International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 • Website: www.ijfmr.com • Email: editor@ijfmr.com

Investigating the Prediction of Semiconductor Wafer Production Through Classification AI Models

Ahan Mathew[1], Rajat Dandekar[2]

Abstract
Semiconductor wafers are the foundation of modern technology to power almost every electronic equipment. Nonetheless, existing processes for manufacturing semiconductors are still susceptible to defects that are not identified until full production is completed, resulting in lost materials and excessive cost. We wanted to build a machine learning classification framework able to predict defective semiconductor wafers early in the manufacturing process. We speculated that the combination of oversampling methods like SMOTE with cost-sensitive machine learning methods would surpass both techniques on their own in spotting minority-class defects. On the University of California Irvine SECOM data, we compared several machine learning methods under both standard and combined configurations, including logistic regression, random forest, support vector machine (SVM), gradient boost, and XGBoost. Our combined approach improved moderate F1-scores and defective sample recall compared to control methods. Even though the resulting performance was still restricted due to excessive imbalance between classes, the findings show promise for combined AI models in improving prediction for semiconductor yield and quality assurance. Future refinements must prioritize sophisticated resampling and ensemble methods to reduce data imbalance further and increase detection rates for rare manufacturing defects.

INTRODUCTION

The rapid development of technology has made semiconductors indispensable in modern society, allowing for advances in computing, communication, and artificial intelligence. As the demand grows, the manufacturing process must achieve exceedingly high precision to limit costly defects. Nevertheless, manufacturing remains prone to variabilities, where even minor deviations can result in defective chips that are not identified until production. This challenge leads to the development of machine learning–based defect detection applications that can label defective wafers at a previous point in the manufacturing process (1,2).

In the past years, machine learning and artificial intelligence structures have shown great promise for predictive quality control applications in manufacturing. Studies by Tsaousis et al. (3) and Saini et al. (4) applied deep learning and combined methodologies to industrial process data, achieving higher defect classification accuracy. Similarly, García et al. (5) surveyed publicly available datasets and highlighted the problems caused by class imbalance, particularly in semiconductor manufacturing processes. More recent studies by the DLR Institute (6) and the IJISAE Journal (7) further showed that combined models that integrate oversampling methods and cost-sensitive learning methodologies surpassed single-model approaches in minority class detection. Overall, these findings highlight the value of combining sampling techniques like SMOTE with cost-sensitive algorithms for addressing imbalance problems and enhancing predictive ability.

While several studies have investigated such ensemble techniques broadly across various datasets, few have examined such techniques specifically in the context of semiconductor wafer datasets. We therefore conducted an experimental study on the SECOM dataset (8) to compare a number of classifiers both baseline and combined with SMOTE and cost-sensitive weights. We speculated that such combined models would have higher recall and F1-scores for minority-class (defective) wafers than baseline alone.

RESULTS

We evaluate the performance of seven machine learning techniques, i.e., baseline logistic regression, SMOTE + logistic regression, balanced logistic regression, random forest, SVM, gradient boosting, and XGBoost on the SECOM data set. Classification metrics like precision, recall, F1-score, and accuracy are reported in table 1.

Baseline logistic regression was highly accurate (0.9416) and had good recall (0.9352) for the dominating class (pass) but poorly detected defects with 0.1739 precision and 0.1905 recall for the minority class. SMOTE utilization increased minority recall by a minor degree but decreased overall precision with increased false positive rates. Balanced logistic regression was correspondingly performing, suggesting analogous effects for class weighting and oversampling.

Among the ensemble predictors, XGBoost and random forest had the highest combined overall accuracy (approximately 0.93) while recognizing no minority-class events at any point in time (zero recall). SVM and gradient boosting were relatively better, with significantly restricted recall (0.0476) while with correspondingly higher precision for defective wafers.

