International Journal of Engineering & Technology, 7 (3.27) (2018) 325-330

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Correlation Feature Selection (CFS) and Probabilistic Neural Network (PNN) for Diabetes Disease Prediction

K. Kalaiselvi[1*], P. Sujarani[2]

[1]*Associate Professor & Head, Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies (VISTAS) (Formerly Vels University), Chennai.*

[2]*Research Scholar, Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies (VISTAS) (Formerly Vels University), Chennai.*

E-mail:sujiraji873@gmail.com

**Corresponding author E-mail:kalairaghu.scs@velsuniv.ac.in*

Abstract

The healthcare sector is a broad area with the abundance of patient information, which creates enormously large records day by day. Though the scientific industry is rich in information but it is poor in knowledge. Diabetics are considered as a primary health issue of the world. As per the WHO 2014 survey According to WHO 2014 report, over 422 million people are affected from the diabetics globally. In the minimization of massive investigations implied on the patients, the data mining uses many mechanisms and strategies to diagnose the diabetic problem. The main objective of this proposal is to introduce assemble Data Mining based Diabetes Disease Prediction System which provides a detailed analysis of diabetics using the database of diabetics patient. The formulated work comprises of two stages such as feature selection ad prediction methods which are made known to maximize the outputs of diabetes disease prediction. Initially Correlation Feature Selection (CFS) is formulated to identify the salient features for the diabetic repository. The identified features are fed into the classifier named Probabilistic Neural Network (PNN) classifier. As the diabetic of the patient is classified using PNN meanwhile the accuracy can be fine – tuned when using the identified features. Depending on the category of data, the diabetic information is gathered from the learning repository. The outputs are correlated with the current algorithms namely Back Propagation Neural Network (BPNN), *Multilayer Perceptron,* Neural Network (MLPNN) were used to fetch the outputs.

Index Terms: *Data mining, diabetes dataset, healthcare industry, Correlation Feature Selection (CFS), feature selection, Probabilistic Neural Network (PNN), machine learning repository and classifier.*

1. Introduction

Stroke, heart disease, cancer, chronic lung cancer and diabetes are categorized into Non-Communicable Diseases (NCDs) which are the cause of 70% deaths carries out globally where Diabetes mellitus Type II is common to everyone [1]. Diabetes mellitus is considered to be a chronic disease. Diabetes maximizes the critical level of micro-vascular problem and macro-problem. People affected by diabetic are probable of acquiring two to four times cardio vascular diseases**.** This leads to the complex health issues like kidney(renal) failure, heart diseases, paralytic attack and vision problems [2-3].

Diabetes creates problems in the whole body. It can be controlled through medications still it maximizes the heart diseases. It is calculated that 422 million people are suffering from diabetics worldwide and this may be doubled in the next two decades [4]. The diabetic is high in the countries like India, China, USA, Indonesia, Japan, Pakistan, Russia, Brazil, Italy and Bangladesh [4]. In the past 3 decades of developed countries there are increased numbers of diabetics; inhabitants commenced to recognize about the diabetics has acutely rooted into every one's life. In the overall population of diabetics, the growth rates of male diabetic patients are higher than that of the female patients.

At present the enormous information is gathered from the patient records, from the hospitals. An analysis strategy is performed by means of mining of data by using a technique called Knowledge discovery for prediction utilization which helps in formulating inferences. This strategy aids in the process of decision making by using its algorithms where huge amount of data are extracted from clinical centres. Data mining strategies can be enforced in diagnosing the diabetic disease at the initial phase considering the importance of early detection of the disease in order to avoid complex situations.

Weather prediction, market survey, engineering design and customer relationship management are the different areas in the human society where Data mining techniques are successfully imposed. Yet the applications which are used in the disease prediction and the analysis of medical data can still be improved. For instance, all the hospitals have a large number of patient's medical data, and it is significant to revise, complement and gather knowledge from this information in order to enhance in medical analysis and prediction of disease[5-6]. It is a sensible hypothesis that there exist a variety of beneficial patterns which are available for the researchers for exploration.

