

# MEDICAL DATA CLASSIFICATION USING SVM AND NEURAL NETWORK CLASSIFIER-A STUDY

S.VANITHA<sup>1</sup>, Dr.P.BALAMURUGAN<sup>2</sup>

<sup>1</sup> PhD Research Scholar, <sup>2</sup> Assistant Professor

Department of Computer Science, Government Arts College, Coimbatore, (India)

## ABSTRACT

*This paper aimed at study and analyzing the machine learning algorithms and finding out most appropriate algorithm for medical data classification. In this study, designed a classification system using Neural Network and Support Vector Machine for medical data classification with various numbers of attributes and instances. It includes two kinds of classification experiments namely diseased and non-diseased data distribution from the Cleveland heart disease data set and class distribution of benign and malignant from the Breast Cancer Wisconsin (Original) Data Set. The experimental outcomes positively demonstrate that the Neural Network classifier is effective in undertaking medical data classification tasks.*

**Keywords:** Classification, Data Mining, Data Set, Neural Network (NN), Support Vector Machine (SVM).

## I. INTRODUCTION

Medical data classification is a challenging task in the field of medical research. The medical record is very important for a patient as well as the doctor. Generally the medical record will help the doctor to classify the diseases, diagnose and give an appropriate treatment to the patient. In recent days, the volume of medical data is huge in size. So, it is very difficult to classify and understand the severity of the diseases manually. This paper is organised as follows. Section II describes the medical Breast Cancer and heart disease datasets. Section III covers the Results and discussion of the diseased and normal medical data classification. Section IV deals with the conclusion.

## II. MEDICAL DATASETS

This section explains about the medical datasets used for different experimental setup. The medical datasets include Breast Cancer Data and Cleveland Heart Data.

### A. Breast Cancer Wisconsin (Original) Data Set

Breast Cancer Wisconsin (Original) Data Set[15] was created by Dr. William H. Wolberg (physician), University of Wisconsin Hospitals, Madison, Wisconsin, USA and donated by Olvi Mangasarian. It consists 10 attributes, 699 instances collected in the different occasions that are distributed into 8 groups. The prediction field refers to the presence of either benign or malignant. The proposed work experiments with the 250 records of all attributes shown in the table I that have randomly selected from the database. Class distribution for the proposed work is benign (value 0) and malignant (value 1).

## **B. Cleveland Heart Disease Data Set**

Cleveland heart disease data set[4] is a Multivariate data set. It contains 76 attributes and 303 instances that are varying Categorical, Integer and Real. However, the proposed experiments refer to using a subset of 13 of them as shown in the table II. The prediction field refers to the presence of heart disease in the patient. The prediction field concentrated on simply attempting to distinguish presence of diseased (value 1) data from non-diseased (value 0).

## **III. CLASSIFICATION RESULT**

### **Experiment-1: Understanding the classification performance of Benign and Malignant Data of Breast Cancer Wisconsin Data Set using NN Classifier:**

The experimental setup consists of 250 records consisting of both Benign and Malignant Data. Training and Testing sets are formed with random selection of 188 records and 62 records respectively. The 10 attributes/features as shown Table I are given as input to the classification system. These features are used as training parameters for the neural network using the scaled conjugate gradient algorithm with 20 hidden neurons for 1000 epochs[3], [5], [8]. Classification results of Benign and malignant medical data is shown in Table IV & VII. The highest classification rate 95.2% is obtained for 10 features trained with 20 hidden neurons. The confusion matrix obtained for training neural network using 10 features with 20 hidden neurons are shown in Tab. III. Figure 1 shows the Performance function plot of Neural Network classifier for the classification of Breast Cancer Data Set respectively.

### **Experiment-2: Understanding the classification performance of Benign and Malignant Data of Breast Cancer Wisconsin Data Set using SVM Classifier:**

The Breast Cancer Data set used for SVM based classification system is similar to NN classification system defined in the previous section. SVM uses an optimum linear separating hyperplane to separate two set of data in a feature space[8], [10]. The success rate of 91.9% is obtained for 10 attributes/features is shown in Table IV. Table III shows the confusion matrix for classifying the Breast cancer data set with fixed number of neurons.

