Analysis of Hybrid Neural Networks for Improved Performance

Zainab Khalid Awan,
Aamir Khan
COMSATS Institute of
Information Technology, Wah
Cantt. Pakistan

Anam Iftikhar,
COMSATS Institute of
Information Technology, Wah
Cantt. Pakistan

Sadia Zahid,
Anam Malik
COMSATS Institute of
Information Technology, Wah
Cantt. Pakistan

ABSTRACT

In this paper we take a close look at the Hybrid Neural Network Model. Hybrid model is attained by combining two Artificial Neural Networks (ANNs). In which the first model is used to perform the feature extraction task and the second one performs prediction task. This paper explores the classifying ability of the proposed hybrid model. We analyze the performance of the model, which is a compound characteristic, of which the prediction accuracy is the most important component. If the prediction accuracy of the model can be increased it will result into enhanced performance of the model. The model that has been built is under the umbrella of pattern recognition and incorporates some of the data mining techniques. Kernel Principal Component Analysis (KPCA) has been implemented in the pre-processing stage for easier subsequent analysis. By the end of the paper, the key factors that enhance the accuracy of the model have been identified and their role explained. It also has been shown that single ANN model's performance deteriorates on an unseen problem much more as compared to the hybrid model. The aim is to provide a model having better performance and accuracy. The paper focuses on the real world applications of the model.

General Terms

Hybrid Neural Network, Pattern Recognition, Classifier, Cross Validation, KPCA, Confusion Matrix, Prediction Accuracy, Outliers, Feature Extraction.

Keywords

Unseen Test Set (UTS), Seen Test Set (STS).

1. INTRODUCTION

The neural network paradigm stemmed from Artificial Intelligence research. The purpose of the research is to develop the machines with cognitive ability. Neural network is aimed to learn just like a human brain, it may be in the form of a computer program or a hardwired machine.

Neural network is a massively parallel adaptive machine that has an ability of storing experiential knowledge and making it available for the future use [1].

1.1 Artificial Neural Network

An artificial neural network is made up of simple elements operating in parallel, called nodes. A neural network is trained to perform a particular function. Training is done by adjusting the values of the connections between elements.
Neural networks are trained in a way that a particular input leads to a specific output. Network adjustment is made on the

basis of difference between the target and output, until a point comes where the output exactly matches the target [2].

Figure 1 represents the block diagram of a neural network. It represents the pathway of an input. Input is fed to the nodes, and then nodes calculate the output which is then compared with the pre-determined target. If the output fails to match the target, it is fed back to the node and the weights are readjusted. The process goes on until the output exactly matches the input. This is exactly the same as happens in a biological neuron.
Commonly, a large dataset formed from many such inputs and targets pairs is used to train the network. This process of training is called supervised learning.

Artificial Neural Networks due to their ability of storing knowledge and making decisions in the future are extensively used in pattern recognition tasks.

1.2 Pattern Recognition

Pattern recognition is all about assigning labels to the objects. Objects are described by a dataset generally known as features or attributes. As pattern recognition is confronted with the real-life problems, in spite of decades of productive research, elegant modern theories still coexist with ad hoc ideas, intuition and guessing [3].

Below are mentioned the two types of pattern recognition, i.e. the supervised pattern recognition (also called supervised learning) and unsupervised pattern recognition (also called unsupervised learning).

1.3 Supervised Pattern Recognition

In the category of supervised pattern recognition each object in the dataset comes with a pre-assigned class label. A classifier is trained to do the labeling "sensibly." Machine is supplied with the learning skills and the labeled data is presented to it. The prediction accuracy will be the decisive factor of adequacy of the classifier as the knowledge learned in the process might be obscure. Classification is an example of supervised pattern recognition.

1.4 Unsupervised Pattern Recognition

The dataset in this category does not have previously assigned class labels. The primary goal of the unsupervised paradigm is to discover the structure of the data set if there is any. This means that the end user wants to know whether there are clusters in the data, and what characteristics make the objects similar within the cluster and different across the clusters. Clustering is an example of Unsupervised Pattern Recognition.
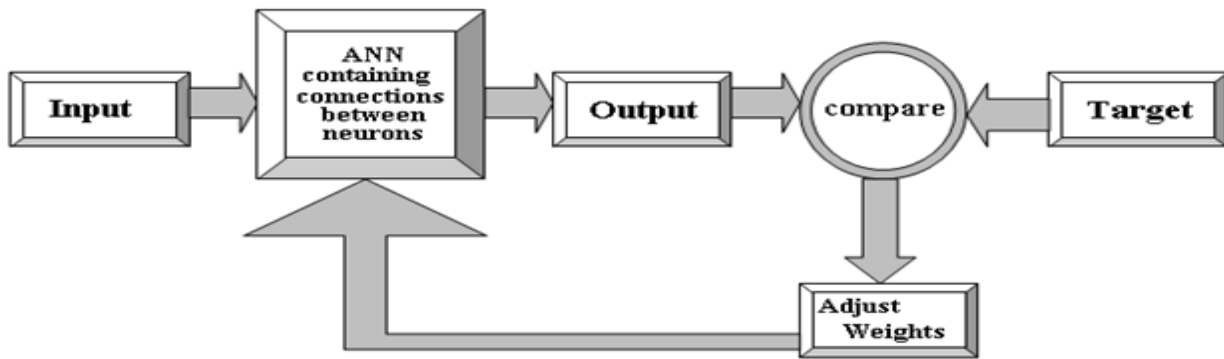
Figure 1: Neural Network block diagram

1.5 The Problem Statement

The most frequently encountered decision making task for the humans is Classification [4]. In finance, science and business, analysts are often faced with the task of classifying items based on historical or measured data. For example, Stock market analysts may wish to categorize a group of stocks as buy, sell, or hold.

