Annals of Emerging Technologies in Computing (AETiC)
Vol. 7, No. 1, 2023

Research Article

Stacked Ensemble-Based Type-2 Diabetes Prediction Using Machine Learning Techniques

Md Abdur Rahim[1], Md Alfaz Hossain[1], Md Najmul Hossain[1], Jungpil Shin[2] and Keun Soo Yun[3,*]

[1]Pabna University of Science and Technology, Bangladesh
rahim@pust.ac.bd; alfaz.cse@gmail.com; najmul_eece@pust.ac.bd
[2]The University of Aizu, Japan
jpshin@u-aizu.ac.jp
[3]Ulsan College, Korea
ksyun@uc.ac.kr
*Correspondence: ksyun@uc.ac.kr

Received: 6th April 2022; Accepted: 22nd December 2022; Published: 1st January 2023

Abstract: Diabetes is a long-term disease caused by the human body's inability to make enough insulin or to use it properly. This is one of the curses of the present world. Although it is not very severe in the initial stage, over time, it takes a deadly shape and gradually affects a variety of human organs, such as the heart, kidney, liver, eyes, and brain, leading to death. Many researchers focus on the machine and in-depth learning strategies to efficiently predict diabetes based on numerous risk variables such as insulin, BMI, and glucose in this healthcare issue. We proposed a robust approach based on the stacked ensemble method for predicting diabetes using several machine learning (ML) methods. The stacked ensemble comprises two models: the base model and the meta-model. Base models use a variety of models of ML, such as Support Vector Machine (SVM), K Nearest Neighbor (KNN), Naïve Bayes (NB), and Random Forest (RF), which make different assumptions about predictions, and meta-models make final predictions using Logistic Regression from predictive outputs from base models. To assess the efficiency of the proposed model, we have considered the PIMA Indian Diabetes Dataset (PIMA-IDD). We used linear and stratified sampling to ensure dataset consistency and K-fold cross-validation to prevent model overfitting. Experiments revealed that the proposed stacked ensemble model outperformed the model specified in the base classifier as well as the comprehensive methods, with an accuracy of 94.17%.

Keywords: *Base and Meta Model; Diabetes Type 2; Machine Learning Techniques; Stacked Ensemble*

1. Introduction

Nowadays, diabetes is gradually becoming one of the leading causes of death. It causes a metabolic disorder in the body that persists for a long time. This disorder is marked by high blood sugar, which can damage organs like the heart, blood vessels, kidneys, eyes, and nerves. There are three different forms of diabetes: type 1 diabetes, type 2 diabetes (T2D), and diabetes that happens during pregnancy. Type 1 diabetes results from an autoimmune reaction that causes the body to stop making insulin and develop faster. However, it affects about 5% to 10% of patients with the disease [1]. Type 2 diabetes (T2D) is characterized by poor use of insulin in the body and an inability to maintain normal blood sugar levels. Currently, the number of patients with T2D is much higher than other types, which is 90% to 95% [2]. The symptoms of gestational diabetes are observed in pregnant women and usually go away after childbirth [3]. It can be a risk factor for T2D in later life. However, early diagnosis of diabetes can save millions of

Md Abdur Rahim, Md Alfaz Hossain, Md Najmul Hossain, Jungpil Shin and Keun Soo Yun, "Stacked Ensemble-Based Diabetes Type-2 Prediction Using Machine Learning Techniques", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 30-39, Vol. 7, No. 1, 1st January 2023, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2023.01.003, Available: http://aetic.theiaer.org/archive/v7/v7n1/p3.html.

lives through proper treatment and lifestyle changes to a healthy life. In this context, this paper proposes an approach for predicting T2D based on the stacked ensemble method (SEM).

