International Journal of Advance Research in Engineering, Science & Technology(IJAREST),
ISSN(O):2393-9877, ISSN(P): 2394-2444,
Volume 2,Issue 7, July- 2015, Impact Factor: 2.125

FAULT PREDICTION BY USING DATA MINING APPROACH

Chanchal Chauhan[1], Dr. Parveen Kalra[2], Dr. C.S. Jawalkar[3]

[1]M.E. Student, Production & industrial engineering Department PEC University of technology
[2]Professor, Production & industrial engineering Department PEC University of technology
[3]Associate Professor, Production & industrial engineering Department PEC University of technology

Abstract

Data mining is an effective tool which can be used in decision making in the organization. In this research, data mining is applied in the JCB Ballabgarh plant for the prediction of faults in the parts of machinery manufactured at JCB. Data is collected and retrieved from the quality department of the JCB and then it is transformed into the format which can be used in data mining tool i.e. rattle. Many variables were added to the dataset and many steps were done to make data useful for data mining. Data mining algorithms were applied on the data which are decision tree, support vector machine (SVM), Artificial Neural Network (ANN). From the results it has been found that ANN showed very high accuracy in the fault prediction while decision tree was good at predicting but SVM showed poor performance in the fault prediction. It have been showed that data mining is an excellent tool for making prediction than the traditional statistical methods.

Keywords- *Data mining, Fault prediction, Artificial Neural Network, Decision tree, Quality control*

1. Introduction

Today, organizations depend heavily on computerized devices and information sources on internet. Consequently, huge amount of data is generated and stored. The IDC Digital Universe Study of 2014 [1] estimates that by 2020, amount of digital information stored will be 44 trillion gigabytes. Most of the data stored is unstructured and organizations have problems dealing with such large quantity of data. One of main challenges of today's organization is to extract meaningful information from stored data e.g. to identify bottlenecks, provide insights, anticipate problems, recommend countermeasures, record policy violations and streamline processes.

This research is done in the JCB ballabgarh, India plant which manufactures Liftall, the pick-&-carry crane. This facility also manufactures the BSIII compliant JCB ecoMAX engine, which is big on fuel savings and high on performance with 16 valve effort. In this plant faults are recorded in a periodic manner i.e. fault in three months from month in which machine is sold. It is denoted as T3. Similarly, fault in year which is denoted by T12. Also, faults are classified into two categories i.e. main fault and consequential faults. Main faults are the faults which occurred from the external factors and consequential faults occurred due to the main faults.

2. Data mining

There are many definitions of data mining available on various sources. The widely accepted definition of data mining is given by Fayyad et al. [2], which is Data mining, which is also referred to as knowledge discovery in databases, means a process of nontrivial extraction of implicit, previously unknown and potentially useful information, such as rules, constraints, and regularities from data in databases.

Data mining serves two goals [3]:

1. *Insight:* identify patterns and trends that are comprehensible, so that action can be taken based on the insight.

2. *Prediction:* a model is built that predicts (or scores) based on input data. For example, a model can be built to predict failure of component of machine. If the prediction is for a discrete variable with a few values, the task is called classification; if the prediction is for a continuous variable, the task is called regression.

3. Data Mining Process

Data mining is process, therefore it is necessary to understand every aspect of it in order to apply. The process consists of six steps or phases [4], as illustrated in Figure 1

It consists of six steps:

1. *Business understanding* – In this phase, we understand the requirements and objectives of the industry. Then define the problem to be solve by data mining
2. *Data understanding* – In this phase, we determine, how data is collected, and then get familiar with data (i.e. attributes etc.).
3. *Data preparation* – Database is reduced/cleaned or converted into final dataset.
4. *Modeling* – we apply different models (decision tree, Neural Network etc.) to the final dataset by changing their parameters.
5. *Evaluation* – In this stage, we built the model and thoroughly evaluate it, then check whether the objective is achieved or not.
6. *Deployment* – in this step, we deploy a report, or again do the whole process to achieve the objective. In many cases, it depends on the user.

