Predictive Maintenance in Semiconductor Manufacturing Using Machine Learning on Imbalanced Dataset

Aziz Ahmad[1], Syed Amir Ali Shah[2], Spogmay Yousafzai[3]

[1]Department of Computer Science, National Research University Higher School of Economics, Moscow, Russia

[2]Department of Information Security & Artificial intelligence, National Research University Higher School of Economics, Moscow, Russia

[3]Department of Computer Software Engineering, University of Engineering and Technology, Pakistan

DOI: https://dx.doi.org/10.51244/IJRSI.2025.1210000018

Received: 14 September 2025; Accepted: 22 September 2025; Published: 28 October 2025

ABSTRACT

Semiconductor manufacturing produces complex high-dimensional data datasets that contain mostly operational records and show product failure occurrences only in a limited portion. Several research studies use machine learning algorithms for predictive maintenance but very few address the issue of SECOM (imbalanced dataset) which contain up to 93% successful outcomes. This paper explains the existing research gap regarding imbalanced data of SECOM dataset and presents an integrated approach with innovative feature reduction and oversampling algorithms and model optimization methods. Our experiments involving the SECOM Semiconductor Manufacturing process dataset with an initial 591 features were reduced to 63 and processed by PCA which led to the Support Vector Classifier (SVC) producing the most accurate results at 98.6% while maintaining robust calibration. The visualization includes both a correlation heatmap showing related features and pie charts showing class distribution before and after data balancing techniques are applied. This research presents implications for predictive maintenance within semiconductor fabs together with future work recommendations.

Keyword- Predictive Maintenance, Semiconductor Manufacturing, SECOM Dataset, Imbalanced Data, Oversampling, Feature Reduction, SVC

INTRODUCTION

The predictive maintenance approach in semiconductor manufacturing proves essential for decreasing operational stoppages along with enhancing production output while maintaining high-quality automatic manufacturing. The main challenge in the SECOM dataset stems from the internal data imbalance since failures exist in only 6.64% of the cases while successful outcomes take up the remaining records. The training of reliable machine learning models faces complexity due to both the high imbalance of this dataset together with its unlabelled sensor data and maintaining high dimensions. Research has shown multiple machine learning techniques succeed in fault detection and maintenance scheduling but insufficient effort aims at correcting the calibration reissues and misclassification risks found in imbalanced datasets. The article provides an all-encompassing approach which optimizes data cleaning and feature reduction and oversampling and hyper parameters adaptation through specialized techniques for the SECOM semiconductor manufacturing dataset with its imbalanced characteristics.

LITERATURE REVIEW

In recent times machine learning prediction applications has attracted in a variety of industrial products. In the field of semiconductor manufacturing, early work by Susto et al. (2015) introduced multiple classifiers for managing high dimensional data within semiconductor manufacturing. This ground breaking study did not