The formulated system uses two phases which are feature selection and prediction methods to maximize the outputs of diabetes disease prediction. Initially Correlation Feature Selection (CFS) is formulated to identify the salient features for the diabetic repository. The identified features are fed into the classifier named Probabilistic Neural Network (PNN) classifier. Depending on many researches which were conducted used a dataset called Pima Indians Diabetes Dataset from the University of California, Irvine

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

(UCI) Machine Learning Database [7]. The formulated research targets to attain knowledge from the diabetes database to extract the ultimate results.

2. Literature Review

Nowadays, the data mining strategies are used to maximize the frequency in the prediction of the probability of disease. Various algorithms and techniques are implied by the researchers. These have enormous potential in the field of research. In this work, some of the significant works are correlated to the formulated work are given below:

Patil et al [8] a formulated Hybrid Prediction Model (HPM) which utilizes Simple K-means clustering algorithm focuses at checking the identified class label of submitted data (wrongly classified instances are eliminated, i.e. pattern are derived from the source data) and meanwhile employing the classification algorithm to the output data. To construct the final classifier C4.5 algorithm is implied by implementing the k-fold cross-validation technique. From the UCI data sets the Pima Indians diabetes data was fetched. 59.4–84.05% was the range of accuracies obtained.

Ahmad et al [9] formulated an innovative technique to correlate the accuracy of prediction using the Multilayer Perceptron in Neural Networks (MLPNNs) along with the tree-based algorithms, specifically Pima Indian diabetes mellitus data set for the ID3 and J48 algorithms. The experiment on classification is carried out with the help of algorithms in WEKA to find out the diabetes class or non-diabetes containing 768 patient data sets. The outputs prove that the pruned J48 tree achieved the greater accuracy of 89.3% when correlated to 81.9% produced by the multilayer perceptron's. On eliminating the pregnant attribute for a specified number of times the prediction accuracy for the pruned J48 tree further raises to 89.7%.

Marcano-Cedeño et al [10] formulated the Artificial Metaplasticity on Multilayer Perceptron (AMMLP) as a model for prediction, for predicting the diabetic disease. The data set used to validate the formulated AMMLP was the Pima Indians database. The outputs achieved by AMMLP were correlated with other existing algorithms of the same set of data. The most effective output yielded by the AMMLP algorithm is 89.93%.

Vijayan and Anjali [11] proposed an innovative approach to review the advantages of various pre-processing strategies for DSS (decision support systems) for diagnosing diabetes which depends on Support Vector Machine (SVM), Naive Bayes classifier and Decision Tree. The pre-processing methods concentrated on this work are Principal Component Analysis and Discretization. The evaluation is done on the accuracy variation with and without pre-processing techniques. Weka was the tool used to carry out this work. The data set selection was made from the UCI repository of machine learning.

Wu et al [12] formulated an innovative pre-processing technique in which the model is classified into two phases. One is the improved K-means algorithm and the other is logistic regression algorithm. To correlate the output of the research with the other research conducted, the Pima Indians Diabetes Dataset and the Waikato Environment for Knowledge Analysis tool were used. The final output achieved the greater accuracy of prediction which is 3.04% when compared to other outputs.This model proves the sufficiency of the data set quality.

Sowjanya et al[13] proposed a mobile/android application based solution to eradicate the insufficiency of consciousness about diabetes. The application utilizes machine learning strategies to find out diabetes levels for the users. Meanwhile, the application also renders knowledge about diabetes and certain recommendations on the disease. Four Machine Learning (ML) algorithms correlations were carried out. The Decision Tree (DT) classifier produces outstanding results when compared to the other 4 ML algorithms. Therefore DT classifier is utilized as a model for machinery of the mobile application for the prediction of diabetics utilizing the real world dataset gathered from the esteemed hospital in the Chhattisgarh, State of India.

