Diabetes mellitus prediction using machine learning within the scope of a generic framework

Nidhi Arora[1], Shilpa Srivastava[2], Ritu Agarwal[3],Vandana Mehndiratta[2], Aprna Tripathi[4]
[1]Department of Computer Science, Kalindi College, University of Delhi, Delhi, India
[2]School of Sciences, Christ (Deemed to be University), Delhi NCR, Ghaziabad, India
[3]Department of Information Technology, Raj Kumar Goel Institute of Technology, Ghaziabad, India
[4]Department of Data Science and Engineering, Manipal University Jaipur, Jaipur, India

Article Info

Article history:

Received Sep 19, 2023
Revised Oct 2, 2023
Accepted Oct 8, 2023

Keywords:

Chronic diseases
Classification
Diabetes
Machine learning
Predictive modelling

ABSTRACT

Artificial intelligence (AI) based automated disease prediction has recently taken a significant place in the field of health informatics. However, due to unavailability of real time large scale medical data, the dynamic learning of prediction models remains principally subsided. This paper, therefore proposes a dynamic predictive modelling framework for chronic diseases prediction in real-time. The framework premise suggests creation of a centralized patient-indexed medical database to dynamically train machine learning (ML) models and predict risk levels of chronic diseases in real time. In this study, comprehensive empirical evaluations to train seven state-of-the-art ML models for diabetes risk prediction are performed in context of phase 2 of the suggested framework. The selected optimal model can then be dynamically applied to predict diabetes in phase 3 of the framework. Various metrics such as accuracy, precision, Recall, F1-score and receiver operating characteristic (ROC) curve are employed for evaluating performances of the trained models. Parameter tunings using different type of kernels, different number of neighbors and estimators are rigorously performed in order to create a suggestive literature for healthcare prediction ecosystem. Comparative analysis indicates high prediction accuracies on diabetes test data records for neural network and support vector machine (SVM) models as compared to other applied models.

Corresponding Author:

Nidhi Arora
Department of Computer Science, Kalindi College, University of Delhi
Delhi, 110008 India
Email: nidhiarora@kalindi.du.ac.in

1. INTRODUCTION

Chronic diseases are quite a concern not only for patients but also for health professionals. As per the definition by MedicineNet [1]: a chronic disease is a disease which may endure beyond three months (as per the definition of the U.S. National Center for Health Statistics) and may last up to a life time; having significant impacts on the regular functioning of many vital parameters of suffering individuals. As per World Health Organization (WHO) report-2016 [2], chronic diseases are not transmittable from one person to another person; last for long duration and progress very slowly. Main chronic diseases that most of the human race suffers are diabetes, heart diseases, cancer and arthritis to name a few [3]. As initially, many of the chronic diseases do not show any noticeable symptoms, medical professionals face quite a challenge in their direct diagnosis and primarily establish their interpretations on varied regularly testable patient's parameters. It's quite later, when such diseases start to impact vital organs behaviors and attract attention of patients as well as medical practitioners [4].

The practice of using automated technology based disease prediction methodologies, therefore have taken their place in medical informatics to assist in diagnostics of such diseases even in absence of visible identifiers [5]. Medical informatics technologies are primarily artificial intelligence (AI) driven and work towards making computers self-trainable to develop ample intelligence and suggest diagnosis for the patient at quite early stages. Such systems were called expert systems and had been in usage since the boom of AI techniques [6]. Machine learning (ML), a branch of AI is now been used widely for predictions in many spheres of applications. Expert systems were data dependent and used to gain suggestive decisions based on focused data sets, whereas ML work on continuously evolving learning models for real time predictions with high accuracy. ML is being successfully utilized in many aspects of medical field such as robotic assistive operative environments, automated sample collections and predictive modelling for predicting diseases [7]–[10]. Medical datasets are now being evolved by saving records of patients on hospital portals. Medical datasets may be collection of prescriptions (textual data), medical test records [11], images of various scans [12], historical parental records, and video or audio recordings.

Many researchers have applied and tested varied ML approaches on specific disease data sets available. Diabetes mellitus prediction using three ML classification algorithms named support vector machine (SVM), random forest (RF) and neural networks were analyzed using PIMA Indian and Laboratory of the Medical City Hospital (LMCH) diabetes datasets by Olisah *et al.* [13]. Authors presented detailed discussion on preprocessing advantages on the datasets to improve accuracy in the prediction of diabetes diseases. Febrian *et al.* [14] applied K-nearest neighbor (KNN) and Naïve Byes classification algorithms to predict diabetes. As per the comparative analysis shown in their paper, Naïve Byes algorithm outperformed KNN algorithm in accurately predicting diabetes for new patients. In another paper by Krishnamoorthi *et al.* [15], diabetes prediction using decision tree (DT)-based RF and SVM learning models was done. Sonia *et al.* [16], applied neural network having multiple layers with no-prop algorithm for diabetes prediction and evaluated their results using sensitivity, specificity, accuracy and a confusion matrix metrics. Many other researchers [17]–[19], also worked on diabetes data sets snippets available on various data platforms and applied specific ML techniques for developing predictive models at various level of accuracies.

For another chronic diseases, such as heart diseases prediction, many authors applied classification based ML techniques such as Shah *et al.* [20] predicted heart diseases using heart disease patient's data sets of Cleveland database on UCI repository. The authors compared the prediction accuracies of Naïve Bayes, DT, KNN, and RF algorithms and reported KNN algorithm as having highest precession in predicting heart diseases. In another paper by Kavitha *et al.* [21], a hybrid of RF and DT ML algorithm was applied on cleveland dataset of heart patients. Many other authors [22], [23] applied varied ML techniques such as SVM, neural networks and reported prediction accuracies at different levels. Not only heart and diabetes chronic diseases, researchers have worked with different diseases datasets snippets to predict parkinson disease [24], liver disease [25], chronic kidney disease [26] and Alzheimer's disease [27] to name a few.

Research gap and main contribution: all the above conducted studies use opensource different diseases medical datasets to present training results of specific ML algotithms with varied accuracies. Medical data, as described above, is primarily patient dependent, country dependent and sometime localized to specific hospitals as well. Therefore, for efficient application of ML in medical field, it is quite necessary to link the medical record keeping systems across hospitals, to apply AI using dynamically evolving real data sets. Therefore, a common framework is required to enable healthcare systems to develop highly accurately trained, tested and continuously evolving automated assistive prediction utilities to benefit all stakeholders. Also, it is observed that presently applied ML algorithms in literature do not present extensive results on hyper-parameter tuning with a single diabetes dataset, availaibility of which as a single source can benefit at large to the stakeholders. This paper, therefore proposes a common framework design for predicting risk for chronic diseases based on various data recorded in the database. The cloud-based framework design could act as a prototype for implementation and development to be an essential part in every healthcare utility. Further, this paper also evaluates seven predictive ML algorithms for the risk prediction focused on diabetes dataset. The empirical results along with hyperparameter tuining results of three algorithms like different type of kernels, neighbors and estimators in classification algorithms namely: SVM, KNN, and ensemble RF respectively. The indepth experimentation on diabetes prediction can give further direction to medical stakeholders by proposing the best parametric and divisive combinations to be utilized for predicting medical risks. Paper organization: next section 2 presents method section which presents design of proposed framework as well as details of ML methods to be applied for training, data availability and its description, implementation details with parameters settings applied. In section 3 discusses results obtained with a detailed comparative analysis along with hyperpaprameter tuinings experimental results, followed by conclusion and references.