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue X October 2025

explicitly address the challenges caused by imbalanced target classes distribution, common problem in real world fabrication environments. Other researchers have directed its investigations towards discovering sensor anomalies and developing physical models for equipment health predictions. Study by Gupta et al. (2022) demonstrate deep learning with ensemble methods achieve high accuracy performance in fault detection. These research works often neglect dataset imbalance or assume balanced data to make conclusions but they do not address the effect of oversampling on model calibration. Chawla et al. (2002) and He et al. (2008) [4] have explored oversampling methods for training data minority class enhancement through SMOTE. Although these methods are widely used in other domains, their integrated application in semiconductor manufacturing predictive maintenance is still underexplored. The research shows that oversampling works well to balance sample distributions yet results in incorrect probability estimation according to van den Goorbergh et al. (2022). The research on anomaly detection and inter-sensor transfer learning conducted by Yan et al. (2024) offers useful information regarding model performance in industrial settings while primarily examining anomaly detection features instead of complete predictive maintenance solutions. The work presented by Wang et al. (2022) together with Lee et al. (2014) showed how statistical filtering and principal component analysis (PCA) succeed in reducing manufacturing data dimensions while maintaining their vital variation. Research is lacking to determine how algorithms that reduce features work together with oversampling techniques to improve both model performance and calibration when used in semiconductor manufacturing operations. Furthermore, several studies have explored how machine learning model calibration performs following the implementation of oversampling techniques. Support vector machines operate well on imbalanced data records according to Farrag et al. (2024), yet probabilistic output calibration needs precise tuning. Studies by Chen et al (2016) examine ensemble techniques and the foundational work from Russell and Norvig (2016) about artificial intelligence but do not connect these approaches directly to preprocessing techniques for semiconductors. The research by Leksakul et al. (2025) evaluates the performance capabilities of ANN Artificial Neural Networks and SVR Support Vector Regression along with MLP Multi-Layer Perceptron and RFR Random Forest Regressor and ARIMA Autoregressive Integrated Moving Average using real manufacturing data from a semiconductor plant. Machine learning methods demonstrate their flexibility by reducing production interruptions and increasing operational efficiency according to the research findings. Farrag et al. (2024) conducted a study which introduces a predictive framework for minority classes within semiconductor manufacturing data that manages noise together with class imbalance concerns. The developed model exhibits positive results through its 0.95 AUC value and its 0.66 precision and 0.96 recall quality metrics which provide details about future maintenance operations and product quality levels. The study by El Mourabit et al. (2020) offers a machine learning predictive system for semiconductor failures which includes data preprocessing and feature selection to enhance prediction accuracy. Guo et al. (2024) examines different detection methods alongside classification methods and location methods in transmission lines and distribution systems while explaining machine learning applications for fault diagnosis. Salem (2018) investigates fault diagnosis systems in semiconductor production which face challenges due to misbalanced and partial data. The authors conduct a study of different machine learning approaches to enhance their ability to detect faults when working with these specific data types. Semiconductor manufacturing problems dealing with rare class predictions in highly imbalanced datasets receive attention through a model employing Particle Swarm Optimization and Deep Belief Networks from Kim et al. (2017). The study by Deb et al. (2020) investigates the Chicken Swarm Optimization algorithm that enables the optimization of imbalanced dataset models in machine learning applications for semi-conductor production. Biau & Scornet (2016) delivers an extensive study of Random Forest algorithms to provide insights about their usage both for predictive maintenance and imbalanced data applications in semiconductor manufacturing. while previous work has provided valuable insights into individual components, oversampling, feature reduction, and machine learning algorithms, few studies have offered an end-to-end framework that simultaneously addresses the imbalance, dimensionality, and calibration challenges in semiconductor predictive maintenance. This research seeks to fill that gap by

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue X October 2025

combining robust preprocessing with advanced oversampling and model tuning, thereby achieving superior predictive performance and calibration.

METHODOLOGY

Data Overview and Preprocessing

The SECOM (Semiconductor Manufacturing Process) dataset contains process control data from a semi-conductor enterprise into a high-dimensional time-series format. The dataset contains 591 numeric features along with a timestamp field while its target variable represents failed products with label -1 and passed products with label 1. A strong class imbalance between failed products at 6.64% and other non-defective products presents major difficulties in supervised learning. The data required preparation through an effective machine learning preprocessing pipeline. The analysis began with feature correlation to exclude attributes whose connection with the target variable was weak. The study removed all features whose Pearson correlation coefficient was lower than 0.05 because researchers wanted to keep variables which demonstrated meaningful predictive aptitude. The elimination of weak relationships during this step fights against overfitting problems along with streamlining computational complexities in later modeling stages. Next, missing data treatment was conducted. Features with more than 20% missing values were eliminated, as excessive imputation could introduce noise and bias. For the remaining features with occasional missing entries, we performed mean imputation based on the missing at random (MAR) assumption. This step preserves data volume while ensuring statistical integrity. The reduced set of 63 features remained dimensionally high and potentially displayed multicollinearity. Next Principal Component Analysis (PCA) underwent implementation to achieve a new feature space transformation. PCA achieves two objectives: first it simplifies multidimensional data by maintaining most data variations and secondly it addresses multicollinearity issues that decrease model interpretability and performance. The analysis of principal components resulted in 32 features which maintained greater than 90% of all original data variations. The data transformations through this method create an informative reduced-dimensional representation that the classification models can utilize. The combination of statistical filtering and unsupervised dimensionality reduction provides a strong foundation for reliable and interpretable model development.

Addressing Data Imbalance

Class imbalance is a defining characteristic of many real-world industrial datasets, especially in fault detection and predictive maintenance contexts. The SECOM dataset contains a failure class that comprises only less than 7% of the data which leads to standard classification models strongly favoring the pass class predictions. Standard classification models can perform well in terms of accuracy when they predict every instance to be non-faulty but they miss all real failures that reduces predictive maintenance effectiveness. For balancing the imbalanced data we used the common oversampling techniques SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling). The SMOTE approach creates synthetic samples through the connection points of minority class examples with their nearest neighbors on the data set dimension. The methodology considers synthetic data more beneficial than basic duplicate data production because standard duplication leads to overfitting. By applying ADASYN on SMOTE technology more emphasis is placed on the classification-intensive points near the decision border to improve training accuracy. The training process included class weighting as a means to decrease bias that results from class imbalance. The model implements a higher punishment for misclassifying minority examples which encourages it to create complex decision boundaries. Visual validation of the class distribution before and after oversampling is provided using pie charts (Figure 2 and Figure 3). Before oversampling, failure cases were heavily underrepresented. The dataset obtains balanced class distribution after applying SMOTE and ADASYN, which ensures more equitable training and leads to enhanced model generalization.

