Date and time at which error occurred
Machine ID’s	Number from 1 to 100 (machine ID)
Error	Errors with its ID’s

The primary objective of this research is to develop a predictive maintenance framework capable of forecasting machine failures approximately 24 hours prior to their occurrence using historical industrial telemetry data. In addition to the selected datasets utilized for analysis and predictive modeling, the original repository also contains maintenance records and machine metadata tables, which are not considered in the current study.

2. EXPLORATORY DATA ANALYSIS AND DATA PREPROCESSING

The dataset used in this work consists of multivariate time-series telemetry data containing continuous measurements of machine operating parameters such as voltage, pressure, rotational speed, and vibration. Exploratory Data Analysis (EDA) was performed to investigate the statistical distribution of these sensor variables and to analyze their relationship with machine failures and operational errors. Understanding the behavior and variability of these parameters is essential for identifying abnormal operational patterns and failure signatures within industrial systems.

To facilitate predictive modeling, the selected datasets were integrated into a consolidated analytical dataset through multiple preprocessing and feature engineering stages. The preprocessing workflow consists of the following steps:

• Initially, the telemetry dataset was merged with the machine error log dataset using timestamp-based alignment. An additional feature was generated to capture the cumulative number of machine errors occurring within the previous 24-hour time window.

• Subsequently, the machine failure dataset was integrated with the extracted telemetry-error dataset at corresponding timestamps. Observations falling within the 24-hour pre-failure interval were assigned the respective failure class labels, whereas all remaining observations were labeled as “No Failure.”

• Temporal features were further extracted from the datetime attribute, including day, month, year, and hour components, to capture time-dependent operational behavior.

During preprocessing, missing values, inconsistent records, and other data quality discrepancies were identified and appropriately handled to improve dataset reliability and model robustness.

The resulting multivariate time-series dataset captures the dynamic variations of pressure, voltage, vibration, and rotational speed across different operational intervals, as illustrated in Fig. 1. Furthermore, the probability distributions of key telemetry variables, including pressure, vibration, and voltage, are presented in Fig. 2 to analyze their statistical characteristics and operational variability.

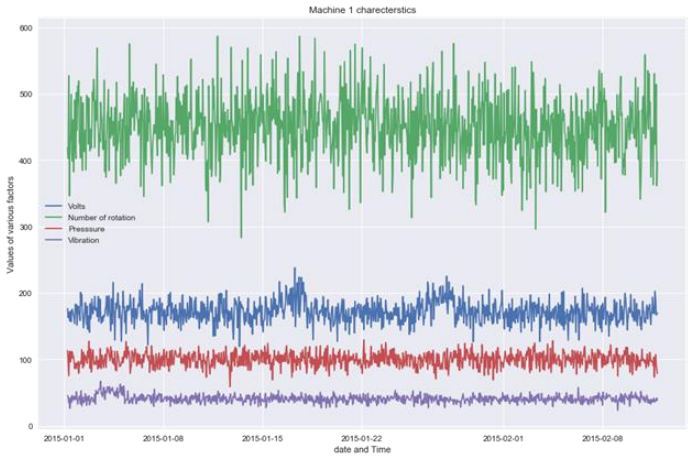


Figure 1: Visualization of telemetry data for machine ID 1

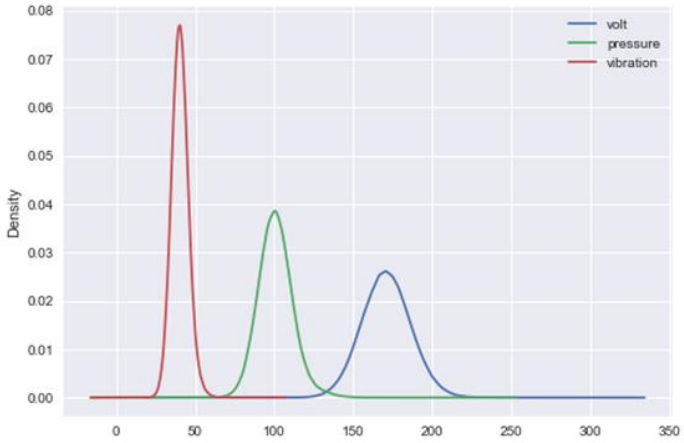


Figure 2: Probability distribution of Pressure, vibration, and voltage

The failure attribute in the dataset contains multiple machine failure categories, including Comp1, Comp2, Comp3, Comp4, and the dominant "No Failure" operational class. Analysis of the class distribution reveals the presence of severe class imbalance within the dataset. The imbalance characteristics are illustrated through the pie chart presented in Fig. 3, while the corresponding bar chart further highlights the disproportionate distribution of failure classes.

