Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 05 Issue 01
2023

Research Article
https://doi.org/10.56979/501/2023

Pre-Diabetic Diagnosis from Habitual and Medical Features using Ensemble Classification

Zeeshan Aamir[1], and Iqbal Murtza[1*]

[1]Department of Creative Technologies, Faculty of Computing & AI, Air University, Islamabad, 44230, Pakistan.
[*]Corresponding Author: Iqbal Murtza. Email: iqbal.murtza@mail.au.edu.pk.

Received: March 01, 2023 **Accepted:** June 03, 2023 **Published:** June 05, 2023.

Abstract: Diabetes is the fastest growing metabolic disorder. It has become a serious global health emergency. This disease if not diagnosed can increase the risk of serious life-threatening diseases like cardiac arrest, brain bleeds also known as brain strokes, kidney damage and many more. In this paper, we considered a challenging problem of" Pre-Diabetes Detection" using different deep learning and machine learning algorithms. The prime target of our study is to detect diabetes at an early age so that we can help the patient to properly diagnose this disease and better take care of their health. This is important because if we can correctly predict and then provide diagnosis of Prediabetes there is a 58% chance that the person can go back to his healthy life. So, considering the power of deep learning and machine learning we intend to exploit/explore parameters/features-space using suitable classification techniques. For this we employed several techniques i.e., LSTMs, MLP and ANN for deep learning and also Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), KNN and Gaussian Naive Bayes (GNB) for machine learning respectively. We also performed majority voting from all Machine Learning and deep Learning models and made an ensemble model from them to evaluate performance. Among these techniques our recommended ensemble model for deep learning and LSTMs outperformed because they were very effective than other standard machine learning and deep learning model. To validate the proposed technique, we used a standard publicly available dataset. We have then applied majority voting and made an ensemble classifier from that and evaluated the results.

Keywords: pre-diabetic diagnosis, ensemble classification, computer aided disease diagnosis, machine learning, majority voting.

1. Introduction

Diabetes is a long-term and incurable disease that happens when the pancreas cannot generate desired amount of insulin and also when our body cannot utilize the produced insulin efficiently. One of the most crucial and important organs in our body is the pancreas, which can also affect the metabolism of our fat, protein, and sugar for daily energy. It is medically stated that concentration of the blood glucose, also called blood sugar, will be large if there is small or no insulin present. Urine containing an excessive amount of sugar will be excreted, and this condition is known as diabetes mellitus [1].

This dangerous illness has no known treatment. It is believed that both hereditary and environmental components play their vital roles in development and progression of diabetes. A person's race, past family details of diabetes, age, obesity, poor intake of food, inactivity, and active and passive smoking both are some few risk factors related to diabetes. Additionally, in case of failing to treat diabetes correctly at early stage has been linked to the emergence of other chronic illnesses such renal disease. Additionally, the patient has a great level of risk because of other non-communicable pre-existing diseases because they are highly likely to get infected by infectious diseases like COVID-19 and are highly vulnerable to them [2].

An overall of about 8.5% of people who were about the age of 18 and bigger had diabetes in the year 2014. Diabetes directly contributed to about 1.5 million fatalities in 2019, and those under the age of 70 accounted for 48% of all deaths. Diabetes was also a vital reason to an additional 460 thousand kidney disease fatalities, & a high concentration of our blood glucose is cause for about 20% of cardiovascular deaths. In 2014, 8.5 percent of those who were 18 years old and older had diabetes [3].

Almost around 1.6 million individuals each year perished as a result of this disease, and it has been ranked as the seventh main reason of early mortality. According to statistical research, there are 463 million diabetics globally as of 2019, and that amount is anticipated to elevate to 578 million by 2030 and also 700 million till 2045. As a result, it is expected that the number of diabetes sufferers will rise rapidly, rising by 25% in the year 2030 and also by 51% around 2045 [4].

Presently, a doctor must manually make an early diagnosis and predict diabetes based on their training, expertise, and observations. Despite the fact that the health-care sector already gathers a considerably huge amount of data, it occasionally hides inherited underlying patterns. Because some parameters may remain concealed so they can have severe effects on the observations and outcomes, these manual judgments can therefore be extremely deceptive and harmful, especially for early diagnosis. So, in order to ensure greater accuracy, improved techniques for early and automated diagnosis are urgently needed. Additionally, since machine learning and currently famous deep learning approaches have demonstrated excellent results in uncovering underlying patterns, they are being applied to a variety of complicated problems to produce effective and reliable results with dependable accuracy [4].

So, if we can efficiently predict diabetes at an early stage then the patient can start its cure early. Studies have shown that there is 58% chance of recovery from prediabetes if we change our diet, do more physical activity and reduce some weight [5]. In an Indian study, it was demonstrated that in those with impaired glucose tolerance (a kind of prediabetes), lifestyle changes and the drug metformin were both beneficial in avoiding the onset of type-2 diabetes. Dietary modifications, physical activities, and behavioral techniques were all part of the lifestyle modification program [6].

Also, in a paper, the overall rate of the type 2 diabetes was 43% lower in lifestyle interceding group than in the controlled group, according to this study, which monitored participants in the Finnish Diabetes Intervention Study for a total of fifteen years. The lifestyle intervention included dietary modifications, increased physical activity, and losing weight. It's important to note that these are only a few of the research that have been done on this subject. However, these studies offer solid proof that persons with prediabetes can delay or avoid the start of type 2 diabetes by making lifestyle modifications [7]. Here we will use different deep learning & machine learning models to predict diabetes. Many classifiers are used here to classify and predict whether or not the person is pre-diabetic or not.

