IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-8, Aug 2018 E-ISSN: 2347-2693

Diabetes Classification Using Machine Learning Techniques With The Help of Cloud Computing

J. Seetha[1*] , T. Chakravarthy[2]

[1]Department of Computer Science, A.V.V.M. Sri Pushpam College, Poondi, Thanjavur, India
[2]Department of Computer Science, A.V.V.M. Sri Pushpam College, Poondi, Thanjavur, India

[*]*Corresponding Author: seetha.twinhami@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 13/Aug/2018, Published: 31/Aug/2018

Abstract - Now a days Diabetes mellitus is a major global public health problems. The machine learning techniques can be applied to help the people in detection of diabetes at an early stage and treatment, which may help in avoiding complications. In our work attempts to propose three kinds of techniques K- Nearest Neighbor (KNN), Naive Bayes (NB) and Artificial Neural Network (ANN) for classifying the individual user as diabetic or non diabetic.Providing diagnostic aid for diabetic by using a set of data that contains only medical information obtained without advanced medical equipments, can help number of people who want to discover the disease or the risk of disease at an initial stage. The experimental system achieves classification accuracy of KNN is 92.59%, NB is 85.71% and ANN is 94.64%. The aim of this study is to classify diabetes disease and deploy in to cloud for cost effective and easy to use.

Keywords - Diabetes mellitus, K-Nearest Neighbor, Naïve Bayes, Artificial Neural Network, Cloud.

I. INTRODUCTION

Diabetes is a disease in which, the body does not produced or properly used the insulin hormone. In our work, an effective machine learning algorithm is proposed for the classification of type Diabetes Mellitus patients. This machine learning algorithms are used for classification will find the optimal hyper-plane which divides the various classes. This can possibly make a huge positive impact on a lot of people lives. The cause of diabetes continues to be a mystery, although both genetics and environmental factors such as obesity and lack of exercise appear to play a major role. The early stage of diabetes is called pre-diabetes. It is also known as impaired glucose tolerance, which is a condition where the blood glucose level increase to a level that is higher than the normal range for most people, but it is still low enough to be considered as diabetes [1]. There are three main types of diabetes: Type 1 diabetes, Type 2 diabetes, Gestational diabetes. Type1 diabetes is an autoimmune disease where the pancreas is unable to produce adequate insulin or produce no insulin at all. This type of diabetes is common within the people under the age of 20years, it occurs in childhood of a person or young adult.

Type 2 diabetes is mostly found among ageing and overweight people. It is known as adult onset diabetes. In this type of diabetes, is often caused by lack of exercises as found among Americans Gestational diabetes is the third types of diabetes, is a condition that women can get, starting from the six months of their pregnancy. About 4% of pregnant women develop gestational diabetes [2].

People with diabetes often stand the risk of developing a number of other serious health problems. Frequent high blood sugar levels can cause serious diseases affecting the heart, blood vessels, eyes, kidney, nerves and teeth. The common symptoms of diabetes include frequent urination with large volume of urine, excessive thirst, extreme hunger, unexplained weight loss, increased fatigue, feeling very tired, feeling ill, blurry vision etc [3].

A. *Overview of Cloud*

Cloud computing is a general term for anything that involves delivering hosted services over the Internet. It can also help people stay more connected to their self-care. It is a new technology and have good performance in storing, managing, sharing and accessing information[4]. The cloud computing based solutions in healthcare can help the physicians to stay in touch with their patients and examine their health condition effectively at a low cost. Cloud services are broadly divided into three categories: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS) .

B. *Cloud Features*:

The additional resources as needed from the consumer requests, and similarly releases these resources when they are not needed. Different clouds offer distinct sorts of resources, e.g., processing, storage, management software, or application services [5]. Clouds are typically erected using large numbers of inexpensive machines. As a result, the cloud vendor can more easily add capacity and can more rapidly replace machines that fail, compared with having machines in multiple laboratories. Generally speaking these machines are as consistent as possible both in terms of configuration and location.

