Advances In Multidisciplinary & Scientific Research
A Multidisciplinary & Interdisciplinary Journal

Vol. 4 No. 2, June 2018

Predictive Analysis In Health Data Using Back Propagation Neural Network (BPNN) and C4.5 Decision Tree

Akinrotimi, Akinyemi Omololu & Oladele, Rufus Olalere
Department of Computer Science
University of Ilorin
Ilorin Nigeria
E-mail: timiakin2011@yahoo.com

ABSTRACT

Abstract: Data mining has played a very important role in reducing simplifying the task of predicting diseases based on certain symptoms. With increasing technology, it is now possible to collect a large data made up of several symptoms that may relate to a particular disease. This has resulted in the increasing popularity of data mining techniques to identify patterns and relationships among large number of variables, thereby identifying the presence of a disease or certain diseases. This paper presents the potential synergies between data mining techniques using the C4.5 Decision Tree technique and biological inspired technique Back Propagation Neural Network (BPNN) in diagnosing breast cancer.

Keywords: - Back Propagation, Cancer, feature selection, Data Pre-Processing, Classification.

Aims Research Journal Reference Format:
Akinrotimi, Akinyemi Omololu & Oladele, Rufus Olalere (2018): Predictive Analysis In Health Data Using Back Propagation Neural Network (BPNN) And C4.5 Decision Tree. Advances in Multidisciplinary & Scientific Research Journal. Vol. 4. No.2, Pp 47-54

1. INTRODUCTION.

Prediction tasks are among the most interesting activities in which to implement intelligent systems. Specifically, prediction is an attempt to accurately forecast the outcome of a specific situation, using as input, information obtained from a concrete set of variables that potentially describe the situation. A problem often faced in clinical medicine, is how to reach a conclusion about the prognosis of cancer patients when presented with complex clinical and prognostic information, since specialists usually make decisions based on a simple dichotomization of variables into a favorable and unfavorable classification (McGuire W.L, 2011). As we enter the new millennium, treatment modalities exist for many solid tumor types and their use is well established. Nevertheless, offset against this is the toxicity of some treatments. As there is a real risk of mortality associated with treatment, it is vital to have the possibility of offering different therapies depending on the patients. In this sense, the likelihood that the patient will suffer a recurrence of her disease is very important, so that the risks and expected benefits of specific therapies can be compared.

2. LITERATURE REVIEW

Among prognostic modeling techniques that induce models from medical data, survival analysis methods are specific both in terms of modeling and the type of data required. Survival models attempt to determine the probability of the event occurring within a specific time, which requires classification models that classify either the occurrence or non-occurrence of the event and optionally model the outcome probabilities. Several tools successfully used in the construction of medical prognosis models have been proposed by the machine learning community (Zupan B, Demsar J., 2013) Neural networks are a form of artificial intelligence that have found application in a wide range of problems and have given, in many cases, superior results to standard statistical models. (Evans C, 2013) demonstrated the predictive reliability of an artificial neural networks model in medical diagnosis. In this case, we utilize the ability of neural networks to recognize complex and highly non-linear relationships, such as are likely to characterize medical circumstances. Artificial Neural Network (ANN) is one of the best artificial intelligence techniques for common data mining tasks, such classification and regression problems.

A lot of research showed that ANN delivered good accuracy in breast cancer diagnosis. However, this method has several limitations. First, ANN has some parameters to be tuned, in the beginning of training process such as number of hidden layer and hidden nodes, learning rates, and activation function. Second, it takes a long time for training process, due to complex architecture and parameters update process in each iteration that need expensive computational cost. Third, it can be trapped to local minima so that the optimal performance cannot be guaranteed. Numerous efforts had been attempted to get the solutions of neural networks limitations. G. B. Huang, 2009, proved that Single Hidden Layer

Neural Networks (SFLN) with tree steps machine learning process could solve those problems by conjugant gradient algorithm. The most important of these classifications are binary classification, either benign or malignant. If the cancer is in benign stage, less invasive and risk of treatments is used than for malignant stage. The reason being that, the chance of survival of the patient is high, and as such, it is not beneficial to increase the speed of recovery at the risk of introducing potentially life-threatening side effects caused by aggressive treatment. On the other hand, a patient with malignant cancer is not so concerned about the kind of treatment or side effect of the treatment (Rajesh Kumar, 2013.). This work proposes a multi approach based on back propagation neural networks and C4.5 decision for the classification medical breast cancer data, the result of which will be compared based on their predictive accuracy, sensitivity, specificity, recall , kappa statistic and other relevant statistical measures.

