**LEVERAGING AI FOR SMART MANUFACTURING**

A minor project report submitted in partial fulfilment of the requirements for the award of the degree of

**Bachelor of Technology**

in

**Mechanical Engineering**

Submitted By

**ARYAN UPPAL (18BME017)**

Under the supervision

Of

**DR. SANJAY SHARMA**

ASSISTANT PROFESSOR

SCHOOL OF MECHANICAL ENGINEERING



**School of Mechanical Engineering**

**SHRI MATA VAISHNO DEVI UNIVERSITY**

**KATRA(J&K)**

**BONAFIDE CERTIFICATE**

This is to certify that the project titled **LEVERAGING AI FOR SMART MANUFACTURING** is a bonafide record of the work done by

ARYAN UPPAL (18BME017)

in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Mechanical Engineering of the SHRI  MATA VAISHO DEVI UNIVERSITY ,KATRA during the year 2020-2021.

DR. SANJAY SHARMA

ASSISTANT PROFESSOR

SCHOOL OF MECHANICAL ENGINEERING

# ABSTRACT

The Applications that leverage the benefits of applying ML in manufacturing industries have been successfully realized.

This project is mainly designed by taking manufacturing problems as my main concern. We all know that manufacturers face problems when it comes to producing goods of higher quality at low cost. One of the major issue that manufacturing industry face is the Failures in Production Lines and this project will help them to Predict Failures in Production Lines by making earlier predictions for those failures so that a manufacturing company can take proper measures in advance to cope up with the problems.

In this project: -

1) Data science methods are applied to this huge data provided by bosh consisting records for products as they progress through different stages.

2) Measurements made for each component along the assembly line to predict internal failures.

3) It was found that it is possible to make a ml model that predict which parts are most likely to fail.

Thus, a failure detection system can be built which is smarter, more accurate and the parts tagged as likely to fail can be salvaged to decrease the operating costs and increase the profit without compromising with the quality.

**ACKNOWLEDGEMENTS**

I take this opportunity to express my profound gratitude and deep regards to school of mechanical engineering for his exemplary guidance, monitoring and constant encouragement throughout the course of this Project. I have immense pleasure in expressing my thanks and deep gratitude to my guide **Dr. Sanjay Sharma** Associate Professor, School of Mechanical engineering SMVDU for his guidance throughout this project. The blessing, help and guidance given by him time to time shall carry me a long way in the journey of life on which I am about to embark.

I thank almighty, my parents, family, friends and teachers for their constant encouragement without which this assignment would not be possible.

Dated: 11/1/2021 Name: ARYAN UPPAL

# TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| **S.No** | **contents** | **Page No.** |
| 1. | Introduction | 7 |
| 2. | **PREDICTION APPROCH** | 8 |
| 3. | DATA INSPECTION | 9 |
| 4. | PREDICTION OVERVIEW | 12 |
| 5. | CLUSTERING | 13 |
| 6. | SUPERVISED LEARNING | 17 |
| 7. | Algorithm Comparison | 19 |
| 8. | Feature Selection | 22 |
| 9. | FUTURE WORK | 25 |
| 10. | REFERENCES | 26 |

**LIST OF TABLES**

**Table Title Page**

**1 Properties Of Each Cluster………………………………...16**

**2 Performance Of Final Model……………………………....23**

**LIST OF FIGURES**

**Figure Title Page**

# 1 Percentage Of Missing Values…………………………………………..11

# 2 Number Of Principle Components…………………………………….13

# 3 Number of Clusters……………………………………………………….….14

# 4 Scatterplot of the Data………………………………………………………16

# 5.1 Algorithm Performance……………………………………………………..20

# 5.2 Algorithm Performance……………………………………………………..20

# 5.3 Algorithm Performance……………………………………………………..21

# 6 Relative Feature Importance……………………………………………….22

# 7 Positive Rates……………………………………………………………………….23

**INTRODUCTION**

As Smart manufacturing is being touted as the next industrial revolution. With real time monitoring of production processes, to increase productivity and stay competitive, the use of data science methods is an obvious next step.