- ***Automated Backup*****:** The automated backup and archival options are offer from different clouds. The cloud may move data or computation to improve responsiveness. Some clouds monitor their offerings for spiteful activity.
- ***Virtualization*****:** In clouds, the Hardware resources are usually virtual; these are shared by multiple users to improve efficiency. That is, the same physical resources are supported to several lightly utilized logical resources.
- ***Parallel Computing*****:** Expressing and executing easily parallelizable computations are using Map/Reduce and Hadoop frameworks, which may use hundreds or thousands of processors in a cloud.

The rest of the paper is organized as follows section2 discuss about various techniques of diabetes classification process, section3 deal with proposed method and section4 discuss about some experimental results.

II. RELATED WORK

Large number of work has been done to find out the efficient methods of medical diagnosis for various disease. In our research work is an attempt to predict efficiently diagnosis of diabetes with accuracy rate.

Clustering noise removal and classification approach: the Self Organizing Map (SOM), Principal Component Analysis (PCA) and Neural Network (NN) are used for Noise removal and classification purpose. SOM is used for clustering purpose and PCA is used noise removal purpose for diabetes disease diagnosis from the real world dataset. SOM clustering is used as an unsupervised classification method to cluster the data of experimental dataset into similar groups, PCA is used for dimensionality reduction and dealing with the multi-co linearity problem in the experimental data. The combination of SOM, PCA and NN, a hybrid intelligent system is increase the predictive accuracy 92.28% of diabetes disease. But this hybrid method used unsupervised clustering technique, this work cannot use large amount of dataset [6].

Data mining techniques: are used like Multi Layer Perceptron (MLP) and the Bayesian Net (BN) classification techniques. MLP is a development from the simple perceptron in which extra hidden layers are added. In this process, more than one hidden layer can be used. The network topology is constrained to be feed forward ie, loop free. Bayesian Net is a statistical classifiers which can predict class membership probabilities, such as the probability that given tuple belong to a particular class or not. And also used information gain feature selection technique, it produce 81.89% accuracy of model with less number of features used only 6 features from Pima Indian Diabetic Data set of UCI [7].

Linear Discriminant Analysis and Support Vector Machine techniques: these techniques are used for feature selection and classification techniques. In feature selection, the Linear Discriminant Analysis (LDA) is used to eliminate the irrelevant features, complexity of data is reduced and saves computational time. There are two techniques are used in the classification (i) Support Vector Machine (SVM), (ii) Feed Forward Neural Network (FFNN). Support Vector Machine is supervised learning approach that constructs hyper plane surface that classify the data with a largest margine. Support Vector Machine classifies the linearly separable data. If the data is non linear, SVM do not attain classification tasks. To overcome this limitation, these support vectors are transformed into higher dimensional feature space, which is Linear SVM. Feed Forward Neural Network is the input signals are directly feed as input to the classifier with any reduction of feature [8]. The feature selection with Linear Discriminant Analysis (LDA) and classification using SVM and FFNN, the accuracy is 75.65%. But this work produce low accuracy rate.

III. PROPOSED WORK

In our approach use Pima Indian Diabetic Dataset from UCI repository, which is classified under two methods diabetic and non diabetic. This dataset consist of eight attributes and one class, which is number of time pregnant, plasma glucose, blood pressure, skin fold thickness, serum insulin, body mass index, diabetic pedigree function, age and diabetic or non diabetic [9]. The classification outcome thus obtained is evaluated for classification accuracy along with sensitivity and specificity measures [10]. Finally, the results attained are send to cloud storage repository for public usage. The table 1 shows attribute ID and attribute name of the Pima Indian diabetes dataset.

Table 1: Pima Indian Diabetic Dataset

Attribute Id	Attribute Name
F1	Pregnant
F2	Plasma Glucose
F3	Diastolic Blood Pressure
F4	Triceps Skin Fold Thickness
F5	Serum-Insulin
F6	Body Mass Index

F7	Diabetes Pedigree Function
F8	Age
Class	Diabetic or Non-Diabetic

At the infrastructure level, we use implement resource elasticity mechanisms that scale the underlying resource infrastructure for optimal performance of Cloud Based Health Care Service (CBHCS) and thus reducing both data costs and time. Finally to address data security mechanisms at multiple levels and provide role based access control to ensure the protection of critical medical data of patients. So our work is cost efficiency, secure and quick access time.

In our proposed method attempts three kinds of techniques such as (i) K-Nearest Neighbor,(ii) Naive Bayes and (iii) Artificial Neural Network for classification of diabetes.