Furthermore, diabetes affects over 422 million people globally, resulting in approximately 1.5 million deaths annually [4]. It also affects most individuals in poor and moderate-income countries such as Bangladesh [5]. Suburbanization, population aging, growing unhealthy cultures, and unpreparedness for afforestation and control may be the main causes and challenges of diabetes, which has become a major health problem in poor and moderate-income countries. Consequently, T2D is growing at an alarming. The activities of many researchers have been able to diagnose diabetes with various machine and deep learning algorithms [6-7]. The voting classifier technique was proposed to predict the diabetes in [8]. The proposed method gives the highest accuracy of 79.04% in the PIMA Diabetes Dataset. In [9], the model of diabetes prediction and classification was proposed, and the authors achieved a classification accuracy of 92.28%. However, raising the classification accuracy level can help forecast diabetes for effective treatments. In [10], the author proposed a deep neural network for categorizing diabetes data that combines stacked auto-encoders to extract features. However, the accuracy was 86.26% when the network was fine-tuned using backpropagation in supervised mode with the training dataset. The stacking-based evolutionary ensemble learning system "NSGA-II-Stacking" was created in [11] to predict the onset of T2D within five years. A multi-objective optimization algorithm was used to improve the classification performance and the simplicity of the ensemble. The proposed system has a maximum accuracy of 83.8%.

Therefore, the primary goal of this research is to reliably identify early-stage diabetes, which may contribute to a healthy lifestyle. The PIMA Indian Diabetes Dataset (PIMA-IDD) has been used to predict diabetes in this study. There are 768 observations in this dataset, containing 9 variables. We pre-processed the dataset and trained four machine learning algorithms, such as K nearest neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), and Naïve Baizes (NB), as base learners for the proposed model. The prediction of these four algorithms is used as input features for predictive meta-learner algorithms such as Logistic Regression (LR). However, this Meta learner produces the final output.

The following are the aims and main contributions of this study:

- The PIMA-IDD is used to evaluate our proposed model in order to determine prediction accuracy and compare it to state-of-the-art machine learning classifiers. To ensure dataset consistency, we used linear and stratified sampling.
- The deficit, conflicting, inappropriate, and incorrect datasets have been processed for better prediction. In the preprocessing step, we examine the zero values of the whole dataset. If a zero value exists, we replace it with the mean values of those specific columns. Additionally, the proposed model was evaluated using a total of 9 features.
- We proposed a robust model using the Stacked Ensemble method (SEM), which works in two levels where the first level is the base model, which gives a prediction, and then the second level where the meta-model takes the first level prediction as input and then the final prediction is given.

2. Related Work

Many academics have proposed and evaluated various prediction models in various healthcare datasets, employing data, machine learning (ML), and deep learning techniques or a combination of these approaches. In [12], the authors proposed a computer-based methodology for predicting diabetes patients and suggested preventive interventions. The authors obtained an accuracy of 84%; nevertheless, enhancing accuracy in early-stage diabetes remains a significant problem. A review was conveyed to detect diabetes based on PIMA-IDD using different ML techniques [13]. The authors compared and discussed the outcomes of several ML algorithms and their pros and limitations. In [14], three different classifiers such as RF, Multilayer Perceptron (MLP), and LR used for diabetes classification. The authors used the PIMA-IDD and achieved the highest accuracy using MLP at 87.26%. A deep neural network is applied to PIMA-IDD, where a dropout method is used to solve overfitting problems [15]. The diabetes prediction was presented in [16] using ML and data mining methodologies. To increase the classification performance of the deep classification models, however, larger datasets are required. A data mining approach was developed to predict T2D in [17]. The authors used K-mean algorithms and LR to analyse their predictions and obtained 90.7% accuracy, which is 3.04% better than comparative approaches. The authors of [18] employed the WEKA technique to predict diabetes patients. Various ML approaches were used, including NB, SVM, RF, and generic CART algorithms, and finally, SVM achieved a maximum accuracy of 79.13%.

Furthermore, three distinct datasets from China, Japan, and Iran were evaluated with stacked ensemble classifiers such as MLP, Decision Tree (DT), SVM, and LR [19]. This model has demonstrated acceptable levels of accuracy and may aid in the early detection of diabetic patients. Deep learning techniques are also essential for the diagnosis of many diseases, including the classification of mental functions [20], the diagnosis of diabetes [10], and automated skin disease prognosis [21]. In this study, we compare the proposed stacked ensemble system with deep and machine learning.

