ARTIFICIAL INTELLIGENCE FOR ENHANCED SEMICONDUCTOR MANUFACTURING: FEATURE SELECTION FOR YIELD IMPROVEMENT

Dr.Y.Ravisankaraiah[1], Ankinapalli Chandrasekhar[2], Jaladi Jeevanadh[2], Annareddy Manjunadha Reddy[2]

[1]Department of Electronics and Communication Engineering, Geethanjali Institute of Science and Technology, Nellore.

ABSTRACT

In semiconductor manufacturing, ensuring high yield rates is critical for optimizing production efficiency and minimizing costs. However, the vast number of signals collected from sensors and process measurement points often contain a mix of relevant information, noise, and irrelevant data, making it challenging for engineers to identify the key factors affecting yield. Feature selection techniques are instrumental in addressing this challenge, as they help identify the most relevant signals that significantly impact yield. In conventional semiconductor manufacturing, engineers are inundated with an extensive array of signals, making it cumbersome and time-consuming to pinpoint the critical factors influencing yield excursions. This data overload often results in suboptimal process efficiency and increased production costs. Traditional approaches may not effectively distinguish between useful information and noise, leading to inefficient troubleshooting and reduced yield rates. Additionally, manual feature selection processes are labour-intensive and may not uncover complex causal relationships between variables, limiting their effectiveness in enhancing semiconductor manufacturing operations. To overcome the limitations of the conventional approach, this study proposes the use of artificial intelligence-based feature selection techniques. By leveraging sophisticated algorithms, the proposed system will rank features according to their impact on semiconductor manufacturing yield. These techniques will not only streamline the identification of crucial variables but also unveil causal relationships within the data, providing a deeper understanding of the production process. The application of cross-validation ensures robustness and reliable evaluation of feature relevance for predictability using error rates. The goal is to empower engineers with a more efficient and data-driven approach to semiconductor manufacturing, resulting in increased yield rates, reduced production costs, and shorter learning cycles. The preliminary results presented here demonstrate the potential of this approach, highlighting the promise of artificial intelligence in revolutionizing semiconductor manufacturing optimization.

Keywords: Semiconductor Manufacturing, Yield Improvement, Signal Processing, Process Efficiency, Production Costs, Data Overload, Troubleshooting.

INTRODUCTION

1.1 OVERVIEW

Manufacturing processes do not always generate out comes that meet the desired quality specifications. Poor quality creates unnecessary scrap and rework, thereby having a substantial negative impact on financial performance. Moreover, it can lower schedule ad hearence, increase inventory levels, and make other improvement opportunities less apparent (Ittner 1994). The American Society for Quality estimates that poor quality generates 10%–15% of the operating expenses in manufacturing companies. Motivated by such figures, manufacturers continuously seek to improve their process performance. To that end, quality management theory suggests to identify and eliminate sources of quality variation (Taguchi 1986, Schmenner and Swink 1998, Zantac et al. 2002, Field and Sinha 2005, Hoppe and Spearman 2011)

Quality improvement has long been supported by statistical methods (e.g., Shewhart 1926). Existing approaches for identifying sources of quality variation focus on linear associations. However, modern manufacturing settings are characterized by high-dimensional data (Kusiak 2017), which frequently involve nonlinear relationships. For instance, in semiconductor fabrication, manufacturers can collect several thousand interrelated measurements for each individual product unit. When neglecting nonlinearities under high-dimensional conditions, manufacturers may not identify important drivers of process quality. Consequently, there is a need for methods that better accommodate nonlinearities in manufacturing data. semiconductor manufacturing, engineers are inundated with an extensive array of signals, making it cumbersome and time-consuming to pinpoint the critical factors influencing yield excursions. This data overload often results in suboptimal process efficiency and increased production costs. Traditional approaches may not effectively distinguish between useful information and noise, leading to inefficient troubleshooting and reduced yield rates. Additionally, manual feature selection processes are labour intensive and may not uncover complex causal relationships between variables, limiting their effectiveness in enhancing semiconductor manufacturing operations.

2.2 RELATIVE WORK

Lee et. al [1] proposed a MXene-coated ceramic nanofiltration (NF) membranes significantly improved the removal efficiency of representative organic contaminants in semiconductor wastewater when they compared to the pristine ceramic membrane. It enhanced the removal efficiencies by 4.2 times for dissolved silica, 3.7 times for dimethyl sulfoxide (DMSO), and 2.5 times for isopropyl alcohol (IPA). Furthermore, the MXene ceramic NF membrane exhibited exceptional fluoride removal capability, achieving approximately 99.1% removal, in addition to organic solvents. Those findings not only help meet regulatory standards for wastewater discharge but also provided valuable insights into

the transport behaviours of contaminants and their separation mechanisms. That innovative approach brought one step closer to efficient and sustainable solutions for semiconductor wastewater treatment.