A. K- Nearest Neighbor (KNN)

The first approach of diabetic classification is K- Nearest Neighbor (KNN), it is a simple classifier that works well on basic recognition problems [11]. This classifier works with R tool it gives the better accuracy 92.59% but it has some drawback, it doesn’t have any training phase and all the training data is utilized during the testing phase. Therefore it is considered as a lazy learning algorithm as it defers computation until classification is performed.

There are many distance functions but Euclidean distance is the most commonly used measure. It is mainly used when data is continuous. Manhattan distance is also very common for continuous variables as in Eqn (1) and Eqn (2).

Euclidean Distance Measure:

$$d(x,y)=\sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \qquad (1)$$

Manhattan Distance Measure with module function:

$$d(x,y)=\sqrt{\sum_{k=1}^{n}|x_k - y_k|^2} \qquad (2)$$

where x_k and y_k are the kth attribute of x and y.

B. Naïve Bayes Classifier (NB)

The second approach of classification is Naive Bayes classifier, it is considered to be one of the most powerful probabilistic classification technique that classifies high dimensional input data [12][13]. It utilizes Bayesian theorem to calculate the probability “ P ” that an unknown instance “ Y ” is classified as class “ C ” among a set of possible outcomes C = {c_1, c_2 .... c_n}. This probability is known as posterior probability and is expressed using the Bayes rule as in Eqn (3).

$$P(C/Y) = \frac{P\left(\frac{Y}{C}\right).P(C)}{P(Y)} \qquad (3)$$

In Naive Bayes classifier gives the result in quick access time but it produce the low accuracy rate 85.71%. So we select another approach called Artificial Neural Network (ANN).

C. Artificial Neural Network (ANN)

It is one of the powerful method in an intelligent field for classifying the diabetic patients into two classes. For achieving better results than both KNN and Naive Bayes classifier. ANN is a set of connected input output network in which weight is associated with each connections [14]. It consist of three layers, input layer, hidden layer and output layer. In supervised learning the neural network is performed by adjusting the weight of connection. By updating the weight iteratively performance of network is improved. On the basis of connection Artificial Neural Network (ANN) can be classified into two categories, (i) feed forward network and (ii) recurrent network. Feed forward neural network, do not form cycle whereas in recurrent neural network connection form cycle. In our approach uses recurrent neural network the neurons of neural network are activated by the weighted sum of input. During training, the inter connection weight are optimized until the network reaches the specified level of accuracy. It has many advantages like parallelism, less affected with noise, good learning ability and high accuracy rate. In ANN works with R tool for classification of diabetes produce the accuracy is 94.64%. Fig 1 shows basic structure of artificial neural network with one hidden layer. Flow chart of Artificial Neural Network is shown in Fig 2.

