**How to Cite**:
Vinutha, D. C., Sushma, S. J., Assegie, T. A., Paramesha, K., & Geethashree, A. (2022). Optimization of RFM for automated breast cancer detection. *International Journal of Health Sciences*, *6*(S1), 5973–5982. https://doi.org/10.53730/ijhs.v6nS1.6218

# Optimization of RFM for automated breast cancer detection

**Vinutha D C**
Professor, Department of CSE (Artificial Intelligence and Machine Learning), Vidyavardhaka College of Engineering, Mysuru, Karnataka 570002, India
Email: vinuthadc@vvce.ac.in

**Sushma S J**
Associate Professor, Department of Electronics and Communication Engineering, GSSS Institute of Engineering and Technology for Women, Mysuru, Karnataka 570002, India
Email: enggsush@gmail.com

**Tsehay Admassu Assegie**
Department of Computing Science, 40 Injibara University, Injibara, Ethiopia,
Email: tsehayadmassu2006@gmail.com

**K Paramesha**
Professor, Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka 570002, India
Email: paramesha.k@vvce.ac.in

**Geethashree A**
Associate Professor, Department of Electronics and Communication Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka 570002, India,
Email: geethashree.a@vvce.ac.in

***Abstract*---**Recent years have seen an upsurge in the acceptance of illness diagnosis and prediction utilizing ML algorithms. A ML model can be employed in the diagnosis of breast cancer illness. In this research, an effective breast cancer prediction model with grid search approach is provided. Using the random forest approach, grid search is used to find the best n-estimator, which may provide the highest possible accuracy for predicting breast cancer. The accuracy of the suggested model can then be utilised to contrast its effectiveness to that of a standard RFM. The experimental result analysis demonstrates that the optimized model has 97.07 percent accuracy whereas the regular random forest technique has an accuracy of 94.73 percent in breast cancer detection.

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
*Manuscript submitted: 09 Feb 2022, Manuscript revised: 27 March 2022, Accepted for publication: 18 April 2022*

***Keywords*---**Breast cancer prediction, Random forest, Optimization, ML (Machine learning), Grid search.

**Introduction**

ML plays a big role in figuring out if someone has breast cancer. This helps cut down on the number of people who die from breast cancer. Recent years have seen a steady drop in breast cancer deaths because of the work that is done in primary prevention, detection, and treatment. Doctors employ a variety of medical imaging techniques to make a "non-invasive" identification of breast tumors. Breast cancer imaging techniques like mammography, MRI, and ultrasound were examined by "Md.Shafiul Islam et al." for their effectiveness, benefits, and drawbacks in early stage breast cancer detection. When it comes to artificial intelligence and machine learning, "deep learning" is still in its infancy. It looks at these areas in order to figure out what features are in data directly. To classify different deep learning systems, Li Deng and Dong Yu looked at how they were classified into supervised, unsupervised, and hybrid deep networks based on the topologies and applications they were used for. It was done by Jawad Nagi and his team with the help of morphological preprocessing as well as the seeded region growing (SRG) method. The pectoral muscle was removed from the breast profile area. Handmade characteristics are retrieved from photographs using traditional methods. According to Maxine Tan et AL examinations of feature variations in a series of mammograms, a patient is likely to be diagnosed with breast cancer in the near future. Gray Level Co-occurrence Matrix (GLCM) and (RBFNN) were used by MellisaPratiwi et al. to help them classify breast cancer using mammography, and they used both. Traditional handcrafted features sometimes require a great deal of manual labour as well as often depend on the knowledge of a specialist to complete. This encourages the design of accurate and suitable feature learning algorithms. Grigoryet al. demonstrated that when the size of the datasets was limited, the performance of both the learning characteristics as well as the handcrafted features was consubstantial. Deep neural networks use auto encoders to learn features automatically and unsupervised. Kersten Petersen et al. automated mammographic risk scoring by estimating breast cancer risk based on the retrieved mammographic features. CSAE (Convolutional Sparse Auto Encoder) is used in this paper to train the features using sparse auto encoders within a convolutional architecture. An intelligent system which can detect breast cancer in its earliest stages is built using Algorithms. Early detection helps patients avoid serious complications and increases their chances of survival. For a disease to be correctly identified and precisely recognised, ML algorithms are widely used in medical research to identify it or predict illness [2]. However, the accuracy of these learning algorithms must be as high as is reasonably possible. Efficient rule sets are created in a specific manner. Random Forest-based rule generation is among the most efficient and reliable techniques to build rules. These models can undoubtedly demonstrate how people make judgments. The entire process can be demonstrated using the rules they devise. Random Forest is a good approach for learning about groupings of things. It excels at generalization, and there is still potential for development in terms of comprehension. In comparison to other decision tree ensembles, RF is believed to be more noise resistant. Bootstrap sampling and randomizing features are used to create RF, a classifier ensemble based on the Gini index. The use of decision trees to generate a random sample of feature splittings is also a possibility. Bootstrap sampling is used to build each tree, and as a result, no two trees are the same. However, standard ML algorithms use global variables that result in reduced sensitivity and efficiency in disease detection [3]. In this work, the standard ML algorithm's default parameter is modified using a grid search technique to uncover the ideal n-estimator for a RFA which is utilised in training to maximize the RFA's efficiency on breast cancer detection. Additionally, the purpose of this study is to address the following research questions:

1) Using a RFA, how can we make it better at predicting breast cancer?
2) With the help of a RFM, what is the best n-estimator to use for detecting breast cancer?
3) How can the RFM's parameters be selected to maximise model performance in the diagnosis of breast cancer?

**Literature Review**

This section focuses on examining scientific material relevant to the optimization of ML algorithms towards disease prediction and detection. Parameter tuning using grid search as well as random search approach is the optimization the. Most widely applied technique for parameter tuning [4-11]. The goal of parameter tuning is to find the most effective parameter for a machine-learning system. In [5], the authors have proposed a vector support as well as decision tree-based breast cancer prediction model. The presented model is evaluated on the test set with accuracy metric as well as the result of the experimental test demonstrates that the model worked well for breast cancer diagnosis. Support vector machines and decision tree models were evaluated to detect breast cancer, and results show that support vector machines outperformed decision tree models in this test. Overall, a classification accuracy of 91.92 percent is attained using the support vector machine, which shows that there is still potential for enhancing the accuracy of ML algorithms for breast cancer diagnosis. Weighted decision trees are used in another study [6]. The breast cancer dataset is used, and the model for finding breast cancer is then suggested. The experimental test on the suggested model demonstrates that the model worked very well breast cancer detection however there is headroom for enhancing the accuracy to a greater level. The best accuracy achieved by the suggested weighted decision tree model is 94.03 percent. The author of [8] applied the KNN algorithm to the breast cancer dataset and suggested a KNN-based breast cancer diagnosis model. The suggested model is evaluated on a test dataset as well as the performance evaluation appears to reveal that the KNN model has an accuracy of 94.35 percent to detect breast cancer. An accuracy score of 94.35 percent to detect breast cancer s suggests that there is still wider potential for enhancing the model accuracy to a greater level for better results.

**Research Methodology**

Datasets utilised for testing and training, programming language used to build the suggested model, and optimization approach used to select the appropriate n-estimator for the RFA are all covered in this part. Random forests are used to train and test for the identification of breast cancer using an online dataset. The data repository comprises 569 observations and 31 characteristics. Three hundred and fifty-seven normal observations and two hundred and twenty-two malignant or breast cancer patient observations compose the repository. A random forest technique is applied to construct a ML model for breast cancer prediction. For the most accurate model, a strategy known as grid search is used to find the best n-estimator. The python programming language is used to implement the proposed model.

In the evaluation of the suggested model performance measures such as prediction accuracy, learning curve, confusion matrix, and precision is utilised for validating the model on the test set. To determine the model's accuracy, divide its total number of forecasts by the number of correct predictions. The predictive accuracy of the model is calculated by the formula stated in equation (1).

$$\text{Predictive Accuracy} = \frac{\text{T}}{\text{N}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots .. (1)$$

Where: T= the total number of correct predictions, TP (true positive and TN (true negative), N= total number of predictions (incorrect and correct predictions).

