COMPARATIVE ANALYSIS OF DIMENSION REDUCTION TECHNIQUES FOR CLASSIFICATION OF RADAR RETURNS FROM IONOSPHERE USING A MODEL BASED ON MACHINE LEARNING TECHNIQUE

Ali Raza* (engr.aliraza@hotmail.com), Ali Zaidi* (alixedi@yahoo.com),
M. Umair Khalid* (mumairk@live.com)

*NESCOM, Karachi, Pakistan.

Abstract— Advances in technology have immensely increased data collection and storage capabilities during the past decades. This progression has overloaded many science fields with huge amount of raw data. Scientists and researchers working in domains as diverse as engineering, medicine, economics etc are facing problems handling and performing analysis on the large database gathered from different resources. This data overload is presenting new challenges in data analysis. Traditional statistical methods fail in such scenarios because, as the data dimensionality increases the number of variables also increase which in turn increases the complexity of the system. The processing of high-dimensional datasets is computationally expensive. To cope up with this problem, Dimension Reduction Techniques are applied. Some of the applications in which dimension reduction techniques are incorporated are pattern recognition, image & video processing, signal processing, classification & regression problems, defense applications, bioinformatics etc. In Pattern Recognition, Object Recognition and classification problems, dimension reduction techniques are used as a pre-process. The reduced dataset is then applied to a classifier for recognition. In this research paper, a comparative study is conducted of two dimension reduction techniques for the binary classification problem in order to analyze the efficiency of HF (High Frequency) Wireless Communication in a given environment, using a benchmarked Ionosphere data set from the UCI Machine Learning repository. The two dimension reduction techniques compared are PCA (Principal Component Analysis) and ICA (Independent Component Analysis). Performance based analysis is performed on the results generated after the application of ICA and PCA - by implementing both the outcomes on a model based on Machine Learning Technique.

I. INTRODUCTION

Since wireless communication is becoming more and more convenient way of communication, many High Frequency (HF) wireless devices are rapidly researched and developed to be applicable commercially as well as in the field of Defense. Most long-distance High Frequency (HF) radio communication (between 3 and 30 MHz) is the result of skywave propagation. The phenomenon of Skywave Propagation deals with the refraction of Electromagnetic Waves back to Earth's surface by the ionosphere. This High Frequency Radio propagation via the ionosphere is an important and crucial means of long-distance radio communication which is especially used by ships and other sea vessels thousands of miles apart from each other. This Communication depends on several factors of the channel it is using which in turn makes it very critical to analyze the state of ionosphere in order to ensure successful communication. Several decision support systems and classification processes may be used in the course of carrying out this critical task.

Several international research programs are being conducted in order to analyze the ionospheric returns which may in turn enhance the efficiency of Radio Communication and Surveillance activities through Skywave Propagation via ionosphere. One of such mega research programs is the High Frequency Active Auroral Research Program (HAARP) [1]. It is an Ionospheric Research Program jointly funded by the US Air Force, the US Navy, the University of Alaska and the Defense Advanced Research Projects Agency (DARPA). The objective of HAARP is to analyze the ionosphere and research the potential for developing ionospheric enhancement technology for radio communications and surveillance purposes [2].

The purpose of this paper is to identify statistically superior approaches intended for the classification of RADAR returns from the ionosphere. This is done by comparing two dimension reduction techniques using a system based on Machine Learning technique. A detailed analysis of the techniques used with experimental results will also be produced.

The two major factors of this decision support system which is based on Machine learning Algorithms are: Dimension Reduction and Classification [3]. Dimension reduction is critical for the classification of the RADAR returns from the ionosphere. The process dimension reduction should be able to discriminate between the useful and the unwanted data and should be able to discard the superfluous data while preserving the data mandatory for efficient classification.

In this paper two of the most widely used dimension reduction techniques PCA (Principal Component Analysis) and ICA (Independent Component Analysis) are used to reduce the dimensions of ionosphere dataset acquired from the UCI Machine Learning repository. The reduced datasets from both techniques are applied to Feed Forward Back Propagation Neural Network (FFBP-NN). Then a comparison of results of both models is discussed.

There are many problems faced by former decision support systems based on Artificial Intelligence. It requires a large amount of database to compare the

characteristics of the data and fails when new circumstances or changing environment is encountered by the system. In the proposed System, Neural Networks are used which have the capability to intelligently classify the RADAR returns from the ionosphere and are capable to learn from the new circumstances which ultimately results in an efficient and reliable system.