Diabetes is very dangerous for human health. In the past, the classification and diagnosis of diabetes were obtained by utilizing different diversity of techniques and methods. Deep learning and Machine Learning (ML) methods are getting lot of prominence and also gaining a lot of enthusiasm in the field of modern medicine. The capability to handle huge number of variables and also producing high and efficient results. Random Forest, Decision tree, KNN, LSTMs, Multilayered Perceptron, and Artificial Neural networks are just a few examples of the many techniques used in deep learning and machine learning.

Several studies have investigated the application of machine learning (ML) techniques for early detection of the diabetes using numerical datasets. For instance, in a study[1] a deep learning (DL) based algorithm was used to predict diabetes using a combination of demographic and clinical features, such as age, sex, BMI, and blood glucose levels. The study shows that deep learning algorithm gave a high accuracy of 86.29 for DCNN and 83.65% for LSTMs demonstrating its potential as a tool for diabetes detection.

In another paper, MLP architecture outperforms QML by following parameters. The results show that OR & OR + MV substantially improve the developed model's evaluation compared to the real raw data by 0.10 (16.43%), 0.29 (47.64%), 0.06 (8.16%), 0.18 (23.51%), 0.05 (7.14%), and 0.25 (35.71%), 0.08(12.37%), 0.28 (42.33%), 0.06 (7.54%), and 0.15 (18.86%) for precision, accuracy, recall, and F1 score, respectively. They also show significant improvements in specific[4].

Additionally, the random forest model efficiently outplayed both Naive Bayes & also J48 decision trees on the entire Pima Indian Diabetes dataset in terms of accuracy metrics (79.57%), specificity (75.00%), AUC (86.24%), precision (89.40%) and f-score (85.17%), with the J48 having the highest sensitivity (88.43%)

from the other three. Samples of classes 0 and 1 show a variation is presumably the cause of significant disparity between the sensitivity and the specificity [2].

Similarly, in a study by [8], a machine learning based algorithm was developed to predict diabetes using a combination of clinical and biochemical features, such as age, sex, FPG levels, and triglyceride levels. The study shows us that the machine learning algorithms gave a high accuracy of 87.26% with LSTMs outperforming traditional diagnostic tests.

Table 1. Summary of some relevant researches

Year/Author	Year	Methodology	Results
Chang et al. [2]	2022	RF	79.57%
		Naive Bayes	79.13%
		J48	75.65%
Gupta et al. [4]	2022	QML	84%
Alex et al. [1]	2022	LSTM	80.2%
		DBN	76.3%
		DNN L-BFGS	79.1%
		DCNN	84.2%
Butt et al. [8]	2021	MLP	86.06%
		LSTM	87.26%
Kaur et al. [9]	2020	SVM	89%
Kannadasan et al. [10]	2019	DNN using stacked auto-encoders	86.26%

2. Materials and Methods

Machine learning has advanced a lot thanks to deep learning. The feature extraction and classification processes are not directly performed by deep learning. All of these tasks are implicitly carried out by the deep learning hidden layers without the involvement of an impartial researcher. An overview of deep learning is provided below [11]. Firstly, the dataset needs to be preprocessed and then normalized then afterwards we will apply different machine Learning and deep learning models like Decision Tree, Logistic Regression, Ada boost, ANN, MLP and LSTMs to predict pre-diabetes. The outcomes will then be compared with other cutting-edge methods.

2.1. Dataset

Of the most common persistent illnesses in the US, diabetes affects multiple millions every year & has significant monetary impact on the economy. A person with diabetes struggles to maintain a normal blood sugar level, which can shorten their expected lifespan and negatively impact their quality of life. Diabetes is a dangerous chronic disorder. During digestion, sugars from a variety of foods are transformed and then enter the circulatory system. Figure 1 shows the heatmap of dataset used.

As a result, the pancreas is stimulated to release insulin. Insulin helps make it feasible for cells in the body to use bloodstream's carbohydrates as a source of fuel. In diabetes, the body. either does not create enough need level of insulin or does not use the insulin which is created as well as needed. The dataset is available on Kaggle on the following link" https://www. kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset". The Behavioral Risk Factor Surveillance System is a yearly telephone survey on health-related issues that the CDC performs. (BRFSS). Each year, around 0.4 million Americans take part in this annual survey, which gets data on risky behaviors, long termed health condition, and use of protective treatments. It had been done annually since the year 1984. For this, a excel file containing the data set from

Kaggle of the entire year of 2015 was used. There are about 330 features in the original dataset & respondents provided about 441,455 responses. Those qualities were either generated variables which were based on specific participant replies or the direct participant questioning. You can get diabetes 012 health indications BRFSS2015 excel file, a cleaned data with 253,680 survey results from the CDC's BRFSS2015. The targeted variable of Diabetes 012 is made up of three classes. 0 means you don't have diabetes, 1 means you do, and 2 means you do. There is class imbalance in this dataset. There are about 21 variable features in this dataset [12]. The correlation matrix of these features is as following.