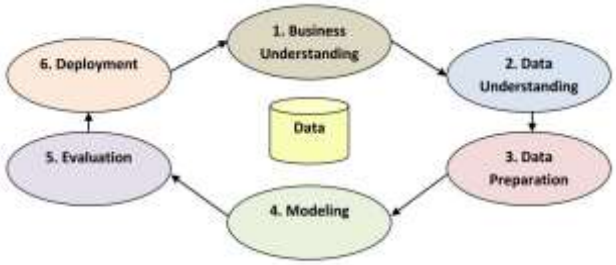

Figure 1. Phases of data mining Process

4. Data mining in quality improvement

Quality improvement through data mining is done by examine what has happened in past and then predict the future problems in the manufacturing system, by thoroughly

understanding the process. Through data mining we can identify the pattern or sequence of events that lead to the failure of product. We can also find out factors which influence the quality of product, thus, quality of product can be improved by managing these factors.

Ferreiro *et al.* [5] obtained a reliable model to detect the generation of burr during high speed drilling in dry conditions on aluminum Al 7075-T6. By using data mining he was able to eliminate the unproductive operations in order to optimize the process and reduce economic cost.

Apte *et al*. [6] applied rule induction, neural network, decision tree, and k-nearest neighbor in disk-drive manufacturing line to reduce the number of expensive tests while meeting the performance criteria. Lee *et al*. [7] used data mining to determine the optimal areas of inspection for a manufactured wafer was able save considerable time that is used in carrying out a 100% inspection of the semiconductor wafer. Oh *et al.* [8] used data mining in the process industry for quality improvement. They established the relationships between quality parameters and process, and identified the main causes of defects, which also provided optimized parameter adjustments.

Shi *et al.* [9] applied data mining in PCB manufacturing process to establish a relationship in nonlinear causes and effect**.** Lieber *et al*. [10] used data mining for predicting the physical quality of intermediate products in interlinked manufacturing processes. They were able to identify most striking operational patterns, promising quality-related features and production parameters. Liao *et al.* [11] applied data mining to model radiographic welding data. Shen *et al.* [12] used data mining to distinguish the fault type or to inspect the dynamic characteristics of the machinery. They demonstrated their approach for the identification of valve faults in a multi-cylinder diesel engine.

Kurasek *et al.* [13] used data mining to solve the quality engineering problems (solder ball defects) in the manufacture of printed circuit boards (PCB). They applied rough set theory to determine the causes of defects which needed further investigation. Skormin *et al.* [14] applied data mining for accurate assessment and forecasting of the probability of failure of hardware, such as avionics based on the historical data of environmental and operational conditions**.** Chen *et al*. [15] generated association rules for defect detection in semiconductor manufacturing. They determined the association between different machines and their combination with defects to determine the defective machine.

Kusiak *et al*. [16] applied neural network to analyze High temperature impacts the performance of turbine bearings. The five over-temperature events were tested using the normal behavior model. The over-temperature events were predicted 1.5 hr. ahead of the fault occurrence.

5. Methodology

Data mining is a method to find patterns in data and build predictions using these patterns. The general steps are first to describe the data in terms of its statistical attributes, visually look at charts and graphs to identify meaningful relationships among the variables, build predictive models based on the patterns found, test if the model appropriately predicts variables using known data separate from the data used to build the model, and finally verify the model with real data.

For this research, a similar approach is used and it is summarized in the following steps:

a) Define a data mining goal.
b) Clean the data and prepare it for analysis.
c) Load datasets to analyze.
d) Look for patterns and relationships.
e) Modify dataset as necessary.
f) Apply models to the dataset.
g) Assess how well these models fit the data.
h) Compare the data mining results to previous work.

Steps 2 through 6 can be repeated as necessary to develop a reasonable model and details of the steps are provided later.

5.1 Define a data mining goal

Before beginning the data mining process, a clear, well-defined objective must be established. The aim is to build a model that predict the percentage increase in fault of parts assembled in the machine. There are five data mining algorithms used for the fault prediction which are Decision tree, Random forests, Boosting, ANN, and SVM. Total 1013 parts are considered in study and 30 parameters are used to train the model. The decision was made to identify the parts in which 250 percentage increase in the fault in T12 with respect to T3.