Approximately 97.33% of the total observations belong to the "No Failure" category, whereas only nearly 2.66% of the observations correspond to actual machine failure events distributed across different component failure classes. Such an extreme imbalance significantly affects the learning capability of conventional classification algorithms, as predictive models tend to become biased toward the majority class while underperforming on minority failure classes.

This imbalance condition necessitates the use of specialized imbalance-aware machine learning techniques, ensemble learning strategies, and resampling approaches to improve minority-class prediction performance and enhance the reliability of rare-event classification in predictive maintenance systems.

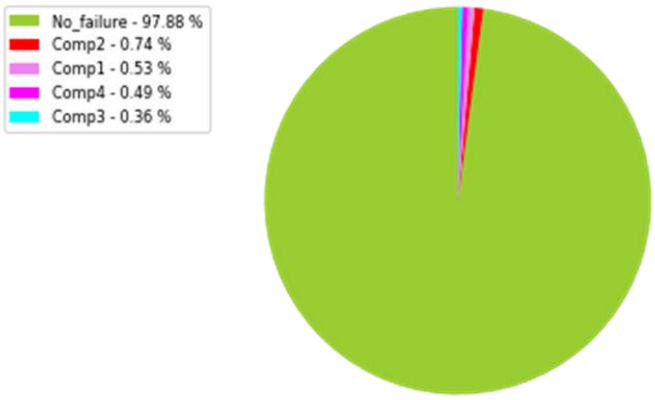


Figure 3: Pie chart for percentage instances of each class in the dataset

From the above observations, it is evident that the dataset exhibits extreme class imbalance, where failure events represent only a very small proportion of the total observations. In the following sections, various machine learning approaches for handling imbalanced classification problems are discussed, starting from conventional methods such as Logistic Regression and extending toward advanced ensemble-based techniques. Different over-sampling and under-sampling strategies are also explored to improve minority-class prediction performance. Furthermore, a hybrid two-step classification framework is proposed to enhance predictive accuracy and overall classification performance for rare-event machine failure prediction.

3. Available methods and their results

In this section, various machine learning approaches for predictive maintenance and rare-event classification are comparatively evaluated. The analysis includes conventional statistical models such as Logistic Regression along with advanced ensemble-based techniques including XGBoost, Balanced Random Forest, and Easy Ensemble classifiers for handling extreme class imbalance in industrial datasets.

3.1. Logistic Regression

The Logistic Regression model [1] was implemented using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) optimization algorithm, which is widely used for parameter estimation in machine learning applications [2]. Since the problem involves multiclass classification, the model was trained using the One-vs-Rest (OvR) strategy instead of multinomial logistic regression.
The performance of the Logistic Regression model is presented in Fig. 3 and Fig. 4. Although the model achieved a high overall accuracy of 97.34%, the ROC curve indicates poor minority-class prediction capability and behavior close to random classification. This observation demonstrates that accuracy alone is not a reliable evaluation metric for highly imbalanced datasets, and additional metrics such as ROC analysis are essential for assessing model performance in rare-event classification problems.