II. PRELIMINARIES

A. Machine Learning Technique

Formerly, several memory based Decision Support Systems and classifiers were used to achieve the recognition or classification processes. Such memory based systems, like the ones implemented on the principles of Artificial Intelligence, have several drawbacks as compared to the systems based on Memory-less adaptive, Machine Learning Models. Major advantages of Memory-less Adaptive Machine Learning Models over the Memory based systems are as follows:

- They do not require a large database of examples to carry out the decision making process once the training is completed.
- Machine Learning Models have the capability of accommodating the data compression and Dimension Reduction algorithms.
- No large computational demands and large memory requirements due to accessibility to several Data Preprocessing Techniques.
- Machine Learning Models can be designed to be adaptive (like Neural Networks) and may have the ability to learn in the pre as well as post classification phase.

The process of designing a system based on Machine Learning Technique is illustrated in Fig.1 and generally involves:

- Acquiring the data and importing it into the system.
- Normalizing and removing the noise from the data by preprocessing.
- Discriminating between the useful part of the data and clutter (it is the part of the data which is not required and will only provide resistance in the classification process).
- Feature extraction from the processed data in order to provide basis for classification.
- Classification process based on the extracted features.
- Supervised learning which involves establishment of a rule by which we can classify new observations into one of the existing classes.

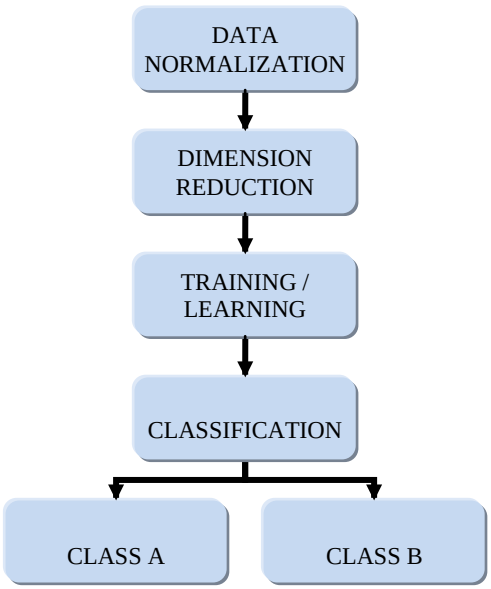

Figure 1. Block Diagram illustrating Machine Learning Process

B. Dimension Reduction Technique

Dimension Reduction or feature extraction is an important preprocessing step to reduce the complexity of the data by reducing its dimensionality. The process of Dimension Reduction has to be implemented with great precision in order to preserve the important components or attributes of data which are necessary for the classification. When the dataset is small with less attributes, the classification performance of several Neural Networks is very efficient and the Network itself is very simple. But when we have to consider larger, high dimensional data then the classification performed by the Neural Networks is usually not that efficient due to the fact that the Network becomes very complex in order to manipulate large number of attributes of the data [4].

So for better classification of high dimensional data, it is usually considered to apply some techniques which could reduce the dimensions of data while preserving important components of it which are mandatory for efficient classification. Dimension reduction techniques like PCA (Principal Component Analysis) or ICA (Independent Component Analysis) or several other preprocessing techniques may be applied for reducing the attributes of the ionosphere data for better classification of the RADAR returns in order to ensure efficient communication of wireless High-Frequency Radio Wave Propagation via ionosphere. In this paper, PCA (Principal Component Analysis) and ICA (Independent Component Analysis) Techniques are used for Dimension Reduction, and then compared for computing better classification results.

C. Principle Component Analysis (PCA)

PCA (Principal Component Analyses) is a way of determining patterns in a data and identifying the similarities and dissimilarities [5] in it. PCA is a very helpful technique of detecting several patterns in those data which cannot be easily represented and analyzed graphically.