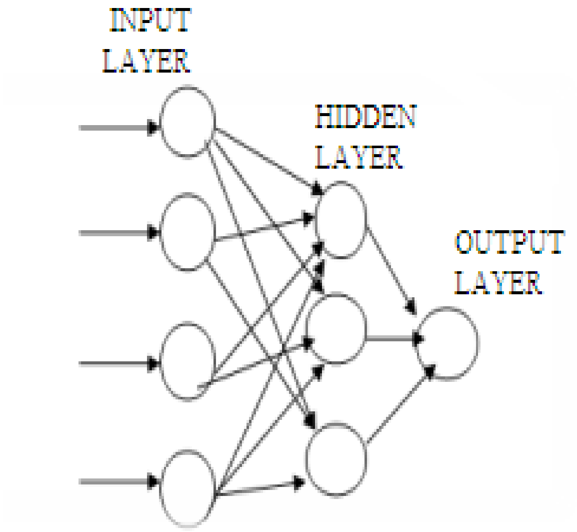

Figure 1: Structure of Artificial Neural network

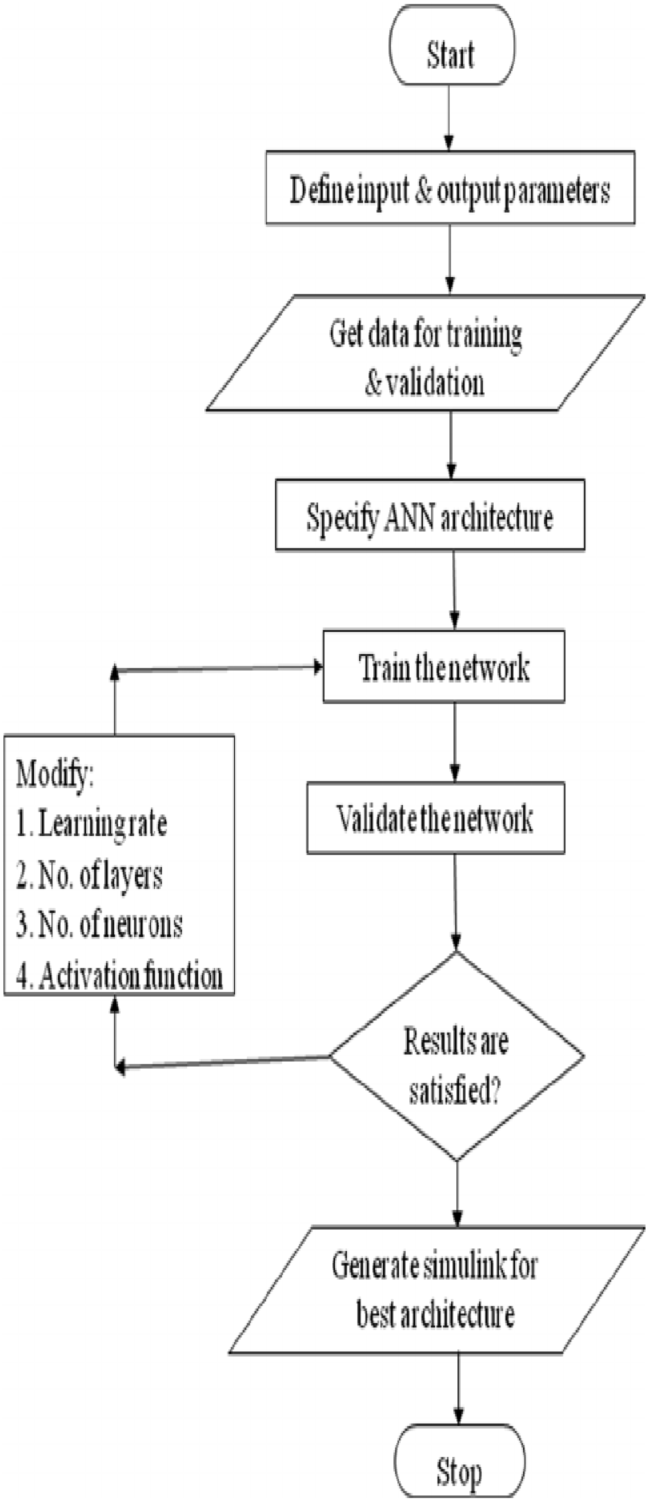

Figure 2: Flow Chart of Artificial Neural Network

The overview of diabetes classification process is shown in Fig 3.

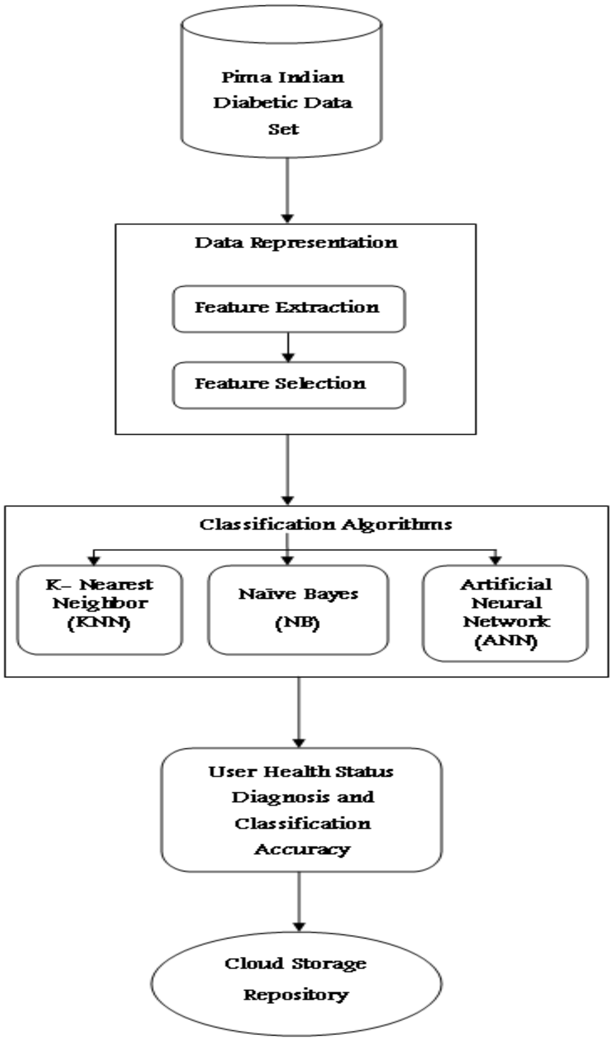
Pima Indian Diabetic Data Set
Data Representation
Feature Extraction
Feature Selection
Classification Algorithms
K– Nearest Neighbor (KNN)
Naïve Bayes (NB)
Artificial Neural Network (ANN)
User Health Status Diagnosis and Classification Accuracy
Cloud Storage Repository