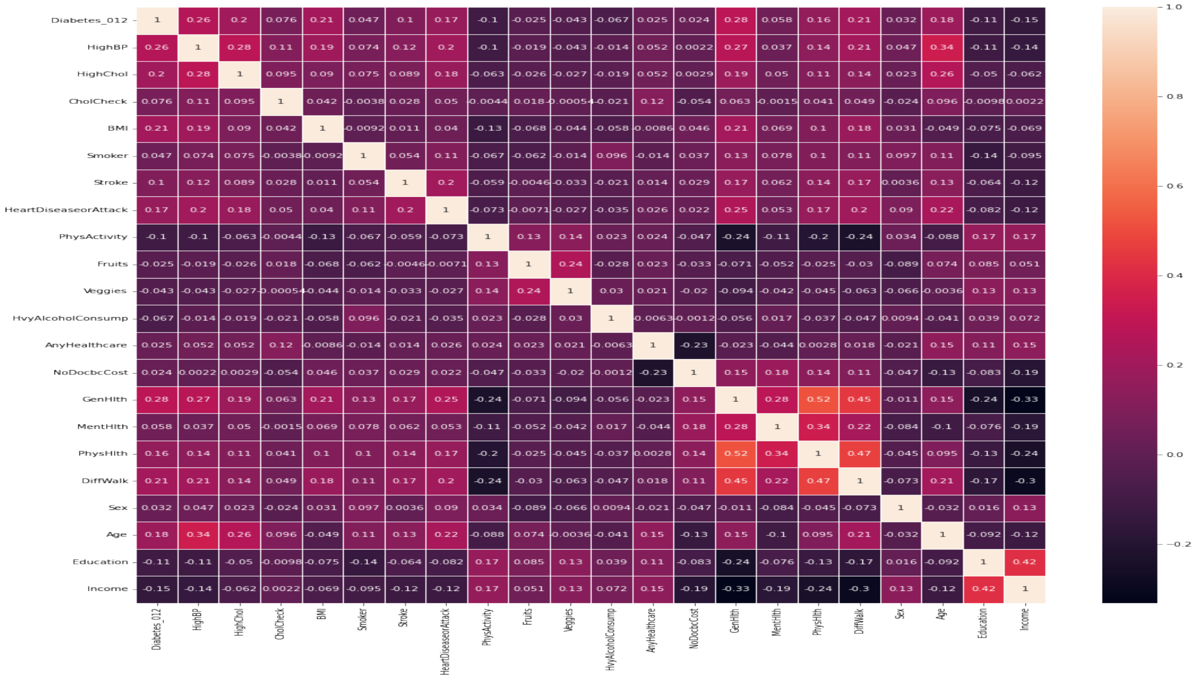

Figure 1. Heat map of the dataset which shows correlation of every feature with the other.

Figure 2 shows the correlation of my dataset. It shows that the bigger the graph is the more is important. So based on this we can drops those features which have very less relation or has a very small impact on diabetes in this case we will drop the features in preprocessing like Fruits, Veggies, AnyhlyCare, NoDocBcCost etc.

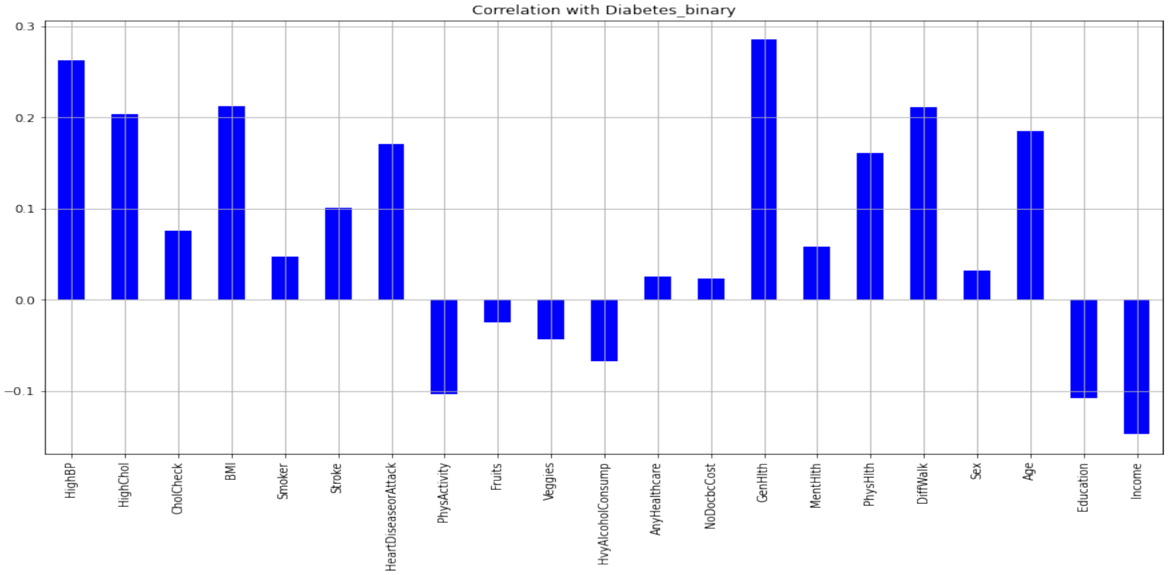

Figure 2. Correlation of diabetes dataset showing importance of every feature separately.