### **Experiment-3: Understanding the classification performance of Heart Disease and Non-Heart Diseased Data using NN Classifier:**

The experimental setup consists of 303 records consisting of both Heart Disease and Non-Heart Diseased Data. Training and Testing sets are formed with random selection of 228 records and 75 records respectively. Each record has 13 attributes/features as shown Table II, are given as input to the classification system. These features are used as training parameters for the neural network using the scaled conjugate gradient algorithm with 20 hidden neurons for 1000 epochs[3], [5], [8]. Classification results of Heart Diseased and Non- Heart Diseased medical data is shown in Tab. VI & VII. The highest classification rate 85.5% is obtained for 13 features trained with 20 hidden neurons. The confusion matrix obtained for training neural network using 13 feature with 20 hidden neurons are shown in Tab. V. Figure 2 shows the Performance function plot of Neural Network classifier for the classification of Cleveland Heart Disease Data Set respectively.

#### Experiment-4: Understanding the classification performance of Heart Diseased and Non-Heart Diseased Data using SVM Classifier:

The experimental setup used for Classification system using SVM is similar to NN classification system formed in the previous section. Here support vector machine classifier trained using the training data set taken from two groups(Heart Diseased and Non-Heart Diseased) given by prediction column. SVM structure contains information about the trained classifier, including the support vectors, that is used by SVM Classifier for classification. Prediction is a column vector of values of the same length as Training set that defines two groups. Each element of Prediction column specifies the group the corresponding row of training belongs to. Prediction column can be a numeric vector, a string array, or a cell array of strings. SVM Training treats NaNs(Not a Numeric) or empty strings in Prediction as missing values and ignores the corresponding rows of training. SVM uses an optimum linear separating hyperplane to separate two set of data in a feature space. The success rate of 78.7% is obtained for 13 attributes/features is shown in Table VI. Table V shows the confusion matrix for classifying the Heart Diseased and Non-Heart Diseased Data set of Cleveland Heart Disease Data with fixed number of neurons.

#### IV. FIGURES AND TABLES

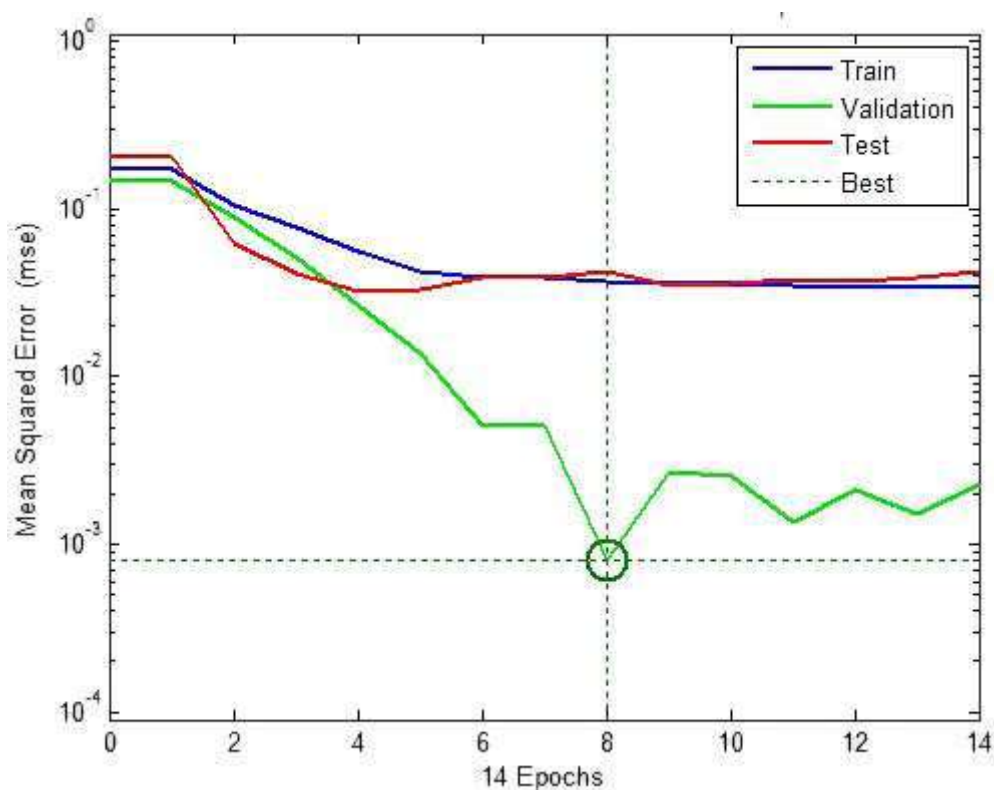


Fig. 1. Neural Network Classifier Performance Function Plot for classifying the benign and malignant data of Breast Cancer Data Set.