3. Proposed Methodology

The main goal of this study is to illustrate a promising strategy for diabetes prediction by analysing important characteristics. We pre-process the dataset using sampling techniques such as linear and stratified and propose a stacked ensemble method (SEM) technique based on various machine learning techniques. The procedure for predicting diabetes using the SEM is described in this section. It is divided into base and meta models. In the base model, various ML methods such as KNN, SVM, RF, and NB were used for predictions which were used as a new training set for the meta-model. The LR model was used as a meta-model for model evaluation to make the final prediction. Figure 1 shows the basic architecture of SEM for diabetes type 2 predictions.

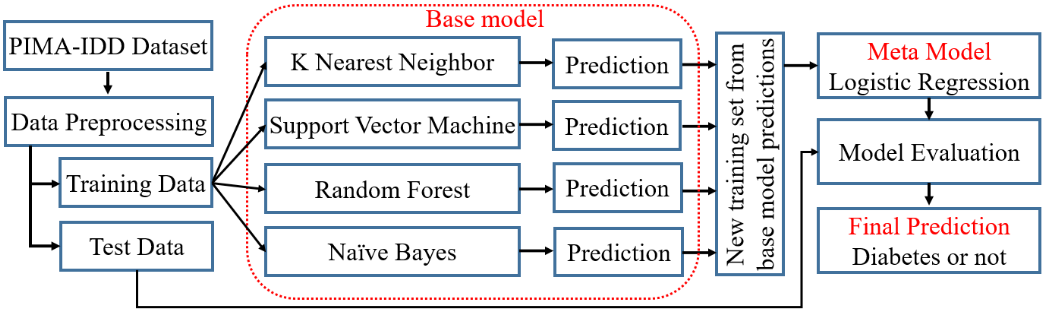

Figure 1. Block diagram of SEM for diabetes prediction.

3.1. Dataset Description and Pre-processing

In this paper, we have considered Pima Indians Diabetes Dataset (PIMA-IDD) [22] to evaluate our proposed model. This was originally published by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). It is used to diagnose and predict whether or not a person has diabetes. The main reason for selecting the PIMA-IDD dataset is that most people worldwide have similar lifestyles that are excessively reliant on processed foods and low levels of physical activity. As a result, people may be more susceptible to diabetes. This dataset was developed by the NIDDK's long-term cohort study on diabetes risk factors. It also includes diagnostic measurements and characteristics that can be used to forecast approaching diabetes or chronic illness. The PIMA-IDD dataset has 9 (nine) variables to determine if a person's diabetes is positive or negative. It comprises information from 768 people, 500 non-diabetic diabetics and 268 diabetics, respectively. The database contains a target variable and 8 (eight) different attributes, as shown in Table 1.

Table 1. Descriptions of PIMA-IDD characteristics.

Feature	Descriptions
Pregnancies	Pregnancy count (Numeric)
Glucose	The density of glucose in plasma
High Blood Pressure	Pulse diastolic (mm Hg)
Thickness of the Skin	The depth of the skin crease (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index
Diabetes Pedigree Function	The pedigree function of diabetes
Age	Age
Outcome	Target outcome (Diabetes or not diabetes)

In the pre-processing step, incorrect or incomplete datasets are considered pre-processed for better prediction. The minimum value of variables of glucose, blood pressure, skin thickness, insulin, and BMI is zero, which makes no sense. Therefore, the mean values of those specific variables in the dataset replaced zero values. In order to achieve more accurate predictions, the datasets must be evenly split between test and training data. Sampling is the practice of selecting a sample of data to reliably extract features and parameters from a larger dataset, allowing it to make a more meaningful contribution to developing machine-learning algorithms. We used sample methods like LS (linear sampling) and SS (stratified sampling) to ensure uniformity. To further analyze the data, LS segments it into subsets. The tuples and fields within the subset are also preserved in their original order. Additionally, SS constructs subsets of the collection through arbitrary division. It also guarantees a uniform class distribution throughout the entire dataset.