In this project I used the large-scale data set of a production line provided by Bosch for hosting a challenge on Kaggle aiming to predict the manufacturing failures using the anonymized features. I decided to use a two-stage method for the prediction by first to cluster the data into groups based on the process and then use supervised learning to predict the failed product in each cluster. This approach helps reducing the sparsity of the data. Various supervised algorithms were compared (Logistic regression, naive bayes, decision tree, gradient boosting and random forest). The random forest algorithm was seen achieving the highest performance score So, it was chosen as the final model.

The recent trend has shifted to harnessing the power of huge data using ML and reducing the effort of manual feature engineering. To cope with large data sets, clustering and PCA (The Principal Component Analysis) have commonly been used to reduce data size and select important features. For this project, a two-stage approach is proposed with an aim of using anonymized features to predict failed products.

From this project, I present my findings from the dataset then explore the challenges faced due to the size of the dataset, the kind of data recorded, and machine learning algorithms that are suitable for these kinds of problems.

**Flowchart of the prediction approach**

**DATA INSPECTION**

*A. Manufacturing Line Dataset*

A competition was hosted by Bosch on Kaggle aiming to predict the failed products in a production line. Each observation represents a product that moves through a manufacturing process. The naming of the feature is done by the convention of “L#\_S##\_F####”, which represents the line, station, and feature number. To know if a product may be a failure or not is provided as a binary class, with 1 representing failure and 0 representing pass. The aim is to use the features to predict the occurrence of a failed product.

The given dataset has numerical, categorical and time data. The categorical data is extremely sparse (more than 99%) and thus not included in this project. The label of the preceding data was required for the time data, which can not be availed in a production line and thus isn’t not considered in this project. Only the numerical dataset is considered.

Summary of the dataset is as follows:

•No. of features: 968.

