International Journal of Civil Engineering and Technology (IJCIET)
Volume 8, Issue 8, August 2017, pp. 731–739, Article ID: IJCIET_08_08_074
Available online at http://http://www.iaeme.com/ijciet/issues.asp?JType=IJCIET&VType=8&IType=8
ISSN Print: 0976-6308 and ISSN Online: 0976-6316

FEATURE SELECTION BASED ON CHI SQUARE IN ARTIFICIAL NEURAL NETWORK TO PREDICT THE ACCURACY OF STUDENT STUDY PERIOD

Otong Saeful Bachri
Artificial Intelligence Program, Department of Information Systems,
Faculty of Engineering, Diponegoro University, Jl. Prof. Soedarto, SH,
Tembalang, Semarang, 50275, and STIKOM POLTEK Cirebon, Indonesia

Kusnadi, Muhammad Hatta
Department of Information Systems, Diponegoro University,
Jl. Prof. Soedarto, SH, Tembalang, Semarang, 50275, and STIKOM Catur Insan Cendekia
Cirebon, Indonesia

Oky Dwi Nurhayati
Department of Information Systems, Diponegoro University,
Jl. Prof. Soedarto, SH, Tembalang, Semarang, 50275

ABSTRACT

A graduation prediction of student can be regarded as a nonlinear classification problem involving some academic parameters that has a goal to predict whether a student can graduate or not. One of the best classification methods which can be used to solve the problem is an Artificial Neural Network (ANN). Certainly, ANN needs a sufficient training data to obtain knowledge in order to classify the data since the inadequate training data can decrease the accuracy of the ANN. The accuracy can be increased by conducting a feature selection which becomes an input for ANN and eliminating the unimportant features which do not have correlation with the output of the ANN. In this research, the statistic approach using the Chi Square method was proposed to select the feature in student academic record data to be the input of ANN. The use of the Chi square succeeds to show which features having a significant influence towards the output of the ANN.

Key words: Neural Network, Feature Selection, Chi Square, Classification, Educational Data Mining.

Cite this Article: Otong Saeful Bachri, Kusnadi, Muhammad Hatta and Oky Dwi Nurhayati, Feature Selection Based On Chi Square In Artificial Neural Network To Predict The Accuracy of Student Study Period, International Journal of Civil Engineering and Technology, 8(8), 2017, pp. 731–739.
http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=8&IType=8

1. INTRODUCTION

The graduation rate of student is one of the indicators of the education process efficiency in a university, so that a serious effort should be done to push the student in finishing the study based on the assigned duration. One of the ways to realize the effort is by doing an early prediction towards the possibility of study success of each student. The prediction of student graduation involves many variables, such as gender, year of admission, GPA of every semester, and others, so as the problem is regarded as a nonlinear problem. One of the best methods in conducting the classification process and nonlinear regression is an Artificial Neural Network (ANN). The ANN can learn the pattern of certain data (training data) and do generalization process to the new data (testing data) that has not been studied previously [1]. However, the performance of ANN is influenced by some parameters, one of them is the input of variable (feature). Not all the data in the training data are used to train the ANN because not all the feature inside has a correlation with the target or output of the ANN. Therefore, the choice of the feature should be done to get an appropriate feature that is preferably used as the input of ANN. Indeed, according to May, et.al. [2], the choice of the feature is very fundamental in a statistic modeling like ANN. Moreover, the choice of the feature is certainly potential to increase the accuracy of the ANN and decrease the computational cost, especially in the process of ANN training [3]. The incorrect choice of the feature, that is for the feature which does not have correlation with the output system, will certainly decrease the accuracy of ANN [4]. In previous research, Bailly, et.al [5] proposed a method called Fuzzy Functional Criterion (FFC) to do the feature selection process in image data in order to finish the head pose estimation problem. The method merged with Generalized Regression Neural Network (GRNN) succeed to increase the recognition accuracy towards the picture, even it surpasses Convolutional Neural Network (CNN). Hassan, et.al [4] proposed the use of Input Significance Analysis (ISA) to select the feature in defective prediction data in a metal. ISA can do a manipulation of network weight in ANN and determine which features having a strong correlation to the target or output of the ANN. An approach uses Evolutionary Computation, that is by utilizing Harmony Search (HS) method [6] or Genetic Algorithm (GA) [7]. Harmony Search (HS) or Genetic Algorithm (GA) can do a combinatorial optimization to find an appropriate combination of feature, so that it can improve the classification accuracy in ANN.