2. METHOD

The proposed framework is presented in Figure 1. The framework’s prototype is divided into 4 main phases. The framework’s prototype is described below.

Phase 1: phase 1 focusses on creation of a centralized patient unique identifier (PUID) indexed centralized database on a cloud platform. Individuals go for regular tests at different designated test centers in any state/country. These tests contain values for different vital parameters, which then combined in a single table stored on cloud-based storage platforms. The storage may be localized or distributed. This centralized database of PUID indexed records grow dynamically, in order to create a big data sets of medical testing records. The centralized table is initialized with pre classified historical data of patients with multiple class attribute columns representing diseases D1, D2…Dn. The tables get regularly appended with unclassified records of tests of individuals.

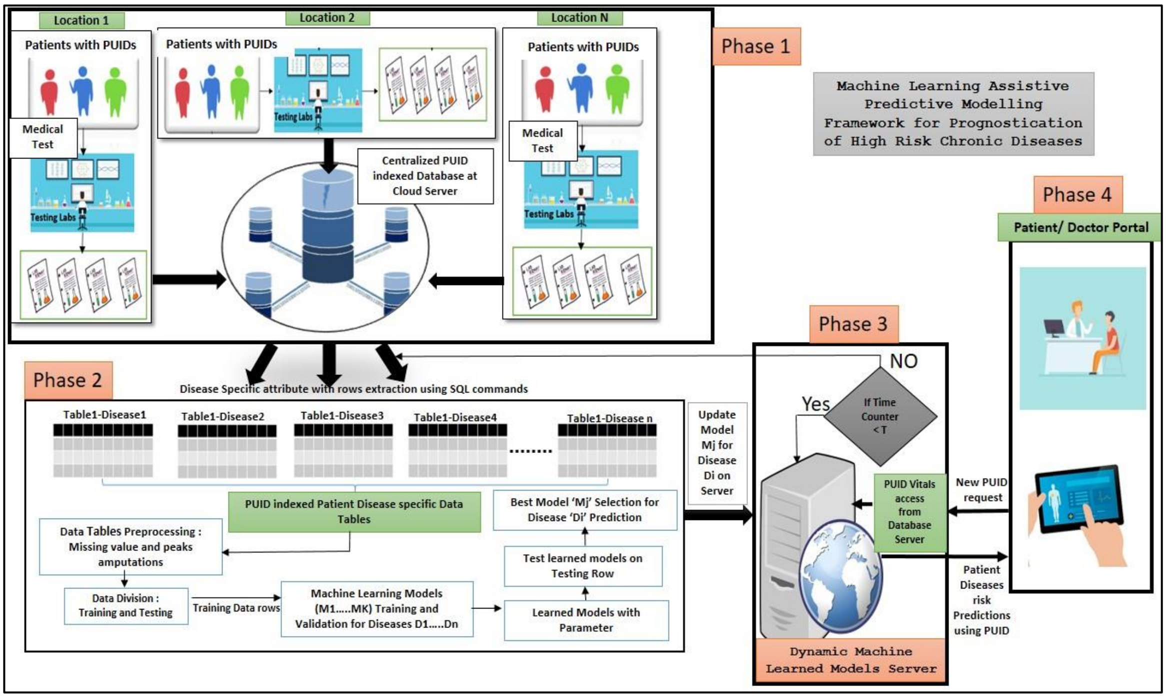

Location 1
Location 2
Location N
Patients with PUIDs
Patients with PUIDs
Patients with PUIDs
Medical Test
Medical Test
Testing Labs
Testing Labs
Testing Labs
Centralized PUID indexed Database at Cloud Server
Phase 1
Machine Learning Assistive Predictive Modelling Framework for Prognostication of High Risk Chronic Diseases
Phase 4
Patient/ Doctor Portal
Phase 3
NO
Yes
If Time Counter <T
Phase 2
Disease Specific attribute with rows extraction using SQL commands
Table1-Disease1
Table1-Disease2
Table1-Disease3
Table1-Disease4
Table1-Disease n
Update Model Mj for Disease Di on Server
PUID indexed Patient Disease specific Data Tables
Best Model ‘Mj’ Selection for Disease ‘Di’ Prediction
Data Tables Preprocessing : Missing value and peaks amputations
Test learned models on Testing Row
Data Division : Training and Testing
Training Data rows
Machine Learning Models (M1.....MK) Training and Validation for Diseases D1.....Dn
Learned Models with Parameter
PUID Vitals access from Database Server
New PUID request
Patient Diseases risk Predictions using PUID
Dynamic Machine Learned Models Server

Figure 1. ML based modelling framework for prognostication of chronic diseases

Phase 2: phase 2 builds disease relevant sub-tables by extracting disease specific parameters from the centralized tables. These sub tables are preprocessed for missing values and peak values amputations; divided into train and test set. The training data subset is used to train ‘k’ ML models M1…Mk. The learned models undergo testing on test subsets and their accuracies are quantified using different metrics such that the best model Mj is chosen to predict disease Di: 1<j<k and 1<i<n.

Phase 3: the selected model and its parameters are updated on the server containing learned models. The trained ML models stored on the server are regularly upgraded after a time interval ‘T’. After some fixed time, interval ‘T’; phase 2 is iteratively repeated with new and upgraded set of classified records extracted from centralized database table, the models are again trained and best model selected with newly learned model parameters are updated in the ML model server.

Phase 4: once the new unclassified test record is added in the centralized table with a unique PUID, it acts as a test data and passed through the learned model. The output predictions are prompted to patient and doctor’s portal. For diseases classified to positive categories with high accuracies, medical practitioners recommend thorough evaluation and may confirm the disease class for that PUID in central database server. Such records are then added to training data set. The framework architecture is focused on cloud storage usage for maintaining the centralized and dynamic data base of patient’s records from varied places. The records are regularly updated and classified based on which ML models are also regularly updated in order to always predict diseases with high accuracy based on new training examples added dynamically.

