1. **Abstract**

In this paper we will learn how do we build a model to make decisions pursuant to manufacturing process of semiconductors. One of the most complex and technologically advanced production processes is that of semiconductors. Uni-variate and multivariate analyses, two common traditional machine learning algorithms, have long been used as a tool for developing predictive models to find errors. Predictive modelling has been the subject of significant collaborative research projects over the last ten years between academia and fab industry. We cover a few of these study areas in this work and then suggest machine learning techniques to automatically produce an accurate prediction model to forecast equipment defects throughout the wafer fabrication process of the semiconductor industries. To maintain high process yields in manufacturing, this research article intends to build a decision model that would assist in immediately identifying any equipment defects. In this paper, we implement various methods of carrying out tests with few tools and python.

1. **Introduction**

Semiconductor manufacturing processes are highly technologically complicated processes with over a thousand steps, inevitably being prone to error due to how delicate nature and repetitive stages they have to undergo. Normally wafers are built out of different layers tangled with circuits and, if the final product is defective, it results in waste for the companies in terms of not only money but also time.

The need for a practical and automated solution to detect failures in early stages is evident, not only to avoid material waste but mainly to avoid releasing faulty products to the market. Therefore, in this work statistical analysis and modelling algorithms were used to develop such a solution, testing many different approaches and methods to define the ideal option for maximizing the quality of the prediction model.

In recent years, numerous manufacturing tools have been fitted with sensors to enable real-time production process monitoring. These sensor data on the production status and equipment state offer a chance for effective control and optimization. Unfortunately, such measurement data are so voluminous that it is challenging to identify any production-related faults in a timely manner. In this research, we investigate the issue of precise equipment failure condition detection during wafer manufacture. The SECOM dataset, which was made available by Michael McCann and Adrian Johnston, is the basis for the data set's descriptions on Kaggle and Research Gate. Using machine learning techniques, the fault detection model can be automatically created from the sensor data already there.Industries and research communities have long been interested in developing a method that is efficient and effective for tracking the condition of equipment and identifying impending failure. Predictive maintenance has become a reality thanks to BIG data. Predictive modelling does not follow a standard methodology; instead, it is frequently customized to address a particular business need.

Machine learning algorithms have traditionally been focused on simple prediction modeling. Given observations that have been generated by an unknown stochastic dependency, the goal is to infer a law that will be able to correctly predict future observations generated by the same dependency. Statistics, in contrast, has traditionally focused on “data modeling”, i.e., on the estimation of a probability law that has generated the data. During recent years, the boundaries between the two disciplines have become blurred and both communities have adopted methods from the other.

The creation and study of algorithms that can learn from data are explored by the scientific field of machine learning. Instead of just executing explicit instructions from a program, these algorithms operate by creating a model from inputs and using that to make predictions or judgments.

Machine learning is the process of using learning algorithms to create models that can reliably predict the class of unlabeled data and characterize the underlying data the best. A significant amount of data has been automatically collected in modern industry with its many automated machines. Engineers might be able to use this raw data to find certain hidden patterns, like the process fault model, to help with the timely examination of the fault’s root causes*.* Further in this paper, we will discuss different methods of predicting failure in the manufacturing process.

1. **Literature Review**

Based on the descriptions in Kaggle and Research Gate, the SECOM dataset, provided by Michael McCann and Adrian Johnston, contain real data on the semi-conductor manufacturing process. The data comes from 590 sensors and process measurement points spread through the whole production line, in this case consisting of 1567 instances. The objective of carrying out this process is to Lower down cases of defective wafers so as to increase productivity and reliability by means of prediction using data provided by sensors installed on semiconductor process.

In the due process of doing this Research. One shall work towards the goal of coming up with a feature selection technique that chooses from over the 500 features, which are provided by sensors scattered across different places of the semiconductor fabrication units. Those features that potentially increase model’s processing time are disregarded. Since Data is imbalanced due to the amount of pass and fail cases, it is necessary to look for a technique that could help to compensate and help the model to increase accuracy. According to statistics show that most semiconductor manufacturing equipment suffers at least 8% unscheduled downtime and loses another 7% to scheduled maintenance in the process.

The production of modern semiconductors has long been hampered by the need to maximize process tool uptime. The technique of determining when equipment needs maintenance in order to prevent a catastrophic failure is known as predictive maintenance. The PdM difficulties typically suffer from an insufficient number of observations to create a solid statistical model. The research study offers a procedure for performing preventive maintenance on semiconductor equipment that entails the processes of gathering and preprocessing raw data, followed by classification using a statistical classification model.

Cost, quality, and delivery time are crucial considerations in the majority of manufacturing processes if businesses want to stay competitive over the long run (Governi 2006), (M. Gallo n.d.). Process control is essential for the semiconductor industries, which run multistage manufacturing processes on a smaller 300 nanometer product size (S. J. Qin 2006). Pham and Afify studied manufacturing-related machine learning approaches. They assessed the various machine learning approaches and looked at the applications where they have been successfully used.

Real-time process control is made possible in semiconductor manufacturing by modern technology using measurement data from equipment sensors and the final electrical test. Process engineers have challenging work since there is so much data that is collected during the entire production process, making effective monitoring and optimal process management by looking into and interpreting this data is tough. Univariate and multivariate control charts, which were once common in process control, are no longer an effective way to manage manufacturing systems with numerous processing stages. It is necessary to use automatic and sophisticated process control methods.