When patterns are identified in a data by using PCA, it's easier to reduce the dimensions of data without much loss of information. Principal Component Analysis involves:

- Computing the Eigen Values and Eigen Vectors from the original Data
- Deciding which Eigen Vectors are significant and forming a Feature Vector
- The Eigen Vector with the highest Eigen value is termed as Principle Component of the data set
- The higher the Eigen value, more significant will be the corresponding Eigen Vector
- Forming a new coordinate system based on the Feature Vector
- Mapping data to the new space
- Reduce complexity of data by reducing its dimensionality

Forming the Feature Vector and extracting the most important components of the data is the task which could be easily and efficiently performed by applying PCA (Principle Component Analysis) Technique. In almost all the datasets having large number of attributes or dimensions, it's very important to extract feature vectors efficiently for accurate classification [6]. The Eigen values computed by this technique help in the formation of a Feature Vector which contains the most important attributes of data which are required for accurate classification.

D. Independent Component Analysis (ICA)

ICA (Independent Component Analyses) is a technique in which an independent condition is achieved which may results in more optimized components as compared to those achieved from the variance optimization which is a part of Principle Component Analysis (PCA).

The objective of ICA is to lessen the statistical dependency between the basis vectors it generates. Mathematically, this may be presented as WXT = U, where ICA looks for any linear transformation W that diminishes the statistical dependency among the rows of U, provided a training set X (as before). In contrast to PCA, the basis vectors in ICA are neither orthogonal nor ranked in order. Also, there is no closed form expression to find W. Alternatively, several iterative algorithms have been proposed based on various search criteria [7]. In this paper, we will concentrate on FastICA algorithm [8] and will generate several experimental results based on it.

E. Classification

After Feature Extraction, the process of classification is performed for analyzing and classifying the RADAR returns from the ionosphere. The process of classification involves discriminating between several objects and to generalize them in several classes.

When training the network and going through the learning process, a distinction can be made between supervised and unsupervised learning for classification. In Supervised learning for classification, each training data is labeled according to the class of events that the data represents [9]. Perceptron and Multi-Layer Perceptron (MLP) are such models which involves Supervised Learning for Classification.

While in Unsupervised learning for classification, each dataset is not accompanied with its class label [10]. It models the structure of the data either in the form of probability density function or by structuring the data in form of cluster centers and widths. Gaussian Mixture Models and Kohonen Networks are such models which involves Unsupervised Learning for Classification.

The process of classification is an important step in determining the ionospheric RADAR returns as Positive (Good) or Negative (Bad). So it may require a lot of computation if the data in hand is huge and high dimensional. On the other hand, the performance demand of Radio Communication via ionosphere in the field of Defense continuously requires new and innovative techniques to be developed for efficient High Frequency Communication with high accuracy and low error rate. In order to achieve this goal, extremely efficient and fast system is needed in order to classify the ionospheric returns in real-time to analyze and determine the feasibility of further communication in that environment.

Neural Networks have shown some promising results where computing large data in real-time environment is required. Neural Networks are found to be efficient classifiers with adaptive capabilities.

F. Neural Network

Neural Computing is the technology which is based on networks of "neuron-like" units [11] .The technique of Neural Networks has shown some promising results where the task of prediction and recognition is required. The feature that distinguishes Neural Network from other techniques is the ability to internally develop and learn the algorithms and to continuously improve itself for better classification by adjusting its weights. Their learning capability depends on the network topology, learning algorithm and the problem which is to be analyzed.

The architecture of the neural network involves densely interconnected nodes embedded in layers and an arrangement of interconnected Neurons. These networks also contain simple Computational Units like Summation Unit. The Neural Network topology may consist of two or more layers containing several nodes. The Input Layer accepts the data for learning or testing while the Output Layer generates or transfers the outcomes of the computation performed in the layers which resides between Input and Output Layers [12].

G. Feed Forward Back Propagation Neural Network (FFBP-NN)

The most common learning algorithm of Neural Network which provides efficient learning environment and accurate classification is Feed Forward Back

Propagation (FFBP) Neural Network. It consists of an Input layer, Hidden Layer and an Output Layer.

The Structure of Feed Forward Back propagation Neural Network is displayed in figure 2.

During the training phase of Feed Forward Back Propagation (FFBP) Neural Network, the training dataset is fed in to the Neural Network via Input Layer. The data is then propagated through Hidden layer and comes out of the Neural Network through Output Layer. This process is called the Forward Pass of Feed Forward Back Propagation Algorithm.

The output values originated from the Output Layer are then compared with the actual target output values.

The error between the target output values coming from the Output Layer and the actual output values is calculated and propagated back towards the Hidden Layer. This process is called the Backward Pass of the Feed Forward Back Propagation Algorithm.