5.2 Data Cleaning and Preparation

Even though data mining techniques are the core of knowledge discovery, most researchers agree that this core only takes 20% of the effort of the whole process while the other 80% effort is endeavored into data preparation [2], [17], [18]. Data cleaning and preparation is the second step in the data mining process. Relative to this research, step 1 entails consolidating the data into a single flat file with each row of data representing a single event.

The following are a list of tasks performed to prepare the Training data for analysis were:

1. Data from year 2012, 2013, and 2014 is taken for the analysis.
2. Zone in which fault is noted is added from the dealers details. Zones are classified into 8 categories.
3. Month in which fault occurred is added from warranty claim data.
4. Year in which fault occurred is added from warranty claim data.
5. Calculated numbers of day's machine is used in field.
6. Average usage of machine per day is calculated in hours per day.
7. Faults in T3 and T12 are added.
8. Main faults and consequential faults are added.
9. Faults in five variants of JCB are added.
10. Deleted parts with 60% or more missing values.
11. Replaced missing values in the remaining parts with zero after consulting authorities at quality department.
12. Percentage increase in the fault of each part in T12 with respected to T3 was calculated.
13. "Prediction" variable was added.
14. Parts with increase of 250 percent are classified as "YES" in the prediction variable.

15. Remaining parts are classified as "NO" in the prediction variable.

For testing dataset, data from 2105 year is considered. Step 2 to 11 were repeated and prediction variable was added and left blank for the testing.

5.3 Data Analysis

The data analysis steps 3 to 7 in this research's procedure are adapted from a process of Logical steps CRISP-DM developed. This iterative process of logical steps is designed to help the user apply the data mining tools in the Rattle software. These steps are load dataset, Explore, Modify, Model, and evaluate which are explained below:

- Sample entails choosing a subset of data that is large enough to contain all pertinent information, but small enough to process quickly. This subset is then divided into three subsets—training, validation, and test sets. The training set of data is used to fit the model, the validation set of data is used to prevent over fitting a model, and the test set is used to evaluate how well the model fits the data.
- Explore is the step to gain a better understanding of the data by identifying trends or anomalies in the data either visually or using statistical methods like cluster analysis.
- Modify entails changing the dataset by performing tasks such as creating new variables, eliminating other variables, and eliminating anomalies. The changes made in this phase are based on the discoveries made in the explore phase.
- Modeling the data is the step where different types of models are chosen for the software to fit to the data automatically.
- Assessing the data is the final step in the iterative process where one checks the validity of the results. This assessment is done by taking a test dataset and applying the model to these data to test if the model predicts the correct result.
- This process continues until the data miner is satisfied with the results.

6. Results

This section discusses the obtained results using Rattle (R) software, for the JCB Company. The results are analyzed and discussed in detail, in the form of graphs and tables.

Decision Tree:

Table 1.1 Confusion matrix for decision tree

	Predicted	
Actual	NO	YES
NO	80	10
YES	17	44

Confusion matrix in table 1.1 shows algorithm predicted correctly 124 parts correctly and 27 parts incorrectly. There are 17 parts in which 250 percentage increase in fault happened and 10 parts in there is no increase in fault, but predicted opposite.

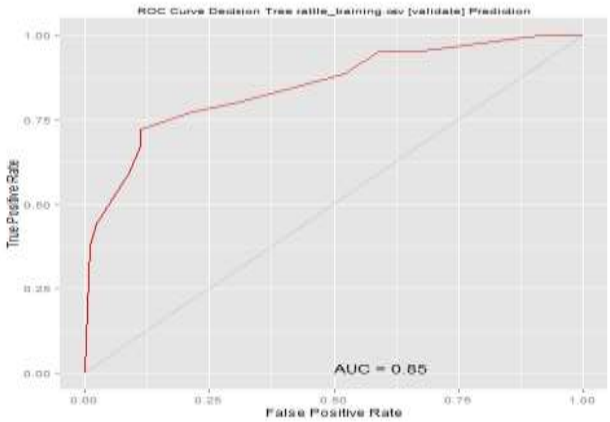

Figure 2.ROC Curve of decision tree

In figure 2 Receiver Operating Curve (ROC) of decision tree is shown. The area under curve (AUC) is 85% for the decision tree.