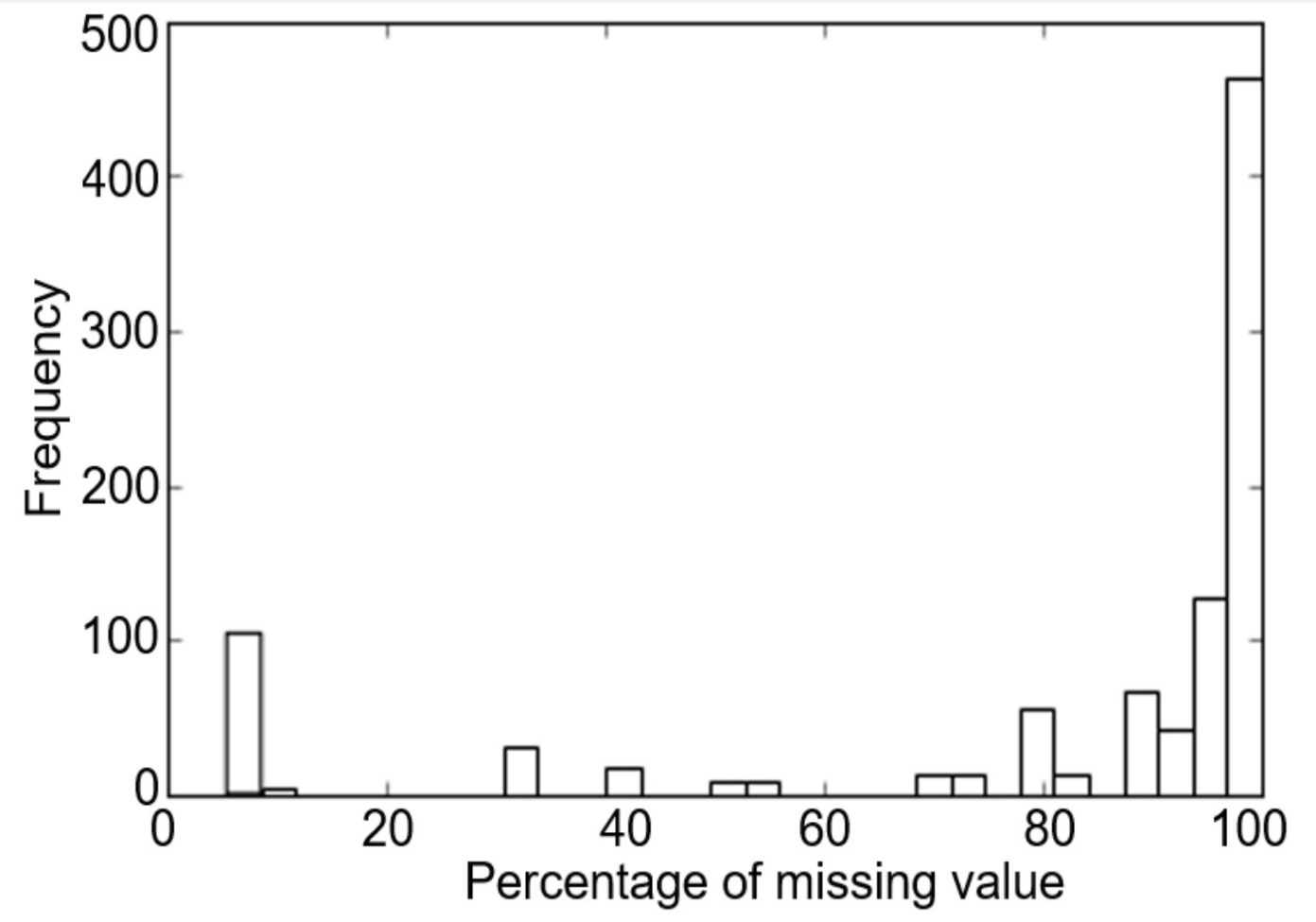
•No. of observations: 1183747

•Percentage of a failed product: 0.58%

•Percentage of missing values: 78.5%

**Several challenges are present in this dataset:**

The high sparsity in the dataset brings problems. Noises are introduced when replacing the missing values. And, it often creates a larger data size which slows down the ml algorithm. So, clustering will be conducted to divide data into similar process groups. In the second stage, supervised learning will be used for the prediction of the failed product in each cluster.

****

**PRODUCT FAILURE PREDICTION OVERVIEW**

The prediction method is conducted in 2 stages: -

**Stage I clustering:** This step is done to cluster data with similar processes together into process groups. In each cluster, the empty and constant features are deleted to scale back the data size.

**Stage II supervised learning:** This step uses supervised learning technique to predict the failed products. Every cluster is treated as an independent dataset and has its own classifier. Throughout the prediction step, the data is 1st classified with relation to that clusters it belongs to. Then the classifier in the corresponding cluster is used to predict whether or not the data is a failed product.

**A. Data Preprocess**

The practice of three-fold cross-validation was followed. Before the data set was processed, it was randomly shuffled and divided into training, cross-validation, and test sets. The training and the cross-validation sets were used for making and evaluating the models. The test set was used only for the evaluation of the final model. The percentage of the 3 sets out of the total dataset are as follow:

•Training: 50%

•Cross-validation: 25%

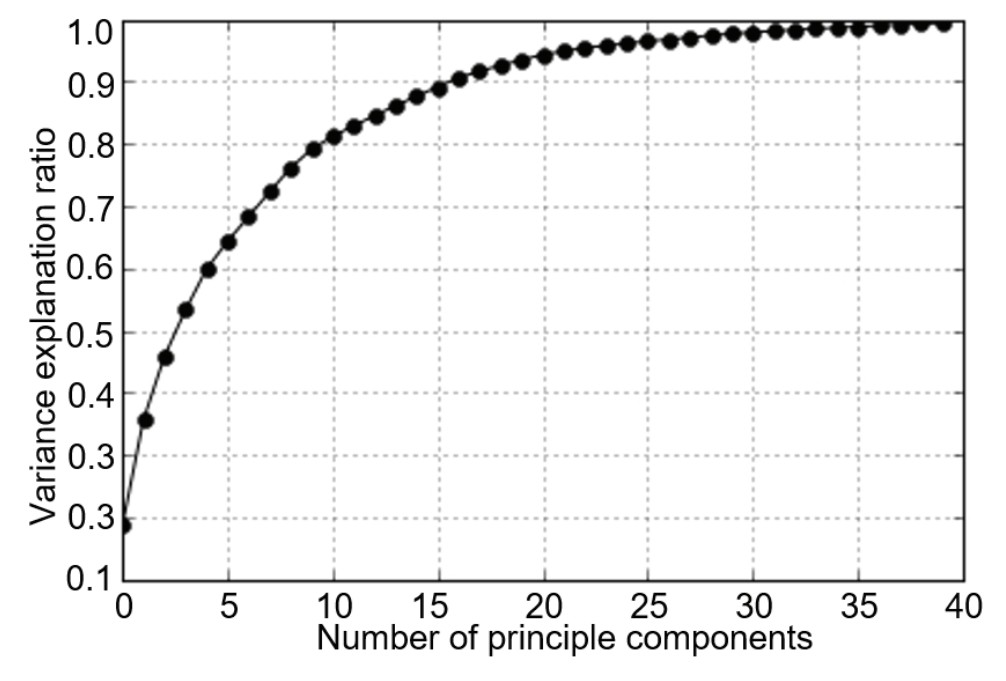
•Test: 25%

**STAGE I: CLUSTERING**

The goal of this step is to cluster data into groups by manufacturing process. The values of the features are first converted into binary, where 1 indicates a value and 0 a missing feature.

The dataset contains 968 features. However, many of the features are correlated because they are measured in the same production stations or production lines. The Principal Component Analysis (PCA) is conducted to reduce the dimension of the features. The PCA transforms the original correlated features into linearly uncorrelated principal components in reduced dimension, while preserving the largest variation.

The variance explanation rate of the PCA is given. Although the dataset has about 1000 features, the processes are greatly correlated. 95% of the variance is explained by 22 principal components. The reduced dimension is chosen to be two for the needs of data visualization.



The next step is to use the unsupervised algorithm so that data can be divide into clusters. The K-mean algorithm is selected. The algorithm iteratively calculates the distance between the data to centroids and assigns the data to the nearest centroid, until it converges to a local optimum. The value of K is determined by measuring the inertia within the clusters. As shown in Fig. below, the mean inertia decreases rapidly at first, then the rate of decrease slows after six clusters. Therefore, the number of clusters are chosen to be six.