**Description of Dataset**

The breast cancer dataset utilised in this work has two classifications, namely breast cancer positive class as well as breast cancer negative class. The association between the breast cancer dataset attributes is displayed in fig 1, utilizing a correlations circle graph as well as the dataset characteristics is described in table.

(Table 1)
Characteristics of datasets

| | |
|---|---|
| Number of instances | 569 |
| Malignant (cancerous observations) | 212 |
| Benign (non-cancerous observations) | 357 |

**Correlation model**

To understand the predictable relationship among the breast cancer dataset features we employed a correlation circle for exploring the feature relationship.

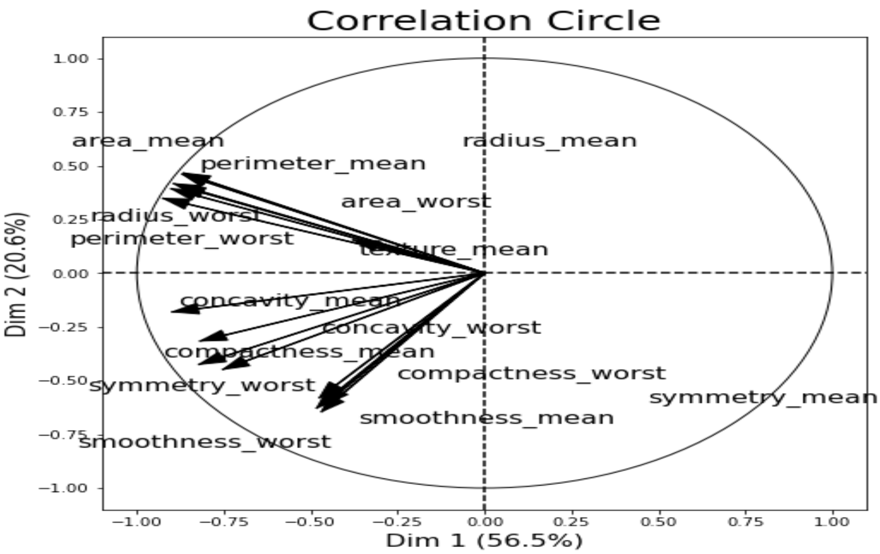


Figure 1. Correlation graph for breast cancer features

Figure 1, demonstrates the relationship between breast cancer dataset set features. Looking at the correlation circle shown in figure 1, we understand that the breast cancer features such as area, perimeter, symmetry, and smoothness have a strong positives relationship.

## Analysis of Results and Discussions

This section discusses the research's findings and experiments. The accuracy and confusion matrix of the suggested model is studied as well as the efficiency of the standard as well as the proposed optimal breast cancer detection model is examined.

## Accuracy of breast cancer prediction model

Figure 2 shows the proposed random forest-based model's accuracy in detecting breast cancer. The model is tested randomly on the test dataset but also the prediction outcome of the RFM is assessed for each new observation as well as the accuracy score is reported in fig 2 for five random tests conducted on the model.

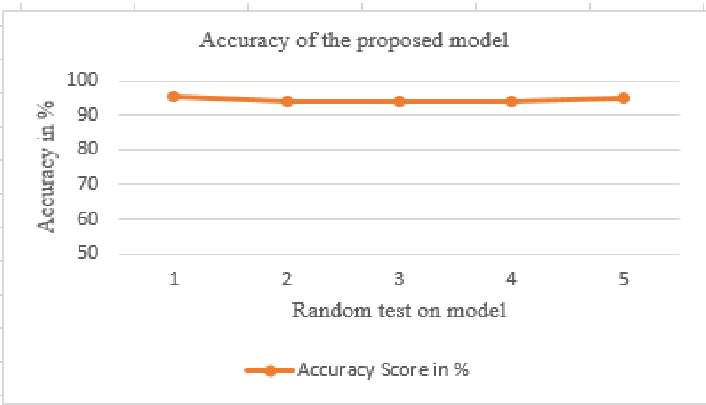


Fig 2. Accuracy of the proposed model

As demonstrated in fig 2, the proposed RFM performed efficiently to detect breast cancer. The accuracy of the model lies in the range of 94%-100% to detect breast cancer.