Zhao et.al [2] proposed a design framework that can fully eliminate cold-heat offset, simultaneously reduce cooling/heating loads and enhance cooling efficiency under full-range semiconductor applications. By detailed modelling and simulations, their proposed design framework was validated and tested under various indoor cooling loads, ventilation rates, and surrounding weather and climate conditions. Results show that 2.3–33.1 % energy savings are achieved and up to 15.8GJ/m2 annual primary energy was saved, compared with the conventional design. It was also observed that cities in cold and mild climates have higher energy-saving potentials than those in hot climates.

Eunseo et.al [3] proposed a study that modelled missing values estimated using Gaussian progress regression by considering those fine movements and noises of data using a quantum mechanics-based stochastic differential equation and corrected them using Ito's lemma. Estimating the missing data in this manner more closely simulates the attributes of data compared to existing estimation methods, enabling more accurate analysis. To demonstrate the excellence of the proposed framework, missing values was estimated using vehicle operation data and semiconductor manufacturing data with multiple missing values, and the estimation results was compared with those of existing algorithms.

Hogg et.al [4] designed to use epitaxial regrowth to enhance semiconductor laser performance and unlock new markets? First, we need to take a closer look at the relatively mature, epitaxial regrowth process for indium phosphide (InP) lasers, commonly used in 5G, datacomms, telecoms and co-located optics applications. Tough we can consider gallium arsenide (GaAs), a less mature material system, which suits a multitude of emerging industrial, biomedical imaging, datacomms, and sensing laser applications.

Xuan et.al [5] proposed domination criterion it was a effective inequality, which can deterministically optimize the objective value of a partial sequence even if the scheduling sequence of subsequent jobs is unknown. In the constructive solution method, a set covering model is designed to capture those effective factory allocation patterns for the groups hidden in the historical solutions to speed up the search for the IG. The comprehensive experiments on 810 test instances demonstrate the effectiveness of HIG.

Chao et.al [6] implemented in a standard Taiwan Semiconductor Manufacturing Company 28 nm process occupies a core area of 0.4 mm2 and generates a frequency ranging from 4 to 5.2 GHz using 50 MHz oscillator input, the power consumption is 40 mW. It achieves 74.8 fs root-mean-square jitter integrated from 10 KHz to 30 MHz. When the PLL output is divided by 2, the output phase noise at 1 MHz frequency offset is −125.75 dBc/Hz, and the PLL FoM is −246.

Saini et.al [7] was Considering these loopholes, the main impetus of the current review is the analysis of the patents granted in nanotechnology and its associated domains, by USPTO during a definite span of 2016-2021. Notably, 50.62% of nanotechnology patent's share in USPTO belongs to US investors and companies, indicating a kind of 'home advantage'. Considering all those stats and facts, the assessment of USPTO nanotechnology patents might help to guide policies and strategic recommendations. Alongside, the SWOT analysis of nanotechnology patents of USPTO predicts the unseen future threats and opportunities, and measures to prevent technological weakness, enhancing development.

Haider et.al [8] proposed for specificity and for facilitating concrete analysis during the still unfolding fourth industrial revolution, digital technologies based on semiconductor material foundation and the development of Artificial Intelligence (AI) are analyzed for China. Our framework of combining efficiency and equity in specific ways used for such concrete applications can be called socially embedded capabilities enhancing national innovation system or SECENIS. Those Chinese SECENIS that was being built for the 21st century has important regional and geoeconomic implications for the future. Those Chinese path shows that innovation systems are critical in building a region such as the New Eurasia. Russia is also trying to follow such a path. Other (East) Eurasian countries will need to follow such a strategy if they are to benefit from the New Eurasian and expanded BRICS arrangements

3. PROPOSED METHODOLOGY

3.1 OVERVIEW

The semiconductor industry stands at the forefront of technological advancement, powering innovations across various sectors. However, optimizing semiconductor manufacturing processes presents formidable challenges, particularly in ensuring high yield rates while minimizing costs. Traditional approaches often struggle to cope with the vast amount of data generated during manufacturing, leading to inefficiencies and suboptimal outcomes. In response to these challenges, a proposed solution leveraging artificial intelligence (AI)-based feature selection techniques offers a promising avenue for revolutionizing semiconductor manufacturing optimization. The proposed