Songthung and Sripanidkulchai [14] formulated a classification to extract a broad set of data collected 12 hospitals in Thailand during the year 2011-2012 containing 22,094 records of selected data set consisting of females of age 15 years and more than that. RapidMiner Studio 7.0 was also embedded with Naive Bayes and Chi-squared Automatic Interaction Detector (CHAID) Decision Tree (DT) classifiers in order to find out the females with greater complication and correlate the outputs to the current hand-computed techniques to predict the diabetes complexity. The target of predicting the risks facilitates to find out the individuals who are detected with diabetes. The outputs show that the classification formulated maximizes the coverage in the prediction of diabetics when compared to hand-computed scoring.

Chandrakar and Saini [15] formulated an innovative Indian Weighted Diabetic Risk Score (IWDRS). Distance based clustering with Euclidean distance; k-means algorithm and discretization are the Machine Learning Algorithms which were utilized to fetch weighted risk score for diabetes prediction with the risk factors like age, Body-Mass-Index, waist measurement, personal details, family details, food diet, physical activities, stress and life style. The outputs prove that the formulated work is better than current techniques.

Chetty et al [16] handled PIMA and Liver-disorder databases. Several researchers have formulated the utilization of K-Nearest Neighbour (KNN) algorithm for the prediction of diabetes disease. Certain researchers have also designed a contrast approach by utilizing K-means clustering for the purpose of pre-processing and KNN for the purpose of classification. These study lead to under privilege classification accuracy and prediction.

Another work was refined implying two contrast methods, out of which, one is Fuzzy C-Means (FCM) clustering algorithm using KNN as a classifier and second one is FCM clustering algorithm using fuzzy KNN as a classifier to enhance the classification accuracy. It was proved to yield successful results when correlated with the current algorithms for the data set provided. The second approach yielded a good output than the earlier technique. Classification is performed with the help of ten folds cross-validation technique.

Priyadarshini et al [17] applied the concept of modified Extreme Learning Machine (ELM) to determine the patients being affected by diabetic or not depending on the information provided, which facilitates the clinical people in identifying the diabetic and non – diabetic patients. It characterizes and correlates the application of two famous machine learning techniques: One is the Back Propagation Neural Network (BPNN) and the other is modified ELM which in turn acts as binary classifiers that help in the prediction of diabetics. Both these techniques are implied on similar category of multi class classification datasets and the proposed work tends to extract certain correlative inferences from training and validating outputs. The set of data which were utilized for this work is derived from UCI learning repository.

3. Proposed System

The main objective of this proposal is to introduce assemble Data Mining based Diabetes Disease Prediction System which provides a detailed analysis of diabetics using the database of diabetics patient. The formulated work comprises of two stages such as feature selection ad prediction methods which are made known to maximize the outputs of diabetes disease prediction. Initially Correlation Feature Selection (CFS) is formulated to identify the salient features for the diabetic repository. The identified features are fed into the classifier named Probabilistic Neural Network (PNN) classifier. This is helpful for obtaining immense knowledge resulting in an innovative assumption focusing in intense perceptive and to carry out further analysis in data mining techniques. This section is contains the description of the data set,

the pre-processing step and the feature selection as well as the algorithm used for classification. The formulated experimental processes have been done utilizing the MATLAB environment. The formulated architecture is shown in Figure 1.

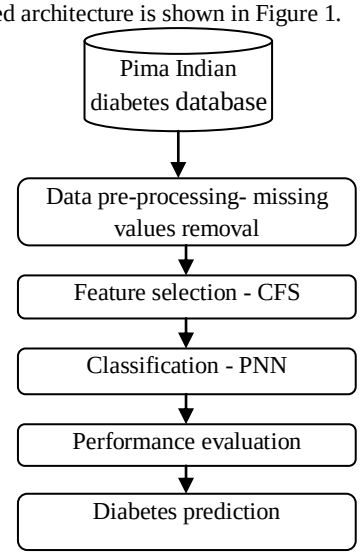

Figure 1: Proposed architecture

Data Pre-Processing

To a maximum extent, the data quality affects the prediction results. This has the sense that pre-processing of data marks a significant criterion in this model [8]. In this formulated work, the identification of exact techniques are used to sharpen the real set of data. Initially a clear analysis made on each attribute's medical indication and its relation to DM. Moreover, identification of number of pregnancies has a minor connectivity with DM [18]. Therefore, conversion of numeric attribute to a nominal attribute is accomplished. The numerical value 0 implies non-pregnant and the numerical value 1implies pregnant. The reduction of complexity is carried out by this process.