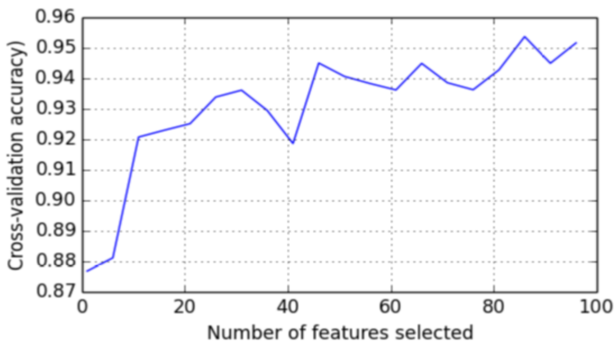


Fig 3. Accuracy vs number of features

Fig 3 demonstrates the accuracy of the proposed RFM against the number of features selected for training. As demonstrated in fig 3, the highest accuracy is achieved with 86 of the original 569 features (at the 20 percent percentile).

**Confusion matrix of the proposed model**

The confusion matrix for the suggested model is given in figure 4. The confusion matrix demonstrates the performance of the proposed model to detect breast cancer in terms of classification accuracy on each class of the observations in the breast cancer dataset. As seen in figure 4, the number of correct classifications performed by the proposed model is 109 and 4 observations are mistakenly identified as a breast cancer patient. The observations wrongly categorized as breast cancer negative is 1. The suggested model correctly classified just one of the 37 breast cancer-negative and 114 breast cancer-positive data that were included in the model's validation study.

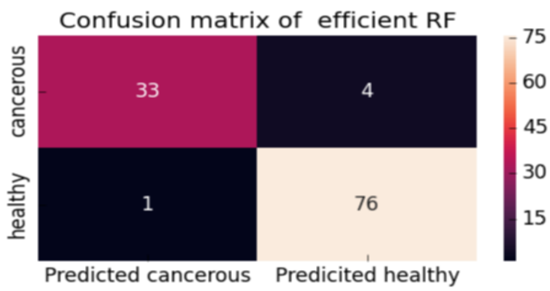


Fig 4. Confusion matrix for the proposed model

**Precision recall curve of the proposed model**

On the graph in figure 5, we can see the precision-recall curve for our model that predicts breast cancer well. A graph like this one, figure 5, shows that the model did a good job of predicting breast cancer. It shows that 96.2 percent of the model's precision was used to predict breast cancer positive class (class 1), and 97.1 percent of the model's precision was used to predict breast cancer negative class (class 2). (class 0). There is a better level of precision in the positive class than there is in a lower level of precision in a lower class.

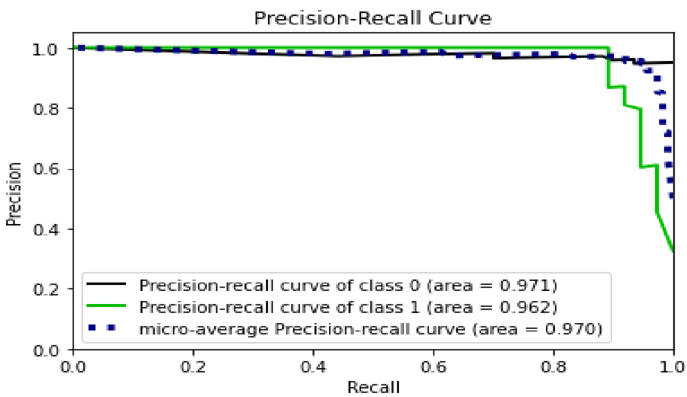


Fig 5. Precision recall curve

**Learning curve**

The learning curve is used to fig out if the model is over fitting or under fitting, so you can fig out if it is. Another thing to note is that the learning curve shown in figure 6 shows that the model's performance improves as the number of training samples grows. Overall, the best prediction accuracy is achieved when 70% of the samples are used for training. We can tell that the model learns better than it does when it comes to predicting unknown samples. Thus, the model doesn't have to deal with over fitting. Fig 6 shows the learning curve of the proposed model. We can figure out that the best split size for the train tests is 70 percent to 30 percent.