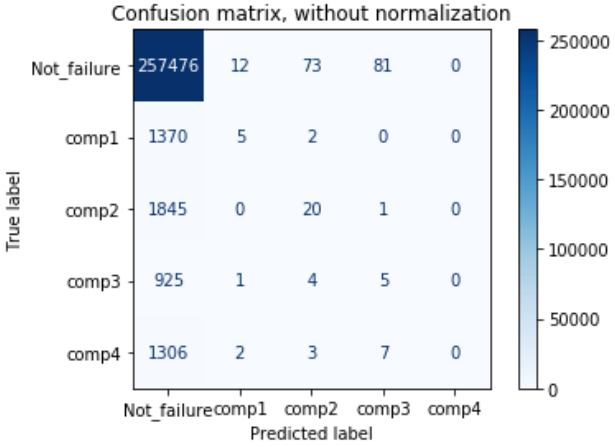


Figure 4: Confusion matrix for losgistic regression

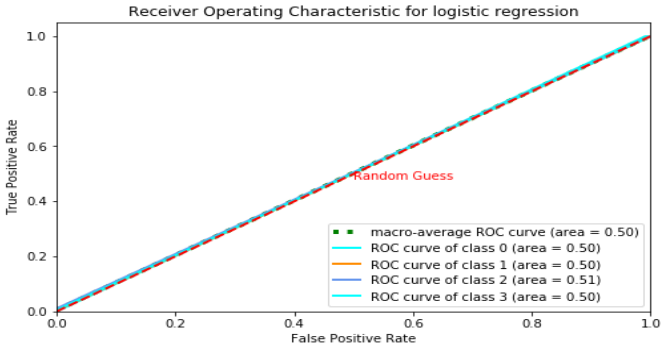


Figure 5: ROC curve for logistic regression

3.2. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an advanced ensemble learning algorithm based on gradient boosting decision trees and is widely recognized for its computational efficiency, scalability, and predictive performance in machine learning applications [3]. XGBoost sequentially constructs multiple weak learners, where each subsequent model attempts to minimize the errors generated by the previous learners. In addition, the algorithm incorporates regularization techniques to reduce overfitting and improve model generalization. XGBoost is particularly effective for handling complex nonlinear relationships and high-dimensional industrial telemetry data. In this study, the model was applied to the highly imbalanced predictive maintenance dataset for machine failure classification. The experimental results demonstrate significant improvement over

the Logistic Regression model in terms of ROC characteristics and minority-class prediction capability. Although the overall accuracy decreased to 83.41%, the confusion matrix and ROC analysis indicate considerably better rare-event detection performance, highlighting the effectiveness of XGBoost for imbalanced industrial classification problems.

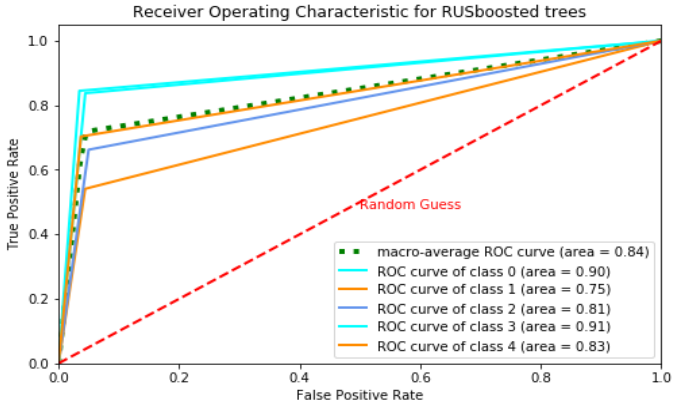


Figure 6: ROC curve for XGBoost (Improved ROC curve)

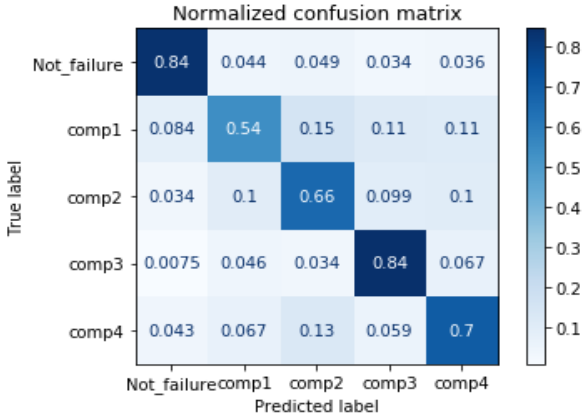


Figure 7: Normalized Confusion matrix for XGBoost Trees

3.3. Balanced Random Forest Classifier

The Balanced Random Forest classifier addresses the class imbalance problem by constructing balanced bootstrap samples during each iteration of the Random Forest training process [5]. In this approach, a bootstrap sample is first generated from the minority class, followed by random sampling of an equal number of observations from the majority class with replacement. This balanced sampling strategy improves the learning capability of the classifier for minority failure classes.

The experimental results presented in Fig. 7 and Fig. 8 demonstrate that the Balanced Random Forest classifier achieves performance comparable to the XGBoost model while significantly outperforming the Logistic Regression approach. The model achieved an overall classification accuracy of 84.27% along with improved minority-class prediction performance and better classification robustness for imbalanced industrial datasets.