Figure 3: Process of Diabetic Classification

IV. EXPERIMENTAL RESULTS

A. Performance Measures

Performance can be evaluated various measures such as classification accuracy, sensitivity and specificity. These measures are evaluated using true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

In our research work have taken 65 subjects as sample data, 40 subjects are “diabetic” and 25 are “non diabetic” out of 65 sample data. In diabetic, the 23 subjects are selected for training and remaining 17 subjects are used for testing. In non diabetic, the 14 subjects are utilized for training while the remaining 11 subjects are used for testing.

K-Nearest Neighbor:

In diabetes, the True Positive (TP) result is 15 and False Negative (FN) value is 0.

In non diabetes, the False Positive (FP) rate is 2 and True Negative (TN) rate is 11 is shown in table 1.2.

Naïve Bayes:

In diabetes, the true positive (TP) result is 14 and false negative (FN) value is 1.

In non diabetes, the false positive (FP) rate is 3 and true negative (TN) rate is 10 is shown in table 1.2.

Artificial Neural Network:

In diabetes, the True Positive(TP) result is 15 and False Negative (FN) value is 0.

In non diabetes, the False Positive (FP) rate is 1 and True Negative (TN) rate is 13 is shown in table 1.2.

- ***Accuracy:***

It refers to the closeness of a measured value to a standard or known value.

Accuracy = (TP+TN) / (TP+FP+TN+FN)

- ***Sensitivity:***

It refers to the ability of a test to correctly identify those with the disease (true positive rate).

Sensitivity = TP/ (TP+FN)

- ***Specificity:***

It is the ability of the test to correctly identify those without the disease (true negative rate).

Specificity = TN/ (TN +FP)

The classification accuracy and the sensitivity and specificity measures of the three classifiers, K-Nearest Neighbor, Naïve Bayes and Artificial Neural Network. The results are shown in table 2.

Table 2: Classification accuracy, sensitivity and specificity measures of the classifier.

Classifier	Diabetes	Non-diabetes	Accuracy %	Sensitivity/ Specificity

K-NN	TP=15
	FN=0
		FP=2
		TN=11
			92.59	1.00/0.84

NB	TP=14
	FN=1
		FP=3
		TN=10
			85.71	0.93/0.76

ANN	TP=15
	FN=0
		FP=1
		TN=13
			94.64	1.00/0.92

V. CONCLUSION AND FUTURE SCOPE

In this work presented a cloud based Health Care Service (CBHCS) that performs Pima Indian Diabetic dataset from UCI repository. It applies K-Nearest Neighbor (KNN), Naïve Bayes (NB) and Artificial Neural Network (ANN) for diabetes classification. In this approach classified the user as diabetic and non-diabetic KNN achieves 92.59% accuracy, the sensitivity is 1.00 and the specificity measure is 0.84. NB classifier achieves 85.71% accuracy, the sensitivity is 0.93 and the specificity measure 0.76. Finally ANN achieves better classification accuracy 94.64%, the sensitivity is 1.00 and the specificity is 0.92. This work is cost effective, globally accessible and a highly converged health care solutions because our data are stored and retrieved from cloud.