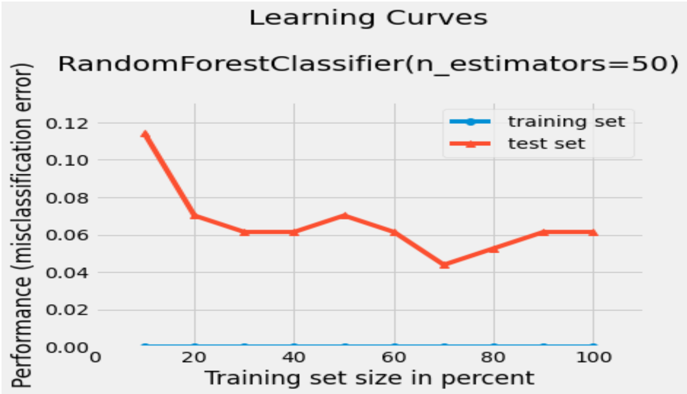

Fig 6. Model learning curve

## Conclusion

In this study, the strategies to develop a random forest-based model to identify breast cancer are looked at. The RFM was given a grid search to see if it could be made to perform better. It utilised this strategy to identify the optimal n-estimator for the model. Grid search is utilised to determine the best n-estimator, which is used to retrain the RFA on the breast cancer test set. On the test set, the model's performance is evaluated based on its accuracy and confusion. The RFM does well when it is trained on the test set using the best estimator. The new model's performance is compared to that of the classic RFM with this in mind. The new model's accuracy is 97.07 percent, whereas the traditional RFM's accuracy is 94.73 percent with the default settings. This means the new model is superior. Mammographic images and classification techniques like KNN, SVM, decision trees and random forest are used to detect breast cancer. This study makes use of publicly available mammograms from the Mini MIAS database. To improve the deep network's ability to identify cancer, mammograms are first normalized to remove artefacts such as digitization noise, radio-opaque artefacts, background, and pectoral muscle. In the classification of dense mammography pictures, the suggested model has achieved efficiencies of more than 98% for SAE spiralled over random forest.

## References

[1] P. Sahity Anaray Anan, Identification of Breast Cancer Using The Decision Tree Algorithm, proceedings of an international conference on systems computation, automation and networking, IEEE, 2019.

[2] R. Arulmozhiyal and K. Baskaran, "Implementation of a Fuzzy PI Controller for Speed Control of Induction Motors Using FPGA," *Journal of Power Electronics,* vol. 10, pp. 65-71, 2010.

[3] Shahrbanoo Goli, Hossein Mahjub, Javad Faradma, Hoda Mashayekhi, Ali-Reza Soltanian, Survival Prediction and Feature Selection in Patients with Breast Cancer Using Support Vector Regression, Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine Volume 2016.

[4] James Bergstra, Yoshua Bengio, Random Search for Hyper-Parameter Optimization, Journal of MLResearch, 2012.

[5] Assegie, T.A, Sushma S J, A Support Vector Machine and Decision Tree-Based Breast Cancer Prediction, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-3, February 2020.

[6] Assegie, T.A, Sushma S.J, Dr. Prasanna Kumar S C, Weighted Decision Tree Model to detect breast cancer , Technology Reports of Kansai University, Volume 62, Issue 03, January 2020.

[7] Assegie, T.A, An optimized K-Nearest Neighbor based breast cancer detection, Journal of Robotics and Control (JRC) Volume 2, Issue 3, May 2020.

[8] Assgie, T.A, Dr. R. Lakshmi Tulasi, N. Komal Kumar, Breast cancer detection model with decision tree and adaptive boosting, IAES International Journal of Artificial Intelligence (IJ-AI) Vol. 10, No.1, March, 2021.

[9] Assegie, T.A, Nair, P.S, The Performance of Different MLModels On Diabetes Prediction, International Journal Of Scientific & Technology Research Volume 9, Issue 01, January 2020.

[10] Assegie, T.A, An optimized K-Nearest Neighbor based breast cancer detection, Journal of Robotics and Control (JRC) Volume 2, Issue 3, May 2020.

[11] Chong Zhang, Kay Chen Tan, Haizhou Li, Fellow, IEEE, Geok Soon Hong, A Cost-Sensitive Deep Belief Network for Imbalanced Classification, IEEE Transactions On Neural Networks And Learning Systems, 2018.