2.1. Applied ML classification models

Following ML classification models are being trained on diabetes data in this paper:

- Naïve Bayes: Naïve Bayes classification model is based on Bayes theorem of conditional probability [28]. The model finds out the posterior ($Pr[H|E]$) and conditional probabilities ($Pr[E|H]$) of events event E for hypothesis H in the situation, and predicts output of any event based on the maximum likelihood (posterior probability) of that event based on conditional probabilities. The model uses (1).

$$Pr[H|E] = \frac{Pr[E|H]\,Pr[H]}{Pr[E]} \tag{1}$$

- Logistic regression: logistic regression is a supervised ML classification model [29]. Developed by applying a logistic or sigmoidal function ‘g’ on the hypothesis equation of linear regression, so as to get the predicted output in the range [0,1]. The output values>=0.5 are quantified to positive class and values <0.5 are quantified to negative class. The hypothesis equation of logistic regression is given by (2) and (3) respectively.

$$h_\theta(x) = g(\theta^T x) \tag{2}$$

$$g(z) = \frac{1}{1+e^{-z}} \tag{3}$$

- Neural network: neural network ML classification models are based on the design of network of human neurons [30]. There are many layers through which input values are fed and forwarded to another layers using activation functions and weight combinations. The weights in the layers are learned using many iterations of feed forward and back propagation algorithms. An example of visualization of a neural network model is shown in Figure 2.

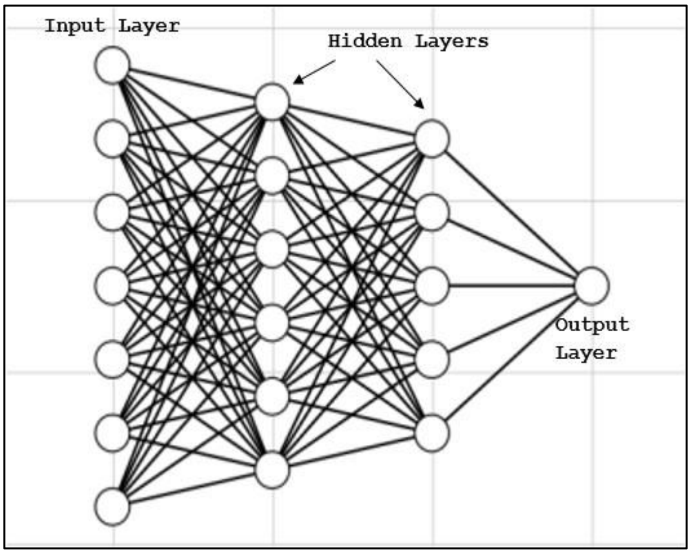


Figure 2. A neural network model with input, hidden, output layer and weights

- DT: DT models are used for classification as well as regression in ML and are developed by application of simple decision rules on the input variables of the data set using metrics like entropy and information gain [31]. The DT algorithm builds a hierarchical decisive structure of nodes, where each node represents an input variable and paths through that node are based on decision rules formed based on training data sets.
- SVM: SVM are primarily used for classification problems. The input data points are treated as vectors plotted in n-dimensional plane. The algorithm develops a line/plane equation to separate the vector points of different classes as far as possible from that line [32]. The distance is quantified by identifying the support vector points in positive and negative class planes.

- KNN: K-NN classification algorithm classify new objects based on the maximum class match with the classification of its K number of nearest neighboring objects based on a distance metric [33]. This algorithm is a lazy learner as it does not develop a model, rather it calculates distances for each new object with all training data set objects, and pics top K number of nearest neighbors out of them.
- RF ensemble classifier: RF ensemble classifier is another quite efficient classification model generation technique, which combines the output of various DTs to conclude a class for a new instance [34]. This methodology develops different DTs using different sample sets, and use them all to classify a new instance.

2.2. Metrics used for prediction accuracy evaluation

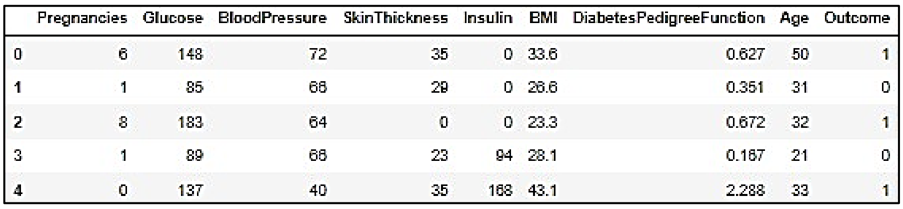
The problem to classify chronic diseases is a classification problem. The mathematical ML models predict outcomes of training and testing data set in discrete classes. For quantifying the efficiency of a developed ML model, confusion matrix is used. Confusion matrix gives the count of correctly predicted outputs (true positives, true negatives) and the incorrectly predicted outputs (false positives and false negatives) of a ML model. Using these counts, various efficiency indication metrics have been developed such as precision, recall, accuracy, F1-score and receiver operating characteristic (ROC) curve.

- Precision: precision value for the predictions done by any ML model specifies the number of correct true positive predictions w.r.t total positive predictions.
- Recall: recall value for the predictions done by any ML model specifies the number of true positive predictions w.r.t total expected positive predictions. Recall is also called as true positive rate (TPR) or sensitivity. False positive rates (FPR) can be found in the same manner.
- Accuracy: accuracy metric value signifies the fraction of correct predictions with respect to total predictions.
- F1-score: F1-score is the harmonic mean of precision and recall. F1-score gives better indication for a data set which has a class imbalance.
- ROC curve: one more useful graphical methodology compare efficiency of different ML models is ROC curves [35]. ROC curves are the graphical depictions between true positive rate and false positive rate of a chosen ML model. The plot shows higher area under curve (AUC) values between 0 and 1, with values of AUC=0.0 showing 100% wrong predictions and AUC=1.1 showing 100% correct predictions.

2.3. Applied experimental settings and dataset description

This paper shows the results of seven ML algorithms on PIMA Indian diabetes data set downloaded from Kaggle (https://www.kaggle.com/uciml/pima-indians-diabetes-database) available via a CC0: public domain license [36]. The description of the data set is given below. The implementation for training and testing ML models is done using Python 3.0 on Anaconda Jupyter. Results are the best of 10 iterations of learning cycles with randomized training (90%) and testing (10%) data sets.