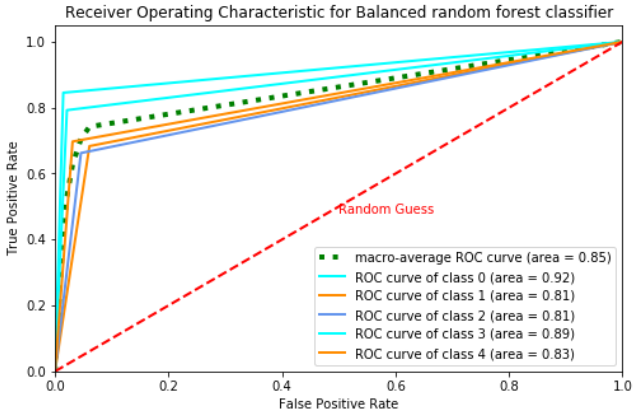


Figure 8: ROC curve for balanced random forest classifier

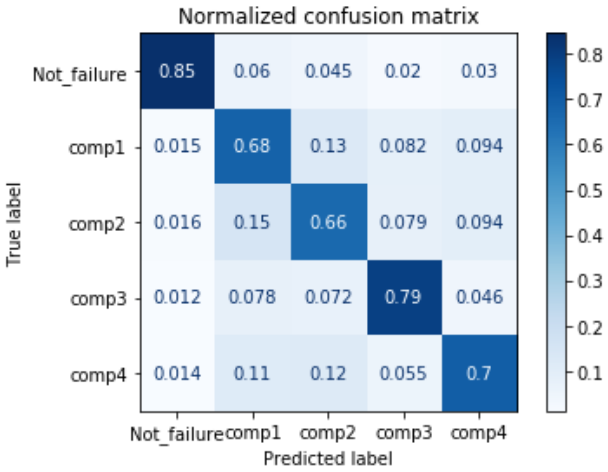


Figure 9: Normalized Confusion matrix for balanced random forest classifier

3.4. Easy Ensemble Classifier

The Easy Ensemble classifier addresses class imbalance by generating multiple balanced subsets from the majority class and training individual learners on each subset independently [7]. In this approach, AdaBoost is used as the base learner for each sampled subset, and the outputs from all weak classifiers are combined to construct the final ensemble model.

The experimental results obtained from the Easy Ensemble classifier are presented in Fig. 9 and Fig. 10. The model achieved an overall classification accuracy of 83.24%. Although the performance is comparable to the previous ensemble-based techniques, significant improvement over the Balanced Random Forest and XGBoost models was not observed. However, the method demonstrated stable minority-class prediction capability for the imbalanced predictive maintenance dataset.

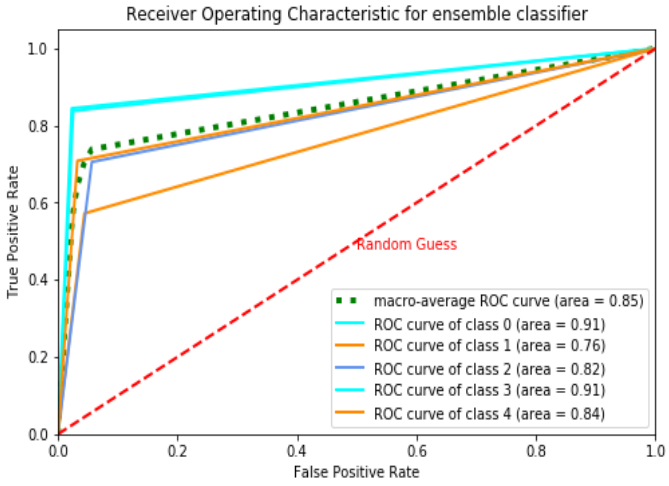
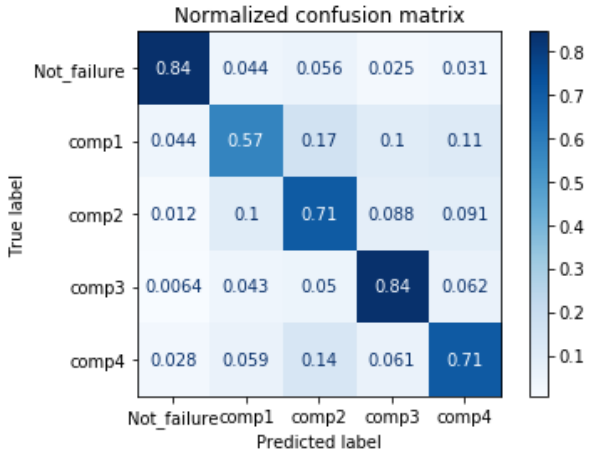


Figure 10: ROC curve for Easy ensemble classifier

The Easy Ensemble classifier is computationally efficient during the training phase and provides effective multiclass resampling by independently generating balanced subsets for each class. This independent sampling strategy improves the classifier's ability to learn minority-class patterns while maintaining training stability for highly imbalanced datasets.