Neural networks are typically used for classification. Classification problem involves an input, that the classifier has to identify, to which output class the input most likely belongs.

For classification problems, the network training is a supervised one, in which the desired or target response of the network for each input pattern is always known a priori [5].

For various orthodox classifications methods neural networks are a promising alternative. The usability of neural networks lies in the following facts [6],
1) Firstly, they are adaptive systems, in which they can conform themselves to the data without any explicit specification of functional form for the underlying model.
2) Secondly, they are non-parametric, in which neural networks can approximate any function with arbitrary accuracy.
3) Furthermore, neural networks are nonlinear models, which make them flexible in modeling real world complex relationships.

1.6 The Objective

The objective of this paper is to study the theory of hybrid neural network model. And then develop a model whose prediction accuracy is significantly higher than the baseline model. Prediction accuracy is calculated with the help of a confusion matrix.

Both the models have been built and compared by using MATLAB Neural Network Toolbox. The reason that we opted to work on MATLAB for carrying out the research methodology lies in the fact that its toolboxes allow one to learn and apply the specialized knowledge. Toolboxes are immersive collections of MATLAB (M-files)

1.7   The Research Question

The following research questions allow the research to meet the objectives proposed:

- Up to what extent the neural networks can predict accurately?
- Can neural networks be used as a practical classification tool by individuals?

1.8 Scope and Limitations

The potential architectures, training algorithms and data reduction techniques are virtually limitless. For this reason, the methodology is somehow limited; one training algorithm i.e. the back propagation is used. The most common architecture is the multi-layer feed forward network, which is used mainly for pattern recognition and classification. This allowed the research to validate the previous research as well as provided an insight to the future work.

The datasets used are the benchmarks, obtained from UCI Machine Learning Repository. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. As an indication of the impact of the archive, it has been cited over 1000 times, making it one of the top 100 most cited "papers" in all of computer science. Those datasets are typically available for the research purposes. The limitation they impose is that, they are extracted in carefully controlled environment and the results obtained may differ from those of obtained in real world applications.

2. HYBRID MODEL/ ARCHITECTURE

A hybrid model incorporates the data mining technique. The first ANN performs feature extraction task and the second ANN performs prediction task. The first step is meant to remove the outliers and to select features. The second one then makes predictions far more accurate than single ANN.

2.1 Baseline Model

Baseline model is the one that provides basis for making a hybrid model. It is the first step in the hybrid model. It is aimed at performing the feature extraction task and removing outliers. As there is no 100% accuracy, misclassification is inevitable. However, it can be minimized. Therefore, first model is used to do pre-processing. In pre-processing the dataset is made clean free from noise. As the model is extremely sensitive to outliers so they must be discarded. The output of this model is the input for the next model [21].

2.1.1   Feature Extraction

The feature extraction task is performed by using Kernel Principal Component Analysis (KPCA). KPCA is an extension of the conventional Principal Component Analysis

(PCA). PCA is meant only for linearly separable problems, which can be separated by a single line on hyperplane. But for linearly non-separable problems, which cannot be separated by a single line on hyperplane kernel trick is used. Kernel tricks are a way of mapping a dataset of observations into a high dimensional space such that the observations become linearly separable. Figure 8 represents the process for converting linearly non-separable to linearly separable problem.
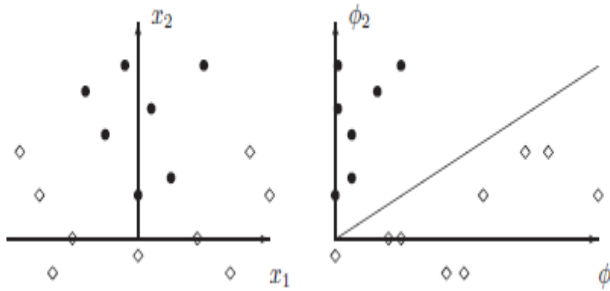

**Figure 2: Kernel trick, linearly non-separable to a linearly separable**

In figure 2, mapping from linearly non-separable to a linearly separable data is shown. On the left side in the figure 8 data is shown that cannot be separated by drawing just a single line. There is no clear decision boundary on which the decision can be made. Therefore, kernel trick is applied to it in order to attain a clear decision boundary [22] [23].

## 2.2 Hybrid Neural Network Model

Hybrid means "*something having two kinds of components that produce the same or similar results, such as a vehicle powered by both an electric motor and internal combustion engine as sources of power for the drive train*".

The output dataset from the first model is used to train the next ANN as it is the clean dataset so training with dataset will result into a better prediction performance. The whole process of forming the hybrid model is represented in the figure 9 (a). In figure 9 (b) steps involved in cross validation are shown.