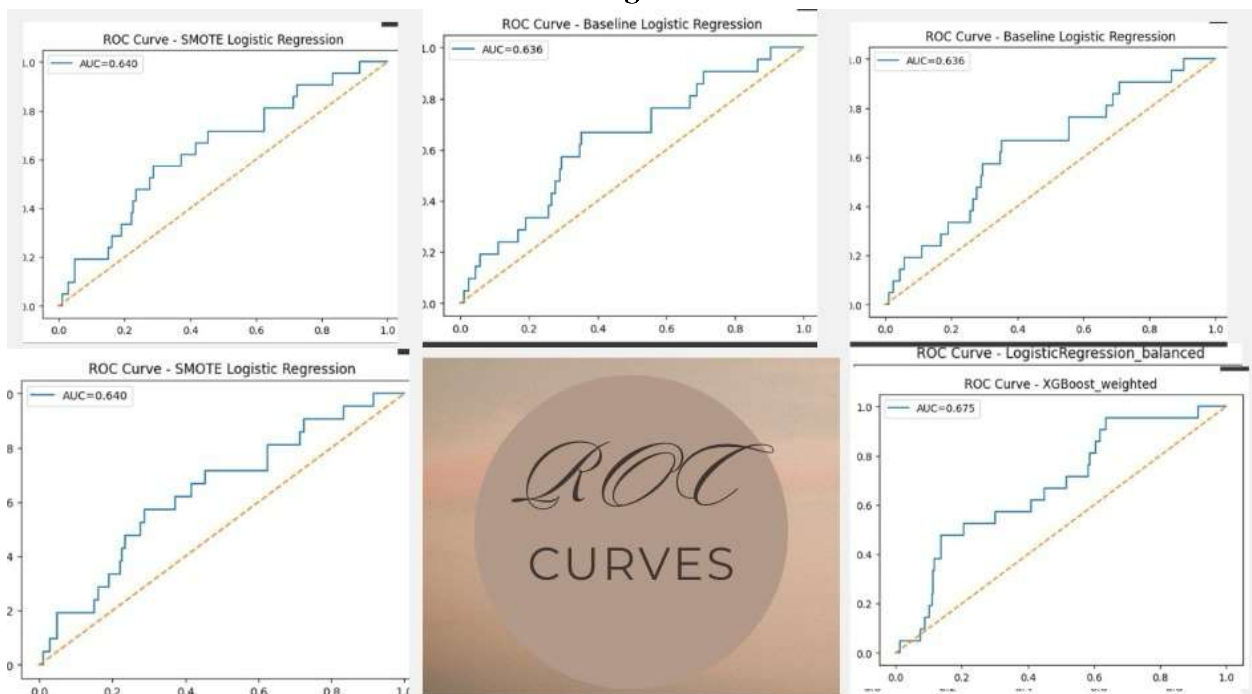
All the models performed poorly with AUC values (≈0.52–0.56) indicating little discriminability between defective and non-defective samples. Since the hybrid logistic regression methods had small increments in minority recall, no methods had workable defect detection with the imbalance inherent in the provided dataset.

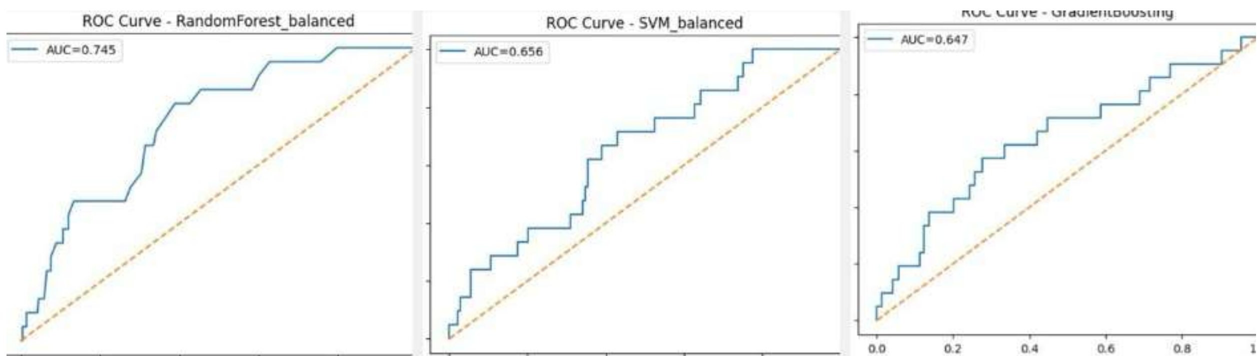
Fig 1

Model	Class	Precision	Recall	F1-score	Support	Accuracy	Macro Avg (Precision/Recall/F1)	Weighted Avg (Precision/Recall/F1)
Baseline Logistic Regression	0	0.9416	0.9352	0.9384	293	0.8854	0.5577 / 0.5628 / 0.5601	0.8902 / 0.8854 / 0.8878
	1	0.1739	0.1905	0.1818	21			
SMOTE + Logistic Regression	0	0.9388	0.8908	0.9142	293	0.8439	0.5250 / 0.5406 / 0.5273	0.8835 / 0.8439 / 0.8624
	1	0.1111	0.1905	0.1404	21			