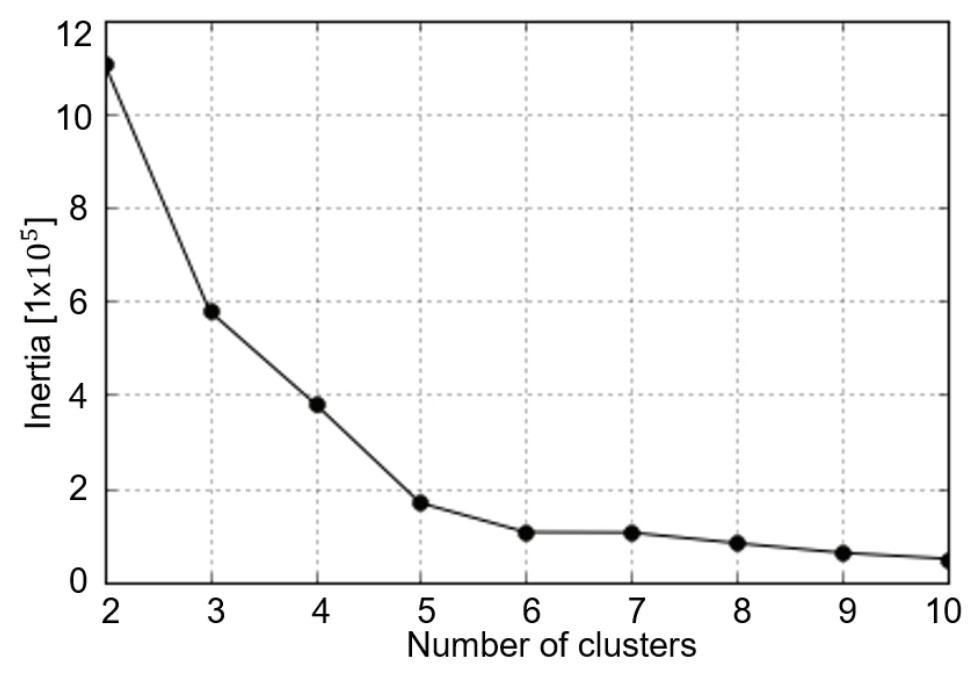
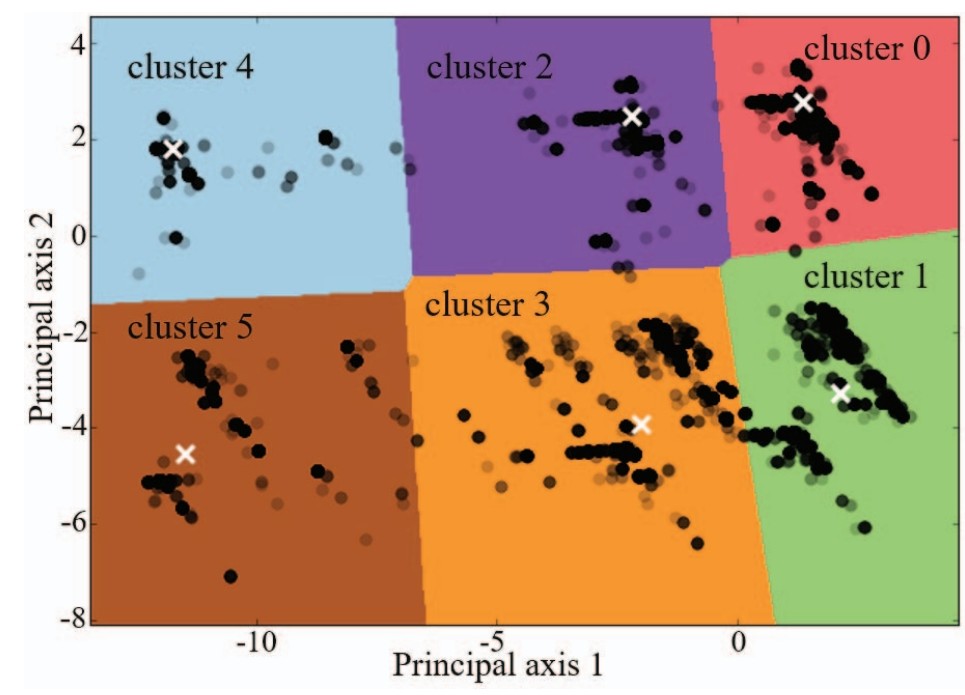
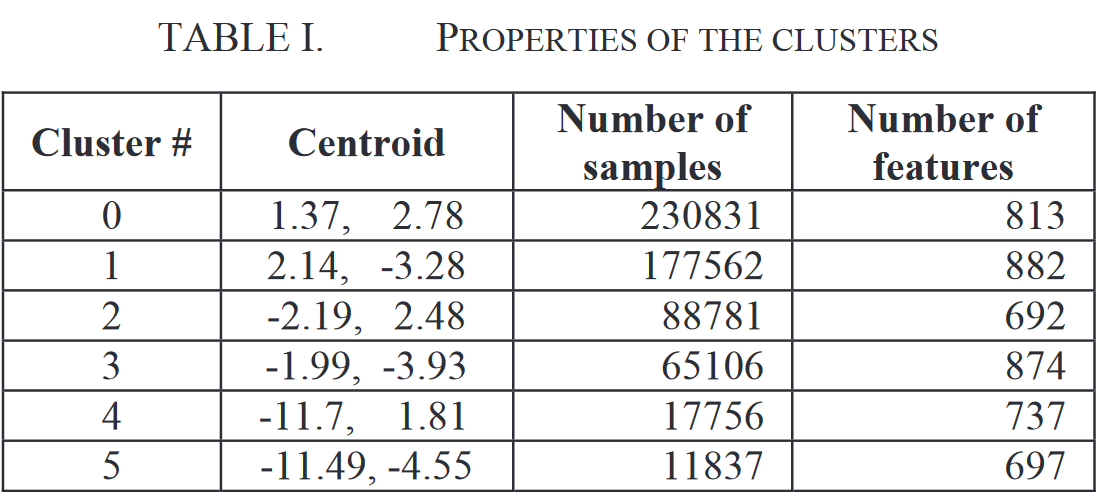


Fig below shows the scatterplot of the data in the first two principal dimensions. Each dot represents a unique process, which involves a different combination of features. The transparency of the data indicates the quantity of data in each process. The darker the colour, the more data is in the process. The centroids of the clusters are shown as the cross marks. The regions of the different clusters are denoted with different colours.



The features having all missing and constant values in each cluster are eliminated. 15.8% of the total dataset is removed. The properties of each cluster are shown in Table1 below. The amount of data in each cluster varies greatly. The first two clusters consist of nearly 70% of the total data.