4. HYBRID TWO-STEP MODEL FOR EXTREME CLASS IMBALANCE

Since the previously discussed models were unable to provide satisfactory performance for highly imbalanced failure data, a hybrid two-step classification framework is proposed. The proposed approach addresses the extreme class imbalance problem by performing binary classification in the first stage followed by multiclass failure classification in the second stage, similar to the methodology discussed in [8].

To implement this framework, the original dataset was transformed by introducing a binary target column containing “Failure” and “No Failure” labels. Initially, binary classification was performed using this modified target variable, and subsequently the Easy Ensemble classifier was applied to classify the specific machine failure category.

In the proposed framework, the first stage performs binary failure detection using an ensemble-based Random Forest architecture consisting of nearly 100 Random Forest models trained using random over-sampling and under-sampling strategies. The second stage performs multiclass failure prediction using the Easy Ensemble classifier.

4.1. Binary Classification Steps

The binary classification stage was implemented using the following procedure. The methodology was inspired by the study on extreme class imbalance discussed in [10].

- Randomly shuffle and partition the majority-class observations into 100 independent subsets.
- Combine each majority-class subset with minority-class failure observations to generate balanced training datasets.
- Apply over-sampling and under-sampling techniques to rebalance the class distribution and train Random Forest classifiers on each balanced subset.
- Predict the class labels using all 100 Random Forest models. If more than 80% of the classifiers predict a failure event, the corresponding sample is forwarded to the second-stage multiclass classification framework.
- In the second stage, the identified failure observation is classified into its corresponding failure category, as illustrated in Fig. 9.

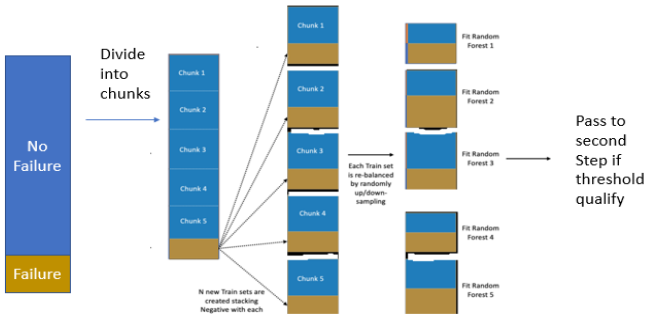


Figure 11: Binary classification stages (help from Franncesco Pochetti, 2019)

4.2. Multilabel Classification Step

The second stage of the proposed framework focuses exclusively on multiclass failure classification using the failure labels present in the dataset. In this stage, the model is trained only on the component failure classes, namely Comp1, Comp2, Comp3, and Comp4. During the testing phase, only those observations identified as failure instances in the first-stage binary classification are forwarded to the second-stage classifier for determining the exact failure category.

The classification accuracy of the proposed framework is evaluated based on this hierarchical prediction process. The corresponding bar chart illustrates the class distribution for individual component failure categories. Since the extreme class imbalance problem is primarily addressed during the first-stage binary classification, the second-stage classifier achieves improved failure categorization performance. The proposed hybrid framework achieved an overall classification accuracy of approximately 92%.