However, those methods are not capable of displaying the significance of a feature to the target. Therefore, the statistical approach, by using the Chi Square method, is believed to be the right choice to get the significance value of each feature. The Chi Square method is also proposed by researcher to select the feature in some cases, such as text classification [8-11], intrusion detection in a computer network [12], and predictions of student performance [40].

The main purpose to be achieved in this research is to analyse what academic feature that has a significant influence on the graduation of student. In this research, Chi Square was used as a method for feature selection process in training data and testing data of the ANN to predict the accuracy of the student study period. Chi Square calculated the significant value of each feature towards the target (output of the ANN). A feature having a significant value of 0.05 would be chosen, while the other would be eliminated.

This research did not discuss the process of classification of data, but only focused on the stage of selection and analysis of the significance of the feature, in which the data used is the academic record of diploma students.

This research is expected to contribute to the application of Chi Square method to feature selection process in educational data mining area, especially to predict students' graduation.

2. MATERIAL

Artificial Neural Network

Artificial Neural Network (ANN) is a method in a Machine Learning area that has a significant development. ANN, basically, is a mathematical model inspired by the the way of brain nerve in processing information [1, 13-15]. ANN has been implicated in some fields, such as handwriting recognition [16-18], face recognition [19-22], voice recognition [23-27], weather forecast [28-33] and etc. An ANN architecture contains some sequence layers, such as input layer, hidden layer, and output layer. Each of the layers contains some neurons functioning as a data processor unit that are interconnected with neurons on the front layer. Figure 1 illustrates an archictecture of a Multi Layer Perceptron (MLP) which is the most popular of the ANN model and has an excellent performance [34].

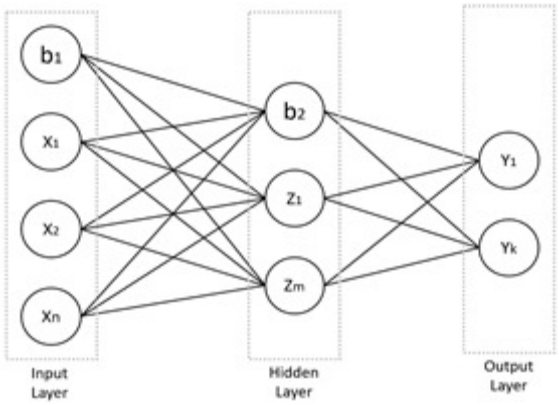

Figure 1 Multi Layer Perceptron (MLP) Architecture

Architecture The ANN can learn and recognize a pattern based on the acquired knowledge through training process called learning algorithm [1]. In every time the ANN does a training process, the network weight, which connects the neurons to every layer will be updated. Every neuron in the hidden layers and output layers does a calculation process to obtain an output in the form of the activation value using sigmoid activation function formulated by equation (1)

$$f(net) = \frac{1}{1+exp^{-net}} \quad (1)$$

where,

$$net = \sum_{i=1}^{n} x_i . w_{ij} + b_i \quad (2)$$

x is a vector input, w is a weight vector connecting two layers, and b is a value bias. The calculation process is conducted in stages, started by counting the activation value of each neuron in the hidden layer, then the value is used for calculation process of the activation value in the output layer.

Feature Selection

Feature selection is a fundamental means to minimize the total variable used in an evaluation or data analysis, especially in the multi-feature data. The main purpose is to increase the system accuracy, and to decrease the computational cost because of the overuse data. The selection process is conducted by selecting every relevant feature, that is for the input feature having a correlation to the target (output) from the system [35-36]. Statistic approach is one effective way to do a feature selection process within the data. In this research, Chi Square is chosen as a feature selection method since it has an excellent performance, especially in multi-class data [12]. The method has been used in some applications, such as tumor classification [37], network intrusion detection [12, 38] , text classification [11], disease diagnosis [39] and etc. Chi Square calculates the correlation strength of each feature individually by calculating the statistical value showed in equation (3).

$$\chi^2 = \sum_{i=1}^{n} \frac{(E_i - O_i)^2}{E_i} \quad (3)$$

where, EI is an expectation value of the –I feature appearance in a certain class, while oil is an actual appearance value of the –I feature in a certain class. Furthermore, the correlation of the significant value can be calculated based on $\chi 2$ value referring to the Chi Square distribution table. If the signed value is smaller than a crisis point, that is 0.05, then the feature has a strong relevance in in data, in other words, the feature is an important feature. Figure 2 shows a Chi Square distribution graphic for df = 1.The df value is obtained by decreasing the amount of the target class (in this case, target class = ‘’graduated’’ or ‘’not-graduated’’) with 1 (df = n – 1). Figure 2. Chi Square distribution graphic for df = 1.

