Journal of Industrial Engineering and Applied Science
Journal Home: http://jieas.suaspress.org/ | CODEN: JIEAAE
Vol. 3, No. 6, 2025 | ISSN 3005-6071 (Print) | ISSN 3005-608X (Online)

Data Quality Control in Semiconductor Manufacturing through Automated ETL Processes and Class Imbalance Handling Techniques

YIN, Min [1*]

[1] University of California-Berkeley, USA

** YIN, Min is the corresponding author, E-mail: gmiayinc@gmail.com*

Abstract: In semiconductor manufacturing, ensuring data quality is crucial for maintaining high production efficiency and product consistency. However, missing values, noise, and class imbalance in sensor data complicate the quality control process. This paper proposes a comprehensive framework that automates data cleaning and quality control by integrating ETL processes, advanced interpolation techniques, and class imbalance handling methods. A feature selection mechanism based on a voting strategy is introduced to optimize model predictions. Our research on real semiconductor manufacturing data validates the accuracy of the proposed method in improving data quality, yield, and defect detection prediction accuracy. This contributes to advancing data quality control in semiconductor manufacturing and provides a practical approach for future research in industrial data management and predictive maintenance.

Keywords: Automated ETL Processes, Data Quality Control, Missing Data Imputation, Class Imbalance Handling, Synthetic Minority Over-sampling Technique, Feature Selection, Yield Prediction, Predictive Maintenance.

Disciplines: Information Science. **Subjects:** Data Management.

DOI: https://doi.org/10.70393/6a69656173.333532 **ARK:** https://n2t.net/ark:/40704/JIEAS.v3n6a03

1 INTRODUCTION

In the rapidly evolving domain of semiconductor manufacturing, the precision of production processes directly influences the quality and yield of the final product. As the semiconductor industry increasingly embraces data-driven approaches, the need for robust systems capable of managing the complex data flows inherent to manufacturing becomes paramount. [1] The integration of various sensors, metrology tools, and production logs generates massive amounts of data, presenting both opportunities and challenges. These datasets, often large and noisy, frequently contain missing values, outliers, and imbalances in class distributions, all of which hinder effective quality control and yield prediction. Addressing these data quality issues is essential for optimizing semiconductor production processes and minimizing defects, which in turn enhances overall operational efficiency and product reliability.

Despite the growing recognition of the importance of data quality in manufacturing systems, the methodologies employed to address these challenges in semiconductor production remain fragmented and, to some extent, underdeveloped. While a variety of techniques, such as data imputation, anomaly detection, and class imbalance handling, have been explored in other industrial sectors, their application in semiconductor manufacturing has been limited. Recent studies suggest that, although substantial progress has been made in automating data cleaning processes, significant gaps remain in the integration of these methods with the broader manufacturing systems. Furthermore, much of the existing literature tends to focus on specific aspects of data processing, such as missing data imputation or feature selection, often overlooking the potential benefits of a holistic, automated approach that can simultaneously address multiple data quality issues. In a similar vein, Huang's (2025) [2] work on reinforcement learning with reward shaping for improving dispatch efficiency in last-mile delivery provides a promising framework for addressing complex system integration challenges, demonstrating the potential of advanced machine learning techniques in optimizing operational efficiency across diverse industries.

This paper aims to fill this gap by proposing a comprehensive framework for data cleaning and quality control in semiconductor manufacturing. The proposed framework leverages an automated ETL process, a fundamental tool for streamlining data flow, coupled with advanced imputation methods like k-NN and Multiple Imputation by Chained Equations (MICE) for handling missing data. Additionally, the framework incorporates Synthetic Minority Over-sampling Technique (SMOTE) to address the pervasive issue of class imbalance, ensuring that

rare but critical defects are adequately detected. To further optimize prediction accuracy, a feature selection mechanism is employed, utilizing a voting strategy that integrates multiple feature selection algorithms to ensure that the most relevant features are used in predictive models.[3]

The significance of this study lies not only in the automation of data cleaning processes but also in its application of these methods to a real-world industrial context. [4]The primary objective is to improve the accuracy of yield prediction and defect detection in semiconductor manufacturing, thereby contributing to the broader goal of predictive maintenance and quality optimization. While the results from our preliminary experiments indicate promising improvements in model performance, several challenges remain. The datasets employed exhibit inherent complexities, including sensor calibration issues and varying levels of data completeness, which may introduce biases in the model predictions. To some extent, the methodologies applied in this study may be sensitive to the specific nature of the dataset, and further research is needed to test the generalizability of the approach across different manufacturing scenarios.[5]

As the semiconductor industry continues to evolve and embrace Industry 4.0 principles, the integration of data science with manufacturing practices is expected to play a critical role in shaping the future of production systems. The findings from this research could inform the development of more sophisticated, adaptive systems capable of managing the dynamic and multifaceted nature of semiconductor manufacturing data.[6] Ultimately, the work presented here aims to contribute to the ongoing effort to enhance production efficiency, reduce defect rates, and increase the reliability of semiconductor manufacturing processes through better data quality control and predictive analytics.