Secondly, due to certain flaws and de-regulations, there may be some lacking and unreliable values in the set of data. These meaningless and erroneous values lead to inappropriate experimental outputs. For instance, in the primitive set of data, the diastolic blood pressure values and body mass index may not be 0, which reveals the absence of the real value. To minimize the impact of meaningless values, the means are utilized that are taken from the training set of data in order to replace all missing values.

Feature Selection

Feature selection can be defined as a pre-processing phase used in machine learning which is powerful in dimensionality reduction, elimination of data which is irrelevant, maximizing the accuracy of learning and enhancing the output comprehensibility .A filtration algorithm is used namely Correlation based Feature Selection (CFS) that stands in Pima Indian diabetes feature selection subsets as per correlation oriented heuristic validation operation. The contradiction in the validation operation is toward subsets that consist of diabetes features that are greatly correlated with the class and uncorrelated with each other. Irrelevant diabetes features may be neglected as they will have inferior correlation with the class. Reiterating diabetes features should be removed out because they will have great correlation with one or more of the left-over features. CFS consists of two significant stages; the initial stage is evaluating the feature-feature and feature-class matrix correlations. The next stage is an exploration process which s implied to find out the feature space and obtain the feasible subset. To verify all the probable subsets and identify the finest is exorbitant because of the high feature space. Many heuristic searching techniques are available such as best first that are adequately implied to identify the feature space in meaningful time interval. The correlation CFS oriented heuristic validation operation is defined as follows [19-21].

$$M_S = \frac{kr_{cf}}{\sqrt{k+k(k-1)r_{ff}}} \quad (1)$$

All the variables are standardized in the Pearson's correlation. The heuristic advantage of a diabetes feature subset S is M_s that consist of k features, the mean diabetes feature-class correlation is $r_{cf}$and the average diabetes feature is r_{ff} diabetes feature inter-correlation. The equations' numerator renders an implication of class prediction of a set of diabetes features; and the equations' denominator shows the redundant features within the diabetes features.

Classification

The classification algorithm targets to stabilize a model that can fetch Pima Indian diabetes data items to a selective class depending on the current Pima Indian diabetes. This is used to elicit important Pima Indian diabetes items from the model or it can even be used in the prediction of tendency of Pima Indian diabetes. The majority of the cases, the binary classifier become the dependent variable in the Probabilistic Neural Network (PNN) classifier.

PNN is widely followed for its several merits [22]. When compared to BP network, in terms of speed PNN is many times faster. The optimal output produced by Bayes classification can be easily achieved by PNN when certain conditions are enforced [22]. PNN can approach a Bayes optimal result under certain easily met conditions [22]. Moreover it acts as robustious in terms of noise. It is selected for a basic structure and for training. The ultimate merit of PNN is that the training proves to be easy and immediate [23]. Here the weights are certainly "assigned" and definitely "not trained". The existing weights will not be exchanged but very few contemporary vectors are infused into weight matrices during the training process. Hence it can be helpful in real-time. The momentum of PNN is accelerating as the training and running process can be implemented by matrix manipulation.

The input vector is classified into a particular class by the network as the class has the highest possibility to be optimal. In the proposed work, the PNN consists of three layers which are namely; the Input layer, Radial Basis Layer and the Competitive Layer. The vector span between input vector and row weight vectors belonging to the weight matrix is validated by the Radial Basis Layer. The span area can be measured by Radial Basis Function nonlinearly. Next the Competitive Layer finds out the shortest span area within them, and thus identifies the training pattern nearest to the input pattern depending on their distance.