3.2. Base and Meta Learner Model

Base models employ several machine learning models that make various assumptions about predictions, which are subsequently used as training sets for meta-models to create final predictions.

3.2.1. K Nearest Neighbors (KNN)

The KNN approach is one of the most widely used classification algorithms in machine learning [23]. It works by computing the Euclidean distance between a set of K neighbors. We estimated the number of observations in each category among all of these K neighbors. However, the new observations were allocated to the categories with the most neighbors. Equation (1) can be used to compute the Euclidean distance.

$$d(j,k) = \sqrt{\sum_{i=1}^{n} (k_i - j_i)^2} \quad (1)$$

3.2.2. Support Vector Machine (SVM)

The SVM is a well-known and effective supervised learning method for classification and regression and outlier detection [24]. SVM chooses extreme points/vectors that help to create hyperplanes. This allows us to place new data points in the appropriate category by separating n-dimensional space in the class. The amount of features in the dataset determines the size of the hyperplane; for example, if there are just two essential features, the hyperplane will be represented by a straight line. When the number of features is three, the hyperplane is depicted as a two-dimensional plane.

3.2.3. Random Forest (RF)

RF is a classifier that uses numerous decision trees on different subsets of a dataset to improve prediction accuracy [25]. It comprises two stages: the first is to merge the N Decision trees into a random forest, and the second is to make predictions for each tree generated in the first phase. The RF takes less time to train than other algorithms and predicts output with maximum accuracy; it also operates quickly on large datasets. The RF algorithm is depicted in Figure 2.

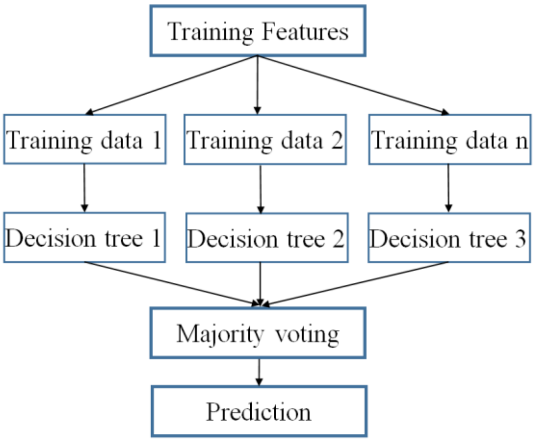

Figure 2. Procedure of the RF algorithm.

3.2.4. Naïve Bayes (NB)

The simplest and most successful classification method is the Naive Bayes Classifier, which helps to create machine learning models that can quickly generate predictions [26]. It is based upon the Bayes' theorem's assumption and implies that one characteristic's presence is unrelated to the presence of other features. Equation (2) illustrates the formula for Bayes' theorem (2).

$$P(J|K) = \frac{P(K|J)*P(J)}{P(K)} \quad (2)$$

Where $P(J)\, and \;\; P(J|K)$ are defined probability and the next probability of the class. $P(K|J)\; and\; P(K)$ is expressed as the probability of a given class and predictor.

3.2.5. Logistic Regression (LR)

Logistic regression can quickly determine the most efficient criteria for classification and classify the observations based on a wide variety of data [19]. It is an approach for determining categorical variables from independent factors. In order to categorize the data, we used an S-shaped logistic function that predicts two maximum values (0 or 1). Equation (3) is used to derive the logistic function. However, the LR focuses on the boundaries between existing classes and further reveals the class potential, which depends on the distance from the boundaries in a certain way. When the data set is too large, it moves toward the limit. These claims about probability make logistical regression more than a classifier.

$$S(x) = \frac{1}{1+e^{-x}} \quad (3)$$

3.2.6. Stacked Ensemble Model

In our proposed method, we have used stacked ensemble for various machine learning algorithms. It has divided into two steps. First, it uses basic classification algorithms that are expressed as base models. Each base model is trained in a given dataset and gives an intermediate prediction. Second, the meta-model takes the intermediate prediction as an input feature and gives the final output for the target value. There is a huge potential for overfitting as we train the proposed model with the same data set. Therefore, to overcome this overfitting problem, we used K-fold cross-validation. The SEM process is illustrated in Algorithm 1.