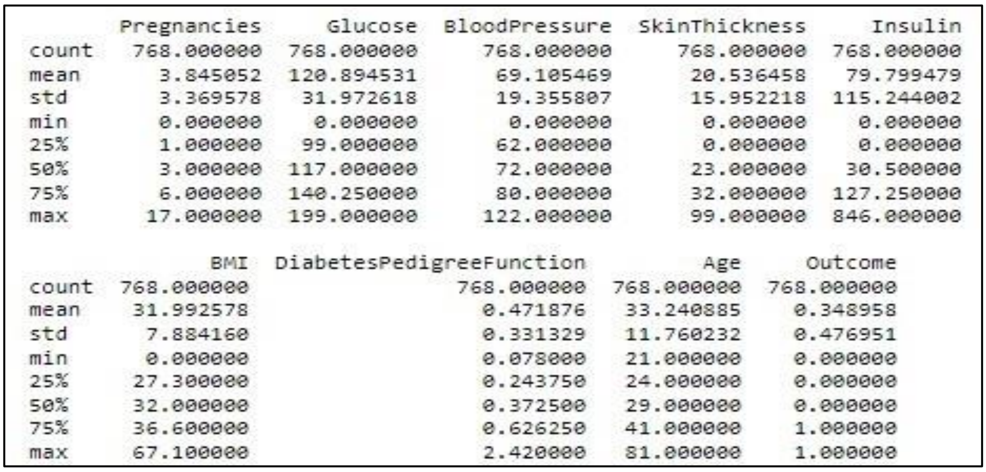
Diabetes data set: the dataset consists of data of 768 females all aged 21 years and above of the PIMA Indian heritage. The data was collected for 9 parameters for all the female patients. A snippet of top 5 rows of the data is shown in Figure 3.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 3. Snippet of diabetes data set

The description of 9 parameters are as follows: pregnancies store the number of pregnancies; glucose stores glucose levels in blood; blood pressure stores the blood pressure measurement; skin thickness stores the thickness of the skin; insulin to express the insulin levels in blood; BMI to store the body mass index; diabetes pedigree function stores the diabetes percentage; age stores the age; outcome stores the final result 1/0 if detected with diabetes or not respectively. The statistical distribution of data in all 9 columns is given in Figure 4 with maximum, minimum, mean and standard deviation of each parameter column. Since the data appears to at different ranges, need to normalize or standardize the data is there.



```
       Pregnancies     Glucose  BloodPressure  SkinThickness     Insulin
count   768.000000  768.000000     768.000000     768.000000  768.000000
mean      3.845052  120.894531      69.105469      20.536458   79.799479
std       3.369578   31.972618      19.355807      15.952218  115.244002
min       0.000000    0.000000       0.000000       0.000000    0.000000
25%       1.000000   99.000000      62.000000       0.000000    0.000000
50%       3.000000  117.000000      72.000000      23.000000   30.500000
75%       6.000000  140.250000      80.000000      32.000000  127.250000
max      17.000000  199.000000     122.000000      99.000000  846.000000

              BMI  DiabetesPedigreeFunction         Age     Outcome
count  768.000000                768.000000  768.000000  768.000000
mean    31.992578                  0.471876   33.240885    0.348958
std      7.884160                  0.331329   11.760232    0.476951
min      0.000000                  0.078000   21.000000    0.000000
25%     27.300000                  0.243750   24.000000    0.000000
50%     32.000000                  0.372500   29.000000    0.000000
75%     36.600000                  0.626250   41.000000    1.000000
max     67.100000                  2.420000   81.000000    1.000000
```

Figure 4. Description of data in all 9 columns

As it is evident from data visualization that the data in all columns are having different ranges. Such data require normalization before effective application of ML algorithms. Therefore, diabetes data set is normalized using min-max normalization to rescale data in range between 0 and 1 using (4) for each ith value of jth column X_{ij} to obtain a normalized value X'_{ij} such that $1 \leq i \leq n \; and \; 1 \leq j \leq c$, where n is the number of rows in data set and c are the number of columns in the data set.

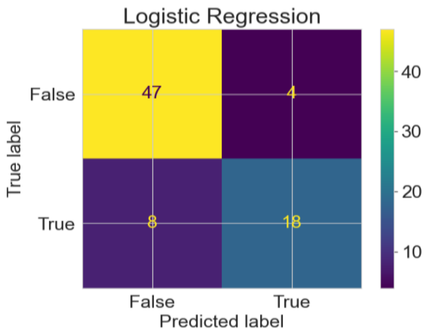
$$X'_{ij} = \frac{X_{ij} - X_{j\_min}}{Xj_{\_max} - X_{j\_min}} \qquad (4)$$

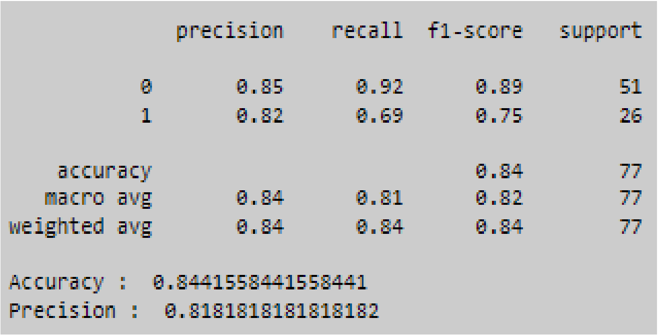
Here, X_{j_max} and X_{j_min} are the maximum and the minimum values of the jth columns respectively. Python MinMaxScaler utility to normalize the data is used in the current implementation. Diabetes data set, after preprocessing, training and test set divisions, is used in training and testing of 7 different ML models namely: logistic regression, neural network, SVM, Naïve Bayes, DT, KNN, and ensemble RF.

3. RESULTS AND DISCUSSION

The detailed empirical comparative results of all ML models are shown here using various accuracy level indicating metrics such as precision, recall, accuracy and F1-score. The results shown here are the best of 10 iterations of learning cycles with randomization done to generate training and testing data sets at the division ratio of 90:10 respectively with implementation in Python 3.0 on Anaconda Jupyter.

– Logistic regression: the classification report of logistic regression based trained model for testing data is shown in Figure 5 depicting precision, recall, F1-score and accuracy values. Logistic regression trained model could achieve an accuracy of 84% and precesion of 82% with a weighted average of F1-score as 84%.

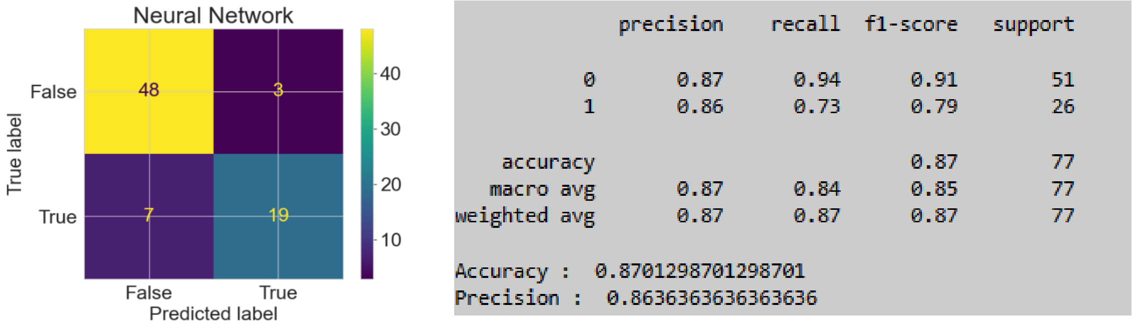




```
              precision    recall  f1-score   support

           0       0.85      0.92      0.89        51
           1       0.82      0.69      0.75        26

    accuracy                           0.84        77
   macro avg       0.84      0.81      0.82        77
weighted avg       0.84      0.84      0.84        77

Accuracy :  0.8441558441558441
Precision :  0.8181818181818182
```