In this way, a properly trained Feed Forward Back Propagation (FFBP) Neural Network tends to predict accurately the results of the inputs which it has never seen before. And this Network keeps on adjusting itself during the training phase by comparing the actual and processed Outputs and forming and efficient Network which can handle undefined inputs with low error rate.

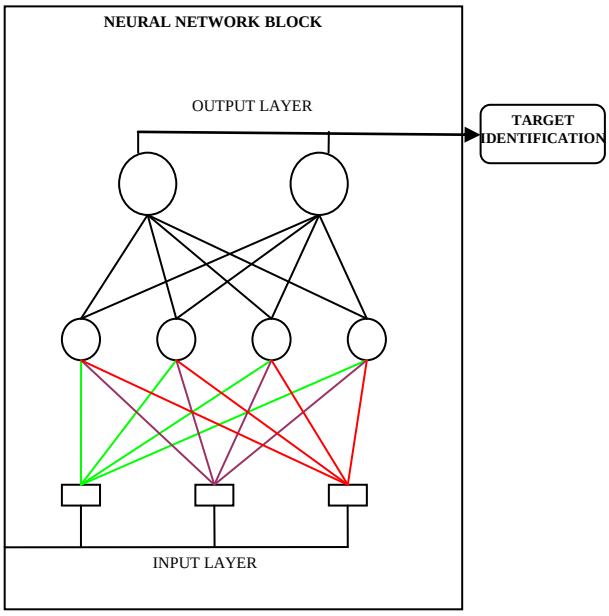

Fig 2. Feed forward back propagation neural network model

III. METHODOLOGY

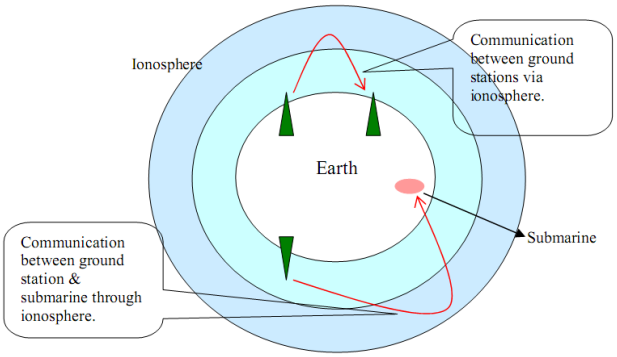

Fig 3. Skywave propagation via ionosphere

The experimental results are generated by analyzing a benchmark Ionosphere dataset from the UCI Machine Learning Repository [13] which was generated by a system of 16 high-frequency antennas used for analyzing the characteristics of ionosphere to ensure efficient High Frequency Radio Communication.

The ionosphere is so termed due to the fact that it is a region within the atmosphere of the Earth in which ion are present. It serves as a medium for long range communication especially with ships and other sea vessels thousands of miles away. Ground station communication medium range and long range is also viable. Most long-distance High Frequency (HF) radio communication (between 3 and 30 MHz) is the result of skywave propagation via ionosphere. Fig 3. illustrates the skywave propagation via ionosphere layer of the Earth's atmosphere.

The task is to distinguish 2 kinds of signals – "positive" that happen to be reflected by means of free electrons in ionosphere and carry beneficial details regarding ionosphere composition, and "negative" which passed through ionosphere without reflection. The electromagnetic signals are characterized by a set of 17 pulsations with every single pulse acquiring two attributes. Therefore the sum of the number of features is of a single sample is 34. These are considered as dimensions of data. The dataset was divided into two subsets: 50% of the patterns were used for learning and the remaining 50% for validation and testing.

A. Principle Component Analysis (PCA)

When we have large high dimensional data then the classification performed by the Neural Networks according to this data is usually not that efficient and the Network becomes very complex due to large number of attributes of the data. So for better classification of high dimensional data, PCA (Principle Component Analysis) Technique is applied in order to reduce the dimensions of

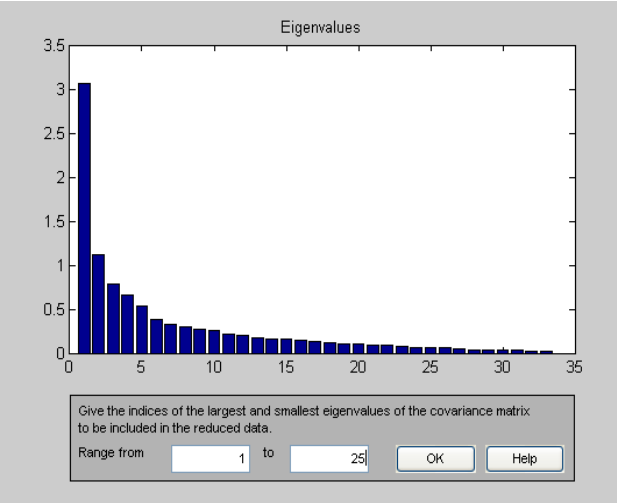

Fig 4. PCA illustration-eigen value plot of the 34 dimensional ionosphere data

data while preserving the important components of it which are mandatory for efficient classification.