Chin-Lin Chi, 2010 presented an article on survival analysis of breast cancer on two breast cancer datasets. This article applies an Artificial Neural Networks (ANNs) to the survival analysis problem. Because ANNs can easily consider variable interactions and create a non-linear prediction model, they offer a more flexible prediction of survival time than traditional methods. This study compares ANN results on two different breast cancer datasets, both of which use nuclear morph metric features. The results show that ANNs can successfully predict recurrence probability and separate patients with good and bad prognosis. Tubakiyan, 2012 has discussed that statistical neural networks can be used to perform breast cancer diagnosis effectively. The scholar has compared statistical neural network with Multi-Layer Perceptron on WBCD database. Radial basis function (RBF), General Regression Neural Network (GRNN), Probabilistic Neural Network (PNN) were used for classification and their overall performance were 96.18% for Radial Basis Function (RBF), 97% PNN, 98.8% for GRNN and 95.74% for MLP. Hence it is proved that these statistical neural network structures can be applied to diagnose breast cancer.

3. PROPOSED SYSTEM

At macro level, breast cancer classification is usually done using the data gathered. Cell parameters like Thickness of Clump, Weak Adhesion, Cell Size Uniformity, uniformity of shape size, single epithetical cell size, barei nuclei, bland chromatin, mitosis are projected using selection techniques. The system will see the effect of classifying breast cancer dataset into their class group using the back propagation neural network and C4.5 decision tree model.

3.1 Technique Procedure

3.1.2 Data Collection

The database used in our study is the Wisconsin breast cancer database. It has been done in the University of Wisconsin by Dr. William H. Wolberg (UCI Machine Learning Repository). The same database has been used by researchers for the purpose of classification and testing algorithms in the world of data mining. 699 patients form the total available database.

The data related to the predictors will be provided to the network input layer. The data will be classified as input and target. The parameters for the classification include

1. Thickness of Clump
2. Cell Size Uniformity
3. Uniformity of Cell Shape
4. Weak Adhesion
5. Single Epithelial Cell Size
6. Barei Nuclei
7. Bland Chomatin
8. Normal Nucleoli
9. Mitosis

3.1.3 Data Preprocessing

After collection of Data from the domain expert, the data will undergo preprocessing for the purpose of data scaling and normalization.

3.1.4 Feature Ranking

The correlation feature selection will be used to rank the breast cancer dataset according to their predictive power.

3.1.5 Classification and prediction

At this phase the data with synthetic information will be passed to the C4.5 decision tree model for classification and also the back propagation neural network.

3.1.6 Performance Evaluation

The hold out technique will be used for splitting the data into training set and testing set, the percentage held out will be used to validate how efficient our developed models thereby creating a statistical metrics for system and algorithm measures.

3.2 SYSTEM FLOW CHART

The flow chart below represents the application flowchart.