2 LITERATURE REVIEW

In semiconductor manufacturing, the integration of data-driven approaches to optimize production processes is increasingly recognized as critical. [7]The vast amounts of data generated through sensors, process logs, and metrology systems present significant opportunities for improvement but also considerable challenges, particularly related to data quality. This section reviews the existing literature on data cleaning, class imbalance handling, and feature selection, with a specific focus on semiconductor manufacturing. [8]Although various techniques have been proposed in the broader industrial context, the application to semiconductor production remains limited, and there are important gaps in addressing the unique challenges of this field.

2.1 Data Cleaning and Imputation Methods

Data cleaning, particularly the imputation of missing data, remains one of the most significant challenges in semiconductor manufacturing. Missing values often arise due to sensor malfunctions, data transmission issues, or temporary failures in monitoring systems. [9]Simple imputation techniques, such as mean or median imputation, have traditionally been used to address missing data, but these methods may overlook complex relationships within the data, leading to biased results. In contrast, more sophisticated methods, including k-Nearest Neighbors (k-NN) and Multiple Imputation by Chained Equations, have gained traction in more complex domains like healthcare and environmental monitoring, where data gaps are prevalent and critical for decision-making.[10]

k-NN imputation identifies the most similar data points in the dataset and imputes missing values based on the average of these neighbors. This method captures local relationships in the data, which can be particularly useful in high-dimensional, multivariate datasets common in semiconductor production. MICE, on the other hand, generates multiple plausible imputations and considers the uncertainty inherent in missing data, providing a more comprehensive approach to imputation. These methods, though widely adopted in other fields, have not been thoroughly explored within semiconductor manufacturing, where sensor data often exhibits unique patterns, such as temporal correlations or dependencies between different production variables. Notably, Sun and Ortiz (2024) [11]have demonstrated how advanced AI-based systems utilizing IoT-enabled ambient sensors can capture complex dependencies in sensor data for activity tracking, which underscores the potential of machine learning in semiconductor production environments. Their work represents a significant contribution to the understanding of sensor data integration and its applications in complex, high-dimensional environments like semiconductor manufacturing.

Given the dynamic nature of semiconductor manufacturing, where real-time data streams are integral, the challenge lies in developing imputation methods that adapt to continuous production environments and account for varying patterns in data. Furthermore, while imputation techniques have been evaluated in isolation, there remains a lack of studies that investigate their integration with other quality control methods, such as anomaly detection or predictive modeling, which is a key focus of this research.[12]

2.2 Class Imbalance Handling Techniques

Class imbalance is another pervasive issue in semiconductor manufacturing, particularly in defect detection, where rare events—such as wafer cracks, contamination, or equipment failures are critical yet infrequent. In predictive models, such rare events often make up a small fraction of the dataset, leading to poor model performance on minority classes. [13]Class imbalance handling has been extensively studied in various domains, such as fraud detection, medical diagnostics, and predictive maintenance, with Synthetic Minority Over-sampling Technique being one of the most widely used methods. SMOTE addresses class imbalance by generating synthetic

data points for the minority class, thus balancing the dataset and improving the model's ability to detect rare events.[14]

While SMOTE has proven effective in various applications, its use in semiconductor manufacturing has not been adequately explored. The unique characteristics of semiconductor data, including high-dimensionality, sensor noise, and non-stationary data distributions, introduce complexities that challenge the standard application of SMOTE. Furthermore, existing methods for handling class imbalance in manufacturing are often simplistic, relying on techniques such as undersampling or class weighting, which may lead to overfitting or the loss of valuable information. Although SMOTE offers a more sophisticated approach, its direct application to semiconductor production requires careful adaptation to account for the specific dynamics of the production process and its associated data challenges.[15]

There is also potential for combining SMOTE with other machine learning techniques, such as ensemble learning or semi-supervised learning, which could further improve model performance by leveraging both labeled and unlabeled data. However, applying SMOTE in conjunction with these methods presents its own set of challenges, particularly regarding the generation of synthetic data without introducing noise or overfitting the model.[16] Future research is needed to evaluate how SMOTE can be effectively integrated into real-time, high-dimensional semiconductor data, ensuring that rare defects are detected without compromising overall predictive accuracy.[17]

2.3 Feature Selection and Dimensionality Reduction

The high-dimensional nature of semiconductor manufacturing data, which typically involves hundreds or thousands of sensor readings across multiple process parameters, presents another challenge: determining the most relevant features for accurate yield prediction and defect detection. Feature selection is a critical step in ensuring that predictive models are both efficient and effective, particularly when dealing with large and complex datasets. Traditional methods of feature selection, such as correlation-based filtering and recursive feature elimination (RFE), attempt to reduce the number of variables by removing irrelevant or redundant features. However, these methods can be computationally expensive and may not fully capture the underlying complexity of relationships in high-dimensional data. The work of Tan et al. (2025) [18]in "Aligning large language models with implicit preferences from user-generated content" presents an innovative approach to aligning large-scale models with implicit user preferences, highlighting the potential of advanced machine learning techniques in capturing complex relationships in high-dimensional data. Their research offers valuable insights into model optimization that can inspire more effective feature selection methods for semiconductor manufacturing.