Figure 5. Confusion matrix and classification report for logistic regression

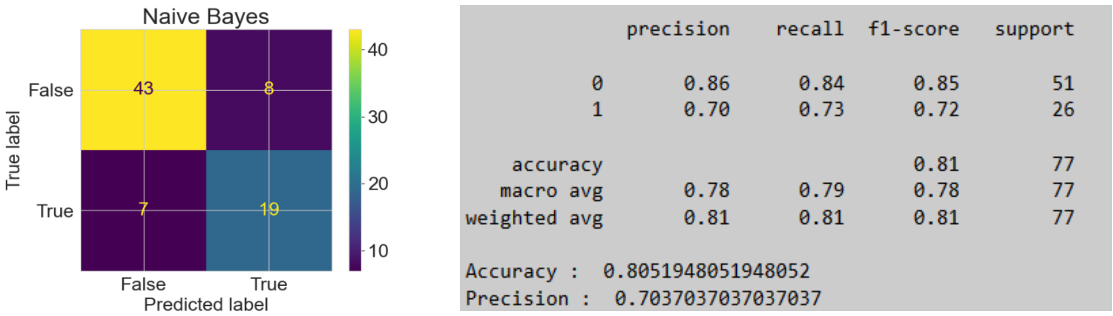
– Neural network: neural network model was trained using 2,000 iterations of back propagation algorithm with a default hidden layer size of 100 nodes; ReLu activation function of multilayer perceptron classifier (MLP classifier) using sklearn neural_network module in python. The model achieved a good 87% accuracy on test data with a precesion of 86% and F1-score 87% as shown in Figure 6.



```
              precision    recall  f1-score   support

           0       0.87      0.94      0.91        51
           1       0.86      0.73      0.79        26

    accuracy                           0.87        77
   macro avg       0.87      0.84      0.85        77
weighted avg       0.87      0.87      0.87        77

Accuracy :  0.8701298701298701
Precision :  0.8636363636363636
```

Figure 6. Confusion matrix and classification report for neural network

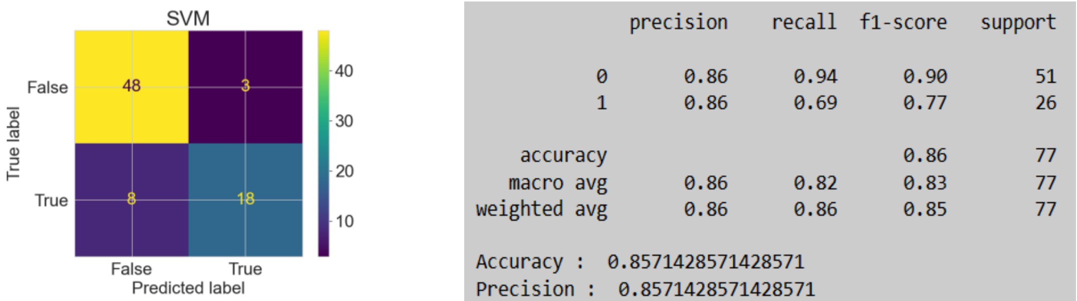
– Naïve Bayes: Naïve Bayes classifier trained model's accuracy however turned out to be 81%; precesion of 70% with a F1-score of 81%. The results are not promising. The confusion matrix and classification report are shown in Figure 7.



```
              precision    recall  f1-score   support

           0       0.86      0.84      0.85        51
           1       0.70      0.73      0.72        26

    accuracy                           0.81        77
   macro avg       0.78      0.79      0.78        77
weighted avg       0.81      0.81      0.81        77

Accuracy :  0.8051948051948052
Precision :  0.7037037037037037
```

Figure 7. Confusion matrix and classification report for Naïve Bayes

– SVM: SVM classifier was trained using different kernel functions to identify the best kernel function on diabetes data. Kernel functions used are: sigmoid, linear, polynomial and radial basis function (RBF). Best accuracy achieved using linear kernel function is 86% as shown in the confusion matrix and classification report for test data classification in Figure 8. Variations in the accuracies of SVM at different kernel functions on test data is shown in Figure 9. Poly kernel function also reported 86% accuracy, however to simplify complexity in medical data prediction domain, we employed linear kernel function.



```
              precision    recall  f1-score   support

           0       0.86      0.94      0.90        51
           1       0.86      0.69      0.77        26

    accuracy                           0.86        77
   macro avg       0.86      0.82      0.83        77
weighted avg       0.86      0.86      0.85        77

Accuracy :  0.8571428571428571
Precision :  0.8571428571428571
```

Figure 8. Confusion matrix and classification report for SVM

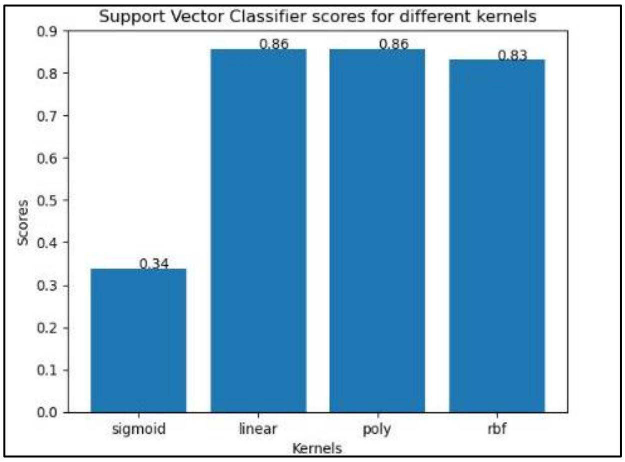
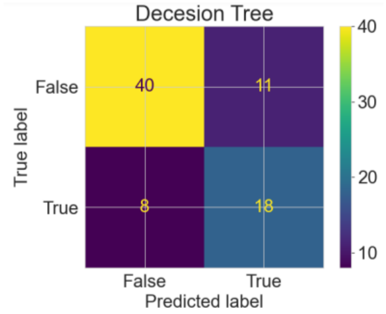


Figure 9. Accuracy scores for SVM at different kernel functions

– DT: the results of DT classification based generated model are shown in Figure 10. It is evident that this trained model could only achieve accuracy of 75% on the test data with an F1-score of 76%.