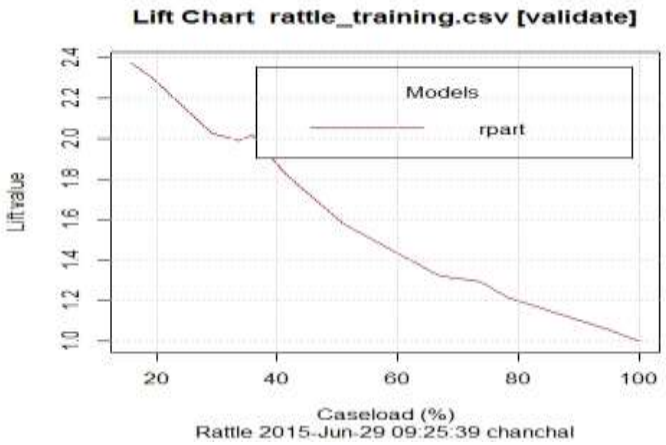

Figure 3. Lift Chart for decision tree

Lift chart of decision tree is shown in figure 3. Which shows that at 20% of dataset model have highest lift value and it decreases gradually.

Support Vector Machine (SVM):

Table 1.2 Confusion matrix for SVM

	Predicted	
Actual	NO	YES
NO	88	2
YES	32	4

Confusion matrix in table 1.2 shows algorithm predicted correctly 92 parts correctly and 36 parts incorrectly. There are 34 parts in which 250 percentage increase in fault happened and 2 parts in there is no increase in fault, but predicted opposite.

In figure 4. ROC curve for SVM is shown and it is observed that AUC is 68%.