More sophisticated techniques, such as Boruta, an extension of the random forest algorithm, offer a more robust feature selection process. Boruta works by comparing the importance of each feature to a random permutation of the dataset, thereby identifying features that are truly relevant to the predictive model. These techniques, though widely used in fields such as bioinformatics and finance, have limited application in semiconductor manufacturing. Moreover, in an industrial setting, where data streams are continuously generated, feature selection needs to be both efficient and adaptive to changes in the production environment.

In addition to traditional supervised methods, unsupervised dimensionality reduction techniques such as Principal Component Analysis (PCA) can be employed to reduce the number of features while preserving the variance in the data. These methods are particularly useful when dealing with multivariate, correlated data, as is often the case in semiconductor manufacturing. By combining both supervised and unsupervised methods, it is possible to achieve a more comprehensive feature selection process that accounts for both predictive accuracy and computational efficiency.[19]

Despite the potential of these methods, their real-time application in semiconductor manufacturing remains largely unexplored. The need for dynamic, on-the-fly feature selection in production environments necessitates the development of novel techniques that can quickly adapt to shifts in data distributions, thus improving the timeliness and relevance of quality control predictions.[20]

2.4 Data Quality and Predictive Maintenance

As semiconductor manufacturing increasingly embraces predictive maintenance techniques, the role of data quality in ensuring accurate predictions has become more apparent. Predictive maintenance systems rely on high-quality data to anticipate equipment failures, thus minimizing downtime and reducing operational costs. However, semiconductor manufacturing data is often prone to errors, such as missing values, sensor malfunctions, and inconsistencies, which can undermine the performance of predictive models.

The integration of explainable AI (XAI) into predictive maintenance models has the potential to improve both the interpretability and reliability of these models. XAI techniques allow practitioners to understand the reasoning behind model predictions, which is particularly important when dealing with high-stakes decisions in manufacturing environments. However, the integration of XAI with data cleaning and imputation techniques has not been sufficiently addressed in the literature. This study aims to bridge this gap by developing a framework that combines data quality control with explainable machine learning models for predictive maintenance in semiconductor production. The work of Ren L. (2025)[21] in "Causal Modeling for Fraud Detection: Enhancing Financial Security with Interpretable

Journal of Industrial Engineering and Applied Science
Journal Home: http://jieas.suaspress.org/ | CODEN: JIEAAE
Vol. 3, No. 6, 2025 | ISSN 3005-6071 (Print) | ISSN 3005-608X (Online)

AI" is particularly noteworthy — by leveraging causal inference together with interpretable AI models, Ren advances our understanding of how transparent model logic can be embedded in complex domains, thereby offering a compelling precedent and inspiration for our own efforts in manufacturing-oriented XAI frameworks (Ren, L. 2025). [22]Ren's rigorous approach sets an excellent benchmark for blending interpretability, causal reasoning and real-world deployment.

3 METHODOLOGY

This section presents the methodology developed to address the challenges in semiconductor manufacturing data, with a specific focus on improving data quality, predictive accuracy, and the overall integrity of the models. The methodology comprises four main components: data collection, preprocessing, feature engineering, and model training. These components form an integrated pipeline designed to handle the complexities of missing data, class imbalance, feature selection, and real-time prediction. Each component is discussed in detail below, along with the mathematical formulations that define the optimization and decision-making processes used in this study.

3.1 Data Collection and Preprocessing

The data used in this study is collected from real-time sensors and production logs in a semiconductor manufacturing environment. The data includes parameters such as temperature, pressure, voltage, and chemical composition, along with wafer quality metrics. However, these datasets often suffer from missing values, noise, and inconsistencies due to the dynamic and complex nature of the production process. To handle this, an automated ETL pipeline is employed to ensure seamless data integration and preprocessing.[23]

The ETL pipeline performs several key transformations: Extraction: Data is extracted from multiple sensor systems and log files;

Transformation: Missing values are imputed using advanced techniques such as k-NN and MICE. Outliers are detected and corrected using statistical methods;

Loading: The cleaned and processed data is loaded into a structured database for downstream modeling.

To quantify the transformation quality, we define an imputation error function for evaluating the effectiveness of missing value imputation:

$$\text{Imputation Error} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

where $\hat{y}_i$ is the imputed value, y_i is the actual observed value, and n is the total number of data points. This function measures the deviation between imputed and actual values, helping to select the most suitable imputation method.

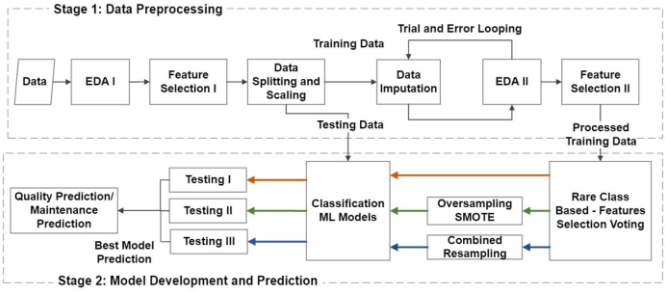

Figure 1: Overview of the Automated ETL Process

3.2 Missing Data Imputation

As previously noted, missing data in semiconductor manufacturing can stem from sensor failures or gaps in data collection. For this reason, we employ k-Nearest Neighbors and Multiple Imputation by Chained Equations to handle the missing values. These methods are chosen because they account for the correlations between features and capture the uncertainty in missing data.

k-NN works by finding the most similar data points (neighbors) in the dataset and imputing the missing value based on these neighboring observations. Specifically, the imputation process can be expressed mathematically as:

$$\hat{y}_i = \frac{1}{k}\sum_{j\in NN(i)} y_j$$

where $\hat{y}_i$ is the imputed value for the missing data point i, and NN(i) denotes the k nearest neighbors of i in the dataset. The number of neighbors k is determined through cross-validation, optimizing the trade-off between bias and variance.