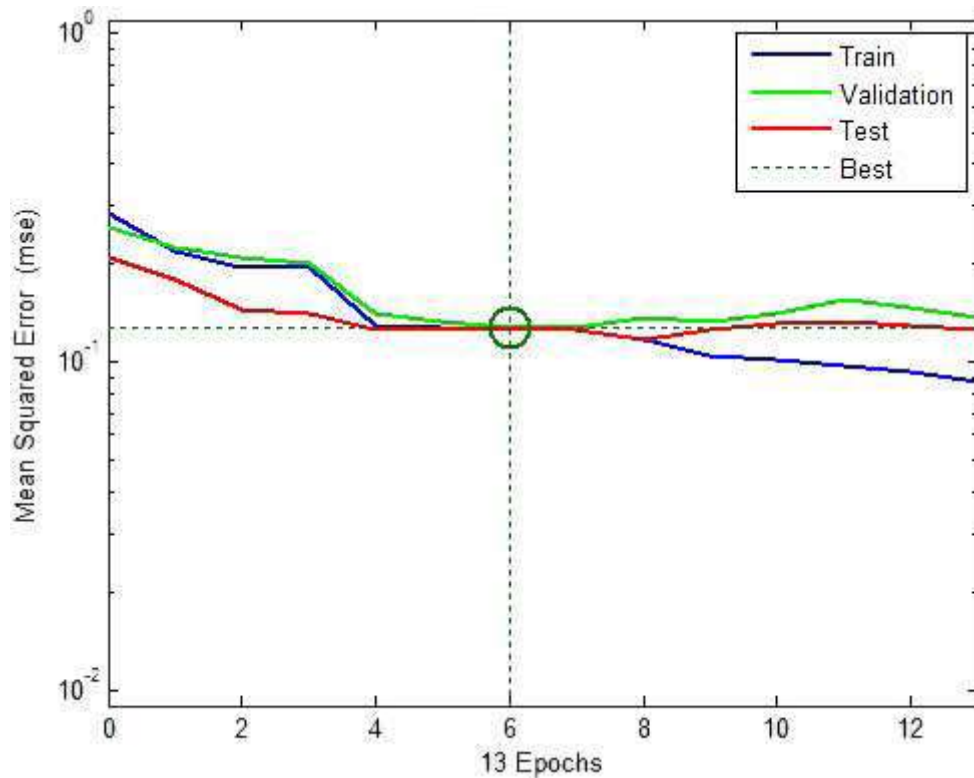


Fig. 2. Neural Network Classifier Performance Function Plot for classifying the non-diseased and diseased data of Cleveland Heart Disease Data Set

**TABLE I**  
**ATTRIBUTES OF BREAST CANCER WISCONSIN (ORIGINAL) DATA SET**

| S.No. | Attributes                  | Values    |
|-------|-----------------------------|-----------|
| 1.    | Sample code number          | Id Number |
| 2.    | Clump Thickness             | 1-10      |
| 3.    | Uniformity of Cell Size     | 1-10      |
| 4.    | Uniformity of Cell Shape    | 1-10      |
| 5.    | Marginal Adhesion           | 1-10      |
| 6.    | Single Epithelial Cell Size | 1-10      |
| 7.    | Bare Nuclei                 | 1-10      |
| 8.    | Bland Chromatin             | 1-10      |
| 9.    | Normal Nuclei               | 1-10      |
| 10.   | Mitoses                     | 1-10      |