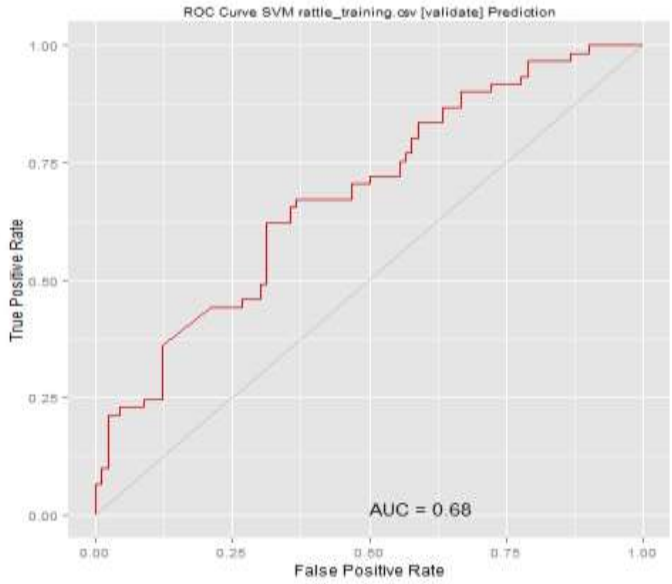

Figure 4. .ROC Curve of SVM

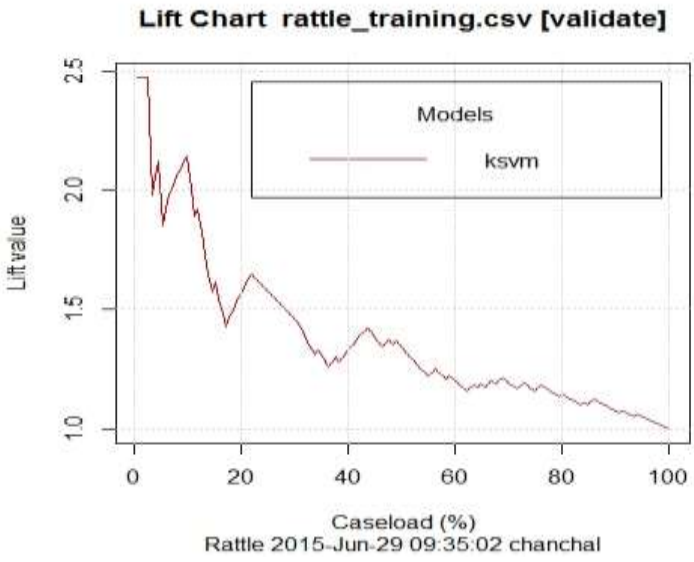

Figure 5. Lift Chart of SVM

In figure 5. Lift chart shows that at 15% caseload model have lift value 2 and then it decreases to 1.6 at 20 % caseload.

Artificial Neural Network (ANN):

The confusion matrix obtained is shown in Table 1.3. Confusion matrix shows algorithm predicted correctly 137 parts correctly and 14 parts incorrectly. There are 8 parts in which 250 percentage increase in fault happened and 6 parts in there is no increase in fault, but predicted opposite.

Table 1.3 Confusion matrix for ANN

	Predicted	
Actual	NO	YES
NO	84	6
YES	8	53

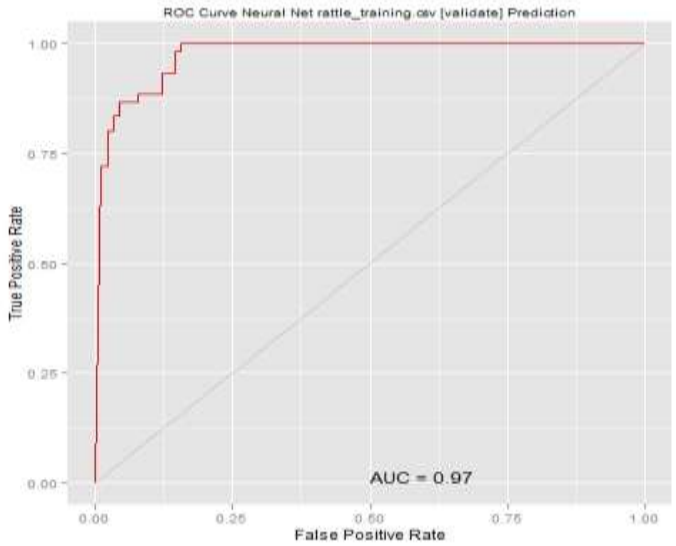

Figure 6. ROC Curve of ANN

In figure 6. ROC curve for SVM is shown and it is observed that AUC is 97%.

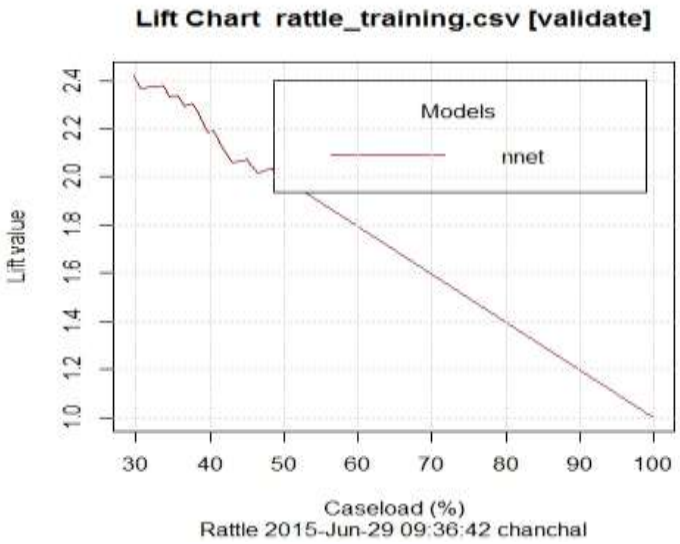

Figure 7. Lift Chart of ANN

In figure 7. Lift chart shows that at 30% caseload model have lift value 2.5 and at 50% caseload it is 2. Then it decreases linearly.

7. Analysis and Discussion on Performance of Algorithms

Results of algorithms have been shown above in terms of confusion matrix and charts. The accuracy of the models were compared by applying the models to all data subsets (train, validate, and test) in table 1.4 below.