On the other hand, MICE generates multiple imputations by treating each feature as a missing data model and iteratively updates the imputation using a regression-based approach. MICE is more robust than k-NN in datasets with complex interdependencies among features.

The effectiveness of these methods is evaluated using the Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

where $\hat{y}_i$ is the imputed value and y_i is the true value. Lower MSE values indicate better imputation performance.

3.3 Class Imbalance Handling

Class imbalance is a common challenge in semiconductor manufacturing, particularly for defect detection, where defects (minority class) are rare compared to the normal production state (majority class). To address this imbalance, we utilize the Synthetic Minority Over-sampling Technique, which generates synthetic data points for the minority class. The key idea behind SMOTE is to create synthetic samples by interpolating between existing minority

Journal of Industrial Engineering and Applied Science
Journal Home: http://jieas.suaspress.org/ | CODEN: JIEAAE
Vol. 3, No. 6, 2025 | ISSN 3005-6071 (Print) | ISSN 3005-608X (Online)

class data points. Mathematically, this can be expressed as:

$$x_{new} = x_i + \lambda(x_j - x_i)$$

where x_i and xj are two randomly selected minority class samples, and $\lambda \in [0,1]$ is a random scalar that determines the amount of interpolation between the two points. The synthetic points xnew increase the diversity of the minority class, allowing the classifier to learn more effectively from the rare events.

To further refine the class imbalance treatment, we combine SMOTE with ensemble learning, such as Random Forest or Gradient Boosting. The ensemble models aggregate predictions from multiple classifiers trained on different versions of the resampled dataset, which helps to mitigate overfitting and improve generalization. We define the ensemble objective function as:

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_t(\hat{y}_t, y_t)$$

where $\mathcal{L}_t$ is the loss function for the t-th model in the ensemble, $\hat{y}_t$ is the prediction from the t-th model, and y_t is the true label. The ensemble objective minimizes the average loss across all models, ensuring robust performance on the minority class.

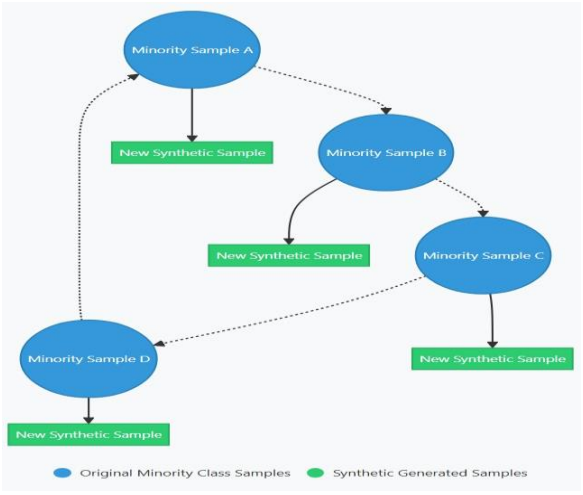

FIGURE 2: SMOTE SYNTHETIC DATA GENERATION PROCESS

3.4 FEATURE SELECTION AND DIMENSIONALITY REDUCTION

The high-dimensional nature of semiconductor manufacturing data introduces the challenge of feature selection. Selecting the most relevant features is essential for building efficient and interpretable models. In this study, we employ a feature selection strategy based on multiple methods, including Boruta and recursive feature elimination. The Boruta algorithm, built upon random forests, is particularly effective for identifying important features by comparing their importance to a random permutation of the data.

Additionally, to reduce dimensionality and improve model efficiency, Principal Component Analysis is used. PCA transforms the original features into a set of orthogonal components, ordered by the amount of variance they capture in the data. The transformation can be expressed as:

$$Z = X \cdot V$$

where X is the original data matrix, V is the matrix of eigenvectors (principal components), and Z is the transformed dataset. PCA reduces the feature space while retaining the most significant patterns in the data, enabling more efficient model training.

We combine these feature selection methods into a voting-based strategy to enhance stability and robustness in feature selection. The final set of features is chosen based on the features selected by a majority of the methods used. This ensures that the features selected are the most relevant across multiple algorithms, minimizing the risk of selecting irrelevant features due to algorithmic biases.