solution addresses the fundamental issue of data overload by employing sophisticated AI algorithms to streamline feature selection. These algorithms sift through the myriad of signals collected from sensors and process measurement points, identifying the most relevant factors influencing yield amidst noise and irrelevant data. By automating this process, engineers can focus their efforts on interpreting results and implementing targeted optimizations, rather than being bogged down by manual data analysis. Central to the effectiveness of the proposed solution is its ability to uncover complex causal relationships within the data. Traditional methods often struggle to discern these intricate connections, limiting their effectiveness in enhancing manufacturing operations. By leveraging AI algorithms capable of identifying and interpreting these relationships, the proposed system provides engineers with a deeper understanding of the production process. This deeper insight enables more informed decision-making and facilitates the implementation of targeted improvements to enhance yield rates and reduce production costs.The implementation of the proposed solution is facilitated through a user-friendly graphical interface (GUI), developed using Python's Tkinter library. The GUI allows engineers to easily upload datasets, preprocess data, and apply various techniques, such as Synthetic Minority Over-sampling Technique (SMOTE) for class balancing. Additionally, the GUI provides options for selecting and evaluating different classifiers, enabling engineers to assess model performance and make informed decisions regarding model selection and deployment. Two classifiers are integrated into the system: Multilayer Perceptron (MLP) and Random Forest. These classifiers are trained on preprocessed data to predict manufacturing outcomes and evaluate performance metrics such as precision, recall, F1-score, and accuracy. By providing engineers with the tools to assess classifier performance, the system enables them to make data-driven decisions and identify areas for further optimization.The system allows engineers to make predictions based on the trained models and visualize performance comparisons through bar graphs. This functionality enhances the interpretability of results, enabling engineers to gain insights into model performance and identify potential areas for improvement. Additionally, the system's ability to visualize performance comparisons facilitates communication and collaboration among team members, fostering a culture of continuous improvement and innovation.The proposed AI-based feature selection approach offers a comprehensive solution to enhance semiconductor manufacturing operations. By leveraging AI algorithms to streamline feature selection and uncover complex causal relationships within the data, the system empowers engineers with the tools and insights needed to optimize manufacturing processes. Through a user-friendly GUI and integrated classifiers, the system facilitates data-driven decision-making and enables engineers to drive improvements in yield rates, production costs, and overall efficiency. By embracing this innovative approach, semiconductor manufacturers can position themselves for success in an increasingly competitive and dynamic industry landscape.

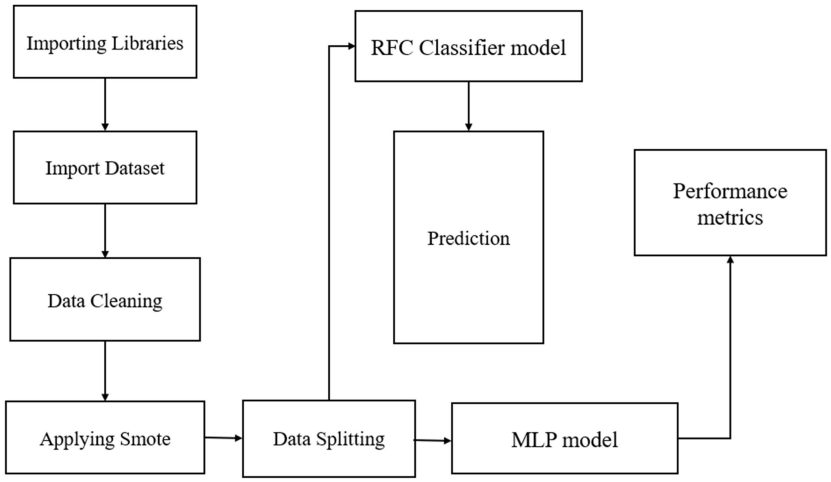

Figure 1: Block Diagram of Proposed System.

3.2 Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

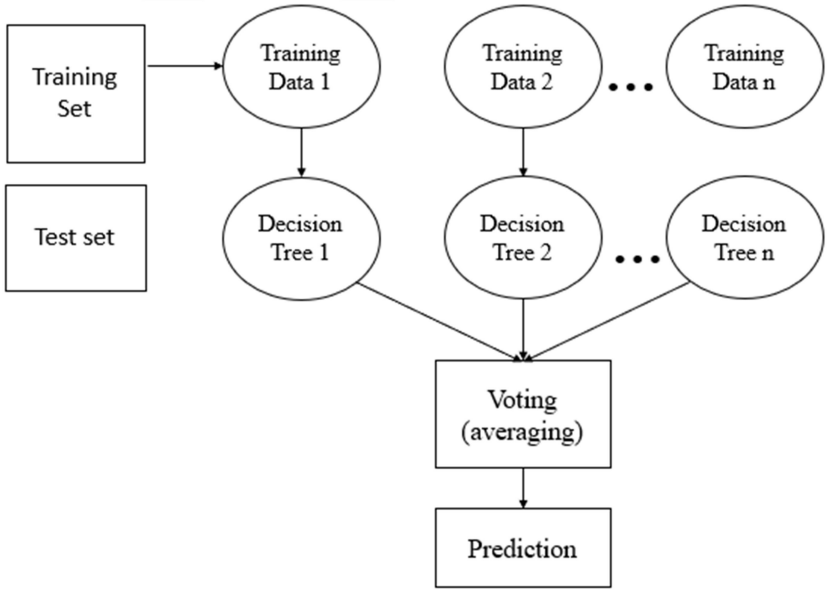

Figure: Random Forest algorithm

Random Forest algorithm

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

Important Features of Random Forest

- **Diversity**- Not all attributes/variables/features are considered while making an individual tree, each tree is different.
- **Immune to the curse of dimensionality**- Since each tree does not consider all the features, the feature space is reduced.
- **Parallelization**-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.
- **Train-Test split**- In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
- **Stability**- Stability arises because the result is based on majority voting/ averaging.

