National College of Ireland

Data Mining for Enhancing Silicon Wafer Fabrication

MSc Research Project
Data Analytics

Omkar Doke
Student ID: x18179525

School of Computing
National College of Ireland

Supervisor: Dr. Muhammad Iqbal

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Omkar Doke
Student ID:	x18179525
Programme:	Data Analytics
Year:	2020
Module:	MSc Research Project
Supervisor:	Dr. Muhammad Iqbal
Submission Due Date:	17/08/2020
Project Title:	Data Mining for Enhancing Silicon Wafer Fabrication
Word Count:	7028
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	22nd September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Data Mining for Enhancing Silicon Wafer Fabrication

Omkar Doke
x18179525

Abstract

Gordon E. Moore found that density of transistors doubled every two years on a microchip. However, now it is doubling in every 18 months[1] thereby making semiconductor manufacturing one of the most complicated technological process. With increasing density, the transistor dimension is reducing thereby demanding rigorous physical and electrical testing to ensure high die yield quality which majorly depends on smooth functioning of equipment's. In the past, various research projects were undertaken on wafer image data in semiconductor manufacturing field to improve the quality and productivity by reducing the impact of contaminants. This research aims at using machine learning techniques on numerical data obtained from sensors in equipment's to predict wafer failure in manufacturing process thereby reducing equipment failure by providing timely maintenance (i.e. predictive maintenance) which in turn would enhance productivity and improve die yield quality. To achieve this, models like XGBoost, Decision Tree, Logistic Regression, Support Vector Machine, Random Forest, K-Nearest Neighbor and Neural Network are used for classification. Various case studies were conducted wherein these models were evaluated for their performance based on their accuracy and precision. Random Forest outperformed all other models with both accuracy and precision over 98% thereby confirming that machine learning techniques can be used to implement predictive maintenance in production line with an aim to improve the productivity by making optimum use of equipment's.

Keywords: Semiconductor Manufacturing, Die Yield Quality, Contaminants, Predictive Maintenance

1 Introduction

1.1 An Overview of Semiconductor evolution

The first point contact transistor was invented in 1947 by Bardeen and Brattain at Bell Laboratories in the US followed by invention of junction transistor by Shockley in 1948 which marked the beginning of transistor era[2]. Prior to that computers were made using vacuum tubes which required huge storage space, consumed lot of electricity and generated tremendous amount of heat. But with the invention and development of semiconductors, computers have seen exponential development. World has moved from handicraft age to big-data age and this swift evolution was possible because of constant

[1] https://www.investopedia.com/terms/m/mooreslaw.asp

[2] https://www.hitachihightech.com/global/products/device/semiconductor/history.html:text=History%20of%20semiconductors,the%20junction%20transistor%20in%201948

architectural framework provided by semiconductor industry as even in this big-data age, large amount of data generated on daily basis is stored in small chips itself. This shows how the capacity and functioning (i.e. quality) of integrated chips has enhanced over a period.

1.2 Motivation

Due to overall development in technological world, electronic gadgets are finding their applications in every aspects of life. The commercial and personal use of all devices has increased tremendously which in turn has increased the demand for semiconductors as they form the backbone of any hardware system. From the time of their birth till now, primary focus in the field of semiconductor manufacturing was on improving their quality. Companies trying to enhance the productivity using same old techniques of 24*7 use of production line with bigger manufacturing plants, has made costly equipment's/machines of production line vulnerable to malfunctioning, damage, increasing the cycle time (CT) etc. This means that even if productivity can be enhanced using the basic techniques, it would increase the production cost as maintaining these equipment's is costly and building a big production plant also requires huge funding thereby reducing the profit as well as making the final product expensive for consumer market. Thus, the primary bottleneck for semiconductor manufacturing industries is to enhance the productivity without affecting the quality and hampering the production cost. As per the statistics, wafer fabrication equipment's undergo 15% of downtime of which 8% is because of unscheduled maintenance and 7% is due to scheduled maintenance. Apart from that, the equipment setting up takes around 27% of total time and for another 14% of overall time, equipment's sit idle (Lizaranzu and Rojo; 2012).

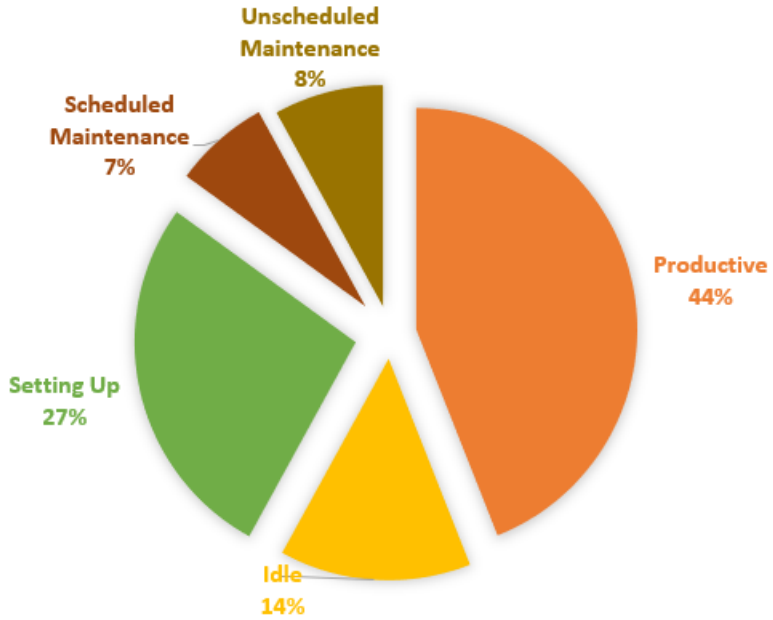

Figure 1: Equipment Utilization

Figure 1 shows that wafer fabrication equipment's are productive only for 44% of overall time. Thus, to improve the productivity, industries need to work on reducing the

unscheduled maintenance thereby making equipment's/machines productive for 52% of total time. This unscheduled maintenance is mainly caused due to lack of equipment's health tracking which is nearly impossible for engineers to do manually. Machine learning models can be implemented here to minimize the impact of unscheduled maintenance by analyzing the data generated from the sensors mounted in the equipment's of production line to predict which equipment/machine is going to need maintenance (i.e. predictive maintenance) by predicting whether the fabricated wafer using that equipment would pass or fail the final test (Khakifirooz et al.; 2018). This would increase the productivity as maintenance time is reduced and due to predictive maintenance, operational cost would reduce thus increasing the profit margin.

1.3 Research Question

"Can Machine Learning models be trained using data mining tools to predict the test result of wafer produced thereby predicting if there is need of maintenance for any equipment on production line?"
The motive of this research is to answer the above question which would help in gaining useful knowledge benefiting semiconductor manufacturing industries. Since the target variable to be predicted in this research is whether the wafer produced on a particular production line is going to pass or fail the tests (i.e. a binary variable), classification algorithms are used while implementing wafer test prediction model.

1.4 Research Objective

To meet the above research question, following objectives are defined:

- Data-preprocessing of two data-sets with wafer fabrication data recorded from sensors mounted in equipment's of production line.
- Normalization of data for transformation and scaling of features.
- Extracting best features contributing to wafer test prediction model using Principal Component Analysis (PCA).
- Performing feature selection using Analysis of Variance (ANOVA) technique.
- Merging of two pre-processed and feature engineered data-sets.
- Handling class imbalance through various experiments.
- Implementing Machine Learning models on final data-set.
- Cross validation of models using K fold technique.

1.5 Contribution

Major contribution of this research is in reducing production cost by application of predictive maintenance. This technique increases the productivity without impacting the quality of wafers thus improving the die yield. Implementing predictive maintenance would avoid sudden failure of equipment's and would reduce the time taken for manual tracking of equipment's by engineers which can be then utilized in an efficient manner to

accomplish other objectives.

Further, the paper is organized as follows, the review of related work is presented in Section 2, proposed approach for equipment fault detection is explained in Section 3, design flow of implementation is discussed in Section 4, machine learning model implementation is illustrated in Section 5, evaluation of obtained results is illustrated in Section 6, finally Section 7 concludes the expected findings from the paper.