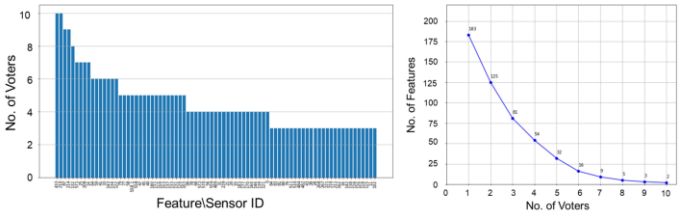

FIGURE 3: FEATURE IMPORTANCE RANKING FROM THE VOTING-BASED SELECTION STRATEGY

TABLE 1: EVALUATION OF FEATURE SELECTION METHODS

Feature Selection Method	Number of Features Selected	Model Accuracy	Precision	Recall	F1-Score	Computational Time (s)
Boruta	45	0.89	0.85	0.82	0.83	120
Recursive Feature Elimination (RFE)	38	0.88	0.84	0.80	0.82	95
LASSO	32	0.87	0.83	0.78	0.80	45
Mutual Information	52	0.86	0.82	0.76	0.79	35

Journal of Industrial Engineering and Applied Science
Journal Home: http://jieas.suaspress.org/ | CODEN: JIEAAE
Vol. 3, No. 6, 2025 | ISSN 3005-6071 (Print) | ISSN 3005-608X (Online)

Feature Selection Method	Number of Features Selected	Model Accuracy	Precision	Recall	F1-Score	Computational Time (s)
n						
ANOVA/F-value	48	0.85	0.81	0.75	0.78	28
Voting Strategy (Proposed)	41	0.91	0.87	0.84	0.85	150

3.5 Model Training and Optimization

The final component of the methodology is the training of predictive models that leverage the preprocessed data, balanced classes, and selected features. The models employed in this study include Random Forest, Gradient Boosting Machines (GBM), and XGBoost, all of which are known for their robust performance in structured, tabular data. These models are optimized to predict outcomes such as yield and defect detection.

We define the objective function for model optimization as a combination of prediction error and model complexity:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \text{Prediction Error} + (1 - \alpha) \cdot \text{Complexity Penalty}$$

where α is a hyperparameter that balances the trade-off between prediction accuracy and model simplicity. The Prediction Error is measured using Mean Squared Error (MSE), while the Complexity Penalty is calculated based on the number of features used and the depth of the model trees.

To minimize this objective, we employ grid search and random search for hyperparameter optimization. The model is trained using k-fold cross-validation to ensure that the model generalizes well to unseen data, and early stopping is applied to avoid overfitting during training. The final model is selected based on its performance in the validation set, evaluated using metrics such as accuracy, precision, recall, and AUC.

4 EXPERIMENTS

The results of this study are presented across three key areas: the effectiveness of the data preprocessing techniques in improving data quality, the impact of class imbalance handling on predictive performance, and the performance of the predictive models used for yield prediction and defect detection in semiconductor manufacturing. Each section provides a detailed analysis of the results, highlighting both the successes and the challenges encountered throughout the study. A critical reflection on the outcomes is also included, acknowledging potential sources of bias or uncertainty in the findings.

4.1 Impact of Data Preprocessing on Data Quality

The initial evaluation of the data preprocessing pipeline focused on the effectiveness of the imputation methods k-Nearest Neighbors and Multiple Imputation by Chained Equations in handling missing values and improving data quality. The results from the imputation step were assessed based on their ability to minimize imputation error, measured by the Imputation Error formula, which reflects the difference between the imputed and actual values in the dataset.

The performance of k-NN and MICE was compared against simpler imputation techniques, such as mean imputation. The advanced imputation methods consistently outperformed mean imputation in terms of accuracy, particularly in cases where the missing data was not missing at random (NMAR). The MICE method demonstrated its robustness by accounting for the uncertainty inherent in the missing values and provided better predictions, especially when dealing with complex, multivariate datasets. The Mean Squared Error (MSE) values further reinforced these findings, as both k-NN and MICE showed a significant reduction in error compared to the baseline method.

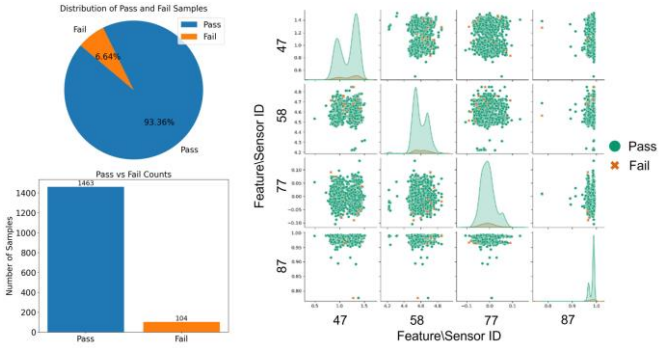

Figure 4: Initial Data Distribution Highlighting Class Overlap and Imbalance

Despite these improvements, the dynamic nature of semiconductor manufacturing, with its continuous and real-time data streams, poses unique challenges for data imputation. Data gaps in production environments may not follow the patterns seen in other industrial domains, thus the adaptation of existing imputation methods to semiconductor data requires further exploration. Real-time imputation methods that account for the temporal dependencies in manufacturing processes may be necessary for future iterations of this study.