Model Training and Evaluation

Once the data was appropriately preprocessed and rebalanced, we conducted an extensive comparative analysis of several supervised learning algorithms. Which includes both traditional classification models and more advanced ensemble-based methods. Specifically, we trained and evaluated Logistic Regression, K-Nearest

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue X October 2025

Neighbors (KNN), Gaussian Naive Bayes, Decision Trees, Bagging Classifier, AdaBoost, Gradient Boosting, Random Forest, and Support Vector Classifier (SVC). We evaluated these models performance in regard to working with high-dimensional data streams and imbalanced distributions while performing predictive maintenance operations. Each model required optimization for its performance which was achieved through the use of Bayesian Optimization for hyperparameter tuning. Bayesian Optimization proved to be a superior choice compared to traditional grid search and random search methods because it excels in finding optimal solutions efficiently for complex search domains. Bayesian Optimization estimates probabilistic behavior of the objective function thus carefully determining which set of hyperparameters requires testing next. The method decreases evaluation requirements and increases the chances of identifying optimal or near-optimal configurations for each classifier. The dataset was divided into training and testing parts after oversampling ended by using 80% training data and reserving 20% for testing. The model evaluation relied on five metrics that included accuracy along with precision, recall and F1 score and Area Under the Receiver Operating Characteristic Curve (AUC). The accuracy rating presents how well the system correctly predicts situations but precision focuses on determining correct failure predictions. The model's ability to detect actual failures appears as Recall since it reflects its sensitivity performance. When dealing with imbalanced classes the F1 score computes precision and recall through harmonic mean evaluation to provide balanced results. AUC delivers performance evaluation through discrimination analysis of model behavior at different threshold points to identify superior classification ability. Based on the performance metrics summarized in Table 1, the Support Vector Classifier (SVC) is the most effective as the top-performing model. It achieved an accuracy of 98.6%, along with high precision, recall, and an AUC score of 0.99. This exceptional performance is primarily attributed to the SVC kernel-based learning mechanism, which enables it to find complex nonlinear patterns within high dimensional data spaces at its best level when combined with PCA for dimension reduction. In comparison, simpler models like Logistic Regression and Gaussian Naive Bayes exhibited relatively lower performance, underscoring the necessity of employing sophisticated classifiers that are capable of modeling intricate patterns in high dimensional, imbalanced datasets typically encountered in semiconductor manufacturing processes.