Table 2. Feature description of dataset

Feature	Description
HighBP	A condition when pressure levels is high.
HighChol	Elevated blood levels of cholesterol leading to increased risk of heart disease and other cardiovascular problems.
CholChk	It is a medical assessment or test performed to measure the levels of cholesterol in the blood.
BMI	A measurement that calculates an individual's body fatness based on their height and weight.
Smoker	A person who regularly engages in the act of smoking tobacco or other substances.
Stroke	Brain damage due to interrupted blood flow, causing sudden neurological impairments.
Heart Disease / Attack	Number of heart diseases or attacks
PhysActivity	Does the patient perform physical activity or not
Veggies	Does the patient eat healthy veggies or not
HvyAlcoholConsump	Heavy drinkers
AnyHealthcare	possess health insurance or other prepaid plans like an HMO, or possess some other form of health coverage, etc. 0 = no, 1 = yes
NoDocbcCost	Was there been an occasion in the previous 12 months where you need medical attention but were unable to do so due to cost? 0 = no 1 = yes
GenHlth	General Health: scale 1-5 1 =excellent, 5 = poor
MentHlth	Mental health
PhysHlth	Physical health, which includes physical illness and injury, during the last 30
DiffWalk	difficulty in walking or using stairs? 1 = yes, 0 =no
Sex	1 = male, 0 = female
Age	Age of the patient
Education	Education level scale 1-6
Income	Income scale, scale 1-8, 1 = less than $10,000, 8 = $75,000 or greater
Diabetes	2 = diabetes, 0 = no diabetes, 1 = prediabetes

2.2. Models

Using deep learning and machine learning algorithms is how we diagnose pre-diabetes. We employ a set of health indicator data that consists of demographic data, clinical measurements, and lab test results. The dataset is preprocessed to eliminate missing values, perform feature scaling, and normalize the data. To determine who is at risk of acquiring diabetes, we employ a two-stage categorization method. We utilize a binary classification system in the initial stage to discriminate between those with and without pre-diabetes. Using metrics like accuracy, recall, precision, and F1-score, we test a various amount of machine learning methods, which included logistic regression, Gaussian NB, random forest, KNN and decision trees etc. We compared how well they perform. In the second stage, we employ a deep learning methodology to forecast the likelihood that patients identified as pre-diabetic in the first stage will develop diabetes.

We employ a neural network with Relu activation functions and many hidden layers. We experiment with several designs and hyperparameters, including batch size, learning rate, learning rate per layer, and number of layers. Using metrics like the area below the curve (AUC), precision, recall, and confusion matrix, we assess the model's performance. We train and evaluate our models using k-fold cross-validation to prevent overfitting and enhance generalization performance. In order to minimize the dataset's dimensionality and find the most useful features, we also do feature selection. We test various feature selection techniques and assess their effectiveness, including correlation-based feature selection, chi-squared test, and recursive feature elimination. Using the Python programming language and well-known libraries like TensorFlow, Scikit-Learn, and Kera's, we put our strategy into practice. To speed up training, the models are trained on a system with a GPU.

Overall, our method combines the strengths of deep learning and machine learning to enable precise and early identification of diabetes and pre-diabetes, which can aid in prompt interventions and stop or delay the onset of problems from diabetes.

2.2.1. Logistic Regression

A numerical method called as logistic regression, widely used for examining the connection between one or many independent variables & a dependent variable (typically binary or categorical) (also called as predictors or explanatory variables). It is widely applied where the goal is to forecast that class and labels for a fresh point of data based on its features. The dependent variable in the logistic regression framework is assumed to have a binomial distribution.

2.2.2. Decision Tree

A decision supporting tool known as a decision tree employs a framework in the form of a tree or models which describe decisions and all of their results, including usefulness, resources, and outcomes from chances events. It is frequently used in data analysis and statistics to assist in determining the approach most likely to succeed in achieving a target. A decision tree's root node, which symbolizes the decision or beginning point, is followed by branches, which stand for potential decisions or actions. In machine learning, they are used for predictions and classification tasks. Here, they can aid in locating the most crucial characteristics or variables in a data and produce guidelines or standards for categorization [13].

2.2.3. K Nearest Neighbour

A simple yet effective and parameter less way is the classification approach is the K Nearest Neighbor (KNN). It is like a lazy learning / instance-based learning where the model stores all of the training dataset's instances rather than being explicitly trained. The algorithm finds the k-nearest neighbor for the new input data in the training data and labels the input according to the dominant class amongst k-nearest neighbors when a new input is received [14]. The value of 'k' in the K nearest neighbor shows the number of nearest neighbors that are taken into account when deciding how to classify anything. Whereas in the regression mode, the largest class among the k nearest neighbors are assigned into the forecast class for new input is simple to use and analyze, because it makes no assumptions about the distribution of the underlying data.

2.2.4. Random Forest

Random Forest is a very famous machine learning method which specialized for regression, classification & other purposes. RF is a type of method of ensemble learning which combines various decision trees to develop a much more accurate & dependable model. Using Random Forest, a collection of decision trees is built, with each tree training on the subset of the data and features that are available. Because of this, each tree is unique from the others and complements the strength & weakness of the others [15]. In order to create a Random Forest model, you must first: The important steps involve Cleaning, Preprocessing, Random Subset Selection, Feature Selection, Creating Decision Tree and then voting.

2.2.5. Gaussian Naïve Bayes

A popular probabilistic machine learning technique which is suitable for classification tasks is Gaussian Naive Bayes (GNB). When the characteristics are continuous and normally distributed, this Naive Bayes method variant is especially helpful. The likelihood function of each class, given the value of input features is calculated by the GNB method. The computation of the joint probability is made easier by the presumption that input features are all independent. All things considered, Gaussian Naive Bayes (GNB) is popular and powerful technique for classification tasks, especially when the inputs are continuous and regularly distributed [16].