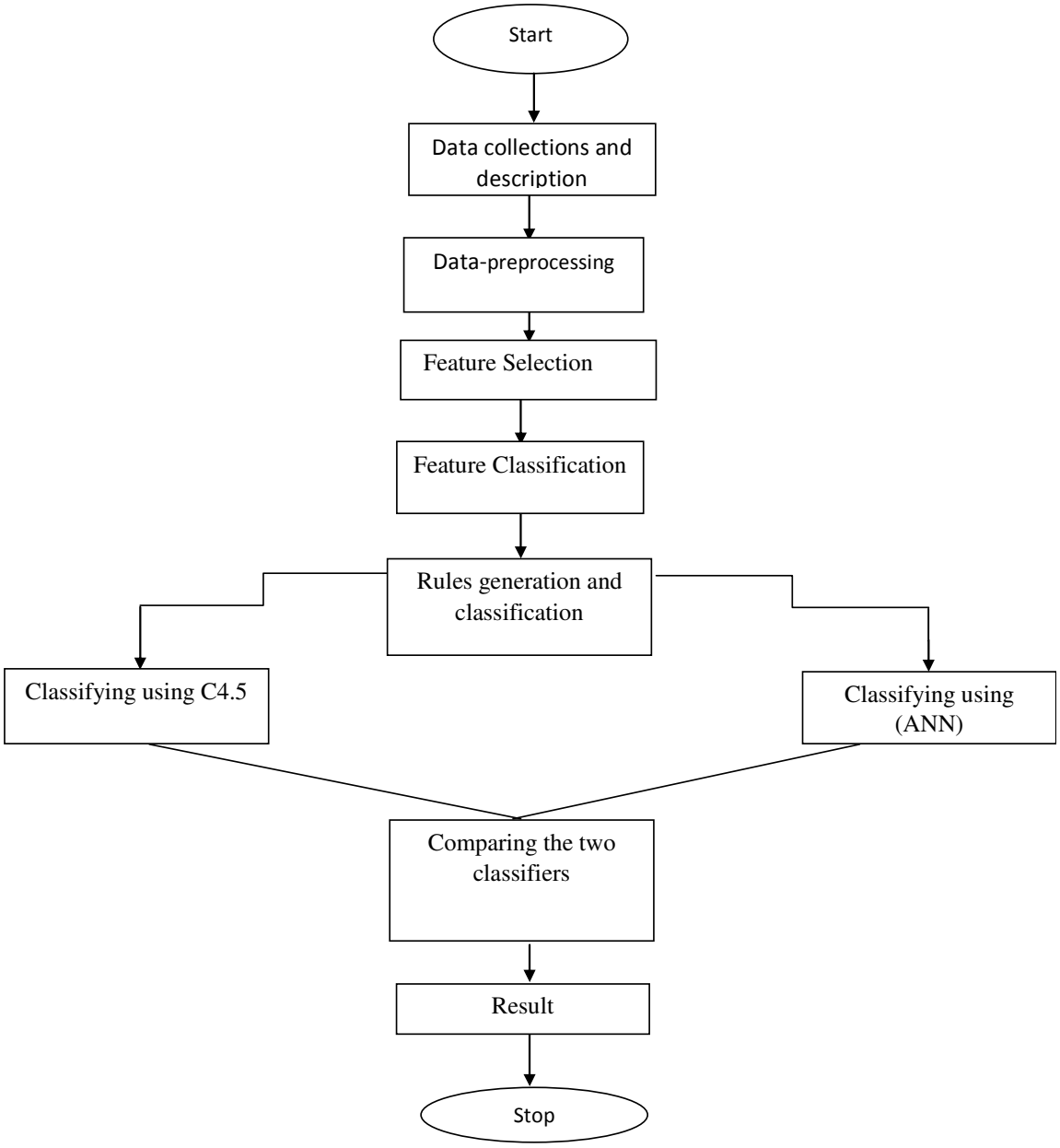

Figure 1.0: System Flow Chart

4. IMPLEMENTATION AND RESULT

4.1 Feature Selection With Correlation Feature Selection

After data filtering, the data was passed on to the select attributes with high predictive ability section. In respect to the class label, the correlation attribute evaluation was used. This helped to evaluate the worth of an attribute by measuring the correlation (Pearson's) between it and the class label. The table below shows the strength of the selected attributes with their correlation factor. The Thickness of Clump and Bland chomatin was discarded from the further classification analysis due to their negligible correlation factor.

Table 1: Ranked features

No.	Correlation Factor	Attributes
1	0.579	Bare nuclei
2	0.534	Cell Size Uniformity
3	0.528	Normal nucleoli
4	0.473	Single epithelial cell size
5	0.473	Uniformity of cell shape
6	0.469	Mitoses
7	0.46	Weak Adhesion
8	0.29	Bland chromatin
9	0.215	Thickness of Clump

4.2 Summary for ANN Classification

The summary table shows the summary statistic of the neural network classifier, with a classification accuracy and training time of 46.9 seconds. The obtained results shows that our MLP behaved generally well with an accuracy of 93.8095% and very low mean square error of 0.0542.

Table 2: Summary of ANN Classification

Correctly Classified Instances	Accuracy
197	93.8095 %
Incorrectly Classified Instances	6.1905 %
13	
Kappa statistic	0.8664
Mean absolute error	0.0542
Root mean squared error	0.2209
Relative absolute error	11.953 %
Root relative squared error	46.2294 %
Total Number of Instances	210
Training Time	46.9secs

4.2.1Confusion Matrix for ANN

The confusion matrix is an indication of the correctly and incorrectly classified class. The class A shows that 127 were actually classified as correctly A and 9 were classified as incorrectly b. The Class B shows that B were correctly classified as 70 and 4 incorrectly classified as A

.