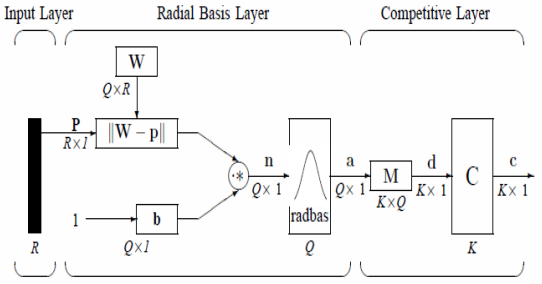

Figure 2: PNN classifier

1) Input Layer: The data vector of the input diabetes is denoted as p which is shown as the black vertical bar in Figure 2. The dimension of this will be R × 1. In the proposed work, R = 10.

2) Radial Basis Layer: In Radial Basis Layer, the data vector of the diabetes, the span area between input diabetes data vector p and the weight diabetes data vector are constructed with each row of weight matrix W are evaluated. The dot product between two

diabetes data vectors is characterized as the data vector of the diabetes span area [23]. Q×R is hypotheses, which is the dimension of W. The i-th element of the distance vector ||W−p|| is produced by the dot product between p and the i-th row of W, where Q×1 is the dimension, as shown in Figure 2. The span area between the diabetes data vector is denoted by the minus symbol, "–". Then, the contradictory diabetes data vector b is merged with ||W− p|| by performing an element-by-element multiplication, that can be shown as "·∗" in Figure 2. The output is denoted as n = ||W− p|| · ∗p. With regard to the centre, the transfer operation in PNN has constructed a distance criterion. In this proposed work, it is defined as

$$radbas(n) = e^{-n^2} \quad (2)$$

Every element of n is given into above equation and yields the relevant of a, the output vector of Radial Basis Layer. The i-thelement of a can be given as

$$a_i = radbas\left(\left||W_i - p|\right|.* b_i\right) \quad (3)$$

Where$W_i$is the vector built with the i-throw of W and b_i is the i-thelement of bias vector b.

3) Certain features of Radial Basis Layer: The i-th element tends to be 1, if the input p is similar to the i-th row of input weight matrix W. A radial basis neuron with a weight vector is nearer to the input vector p generates a numerical value near 1 and its result weights in the competitive layer shall transfer their values to the competitive function. There is a possibility that many elements of 'a' are near to 1 as the patterns o input are close to many training patterns.

4) Competitive Layer: There is no contradiction in the Competitive Layer. In this layer, the vector 'a' is initially multiplied with layer weight matrix M, yielding the resultant vector d. The competitive function shown in Figure 2, denoted as C, yields one corresponding to the greatest element of d, and 0's nowhere. C is the resultant vector of competitive operation. The number of plant which can be classified in C, has the index value 1. It can be utilized as index value to identify the name of this plant scientifically. 64 is the output vectors' dimension, K.

4. Results and Discussion

It is comfortable to study the result of the experiment through a visualized interface using MATLAB. The PNN classifier is analyzed and validated considering the following issues. The Pima Indian Diabetes set of data consists of information about 768 patients (tested positive instances are 268 in number and tested negative instances are 500 in number) which are collected near Phoenix, Arizona and USA. Tested_ positive and tested negative diagnoses the patient to be diabetic or not. Each instance has8 attributes, which all are numeric. These information are about health details and the results from medical examinations. The description of attributes in the set of data is listed as follows, and Table 1 shows few samples collected from the dataset.

- Number of times pregnant (preg)
- Plasma glucose concentration at 2 h in an oral glucose tolerance test (plas)
- Diastolic blood pressure (pres)
- Triceps skin fold thickness (skin)
  2-h serum insulin (insu)
- Body mass index (bmi)
- Diabetes pedigree function (pedi)
- Age (age)
- Class variable (class)

Table 1: Samples of Dataset

Preg	plas	pres	skin	Insu	Mass	pedi	Age	class
6	148	72	35	0	33.6	0.627	50	tested_positive
1	85	66	29	0	26.6	0.351	31	tested_negative
8	183	64	0	0	23.3	0.672	32	tested_positive
1	89	66	23	94	28.1	0.167	21	tested_negative
0	137	40	35	168	43.1	2.288	33	tested_positive
5	116	74	0	0	25.6	0.201	30	tested_negative
3	78	50	32	88	31	0.248	26	tested_positive
10	115	0	0	0	35.3	0.134	29	tested_negative
2	197	70	45	543	30.5	0.158	53	tested_positive
8	125	96	0	0	0	0.232	54	tested_positive

In general, the process of prediction contains four different results called True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The precision is calculated by follow

Precision= TP/ (TP+FP) (4)

The recall, also known as the specificity, is calculated by equation (5).