The dataset taken is split into two proportions of 20% and 80%. 20% of the dataset is retained for testing the model and 80% is used for training the model. After splitting the dataset 5-fold cross-validation (CV) is performed on it in order to prevent overfitting. From the results of cross-validation baseline model is selected followed by the operation of feature extraction on the original dataset. Using the refined dataset the baseline model is trained and then tested on seen as well as unseen data.
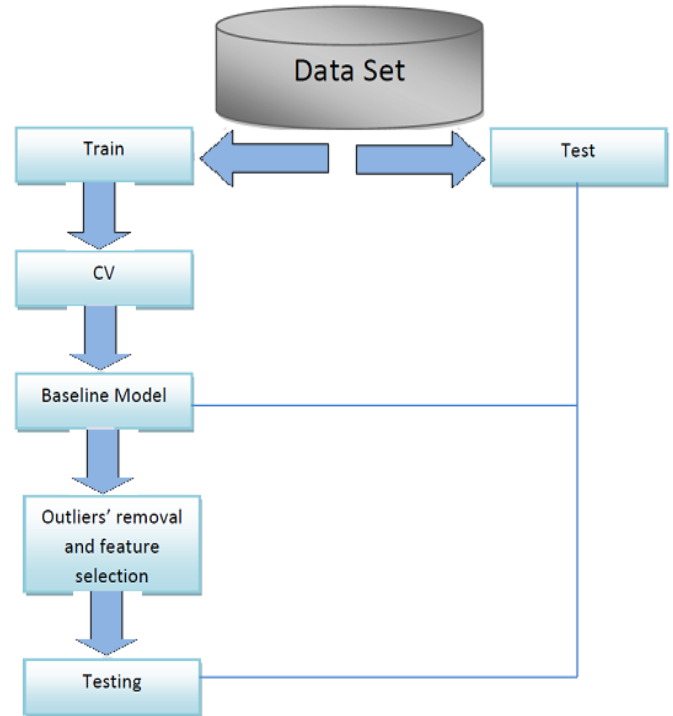
(a)
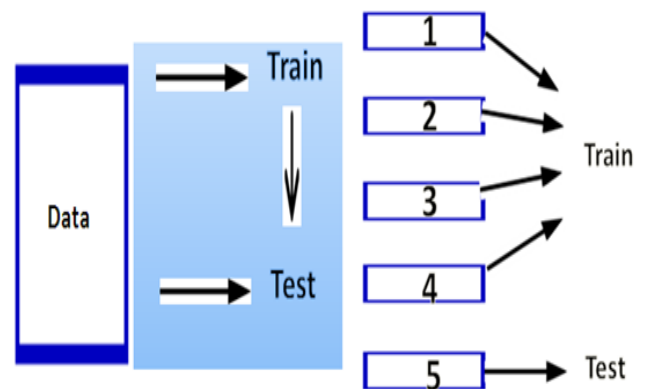
(b)

Figure 3: Process of combining two ANNs

3. RESEARCH METHODOLOGY

Apart from discussing the procedural framework of the model it takes a closer look at the datasets on which the experiment is performed. This first describes the datasets used, then discusses the steps in development of the model, followed by the evaluation and concludes with the end results.

3.1 Datasets

For the purpose of this paper, two different datasets have been considered. The datasets have been taken from the UCI Machine Learning Repository. Specifications of all the datasets are provided under the following headings.

3.2 Two-way Split

The datasets are divided into two parts 80% and 20%. Training set is made up of 80% of the dataset that are presented to the network during training, and the network is adjusted according to its error. The remaining 20% has no effect on training and so provide an independent measure of network performance during and after training [24].

3.3 Model Development- Step 1

The first prepared model is called the baseline model. The baseline model is selected from different models acquired by different parameter settings. The model with the highest accuracy is selected as the baseline model. Apart from this, five-fold cross validation is also performed to prevent the phenomenon called "*overfitting*". The accuracy calculated is based on the confusion matrix.

In the following sections the baseline model and the hybrid model (ANN+ANN) are discussed along with the details of intermediate steps involved.

3.4 The baseline model

At first, we use the original datasets to train a MLP neural network as the baseline ANN model for comparisons [25] [26]. In addition, four different learning epochs (50, 100, 200, and 300) and five different hidden layer nodes (8, 12, 16, 24, and 32) are used in order to obtain the best ANN baseline model. Table 1 (chapter 5) shows the settings of the learning epochs and numbers of hidden layer nodes. As a result, there are twenty different ANN models developed for comparisons [27].

Selected baseline model will be further used to develop the hybrid model. The accuracies in the tables (1 and 2) refer to chapter, are very high. This is due to the fact that testing is performed using the same dataset. The dataset that was originally used to train the model is then used to test the model. This is a terrific idea to test the model on the same dataset which was previously used for training. If this practice is carried on continuously then it will eventually lead a phenomenon called "*overfitting*" [28].

3.5 Overfitting

The concept of overfitting is crucial to machine learning. The model is presented with exemplary dataset in order to learn the patterns and generalize to the unseen situations. Overfitting occurs when a neural network is trained for so long that it begins to memorize the dataset. It then performs excellently on the same dataset but when it is tested using an unseen dataset it simply fails drastically. The model that has been overfit will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data.

The efficacy of the model is determined by its ability to perform well on unseen data. When a model begins to memorize training data rather than learning to generalize from the trend overfitting occurs [29].

Overfitting has to be avoided somehow. The phenomenon that handles the problem of overfitting most effectively is known as "*cross validation*". Therefore the next step taken in the paper is the cross validation [30].

3.6 Cross Validation

Cross validation concept is closely related to overfitting. The model is subjected to a training set and it is assumed that the

model will reach a state whereby it will be able to generalize to the unseen problems well. In the cases where training dataset is not the representative of every situation that may be encountered or the training was performed for too long the learner may adjust to very specific random features of the training data that have no causal relation to the target function.
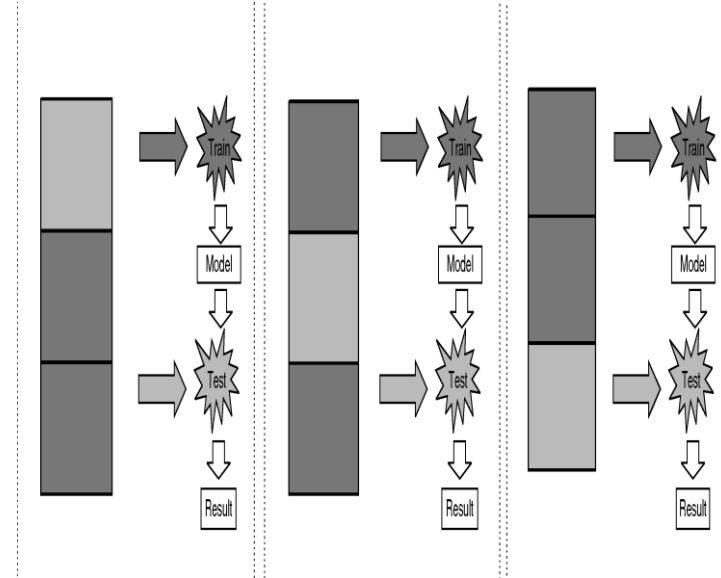