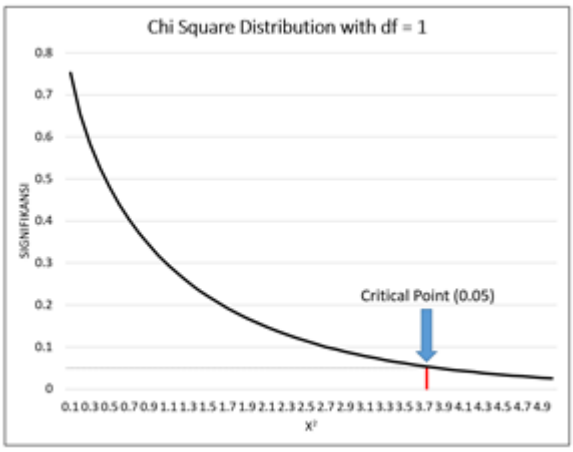

Figure 2 Chi Square distribution graphic for df = 1.

3. DATA

The data used in this research were the student academic record of STIKOM Poltek Cirebon containing 292 records. The raw data consisted of eight features involving Student Registration Number, Name, Gender, Year of Admission, the 1st semester GPA (IP 1), IP 2, IP 3, and graduation status (graduated/not).

4. RESULTS

The first step in order to do Chi Square calculation was by cleaning the data from features that did not have a correlation with the data analysis process, so that the Student Registration Number and Name were eliminated because they did not influence the duration of the study period. Therefore, there were only six features processed to the next step. Of the six features,

five features were the variable input whose correlation was calculated towards the sixth feature (graduation status) which became the target (variable output) using the Chi Square method. The correlation analysis was conducted independently in every variable input towards variable output (graduation status) to see whether each variable input was correlated with the expected output. The existence of the correlation between variable input and variable output was marked by the significant value of <0.05. Table 4.1 shows the significant value of the correlation between each variable input calculated by Chi Square.

TABLE I Significant Value of Each Variable Inputs Towards Variable Outputs

No	Variable Input	Value χ^2	Significance	Conclusion
1	Gender	0.038	0.8454	Uncorrelated to the output
2	Year of Admission	65.80	1.74×10^{-13}	Strongly correlated to the output
3	GPA of Semester 1	51.77	3.35×10^{-11}	Strongly correlated to the output
4	GPA of Semester 2	122.49	2.25×10^{-26}	Strongly correlated to the output
5	GPA of Semester 3	145.40	2.59×10^{-31}	Strongly correlated to the output

Of the five variable inputs processed by Chi Square, only gender feature that does not have correlation towards graduation status. It is marked by the significant value of > 0.05. Besides, it can be observed that year of admission evidently has a strong influence to the determination of student graduation showed by a large $\chi 2$ value and significant value of < 0.05. The larger $\chi 2$ value is, the smaller significant value is resulted. The $\chi 2$ value in every semester experiences an increase along with the increase of the semester, meaning that GPA in every higher semester has a stronger correlation to determine the student graduation. If it is seen on the whole, GPA of semester 3 has the strongest $\chi 2$ value of the other input variables. This indicates that the variable has a strong influence in determining the student graduation from the whole existence of input variables. The $\chi 2$ value above is totally different compared to the $\chi 2$ value in a gender variable which only 0.038.

Figure 3 can explain why the $\chi 2$ value in gender variable is very small and conversely, why the value is very large in IP 3.

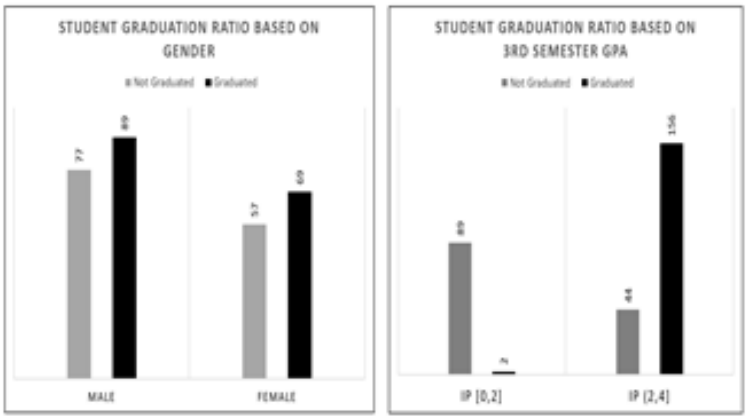

Figure 3 The comparison of the student graduation ratio based on gender and GPA of semester 3 (IP 3).