Visualization of Insights

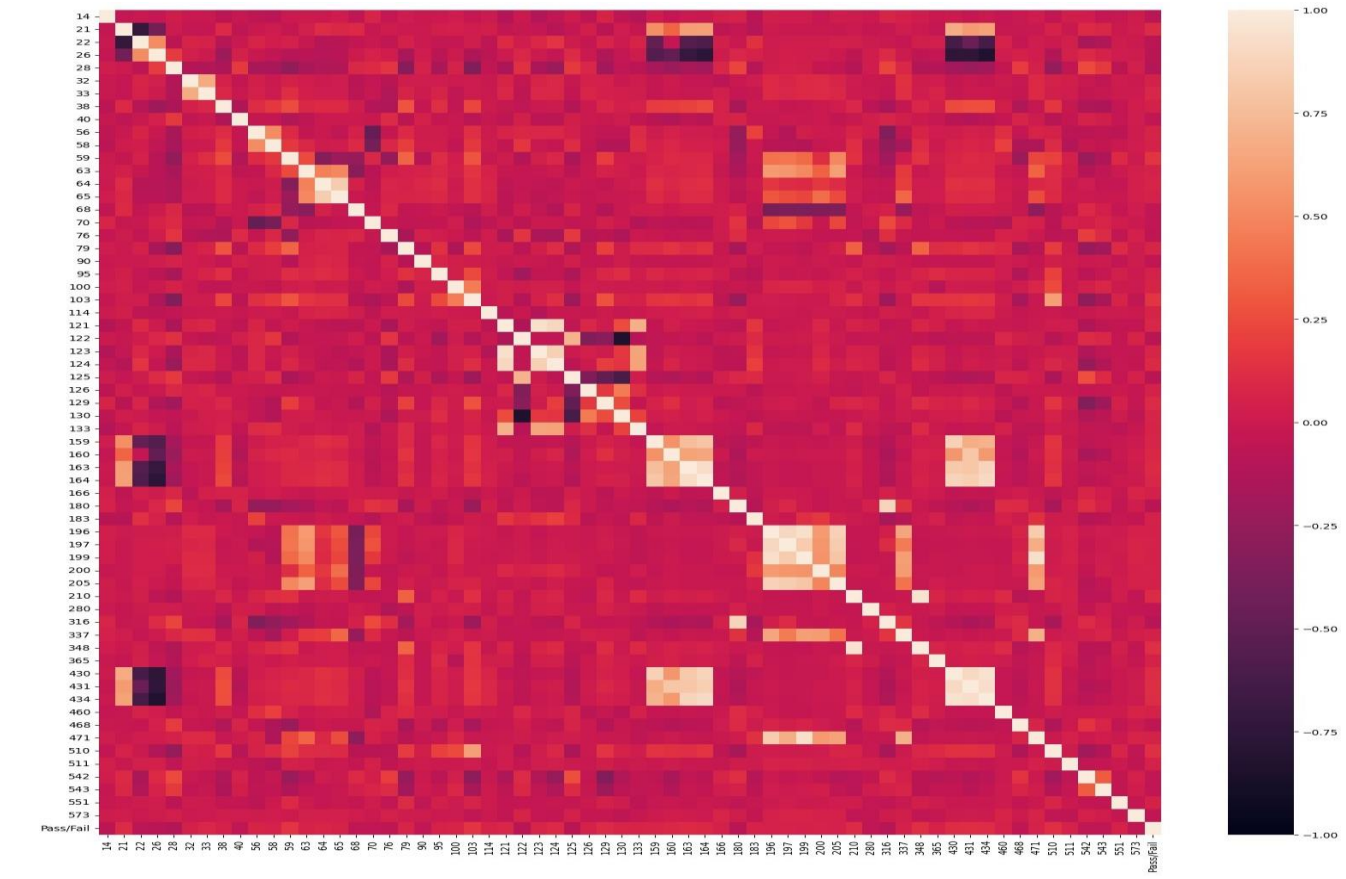

Figure 1: Heatmap visualize the correlation between the features

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue X October 2025

The heatmap (Figure 1) visualization depicts the interrelations of selected features. A solid positive linear relationship exists between attributes '180' and '316' and a profound negative relationship appears between '122' and '130' while '59', '103', '210', '348' strongly affect the target variable. The visual presentation identifies certain attributes that provide the strongest predictive power for determining system failures.