2.2.6. Artificial Neural Network

Another categorization method known as ANN. It is a deep learning algorithm which yields much more accurate and precise results (accuracy, precision etc.) than the current techniques. It is an efficient mathematical representation that draws inspiration from the structure and operation of organic neurons. Artificial neurons have mesh connectivity for functional connectivity, and each neuron has an equal weight [17]. The learning rate, which modifies weights at every step and is in charge of the model's fundamental learning properties, is another important ANN parameter. It must be chosen to power of 10, specifically to the powers of 0.001, 0.01, & 0.1. In the model, the learning rates are set to 0.1.Sigmoid and Relu activation functions will also be applied.

2.2.7. Long Short-Term Memory (Lstm)

Long Short-Term Memory, also known as LSTMs, is like a recurrent neural network (RNN) structure which is frequently employed in tasks involving sequence prediction and natural language processing (NLP). Since their first introduction by Hochreiter & Schmidhuber in 1997, LSTMs have gained popularity for a variety of tasks, including sentiment analysis, machine translation, and speech recognition. The vanishing gradient issue, which standard RNNs experience and makes it challenging to learn long-term dependencies, is addressed by LSTMs. Three gates the input gates and then there is the forget gates & also one last output gates which control the memory cell. LSTMs. Also, the weights complexity is also reduced while update to O (1). Just like back propagation through time which is also an advantage [18].

2.2.8. Ensemble Model From Majority Voting Of Machine And Deep Learning Models

Predictions from various different models are combined to get a final forecast using the potent machine learning and deep learning technique known as ensemble modelling. A common technique for ensemble modelling is voting by majority, where every model in the ensemble votes, and the result is decided by a majority vote. In this instance, an ensemble model was developed using majority voting on a set of deep learning and machine learning (ML) models.

Here in my paper, several algorithms, namely Random Forest, Gaussian Naive Bayes, Decision Tree, Logistic Regression, and K-Nearest Neighbors, were used in the ensemble's machine learning component. These models were each trained using the data set and the majority voting method was used to integrate their predictions. This strategy is predicated on the notion that each model has strengths and flaws of its own, and that by integrating their forecasts, we may take use of their combined knowledge.

The LSTM (Long Short-Term Memory) networks, multi-layer perception (MLP), and artificial neural networks (ANN) were used in the ensemble's DL component, in contrast. These DL models have a reputation for being able to understand intricate connections and the complex patterns in data, which makes them very useful for variety of tasks. To create the final prediction of the ensemble, the predictions of these models were also pooled using majority voting.

In summary, the ensemble model produced in this situation by majority voting among the DL models, particularly the LSTM, produced the best accuracy [19-23]. It is crucial to remember that the ensemble model, which consists of all the ML models, also accomplished excellent accuracy [24-28]. This demonstrates the potency of ensemble modelling in utilizing the advantages of several models and utilizing their collective wisdom as follows:

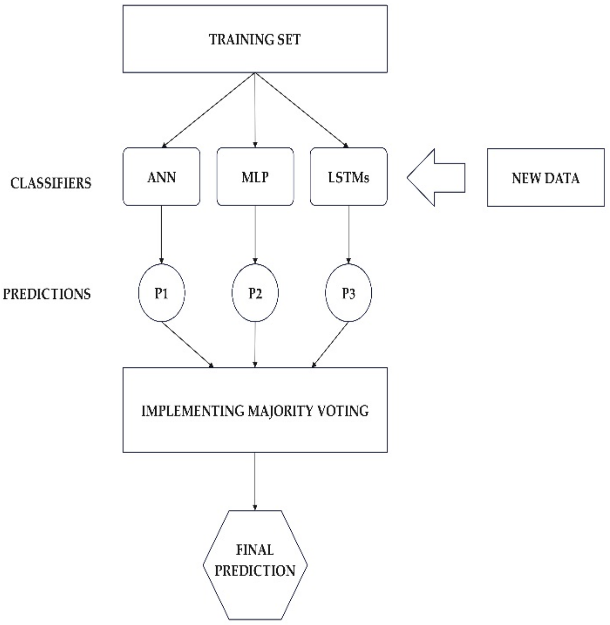

Figure 3. Ensemble model from majority voting of deep learning models

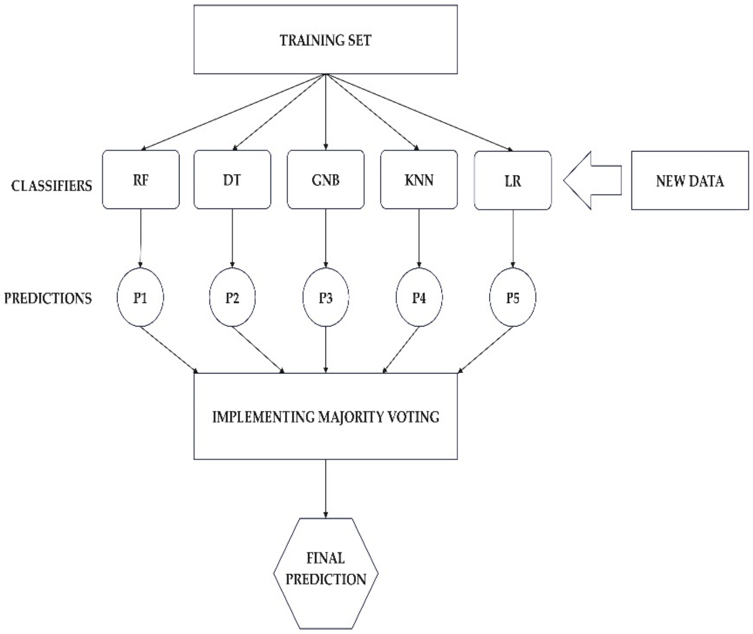

Figure 4. Ensemble model from majority voting of machine learning models

3. Results

The purpose of this study was to create a model which based on deep learning and machine learning for early diagnosis of diabetes in pre-diabetic patients. 253,680 records made up the dataset, of which 213,703 records were normal, 4,631 records were pre-diabetic, and 35,346 records were diabetic. The data was initially pre-processed by removing outliers and missing values. After that, a 75:25 split for the train and the test sets were applied to pre-processed data. On the pre-processed dataset, almost about of five machines learning algorithms and 3 deep learning algorithms were trained and assessed [29-33]. The results clearly shows that the deep learning-based model and our ensemble model made from deep learning technique outperformed the ML models, our ensemble model of Deep Learning and LSTMs got the best outcome with an accuracy of 85.03% and 85.00% respectively. The best-performing machine learning model was the ensemble model of ML models and Random Forest with the accuracies of 84.38% and 82.52% respectively, which achieved the highest accuracy and other performance evaluation parameters.

Table 3. Performance of the deployed individual and ensemble classifications

Sr. #	Model Name	Accuracy	Precision	Recall	F1 Score
1	Deep Learning Ensemble	85.03	80.51	85.00	81.01
2	LSTM	85.00	80.64	85.00	81.46
3	ANN	84.56	79.75	84.56	80.64
4	Conventional Learning Ensemble	84.45	79.03	83.45	80.32
5	MLP	84.27	79.79	84.27	81.08
6	Random Forest	82.52	80.12	82.52	82.13
7	Decision Tree	76.58	78.05	76.58	77.29
8	KNN	67.90	80.72	67.90	72.72
9	GNB	66.36	83.63	66.36	72.63
10	LR	64.47	85.10	64.47	71.99

In summary, the study shows the potential of machine & deep learning models for detection of diabetes in pre-diabetic patients at an early stage. The most crucial features for predicting diabetes were found by the deep learning-based model, which performed better than all machine learning models. These findings may help medical practitioners spot pre-diabetic patients who are at very large risk of acquiring the diabetes and should start treatment firstly.