Logistic Regression Balanced	0	0.9388	0.8908	0.9142	293	0.8439	0.5250 / 0.5406 / 0.5273	0.8835 / 0.8439 / 0.8624
	1	0.1111	0.1905	0.1404	21			
Random Forest Balanced	0	0.9331	1	0.9654	293	0.9331	0.4666 / 0.5000 / 0.4827	0.8707 / 0.9331 / 0.9008
	1	0	0	0	21			
SVM Balanced	0	0.9359	0.9966	0.9653	293	0.9331	0.7179 / 0.5221 / 0.5261	0.9067 / 0.9331 / 0.9065
	1	0.5	0.0476	0.087	21			
Gradient Boosting	0	0.9355	0.9898	0.9619	293	0.9268	0.5927 / 0.5187 / 0.5209	0.8896 / 0.9268 / 0.9029
	1	0.25	0.0476	0.08	21			
XGBoost Weighted	0	0.9329	0.9966	0.9637	293	0.9299	0.4665 / 0.4983 / 0.4818	0.8705 / 0.9299 / 0.8992
	1	0	0	0	21			

Fig 2





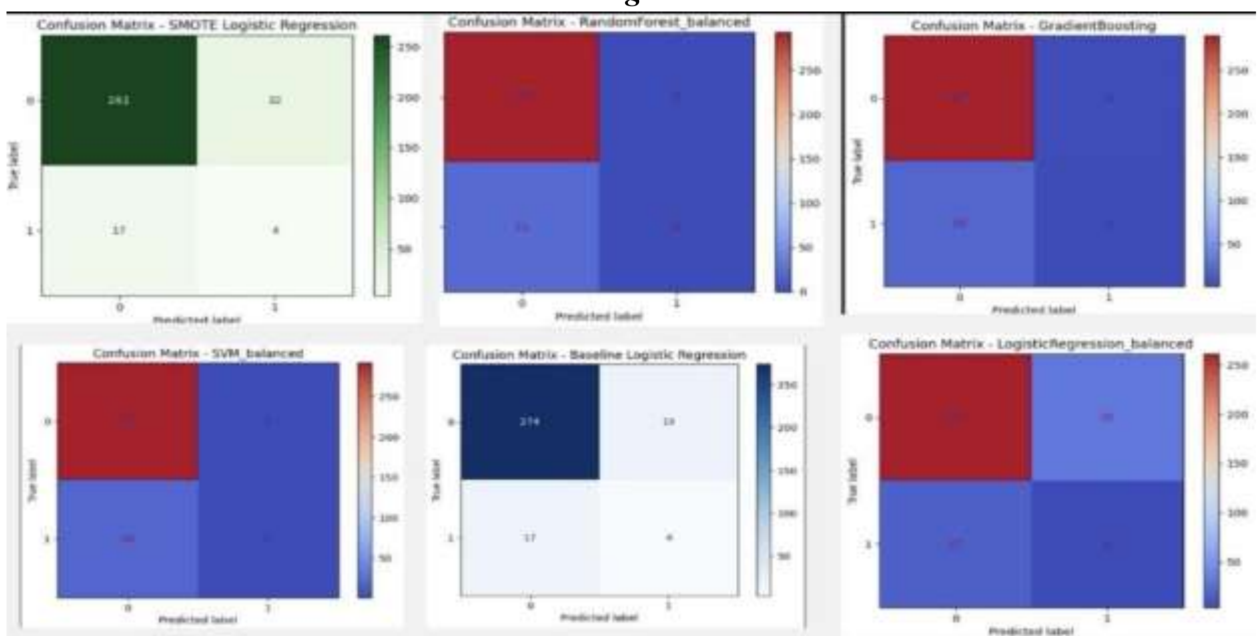
DISCUSSION

Experiment outcomes show that there is still difficulty in predicting rare manufacturing failure with this dataset that is highly imbalanced. The logistic regression base model had a significant bias in favoring the majority class, which is expected given that defective wafers formed a tiny proportion of the dataset. Use of SMOTE and weighing of classes improved the model's reactivity to minority cases, yet the same correction also caused excessive rates of false positives, thereby reducing the overall accuracy.