**TABLE II**  
**ATTRIBUTES OF CLEVELAND HEART DISEASE DATA SET**

| S.No. | Attributes | Description  |
|-------|------------|--|
| 1.    | Age        | Age in years   |
| 2.    | Sex        | Sex (1=male;0=female)  |
| 3.    | cp         | Chest pain type<br>(1=typical angina; 2=atypical angina; 3=non-anginal pain; 4= asymptomatic)  |
| 4.    | trestbps   | Resting blood pressure (in mm Hg on admission to the hospital)   |
| 5.    | chol       | Serum cholestoral in mg/dl   |
| 6.    | fbs        | Fasting blood sugar (> 120 mg/dl) (1 = true; 0 = false)  |
| 7.    | restecg    | Resting electrocardiographic results<br>(0 = normal; 1 = having ST-T wave abnormality<br>(T wave inversions and/or ST elevation or depression of > 0.05 mV)) |
| 8.    | thalach    | Maximum heart rate achieved  |
| 9.    | exang      | Exercise induced angina (1 = yes; 0 = no)  |
| 10.   | oldpeak    | ST depression induced by exercise relative to rest   |
| 11.   | slope      | Slope of the peak exercise ST segment<br>(1 = upsloping; 2 = flat; 3 = downsloping)  |
| 12.   | ca         | Number of major vessels (0-3) colored by flourosopy  |
| 13.   | thal       | (3 = normal; 6 = fixed defect; 7 = reversable defect)  |

**TABLE III**  
**CONFUSION MATRIX FOR CLASSIFYING THE BENIGN AND MALIGNANT DATA OF BREAST CANCER DATA SET WITH FIXED NUMBER OF NEURONS (20 NEURONS)**

| Classifier                   | Actual Class | Predicted Class |          |
|------------------------------|--------------|-----------------|----------|
|                              |              | Non-Diseased    | Diseased |
| Support Vector Machine (SVM) | Non-Diseased | 36              | 2        |
|                              | Diseased     | 3               | 21       |
| Neural Network (NN)          | Non-Diseased | 30              | 2        |
|                              | Diseased     | 1               | 30       |

**TABLE IV**  
**CLASSIFICATION RATE (%) OF BENIGN AND MALIGNANT FOR BREAST CANCER DATA (ACC-ACCURACY; TPR-TRUE POSITIVE RATE; TNR-TRUE NEGATIVE RATE)**

| Data Sets                    | ACC     | Recall (TPR) | TNR    | Precision | Negative Predictive |
|------------------------------|---------|--------------|--------|-----------|---------------------|
| Support Vector Machine (SVM) | 91.9355 | 0.9231       | 0.9130 | 0.9474    | 0.8750              |
| Neural Network (NN)          | 95.2381 | 0.9677       | 0.9375 | 0.9375    | 0.9677              |

**TABLE V**

**CONFUSION MATRIX FOR CLASSIFYING THE NON-DISEASED AND DISEASED DATA OF CLEVELAND HEART DISEASE DATA WITH FIXED NUMBER OF NEURONS (20 NEURONS)**

| Classifier                   | Actual Class | Predicted Class |          |
|------------------------------|--------------|-----------------|----------|
|                              |              | Non-Diseased    | Diseased |
| Support Vector Machine (SVM) | Non-Diseased | 34              | 7        |
|                              | Diseased     | 9               | 25       |
| Neural Network (NN)          | Non-Diseased | 28              | 3        |
|                              | Diseased     | 8               | 37       |

**TABLE VI**

**CLASSIFICATION RATE(%) OF NON-DISEASED AND DISEASED FOR CLEVELAND HEART DISEASE DATA (ACC-ACCURACY; TPR-TRUE POSITIVE RATE; TNR-TRUE NEGATIV RATE)**

| Data Sets                    | ACC     | Recall (TPR) | TNR    | Precision | Negative Predictive |
|------------------------------|---------|--------------|--------|-----------|---------------------|
| Support Vector Machine (SVM) | 78.6667 | 0.7907       | 0.7813 | 0.8293    | 0.7353              |
| Neural Network (NN)          | 85.5263 | 0.7778       | 0.9250 | 0.9032    | 0.8222              |

**TABLE VII**

**CLASSIFICATION RESULTS: NEURAL NETWORK CLASSIFIER WITH FIXED NUMBER OF NEURONS (20)**

| Data sets               | Phase      | Samples | Accuracy | MSE    | Error % |
|-------------------------|------------|---------|----------|--------|---------|
| Cleveland Heart Disease | Training   | 182     | 89%      | 0.126  | 15.93   |
|                         | Validation | 45      | 82%      | 0.128  | 17.78   |
|                         | Testing    | 76      | 85.5%    | 0.125  | 14.47   |
| Breast Cancer           | Training   | 149     | 96%      | 0.0364 | 4.027   |
|                         | Validation | 38      | 100%     | 0.0799 | 0       |
|                         | Testing    | 63      | 95.2%    | 0.0417 | 4.76    |