Table 3.0 ANN Confusion Matrix

A	B	
127	9	a=2
4	70	b=4

4.3 C4.5 DECISION TREE CLASSIFICATION SUMMARY TABLE

The summary table shows the summary statistic of the neural network classifier, with a classification accuracy and training time of 46.9 seconds. The obtained results shows that our MLP behaved generally well with an accuracy of 93.8095% and very low mean square error of 0.0542.

Table 4: C4.5 Summary Table

Correctly Classified Instances	**Accuracy**
198	94.2857 %
Incorrectly Classified Instances	5.7143 %
12	
Kappa statistic	0.8756
Mean absolute error	0.0834
Root mean squared error	0.0834
Relative absolute error	18.3998 %
Root relative squared error	46.7512 %
Total Number of Instances	210
Training Time	0.16 secs

4.3.1 Confusion Matrix for C4.5

The confusion matrix is an indication of the correctly and incorrectly classified class. The class A shows that 129 were actually classified as correctly A and 7 were classified as incorrectly b. The Class B shows that B were correctly classified as 69 and 5 incorrectly classified as A.

Table 5: Confusion Matrix

A	B	
129	7	a=2
5	69	b=4

4.4 COMPARATIVE GRAPHICAL ILLUSTRATION

A. Classification accuracy.

The graph below shows in percentage the Classification rate of the C4.5 and Neural Networks Classifier. It shows the rate at which each classifier was able to classify correctly.

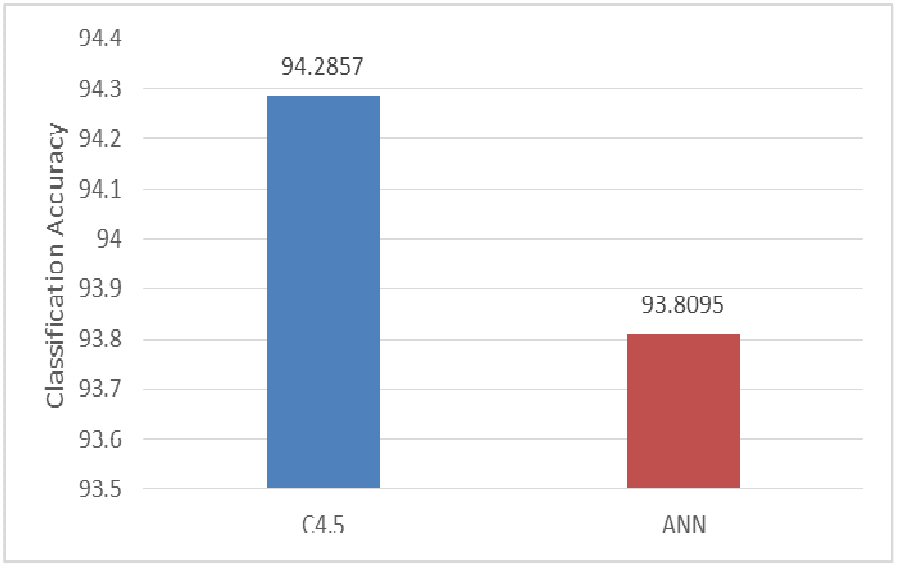

Figure 2.0: Classification Accuracy

B. Training Time.

The graph below shows the time taken to build each of the model, it shows that the C4.5 decision tree performed excellently faster than the neural networks.