The Eigen Values and their corresponding Eigen Vectors for the Ionosphere data are illustrated in Fig. 4.

By analyzing the plot of the Eigen Values in Fig 3 and by several experiments it may be determined that the accuracy of the results does not decline when dimensions more then 25 were considered for this particular dataset under the proposed scenario. On the basis of this analysis, it is found safe to consider 25 dimensions to be processed for classification and the remaining to be neglected without losing much information.

B. Independent Component Analysis (ICA)

Several ICA Algorithms may be suitable for a particular problem. In our experimentations as shown in fig. 5, we have preferred to apply the FastICA Algorithm [8] due to its captivating convergence capability in addition to its quality of furnishing results with substantial computational swiftness especially for high dimensional data.

It may also be noticed that in contrast to ICA, PCA functions solely upon second-order statistics i.e variances that in fact corresponds to the most significant eigen-vectors associated with the particular sample covariance matrix.

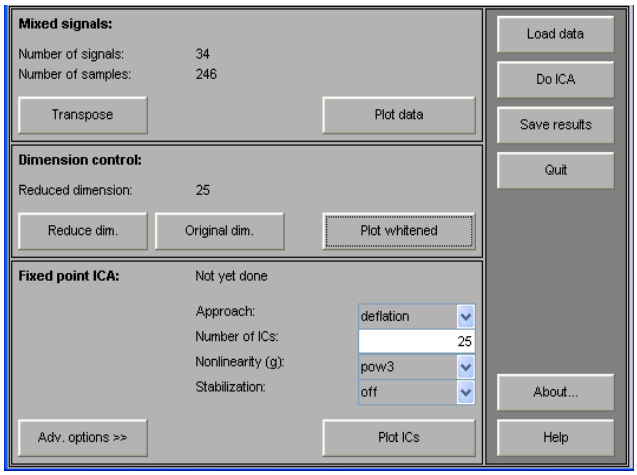

Fig 5. ICA illustration applied on ionosphere data

C. Feed Forward Back Propagation Neural Network

Here, for the Classification purpose, FFBP (Feed Forward Back Propagation) Neural Network is implemented in MATLAB on the Ionosphere Data. The Ionosphere data has been preprocessed by applying PCA and ICA techniques. Both the scenarios are considered separately and results are produced to compare the efficiency of the classification by feeding the preprocessed data into FFBP (Feed Forward Back Propagation) Neural Network.