Recall= TP/ (TP+FN) (5)

F-measure is defined as the mean of precision and recall is calculated by equation (6)

F-measure = 2.P.R/ (P+R) (6)

The confusion matrix displays these four results of BPNN, MLPNN and PNN in Table 2(a), (b) and (c).

Table 2(a): Confusion matrix of BPNN

Class		Actual Classes	Actual Classes
		Yes	No
Predicted Classes	Yes	144	12
	No	29	71
		173	83
			256

Table 2(b): Confusion matrix of MLPNN

Class		Actual Classes	Actual Classes
		Yes	No
Predicted Classes	Yes	156	9
	No	17	74
		173	83
			256

Table 2(c): Confusion matrix of PNN

Class		Actual Classes	Actual Classes
		Yes	No
Predicted Classes	Yes	169	6
	No	4	77
		173	83
			256

The results of BPNN, MLPNN and PNN with precision, recall, f-measure and accuracy are shown in Table 3.

Table 3: Results Comparison with classifiers

	Back propogation Neural netw...	MLP Neural network	Probabilistic Neural Network
Accuracy	83.5938	89.4531	95.3125
Precision	0.8113	0.8740	0.9405
Recall	0.8316	0.9063	0.9559
F- Measure	0.8213	0.8898	0.9481

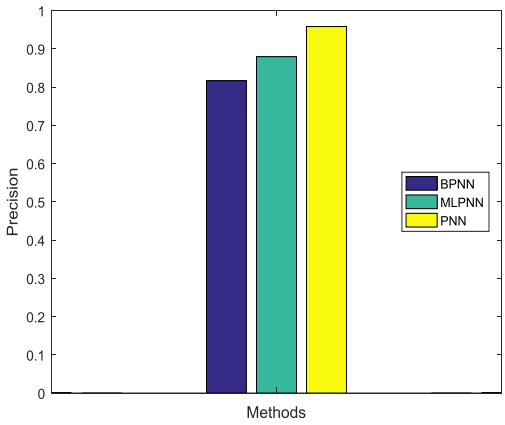

Figure 3: Precision comparison vs. classifiers

In the figure 3 shows the precision comparison results of BPNN, MLPNN and PNN classifiers. The proposed PNN produces higher precision results of 94.05%, whereas other classifiers such as BPNN, and MLPNN produces precision results of 81.13% and 87.40%.

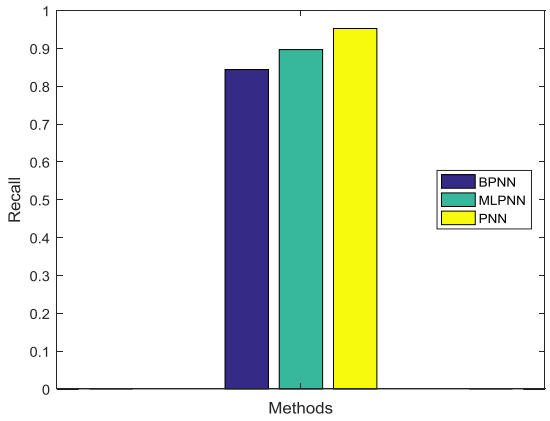

Figure 4: Recall comparison vs. classifiers

In the figure 4 shows the recall comparison results of BPNN, MLPNN and PNN classifiers. The proposed PNN produces higher recall results of 95.59%, whereas other classifiers such as BPNN and MLPNN produces precision results of 83.16% and 90.63%.