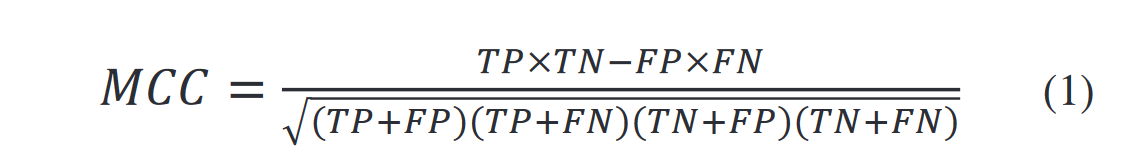
The following step is done to replace the values which are still missing. Since the missing value is caused by different manufacturing processes and is not missing at random, they are replaced with the special value “-1”.

**STAGE II: SUPERVISED LEARNING**

In the first step, the data is divided into several clusters. The next step is to utilize the supervised learning algorithms in order to predict the failed product in each cluster. Each cluster is treated as an independent dataset, where different algorithms and parameters can be chosen independently of that of other clusters.

A. Performance Metrics

The performance metrics need to accommodate for the largely imbalanced data. The Matthew's Correlation Coefficient (MCC) is chosen as the metric, which is defined as:

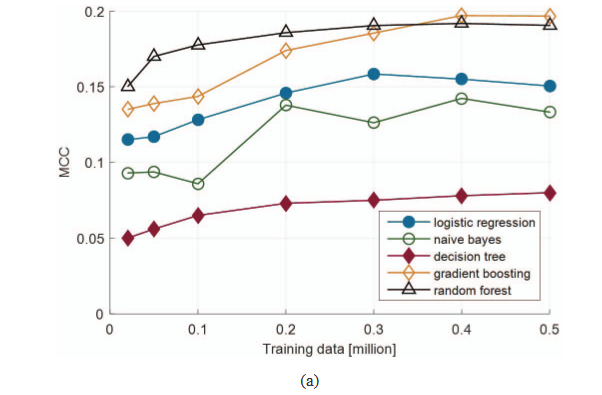


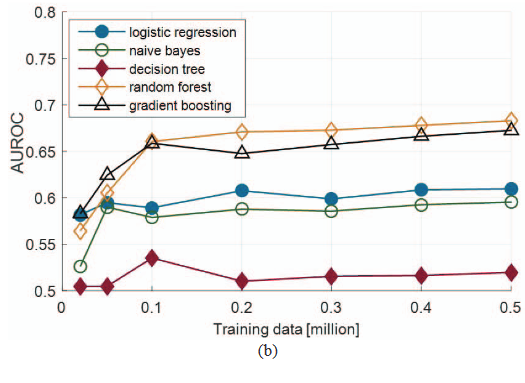
Where TP, TN, FP, FN represent True Positive, True Negative, False Positive and False Negative values in the confusion matrix. The MCC score ranges from zero (random guess) to one (perfect classification). MCC is calculated from the binary prediction results directly. A threshold is needed to convert the probability into the binary result. The threshold was optimized to the highest MCC score on the cross-validation set.

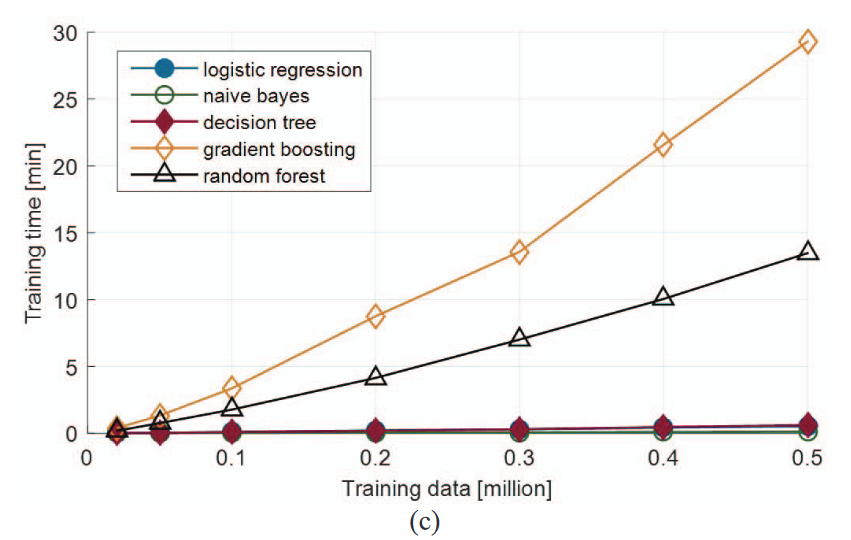
The Area Under the Receiver Operating Characteristic Curve (AUROC) [15] is used as an additional metric for the robustness of the algorithm. The classification results are provided as the probability of whether or not the data belongs to one class. The ROC curve is made by plotting the true positive (TP) rate against the false positive (FP) rate at various threshold settings. The AUROC mainly ranges from 0.5 (random guess) to one (perfect classification.).