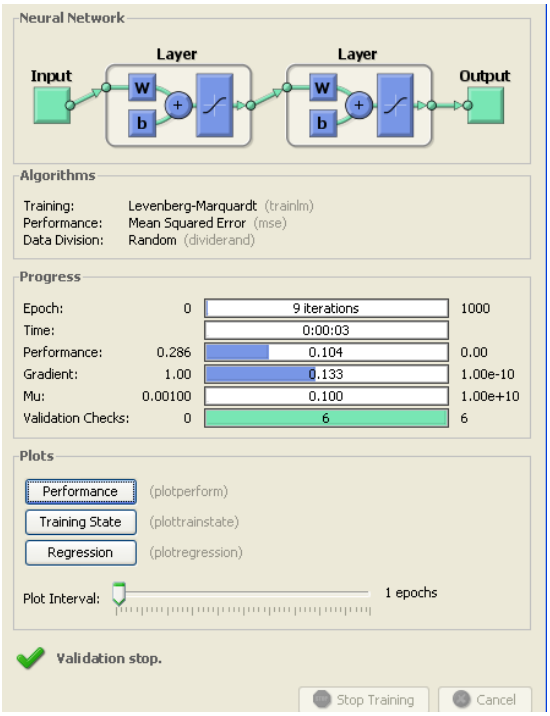

Fig 6. Making and training of FFBP neural network

Implementing FFBP Neural Network With the Preprocessing PCA Technique

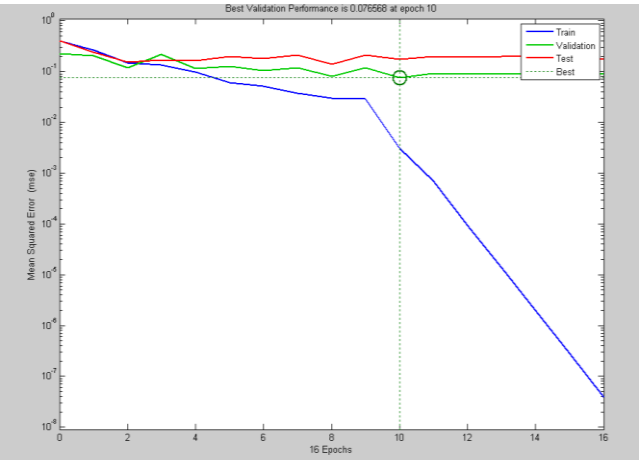

Fig 7. Performance of FFBP-NN with PCA

When the dimension reduction technique namely Principle Component Analysis (PCA) was applied to the data set as shown in Fig. 4, it was found that the dataset could be transformed from 34 dimensions to 25 dimensions without loosing the important attributes of data which are necessary for optimal classification.

The performance graph of the FFBP-NN when fed with preprocessed (PCA) 25 dimensional Data is shown in Fig 7.

Implementing FFBP Neural Network With the Preprocessing ICA Technique

The dimension reduction technique namely Independent Component Analysis (ICA) was applied to the Ionosphere data set as illustrated in Fig. 5, and then imported that preprocessed data in the FFBP - NN (Feed Forward Back Propagation – Neural Network) whose performance graph is shown below in Fig 8:

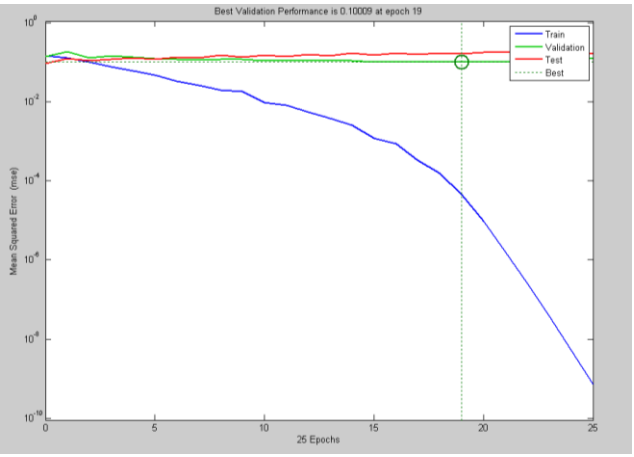

Fig 8. Performance of FFBP with ICA

IV. EXPERIMENTAL RESULTS

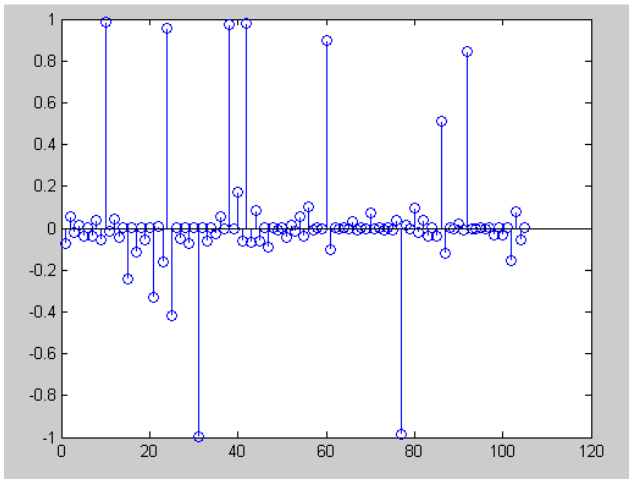

Fig 9. Classification results of ionosphere data with PCA

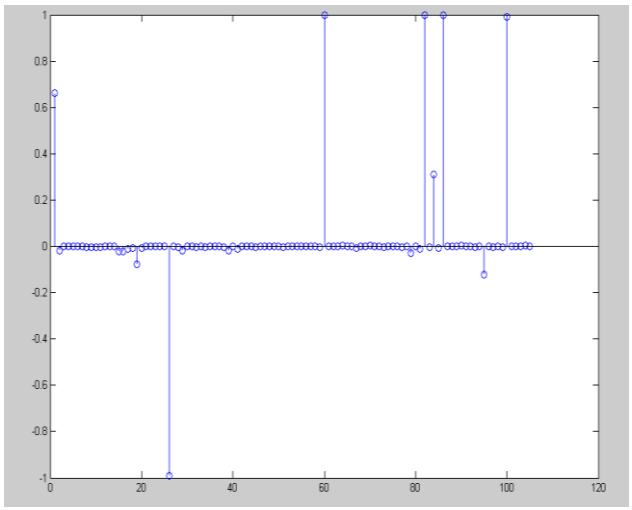

Fig 10. Classification results of ionosphere data with ICA

The experimental classification results of the FFBP-NN (Feed Forward Back Propagation – Neural Network) are shown in this section. Below, in Fig 9, classification results of the preprocessed 25 dimensional PCA data are shown.

Similarly, in Fig. 10, classification results of the preprocessed data gone through FastICA technique is shown.

The comparison of the percentages of correct classification for FFBP-NN Using PCA and ICA is tabulated in Table I:

TABLE I.
Percentages of Classification for FFBP-NN using PCA and ICA

Compression Technique	Classification Technique	Classification Efficiency
PCA	FFBP-NN	90.47
ICA	FFBP-NN	95.23%

V. CONCLUSION

Presented were the two statistically superior approaches intended for improving the classification of RADAR returns from the ionosphere to ensure successful long-distance radio communication. This was done by comparing two dimension reduction techniques namely PCA and FastICA using a classification system based on FFBP-NN (Feed Forward Back Propagation-Neural Network).

Comparison in between the dimension reduction techniques namely PCA and ICA is very complex as several factors like the type of data, its attributes, amount of noise and varying conditions may be taken into account. From detailed analysis of the techniques used with experimental results tabulated in Table 1, it may be depicted that the ionosphere data has shown better classification performance using FFBP-NN when preprocessed with FastICA as compared to PCA.

VI. REFERENCES

[1] "HAARP Fact Sheet". HAARP. 15 June 2007. Retrieved 2009-09-27.

[2] "Purpose and Objectives of the HAARP Program". HAARP. Retrieved 2009-09-27.

[3] Yongzeng Shen, Qicong Wang and Shiming Yu, "A Target Recognition of Wavelet Neural Network Based on Relative Moment Features" IEEE Intelligent Control and Automation , pp-4089-4092, 2004

[4] Ethem Alpaydm, "Introduction to Machine Learning", Massachusetts Institute of Technology Press, 2004.

[5] "Handbook Of Neural Network Signal Processing", CRC Press LLC, edited by Yu Hen Hu and Jenq-Neng Hwang, 2002.

[6] Ruiz-del-Solar, J., Kottow, D. "Neural-based architectures for the segmentation of textures", Proceedings 15th International Conference on Pattern Recognition, vol. 3, pp.1080- 1083, Barcelona, Spain, 3-7 Sept. 2000.

[7] Hyvärinen, A. and E. Oja, *Independent* Component Analysis: Algorithms and *Applications.* Neural Networks, 2000. 13(4-5): p. 411-430.

[8] Laboratory of Computer And Information Science Adaptive Informatics Research Centre [http://www.cis.hut.fi/projects/ica/fastica/code/dlcode.shtml]

[9] "Machine Learning, Neural and Statistical Classification" (Ellis Horwood Series in Artificial Intelligence), Prentice Hall, edited by D. Michie, D.J. Spiegelhalter, C.C. Taylor, July 1994.

[10] "The Handbook of Brain Theory and Neural Networks", Massachusetts Institute of Technology Press, edited by Michael A. Arbib, pp-1-4, 2003.

[11] Gang Liu and Robert M. Haralick "Optimal matching problem in detection and recognition performance evaluation, " Pattern Recognition Volume 35, Issue 10, October 2002, Pages 2125-2139

[12] Ruiz-del-Solar, J., Kottow, D. "Neural-based architectures for the segmentation of textures", Proceedings 15th International Conference on Pattern Recognition, vol. 3, pp.1080- 1083, Barcelona, Spain, 3-7 Sept. 2000.

[13] UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets.html]