The small difference between the number of graduated and not-graduated students makes the gender variable does not have influence in determining the student graduation. It is different from the graduation ratio based on IP 3 which has a large difference between the number of graduated and not-graduated students. Students whose GPA is from 0.00 to 2.00 tend to not graduate, while the other whose GPA are more than 2.00 tend to graduate (on time). This is the reason why the gender variable does not have correlation towards variable output.

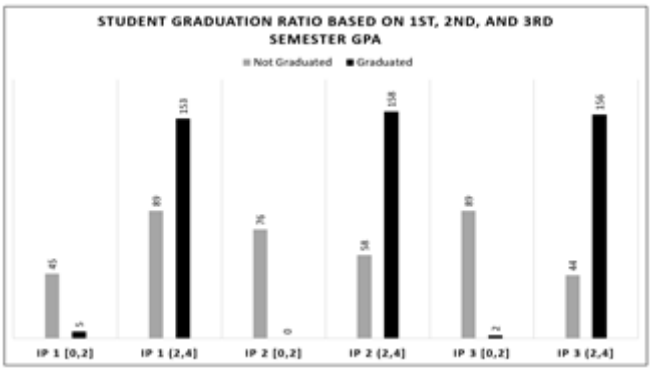

Figure 4 Student graduation ratio based on the 1st, 2nd, 3rd semester GPA (IP 1, IP 2, IP 3).

The comparison of student graduation ratio based on IP 1, IP 2, IP 3 presented in figure 4 shows a large distance between the graduated and not-graduated student based on each GPA.

In general, the results of this study have different results compared with similar research in the field of educational data mining. These differences caused mainly by the different feature on the data used from the beginning phase. For instance, Doshi, et.al.' [40] study applying Chi Square as one of the feature selection methods in their research on prediction of student performance, is more emphasis on the process of admission of students in a college. While other studies have not been discussed about feature selection for the determination of student graduation as discussed in this paper.

5. CONCLUSION

Chi Square succeeds to show a significant value of every feature by conducting a correlation test between each variable input and variable output. Based on the conducted test, it can be concluded that variable which can be used as the input of Artificial Neural Network (ANN) is year of admission, IP 1, IP 2, and IP 3, while gender variable can be ignored. Besides, each variable which has a larger $\chi 2$ value, then the significant value of the variable towards the variable output also becomes larger, and so does it conversely. A large $\chi 2$ value also shows the ratio between one class to the other class (in this case, ''graduated'' and ''not-graduated'' class), meaning the different number of data in one class to the other class is large.

REFERENCES

[1] Haykin, S. (2009). *Neural Network and Learning Machines – Third Edition*. New Jersey: Pearson Prentice Hall.

[2] May, R., Dandy, G., Maier, H. (2011). *Review of Input Variable Selection Methods for Artificial Neural Networks, Artificial Neural Networks - Methodological Advances and Biomedical Applications*. IntechOpen.

[3] Miao, J., Niu, L. (2016). *A Survey on Feature Selection*. Information Technology and Quantitative Management 2016, 919-926. doi: 10.1016/j.procs.2016.07.111.

[4] Hassan, R., Hassan, W.H., Al-Shaikhli, I.F.T., Ahmad, S., Alizadeh, M. (2014). *Feature Ranking through Weights Manipulations for Artificial Neural Networks Based Classifiers*. 2014 5th Int'l Conf. On Intelligence Systems, Modelling, and Simulation, 148-153. doi: 10.1109/ISMS.2014.31.

[5] Bailly, K., Milgram, M. (2009). Boosting Feature Selection for Neural Network Based Regression. *Neural Networks, Elsevier, 22* (5-6), 748-756. doi: 10.1016/j.neunet.2009.06.039.

[6] Das, S., Singh, P.K., Bhowmik, S., Sarkar, R., Nasipuri, M. (2016). *A Harmony Search Based Wrapper Feature Selection Method for Holistic Bangla Word Recognition.* 12th Int'l Multi-Conference on Information Processing 2016, 395-403. doi: 10.1016/j.procs.2016.06.087.