2 Related Work

2.1 Improving Manufacturing Process Efficiency using IoT and Cloud

Availability of abundant data from production line in this big data era has made it easy to extract quality information and understand important aspects of process using the available data. CT is defined as the time taken to process a wafer lot from start to end. It is an important aspect to be considered while increasing productivity as it emphasis on efficient use of production lines to maintain the required quality standards. Wang et al. (2020) in their research proposed a feature selection method which would select all those features from big data which explain fluctuations in CT and can be used to reduce CT. This method was tested on the data-set from Singapore based Silicon Wafer Fabrication System (SWFS) which provided prediction accuracy of more than 60%. However, they did not take into consideration the correlation between various parameters that significantly contributes in controlling the CT.
Another research with an aim of CT reduction was conducted by (Wang and Zhang; 2016) wherein they designed Big Data Analytics technique (BDA) to improve production line reliability by detecting the failures using Hadoop. Although, this technique was effective, but it worked well with a limited number of features. This technique thus needs fine tuning to incorporate the factors impacting wafer production based of their correlation matrix. Reducing device dimension has demanded for precision in producing chips as it poses challenge of physical failure due to nano dust particles, contaminants etc. Therefore, (Chien et al.; 2015) in their research proposed an approach which would identify root cause of excursion from big data thereby improving the circuit probing yield. Empirical study showed that proposed methodology has an accuracy of over 70%. However, major drawback is that it works well when studied for excursion caused my single root cause but fails when implemented to identify multiple root causes.
March and Scudder (2019) used Internet of Things (IoT) for proposing predictive maintenance technique which can be used on big data to accomplish the maintenance requirements. The proposed technique focused on improving industries production efficiency with collaborative use of Big Data and IoT. The evaluation of this technique highlighted that although model took more time for execution but provided with better knowledge quality. Cloud is reshaping industrial dynamics, as large amount of data can be processed at once. Predictive maintenance can be achieved with the help of cloud as well. Mobile agent was thus designed with the help of cloud computing by (JOUR et al.; 2017) that would act as a remote system for predictive maintenance of production line. Technological giants of this field would be benefited by this system as they have their headquarters and production plants set up in different geographical locations thereby reducing the need of

centralized maintenance units. This approach however has safety concerns as to taking responsibility of any failure and completely depends on resource reliability, availability and adaptability.
Tao et al. (2018) in their research aimed to create smart semiconductor manufacturing process by simultaneous use of artificial intelligence (AI), IoT and Big Data. Cloud computing along with Hadoop was used to enhance flexibility of semiconductor fabrication wherein the smart system developed would provide timely maintenance to the equipment's after analyzing the data stored in cloud using Hadoop. However, this technique failed when implemented on data collected from devices performing multiple functions. With availability of abundant data, how big data influences production process was studied by (Kuo; 2019) wherein they proposed an approach for monitoring and improving equipment's in production line and system as a whole. Since it took into consideration all the aspects of production line to improve the system, it worked as designed but was not that accurate as compared to data mining approach.

2.2 Improving Manufacturing Process Efficiency using Simulation

Simulation of massive production plant helps engineers to operate and track all equipment's simultaneously. This approach was undertaken by (Qiao et al.; 2020) in their research where they proposed a simulation model consisting of scheduler and criteria selector which would help in rescheduling the production and change the delivery date in order to have minimum impact and stress on equipment's by providing them partial maintenance. This simulation model when evaluated proved to be more stable and robust than previous simulation models as it is adaptable to dynamic and unplanned changes.
Equipment's breakdown is one of the major reasons for dynamic nature of production plant. This was addressed by (Wang et al.; 2019) in their research wherein they proposed a simulation model which aimed at increasing the throughput by focusing on ways to increase the availability of equipment's. This simulation model took into consideration the production constraints, dispatching rules, and equipment behavior to classify equipment's in various groups making it easy for their timely maintenance. The equipment management and maintenance problem was addressed by (Gallo et al.; 2007) by proposing a simulation model which worked on two predefined viz. event driven and time driven models which provided alternative decision for maintenance and management of equipment's by analyzing data. This simulation model when evaluated proved to be effective but didn't consider the effect of collinearity as it worked on two pretrained models.
Shi and Zhou (2009) addressed this issue by accounting collinearity between error and requirement with the help of mathematical equations instead of using any pretrained model. This helped in improving the manufacturing process and provided quality control check with increasing productivity by efficient equipment utilization. (Lizaranzu and Rojo; 2012) in their research designed a Java based software to track equipment's activity in order to identify the fatigue level to provide maintenance and avoid failure. Their approach works well in final stage of production line where wafer is tested for their quality. Although the approach was innovative wherein equipment activity was tracked but it had a major drawback of not notifying the need of maintenance for equipment based on its fatigue.
Advanced process control (APC) was designed with an aim to predict yield and provide timely maintenance by (Moyne et al.; 2016). Better troubleshooting performance was

achieved as they made use of Impala which provided large storage. However, the primary objective of providing predictive maintenance was not achieved by their model. Yield prediction was also aimed by (Bomholt et al.; 2018) in their research where they developed TCAD application to predict and enhance wafer production. This application works best prior to initialization of fabrication as it takes into consideration the electric specifications provided by client and then calculates the specifications for fabrication equipment in order to increase the die yield, productivity and keeps the failure to bare minimum.
Photolithography is used for etching and implanting conducting layers onto silicon wafer and minor fault in any of this layer's results into lower die yield. This was addressed by (Ishida et al.; 2014) by proposing a model based volume diagnosis (MVD) which uses support vector regression (SVR) to handle noise and R square for silicon design mismatch. This model tracks and detects any faults caused during the etching process. When evaluated, their model achieved an accuracy of 91% but it is observed that this result is because of the use of data with predefined faults and high R square value.

2.3 Improving Manufacturing Process Efficiency using Statistical Approach

Statistical approach helps in achieving an optimal solution between enhanced semiconductor production and enhanced quality as it addresses this bottleneck issue with a mathematical perspective. Khakifirooz et al. (2018) in their approach made use of Gibbs sampling, Cohen's Kappa coefficient and Bayesian Inference to analyze data for identifying influential factors and detecting defect to provide smart manufacturing process. Even though this approach accounted for the collinearity between factors, it had no function for auto calculation of external factors and process impacting the die yield and quality. This led to manual calculation of this values which were then feed to system making it highly sensitive to an individual's approach.
Ge and Song (2010) adapted statistical approach in their research to design adaptive PCA to minimize the negative impact from any factor in data which was then monitored, processed and analyzed by support vector data description (SVDD) without future value estimation. The proposed approach when evaluated worked well on data with linear and stationary characteristics whereas failed when implemented on non-linear data. Smart decision making statistical model was designed by (Chien et al.; 2016) which focused on lowering the energy consumption in fabrication by making efficient use of equipment's with the help of overall process energy efficiency (OPE) indicator. This approach when evaluated worked well for implementing efficient use of equipment's in their active state but didn't address the underline objective of how to increase the usage of equipment's by avoiding the failure.

2.4 Improving Manufacturing Process Efficiency using Machine Learning Approach

Traditional statistical process control charts (SPC) used in semiconductor manufacturing plant have their limitation when used on today's multi-dimensional and vast data for monitoring of production line as SPC fails in detecting outliers thereby resulting in false alarms and reducing the efficiency. Machine Learning models can thus be used to handle large datasets generated in production plants on daily basis for monitoring and improving the productivity of plant. Table 1 illustrates the performance of machine learning

Table 1: Classification studies in semiconductor manufacturing field

Author	**Model Used**	**Best Accuracy**
Fernandes et al. (2020)	Neural Network	89.64%
Chien et al. (2012)	KNN	75%
Adly et al. (2015)	SVM-RBF	87.5%
Braha and Shmilovici (2002)	Decision Tree	77%

models implemented in previous research projects on wafer image dataset by various researcher's.

Yu and Kuo (2016) in their approach made use of Back Propagation Neural Network which determined the resources requirement and optimized their allocation to the process. This assured timely allocation of resource and tools to active equipment's on production line after forecasting the scheduled maintenance for each of them. This in turn reduced the cycle time because of uninterrupted functioning thereby increased the productivity.

Identifying the main cause of excursion was aimed by (Wei-Chou Chen et al.; 2004) in their research with the help of exploratory data analysis (EDA). Their approach identified whether the excursion is because of any contaminant or equipment's fault, but since this approach didn't include any use of models, engineers had to intensely study the charts and graphs obtained after EDA to come to a conclusion which proved to be time consuming. Machine learning tools were used by (Shan et al.; 2017) on image data-set obtained at the end of fabrication process to detect faults and identify their cause. They also made use of Chi-square test of independence to identify relation between detected and actual defect. The model was evaluated, and it worked well when implemented only on impacted region of wafer image to detect defects but failed when implemented on entire wafer image.

Increased production increases the energy consumption which goes against the climate protection policy of any manufacturer. This issue was addressed by (Yu et al.; 2017) in their research wherein they made use of support vector regression and neural network to develop a smart decision making model which would provide with efficient energy saving ways. Their approach could develop the relation between energy consumption of tools and input features/factors at each stage but couldn't provide an efficient energy usage prediction model which they aimed for. Chien et al. (2015) made use of logistic regression and random forest in their research with an aim to identify main cause of excursion for increasing die yield. Logistic regression clustered features with high correlation thereby avoiding collinearity followed by random forest classifying continuous and categorical features. Their approach reduced troubleshooting time and provided high accuracy when implemented to identify only one major cause of excursion.