3.3.1 Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Below are some points that explain why we should use the Random Forest algorithm

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

Types of Ensembles

Before understanding the working of the random forest, we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model. Ensemble uses two types of methods:

Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest. Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.

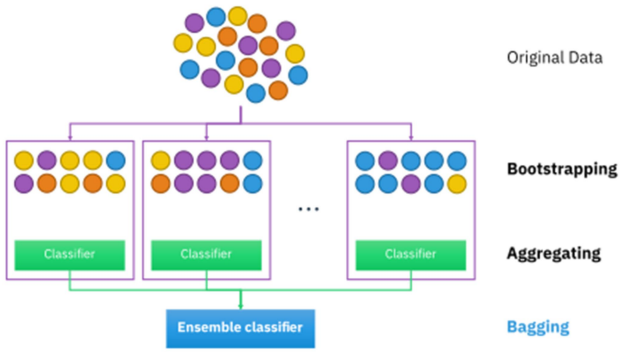

Figure. RF Classifier analysis.

Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

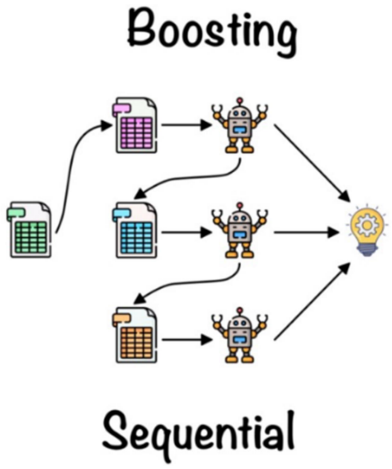

Figure. Boosting RF Classifier.

3.3.2 ADVANTAGES OF RANDOM FOREST CLASSIFIER

High Accuracy: Random Forest Classifier typically yields high accuracy compared to many other classification algorithms. It is less prone to overfitting, especially when the number of trees (n_estimators) is large.

Robust to Overfitting: Random Forest Classifier constructs multiple decision trees during training and combines their predictions through averaging or voting. This ensemble approach helps to reduce overfitting by averaging out biases and errors present in individual trees.

Handles Missing Values and Outliers: Random Forest Classifier can handle missing values and outliers effectively. It does not require data preprocessing like imputation of missing values or outlier removal, making it convenient for datasets with incomplete or noisy data.

Non-Parametric and Versatile: Random Forest Classifier is a non-parametric algorithm, meaning it makes no assumptions about the distribution of the data. It can handle both numerical and categorical features without the need for feature scaling or transformation.

Implicit Feature Selection: Random Forest Classifier performs implicit feature selection by ranking features based on their importance in reducing impurity (e.g., Gini impurity or entropy) across all decision trees. This helps in identifying the most relevant features for the classification task.

Efficient on Large Datasets: Random Forest Classifier is computationally efficient and can handle large datasets with many features and instances. It can be parallelized easily, allowing for efficient training on multicore processors or distributed computing platforms.

Estimation of Feature Importance: Random Forest Classifier provides a measure of feature importance, which can be useful for feature selection and understanding the underlying relationships between features and the target variable.

Less Sensitive to Noise: Random Forest Classifier is less sensitive to noise and outliers compared to individual decision trees. Since it aggregates predictions from multiple trees, it can filter out noise and focus on the most prevalent patterns in the data.

4. RESULT AND DISCUSSION

Figure 1: shows a graphical user interface (GUI) designed for uploading datasets related to semiconductor manufacturing. Users interact with this interface to load their dataset into the system for further analysis and processing. Figure 2: presumably displays the structure or initial state of the dataset before the Synthetic Minority Over-sampling Technique (SMOTE) is applied. The dimensions of the dataset are provided, indicating the number of rows and columns.