[7] Emmanoulidis, C., Hunter, A., Macintyre, J., Cox, C. (2001). A Multi-Objective Genetic Algorithm Approach to Feature Selection in Neural and Fuzzy Modeling. *Evolutionary Computation: An International Journal on the Internet*, *3*(1), 1-26.

[8] Zareapoor, M. (2015). Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection. *Int'l. Journal of Information Engineering and Electronic Business*, *2*, 60-65. doi: 10.5815/ijieeb.2015.02.08.

[9] Sarkar, S.D., Goswami, S. (2013). Empirical Study on Filter Based Feature Selection Metdhods for Text Classification. *International Journal of Computer Applications*, *81*(6), 38-43.

[10] Thabtah, F., Eljinini, M.A.H., Zamzeer, M., Hadi, W.M. (2009). Naive Bayesian Based on Chi Square to Categorize Arabic Data. *International Business Information Management Association*, *10*, 158-163.

[11] Mesleh, A.M. (2007). Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System. *International Journal of Computer Science*, *3*(6), 430-435.

[12] Ikram, S.T., Cherukuri, A.K. (2016). Intrusion Detection Model Using Fusion of Chi-Square Feature Selection and Multi Class SVM. *Journal of King Saudi Saud University – Computer and Information Science*. doi: 10.1016/j.jksuci.2015.12.004.

[13] Agrawal S., Agrawal, J. (2015). *Neural Network Techniques for Cancer Prediction: A Survey*. 19th Int'l Conf. on Knowledge Based Intelligent Information and Engneering Systems, 769-774. doi: 10.1016/j.procs.2015.08.234.

[14] Zhang, G., Patuwo, B.E., Hu, M.Y. (1998). Forecasting with Artificial Neural Networks: State of The Art. *International Journal of Forecasting*, *14*, 35-62.

[15] Fausett, L.V., Cliffs, E. (1994). Fundamentals of Neural Networks: Architectures, Algorithms, and Applications. New Jersey: Prentice Hall.

[16] Elleuch, M., Maalej, R., Kherallah, M. (2016). *A New Design Based-SVM of the CNN Classifier Architecture with Dropout for Offline Arabic Handwritten Recognition*. Int'l Conf. on Computational Science 2016, 1712-1723. doi: 10.1016/j.procs.2016.05.512.

[17] Biglary, M., Mirzaei, F., Neycharan, J.G. (2014). Persian/Arabic Handwritten Digit Recognition Using Local Binary Pattern. *Int'l. Journal of Digital Information and Wireless Communications*, *4*(4), 486-492.

[18] Kader, M.F., Deb, K. (2012). Neural Network-Based English Alphanumeric Character Recognition. *Int'l Journal of Computer Science, Engineering, and Applications*, *2*(4), 1-10. doi: 10.5121/ijcsea.2012.2401.

[19] Nandini, M., Bhargavi, P., Sekhar, G.J. (2013). Face Recognition Using Neural Networks. *International Journal of Scientific and Research Publications*, *3*(3), 1-5.

[20] Xu, Y., Zhang, X., Gai, H. (2011). Quantum Neural Networks for Face Recognition Classifier. *Advanced in Control Engineering and Information Science*, *15*, 1319-1323. doi:10.1016/j.proeng.2011.08.244.

[21] Agarwal, M., Jain, N., Kumar, M., Agrawal, H. (2010). Face Recognition Using Eigen Faces and Artificial Neural Network. *International Journal of Computer Theory and Engineering*, *2*(4), 624-629.

[22] Reddy, K.R.L., Babu, G.R., Kishore, L. (2010). *Face Recognition Based on Eigen Features of Multi Scaled Face Components and an Artificial Neural Network*. Int'l. Conf. on Biometrics Technology, 62-74. doi:10.1016/j.procs.2010.11.009.

[23] Hu, X., Lu, X., Hori, C. (2014). *Mandarin Speech Recognition Using Convolution Neural Network with Augmented Tone Features.* 9th Int'l Symposium on Chinese Spoken Language Processing, 15-18.

[24] Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., Yu, D. (2014). Convolutional Neural Networks for Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(10), 1533-1545. doi: 10.1109/TASLP.2014.2339736.

[25] Gevaert, W., Tsenov, G., Mladenov, V. (2010). Neural Networks Used for Speech Recognition. *Journal of Automatic Control, University of Belgrade*, *20*, 1-7. doi: 10.2298/jac1001001g.