Improving die yield was also aimed by (Nakata et al.; 2017) in their research for which they used convolution neural network (CNN) and K-mean clustering on wafer image data-set. K-mean clustering was used to form clusters of wafer images with failures and CNN was then implemented to identify a pattern in those clusters. This research helped engineers in identifying and classifying failures using pattern mining which thereby helped in addressing the cause of this fault, mending it and improving die yield. Another research was performed on wafer image data for identifying failure patterns by (Wu et al.; 2015) wherein they initially performed feature reduction in order to reduce the data dimension followed by implementing support vector machine (SVM) to identify and classify failure patterns. However, this approach didn't work well on data with low dimensions.

Adly et al. (2015) undertook research in the same field on wafer image data wherein they made collaborative use of data mining tools and general-regression-network-based (RGRN) model. The model was then trained on values of independently sampled image dataset and thus when evaluated provided 98% accuracy, had low variance, and successfully detected failure patterns in image data. Improving die yield by detecting micro-contaminants was aimed by (Braha and Shmilovici; 2002) in their research wherein they used decision tree along with neural network to build a classification model which was then implemented on data with low dimensions and high multicollinearity. As it works on lower dimension data, the execution time is comparatively less but it acts as a drawback because of increasing data size in today's world on which this model fails to work efficiently.

Researchers usually focus on detecting few causes of low yield and work on overcoming them. However, a very different approach was followed by (Chien et al.; 2012) wherein they aimed at detecting any fault leading to loss in die yield. Feature extraction using PCA followed by machine learning classification tools were implemented for constructing smart manufacturing. The model when evaluated worked well for identifying and classifying faults but failed in providing smart manufacturing as it lacks auto decision making capability. Fernandes et al. (2020) made use of Long Short-Term Memory (LSTM) Neural Network on numerical data obtained from the equipment's to detect failure in advance to apply predictive maintenance on those equipment's. When evaluated their Neural networks were 85-90% precise based on the number of hidden layers in them however it was found that this method of predictive maintenance worked well when applied on data over longer time.

Most of the research projects are either focused of improving the productivity by applying predictive maintenance on equipment's to avoid their failures or on improving die yield (i.e. quality) by detecting root cause of faults to overcome those. Lee et al. (2019) in their research proposed an approach for systematic assignment of wafers to equipment's on production line to maintain predetermined productivity with better die yield. They made use of an innovative statistical model which predicted the path for wafer in production line with ANOVA for feature selection and was successful in obtaining binary optimization. The approach however lacked implementation of machine learning tools apart from regression.

In conclusion, various research projects conducted in the field of wafer fabrication focused on improving die yield by working on wafer image to identify the faults and rectify their root cause whereas very few research were focused on improving the productivity by optimum allocation of wafers to various equipment's of production line. This encouraged an idea on working on sensor data from equipment's to increase the productivity without impacting the die yield by applying predictive maintenance with the help of machine learning tools and techniques.

3 Methodology

Different stages of this research resemble cross-industry standard process for data mining (CRISP-DM) stages. Wafer fault prediction methodology is represented in Figure 2 which covers 6 main stages viz. understanding the wafer fabrication business model, collecting data for research, pre-processing data collected from various resources, modelling, evaluation of machine learning models. The generated results could then be used

for effective decision making.

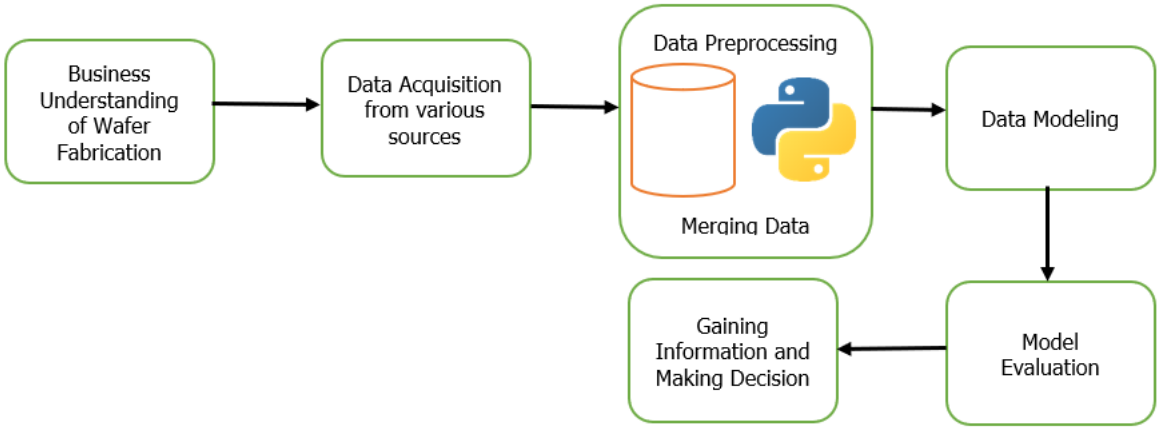

Figure 2: Wafer Fault Prediction Methodology

3.1 Business Understanding

Common factors responsible for lower die yield and moderate productivity are primarily considered to be faults on wafer caused by microcontaminants and non-optimal wafer allocation to equipment's on production line. However, other factors like health and functioning of equipment's which are primary backbone for smooth transition and timely completion of process are often overlooked. One practical approach to enhance productivity and die yield, is to consider a research on gaining insights about equipment's working and enhancing their performance by understanding and predicting the failure patterns for providing them timely maintenance to avoid future failures thereby increasing the availability of equipment's for various tasks. This would help in optimal allocation of wafers to available equipment's thereby increasing productivity.

3.2 Data Acquisition

Two datasets are gathered from two different sources as shown in Table 2. Both the data files were downloaded in .csv format. Both the datasets consist of data recorded from sensors in the equipment's of wafer fabrication production line, with dependent variable classified into 2 classes viz. pass (-1) and fail (+1) depending on the test results of fabricated wafer.

Table 2: Dataset Information

Dataset	Records	Attributes
UCI SEMCOM[3]	1567	591
WAFER[4]	7164	152

3

[3]https://archive.ics.uci.edu/ml/datasets/SECOM
[4]http://www.timeseriesclassification.com/description.php?Dataset=Wafer

3.3 Data Pre-processing

Crucial steps like handling missing values, feature engineering such as normalization and merging of datasets were done in this phase. UCI SEMCOM dataset consists of 591 attributes with 27 attributes having more than 50% of missing values which were dropped as it didn't lead to data loss. Apart from that, attributes with zero variance (i.e. no effect of dependent variable) were dropped as their presence or absence didn't have any impact on research. Attributes with less than 50% of missing values were imputed with median as the attributes had outliers and data has skew symmetric distribution. Thereafter, dataset was normalized because attributes consisted of outliers as well as the attribute values were in different range. The dependent variable of UCI SEMCOM dataset consists of pass category defined as '-1' and fail category as '+1'.
WAFER dataset consists of 154 attributes with no attributes having more than 50% of missing values thereby none of the attributes were dropped. Also, when checked for impact of attributes on dependent variable, it was found that none of the attributes had zero variance. Attributes with less than 50% of missing values were imputed with mean as the attributes didn't have outliers. Thereafter, dataset was normalized for scaling. The dependent variable of WAFER dataset consisted of pass category defined as '+1' and fail category as '-1'. To have standardized definition of pass and fail classes in dependent variable, we interchanged the designation for WAFER dataset thereby assigning '-1' to pass class and '+1' to fail class.