Figure 4: 3-fold cross validation

Cross validation is a statistical practice of randomly dividing the dataset into k folds. Folds may also be referred as the subsets. However, the subsets must be disjoint [31]. After that, one of the k folds is picked and retained for the subsequent testing and rest of the four folds are used for training. The process goes on until all the folds are traversed exactly once. By the end of the cross validation each fold must have been used for testing purpose exactly once while remaining k-1 must have been used for training. This process is described in the figure 4.

The figure 4 represents 3-fold cross validation. Upon completion averaging will be used to obtain an aggregate measure of performance metric from these samples.

For the purpose of this paper 5-fold cross validation is considered. As a result there are twenty different MLP baseline models for comparisons.

Now there are twenty different accuracies for each subset. Each subset/fold of highest accuracy is chosen, five folds with highest accuracy provide basis for developing hybrid model. Then these five folds are used for developing the hybrid model.

4. MODEL DEVELOPMENT- STEP 2

The results obtained from step 1 form basis for the development of hybrid model. The best parameter setting models are selected form twenty different models.

4.1 Hybrid Model

The hybrid model is made by cascading two ANNs, in a way that the first model performs the task of data reduction and the

second one is used to perform prediction hence, determines accuracy.

The term "data reduction" has a very wide meaning that it includes a number of different ways to perform reduction task. It is primarily dependant on the designer's preference that which technique suits best to a specific model [32]. There is no universal measure that what technique will fit best to a particular model. Some of the techniques are: cleaning, normalization, transformation, feature extraction and selection etc. Therefore, it is concluded that the first model performs data reduction referred to as "pre-processing".

4.2 Pre-Processing

The first and foremost thing required for a good predictive model is the quality of data given to it. Wherever the word quality comes a phrase comes into mind automatically and that is *"garbage in, garbage out"*. The phrase says it all. If the model is given a data that is not well organized, not representing every possible scenario, having the missing values, having noise/outliers and a number of other such possible reasons results into poor predictive performance of the model [33].

This happens because the data-collection procedures are loosely controlled and result into out-of-range values (Salary: −200) incorrect data combinations (Gender: Male, Pregnant: Yes). There may be some missing values as well for example, when the questionnaire is returned incomplete or in archaeological survey some parts of the master piece were missing and so on. That is why data pre-processing is the first step performed [34] whose end product is training set.

Data pre-processing is an important step of data mining. Whereas, data mining is the analysis step of "*Knowledge Discovery in Databases*". Knowledge Discovery results in discovery of patterns in large databases.

Traditional PCA is meant only for linearly separable problem. For the purpose of this paper linearly non-separable problems are considered. Linearly non-separable problems are effectively dealt with Kernel PCA (KPCA). KPCA is an advanced form of conventional PCA [35].

4.3 ANN+ANN cascade

The resultant product of the pre-processing stage is now ready to train neural network. This dataset is a fine dataset free from outliers. The network is then trained using the clean dataset. The network when tested will perform much better than the baseline.

As the dataset was initially split into two parts, 80% was used for training and the remaining 20% was retained for further testing of the model. 20% was reserved for testing the model's ability that how much the model has learned to generalize. Unseen Test Set (UTS) provides a means of determining the model's ability to perform on unseen data. In the figure 5 the procedure of cascading two AANs is illustrated along with the Unseen Test Set (UTS).

This is how the hybrid model is developed and performance is analyzed. Below are represented the prediction accuracies recorded after making the model hybrid.
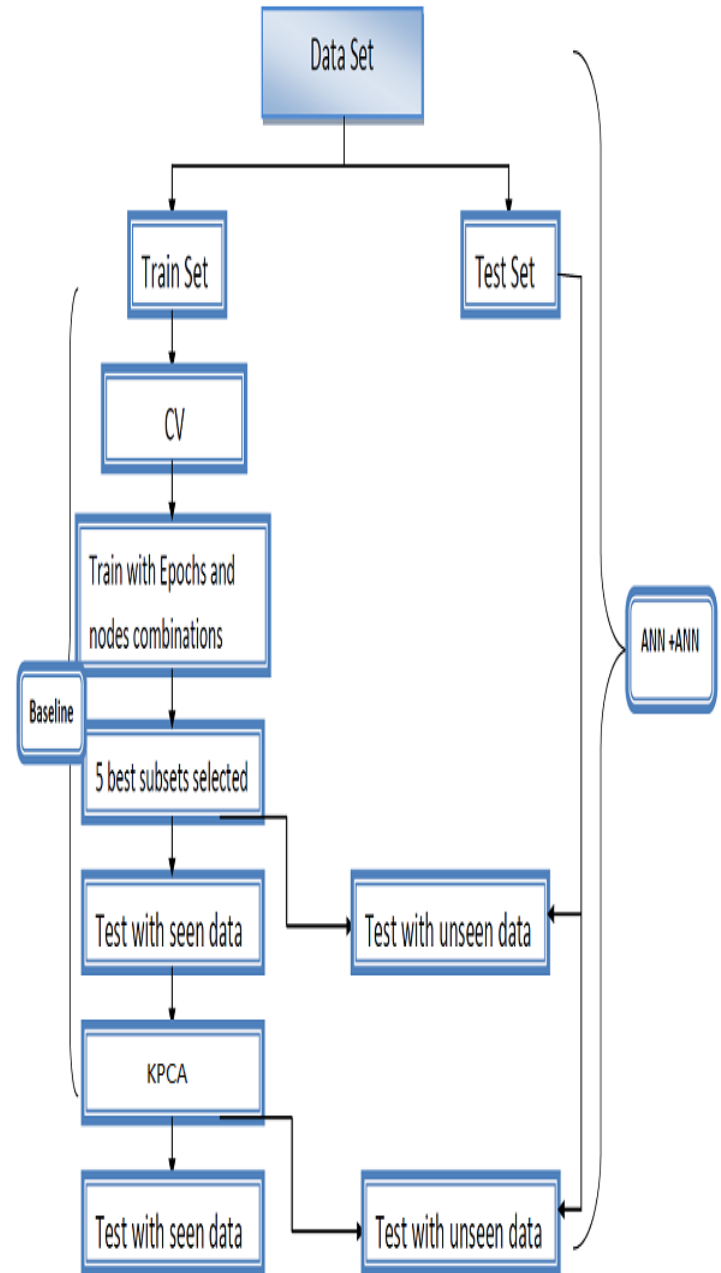
Figure5: Process of combining two ANN model