A decision tree classification model was proposed by Ison and colleagues (A. M. Ison 1997), (Spanos 1996)to identify plasma etch equipment faults. The five sensor signal data were used to create the model. The challenge of fault identification during the etch process was also extensively researched. Building a specific control chart for identifying a certain sort of issue was suggested by Goodlin et al (B. E. Goodlin 2003). Directly from the etcher, they gathered data on tool states. There are 19 variables in these data. The statistical strategy was used in Spitzlsperger and colleagues' (G. Spitzlsperger 2005) research as well. To sustain changes in the mean and standard deviation coefficients by remodelling approach, they employed the multivariate control chart method.

The process control problem with a few aspects of tool-state and process-state measurement data has been studied in the majority of work on fault detection algorithms. In a somewhat different scenario, McCann and his team (M. McCann 2008) suggested that the measurement data from the wafer manufacture process contain as many as 590 characteristics. The prediction and computational performances must be improved by feature selection approaches (Elisseeff 2003) due to the abundance of features or variables.

The last aspect of the wafer fabrication data (SECOM, SEmiCOnductor Manufacturing n.d.) acquired from 590 sensors, with a label declaring pass or fail condition, is also analyzed in this research. There are 104 fail cases compared to 1,463 pass cases in the observed data. In this study, a boosting methodology is developed to deal with the extreme imbalance between pass and fail instances in addition to a feature selection method for extracting the post discriminative sensors.

**Three Fundamental Areas to gain Knowledge from Existing Literature are:**

1. **Imputation Methods**

Many real-world datasets may include missing values for a variety of reasons. In many cases, they are encoded as NaNs, blanks, or other types of stand-ins. Training a machine learning model on a dataset with many missing values might have a major negative influence on the model's quality. Certain algorithms, such as scikit-learn estimators, presumptively assume that every value is numerical and has, and always will have, a meaningful value.

One way to solve this problem is to discard the observations with missing data. You face the chance of losing data points with crucial information, though. The preferable option would be to impute the missing data. To put it another way, we must extrapolate those missing figures from the existing data.

**Types of Imputation methods:**

1. **Imputation Using (Mean/Median) Values:**

This method involves determining the mean/median of the non-missing values in each column before replacing each column's individual missing values independently of the others. Only numerical data may be used with it.

1. **Hot-Deck imputation:**

works by picking the missing value at random from a group of related and comparable variables. In summary, there is no ideal method for making up for missing values in a dataset. Each technique may perform better for some datasets and types of missing data, but much worse for other datasets. There are some predefined guidelines for selecting the best approach to employ for different kinds of missing values, but you should also experiment to see which model fits your dataset the best.

1. **Imputation Using (Most Frequent) or (Zero/Constant) Values**:

**Most Common is a different statistical method for replacing missing values with their most frequent equivalents inside each column. It works with categorical characteristics (strings or numerical representations).**

1. **Imputation Using k-NN:**

A simple categorization procedure is known as the k nearest neighbors. The method predicts the values of any additional data points using "feature similarity." In other words, the value given to the new point depends on how much it resembles the points in the training set. Finding the k's nearest neighbors to the observation with missing data and then imputing them based on the non-missing values in the neighborhood can be very helpful in forming predictions about the missing values. Let's look at some sample code using the Impuyte library, which offers a straightforward method of applying KNN for imputation.

1. **Imputation Using Multivariate Imputation by Chained Equation (MICE)**

This kind of imputation functions by repeatedly filling in the missing data. As it more accurately gauges the uncertainty of the missing values, multiple imputations (MIs) are superior to a single imputation. The chained equations method is likewise very adaptable and can deal with various variables and various data formats (e.g., continuous or binary), as well as complexity such as boundaries or survey skip patterns.

1. **Imbalanced Data**

When locating the minority class has a considerably higher value than finding the majority, you must cope with an unbalanced data set. This will help us to generate higher accuracy model and to reduce fault prediction. The given SECOM dataset can be categorized by moderately imbalanced and it has a binary target variable. To balance the data, we use the following methods:

* 1. **Randomly oversampling Examples (ROSE)**

The idea with ROSE is to generate augmented sample of data by combining techniques of undersampling and oversampling as it is Unlike undersampling, this method leads to no information loss. Another advantage of using ROSE is that it is more efficient than simple over-sampling. The only drawback can be that it is likely to increase overfitting due to replication of data from minority class.

1. **Synthetic Minority Over-Sampling Technique (SMOTE)**

The Idea with SMOTE is that here we are Oversampling by selecting a data point and generating a new datapoint somewhere between it and a random neighbor datapoint. The advantage is that it alleviates overfitting caused by *ROSE* as synthetic examples are generated rather than replication of instances. But on the other hand it has a tendency to cause strong blindness in the selection of nearest neighbors for the synthetic samples.

1. **Adaptive Synthetic Sampling Method (ADASYN)**

The adaptive synthetic sampling approach, or ADASYN algorithm, expands on the SMOTE methodology by giving minority classes that are challenging more weight at the classification boundary. ADASYN uses a weighted distribution for various minority class examples based on how challenging they are to learn, producing more synthetic data for the more challenging minority class instances.

1. **Synthetic Minority Over-Sampling Technique & Edited Nearest Neighbor (SMOTE-ENN)**

This method Similar approach as *SMOTE*, just that sampling and cleaning of data points is performed on both the classes.

1. **Feature Selection**

The crucial phase in the entire predictive modelling process is to identify the parameters that have the greatest influence. It is crucial to remember that the choice of predictor variables to employ as model inputs plays a key role in determining whether a predictive model is successful. The use of a trial-and-error method is the simplest technique for selecting parameters. However, because there are so many parameter variables with this method, the processing is time-consuming and difficult. It is also feasible to avoid this predictor selection stage with the majority of modelling techniques by using all data as input variables, however this strategy could lead to unfavorable model behavior. By doing this, it will help us to avoid over-fitting, we will have more Computational Power and less prediction accuracy.