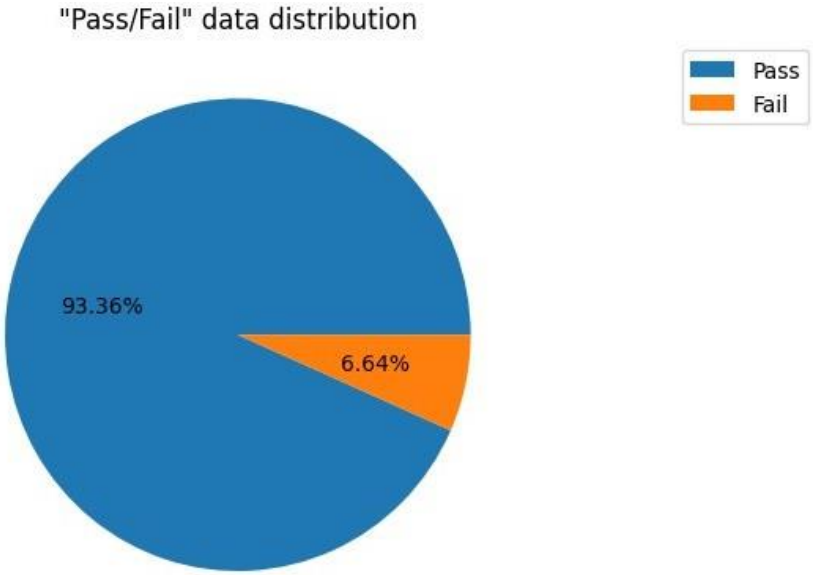

Figure 2: Pie charts shows the target class distribution before oversampling

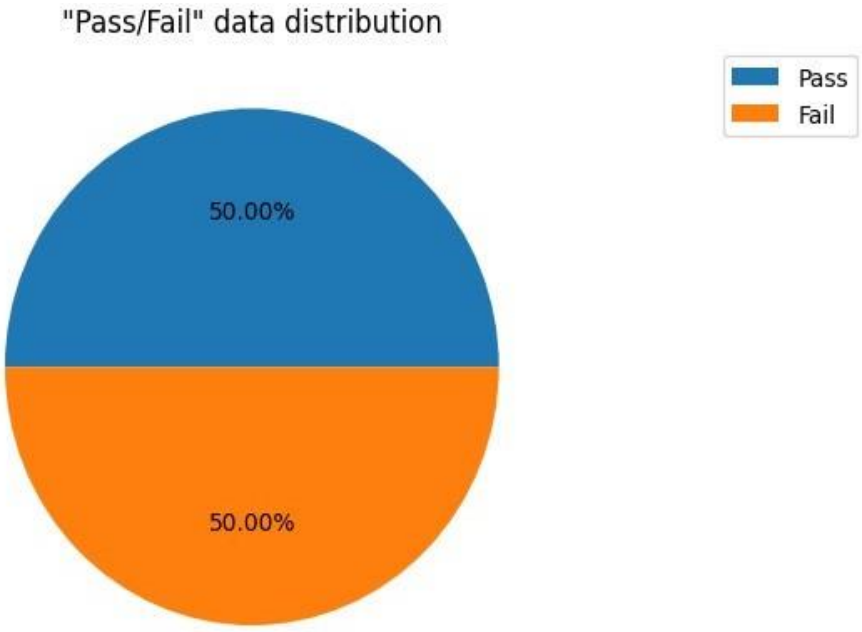

Figure 3: Pie charts shows the target class distribution after oversampling

The distribution of target classes appears in (Figure 2) via a pie chart before implementing oversampling methods. This visualization demonstrates the extreme class imbalance, where only 6.64% of the records represent failed products, emphasizing the need for data balancing techniques to improve model performance and predictive reliability.

The target class distribution displays the results of oversampling in (Figure 3). The balanced dataset distribution appears in the pie chart to show how the previous uneven data was balanced for equal representation between pass and fail target classes. The implementation of balancing processes provides machine learning models with adequate representation from both classes improving their accuracy and robustness.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue X October 2025

RESULTS

Our experimental evaluation confirms that the using of feature reduction and oversampling techniques proved effective for dealing with imbalanced data while decreasing runtime complexity in experimental testing. Support Vector Classifier (SVC) demonstrates the best performance among tested all classifiers since it achieves 98.6% accuracy as well as the best performance for precision, recall and AUC scores according to Table 1. Several features possess high relationship to the target variable according to the heatmap during our analysis and the pie chart data shows class distributions transform significantly due to oversampling.

Table 1. Performance Metrics of Various Models

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	75.9%	76.1%	75.9%	75.9%	0.82
K-Nearest Neighbors	82.4%	87.1%	82.4%	81.9%	0.88
Gaussian Naïve Bayes	76.6%	76.6%	76.6%	76.6%	0.80
Decision Tree	82.6%	82.7%	82.6%	82.6%	0.85
Bagging Classifier	97.3%	97.4%	97.3%	97.3%	0.98
AdaBoost Classifier	81.9%	81.9%	81.9%	81.9%	0.86
Gradient Boosting Classifier	93.7%	94.0%	93.7%	93.7%	0.97
Random Forest Classifier	96.6%	96.7%	96.6%	96.6%	0.97
SVC	**98.6%**	**98.6%**	**98.6%**	**98.6%**	**0.99**

DISCUSSION

This study addresses a critical gap in the reviewed studies by providing an end-to-end framework that integrates oversampling, feature reduction, and model tuning tailored to the imbalanced nature of semiconductor manufacturing data. The preprocessing workflow which begins with feature selection then processes missing data points and applies Principal Component Analysis maintains only the beneficial information from the dataset. The effective balancing of the dataset requires the use of oversampling methods SMOTE and ADASYN because target distribution changes can be seen in the produced pie charts. Furthermore, the heatmap for feature correlations reveals that attributes, such as '180' and '316', have a strong positive correlation, while others, such as '122' and '130', show a negative relationship. These insights suggest that the data contains inherent patterns that, when properly exploited, lead to significant improvements in predictive performance. Our experiments indicate that the SVC, which leverages kernel based learning, is particularly optimal choice for this task, outperforming other classifiers by achieving an accuracy of 98.6% along with excellent precision and recall. The implementation of robust data preprocessing methods with advance oversampling and hyperparameter tuning thus provides a promising framework for reliable predictive maintenance in semiconductor fabs.