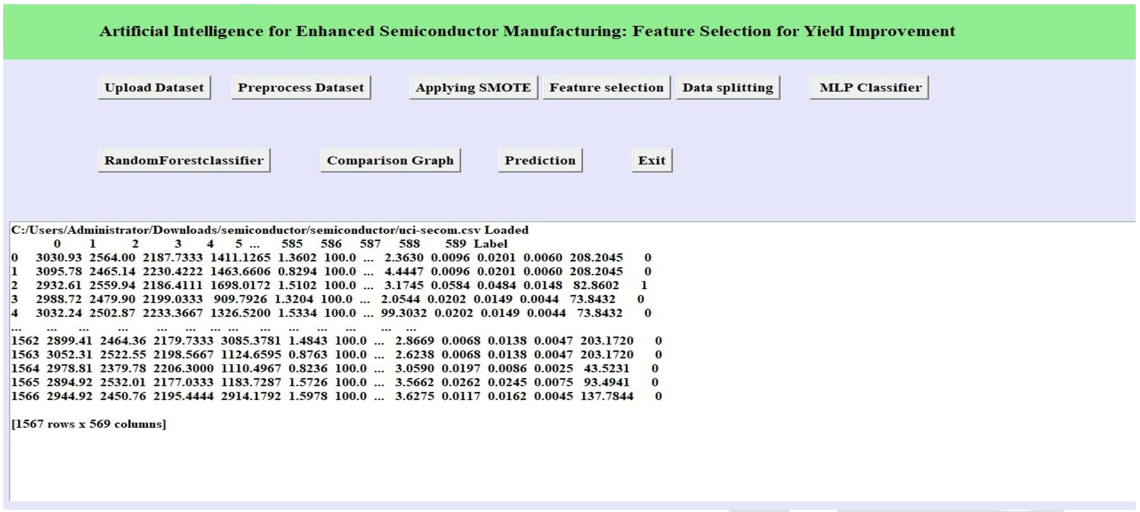

Figure 1: Uploding Dataset in the Semiconductor Manufacturing GUI.

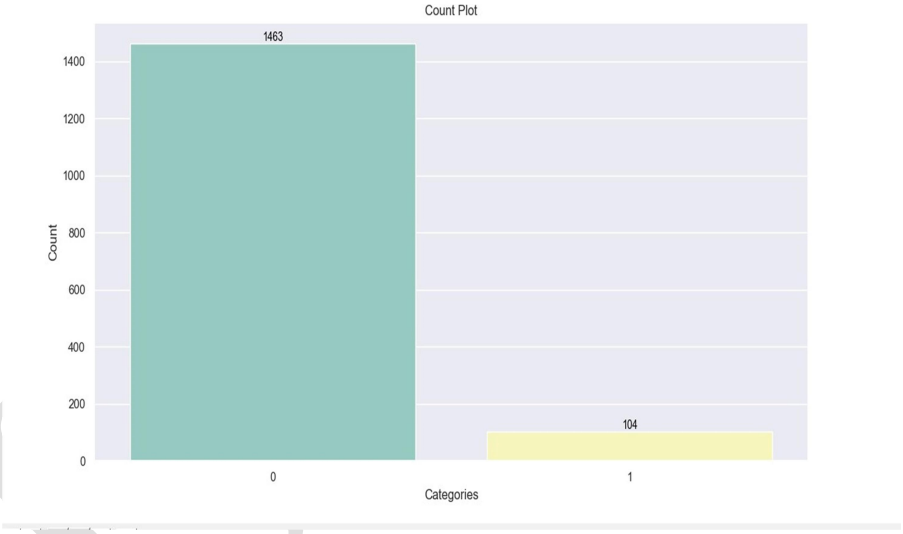

Figure 2: dataset before applying smote [1569 rows x 569 columns].

Figure 3: probably illustrates a count plot, which visualizes the distribution of classes in the dataset after the application of SMOTE. SMOTE is a technique commonly used to address class imbalance by generating synthetic samples of the minority class, thereby balancing the class distribution. Figure 4: represents the application of Principal Component Analysis (PCA) to reduce the dimensionality of the dataset. PCA is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional space while preserving important information.

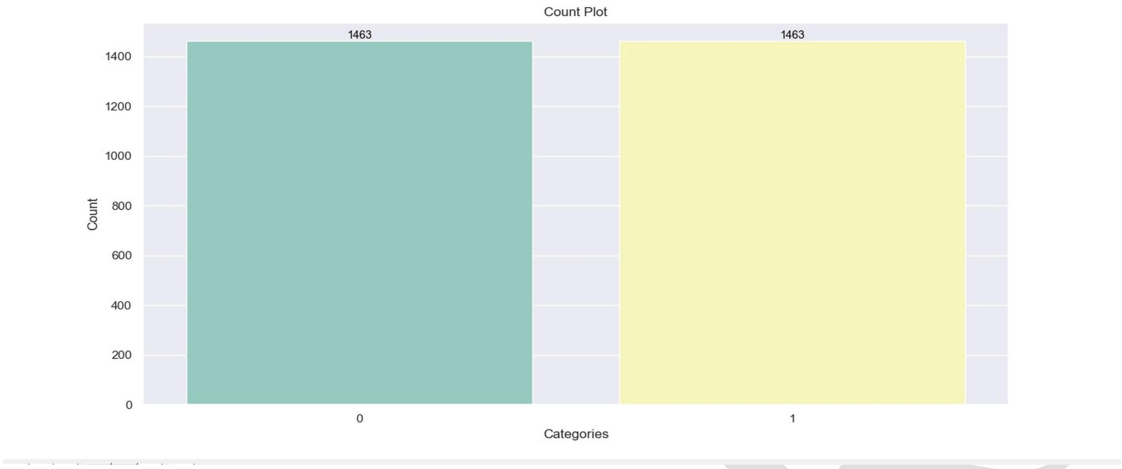

Figure 3: Count Plot after applying SMOTE.