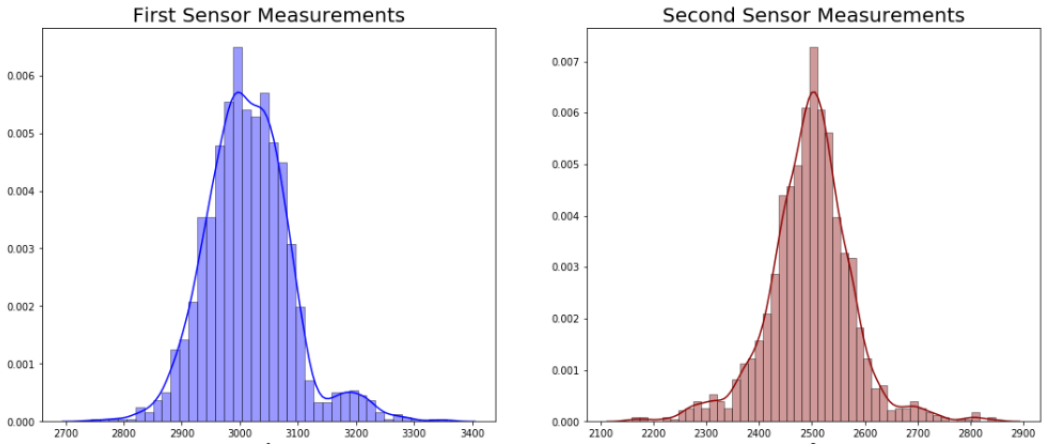

Figure 3: Probability Density Function Curve of Attributes with attribute values on X-axis

Figure 3 shows the probability of occurance of particular sensor value in data whereas Figure 4 illustrates the presence or absence of outliers in both data-sets to address the imputation of missing value by median and mean respectively. Final dataset had class imbalance with minor class just 9.92% of total data. Since the research focused on classification and dependent variable was binary, the class imbalance was handled using Synthetic Minority Over-Sampling Technique (SMOTE) wherein the minority class was oversampled to 50% to that of majority class thereby keeping a ratio of 2:1 (i.e. for every 2 cases of majority class there is 1 case of minority class).
Figure 5 shows the class imbalance in final merged data followed by Figure 6 illustrating various techniques in which this class imbalance was addressed. PCA was then applied on both the datasets after preprocessing to extract top features which explained more

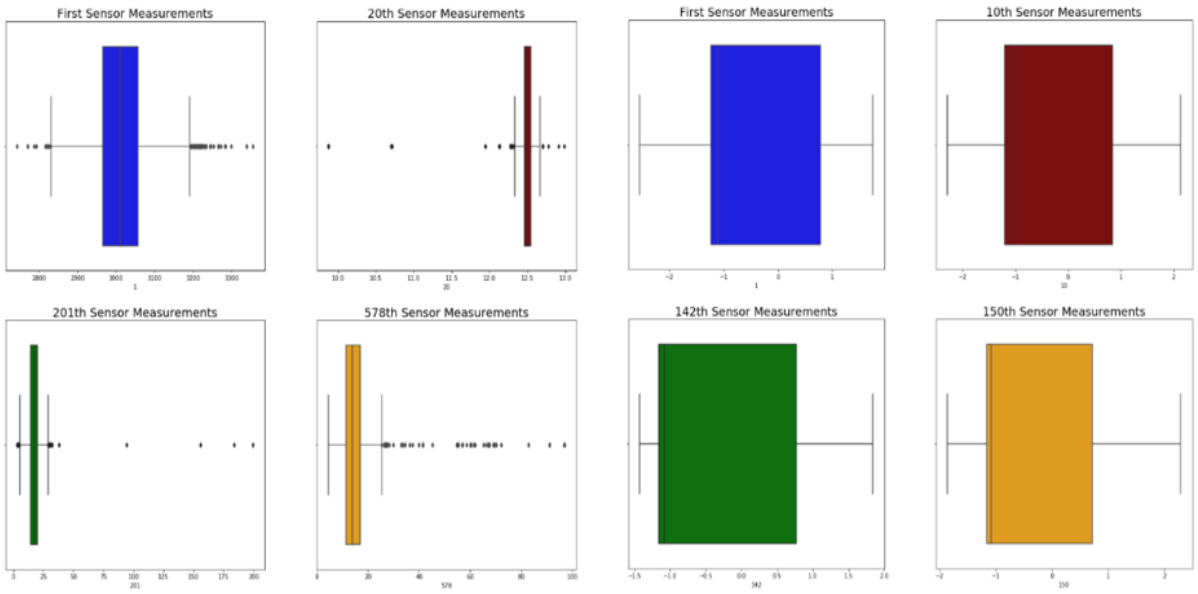

Figure 4: Box Plot of attributes from both dataset

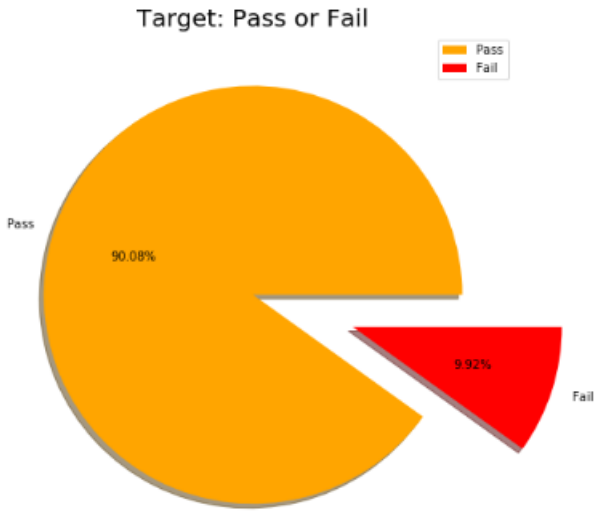

Figure 5: Class Imbalance of final dataset

than 80% of variance of entire dataset. Both the datasets were normalized prior to PCA to improve data integrity and reduce data redundancy. To normalize both the dataset, dependent variable was separated from independent variables which was then joint to the normalized variables of datasets.

Figure 7 illustrates the variance explained by extracted components in both data-sets. For UCI SEMCOM dataset, PCA was applied to extract 250 components from 447 attributes whereas for WAFER dataset PCA was applied to extract 150 components. Then after, variance ratio was calculated and plotted for principal components of both datasets. To merge two PCA data-frame's, top 100 principal components were selected as they explained more than 80% variance of both datasets.