Table 1.4 Accuracy of algorithms on validation and testing data sets.

Algorithms	Validation	testing
Decision tree	0.82	0.81
SVM	0.65	0.60
ANN	0.92	0.92

From table 1.4, ANN have highest accuracy on validation and testing datasets. Decision tree have good accuracy but SVM have lowest accuracy as compare to other algorithms used on datasets.

An ROC curve is the most commonly used way to visualize the performance of a binary classifier, and AUC (Area under curve) is the best way to summarize its performance in a single number**.** ROC shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. In table 1.5, comparison of algorithms is shown in terms of AUC.

Table 1.5 AUC of the algorithms in validation and testing datasets

Algorithms	Validation	testing
Decision tree	0.85	0.86
SVM	0.68	0.66
ANN	0.98	0.94

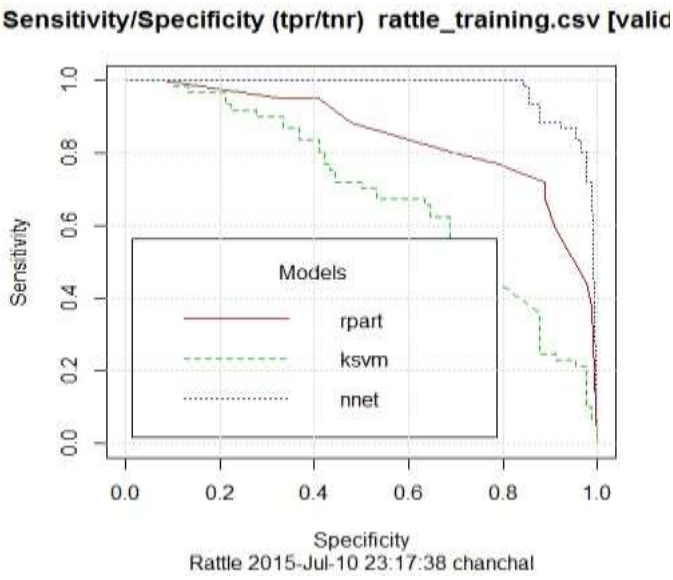

Figure 8. Sensitivity chart of all the models.

Since we cannot expect a classifier to be 100% accurate, some instances will not be correctly classified. These instances fall into two categories: Type I errors and Type II errors, or false positives and false negatives, respectively. A false positive is a non-faulty class erroneously classified as fault-prone, while a false negative is a faulty class that is misclassified as non-faulty. By varying the cut-off, one can to some degree control the ratio of false positives versus false negatives.

Sensitivity and specificity are more fine-grained measures that enable us to assess such a Trade-off between type I and type II errors. The former measure is the percentage of actual Positives that are correctly classified, i.e., in our context, the percentage of faulty classes classified as such. Sensitivity serves as a measure of how many faulty classes we are likely to find (or miss) if we use the prediction model.

From figure 8, it can be observed that ANN shows very good performance. Decision tree shows good performance. SVM shows poor performance than the other algorithms.

8. Conclusion

Data mining proved to be beneficial in both describing the failure events and building a fairly good model to predict the occurrence of fault.

Visually exploring the data via bar graphs did not yield any noticeable patterns. Very high accuracy have been achieved for the prediction of increase in the faults. Data mining provides information from the data bases which are cannot be obtained from other statistical methods. Several variables were considered for the prediction which cannot be done by other methods. ANN algorithm shows an excellent performance in fault prediction. Data mining shows great potential in the field of data analysis and decision making in organization to improve quality and maintenance of the product and services. It was found that, Artificial Neural Network and 94% accuracy. Support vector machine had lowest accuracy among the all the algorithms used in fault prediction. It had 34% overall error. Decision tree algorithm was able to achieve 86% accuracy.

This study focused on identifying parts in which increase in the fault would happen, however, there are other potential studies that would be beneficial to explore such as: Seasonal effect is can be taken into consideration while making prediction. Role of dealers can be studied. Data mining can be applied to the subsystems in machine separately to make prediction for e.g. electrical subsystem, hydraulic subsystem etc.