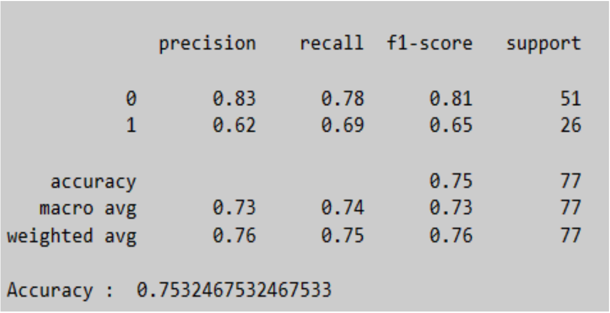

```
              precision    recall  f1-score   support

           0       0.83      0.78      0.81        51
           1       0.62      0.69      0.65        26

    accuracy                           0.75        77
   macro avg       0.73      0.74      0.73        77
weighted avg       0.76      0.75      0.76        77

Accuracy :  0.7532467532467533
```

Figure 10. Confusion matrix and classification report for DT

– KNN: KNN classification algorithm is used to train the model using different values of K (number of neighbors). The accuracy scores achieved at different K values is shown in Figure 11. The best scores achieved at K values 7 and 8. The detailed classification report at K value 8 is shown in Figure 12. The model could only obtain an accuracy of 79%, precision 77% and F1-score of 78% at k value 8.

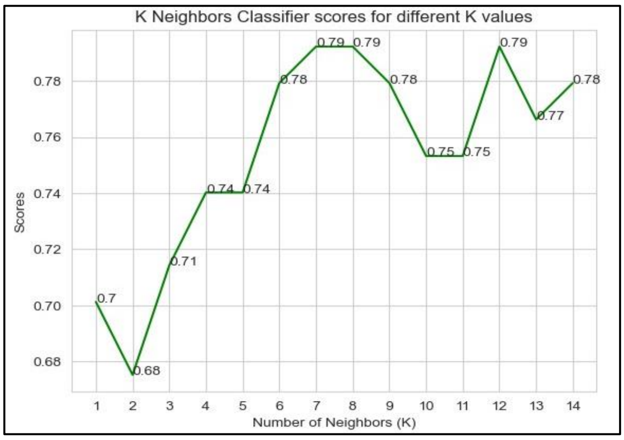
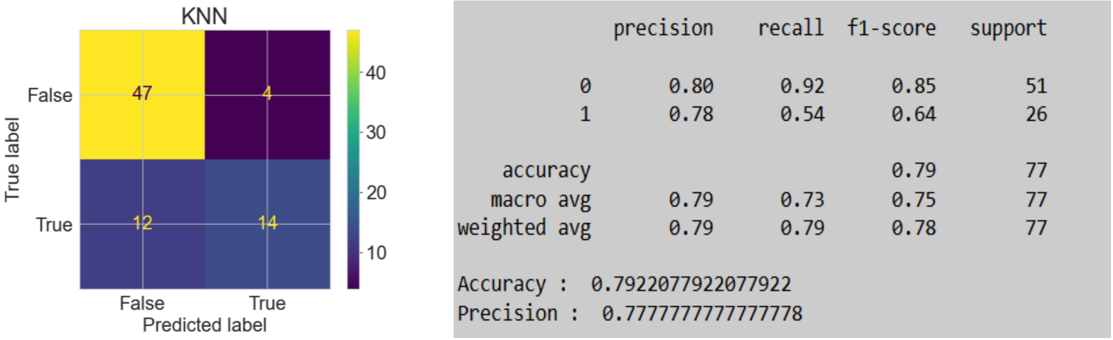


Figure 11. Accuracy scores for KNN using different number of neighbors



```
              precision    recall  f1-score   support

           0       0.80      0.92      0.85        51
           1       0.78      0.54      0.64        26

    accuracy                           0.79        77
   macro avg       0.79      0.73      0.75        77
weighted avg       0.79      0.79      0.78        77

Accuracy :  0.7922077922077922
Precision :  0.7777777777777778
```

Figure 12. Confusion matrix and classification report for KNN classifier

- RF ensemble classifier: RF ensemble classifier is tested using different number of estimators ranging from 10 to 500. The best result of accuracy is achieved in the range 300 to 500, as shown in Figure 13. confusion matrix and detailed classification report at 300 number of estimators is shown in Figure 14. accuracy of 81% with precision of 73% having F1-score of 82% could only be achieved using this model on test data.

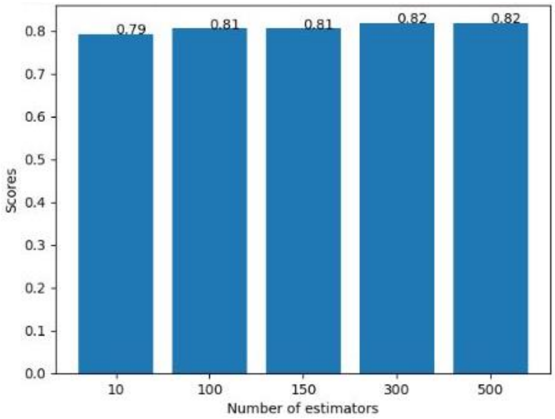
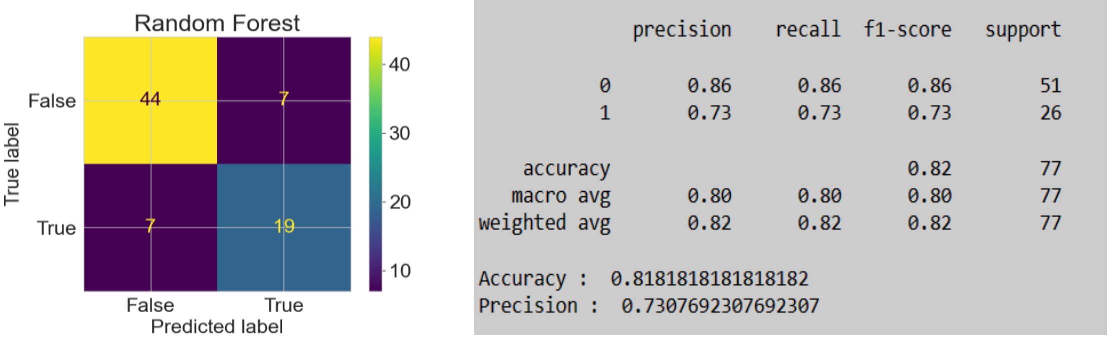


Figure 13. RF classifier scores for different number of estimators



```
              precision    recall  f1-score   support

           0       0.86      0.86      0.86        51
           1       0.73      0.73      0.73        26

    accuracy                           0.82        77
   macro avg       0.80      0.80      0.80        77
weighted avg       0.82      0.82      0.82        77

Accuracy :  0.8181818181818182
Precision :  0.7307692307692307
```

Figure 14. Confusion matrix and classification report for RF classifier

3.1. ROC curve: graphical evaluation metric

ROC curves are plotted for the results obtained for all the 7 learned models at best hyperparameter settings as suggested in above discussion. The plotted ROC curve is shown in Figure 15. Neural network learned model and SVM models show a good bent towards upper left corner, which depicts high accuracies for both the algorithms. The same is supported by their higher values of AUC i.e., 0.817 and 0.836 as compared to all other trained models as mentioned along with the curves in Figure 15. Hence, it is further evident that both neural network and SVM model with linear kernel function can be successfully applied to predict diabetes millitus chronic disease with high accuracy. Further accuracy can be improved with availaibility of bigger dataset, for which it is suggested to follow the design of suggested framework and develop a cloud based medical data collection and prediction model accessible in real time to all stakeholders via their personal handhel mobile applications.