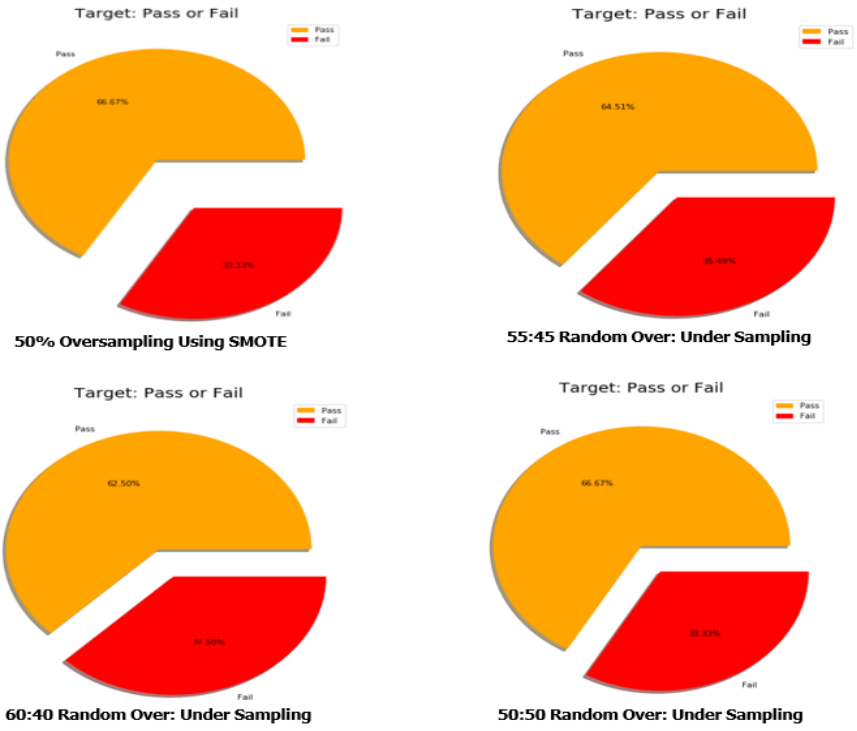

Figure 6: Handling Class Imbalance

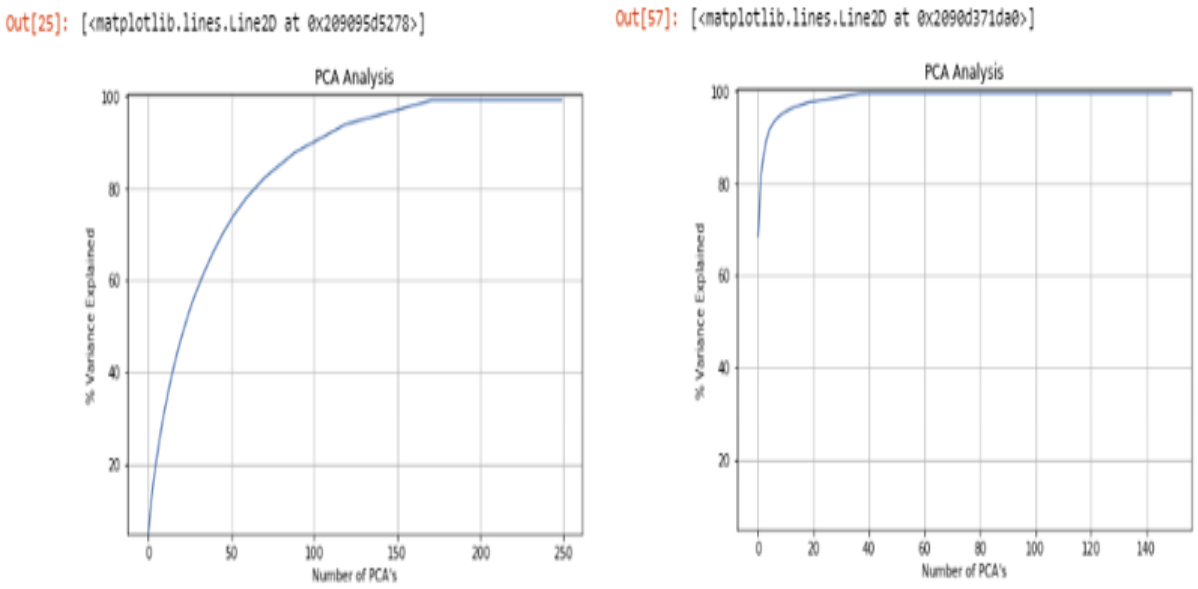

Figure 7: PCA Analysis for components of both datasets

In similar way, feature selection was also performed of both the datasets using ANOVA technique to select top 100 features from both the dataset which were then merged to form a final dataset which was then processed for overcoming class imbalance using SMOTE.

4 Design Specification

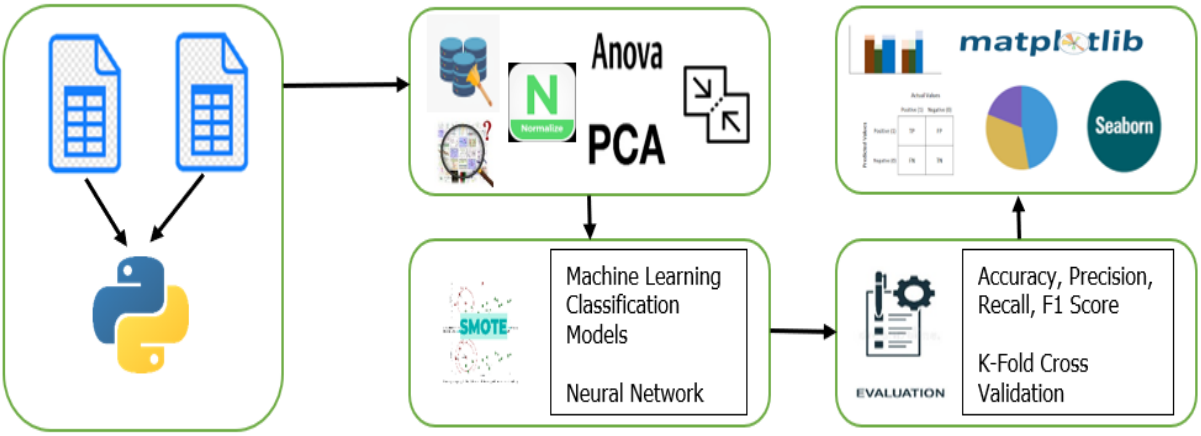

Figure 8: Wafer test Prediction - Design Flow

The architecture followed for this project is illustrated in Figure 8.

- In first stage, 2 datasets related to wafer production line are downloaded from UCI and CMU websites.
- The datasets are then loaded in Python where they are initially preprocessed. Then, exploratory data analysis (EDA) was performed to get some valuable insights from both datasets. Both the datasets are then normalized to perform feature extraction using PCA and featuring selection using ANOVA. The 2 data-frames of PCA are merged to form 1 finalized data frame and similarly 2 data-frames of ANOVA are merged to for other data frame.
- Train-test split is then applied in 75:25 ratio respectively followed by addressing class imbalance in training set using SMOTE oversampling of minority class to 50% of that of majority class, followed by implementation of machine learning classification models and a basic neural network with 5 hidden layers, 1 input and 1 output layer.
- Class Imbalance was also addressed using random under-sampling and oversampling together in the ratio of 60:40, 55:45, 50:50 respectively followed by implementation of models.
- The models were then evaluated using precision, F1 score, recall and classification accuracy. They were also cross validated using stratified K-fold cross validation.
- Lastly, the results were visualized in form of plot using various python libraries.

5 Implementation

The implementation consists of following steps: data preparation, implementation of models, cross validation of models under a range of stratified K-fold (10-50 folds) validation.

5.1 Data Preparation

The final dataset obtained after data cleaning, preprocessing and feature engineering (i.e. feature extraction using PCA and feature selection ANOVA) and merging consist of 100 columns and 8731 cases. This data was then split into 75% training dataset and 25% testing set. The class imbalance in training set was addressed in two ways viz. oversampling the minority class using SMOTE, applying random over-sampling and under-sampling in the ratio of 40:60, 45:55, 50:50 respectively. To reduce overfitting of models while being trained on training dataset, cross validation using K-fold test was performed for all the models. Table 3 illustrates size and proportion of train and test data-set.

Table 3: Dataset Structure Details

	Total Size	**Proportion**
Training Set	6548	0.75
Testing Set	2183	0.25
Total	8731	1

5.2 Decision Tree Classifier

Decision Tree Classifier (DT) is a supervised model as it is trained on the data with its correct output in order to learn the pattern which the model then uses to predict the output of new data on which it is implemented. It is a basic machine learning model with 3 components viz. nodes, edges, and leaf nodes. Node is the classification question whereas edges are the answer to that question (i.e. yes or not) and leaf node marks the exit point. DT classifies data into subplot by identifying lines. This process is performed repetitively as there might be multiple regions of same class. Real time data consist of impurity i.e. the distribution of classes is random and not defined in specific region, this is known as gini impurity (Braha and Shmilovici; 2002). Entropy is randomness of variable i.e. measure of impurity, is used to calculate information gain which then identifies which division would provide less impurity. DT thus selects the decision which has best information gain thereby correctly identifying majority of the classes. Due to comparatively smaller data size and reduction of attributes by feature engineering, DT was used in this research for classification.

5.3 Logistic Regression

Logistic Regression is a predictive analysis model which explains the relation between one dichotomous dependent variable and multiple independent variables with any data type. It is best suitable when multicollinearity of independent variables and impact of outliers is addressed. The working of logistic regression is mainly focused on estimating

log of odds of an event which explain the variance of data, thereby making it prone to overfitting with increasing number of independent variables. As multicollinearity was absent and impact of outliers was addressed by normalizing the data, logistic regression was used in this research.

5.4 Extreme Gradient Boosting Classifier

Extreme Gradient Boosting (XGBoost) Classifier is a popular classifier algorithm as it is efficient, portable, and flexible. It is a boosted decision tree for better performance as it learns from previous predictor variable residuals. XGBoost is used in this research as model outperforms most of the other classification predictive models when implemented on tabular data obtained in comma separated value (.csv) files.

5.5 Random Forest Classifier

Random Forest Classifier is an ensemble algorithm as it builds multiple decision trees by creating subsets of training dataset which are then used for classification by aggregating their votes to decide final class in test dataset (Tsanas and Xifara; 2012). This aggregation of decisions from various decision trees reduced the loss due to noise making random forest more accurate than just single decision tree. Random forest uses 10 decision trees in default setting to calculate the entropy of gini impurity as that of decision tree classifier. Random forest is used in this research due to its better performance and computational efficiency.

5.6 Support Vector Machine

Support Vector Machine (SVM) developed by Vapnik in AT and T Bell Laboratories (1997), is one of the robust prediction methods. SVM constructs an optimal hyper plane creating a separation with help of quadratic programming in hidden feature space to find unique solution (Adly et al.; 2015). SVM supports various kernels (i.e. mathematical functions) which can be used based on the aim of research. In this research, linear and radial basis function (RBF) kernels are used to compare the SVM performance on precision scale as no prior knowledge of data is required for these kernels.

5.7 Gaussian Naïve Bayes

Naïve Bayes Classifier is used for binary or multiclass classification and is famous among researcher's because its hypothesis calculation is tractable. Gaussian Naïve Bayes (NB) is an extension to Naïve Bayes as it works on the principle of normal distribution by estimating mean and standard deviation of training data. Thus, NB uses probabilistic approach for prediction and performance better when implemented on imbalanced dataset (Tao et al.; 2018). Functioning of NB can be beneficial in classifying and predicting equipment fault as the data is highly class imbalanced.

5.8 K Nearest Neighbour

KNN is a supervised machine learning technique which assumes that similar things exist in proximity. It was used in the research as it is easy to implement with no pre-requisite

of parameter tuning. The performance of the same was recorded and displayed in the tables of evaluation.