4.4 Evaluation Method

Up till now it has been shown that how the model is developed. Now is the time to see how the model is evaluated. Before this, the term prediction accuracy has been mentioned time and again without giving any idea about from where does the prediction accuracy is measured and how.

In the following sections the concept of confusion matrix is given along with its calculations. Apart from that, description of the datasets on which testing is performed is also given.

4.5 Confusion Matrix

In order to visualize the state attained by a learner a commonly used tool is known as confusion matrix. In a classification problem there are two cases firstly, the accurate

class and secondly the predicted class. Confusion matrix provides visualization of the results obtained by a predictive model and shows how well or poor is the performance. Confusion matrix as the name indicates is a matrix of the order n×n, where n is the number of classes in the problem. The predicted classes are along the columns whereas the actual classes are represented along the rows. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. It is called confusion matrix because it facilitates to see if the system is confusing classes, which is often called mislabeling of one class as another. As an example for the purpose of explanation see figure 6. Following conclusions are drawn from the above confusion matrix.

- The total number of correct predictions are 1250 (520 + 730)
- The total number of incorrect predictions are 39 (13 + 26)
- There are total of 1289 cases (1250+39).
- The error rate is 39/1289 = 0.0302
- The accuracy rate is 1250/1289 = 0.969

Predicted

Actual

	Banana	Orange
Banana	520	26
Orange	13	730

Figure6: A Confusion Matrix for Two Class Problem

In the same way the prediction accuracies for the glass dataset and cancer dataset are calculated.

4.6 Unseen Test Set (UTS) and Seen Test Set (STS)

After the model is built and training is performed by using clean dataset, testing is done on two different sets. The unseen test set that is the one that was retained initially. When testing is performed on UTS it is seen that the performance of the model deteriorates. The performance of baseline model deteriorates more than the hybrid model. Hence proved that hybrid model outperforms the single one.

Testing is then performed on STS. It has been observed that the performance of hybrid model increases more than the baseline model. Graphical illustration for these two tests is given in figures 13 &14.

It can be seen from the figure 13 that on UTS performance deteriorates. The performance of baseline is much less as compared to the performance of a hybrid model (glass dataset).

It can be concluded from figure 14 that performance of the hybrid model decreases on UTS less than the baseline model (cancer dataset).

5. EXPERIMENTAL RESULTS

The results obtained are given in tables and now is the time to analyze the results and to draw conclusions. The results are discussed one by one and conclusions drawn are mentioned accordingly.

5.1 Results of the baseline model

Refer to table 1 and table 2. As there is no ground reality against which to choose the number of hidden nodes and learning epochs (iterations) four different epochs and five different numbers of nodes have been used. Results are recorded while training and testing both are performed on the same dataset. So nothing can be said about how well the model has learned to generalize.

Table 1: Parameter Settings (Glass Dataset)

Number of Neurons	50 epochs	100 epochs	200 epochs	300 epochs
8	92.5	92.9	96.3	94.4
12	97.6	95.7	93.4	96.2
16	94.8	75.7	94.3	96.7
24	92.9	93.9	94.8	93.9
32	93.5	93.4	98.1	88.3

Table 2: Parameter Settings (Cancer Dataset)

Number of Neurons	50 epochs	100 epochs	200 epochs	300 epochs
8	96.4	95.8	96.9	96.6
12	97.4	97.4	97.7	96.9
16	97.1	97.6	96.8	96.2
24	98.1	96.9	96.8	96.2
32	97.7	97.4	96.9	96.7

After this 5-fold cross-validation is performed; its results are discussed under the next heading.

5.2 Results after cross-validation

To check the model's generalization ability 5-fold cross validation has been performed. From five subsets, each of the subset with the highest accuracy from twenty different models is selected. See table 3 and table 4. Then five best baseline models have been selected. These five models have been chosen because they provided highest prediction accuracy thus giving us a chance to determine best settings.