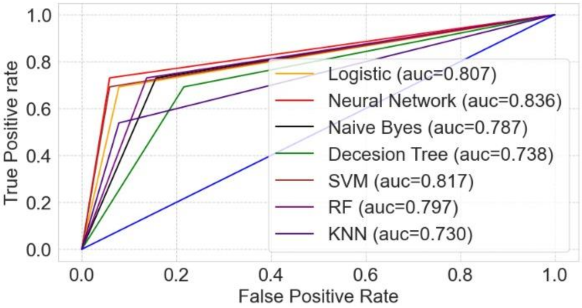


Figure 15. ROC curve for 7 classifiers learned models on diabetes data

4. CONCLUSION

This paper proposed a novel ML assistive modelling framework for prognostication of high risk diseases in real time. The proposed framework is cloud based and dynamically evolve learned models using updated training data of patient's medical records. The framework align its task in four phases and aims as developing a common patient indexed database on a cloud platform at its top most layer. The middle layers aims to train models in regular intervals of time in view of continuously updating training data. The final layer is end user interface, which provide indicative signals to stakeholders for possibility of chronic diseases. The paper further presented a detailed empirical evaluation of seven ML based classification algorithms to train models for the prediction of a highly significant chronic disease diabetes. It was observed that with increase in data sets, variations in models parameters and proportions of training and testing data, model show different level of prediction accuracies. It was observed that out of all seven models, for the current snippet of data set. Neural network model and SVM model showed highest prediction accuracies, and therefor they are highly reliable. It would interesting to evaluate all algorithms on dynamic datasets with different parametric variations in real time as a future work. Evaluation of neural network further is required with different level of hidden layers as future work.

REFERENCES

[1] "Definition of chronic disease," MedicineNet , 2016, [Online]. Available: http://www.medicinenet.com/script/main/art.asp?articlekey=33490. (accessed Sep. 01, 2023).

[2] "Noncommunicable diseases," WHO, 2016, [Online]. Available: http://www.who.int/topics/noncommunicable_diseases/en/. (accessed Sep. 01, 2023).

[3] S. Bernell and S. W. Howard, "Use your words carefully: What is a chronic disease?," *Frontiers in Public Health*, vol. 4, Aug. 2016, doi: 10.3389/fpubh.2016.00159.

[4] M. S. Fragala, D. Shiffman, and C. E. Birse, "Population health screenings for the prevention of chronic disease progression," *The American Journal of Managed*, vol. 25, no. 11, pp. 548–553, 2019.

[5] A. M. Bauer, S. M. Thielke, W. Katon, J. Unützer, and P. Areán, "Aligning health information technologies with effective service delivery models to improve chronic disease care," *Preventive Medicine*, vol. 66, pp. 167–172, Sep. 2014, doi: 10.1016/j.ypmed.2014.06.017.

[6] J. Singla, D. Grover, and A. Bhandari, "Medical expert systems for diagnosis of various diseases," *International Journal of Computer Applications*, vol. 93, no. 7, pp. 36–43, May 2014, doi: 10.5120/16230-5717.

[7] D. Dahiwade, G. Patle, and E. Meshram, "Designing disease prediction model using machine learning approach," in *Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019*, Mar. 2019, pp. 1211–1215, doi: 10.1109/ICCMC.2019.8819782.

[8] K. S. Nugroho, A. Y. Sukmadewa, A. Vidianto, and W. F. Mahmudy, "Effective predictive modelling for coronary artery diseases using support vector machine," *International Journal of Artificial Intelligence (IJAI)*, vol. 11, no. 1, pp. 345–355, Mar. 2022, doi: 10.11591/ijai.v11.i1.pp345-355.

[9] G. L. A. Kumari, P. Padmaja, and J. G. Suma, "A novel method for prediction of diabetes mellitus using deep convolutional neural network and long short-term memory," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 26, no. 1, pp. 404–413, Apr. 2022, doi: 10.11591/ijeecs.v26.i1.pp404-413.

[10] A. Selwal and I. Raoof, "A Multi-layer perceptron based intelligent thyroid disease prediction system," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 17, no. 1, pp. 524–532, Jan. 2020, doi: 10.11591/ijeecs.v17.i1.pp524-532.

[11] M. M. Rahman and D. N. Davis, "Machine learning-based missing value imputation method for clinical datasets," in *Lecture Notes in Electrical Engineering*, vol. 229 LNEE, Springer Netherlands, 2013, pp. 245–257.

[12] M. D. Kohli, R. M. Summers, and J. R. Geis, “Medical image data and datasets in the era of machine learning—whitepaper from the 2016 C-MIMI meeting dataset session,” *Journal of Digital Imaging*, vol. 30, no. 4, pp. 392–399, May 2017, doi: 10.1007/s10278-017-9976-3.

[13] C. C. Olisah, L. Smith, and M. Smith, “Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective,” *Computer Methods and Programs in Biomedicine*, vol. 220, p. 106773, Jun. 2022, doi: 10.1016/j.cmpb.2022.106773.

[14] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, “Diabetes prediction using supervised machine learning,” *Procedia Computer Science*, vol. 216, pp. 21–30, 2022, doi: 10.1016/j.procs.2022.12.107.

[15] R. Krishnamoorthi *et al.*, “A novel diabetes healthcare disease prediction framework using machine learning techniques,” *Journal of Healthcare Engineering*, vol. 2022, pp. 1–10, Jan. 2022, doi: 10.1155/2022/1684017.

[16] J. J. Sonia, P. Jayachandran, A. Q. Md, S. Mohan, A. K. Sivaraman, and K. F. Tee, “Machine-learning-based diabetes mellitus risk prediction using multi-layer neural network no-prop algorithm,” *Diagnostics*, vol. 13, no. 4, p. 723, Feb. 2023, doi: 10.3390/diagnostics13040723.

[17] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, “Likelihood prediction of diabetes at early stage using data mining techniques,” in *Advances in Intelligent Systems and Computing*, vol. 992, Springer Singapore, 2020, pp. 113–125.

[18] P. Sonar and K. J. Malini, “Diabetes prediction using different machine learning approaches,” in *Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019*, Mar. 2019, pp. 367–371, doi: 10.1109/ICCMC.2019.8819841.

[19] M. M. Bukhari, B. F. Alkhamees, S. Hussain, A. Gumaei, A. Assiri, and S. S. Ullah, “An improved artificial neural network model for effective diabetes prediction,” *Complexity*, vol. 2021, pp. 1–10, Apr. 2021, doi: 10.1155/2021/5525271.