**Algorithm Comparison**

Choosing the supervised learning algorithm is totally based on performance and computation time. We compared the several widely-used algorithms. The machine learning algorithms were implemented using the Scikit-Learn package. MCC and AUROC both are used as the performance metric. The training time is been used as the metric of computation time. The results of the whole data set are shown in Fig. 6. Both the MCC and AUROC scores increase as the training data size increases; however, the rate of increase slows down after 100,000 observations. The ensemble methods (random forest and gradient boosting) present better performance than those of simple classifiers (logistic regression, naïve Bayes, and a decision tree). The computation time of the ensemble methods increases linearly with the size of the data, which is being desired for the large dataset.



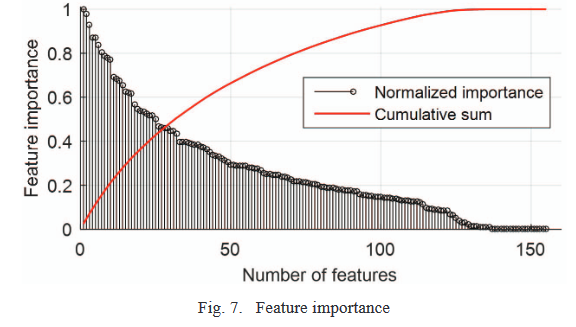




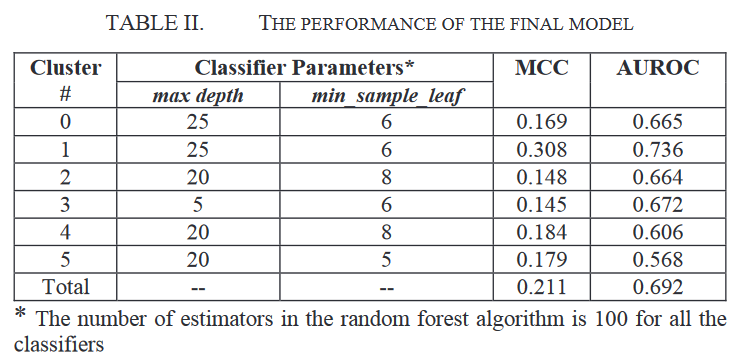
The random forest algorithm is chosen by considering the performance and training time. The algorithm is using multiple decision trees to construct the classifier to overcome the problem with overfitting by an individual decision tree.

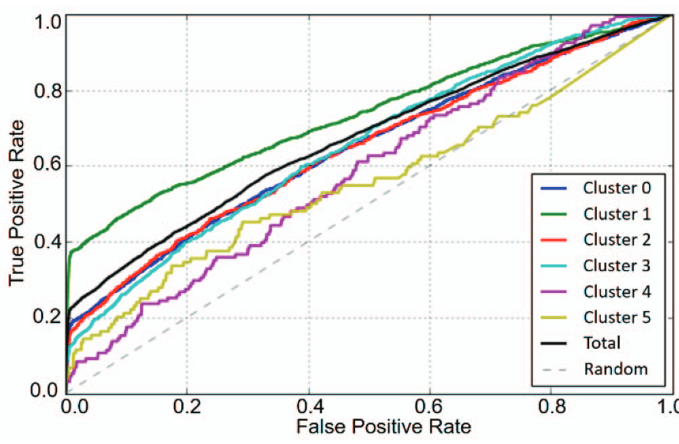
**Feature Selection**

The original dataset contains hundreds of features. However, the features contribute differently to predicting the target and most of the features are irrelevant. To further reduce the data size and speed up the machine learning algorithm, the features with less importance are truncated. The importance of feature can be measured by assessing the frequency of the feature being used as a node in a decision tree. The more frequently a feature is being used as a node in a decision tree, the more important that feature is. The ensemble method, like the random forest, is capable of ranking the feature importance by averaging the feature importance of each tree. Fig. 7 shows relative feature importance ranked up by the random forest classifier for the data in cluster 1. The features are selected until 95% of the cumulative sum of the feature importance has been reached and The same process is conducted for the other five clusters. In addition, the features with relatively high importance often indicate which station is problematic and thus can provide insight for quality assurance.