CONCLUSION

In this article, we have developed a framework for semiconductor manufacturing predictive maintenance which addresses problems caused by unbalanced data distribution. Our method establishes significant model performance improvement thorough data cleaning procedures with advanced feature reduction techniques and optimal oversampling methods and precise model parameter optimization. The experimental demonstrate that Support Vector Classifier (SVC) outperforms to all other models as it achieves an accuracy of 98.6% with robust performance metrics. The visualizations include both a feature correlation heatmap and target class distribution pie charts which provide critical insights into the underlying data structure and the impact of oversampling. Future work will explore further improvements in model interpretability enhancement and real-time sensor implementation tasks for maintenance scheduling dynamic industrial operational settings.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue X October 2025

REFERENCES

1. Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. (2014). Machine learning for predictive maintenance: A multiple classifier approach. IEEE transactions on industrial informatics, 11(3), 812-820.
2. Thomas, J., Patidar, P., Vedi, K. V., & Gupta, S. (2022). An analysis of predictive maintenance strategies in supply chain management. Int J Sci Res Arch, 6(01), 308-17.
3. 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.
4. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) (pp. 1322-1328). Ieee.
5. Van den Goorbergh, R., van Smeden, M., Timmerman, D., & Van Calster, B. (2022). The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. Journal of the American Medical Informatics Association, 29(9), 1525-1534.
6. Yan, P., Abdulkadir, A., Luley, P. P., Rosenthal, M., Schatte, G. A., Grewe, B. F., & Stadelmann, T. (2024). A comprehensive survey of deep transfer learning for anomaly detection in industrial time series: Methods, applications, and directions. IEEE Access, 12, 3768-3789.
7. Wang, S., & Chen, Y. (2024, July). Improved yield prediction and failure analysis in semiconductor manufacturing with xgboost and shapley additive explanations models. In 2024 IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA) (pp. 01-08). IEEE.
8. Lee, J., Wu, F., Zhao, W., Ghaffari, M., Liao, L., & Siegel, D. (2014). Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications. Mechanical systems and signal processing, 42(1-2), 314-334.
9. Farrag, A., Ghali, M. K., & Jin, Y. (2024). Rare Class Prediction Model for Smart Industry in Semiconductor Manufacturing. arXiv preprint arXiv:2406.04533.
10. Chen, K., Huang, C., & He, J. (2016). Fault detection, classification and location for transmission lines and distribution systems: a review on the methods. High voltage, 1(1), 25-33.
11. Norvig, P. R., & Intelligence, S. A. (2002). A modern approach. Prentice Hall Upper Saddle River, NJ, USA: Rani, M., Nayak, R., & Vyas, OP (2015). An ontology-based adaptive personalized e-learning system, assisted by software agents on cloud storage. Knowledge-Based Systems, 90, 33-48.
12. Leksakul, K., Suedumrong, C., Kuensaen, C., & Sinthavalai, R. Predictive Maintenance in Semiconductor Manufacturing: Comparative Analysis of Machine Learning Models for Downtime Reduction.
13. El Mourabit, Y., El Habouz, Y., Zougagh, H., & Wadiai, Y. (2020). Predictive system of semiconductor failures based on machine learning approach. International journal of advanced computer science and applications (IJACSA), 11(12), 199-203.
14. Guo, P., & Chen, Y. (2024, July). Enhanced yield prediction in semiconductor manufacturing: Innovative strategies for imbalanced sample management and root cause analysis. In 2024 IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA) (pp. 1-6). IEEE.
15. Salem, M., Taheri, S., & Yuan, J. S. (2018). An experimental evaluation of fault diagnosis from imbalanced and incomplete data for smart semiconductor manufacturing. Big Data and Cognitive Computing, 2(4), 30.
16. Kim, J. K., Han, Y. S., & Lee, J. S. (2017). Particle swarm optimization–deep belief network–based rare class prediction model for highly class imbalance problem. Concurrency and Computation: Practice and Experience, 29(11), e4128.
17. Deb, S., Gao, X. Z., Tammi, K., Kalita, K., & Mahanta, P. (2020). Recent studies on chicken swarm optimization algorithm: a review (2014–2018). Artificial Intelligence Review, 53(3), 1737-1765.
18. Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227.