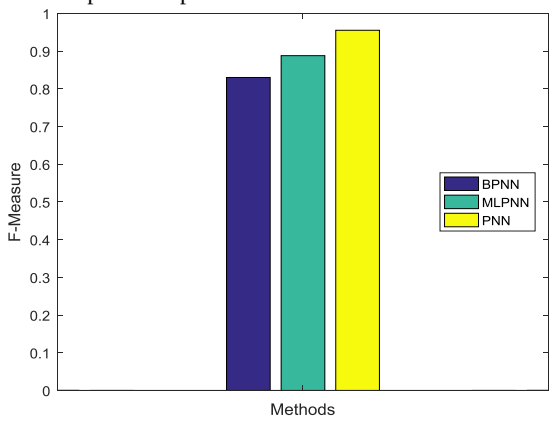

Figure 5: F-measures Comparison vs. classifiers

In the figure 5 shows the F-measure comparison results of BPNN, MLPNN and PNN classifiers. The proposed PNN produces higher F-measure results of 94.81%, whereas other classifiers such as BPNN, and MLPNN produces precision results of 82.13% and 88.98%.

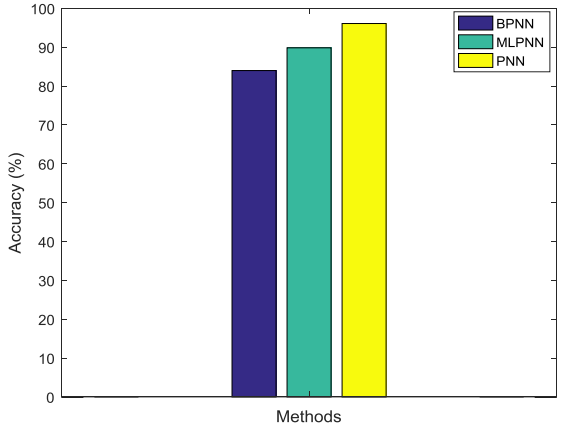

Figure 6: Accuracy Comparison vs. classifiers

In the figure 6 shows the accuracy comparison results of BPNN, MLPNN and PNN classifiers. The proposed PNN produces higher accuracy results of 95.312%, whereas other classifiers such as BPNN, and MLPNN produces precision results of 83.59% and 89.45%.

5. Conclusion and Future Work

Nowadays, the data mining strategies are used to maximize the frequency in the prediction of the probability of disease. These have enormous potential in the field of research. The formulated system uses two phases which are feature selection and prediction methods to maximize the outputs of diabetes disease prediction. Initially Correlation Feature Selection (CFS) is formulated to identify the salient features for the diabetic repository. The identified features are fed into the classifier named Probabilistic Neural Network (PNN) classifier. Pima Indians Diabetes Dataset from the University of California, Irvine (UCI) Machine Learning Database is used for experimentation and the MATLAB environment is used for research. It establishes the experimental data with good quality. [24]The PNN classifier can be applied in the Pima Indian Diabetes Dataset that yields greater accuracy. The outcome of this is a model which can be utilized for the realistic health management of diabetes. In the proposed study, the outputs are validated depending on the precision, recall, f-measure, and accuracy. In the scope for future enhancement, the researchers can collect neighbourhood details and optimization based feature selection methods can be imposed to fetch and fine tune the accuracy and precision details. Few more parameters can be added to the research such as thirst, fatigue, frequency of urination etc for further development and improvement of the research. [25]

References

[1] World Health Organization, Diabetes Program. http://www.who.int/diabetes/en/

[2] Parthiban G & Srivatsa SK, "Applying machine learning methods in diagnosing heart disease for diabetic patients", *International Journal of Applied Information Systems (IJAIS)*, Vol.3, (2012), pp.2249-0868.

[3] Zhiren L, "Machine learning group at the university of Waikato", *Weka*. (2013-12-20), [2015-10-22].

[4] Hina S, Shaikh A & Sattar SA, "Analyzing Diabetes Datasets using Data Mining", *Journal of Basic and Applied Sciences*, Vol.13, (2017), pp.466-471.

[5] Riccardo B & Blaz Z, "Predictive data mining in clinical medicine: current issues and guidelines", *Int J Med Inf*., Vol.77, (2008), pp.81–97.