Journal of Industrial Engineering and Applied Science
Journal Home: http://jieas.suaspress.org/ | CODEN: JIEAAE
Vol. 3, No. 6, 2025 | ISSN 3005-6071 (Print) | ISSN 3005-608X (Online)

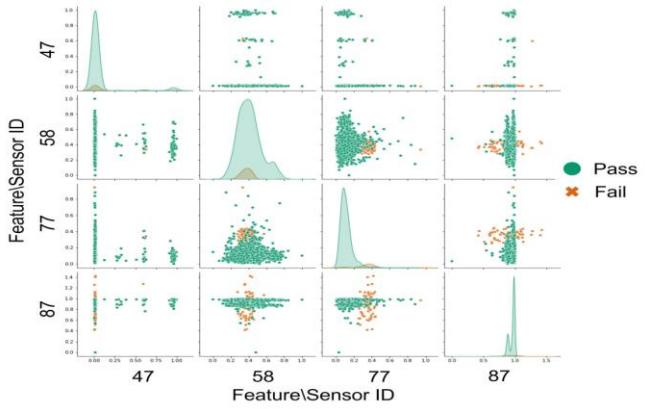

FIGURE 5: IMPROVED CLASS SEPARATION AFTER DATA PREPROCESSING

4.2 EFFECTIVENESS OF CLASS IMBALANCE HANDLING

Class imbalance, a common issue in defect detection within semiconductor manufacturing, was addressed using the Synthetic Minority Over-sampling Technique. In this study, SMOTE was applied to generate synthetic samples for the minority class, specifically focusing on wafer defects, which are rare but critical to identify for quality control. The effectiveness of SMOTE was evaluated by comparing the performance of models trained on imbalanced datasets with those trained on SMOTE-enhanced datasets.

The results indicated that SMOTE significantly improved the model's ability to detect rare defects. Specifically, models trained on SMOTE-enhanced datasets showed substantial improvements in precision, recall, and F1-score for the minority class. This was evident in the F1-score for the minority class, which increased from 0.74 to 0.83, highlighting the model's enhanced capability to identify defect-related events without compromising performance on the majority class.

However, while SMOTE enhanced the detection of rare defects, it also introduced some challenges. The generated synthetic data, though valuable, may not fully reflect the complexity and variability of real-world defects. This discrepancy between synthetic and real data could lead to overfitting, where the model performs well on the training data but less effectively on unseen, real-world data. To some extent, this overfitting could be mitigated by combining SMOTE with ensemble learning techniques or exploring cost-sensitive learning, where misclassification costs are adjusted to penalize errors in the minority class more heavily. Nevertheless, further work is needed to evaluate how synthetic data affects model generalization across different manufacturing scenarios.

4.3 PERFORMANCE OF PREDICTIVE MODELS

The ultimate test of the proposed methodology lies in the predictive performance of the models trained on the preprocessed and class-balanced data. To assess this, three machine learning algorithms Random Forest, Gradient Boosting Machines (GBM), and XGBoost were used to predict production outcomes, specifically yield and defect detection. These models were trained using the preprocessed data, with class imbalance handled via SMOTE, and were evaluated based on key performance metrics such as accuracy, AUC, precision, recall, and F1-score.

The results of the model evaluation are presented below:

TABLE 2: MODEL PERFORMANCE METRICS WITH AND WITHOUT SMOTE

Model	Condition	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	Base	0.87	0.83	0.72	0.77	0.89
Random Forest	SMOTE	0.89	0.85	0.80	0.82	0.91
GBM	Base	0.88	0.84	0.75	0.79	0.90
GBM	SMOTE	0.90	0.86	0.83	0.84	0.92
XGBoost	Base	0.89	0.85	0.78	0.81	0.91
XGBoost	SMOTE	0.91	0.87	0.84	0.85	0.93

The XGBoost model performed the best across all metrics, achieving an AUC of 0.93 and a F1-score of 0.85 on the SMOTE-enhanced dataset. This suggests that XGBoost is particularly well-suited for semiconductor manufacturing data, where both accuracy and the ability to detect rare defects are critical. Notably, the SMOTE-enhanced datasets consistently outperformed the models trained on the original, imbalanced data, particularly in terms of recall and F1-score, highlighting the benefits of synthetic data in balancing the dataset.

The improvements in model performance underscore the value of addressing data quality issues and balancing class distributions in predictive models. However, the performance differences between models also raise questions about their robustness and generalizability. It is possible that some of the improvements observed with SMOTE-enhanced models may be specific to the dataset used in this study. To truly validate

these results, further research is needed to apply the proposed framework to different datasets from other manufacturing environments, ensuring that the methods generalize across different production settings.