Table 3: Prediction Accuracies of twenty different baseline models (Glass Dataset)

Learning Epochs	50					100				
Hidden Nodes	8	12	16	24	32	8	12	16	24	32
Glass Dataset										
Fold 1	87.7	69.6	87.8	91.7	90.9	90.8	91.6	90.9	90.7	87.8
Fold 2	91.2	88.2	90.8	85.2	88.2	94.1	94.2	85.3	91.2	94.3
Fold 3	91.3	91.1	94.1	79.4	82.3	88.2	91.1	92.1	94.0	89.7
Fold 4	94.2	94.2	90.9	82.8	90.0	89.9	94.2	90.0	94.2	94.8
Fold 5	93.9	92.8	93.8	94.8	91.0	92.8	93.7	90.1	91.0	89.9
Average	91.7	87.2	91.5	86.8	88.5	91.16	92.9	89.6	92.2	91.3

Learning Epochs	200					300				
Hidden Nodes	8	12	16	24	32	8	12	16	24	32
Glass Dataset										
Fold 1	93.9	90.9	84.8	81.8	90.9	91.7	93.9	93.9	93.4	84.8
Fold 2	88.2	94.1	88.3	89.9	94.1	94.1	91.7	94.1	94.2	94.9
Fold 3	94.1	93.4	94.5	94.1	91.2	88.2	88.5	91.1	91.2	93.1
Fold 4	91.1	94.2	94.3	94.8	94.5	94.8	91.4	94.2	91.9	94.3
Fold 5	91.6	94.4	36.2	92.9	90.0	92.2	91.9	92.9	92.0	94.4
Average	91.8	93.5	79.6	90.7	92.3	92.3	91.5	93.2	92.5	92.3

Table 3&4 show the results recorded after performing cross-validation. These are the results of baseline model.

Table 4 : Prediction Accuracies of twenty different baseline models (Cancer Dataset)

Learning Epochs	50					100				
Hidden Nodes	8	12	16	24	32	8	12	16	24	32
Cancer Dataset										
Fold 1	91.2	88.3	91.1	90.3	92.0	90.1	89.2	94.6	90.2	94.6
Fold 2	93.7	88.3	90.2	90.3	90.2	91.0	89.2	92.8	93.7	94.6
Fold 3	92.1	89.2	92.9	92.1	92.9	91.9	90.9	88.4	90.2	89.9
Fold 4	91.2	89.3	91.1	92.1	91.1	89.3	90.2	90.2	91.1	89.2
Fold 5	92.1	90.2	94.7	91.3	90.3	89.3	90.2	88.5	90.3	88.3
Average	92.1	89.0	92.0	91.2	91.3	90.3	89.9	90.9	91.1	91.3

Learning Epochs	200					300				
Hidden Nodes	8	12	16	24	32	8	12	16	24	32
Cancer Dataset										
Fold 1	88.3	90.9	94.6	94.6	90.1	91.2	90.2	89.3	94.6	94.6
Fold 2	94.6	90.3	34.5	88.4	88.3	91.2	90.2	94.6	90.3	88.3
Fold 3	92.2	93.9	37.2	90.2	88.3	93.9	91.9	92.0	93.0	90.1
Fold 4	91.3	92.1	37.2	89.4	90.1	92.1	92.0	89.4	91.3	89.3
Fold 5	94.7	90.4	35.5	90.3	92.0	94.7	91.2	89.4	92.2	89.3
Average	92.2	91.5	47.8	90.6	89.8	92.6	90.9	90.9	92.3	90.3

Table 5: Five best baseline model (Glass Dataset) with STS

Test Set	Learning Epochs	Hidden Nodes	Accuracy
Fold 1	200	8	93.9
Fold 2	100	8	94.1
Fold 3	50	16	94.1
Fold 4	50	8	94.3
Fold 5	50	24	94.8

Table 5 shows the best baseline model selected for glass dataset from 100 different settings. Similarly, table 6 shows the selected baselines for cancer dataset. These tables show the accuracies of the model on seen test set (STS). For the purpose of showing the performance deterioration on unseen data the baseline models are tested on Unseen Test Set (UTS). The results after performing tests on UTS are shown in the tables 7&8. It can be seen from tables 7&8 that the prediction accuracies decrease on the data that was already not known to the model. Later on, hybrid model will also be tested on the UTS comparisons will be made between the two.

Table 6: Five best baseline models (Cancer Dataset) with STS

Test Set	Learning Epochs	Hidden Nodes	Accuracy
Fold 1	100	16	94.6
Fold 2	100	32	94.6
Fold 3	200	12	93.9
Fold 4	50	24	92.1
Fold 5	50	16	94.7

When testing was performed on UTS following accuracies were recorded.

Table 7: Five best baseline model (Glass Dataset) with UTS

Test Set	Learning Epochs	Hidden Nodes	Accuracy
Fold 1	200	8	87.7
Fold 2	100	8	87.9
Fold 3	50	16	87.9
Fold 4	50	8	88.0
Fold 5	50	24	88.5

It can be seen that the difference is very high between the results of table 5 & table 7. Same is true for cancer dataset.

For cancer dataset similar test is performed. Again the comparison between table 6 and table 8 shows the significant decrease in accuracies.

5.3 Results after cascading two ANNs

After cascading two ANNs prediction accuracies are recorded again. The first ANN does the task of data reduction (feature extraction) and the second one performs the task of prediction. The comparisons of the accuracies before and after making model a hybrid one are shown. Significant increase in the prediction accuracies has been illustrated.

5.3.1 Comparison between Baseline and Hybrid Model with STS (Glass Dataset)

The difference between the accuracies in the tables 5&9 can be felt easily. The performance of the hybrid model on STS is much higher as compared to the baseline model.