5.9 Neural Network

Neural Network (NN) works as a human brain to recognize the relation and pattern in data. NN architecture consists of three layers viz. input layer, hidden and output layer. Keras library is used to implement sequential NN which uses output of previous layer as input to next layer (Fernandes et al.; 2020). Dense constructor was used to define the layers of basic NN designed for this research. Compiler was used to add loss function and optimizer. Table 4 illustrates the input dimension for input, hidden and output layers along with the activation function used for each of those layers while designing a basic NN deigned for this research. 'Sparse Categorical Crossentropy' is used as loss function and 'adam' as an optimizer function to improve the accuracy. The NN was tested for different batch sizes and epochs.

Table 4: Dataset Structure Details

Layers	Input Dimension	Activation Function
Input	100	
1st Hidden	51	RELU
2nd Hidden	27	RELU
3rd Hidden	15	RELU
4th Hidden	9	RELU
5th Hidden	6	RELU
Output	2	SIGMOID

6 Evaluation

Models were evaluated for their precision over accuracy because to apply predictive maintenance on equipment's of production line, our models need to be precise in detecting the failed wafer category.

6.1 Feature Extraction Experiment using PCA

Dimensionality reduction using PCA was conducted on both datasets as they contained large number of attributes. Top 100 PCA's were selected from both the datasets which explained more than 80% variance of data and were merged to form a final dataset which was split into train and test set in 75:25 ratio respectively. Then after, class imbalance in train set was addressed using SMOTE where the minority class was oversampled to 50% of majority class. Table 5 shows the performance of all the models implemented.
As per the results, Random Forest outperformed all other models in terms of precision followed by SVM-RBF and XGBoost.

6.1.1 Stratified K-fold Cross Validation

The accuracy of all the models was cross validated using K-fold cross validation technique with folds ranging from 10 to 50. It was observed that model's accuracy increased in

Table 5: Results of Feature Extraction

Machine Learning Algorithm	Accuracy	Recall	Precision	F1 Score
Decision Tree	95.28%	84.37%	77.46%	80.77%
Logistic Regression	85.80%	59.82%	37.85%	46.36%
XGBoost	98.03%	86.16%	94.15%	89.97%
Random Forest	98.21%	84.38%	97.23%	90.65%
SVM-Linear	90.52%	58.93%	53.44%	56.05%
SVM-RBF	98.26%	87.50%	95.15%	91.16%
Naive Bayes	50.39%	83.93%	12.22%	25.77%
KNN	93.31%	94.64%	61.27%	74.39%

decimal points when number of folds were incremented from 30 to 50 therefore, Table 6 represents accuracy of all models up to 30 folds. From the table, XGBoost has the best

Table 6: Results of K-Fold Validation

Machine Learning Algorithm	K = 10	K = 20	K = 30
Decision Tree	94.17%	94.69%	95.92%
Logistic Regression	80.64%	80.78%	80.99%
XGBoost	95.86%	97.94%	98.90%
Random Forest	98.76%	98.79%	98.76%
SVM-Linear	84.12%	84.23%	84.39%
SVM-RBF	94.76%	96.83%	98.03%
Naive Bayes	42.43%	54.42%	57.98%
KNN	91.96%	92.25%	92.74%

accuracy followed by Random Forest but the difference between 2 is just 0.14%.

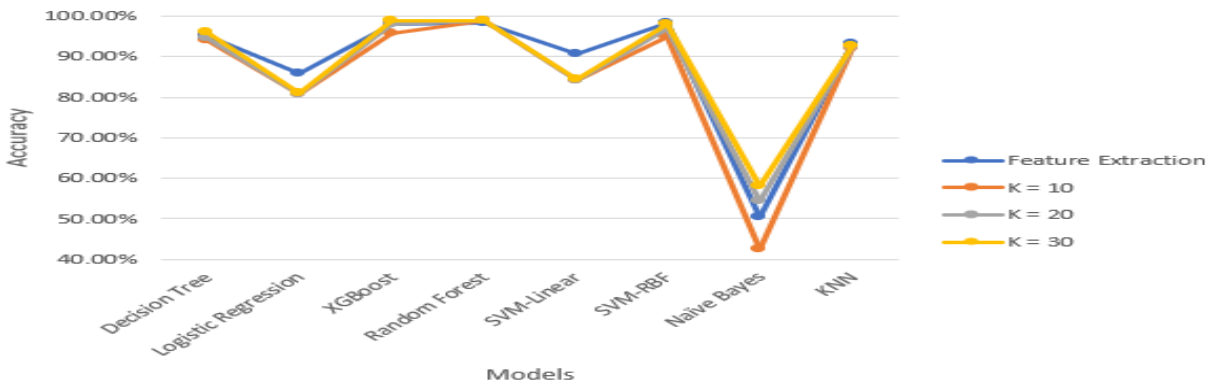

Figure 9: Comparison of Model Accuracy

Figure 9 compares model's for accuracy under feature extraction and K-fold validation.

6.1.2 Experiment with Logistic Regression

From Table 5, logistic regression model is not precise enough even after achieving acceptable accuracy. Logistic Regression threshold for classification of probabilities was identified to understand the cause of low precision.

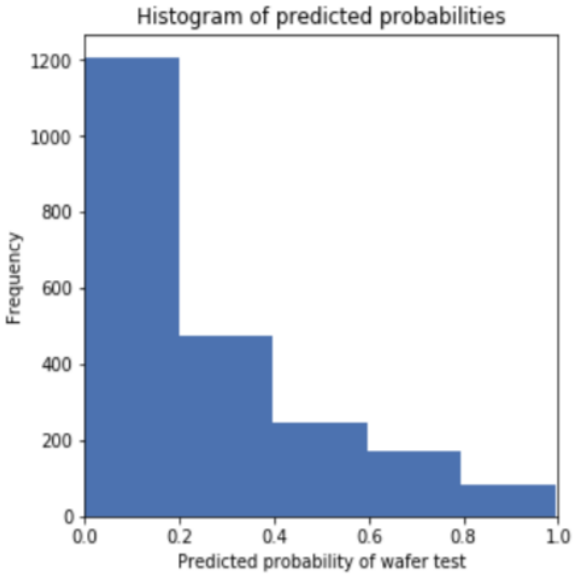

Figure 10: Histogram of predicted probabilities

From the histogram in figure 10 and array of first 10 predicted probabilities for class 1, it was found that it used probability of 0.5 as the threshold above which model classified case to be true positive. The threshold was then adjusted to 0.2 following the histogram and it was found that model's precision fell to 18.72% and accuracy to 61.89% indicating that logistic regression failed in identifying wafer failure.

6.2 Feature Selection Experiment using ANOVA

Feature selection was conducted using ANOVA to reduce dimensions of both datasets. Initially number of features were gradually reduced to identify feature count for which models provide optimum performance, however different models provided optimum performance for different feature count. Then after, top 100 features were selected from both the datasets and merged to form a final dataset to compare model's performance with that of feature extraction technique. Data was split into train and test set in 75:25 ratio respectively and class imbalance in train set was addressed using SMOTE where the minority class was oversampled to 50% of majority class. Table 7 shows the performance of all the models implemented.

Table 7: Results of Feature Selection

Machine Learning Algorithm	**Accuracy**	**Recall**	**Precision**	**F1 Score**
Decision Tree	96.93%	87.95%	83.12%	85.47%
Logistic Regression	86.80%	61.16%	40.53%	48.75%
XGBoost	98.35%	87.05%	96.53%	91.55%
Random Forest	98.58%	87.50%	98.49%	92.67%
SVM-Linear	92.99%	60.27%	67.84%	63.83%
SVM-RBF	98.26%	87.95%	94.71%	91.20%
Naive Bayes	72.61%	58.48%	20.60%	30.47%
KNN	95.37%	89.73%	72.04%	79.92%

From Table 5 and Table 7, it is seen that performance of most of the models improved over all parameters with feature selection technique and Random Forest once again outperformed rest of the models in both accuracy and precision.

6.2.1 Stratified K-fold Cross Validation

The accuracy of all the models was cross validated using stratified K-fold cross validation technique with folds ranging from 10 to 50. It was observed that model's accuracy increased in decimal points when number of folds were incremented from 30 to 50 therefore, Table 8 represents accuracy of all models up to 30 folds from which Random Forest is found to have the best accuracy followed by XGBoost with difference of just 0.09%.

Table 8: Results of Stratified K-Fold Validation

Machine Learning Algorithm	**K = 10**	**K = 20**	**K = 30**
Decision Tree	95.53%	95.41%	95.55%
Logistic Regression	74.11%	78.64%	79.95%
XGBoost	92.78%	97.17%	98.05%
Random Forest	94.00%	97.50%	98.14%
SVM-Linear	77.85%	82.89%	83.95%
SVM-RBF	88.88%	93.48%	94.63%
Naive Bayes	58.03%	61.88%	65.34%
KNN	98.12%	93.89%	94.84%

Figure 11: Comparison of Model Accuracy

Figure 11 compares model's for accuracy under feature extraction and K-fold validation.