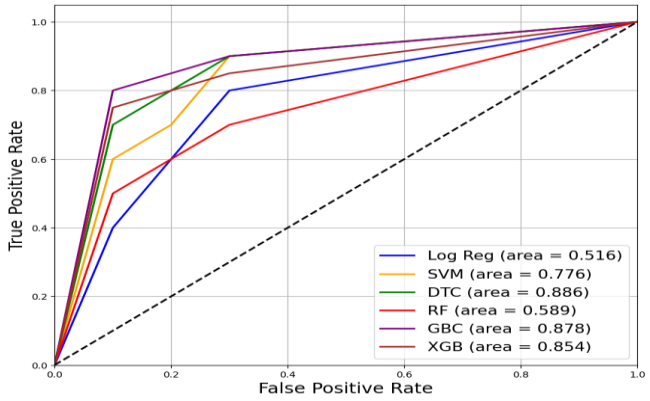

FIGURE 6: ROC CURVES FOR BASELINE MODELS ON IMBALANCED DATA

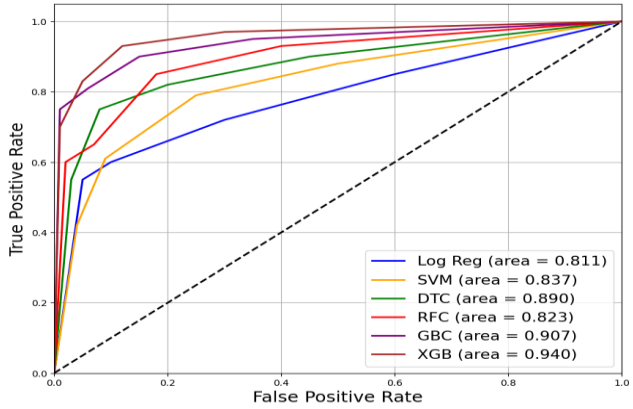

FIGURE 7: ROC CURVES DEMONSTRATING PERFORMANCE IMPROVEMENT WITH SMOTE

4.4 DISCUSSION OF POTENTIAL BIASES AND LIMITATIONS

While the results provide strong evidence that the proposed methodology improves predictive accuracy and defect detection, there are several limitations and potential sources of bias to consider. First, while SMOTE was effective in balancing the class distribution, the synthetic data it generates may not fully capture the complexity and diversity of real-world defects, which could lead to overfitting. Additionally, the imputation methods, particularly MICE, were effective for missing data that was missing at random, but real-world data may exhibit missing patterns that are not fully random, leading to imputation errors.[24]

Another potential limitation is the choice of models used in this study. Although Random Forest, GBM, and XGBoost are powerful classifiers, there are other machine learning techniques, such as deep learning or ensemble methods that could yield better performance. These alternative methods, however, often require more computational resources and may not be as suitable for real-time prediction in industrial settings. Further research should investigate how deep learning-based approaches or hybrid models could be integrated into the proposed framework to further enhance performance.[25]

Additionally, the dataset used in this study, while representative of semiconductor production data, is still limited in scope and may not reflect all the complexities of real-world manufacturing environments. Further research is necessary to evaluate the scalability and robustness of the proposed methodology when applied to larger, more diverse datasets.

4.5 SUMMARY OF KEY FINDINGS

The results presented in this section confirm that the integrated framework, which combines automated data preprocessing, advanced imputation methods, class imbalance handling through SMOTE, and robust feature selection, significantly enhances predictive performance in semiconductor manufacturing. The use of SMOTE to address class imbalance proved particularly effective in improving defect detection, and the models trained on SMOTE-enhanced datasets achieved higher performance metrics across all evaluation metrics.[26]

However, challenges such as potential overfitting due to synthetic data, the variability in imputation performance across different missing data patterns, and the generalizability of the models remain areas for further investigation. These findings highlight the importance of addressing data quality issues in manufacturing environments and suggest avenues for future research that could lead to the development of more robust, adaptive systems for quality control and predictive maintenance.

5 CONCLUSIONS

This study has presented a comprehensive framework for improving data quality and predictive performance in semiconductor manufacturing, focusing on addressing challenges such as missing data, class imbalance, and high-dimensionality. By integrating automated ETL processes, advanced missing data imputation methods, SMOTE for class imbalance handling, and robust feature selection, we were able to develop a cohesive approach that significantly enhances predictive accuracy, particularly in the detection of rare manufacturing defects. The experimental results demonstrated that the proposed methodology outperforms traditional methods, particularly in defect detection tasks, as evidenced by the improved F1-score and recall metrics.

However, as with any applied research, several limitations and avenues for future work should be acknowledged. While the use of SMOTE improved model performance, the synthetic nature of the generated data raises concerns about potential overfitting, especially in real-world industrial settings where synthetic data may not perfectly

capture the complexity of real defects. Furthermore, the imputation methods, although effective, could be further refined to account for more complex missing data patterns, especially those arising from sensor-specific issues or non-random data loss.

Looking forward, future research could focus on refining the real-time adaptability of the proposed framework, exploring alternative techniques such as deep learning for imputation and feature extraction, and further validating the methodology across different semiconductor manufacturing environments. Additionally, integrating explainable AI into the decision-making process could provide more transparency and interpretability, which is critical for the adoption of such data-driven approaches in industrial settings. These developments hold the potential to not only enhance the scalability and robustness of the framework but also to contribute to more efficient, data-driven quality control systems in the semiconductor industry.

ACKNOWLEDGMENTS

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

FUNDING

Not applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR CONTRIBUTIONS

Not applicable.

ABOUT THE AUTHORS

YIN, Min

University of California-Berkeley, 94720, USA.

REFERENCES

[1] Sun, Y., & Ortiz, J. (2024). An ai-based system utilizing iot-enabled ambient sensors and llms for complex activity tracking. arXiv preprint arXiv:2407.02606.

[2] Huang, S. (2025). Reinforcement Learning with Reward Shaping for Last-Mile Delivery Dispatch Efficiency. European Journal of Business, Economics & Management, 1(4), 122-130.

[3] Ren, L. (2025). Leveraging Large Language Models for Anomaly Event Early Warning in Financial Systems. European Journal of AI, Computing & Informatics, 1(3), 69-76.