Combination methods like random forest and XGBoost, although not less robust than others under balanced conditions, were incapable of discovering any minority instances. This means their splitting criteria are still under control of majority patterns even during weighted training. Gradient boosting and SVM demonstrated minor gains in precision during failure but not enough recall to be viable under industrial conditions.

These are consistent with previous work highlighting the shortcomings of standard classifiers in imbalance scenarios in industrial defect prediction (2, 4, 6). Per previous manufacturing AI studies (7, 9), the performance can be improved with more advanced data-level balancing techniques (e.g., ADASYN, Borderline-SMOTE), ensemble stacking, or cost-sensitive, non-standard algorithms specifically developed for imbalance data. In key manufacturing systems where costly false negatives are involved, future work must aim at maximizing minority recall with tolerable false-positive rates.

Fig 3



Conclusion

It explored the applications of various schemes of classification, i.e., Logistic Regression, RandomForest, Support Vector Machines (SVM), Gradient Boosting, and XGBoost, for detecting manufacturing failure given the SECOM dataset. A significant challenge encountered was that the dataset was seriously imbalanced, with manufacturing failure (minority) occurrences far less than successful ones (majority).

Baseline Logistic Regression had a high overall accuracy due to the prevalence of the dominating majority class but performed badly at recognizing the minority class with a very low recall (0.1905) for failure. Using SMOTE oversampling and balanced class weights for Logistic Regression also led to comparable characteristics of their performance, with a modest increase in recall for the minority class at the cost of still high False Positive rates (poor precision).

Ensemble methods such as RandomForest and XGBoost, when trained with balanced class weights or position scale weighting with SMOTE, unfortunately did not capture a single instance of the minority class during the test set, and consequently zero recall and precision were recorded for failures. SVM and Gradient Boosting models also showed correspondingly low capability in failure detection, with low recall even at a moderate increase in precision than the Logistic Regression models. The across-the-board low AUC values for each model also reflect their poor ability to discriminate between the two classes in this imbalanced scenario.

These findings emphasize the natural challenge in directly applying typical classification schemes to severely imbalanced datasets for critical applications such as manufacturing failure prediction, wherein False Negative cost is expensive. Even though certain methods such as SMOTE and class weights were used, their capability to effectively improve minority class recall in a marked manner without overly increasing False Positives was restricted during this study.

Future Steps

From such observations, future studies must aim at overcoming the imbalance within the dataset and shortcomings of the model more efficiently:

Advanced Resampling Methods: Explore and contrast the effect of more complex oversampling methods (e.g., ADASYN, Borderline SMOTE) and undersampling methods (e.g., NearMiss, TomekLinks), possibly in combination, to create a more balanced and comprehensive training set.

Imbalance-Specific Algorithms: Describe classification algorithms that are designed specifically to work with imbalance data, like Balanced Random Forest, EasyEnsemble, or cost-sensitive learning-capable algorithms that have inherent handling for imbalance.

Hyperparameter Optimization: Perform exhaustive hyperparameter tuning for the promising models, applying methods such as cross-validation and optimization methods, keeping an eye on measures appropriate for minority class performance (e.g., minority class F1-score, G-mean).

Feature Engineering and Selection: Search the characteristics for their relevance toward predictive failure forecasting and explore techniques for feature engineering that can yield more informative characteristics. Apply feature selection methods to potentially remove noisy or redundant characteristics that are potentially detrimental to the model's performance.

Ensemble and Stacking Methods: Create ensemble or stacking methods that assemble predictions across several different models together to take advantage of their unique strengths and possibly augment the overall result on the imbalanced set.

Metrics for Evaluation: Adopt a wider set of metrics for evaluation suitable for imbalanced sets, e.g., area under the Precision-Recall curve (AUPRC) that is potentially more informative than AUC in such a situation, along with investigation on the trade-off between precision and recall depending on particular requirements for functionality and the cost due to False Positives and False Negatives incurred.

Investigating Data Characteristics: A closer analysis of the characteristics of the minority class objects and what features are discriminative for them with respect to the majority class can yield information that can be used for developing more selective modeling approaches.