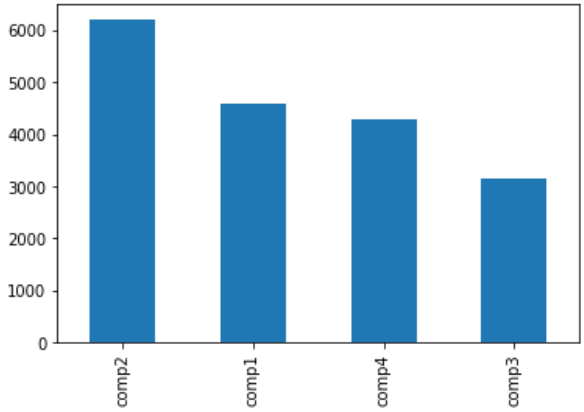


Figure 12: Number of instances of component failure for second step

Now we see the confusion matrix for first step and second step classification

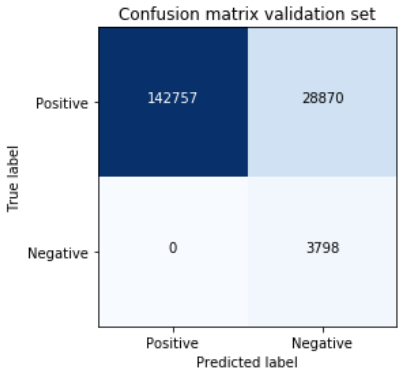


Figure 13: First Step Binary classiifcation confusion matrix

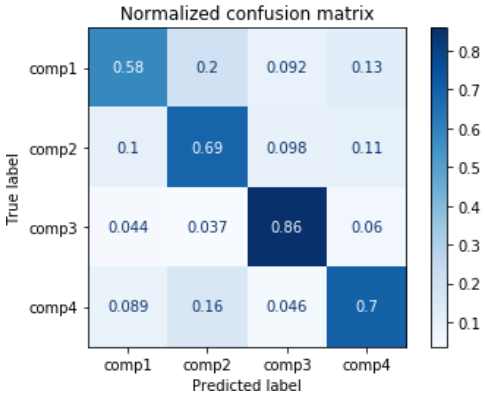


Figure 14: Second Step classiifcation confusion matrix

5. RESULTS

In this section, the experimental results obtained from different machine learning models are comparatively analyzed. The proposed hybrid two-step classification framework demonstrates the best overall performance among all evaluated approaches, primarily due to its ability to effectively handle extreme class imbalance in the predictive maintenance dataset.

The hybrid framework achieved higher average precision and recall values compared to conventional classification models, indicating improved minority-class prediction capability and better rare-event detection performance. In addition, the proposed two-stage architecture significantly improved overall classification precision by combining binary failure detection with multiclass failure categorization.

Table 4: COMPARISON OF METHODS

Method	***Avg Precision***	***Avg recall***	***Accuracy***
RUSBoosted trees	0.26	0.72	84.12
Balanced Random forest	0.28	0.74	84.23
Easy Ensemble	0.27	0.73	84.18
Hybrid two step model	0.32	0.70	92.01

6. CONCLUSION AND FUTURE WORK

Classification problems involving highly imbalanced industrial datasets often require specialized machine learning techniques and imbalance-aware learning strategies. From the experimental analysis, it is evident that classification accuracy alone cannot be considered a reliable performance metric for rare-event prediction problems. Evaluation measures such as precision, recall, ROC analysis, and confusion matrices provide a more meaningful assessment of model performance in highly imbalanced environments.

The proposed hybrid two-step classification framework demonstrated improved predictive performance for machine failure prediction by effectively handling extreme class imbalance. In future work, maintenance history and machine metadata can be incorporated into the predictive framework to further improve model accuracy, failure detection capability, and overall predictive reliability in industrial IoT systems.

References

[1] Menard, Scott Applied Logistic Regression Analysis. 2nd ed. Quantitative Applications in the Social Sciences. Thousand Oaks, CA: SAGE Publications, Inc., 2002. doi: 10.4135/9781412983433.

[2] Robert Malouf. 2002. A comparison of algorithms for maximum entropy parameter estimation. In Proceedings of the 6th conference on Natural language learning - Volume 20 (COLING-02). Association for Computational Linguistics, USA, 1–7. DOI:https://doi.org/10.3115/1118853.1118871

[3] Seiffert, Chris, et al. "RUSBoost: A hybrid approach to alleviating class imbalance." IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 40.1 (2009): 185-197.

[4] Freund, Y. & Schapire, R. (1994). A decision-theoretic generalization of on-line learning and an application to boosting. In Proceedings EuroCOLT'94: European Conference on Computational Learning Theory.LNCS.

[5] Chen, Chao, Andy Liaw, and Leo Breiman. "Using random forest to learn imbalanced data." University of California, Berkeley 110 (2004): 1-12.

[6] X. Y. Liu, J. Wu and Z. H. Zhou, "Exploratory Under sampling for Class-Imbalance Learning," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 39, no. 2, pp. 539-550, April 2009.

[7] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in Proc. 14th Int. Conf. Mach. Learn.,Nashville, TN, 1997, pp. 179–18

[8] Lian, Chunfeng, Su Ruan, and Thierry Denœux. "An evidential classifier based on feature selection and two-step classification strategy." Pattern Recognition 48.7 (2015): 2318-2327.

[9] https://www.kaggle.com/c/mpred-datascience-challenge/data

[10] Franncesco Pochetti (2019, March 14). Extreme label imbalance: when you measure the minority class in basis points [Blog post]. Retrieved from http://francescopochetti.com/extreme-label-imbalance-when-you-measure-the-minority-class-in-basis-points/