[20] D. Shah, S. Patel, and S. K. Bharti, “Heart disease prediction using machine learning techniques,” *SN Computer Science*, vol. 1, no. 6, Oct. 2020, doi: 10.1007/s42979-020-00365-y.

[21] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, “Heart disease prediction using hybrid machine learning model,” in *Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021*, Jan. 2021, pp. 1329–1333, doi: 10.1109/ICICT50816.2021.9358597.

[22] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, “Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison,” *Computers in Biology and Medicine*, vol. 136, p. 104672, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104672.

[23] V. Sharma, S. Yadav, and M. Gupta, “Heart disease prediction using machine learning techniques,” in *Proceedings - IEEE 2020 2nd International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2020*, Dec. 2020, pp. 177–181, doi: 10.1109/ICACCCN51052.2020.9362842.

[24] B. S. Tiwari, A. K. Sahani, A. K. Koulibaly, P. F. Nobili, M. Pagani, and K. Tatsch, “A survey of machine learning based approaches for Parkinson disease prediction,” *International Journal of Computer Science and Information Technologies*, vol. 6, no. 2, pp. 1648–1655, 2015.

[25] A. S. Rahman, F. J. M. Shamrat, Z. Tasnim, J. Roy, and S. A. Hossain, “A comparative study on liver disease prediction using supervised machine learning algorithms,” *International Journal of Scientific and Technology Research*, vol. 8, no. 11, pp. 419–422, 2019.

[26] I. U. Ekanayake and D. Herath, “Chronic kidney disease prediction using machine learning methods,” in *MERCon 2020 - 6th International Multidisciplinary Moratuwa Engineering Research Conference, Proceedings*, Jul. 2020, pp. 260–265, doi: 10.1109/MERCon50084.2020.9185249.

[27] C. Kavitha, V. Mani, S. R. Srividhya, O. I. Khalaf, and C. A. T. Romero, “Early-stage alzheimer’s disease prediction using machine learning models,” *Frontiers in Public Health*, vol. 10, Mar. 2022, doi: 10.3389/fpubh.2022.853294.

[28] K. M. Leung, “Naive bayesian classifier,” *Polytechnic University Department of Computer Science/Finance and Risk Engineering*, pp. 123–156, 2017.

[29] E. Bisong, “Logistic regression. Building machine learning and deep learning models on google gloud platform: a comprehensive guide for beginners,” *CA: Apress*, pp. 59–64, 2019.

[30] S. M. Weiss and I. Kapouleas, “An empirical comparison of pattern recognition, neural nets, and machine learning classification methods,” *In IJCAI*, vol. 28, pp. 781–787, 1989.

[31] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, “An introduction to decision tree modeling,” *Journal of Chemometrics*, vol. 18, no. 6, pp. 275–285, Jun. 2004, doi: 10.1002/cem.873.

[32] D. Meyer and F. T. Wien, “Support vector machines,” *R News*, vol. 1, no. 3, pp. 23–26, 2001.

[33] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, “KNN model-based approach in classification,” in *In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE*, 2003, pp. 986–996.

[34] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001.

[35] A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997, doi: 10.1016/S0031-3203(96)00142-2.

[36] “PIMA Indian diabetes dataset,” Kaggle, 2016, [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. (accessed Sep. 01, 2023).

BIOGRAPHIES OF AUTHORS

Nidhi Arora is an Associate Professor in the Department of Computer Science, Kalindi College, University of Delhi, India. With a teaching experience of 23 years and a Ph.D. in Computer Science from University of Delhi, her primary research areas are social networks, nature inspired computing and ML. She has published many peers reviewed research articles in international journals of repute, book chapters, and has also presented many research papers in various international conferences of ACM, Springer and IEEE. Dr. Nidhi Arora has delivered many talks on latest research topics such as “data sciences”, “e-content development” and “deep learning” to name a few. She has been part of many administrative assignments, committees, on the board of technical programme committee, and acted as a reviewer to in many international journals. She can be contacted at email: nidhiarora@kalindi.du.ac.in.

Dr. Shilpa Srivastava is an Associate Professor and Chairperson of the Committee 'AI Policy recommendations' at Christ University Delhi NCR. She possesses a Ph.D. Computer Science from Uttarakhand Technical University and her areas of interest include application of soft computing in medical domain, theory of computation, algorithms, and ehealth services. Currently she is guiding three Ph.D. Scholars and published many peers reviewed research articles in international journals of repute, one patent, book chapters, and has also presented many research papers in various international conferences of ACM, Springer and IEEE. She has delivered talks on the latest research topics such as "ML", "mobile learning", bring your own device", to name a few. She has also worked as CO-PI in a research project in collaboration with IIT Roorkee and Liverpool Hope University UK. A passionate teacher and an avid researcher with 21 years of teaching experience, has been part of many administrative assignments, committees, curriculum development Incharge, on the board of Programme Committee of reputed international conferences, and acted as a reviewer to review articles in many springer and IEEE journals. She can be contacted at email: shilpa.srivastava2015@gmail.com.

Ms. Ritu Agarwal is an Associate Professor and Head of Department of Information Technology at Raj Kumar Goel Institute of Technology, Ghaziabad. She has over 20 years of rich experience in teaching, research and administration. She has expertise in the field of computer science. She published more than 10 papers in international/national journals and conferences including SCI and Scopus. She is reviewer for several national/international conferences of high repute. She is IBM certified in web development with Java. She has good knowledge of outcome-based education and part of NBA process. She had taken visiting lectures for programming in multiple institutions. She can be contacted at email: agarwalritu7@gmail.com.

Ms. Vandana Mehndiratta has done MCA from Ch Charan Singh University in 2001. She is M.Tech. (CSE) from Sam Higgin Bottom Institute of Agriculture, Technology and Sciences, Allahabad in 2006. She has 17 years of experience in the field of computer science. Currently she is working as an Assistant Professor in CHRIST (deemed to be University) and also a Ph.D. scholar form the same university. Her research area is machine learning. She can be contacted at email: vandana.mehndiratta@christuniversity.in.

Dr. Aprna Tripathi is an Assistant Professor in the Department of Data Science and Engineering, Manipal University Jaipur, Jaipur. She received her bachelor's degree in sciences from Kanpur University, Master's in Computer Applications from HBTI, Kanpur, M.Tech. from Banasthali University, Rajasthan and Ph.D. from NIT Allahabad, Prayagraj. With over 15 years of teaching and research experience, her scholarly contributions can be found in prestigious national and international journals and conferences, including those recognized by SCI and Scopus. Furthermore, she actively contributes as a reviewer for prestigious academic journals. Her areas of specialization include software engineering, software testing, data visualization, and data structures and algorithms. Notably, she has authored a book titled "component-based systems: estimating efforts using soft computing techniques". In addition, she holds membership in the Association for Computing Machinery (ACM) and IEEE. She can be contacted at email: aprna.tripathi@jaipur.manipal.edu.