[6] Gittens M, King R, Gittens C & Als A, "Post-diagnosis management of diabetes through a mobile health consultation application", *IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom)*, (2014), pp.152-157.

[7] http://archive.ics.uci.edu/ml/datasets/PimaþIndiansþDiabetes.
[8] Patil BM, Joshi RC & Toshniwal D, "Hybrid prediction model for type-2 diabetic patients", *Expert systems with applications*, Vol.37, No.12, (2010), pp.8102-8108.
[9] Ahmad A, Mustapha A, Zahadi ED, Masah N & Yahaya NY, "Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for Diabetes Mellitus", *In Digital Information Processing and Communications*, (2011), pp.537-545.
[10] Marcano-Cedeño A, Torres J & Andina D, "A prediction model to diabetes using artificial met plasticity", *International Work-Conference on the Interplay between Natural and Artificial Computation*, (2011), pp.418-425.
[11] Vijayan VV & Anjali C, "Decision support systems for predicting diabetes mellitus A Review", *Global Conference on Communication Technologies (GCCT)*, (2015), pp.98-103.
[12] Wu, H, Yang, S, Huang, Z, He, J & Wang, X, "Type 2 diabetes mellitus prediction model based on data mining", *Informatics in Medicine Unlocked*, (2018), pp.100-107.
[13] Sowjanya K, Singhal A & Choudhary C, "MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices", *IEEE International on Advance Computing Conference (IACC)*, (2015), pp.397-402.
[14] Songthung P & Sripanidkulchai K, "Improving type 2 diabetes mellitus risk prediction using classification", *13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, (2016), pp.1-6.
[15] Chandrakar O & Saini JR, "Development of Indian weighted diabetic risk score (IWDRS) using machine learning techniques for type-2 diabetes", *Proceedings of the 9th Annual ACM India Conference*, (2016), pp.125-128.
[16] Chetty N, Vaisla KS & Patil N, "An improved method for disease prediction using fuzzy approach", *Second International Conference on Advances in Computing and Communication Engineering (ICACCE)*, (2015), pp.568-572.
[17] Priyadarshini R, Dash N & Mishra R, "A Novel approach to predict diabetes mellitus using modified Extreme learning machine", *International Conference on Electronics and Communication Systems (ICECS)*, (2014), pp. 1-5.
[18] Karim M, Orabi1 YMK & Thanaa MR, "Early predictive system for diabetes mellitus disease", *ICDM 2016, LNAI*, (2016), pp.420–427.
[19] Hall M, *Correlation-based feature selection for machine learning*, PhD Thesis, Department of Computer Science, Waikato University, New Zealand, (1999).
[20] Hall M & Smith, L, "Feature Selection for Machine Learning: Comparing a Correlation based Filter Approach to the Wrapper", *Twelfth International Florida Artificial Intelligence Research Society Conference*, (1999), pp.235– 239.
[21] Saeys Y, Inza I & Larranaga P, "A review of feature selection techniques in bioinformatics", *Bioinformatics*, Vol.23, No.19, (2007), pp.2507–2517.
[22] Wu SG, Bao FS, Xu EY, Wang YX, Chang YF & Xiang QL, "A leaf recognition algorithm for plant classification using probabilistic neural network", *IEEE International Symposium on Signal Processing and Information Technology*, (2007), 11-16.
[23] Mishra S, Bhende CN & Panigrahi BK, "Detection and classification of power quality disturbances using S-transform and probabilistic neural network", *IEEE transactions on power delivery*, Vol.23, No.1, (2008), pp.280-287.
[24] G Abilbakieva, M Knissarina, K Adanov, S Seitenova, G Bekeshova (2018). Managerial competence of future specialists of the education system (Preschool education and upbringing) and medicine in the comparative aspect. Opción, Año 33, No. 85. 44-62.
[25] Akhpanov, S. Sabitov, R. Shaykhadenov (2018). Criminal pre-trial proceedings in the Republic of Kazakhstan: Trend of the institutional transformations. Opción, Año 33. 107-125.