[12] Siyuan Zhou, Ya Zhang, Active Learning for Cost-Sensitive Classification Using Logistic Regression Model, IEEE, 2016.

[13] Donghui Shi, Jian Guan, Jozef Zurada, Cost-Sensitive Learning for Imbalanced Bad Debt Datasets in Healthcare Industry, Asia-Pacific Conference on Computer Aided System Engineering, IEEE, 2015.

[14] Todd Perry, Mohamed Bader-El-Den, Steven Cooper, Imbalanced Classification using Genetically Optimized Cost Sensitive Classifiers, IEEE, 2015.

[15] Assegie, T.A, R. Lakshmi Tulasi, N. Komal Kumar, Breast cancer prediction model using decision tree and adaptive boosting algorithm, IAES Journal of Artificial Intelligence, Vol. 10, No.1, March, 2021.

[16] Sushma S. J., S. C. Prasanna Kumar, A novel approach to jointly address localization and classification of breast cancer using bio-inspired approach, International Journal of Electrical and Computer Engineering (IJECE) Vol. 9, No. 2, April 2019.

[17] Achmad Ridok, Nashi Widodo, Wayan Firdaus Mahmudy, Muhaimin Rifa'I, A hybrid feature selection on AIRS method for identifying breast cancer diseases, International Journal of Electrical and Computer Engineering (IJECE) Vol. 11, No. 1, February 2021.

[18] Mohammed Y. Kamil, Computer-aided diagnosis system for breast cancer based on the Gabor filter technique, International Journal of Electrical and Computer Engineering (IJECE) Vol. 10, No. 5, October 2020.

[19] Susama Bagchi, Kim Gaik Tay, Audrey Huong, Sanjoy Kumar Debnath, Image processing and MLtechniques used in a computer-aided detection

system for mammogram screening-A review, International Journal of Electrical and Computer Engineering (IJECE) Vol. 10, No. 3, June 2020.
[20] Shubham Sharma, Archit Aggarwal, Tanupriya Choudhury, Breast Cancer Detection Using MLAlgorithms, IEEE, 2018.
[21] V.Shobana,K.Nandhini, Multiclass Data Imbalance Oversampling Techniques (Mudiot) and Random Selection of Features, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-12, October 2019.
[22] Bin Dai, Rung-Ching Chen, Shun-Zhi Zhu, Wei-Wei Zhang, Using RFA for Breast Cancer Diagnosis, IEEE, 2018.
[23] Abien Fred M. Agarap, On Breast Cancer Detection: An Application of MLAlgorithms on the Wisconsin Diagnostic Dataset, IEEE, 2019.
[24] J. Dheeba and S. Tamil Selvi, "Classification of malignant and benign microcalcification using SVM classifier," 2011 International Conference on Emerging Trends in Electrical and Computer Technology, Tamil Nadu, 2011, pp. 686-690.
[25] Hongpeng Yin, Xuguo Jiao, Yi Chai, Bin Fang, “Scene classification based on single-layer SAE and SVM”, Expert Systems with Applications, Volume 42, Issue 7, 1 May 2015, Pages 3368-3380
[26] A. Bosch, A. Zisserman and X. Munoz, "Image Classification using Random Forests and Ferns," 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, 2007, pp. 1-8.
[27] J. Liu, J. Chen, X. Liu and J. Tang, "An investigate of mass diagnosis in mammogram with random forest," The Fourth International Workshop on Advanced Computational Intelligence, Wuhan, 2011, pp. 638-641.
[28] ShuangBai, Growing random forest on deep convolutional neural networks for scene categorization, Expert Systems with Applications, Volume 71, 1 April 2017, Pages 279-287.
[29] NeerajDhungel, Gustavo Carneiro, Andrew P. Bradley, A deep learning approach for the analysis of masses in mammograms with minimal user intervention, Medical Image Analysis, Volume 37, April 2017, Pages 114-128.
[30] MellisaPratiwi, Alexander, JeklinHarefa, Sakka Nanda, Mammograms Classification Using Gray-level Co-occurrence Matrix and Radial Basis Function Neural Network, Procedia Comp