Algorithm 1. SEM Process

```
Input: Training data $T = \{a_i, b_i\}_{i=1}^{n}$
Output: Classifier of ensemble E
Step 1: Learn base model classifiers
        for $x = 1\ to\ m$ do
          learn $e_t$ based on $T$
        end for
Step 2: new data set of base model predictions
        for $y = 1\ to\ n$ do
        $T_e = \{a_i', b_i\}, where\ a_i^{;} = \{e_1(a_i), \dots, e_m(a_i)\}$
        end for
Step 3: create a meta model
        learn E based on $T_e$
        return E
```

4. Experimental Results and Discussion

This section discusses the effectiveness, efficiency, and satisfaction of the proposed approach for diagnosing T2D.

4.1. Pre-processing and Evaluation Metrics

The PIMA-IDD dataset was used in the experimentation. There are 768 data points in the dataset, with 9 feature columns, as described in section 3.1. According to descriptive analysis, some variables have a minimal value of 0. However, these values are either non-existent or extrinsic. We used mean values to pre-process these 0's. Moreover, we obtain the relationship between different properties of the dataset using a correlation matrix. Figure 3 shows the correlation heatmap, which indicates how strongly a feature is closely related to other features. It also simultaneously fills missing values and rejects outliers. Figure 4 displays the histogram of the essential characteristics.

The dataset was partitioned in a 70:30 (%) ratio for training and testing. We analysed performance matrices like accuracy, precision, recall, and F1 score to confirm the consistency and effectiveness of the proposed methods. True positive (TrP) indicates that the predicted and actual class values are 1. When the predicted and actual class values are both zero, this is referred to as a true negative (TrN). A false negative (FaN) or false positive (FaP) occurs when the expected class differs from the actual class. The essential metric is accuracy, defined as the ratio of correctly predicted observations to total observed observations. Using Equations (4), (5), (6), and (7), we determined the accuracy, precision, recall, and F1 score.