## V. CONCLUSION

Now-a-days, Health care system generates vast amount of information and it is accumulated in medical databases. So, the manual classification of this information is becoming more and more difficult. Therefore, there is an increasing interest in developing automated evaluation methods to follow up the diseases. In this paper, discussed the two class (diseased and non-diseased) classification problem. The proposed work used Neural Network and Support Vector Machine for classifying benchmark medical data sets namely Cleveland Heart Disease Dataset and Breast Cancer Wisconsin (Original) Dataset. Based on the experimental results, Neural Network classifier gives prominent results in the classification of diseased and non diseased data from each medical dataset. The major challenge of any classification problem is finding the effective/meaningful features from the number of attributes present in the data set. This work opens up the new interesting

applications in medical data classification. Future contributions will concentrate on developing a novel and hybrid intelligent system for medical data classification.

#### IV REFERENCES

- [1] Aqueel Ahmed, Shaikh Abdul Hannan, “Data Mining Techniques to find out Heart Diseases : An Overview,” *International Journal of innovative Technology and Exploring Engineering (IJITEE)*, Vol. 1, No. 4, Sep-2012.
- [2] Arun George Eapen, “Application of Data mining in Medical Applications,” 2004.
- [3] Ben Krose, Patrick van der Smagt, “Introduction to Neural Networks,” *The University of Amsterdam*, 8th edition, 1996.
- [4] Cleveland Clinic Foundation Heart disease data set available at: <http://archive.ics.uci.edu/ml/datasets/Heart+Disease>.
- [5] Faissal MILI, Manel HAMDI, “A hybrid Evolutionary Functional Link Artificial Neural Network for Data mining and Classification,” (*IJACSA*) *International Journal of Advanced Computer Science and Applications*, Vol. 3, No. 8, 2012.
- [6] P. Gayathri, N. Jaisankar, “Comprehensive Study of Heart Disease Diagnosis Using Data Mining and Soft Computing Techniques,” *International Journal of Engineering and Technology (IJET)*, 2013.
- [7] R.C. Gonzalez, R.E. Woods, “Digital Image Processing,” *Second Ed. Prentice-Hall*, New Jersey, 2002.
- [8] Jianxin Chen, Yanwei Xing, Guangcheng Xi, Jing Chen, Jianqiang Yi, Dongbin Zhao, Jie Wang, “A Comparison of Four Data Mining Models: Bayes, Neural Network, SVM and Decision Trees in Identifying Syndromes in Coronary Heart Disease,” *Advances in Neural Networks - ISNN 2007, Lecture Notes in Computer Science*, Vol.4491, pp. 1274-1279, 2007.
- [9] Mai Shouman, Tim Turner, Rob Stocker, “Applying k-nearest neighbor in diagnosing heart disease patients,” *International Journal of Information and Education Technology*, Vol. 2, No. 3, pp. 220-223, June 2012.
- [10] Sandhya Joshi, Hanumanthachar Joshi, “SVM Based Clinical Decision Support System for Accurate Diagnosis of Chronic Obstructive Pulmonary Disease,” *International Journal of Engineering Research & Technology (IJERT)*, Vol. 2, No. 4, April 2013.
- [11] Saharon Rosset et.al., “Medical Data mining: Insights from winning two competitions,” *Data Mining and Knowledge Discovery*, 2009.
- [12] Shamsheer Bahadur Patel et.al, “Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques,” *IOSR Journal of Agriculture and Veterinary Science (IOSRJA VS)*, Vol. 4, No. 2, pp. 61-64, 2013.
- [13] B.Sujatha, C.Akila, “ Prediction of Breast Cancer using Data Mining,” *International Journal of Computer Science and Management Research*, Vol.1, No.-3, Oct 2012.
- [14] John A. K. Suykens et.al., “Machine Learning”, *Kluwer Academic Publishers*, Vol. 54, pp. 5-32, Netherland, 2004.
- [15] UCI Machine Learning Repository: <http://archive.ics.uci.edu/ml/datasets.html>