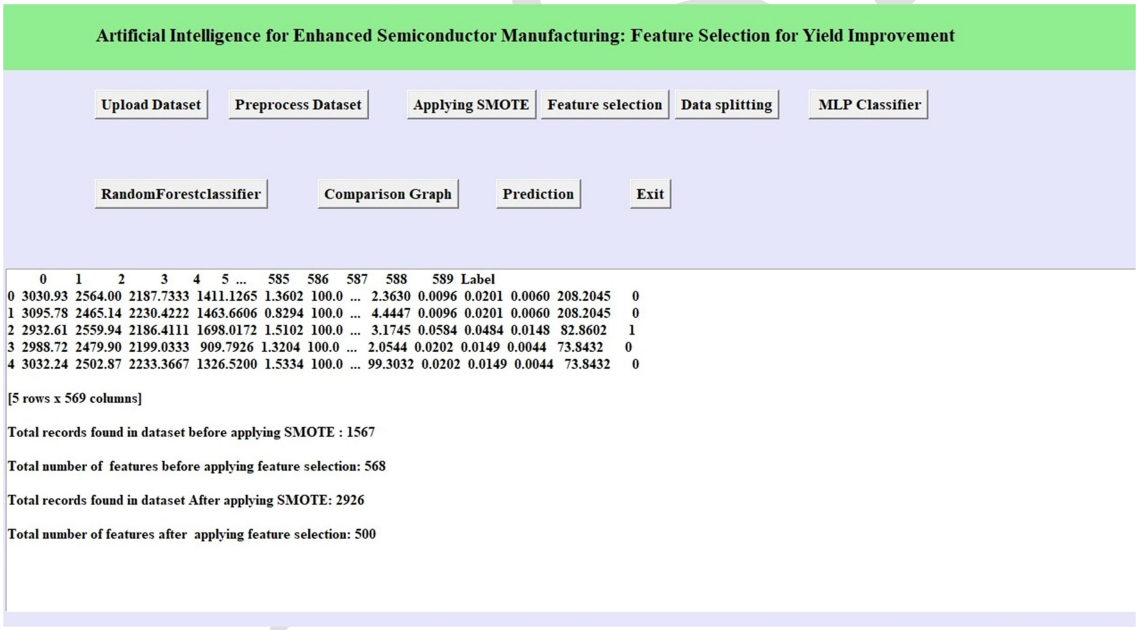

Figure 4: Applied PCA to reduce the dimensionality of the dataset.

Figure 5: depicts the process of splitting the dataset into training and testing sets. In machine learning, it is crucial to separate the dataset into these subsets to evaluate the performance of models on unseen data. Figure 6: figure displays the confusion matrix, a performance evaluation metric, specifically for a Multi-Layer Perceptron (MLP) Classifier. A confusion matrix provides insight into the classifier's true positive, false positive, true negative, and false negative predictions.