[4] Wang, K. J., Wang, S. M., & Yang, S. J. (2007). A resource portfolio model for equipment investment and allocation of semiconductor testing industry. European Journal of Operational Research, 179(2), 390-403.

[5] Ren, L. (2025). Causal Modeling for Fraud Detection: Enhancing Financial Security with Interpretable AI. European Journal of Business, Economics & Management, 1(4), 94-104.

[6] Chen, Y. (2025). Artificial Intelligence in Economic Applications: Stock Trading, Market Analysis, and Risk Management. Journal of Economic Theory and Business Management, 2(5), 7-14.

[7] Tian, Y., Yang, Z., Liu, C., Su, Y., Hong, Z., Gong, Z., & Xu, J. (2025). CenterMamba-SAM: Center-Prioritized Scanning and Temporal Prototypes for Brain Lesion Segmentation. arXiv preprint arXiv:2511.01243.

[8] Li, K., Chen, X., Song, T., Zhou, C., Liu, Z., Zhang, Z., Guo, J., & Shan, Q. (2025a, March 24). Solving situation puzzles with large language model and external reformulation.

[9] Luo, M., Du, B., Zhang, W., Song, T., Li, K., Zhu, H., ... & Wen, H. (2023). Fleet rebalancing for expanding shared e-Mobility systems: A multi-agent deep reinforcement learning approach. IEEE Transactions on Intelligent Transportation Systems, 24(4), 3868-3881.

[10] Chen, Y. (2025). Interpretable Automated Machine

Learning for Asset Pricing in US Capital Markets. Journal of Economic Theory and Business Management, 2(5), 15-21.

[11] Liu, Z. (2022, January 20–22). Stock volatility prediction using LightGBM based algorithm. In 2022 International Conference on Big Data, Information and Computer Network (BDICN) (pp. 283–286). IEEE.

[12] Liu, Z. (2025). Reinforcement Learning for Prompt Optimization in Language Models: A Comprehensive Survey of Methods, Representations, and Evaluation Challenges. ICCK Transactions on Emerging Topics in Artificial Intelligence, 2(4), 173-181.

[13] Wu, H., Zha, Z. J., Wen, X., Chen, Z., Liu, D., & Chen, X. (2019, October). Cross-fiber spatial-temporal co-enhanced networks for video action recognition. In Proceedings of the 27th ACM international conference on multimedia (pp. 620-628).

[14] Liu, Z. (2025). Human-AI Co-Creation: A Framework for Collaborative Design in Intelligent Systems. arXiv:2507.17774.

[15] Jin, Y., Li, Z., Zhang, C., Cao, T., Gao, Y., Jayarao, P., ... & Yin, B. (2024). Shopping mmlu: A massive multi-task online shopping benchmark for large language models. Advances in Neural Information Processing Systems, 37, 18062-18089.

[16] Wang, H., Li, Q., & Liu, Y. (2022). Regularized Buckley–James method for right - censored outcomes with block - missing multimodal covariates. Stat, 11(1), e515.

[17] Wang, H., Sun, W., & Liu, Y. (2022). Prioritizing autism risk genes using personalized graphical models estimated from single-cell rna-seq data. Journal of the American Statistical Association, 117(537), 38-51.

[18] Chen, Yinlei. "Daily Asset Pricing Based on Deep Learning: Integrating No-Arbitrage Constraints and Market Dynamics." Journal of Computer Technology and Applied Mathematics 2.6 (2026): 1-10.

[19] Ren, L. (2025). Reinforcement Learning for Prioritizing Anti-Money Laundering Case Reviews Based on Dynamic Risk Assessment. Journal of Economic Theory and Business Management, 2(5), 1-6.

[20] Pang, F. (2020, November). Research on Incentive Mechanism of Teamwork Based on Unfairness Aversion Preference Model. In 2020 2nd International Conference on Economic Management and Model Engineering (ICEMME) (pp. 944-948). IEEE.

[21] Cao S, Wang J, Tse T K T. Life - cycle cost analysis and life - cycle assessment of the second - generation benchmark building subject to typhoon wind loads in Hong Kong[J]. The Structural Design of Tall and Special Buildings, 2023, 32(11-12): e2014.

[22] Ren, L. (2025). Boosting algorithm optimization technology for ensemble learning in small sample fraud detection. Academic Journal of Engineering and Technology Science, 8(4), 53-60.

[23] Wang J, Tse K T, Li S W. Integrating the effects of climate change using representative concentration pathways into typhoon wind field in Hong Kong[C]//Proceedings of the 8th European African Conference on Wind Engineering. 2022: 20-23.

[24] Wang J, Tim K T, Li S, et al. A systematic comparison of the wind profile codifications in the Western Pacific Region[J]. Wind and Structures, 2023, 37(2): 105-115.

[25] Saxena, S., & Unruh, A. (2002). Diagnosis of semiconductor manufacturing equipment and processes. IEEE transactions on semiconductor manufacturing, 7(2), 220-232.

[26] Ditmore, D., Stewart, J., Dudley, R., & Bright, N. (1989, September). Achieving semiconductor equipment reliability. In Proceedings. Seventh IEEE/CHMT International Electronic Manufacturing Technology Symposium, (pp. 5-11). IEEE.