$$Accuracy = \frac{TrP+TrN}{TrP+FaP+TrN+FaN} \tag{4}$$

$$Precision = \frac{TrP}{TrP+FaP} \tag{5}$$

$$Recall = \frac{TrP}{TrP+FaN} \tag{6}$$

$$F1\ Score = \frac{2*Precision*Recall}{Precision+Recall} \tag{7}$$

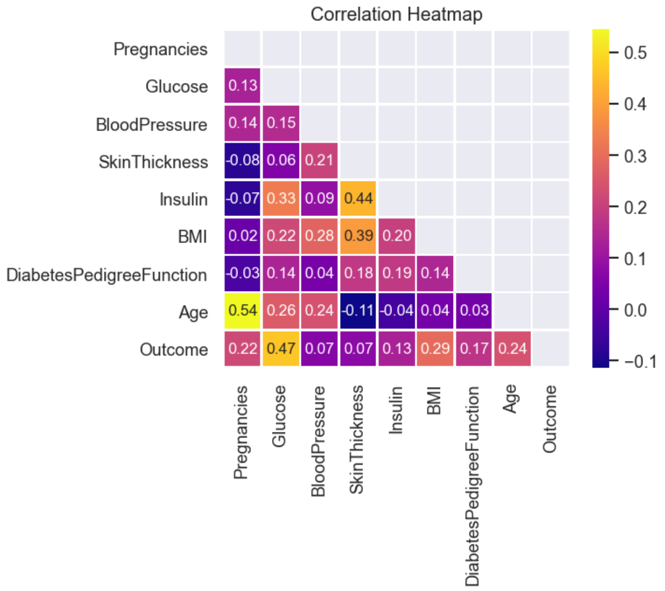

Figure 3. Correlation heatmap

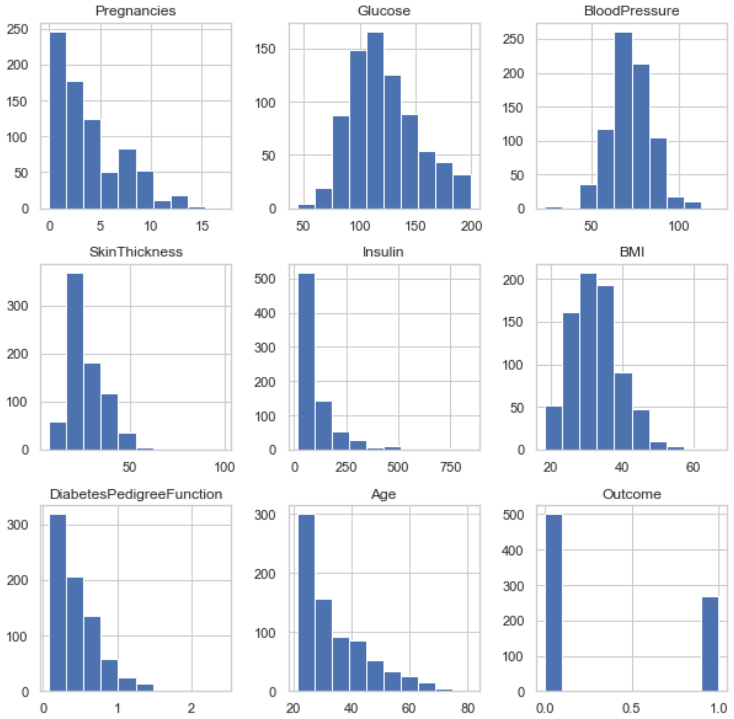

Figure 4. Histogram of the key attributes

4.2. Results and Discussion

The proposed approach employs a stacked ensemble of base and meta models. However, an ensemble of four machine learning models (KNN, SVM, RF, and NB) is used to create the base model, and the predictions from the base model are fed into the meta model (LR). Accuracy, precision, recall, and F1 scores are used to evaluate base and meta-model performance. The performance evaluation of the base classifier and the proposed method is shown in Table 2. Figure 5 compares the recognition accuracy of several machine learning algorithms and the proposed SEM.

Table 2. Performance evaluation for dataset PIMA-IDD

Methods	Evaluation Metrics			
	Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)
KNN	83.16	82.32	82.74	77.99
SVM	91.33	79.20	84.83	79.29
RF	98.15	83.55	90.40	86.73
NB	84.69	83.42	84.05	79.61
LR	86.73	82.52	84.58	79.94
SEM	95.92	94.95	95.43	94.17

Table 2 compares the recognition accuracy of several machine-learning approaches using the PIMA-IDD dataset. The stacked ensemble approach performed remarkably well in terms of accuracy, precision, F1 score, and recall, as shown in Table 2, with 94.17 %, 95.92 %, 95.43 %, and 94.95 %, respectively. The individual performances of each method give an accuracy of KNN (77.99%), SVM (79.29%), RF (86.73%), NB (79.61), and LR (79.94%), respectively. We observed that the LR does not perform well on the given dataset, with an accuracy of 79.94%; however, as a meta-model, it performs betters with an accuracy of 94.17%. The performance comparisons of the base model and the meta models are shown in Figure 6. Moreover, the correlation characteristics with the target variable demonstrate that the correlation coefficient is greatly improved over that presented in Figure 7. The data shown highlights some numbers such as maximum, minimum, standard deviation, mean, and quartiles of 25%, 50%, and 75%. By analysing BMI and pregnancy, as shown in Figure 7, we observe a strong positive relationship between BMI and the number of pregnancies. In addition, as shown in Figure 8, women who tested positive were thought to have a higher BMI for the interquartile range.