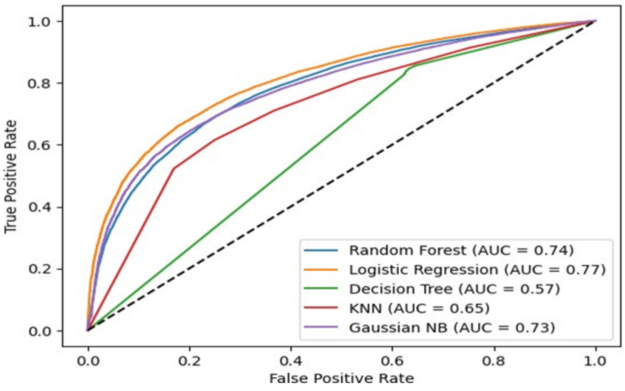

Figure 5. ROC Curve for conventional ML models

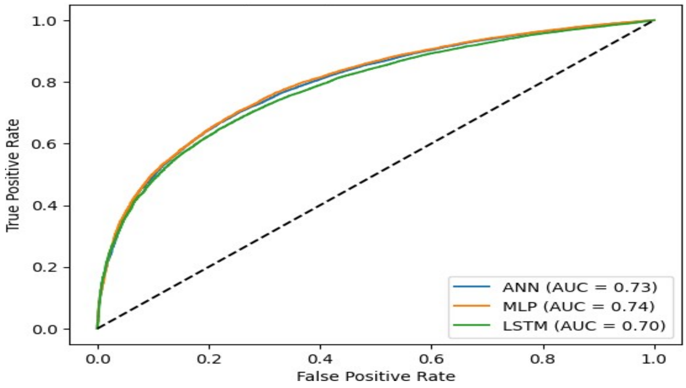

Figure 6. ROC Curve for DL models

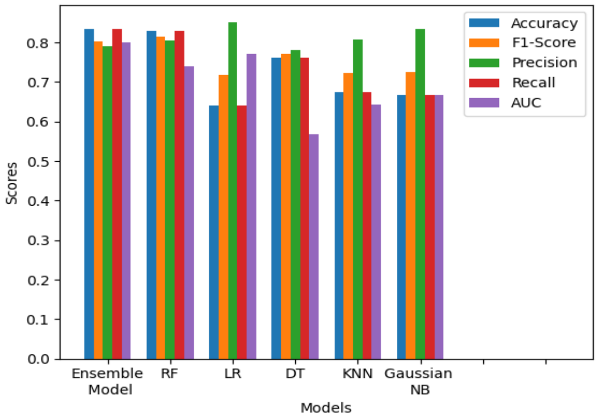

Figure 7. Results of ML models

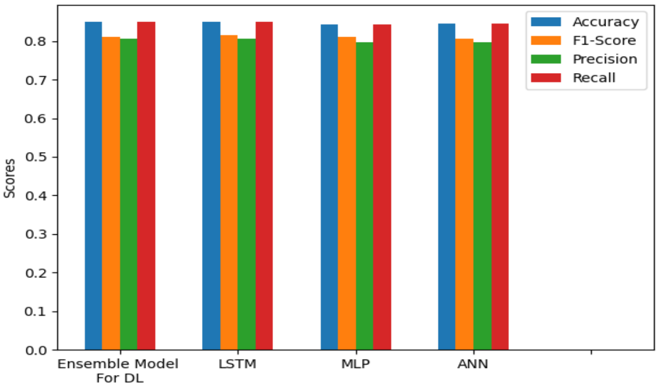

Figure 8. Results of DL model

4. Discussion

In this research, we suggested a machine and deep learning expert system to predict pre-diabetes and in its initial phases of diabetes. This study aimed to increase pre-diabetes detection accuracy, which can aid medical professionals in identifying and stopping the growth of diabetes in individuals. An approximately 253,680-person CDC dataset was used to create the expert system. After handling missing values as well as outliers during preprocessing, the data was split into test and training sets. On the training set, various deep learning and machine learning based models, such as LSTMs, logistic regression (LR), decision tree (DT), random forests, and artificial neural networks (ANN) were trained. So, the outcomes demonstrated that deep learning models performed better in terms of accuracy than machine learning models. The testing set showed that our ensemble model and LSTMs had the highest accuracy, 85.03%, proving that the suggested expert system is quite good at spotting pre-diabetes and the early stages of diabetes. Generally speaking, the suggested expert system using deep learning and machine learning approaches has demonstrated good results in identifying pre-diabetes and the early phases of diabetes. Health-care practitioners can utilize the system to detect high-risk pre-diabetic patients and give prompt interventions to stop the development of diabetes. Future studies could concentrate on creating a web-based platform or a mobile application to increase the system's useability for patients and medical personnel. Also using ensemble models and other techniques, accuracy can still be improved.

5. Conclusions

The expert system created in this work employing machine learning and deep learning approaches demonstrated promising outcomes in early detection and diagnosis of prediabetes at the early stages of diabetes. The algorithm successfully identified people who were at risk of acquiring diabetes with a high degree of accuracy. The findings of this study show the power of deep learning & machine learning approaches to uplist the precision of prediabetes detection, that can eventually aid patients in preventing the advancement of diabetes. The expert system created for this study, can serve as a very useful resource for medical professionals since it can identify people which are at a greater risk for developing diabetes and offer prompt interventions to stop it from happening. Future research can concentrate on improving the deep and machine learning models employed in the intelligent system as well as extending the dataset utilized in this work to encompass a more varied population. The system's impact can also be increased by creating a mobile app or web platform that makes it easier for patients and medical professionals to access the system. Future research can concentrate on improving the deep and machine learning models employed in the intelligent system as well as extending the dataset utilized in this work to encompass a more varied population. The system's impact can also be increased by creating a mobile app or web platform that makes it easier for patients and medical professionals to access the system. Overall, this study's findings show the promise of deep and machine learning approaches in the early diagnosis of pre-diabetes and initial phases of diabetes and emphasize the value of early intervention in stopping the evolution of this condition.