[26] Alotaibi. (2008). Comparative Study of ANN and HMM to Arabic Digits Recognition Systems. *Journal of King Abdul Aziz University: Engineering and Science*, *19*(1), 43-60.

[27] Tebelskis, J. (1995). *Speech Recognition using Neural Networks*. Doctoral Thesis. Carniege Mellon University, Pennsylvania, USA.

[28] Grover, A., Kapoor, A., Horvitz, E. (2015). *A Deep Hybrid Model for Weather Forecasting*. 21st ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 379-386. doi: 10.1145/2783258.2783275.

[29] Malik, P., Singh, S., Arora, B. (2014). An Effective Weather Forecasting Using Neural Network. *International Journal of Emerging Engineering Research and Technology*, *2*(2), 209-212.

[30] Culclasure, A. (2013). *Using Neural Networks to Provide Local Weather Forecast*. Master Theses. Georgia Southern University, Georgia.

[31] Abhishek, K., Singh, M.P., Ghosh, S., Anand, A. (2012). *Weather Forecasting Model Using Artificial Neural Network*. 2nd Int'l. Conf. on Computer, Communication, Control, and Information Technology, 311-318. doi: 10.1016/j.protcy.2012.05.047.

[32] Baboo, S., Sheref, I.K. (2010). An Efficient Weather Forecasting System using Artificial Neural Network. *International Journal of Environmental Science and Development*, *1*(4), 321-326.

[33] Maqsood, I., Khan, M.R., Abraham, A. (2004). An ensemble of neural networks for weather forecasting. *Neural Computation and Application*, *13*, 112-122. doi 10.1007/s00521-004-0413-4.

[34] Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Singapore: Springer.

[35] Kumar, V., Minz, S. (2014). Feature Selection: A Literature Review. *Smart Computing Review*, *4*(3), 211-229. doi: 10.6029/smartcr.2014.03.007.

[36] Bolón-Canedo, V., Seth, S., Sánchez-Marono, N., Alonso-Betanzos, A., Príncipe, J.C. (2011). *Statistical Dependence Measure for Feature Selection in Microarray Datasets*. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 23-28.

[37] Zhang, H., Li, L., Luo, C., Sun, C., Chen, Y., Dai, Z., Yuan, Z. (2014). Informative Gene Selection and Direct Classification of Tumor Based on Chi-Square Test of Pairwise Gene Interactions. *Biomed Research International*, *2014*, 1-9. doi: 10.1155/2014/589290.

[38] Barot, V., Chauhan, S.S., Patel, B. (2014). Feature Selection for Modeling Intrusion Detection. *Int'l. Journal of Computer Network and Information Security*, *7*, 56-62. doi: 10.5815/ijcnis.2014.07.08.

[39] Dharmendra Kumar singh, Pragya Patel, Anjali Karsh, Dr.A.S.Zadgaonkar, Analysis of Generated Harmonics Due To CFL Load On Power System Using Artificial Neural Network, Volume 5, Issue 3, March (2014), pp. 56-68, International Journal of Electrical Engineering and Technology (IJEET).

[40] Dharmendra Kumar singh, Ekta Singh Thakur, Smriti Kesharwani, Dr. A.S.Zadgaonkar, Analysis of Generated Harmonics Due To Single Phase PWM Ac Drives Load On Power System Using Artificial Neural Network, Volume 5, Issue 2, February (2014), pp. 173-185, International Journal of Advanced Research in Engineering and Technology.

[41] Upendra R.S, Pratima Khandelwal, Veeresh A V, Application of Artificial Neural Network Statistical Design (Ann) In Enhanced Production of Biopharmaceuticals, Volume 6, Issue 3, March (2015), pp. 46-52, International Journal of Computer Engineering and Technology

[42] Son, C.S., Jang, B.K., Seo, S.T., Kim, M.S., Kim, Y.N. (2012). A Hybrid Decision Support Model to Discover Informative Knowledge in Diagnosing Acute Appendicitis. *BMC Medical Informatics and Decision Making*, *12*(17), 1-14. Doi: 10.1186/1472-6947-12-17.

[43] Doshi, M., Chaturdevi, S.K. (2014). Correlation Based Feature Selection (CFS) Technique to Predict Student Performance. Int'l Journal of Computer Networks & Communications, 6(3), 197-2016. Doi: 10.5121/ijcnc.2014.6315.