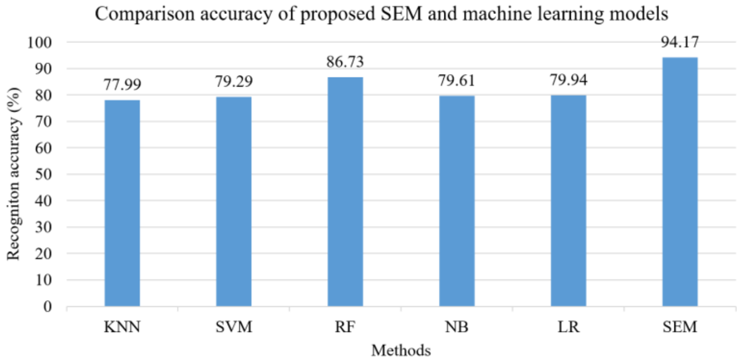

Figure 5. The comparison accuracy of various machine learning methods with the proposed SEM.

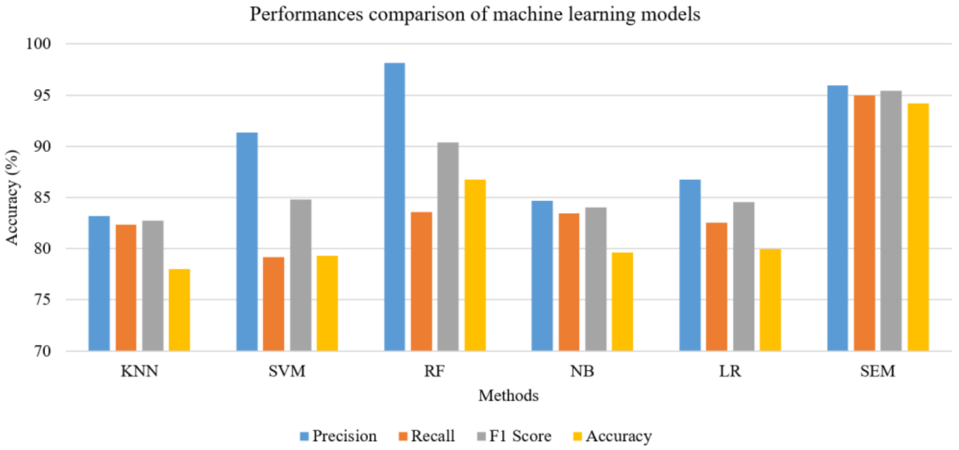

Figure 6. A comparison graph showing the evaluation metrics.

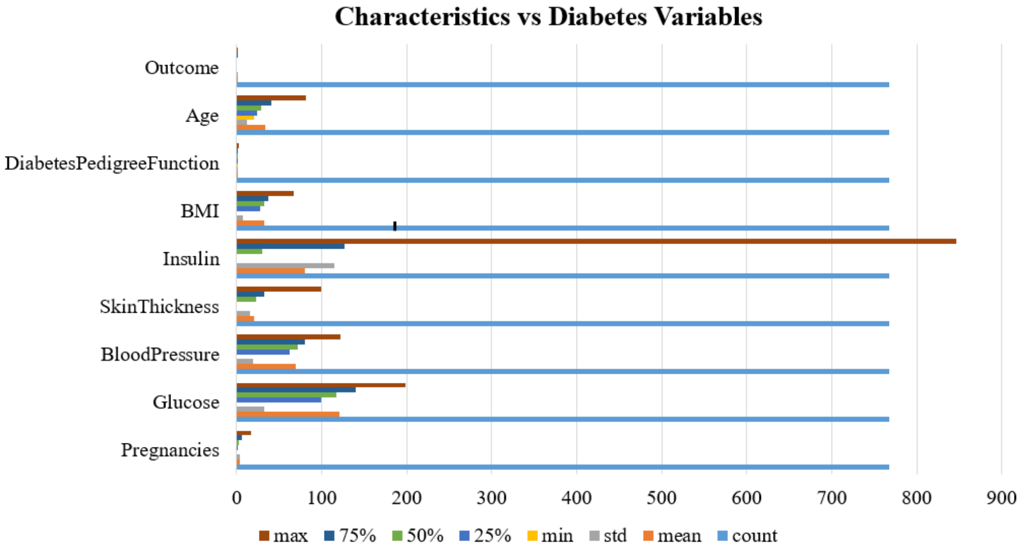

Figure 7. A comparison of BMI, pregnancy, and diabetes characteristics.