REFERENCES

[1] The Digital Universe and Big Data – EMC 2015. [ONLINE]. Available at : http://www.emc.com/leadership/digital-universe/index.htm. [Accessed 29 May 15].

[2] Fayyad U. M., Piatetsky-Shapiro G., Smyth P. and Uthurusamy, R., (1996), Advancesmin Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, CA.

[3] Ron Kohavi (2000). "Data Mining and Visualization". *Journal of National Academy of Engineering.*

[4] Chapman P., Clinton J., Kerber R., Khabaza T., Reinartz T., Shearer C., and Wirth R. (2000), CRISP-DM 1.0 Process and User Guide, www. Crisp-dm. org.

[5] Ferreiro, S., Sierra, B., Irigoien, I. and Gorritxategi, E. (2011). Data mining for quality control: Burr detection in the drilling process. *Journal of ELSEVIER, Computers & Industrial Engineering* 60 (2011), 801–810.

[6] Apte, C., Weiss, S., and Grout, G., 1993, "Predicting Defects in Disk Drive Manufacturing: A Case Study in High Dimensional Classification,"*IEEE Annual Computer Science Conference on Artificial Intelligence in Application, Los Alamitos, pp. 212–218.*

[7] Lee, J. H., and Park, S. C., 2001, "Data Mining for High Quality and Quick Response Manufacturing,"*Data Mining for Design and Manufacturing: Methods and Applications,* D. Braha, ed., Kluwer Academic, pp. 179–205.

[8] Oh, S., Han, J., and Cho, H., 2001, "Intelligent Process Control System for Quality Improvement by Data Mining in the Process Industry,"*Data Mining for Design and Manufacturing: Methods and Applications*, D. Braha, ed., Kluwer Academic, Dordrecht, pp. 289–310.

[9] Shi, X., and Boyd, P. S. D., 2004, "Applying Artificial Neural Network and Virtual Experimental Design to Quality Improvement of Two Industrial Processes," Int. J. Prod. Res., 42(1), pp. 101–118.

[10] Daniel Lieber, Marco Stolpe, Benedikt Konrad, Jochen Deuse, Katharina Morik (2013). Quality Prediction in Interlinked Manufacturing Processes based on Supervised & Unsupervised Machine Learning. *Journal of ELSEVIER, pp. 193-198.*

[11] Liao, T. W., Wang, G., Triantaphyllou, E., and Chang, P. C., 2001, "A Data Mining Study of Weld Quality Models Constructed With MLP Neural Networks From Stratified Sample Data,"*Industial Engineering Research Conference*, Dallas, TX, p. 6.

[12] Shen, L., Tay, F. E. H., Qu, L. S., and Shen, Y., 2000, "Fault Diagnosis Using Rough Set Theory," Comput Ind., 43, pp. 61–72.

[13] Kusiak, A., and Kurasek, C., 2001, "Data Mining of Printed Circuit Board," IEEE Trans. Rob. Autom., 17(2), pp. 191–196.

[14] Skormin, V. A., Gorodetski, V. I., and PopYack, I. J., 2002, "Data Mining Technology for Failure of Prognostic of Avionics," IEEE Trans. Aerosp. Electron. Syst., 38(2), pp. 388–403.

[15] Chen, W. C., Tseng, S. S., and Wang, C. Y., 2004, "A Novel Manufacturing Defect Detection Method Using Data Mining Approach," Lecture Notes in Artificial Intelligence, 3029, pp. 77–86.

[16] Kusiak. A and Verma, A., (2012), "Analyzing bearing faults in wind turbines: A data-mining approach". Renewable Energy, *Journal of ELSEVIER*, pp. 110-116.

[17] Becerra-Fernandez, I., Zanakis, S. H. & Walczak, S. (2002). Knowledge discovery techniques for predicting country investment risk. *Computer & Industrial Engineering*, 43(4), 787-800.

[18] Brachman, R. J, Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G. & Simoudis, E. (1996). Mining business databases. *Communications of the ACM* 3, 9(11), 42-48.