Table 8: Five best baseline models (Cancer Dataset) with UTS

Test Set	Learning Epochs	Hidden Nodes	Accuracy
Fold 1	100	16	88.3
Fold 2	100	32	88.3
Fold 3	200	12	87.6
Fold 4	50	24	85.9
Fold 5	50	16	88.5

Table 9: Prediction accuracies of ANN+ANN model (Glass dataset) with STS

Test Set	Learning Epochs	Hidden Nodes	Accuracy
Fold 1	200	8	96.5
Fold 2	100	8	96.5
Fold 3	50	16	95.6
Fold 4	50	8	97.1
Fold 5	50	24	97.7

5.3.2 Comparison between Baseline and Hybrid Model with STS (Cancer Dataset)

By looking at tables 6& 10 it can be concluded that prediction performance for the baseline model is less than hybrid model even on seen data (STS).

Table 10: Prediction accuracies of ANN+ANN model (Cancer dataset) with STS

Test Set	Learning Epochs	Hidden Nodes	Accuracy
Fold 1	100	16	97.5
Fold 2	100	32	98.2
Fold 3	200	12	97.9
Fold 4	50	24	96.8
Fold 5	50	16	97.7

5.3.3 Comparison between Baseline and Hybrid Model with UTS (Glass Dataset)

By looking at the tables 7& 11 it has been that performance of the baseline model deteriorates more on UTS than a hybrid model.

Table 11: Prediction accuracies of ANN+ANN model (Glass dataset) with UTS

Test Set	Learning Epochs	Hidden Nodes	Accuracy
Fold 1	200	8	95.3
Fold 2	100	8	90.7
Fold 3	50	16	76.2
Fold 4	50	8	95.3
Fold 5	50	24	96.8

5.3.4   Comparison between Baseline and Hybrid Model with UTS (Cancer Dataset)

When tables 8& 12 are compared it has been noticed that performance of hybrid model deteriorates less than the baseline model. Same was true for the first dataset.

Table 12: Prediction accuracies of ANN+ANN model (Cancer dataset) with UTS

Test Set	Learning Epochs	Hidden Nodes	Accuracy
Fold 1	100	16	96.4
Fold 2	100	32	96.7
Fold 3	200	12	97.1
Fold 4	50	24	96.1
Fold 5	50	16	96.2

6. GRAPHICAL ILLUSTRATION

Graphs are drawn for the illustration purpose of the comparison results. Figure 13& 14 represent the glass and cancer dataset respectively. Both the graphs show that on STS performance of hybrid was more than baseline model. After performing the test on UTS performance of both the models undergo degradation. The degree of degradation confronted by hybrid model is far less than that of the baseline model.
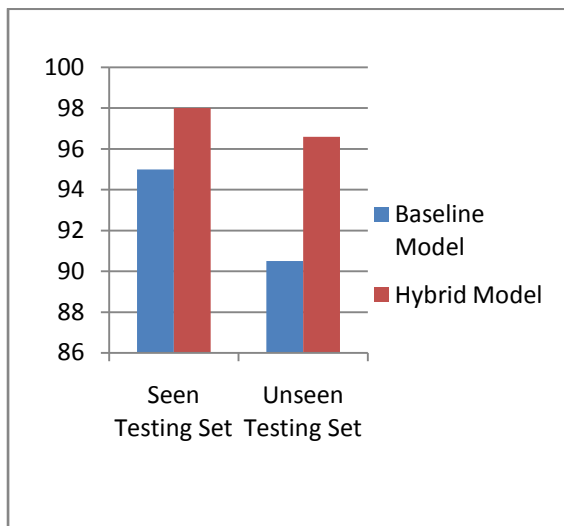

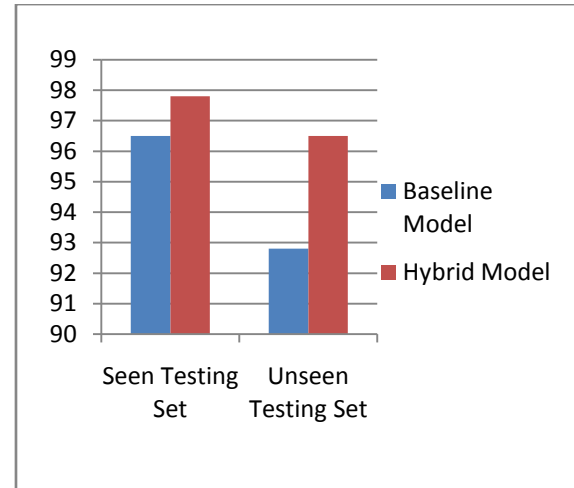**Figure 1: Performance deterioration on UTS (baseline and hybrid comparison) Glass dataset**

Figure 2: Performance deterioration on UTS (baseline and hybrid comparison) Cancer dataset

7. CONCLUSIONS AND FUTURE RESEARCH

The extent to which ANNs can predict accurately can be significantly enhanced by using hybrid approaches rather using a single baseline model. The accuracy is tested on using the dataset that is a sample taken from the entire population. The extent to which the ANNs can predict accurately depends on how well is the dataset extracted. The dataset must cover most of the potential situations in a particular problem. The ultimate goal is to enhance the accuracy on the whole population rather than a sample. So, we conclude by saying that the more accurately the dataset is organized and presented the higher is the accuracy.

There are certain steps that an individual has to meticulously follow before using ANNs for their practical considerations. The model must be tested historically. Model should be tested for what has happened in the past. Only that will tell the adequacy of the model. Other indicators must also be used with ANNs.

7.1 Future Research

For future work several things can be considered. Firstly, the pre-processing step is very important. As it removes the outliers so it enhances the performance of the model. There are a number of different methods proposed till date for outlier removal. Different methods can be applied and results be analyzed. Feature selection may also be applied in comparison with feature extraction and useful conclusions could be drawn. In feature selection a subset of original variables is selected for classifier design.