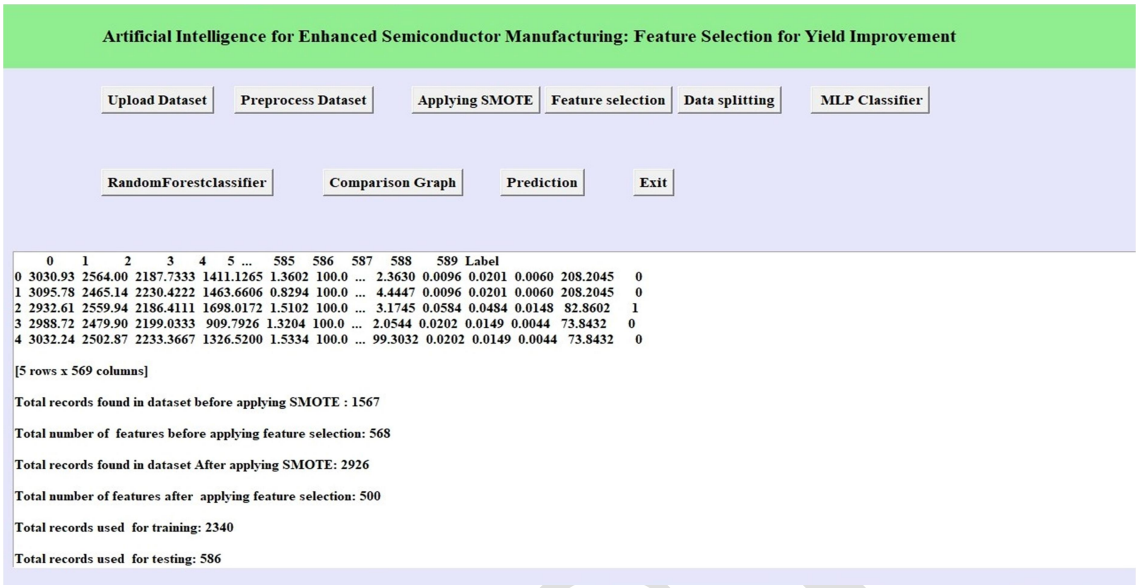

Figure 5: Splits the dataset into training and testing sets

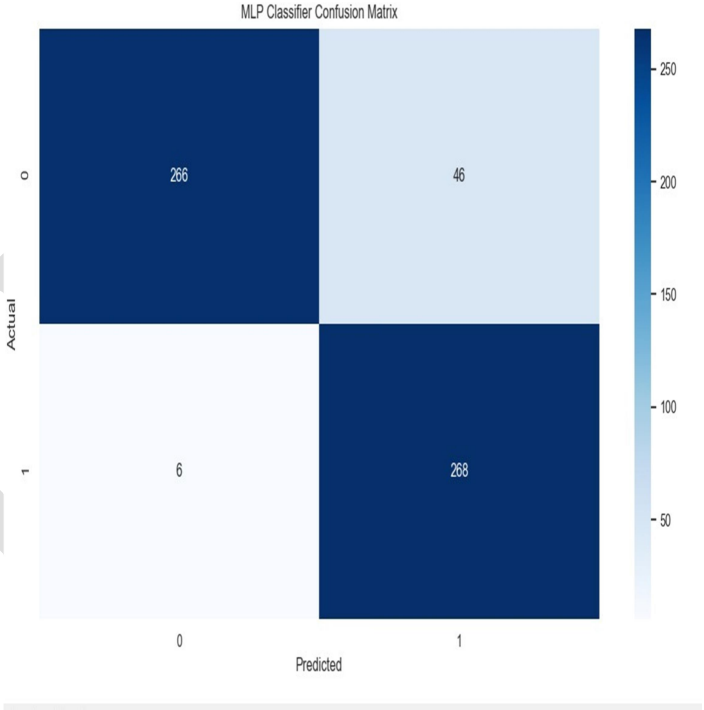

Figure 6: confusion matrix of MLP Classifier.

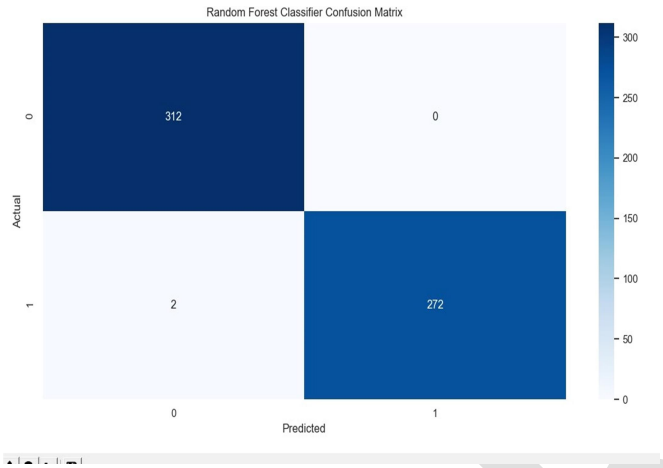

Figure 7: Confusion matrix of RFC model.

Figure 7: shows the confusion matrix for a Random Forest Classifier (RFC) model. Similar to Figure 6, this confusion matrix provides a detailed breakdown of the RFC model's predictions across different classes. Figure 8: presents a graphical comparison of the performances between the MLP Classifier and RFC model. The comparison includes metrics such as precision, recall, F-measure, and accuracy, illustrating how these models perform relative to each other.

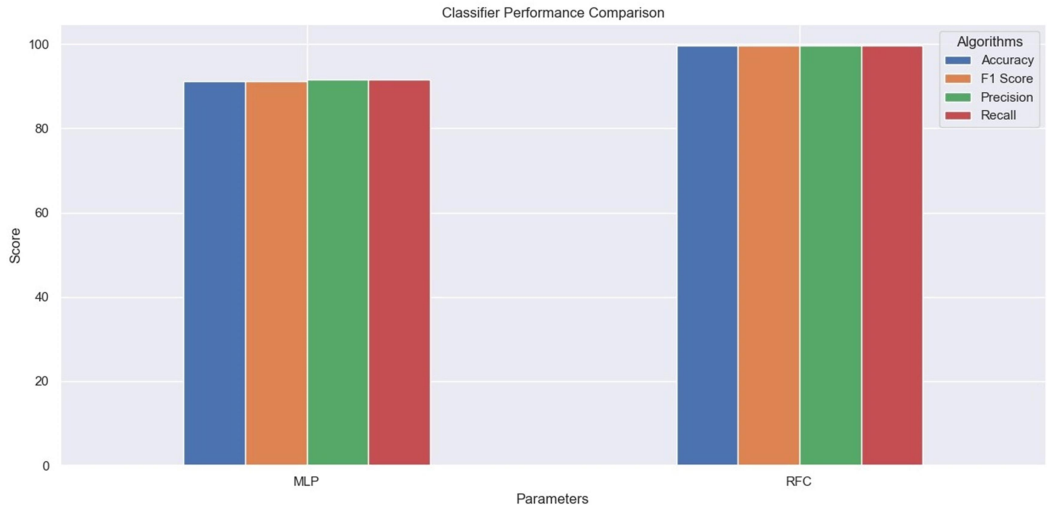

Figure 8: Performances Comparison Graph Between MLP And RFC.