6.3 Implementation of both over-sampling and under-sampling simultaneously

Since feature selection technique provided better model performance, the class imbalance in the train set of feature selected data was addressed in this experiment by random over-sampling of minority class and random under-sampling of majority class in 40:60, 45:55, 50:50 ratio respective. All models were implemented for 3 different ratios and it was found that models provided optimum performance for 45:55 ratio. Table 9 represents model's performance for 45:55 sampling ratio.

Table 9: Results of Random Over-Sampling and Random Under-Sampling

Machine Learning Algorithm	Accuracy	Recall	Precision	F1 Score
Decision Tree	96.47%	85.27%	81.28%	83.22%
Logistic Regression	84.24%	59.37%	34.46%	43.61%
XGBoost	97.98%	85.26%	94.55%	89.67%
Random Forest	98.21%	83.48%	98.94%	90.56%
SVM-Linear	90.56%	60.71%	53.54%	56.90%
SVM-RBF	97.66%	87.05%	89.86%	88.44%
Naive Bayes	85.57%	92.86%	41.02%	56.91%
KNN	96.47%	89.29%	79.05%	83.86%

Table 9 illustrates that Random Forest has outperformed all other models for precision and accuracy. From all the experiment's, Random Forest and XGBoost have proven to have best precision in predicting wafer failure.

6.4 Experiment with Neural Network

Neural Network with 5 hidden layers, 1 input and 1 output layer was implemented on data sets obtained from feature extraction, feature selection and 50:50 ratio of random over-sampling and random under-sampling with 25 and 50 as epoch and batch size of 60. Performance of NN improved in 50:50 sampling on feature selected data with an accuracy

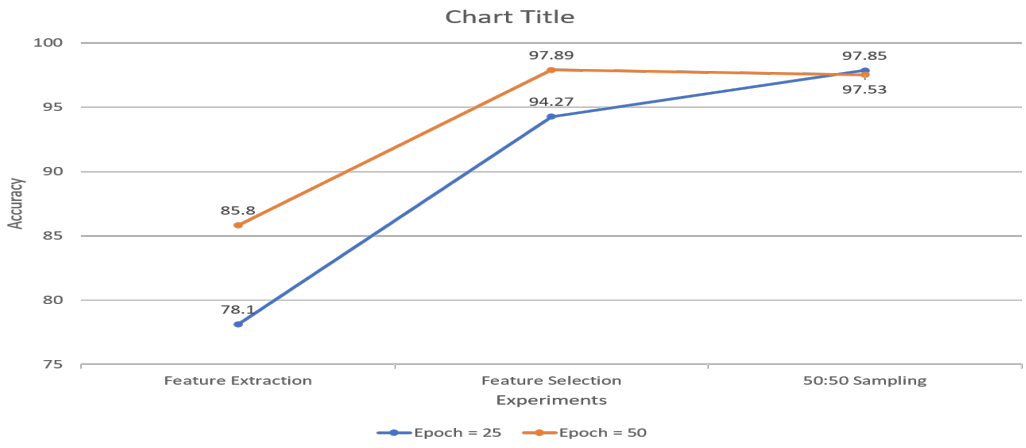

Figure 12: Neural Network Performance

over 97% for both 25 and 50 Epochs. Overall NN with epoch 50 has a better accuracy. Figure 12 illustrates the performance of NN in 3 different experiments for epochs 25 and 50.

6.5 Discussion

This research was conducted with an intention of exploring the relation between sensor data from equipment's and the test result of final wafer fabricated. Unlike the traditional approaches of predicting which equipment caused the failure by analyzing wafer image data for its failures, this approach included use of data from equipment's itself to identify their correlation with wafer test results. Major challenges faced include:

1. Addressing the impact of missing values:
   Due to limited number of cases, attributes with more than 50% of missing values were dropped to keep the data loss minimum. Then after, imputing attributes with median and mean depending on presence or absence of outliers respectively.

2. Addressing the impact of class imbalance:
   Class imbalance was addressed by conducting various experiments over-sampling and under-sampling to assure that models don't over-fit and provide bias results.

Nine machine learning models were implemented and evaluated. Various experiments were conducted to improve the performance of models, such as feature extraction using PCA, feature selection using ANOVA, implementation of random over and under sampling in 40:60, 45:55, 50:50 ratio on train set of feature selected data. All models except binary logistic regression and Naïve Bayes performed with precision of above 80%. Neural Network achieved an acceptable accuracy over 75% in all the experiments with highest of 97.85% for 50:50 ratio of sampling. Random Forest has the highest precision of over 95% in all experiments followed by XGBoost and SVM-RBF. Random Forest used 10 decision trees because of which it achieved better accuracy and precision over others. Figure 13 and 14 illustrates the performance of all models for precision and accuracy respectively.
Comparative analysis illustrates that feature selection using ANOVA performed better than feature extraction using PCA and 45:55 sampling. When compared with results obtained in previous researches conducted using data mining algorithms, KNN algorithm used in research by (Chien et al.; 2012) achieved an accuracy of 75% on wafer image data, while this research produced an accuracy of 95.37% for KNN. Similarly, SVM-RBF used by (Adly et al.; 2015) had an accuracy of 87.5%, while in this research SVM-RBF achieved an accuracy of 98.26%. Decision tree used by (Braha and Shmilovici; 2002) on wafer image data had an accuracy of 77% whereas in this research DT has an accuracy of 96.93%. On the other hand, SVM-Linear used in previous research had a F1 score of 88.9% where as SVM-Linear in this research has a F1 score of 90%.
Logistic Regression even though has a good accuracy but failed in precisely predicting True Positive class (i.e. true wafer failure) because it could not categorize the probabilities of events. Similarly, Naïve Bayes failed because of phenomenon called zero frequency where it assigns 0 to an event if the category is not observed in training dataset. Also, major limitation of Naïve Bayes is its assumption of predictors to be completely independent which is impossible in real world data.