As a part of future scope, we would like to extend our classification process, if the subject as diabetic then the diabetic person as type1 diabetic or type2 diabetic or gestational diabetic. These types of classification is interesting for future work.

REFERENCES

[1] Barrie Sosinsky, “Cloud Computing Bible”, Wiley Publication, India. Pp 083-120, 2011. For Book

[2] American Diabetes Association, “Diagnosis and Classification of Diabetes Mellitus”, American Diabetes Association Journals, Vol 37, Pp. 81-90, January 2014. For Journal

[3] E O Olaniyi, K Adnan,” Onset Diabetes Diagnosis using Artificial Neural Network”, International Journal of Scientific & Engineering Research, Vol 5 Issue 10, Oct 2014. For Journal

[4] Ch Chakradhara Rao, Mogasala Leelarani and Y Ramesh Kumar, “Cloud:Computing Services And Deployment Models”, International Journal of Engineering and Computer Science, Vol. 2, Issue 12, pp.3389 – 3392, Dec 2013. ISSN:2319 – 7242. For Journal

[5] Sean Marston, Zhi Li , Subhajyoti Bandyopadhyay, Juheng Zhang , Anand Ghalsasi, “Cloud computing - The business perspective”, Elsevier, pp. 176–189, 2010. For Journal

[6] Mehrbakhsh Nilashi, Othman Ibrahim, “Accuracy Improvement for Diabetes Disease Classification a Case on a Public Medical Dataset”, Fuzzy Information and Engineering Elsevier 2017.

[7] Amit kumar Dewangan and Pragati Agrawal “Classification of diabetes Mellitus using Machine Learning Techniques”, International journal of engineering and applied science, vol.2, issue 5, may 2015. For Journal

[8] Parashar A, Burse K and Rawat K, “A Comparative Approach for Pima Indians Diabetes Diagnosis using LDA - Support Vector Machine and Feed Forward Neural Network”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, pp. 378-383, 2014. ISSN: 2277 128X. For Journal

[9] Dilip Kumar, Sanchita paul, “Classification of Pima Indian Diabetes Dataset using Naïve Bayes with genetic algorithm as an attribute selection”. Communication and computing system, Dec 2017. For Coference

[10] Akil Bansal, Manish kumar Ahirwar, Piyush kumar sukla, “A Survey on Classification Algorithms used in Healthcare Environment of the Internet of Things”. International journal of Computer Sciences and Engineering, Vol 6, Issue 7, Pp 883-887, July 2018. For Journal

[11] Pankaj Deep kaur and Inderveer Chana ” Cloud based intelligent system for delivering health care as a service”, 2013 Elsevier, Volume 113, Issue 1, pp. 346-359, January 2014. For Journal

[12] Aiswarya Iyer, S.Jeyalatha, Ronak Sumbaly,” International Journal of DataMining & Knowledge Management Process, Vol.5, No.1, January 2015. For Journal

[13] Pooja, Komal kumar Bhatia, "Spam Detection using Naïve Bayes Classifier". International journal of Computer Sciences and Engineering, Vol 6, Issue 7, Pp 712-716, July 2018. For Journal

[14] Manaswini Pradhan, Ranjit Kumar Sahu,"Predict the onset of diabetes disease using Artificial Neural Network (ANN)", International Journal of Computer Science & Emerging Technologies, Vol 2, Issue 2, April 2011. For Journal

Authors Profile

Miss. Seetha.J received her Bachelor's degree in Computer Science during the year 2013 and Master of Computer Science in year 2015. Currently she is pursuing her Ph.D (Full Time) in computer Science from A.V.V.M Sri Pushpam College, Poondi, Thanjavur. She receives stipend from state government and she has published 4 research papers in reputed international journals. Her main research work focuses on cloud computing and machine learning.

Dr.T.Chakravarthy, Associate Professor Department of Computer Science, A.V.V.M Sri Pushpam College, Poondi, Thanjavur. Completed his Ph.D in Computer Science from SIGC, Trichy, Bharathidasan University in 2008. He has published 49 papers in reputed both national and international journals and one text book. He has produced 3 Ph.D Scholars and 85 M.Phil Scholars.