By taking such future actions, the goal is to build a stronger and more consistent predictive model that can readily detect manufacturing failure, thus allowing proactive correction and enhancing manufacturing quality and effectiveness.

MATERIALS AND METHODS

Dataset Preparation

We used the public SECOM dataset at the University of California, Irvine (https://archive.ics.uci.edu/dataset/179/secom). The SECOM data have 1567 samples with 590 numeric features that are the sensor readings during the fabrication process of a semiconductor wafer. Each sample is labeled as "pass" or "fail." The data are highly imbalanced, with the failure making up less than 5% samples.

Preprocessing and Modeling

Preprocessing consisted of handling missing values along with continuous feature normalization. Models were written in Python using the module scikit-learn. We evaluated:

Logistic Regression Introduction

Synthetic Minority Over-Sampling Technique with Logistic Regression

Cost-Sensitive Classifiers:

1. logistic regression (balanced class weights)
2. random forest (balanced
3. SVM (balanced)
4. gradient boosting
5. XGBoost (balanced).

The SMOTE was applied exclusively to the training set for the construction of synthetic minority samples. Training and testing were based on 80% and 20% of information, respectively, for each version of the model.

Evaluation Metrics

The following statistical metrics were used to quantify model performance:

Precision: Number of true positive predictions is divided by the total true positives. This gives an indication of how many predicted defective samples were defective.

Recall (Sensitivity): The proportion of true positive observations that have been properly predicted. It assesses the capability of the model to detect all defective samples.

F1-Score: This is the harmonic mean of both precision and recall and gives a combined measure that encompasses both. A high F1-score indicates a good balance between precision and recall.

Support: The true events per class in the test set, to provide context to metric weighting.

Accuracy: The total ratio of samples that have been classified correctly. This is a misleading measure for imbalanced data collections as a classifier that simply predicts the majority group can seem highly accurate.

Macro Average: The unweighted mean of recall, precision, and F1 across all classes irrespective of frequency.

Weighted Average: The precision, recall, and F1-score weighted by the frequency of classes, providing a more balanced measure of performance with imbalanced data.

Confusion Matrix: A comparison visualization between predicted and actual values that depicts True Positives (accurate defect predictions), True Negatives (accurate non-defect predictions), False Positives (defects not predicted properly), and False Negatives (defects missed).

ROC Curve and AUC: ROC Curve is a plot of True Positive Rate against False Positive Rate at varying thresholds, and AUC (Area Under the Curve) is a summary measure of overall discriminative power. AUC = 1.0 is perfect discrimination, while 0.5 is random guess.

ACKNOWLEDGMENTS

[Insert acknowledgments here — e.g., mentors, institutions, or funding sources.]

REFERENCES

1. IBM Research Blog. “How AI is Improving Chip Design and Production.” *IBM Research*, 2025. https://research.ibm.com/blog/how-ai-is-improving-chip-design-and-production.
2. ArXiv Preprint. “Rare Class Prediction Model for Smart Industry in Semiconductor Manufacturing.” *arXiv*, 2024. https://arxiv.org/abs/2406.04533.
3. Tsaousis, C. et al. “AI-Based Approaches for Fault Detection in Manufacturing Systems.” *PMC*, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC11398254/.
4. Saini, R., et al. “Hybrid Learning Methods for Defect Detection in Semiconductor Fabrication.” *IJISAE*, 2024. https://ijisae.org/index.php/IJISAE/article/download/5897/4648/10980.
5. García, A., et al. “Review of Publicly Available Datasets in Manufacturing Systems.” *DLR Publications*, 2025. https://elib.dlr.de/211380/.
6. “Integration of AI and Machine Learning in Semiconductor Manufacturing for Defect Detection and Yield Improvement.” *ResearchGate*, 2024. https://www.researchgate.net/publication/382087944.
7. InPressCo Journal. “AI Techniques for Imbalanced Industrial Datasets.” *International Journal of Current Engineering and Technology*, 2025. http://inpressco.com/wp-content/uploads/2025/07/Paper6335-344.pdf.

8. UCI Machine Learning Repository. “SECOM Dataset.” *University of California, Irvine*, https://archive.ics.uci.edu/dataset/179/secom.
9. ScienceDirect. “AI for Production Control.” *Production and Operations Management*, 2024. https://www.sciencedirect.com/science/article/abs/pii/S0278612524002218.