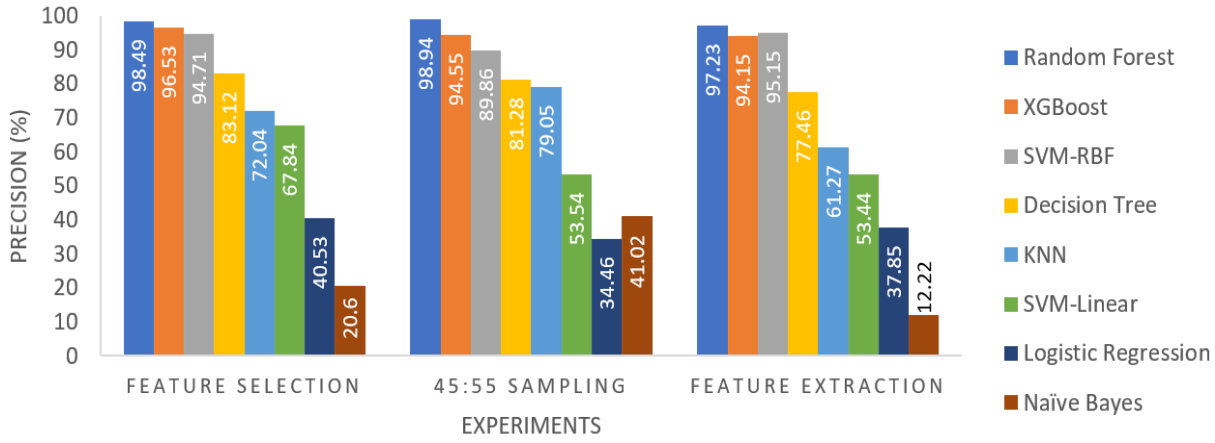

Figure 13: Comparative Analysis for Precision

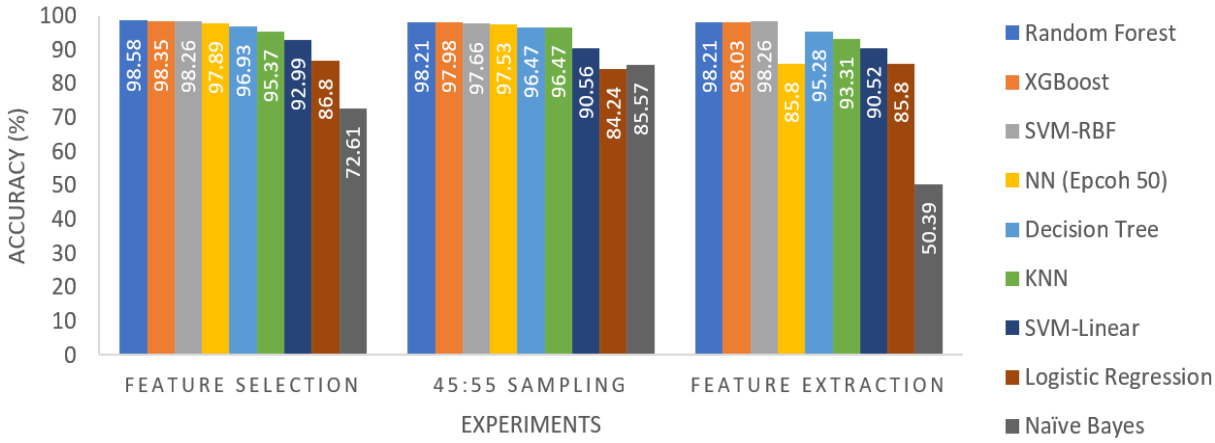

Figure 14: Comparative Analysis for Accuracy

7 Conclusion and Future Work

The proposed research worked has implemented one of the recent approaches for applying predictive maintenance in the manufacturing field. This research made use of 2 data sets consisting of numerical data collected from the sensors in equipment's of production line which were thoroughly processed before implementing nine classification-based machine learning algorithms under various experiments. Feature selection using ANOVA achieved better model performance than other experiments as Random Forest outperformed all models in terms of precision and accuracy, followed by XGBoost and SVM. This research highlighted that, numerical data from wafer fabrication is of equal use as that of image data to predict the wafer failure and identify the equipment responsible for that. This can help in implementing predictive maintenance for equipment's to avoid near future failures, thus increasing the productivity and enhancing the die yield.

In future, classification algorithms developed in this research can be used on real time data of production line and they can be incorporated into company systems to provide timely servicing to all equipment's. The research made use of basic Neural Network as data size was limited, but when working on real time big data, this Neural Network can

be modified by adding hidden layers and nodes and then testing it for various epochs. Alternatively, if working on real time time-series data from wafer fabrication, these models can be further optimized and deployed in end to end deployment to automate the task thereby saving time.

Acknowledgement

I would like to thank my supervisor, Dr. Muhammad Iqbal, for his constant support, guidance, encouragement, and motivation throughout the research work. For consecutive 13 weeks my supervisor assisted me with my queries in each of the one to one meeting. His continuous feedback helped me to present this research project in an insightful and sophisticated manner.

References

Adly, F., Yoo, P. D., Muhaidat, S., Al-Hammadi, Y., Lee, U. and Ismail, M. (2015). Randomized general regression network for identification of defect patterns in semiconductor wafer maps, *IEEE Transactions on Semiconductor Manufacturing* **28**(2): 145–152.

Bomholt, L., Chalmers, J. and Fichtner, W. (2018). Method and system for enhancing the yield in semiconductor manufacturing. US Patent 10018996.
URL: *http://www.google.it/patents/US10018996*

Braha, D. and Shmilovici, A. (2002). Data mining for improving a cleaning process in the semiconductor industry, *IEEE Transactions on Semiconductor Manufacturing* **15**(1): 91–101.

Chien, C.-F., Hsu, C.-Y. and Chen, P.-N. (2012). Semiconductor fault detection and classification for yield enhancement and manufacturing intelligence, *Flexible Services and Manufacturing Journal* **25**.

Chien, C.-F., Liu, C.-W. and Chuang, S.-C. (2015). Analysing semiconductor manufacturing big data for root cause detection of excursion for yield enhancement, *International Journal of Production Research* **55**: 1–13.

Chien, C.-F., Peng, J.-T. and Yu, H.-C. (2016). Building energy saving performance indices for cleaner semiconductor manufacturing and an empirical study, *Comput. Ind. Eng.* **99**: 448–457.

Fernandes, S., Antunes, M., Santiago, A., Barraca, J., Gomes, D. and Aguiar, R. (2020). Forecasting appliances failures: A machine-learning approach to predictive maintenance, *MDPI Journals* .

Gallo, M., Guizzi, G. and Zoppoli, V. (2007). An integrated approach to develop a simulation model in manufacturing processes, *INTERNATIONAL JOURNAL OF SYSTEMS APPLICATIONS, ENGINEERING DEVELOPMENT* **1**.

Ge, Z. and Song, Z. (2010). Semiconductor manufacturing process monitoring based on adaptive substatistical pca, *IEEE Transactions on Semiconductor Manufacturing* **23**(1): 99–108.

Ishida, T., Nitta, I., Banno, K. and Kanazawa, Y. (2014). A volume diagnosis method for identifying systematic faults in lower-yield wafer occurring during mass production, pp. 670–675.

JOUR, Wang, J., Zhang, L., Duan, L. and Gao, R. X. (2017). A new paradigm of cloud-based predictive maintenance for intelligent manufacturing, *Journal of Intelligent Manufacturing* **28**.

Khakifirooz, M., Chien, C. F. and Chen, Y.-J. (2018). Bayesian inference for mining semiconductor manufacturing big data for yield enhancement and smart production to empower industry 4.0, *Applied Soft Computing* **68**: 990–999.

Kuo, Y.-H. (2019). From data to big data in production research: the past and future trends, *International Journal of Production Research* **57**: 4828–4853.

Lee, D.-H., Lee, C.-H., Choi, S.-H. and Kim, K.-J. (2019). A method for wafer assignment in semiconductor wafer fabrication considering both quality and productivity perspectives, *Journal of Manufacturing Systems* **52**: 23–31.

Lizaranzu, M. J. M. and Rojo, F. C. (2012). Equipment utilization tracking and improvement in semiconductor industry in probe and final test areas, *IFAC Proceedings Volumes* **45**: 127–132.

March, S. T. and Scudder, G. D. (2019). Predictive maintenance: strategic use of IT in manufacturing organizations, *Information Systems Frontiers* pp. 1–15.

Moyne, J., Samantaray, J. and Armacost, M. (2016). Big data capabilities applied to semiconductor manufacturing advanced process control, *IEEE Transactions on Semiconductor Manufacturing* **29**(4): 283–291.

Nakata, K., Orihara, R., Mizuoka, Y. and Takagi, K. (2017). A comprehensive big-data-based monitoring system for yield enhancement in semiconductor manufacturing, *IEEE Transactions on Semiconductor Manufacturing* **30**(4): 339–344.

Qiao, F., Ma, Y., Zhou, M. and Wu, Q. (2020). A novel rescheduling method for dynamic semiconductor manufacturing systems, *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **50**(5): 1679–1689.

Shan, C., Babighian, P., Pan, Y., Carulli, J. and Wang, L. (2017). Systematic defect detection methodology for volume diagnosis: A data mining perspective, pp. 1–10.

Shi, J. and Zhou, S. (2009). Quality control and improvement for multistage systems: A survey, *IIE Transactions* **41**(9): 744–753.

Tao, F., Qi, Q., Liu, A. and Kusiak, A. (2018). Data-driven smart manufacturing, *Journal of Manufacturing Systems* **48**: 157–169.

Tsanas, A. and Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools, *Energy and Buildings* **49**: 560–567.

Wang, J. and Zhang, J. (2016). Big data analytics for forecasting cycle time in semiconductor wafer fabrication system, *International Journal of Production Research* **54**(23): 7231–7244.

Wang, J., Zheng, P. and Zhang, J. (2020). Big data analytics for cycle time related feature selection in the semiconductor wafer fabrication system, *Computers Industrial Engineering* **143**.

Wang, L.-C., Chu, P.-C. and Lin, S.-Y. (2019). Impact of capacity fluctuation on throughput performance for semiconductor wafer fabrication, *Robotics and Computer-Integrated Manufacturing* **55**: 208–216.

Wei-Chou Chen, Shian-Shyong Tseng, Kuo-Rong Hsiao and Chia-Chun Liu (2004). A data mining projects for solving low-yield situations of semiconductor manufacturing, pp. 129–134.

Wu, M., Jang, J. R. and Chen, J. (2015). Wafer map failure pattern recognition and similarity ranking for large-scale data sets, *IEEE Transactions on Semiconductor Manufacturing* **28**(1): 1–12.

Yu, C., Chien, C. and Kuo, C. (2017). Exploit the value of production data to discover opportunities for saving power consumption of production tools, *IEEE Transactions on Semiconductor Manufacturing* **30**(4): 345–350.

Yu, C. and Kuo, C. (2016). Data mining approaches to optimize the allocation of production resources in semiconductor wafer fabrication, *2016 International Symposium on Semiconductor Manufacturing (ISSM)*, pp. 1–4.