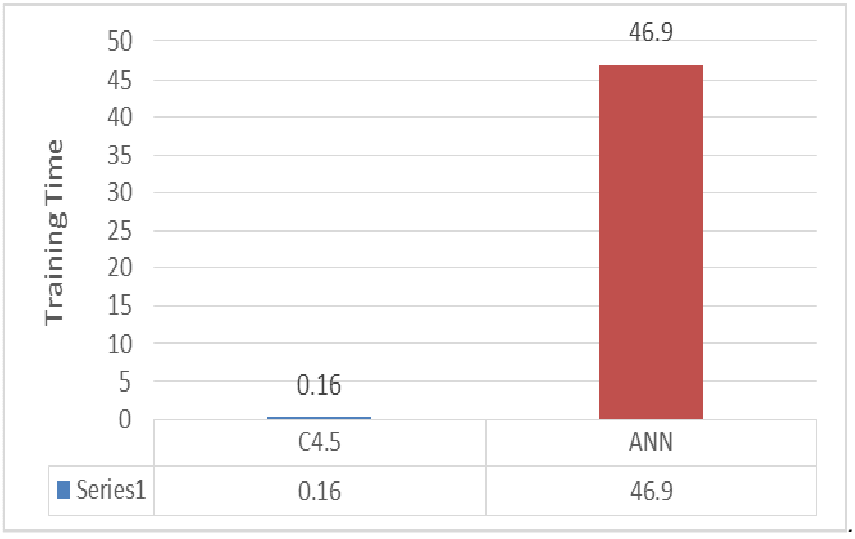

Figure 3.0: Training Time

C. Sensitivity

The Sensitivity (also called the true positive rate, the recall, or probability of detection in some fields) measures the proportion of positives that are correctly identified as such (e.g. the percentage of sick people who are correctly identified as having the condition). The ANN shows a higher rate probability of detection than the C4.5

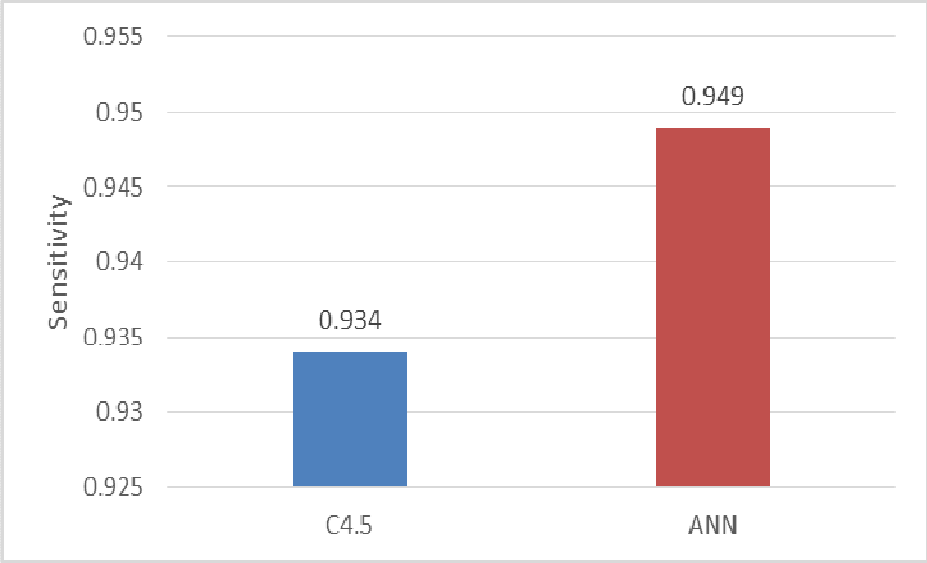

Figure 4.0: Sensitivity

5. CONCLUSION

This paper work presented a comparative classification case study for Breast Cancer Wisconsin Dataset Using the Artificial Neural Network (ANN) techniques and C4.5. From the ANN, a Multi-layer perception neural network along with back propagation algorithm was adopted and the J48 which is class for generating a pruned C4.5 decision tree. The obtained results shows that the C4.5 was more efficient in the knowledge retention rate as its classifier accuracy is quite higher than that of the ANN, the C4.5 also gave a very high computational time and a better probability detection rate than the ANN model. It is quite clear that the C4.5 model performed better than the ANN model for this study.

REFERENCES

1. McGuire WL, Tandom AT, Allred DC, Chamnes GC, Clark GM, (2011) How to use prognostic factors in axillary node-negative breast cancer patients. J Natl Cancer Inst 1990; 82:1006–15.
2. Zupan B, Demsar J, Kattan MW, Beck JR, Bratko I. Machine learning for survival analysis: a case study on recurrence of prostate cancer. Artif Intell Med 20(1):59–75.
3. G. B. Huang and H. A. Babri, (2009) “Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions.,” IEEE Trans. Neural Netw., vol. 9, no. 1, pp. 224– 9, Jan. 1998.
4. Rajesh Kumar Tripathy, (2013) An investigation of the breast cancer classification using various machine-learning techniques, department of biotechnology & medical engineering national institute of technology rourkela-769008, Orissa, India.
5. Chih-Lin Chi, (2012)“Global burden of cancers attributable to infections in 2008: a review and synthetic analysis,” The Lancet Oncology, vol. 13, no. 6, pp. 607–615,
6. Tüba KIYAN, (2012) “BREAST CANCER DIAGNOSI S USING STATISTICAL NEURAL NETWORKS,” Journal of electrical & electronics Engineering, Istanbul University vol. 13, no. 6, pp. 607–615,