Table 1: Performance comparison of quality metrics obtained using MLP Classifier model and random forest classifier (RFC) model.

Model	Accuracy	Precision	Recall	F1 score

MLP model	91.80	92.82	91.80	91.98
RF model	99	99	99	99

Table 1: Performance description for all Models.

1. **Precision**: Precision is a measure of the accuracy of the positive predictions made by the classifier. It is calculated as the ratio of true positive predictions to the sum of true positive and false positive predictions. For the MLP Classifier, the precision is 92.82%, indicating that approximately 90.82% of the positive predictions made by this classifier are correct. For the Random Forest Classifier, the precision is higher at 99.27%
2. **Recall**: Recall, also known as sensitivity or true positive rate, measures the proportion of actual positives that were correctly identified by the classifier. It is calculated as the ratio of true positive predictions to the sum of true positive and false negative predictions. The MLP Classifier has a recall of 91.80%, meaning that it correctly identifies approximately 91.80% of the actual positive instances. The Random Forest Classifier has a slightly higher recall of 99 %.
3. **F-Measure**: The F-Measure, also called the F1 score, is the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall. It is calculated as 2 * (precision * recall) / (precision + recall). The MLP Classifier has an F-Measure of 91.98%, while the Random Forest Classifier has an F-Measure of 99%.
4. **Accuracy**: Accuracy is a measure of the overall correctness of the classifier's predictions. It is calculated as the ratio of the number of correct predictions to the total number of predictions. Both classifiers have high accuracy scores, with the MLP Classifier achieving an accuracy of 91.76% and the Random Forest Classifier achieving an accuracy of 99%.

5.CONCLUSION

In semiconductor manufacturing, optimizing production efficiency and minimizing costs hinge on maintaining high yield rates. However, the overwhelming volume of signals collected from sensors and process measurement points poses a challenge in identifying key factors affecting yield. Traditional approaches are often inefficient, as they struggle to discern relevant information from noise, leading to suboptimal process efficiency and increased production costs. This study proposes the adoption of artificial intelligence-based feature selection techniques to address these challenges. By leveraging sophisticated algorithms, the proposed system effectively identifies the most relevant signals impacting semiconductor manufacturing yield. Moreover, it uncovers complex causal relationships within the data, providing engineers with a deeper understanding of the production process. The application of cross-validation ensures the robustness and reliability of feature relevance evaluation, enhancing predictability using error rates. By empowering engineers with a more efficient

and data-driven approach, this system aims to increase yield rates, reduce production costs, and shorten learning cycles in semiconductor manufacturing.

REFERENCES

[1] Lee, Yoojin, Jihyeon Lee, Yeon So, Soyoun Kim, and Chanhyuk Park. "Mxene-based ceramic nanofiltration membranes for selective separation of primary contaminants in semiconductor wastewater." *Separation and Purification Technology* 331 (2024): 125653.

[2] Zhao, Wenxuan, Hangxin Li, and Shengwei Wang. "A generic design optimization framework for semiconductor cleanroom air-conditioning systems integrating heat recovery and free cooling for enhanced energy performance." Energy 286 (2024): 129600.

[3] Oh, Eunseo, and Hyunsoo Lee. "Quantum mechanics-based missing value estimation framework for industrial data." Expert Systems with Applications 236 (2024): 121385.

[4] Hogg, Richard. "Epitaxial regrowth: How GaAs is the key to unlocking new semiconductor laser markets." PhotonicsViews 21, no. 1 (2024): 53-55.

[5] He, Xuan, Quan-Ke Pan, Liang Gao, Janis S. Neufeld, and Jatinder ND Gupta. "Historical information based iterated greedy algorithm for distributed flowshop group scheduling problem with sequence-dependent setup times." Omega 123 (2024): 102997.

[6] Zhao, Chao, Hao Zhang, Youming Zhang, Xusheng Tang, and Chengyu Pan. "A 4–5.2 GHz PLL with 74.8 fs RMS jitter in 28 nm for RF Sampling Transceiver application." Microwave and Optical Technology Letters 66, no. 1 (2024): e33889.

[7] Saini, Neha, Prem Pandey, Amit Kumar Tiwari, and Atul Kulkarni. "A framework of evolution and potential impact of nanotechnology in USPTO: the SWOT analysis." International Journal of Intellectual Property Management 14, no. 1 (2024): 1-16.

[8] Srinivasarao, G., Penchaliah, U., Devadasu, G. et al. Deep learning based condition monitoring of road traffic for enhanced transportation routing. J Transp Secur 17, 8 (2024). https://doi.org/10.1007/s12198-023-00271-3

[9] Khan, Haider. "Geoeconomics of a New Eurasia during the Fourth Industrial Revolution: The Role of China's Innovation System, BRI and Sanctions from the Global North." (2024)