The next thing would be the choice of classifier. In this paper we predict the class membership of a pattern based on a single classifier induced from the dataset. In ensemble methods classification techniques are combined, which is related in particular to the problem of combining decisions coming from multiple targets. Ensemble methods (or classifier combination techniques) has been an area of growing research in recent years and is related to developments in the data fusion literature where, in particular, the problem of decision fusion (combining decisions from multiple target detectors) is being addressed extensively.

8. ACKNOWLEDGMENTS

9. REFERENCES

[1] S. Wermter. Knowledge extraction from transducer neural networks. Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Techniques, 12:27-42, 2000.

[2] O. Chapelle, B. SchÄolkopf, and A. Zien, editors. Semi-Supervised Learning. MIT Press, 2006.

[3] N. V. Chawla, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer. Learning ensembles from bites: A scalable and accurate approach. Journal of Machine Learning Research, 5:421-451, 2004.

[4] H. Chen and X. Yao. When does diversity in classifier ensembles help generalization? Machine Learning, 2008. In Revise.

[5] H. Chen and X. Yao. Regularized negative correlation learning for neural network ensembles. IEEE Transactions on Neural Networks, 2009b.

[6] H. H. Dam, H. A. Abbass, C. Lokan, and X. Yao. Neural-based learning classifier systems. IEEE Transactions on Knowledge and Data Engineering, 20(1):26-39, 2008.

[7] R. Diaz-Uriarte and S. Andres. Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7(1):3, 2006.

[8] P. C. Pendharkar, Genetic algorithm based neural network approaches, Expert Systems with Applications, 6714-6720, 36, 2009.

[9] Hung, S. Y., Yen, D. C., & Wang, H. Y, Applying data mining to telecomm churn management. Expert Systems with Applications, 31(3), 515–524, 2006.

[10] Tan, P. N., Steinbach, M., & Kumar, Introduction to data mining. Boston, MA: Addison Wesley Publishing, 2006.

[11] T. G. Dietterich. Ensemble methods in machine learning. Lecture Notes in Computer Science, 1857:1-15, 2000.

[12] N. Garcia, C. Hervas, and D. Ortiz. Cooperative coevolution of artificial neural network ensembles for pattern classification. IEEE Transactions on Evolutionary Computation, 9(3):271-302, 2005.

[13] G. Giacinto and F. Roli. Design of effective neural network ensembles for image classification purposes. Image and Vision Computing, 19(9-10):699-707, 2001.

[14] M. M. Islam, X. Yao, and K. Murase. A constructive algorithm for training cooperative neural network ensembles. IEEE Transaction on Neural Networks, 14(4):820-834, 2003.

[15] L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning, 51(2): 181-207, 2003.

[16] X. Liao, H. Li, and L. Carin. Quadratically gated mixture of experts for incomplete data classification. In ICML'07: Proceedings of the 24th international conference on Machine learning, 553-560, 2007.

[17] N. Ueda. Optimal linear combination of neural networks for improving classification performance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(2):207-215, 2000.

[18] Z. Zhou, J. Wu, and W. Tang. Ensembling neural networks: many could be better than all. Artificial Intelligence, 137(1-2):239-263, 2002.

[19] S. Wermter and R. Sun: Hybrid Neural Systems, Springer-Verlag, 1778, 1-13, 2000.

[20] Rajeev Alur and George J. Pappas, editors. Hybrid Systems: Computation and Control, 7th International Workshop, HSCC 2004, Philadelphia, PA, USA, 25-27, 2004, Proceedings, volume 2993 of LNCS. Springer, 2004.

[21] A. van der Schaft and H. Schumacher. An Introduction to Hybrid Dynamical Systems. Springer, London, 2000.

[22] S.W. Al Sayegh. Hybrid Neural Networks, IEEE, IJCNN, 2006.

[23] Chiang, D. A., Wang, Y. F., Lee, S. L., & Lin, C. J. Goal-oriented sequential pattern for network banking churn analysis. Expert Systems with Applications, 25, 293–302, 2003.

[24] Coussement, K., & Van den Poel, D. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications, 34, 313–327, 2008.

[25] Han, J., & Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001.

[26] Laurene Fausett. Fudamentals of Neural Networks.

[27] Hung, C., & Tsai, C.-F. Segmentation based on hierarchical self-organizing map for markets of multimedia on demand. Expert Systems with Applications, 34(1), 780–787, 2008.

[28] Hung, S. Y., Yen, D. C., & Wang, H. Y. Applying data mining to telecom churn management. Expert Systems with Applications, 31, 515–524, 2006.

[29] Kim, H. S., & Yoon, C. H. Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. Telecommunications Policy, 28, 751–765, 2004.

[30] Kim, M., Park, M., & Jeong, D. The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services. Telecommunications Policy, 28, 145–159, 2004.

[31] Acir, N. A support vector machine classifier algorithm based on a perturbation method and its application to ECG beat recognition systems. Expert Systems with Applications, 31(1), 150–158, 2006.

[32] Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2004). A practical guide to support vector classification. Technical Report, Department of Computer Science and Information Engineering, National Taiwan University.

[33] Chin-Fong, Yu-Hsin. Customer Churn prediction by using Hybrid Neural Networks. Expert Systems with Applications. 36, 12547-12553, 2009.

[34] Li, C. T., & Tan, Y. H. Adaptive control of system with hysteresis using neural networks. Journal of Systems Engineering and Electronics, 17, 163–167, 2006.

[35] Zhang, X., Edwards, J., & Harding, J. Personalised online sales using web usage data mining. Computers in Industry, 58, 772–782, 2007. University of Washington.