Data Availability Statement: The dataset is available on Kaggle on the following link" https://www. kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset". The Behavioral Risk Factor Surveillance System is a yearly telephone survey on health-related issues that the CDC performs. (BRFSS). Each year, around 0.4 million Americans take part in this annual survey.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. S. A. Alex, J. J. V. Nayahi, H. Shine, and V. Gopirekha, "Deep convolutional neural network for diabetes mellitus prediction," Neural Computing and Applications, vol. 34, pp. 1319-1327, 2022.
2. V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," Neural Computing and Applications, vol. 35, pp. 16157-16173, 2023.
3. W. H. Organization. (2023). Diabetes. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes
4. H. Gupta, H. Varshney, T. K. Sharma, N. Pachauri, and O. P. Verma, "Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction," Complex & Intelligent Systems, vol. 8, pp. 3073-3087, 2022.
5. D. P. P. R. Group, "Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin," New England journal of medicine, vol. 346, pp. 393-403, 2002.
6. A. Ramachandran, C. Snehalatha, S. Mary, B. Mukesh, A. Bhaskar, V. Vijay, et al., "The Indian Diabetes Prevention Programme shows that lifestyle modification and metformin prevent type 2 diabetes in Asian Indian subjects with impaired glucose tolerance (IDPP-1)," Diabetologia, vol. 49, pp. 289-297, 2006.
7. J. Lindström, P. Ilanne-Parikka, M. Peltonen, S. Aunola, J. G. Eriksson, K. Hemiö, et al., "Sustained reduction in the incidence of type 2 diabetes by lifestyle intervention: follow-up of the Finnish Diabetes Prevention Study," The Lancet, vol. 368, pp. 1673-1679, 2006/11/11/ 2006.
8. U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications," Journal of Healthcare Engineering, vol. 2021, p. 9930985, 2021/10/01 2021.
9. H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," Applied Computing and Informatics, vol. 18, pp. 90-100, 2022.
10. K. Kannadasan, D. R. Edla, and V. Kuppili, "Type 2 diabetes data classification using stacked autoencoders in deep neural networks," Clinical Epidemiology and Global Health, vol. 7, pp. 530-535, 2019/12/01/ 2019.
11. S. G, V. R, and S. K.P, "Diabetes detection using deep learning algorithms," ICT Express, vol. 4, pp. 243-246, 2018/12/01/ 2018.
12. D. H. I. Dataset. (2015). Kaggle. Available: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
13. D. Tree. (2023). GeeksforGeeks. Available: https://www.geeksforgeeks.org/decision-tree/
14. M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," Pattern Recognition, vol. 40, pp. 2038-2048, 2007/07/01/ 2007.
15. S. J. Rigatti, "Random forest," Journal of Insurance Medicine, vol. 47, pp. 31-39, 2017.
16. Y. Resti, E. S. Kresnawati, N. R. Dewi, and N. Eliyati, "Diagnosis of diabetes mellitus in women of reproductive age using the prediction methods of naive bayes, discriminant analysis, and logistic regression," Science and Technology Indonesia, vol. 6, pp. 96-104, 2021.
17. H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," Journal of Diabetes & Metabolic Disorders, vol. 19, pp. 391-403, 2020.
18. S. Larabi-Marie-Sainte, L. Aburahmah, R. Almohaini, and T. Saba, "Current techniques for diabetes prediction: review and case study," Applied Sciences, vol. 9, p. 4604, 2019.
19. Li, J., Hu, Q., Imran, A., Zhang, L., Yang, J. J., & Wang, Q. (2018, July). Vessel recognition of retinal fundus images based on fully convolutional network. In 2018 IEEE 42nd annual computer software and applications conference (COMPSAC) (Vol. 2, pp. 413-418). IEEE.
20. Akhtar, F., Li, J., Guan, Y., Imran, A., & Azeem, M. (2019). Monitoring bio-chemical indicators using machine learning techniques for an effective large for gestational age prediction model with reduced computational overhead. In Frontier Computing: Theory, Technologies and Applications (FC 2018) 7 (pp. 130-137). Springer Singapore.
21. Wajahat, A., Imran, A., Latif, J., Nazir, A., & Bilal, A. (2019, January). A Novel Approach of Unprivileged Keylogger Detection. In 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) (pp. 1-6). IEEE.
22. Bilal, A., Sun, G., Mazhar, S., Imran, A., & Latif, J. (2022). A Transfer Learning and U-Net-based automatic detection of diabetic retinopathy from fundus images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 10(6), 663-674.
23. Imran, A., Li, J., Pei, Y., Mokbal, F. M., Yang, J. J., & Wang, Q. (2020). Enhanced intelligence using collective data augmentation for CNN based cataract detection. In Frontier Computing: Theory, Technologies and Applications (FC 2019) 8 (pp. 148-160). Springer Singapore.
24. Ali, S., Rehman, S. U., Imran, A., Adeem, G., Iqbal, Z., & Kim, K. I. (2022). Comparative Evaluation of AI-Based Techniques for Zero-Day Attacks Detection. Electronics, 11(23), 3934.
25. A., Imran, A., Ullah, I., Alzahrani, A., Alheeti, K. M. A., & Yasin, A. (2022, October). Multi-model Ensemble Based Approach for Heart Disease Diagnosis. In 2022 International Conference on Recent Advances in Electrical Engineering & Computer Sciences (RAEE & CS) (pp. 1-8). IEEE.
26. Imran, A., Li, J., Pei, Y., Mokbal, F. M., Yang, J. J., & Wang, Q. (2019, July). Enhanced Intelligence Using Collective Data Augmentation for CNN Based Cataract Detection. In International Conference on Frontier Computing (pp. 148-160). Springer, Singapore.

27. Imran, A., Li, J., Pei, Y., Akhtar, F., Yang, J. J., & Wang, Q. (2019, December). Cataract Detection and Grading with Retinal Images Using SOM-RBF Neural Network. In 2019 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 2626-2632). IEEE.
28. Imran, A., Faiyaz, M., & Akhtar, F. (2018). An enhanced approach for quantitative prediction of personality in Facebook posts. International Journal of Education and Management Engineering (IJEME), 8(2), 8-19.
29. Imran, A., Aslam, W., & Ullah, M. I. (2017). Quantitative Prediction of Offensiveness using Text Mining of Twitter Data. Sindh University Research Journal-SURJ (Science Series), 49(1).
30. Li, J., Hu, Q., Imran, A., Zhang, L., Yang, J. J., & Wang, Q. (2018, July). Vessel Recognition of Retinal Fundus Images Based on Fully Convolutional Network. In 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC) (Vol. 2, pp. 413-418). IEEE.
31. Asad, R., Imran, A., Li, J., Almuhaimeed, A., & Alzahrani, A. (2023). Computer-Aided Early Melanoma Brain-Tumor Detection Using Deep-Learning Approach. Biomedicines, 11(1), 184.
32. M. R. R. Rana, S. U. Rehman, A. Nawaz, T. Ali, A. Imran et al., "Aspect-based sentiment analysis for social multimedia: a hybrid computational framework," Computer Systems Science and Engineering, vol. 46, no.2, pp. 2415–2428, 2023.
33. Sohaib Latif, Xian Wen Fang, Kaleem Arshid, Abdullah Almuhaimeed, Azhar Imran & Mansoor Alghamdi (2023) Analysis of Birth Data using Ensemble Modeling Techniques, Applied Artificial Intelligence, 37:1.