During the prediction step, the algorithm first determines which cluster the data belongs to and then predicts whether the product is a failure by using the classifier for the responding cluster. The final performance score for each cluster and the total data set is shown in Table II and the ROC curve is shown in Figure.





The AUROC score ranges from 0.5 (random guess) to one (perfect classification); the score of 0.69 is a relatively low score, which indicates the complexity of this production failure prediction problem. The slope of the ROC curve is steep in the beginning but soon flattens, which reveals that a small portion of the positive data is easy to classify, but most of the data is hard to classify. The performance of each cluster also varies highly. The data in Cluster 1 is much easier to classify and the data in Cluster 3 is the most difficult to classify.

**FUTURE WORK**

Large-scale manufacturing data is being generated as processes becoming more intelligent. The emergence of big data provides both the opportunity for using predictive models in quality assurance and the challenge for data processing. In this paper, we proposed a two-stage method to predict manufacturing failures in a production line. During the first step, the data is clustered into similar process groups. Then, supervised learning technique is applied to each cluster to predict product failure. This approach reduces the sparsity of the data set by eliminating irrelevant features. In future work, the parameters of the classifiers can be optimized to increase performance. Meanwhile, additional models will be considered. In addition to the numerical data, the categorical data and the timestamp data can also be used to provide additional information.

**REFERENCES**

[1] A. Zaslavsky, C. Perera, and D. Georgakopoulos, “Sensing as a Service and Big Data,” Proc. Int. Conf. Adv. Cloud Comput., pp. 21–29, 2012.

[2] C. Eaton, D. Deroos, T. Deutsch, G. Laipis, and P. Zikopolos, Understanding big data:Analytics for enterprise class hadoop and streaming data., vol. 11, no. 1. 2011.

[3] J. A. Harding, M. Shahbaz, Srinivas, and A. Kusiak, “Data Mining in Manufacturing: A Review,” J. Manuf. Sci. Eng., vol. 128, no. 4, p. 969, 2006.

[4] A. K. Choudhary, J. A. Harding, and M. K. Tiwari, “Data mining inmanufacturing: A review based on the kind of knowledge,” J. Intell. Manuf.,vol. 20, no. 5, pp. 501–521, 2009.

[5] E. Wolbrecht, B. D’ambrosio, R. Paasch, and D. Kirby, “Monitoring and diagnosis of a multistage manufacturing process using Bayesian networks,”Artif. Intell. Eng. Des. Anal. Manuf., vol. 14, no. 1, pp. 53–67, 2000.

[6] W. Sun, J. Chen, and J. Li, “Decision tree and PCA-based fault diagnosis of rotating machinery,” Mech. Syst. Signal Process., vol. 21, no. 3, pp. 1300–1317, 2007.

[7] S. Cho, S. Asfour, A. Onar, and N. Kaundinya, “Tool breakage detection using support vector machine learning in a milling process,” Int. J. Mach. Tools Manuf., vol. 45, no. 3, pp. 241–249, 2005.

[8] D. Cabrera, F. Sancho, R. V. Sánchez, G. Zurita, M. Cerrada, C. Li, and R.E. Vásquez, “Fault diagnosis of spur gearbox based on random forest and wavelet packet decomposition,” Front. Mech. Eng., vol. 10, no. 3, pp. 277– 286, 2015.

[9] Y. Li, “Building a Decision Cluster Classification Model by a Clustering Algorithm to Classify Large High Dimensional Data with Multiple Classes.” Diss. Hong Kong Polytechnic University, 2010.

[10] D. Feldman, M. Schmidt, and C. Sohler, “Turning Big data into tiny data : Constant-size coresets for k -means , PCA and projective clustering,” Proc.Twenty-Fourth Annu. ACM-SIAM Symp. Discret. Algorithms, pp. 1434–1453, 2013.