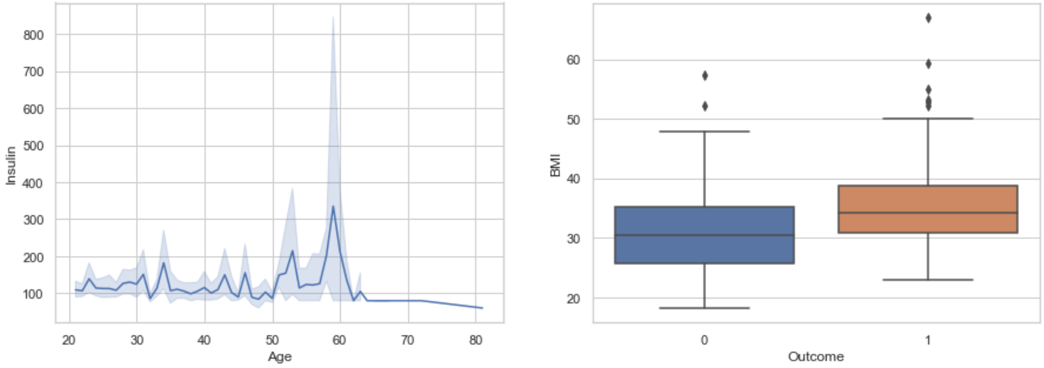

Figure 8. Characteristics associated with diabetes.

Moreover, the proposed stacked ensemble techniques have also been compared with the state-of-the-art methods. Table 3 shows the results of the proposed SEM and state-of-the-art techniques. According to the results of the experiments, the proposed technique outperforms the traditional ones.

Table 3. Comparison with conventional methods

References	Methods	Accuracy (%)
Ref. [6]	Decision tree-based RF and SVM	83
Ref. [9]	K-NN	92.28
Ref. [10]	Deep neural networks	86.26
Ref. [11]	NSGA-II-Stacking	83.8
Ref. [14]	LSTM	87.26
Proposed	Stacked Ensemble Method	94.17

5. Conclusion

This research proposed a stacked ensemble strategy to predict patients with T2D using base and meta-models. As a base model, four machine learning algorithms, such as KNN, SVM, RF, and NB, are considered, whereas the LR is employed for meta models. However, the predictions of the base model were employed as input to the meta-model. Therefore, the Meta-model (LR) was used to obtain the final predictions. The PIMA-IDD dataset was used in the experiment. The experimental findings demonstrated that the proposed stacked ensemble technique outperformed other machine learning methods. The proposed technique attained the highest accuracy of 94.17%, whereas the individual machine learning algorithms obtained KNN (77.99%), SVM (79.29%), RF (86.73%), NB (79.61), and LR (79.94%), respectively. In future research, we will investigate in-depth learning models to predict T2D by including several characteristics for improved performance.

References

[1] Anjali Verma, Rajesh Rajput, Surender Verma, Vikas KB Balania and Babita Jangra, "Impact of lockdown in COVID 19 on glycemic control in patients with type 1 Diabetes Mellitus", *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, ISSN: 1871-4021, Vol. 14, No. 5, pp. 1213-1216, 1st September 2020, Published by Elsevier, DOI: 10.1016/j.dsx.2020.07.016, Available: https://www.sciencedirect.com/science/article/pii/S1871402120302642.

[2] Ralph A. DeFronzo, Ele Ferrannini, Leif Groop, Robert R. Henry, William H. Herman *et al.*, "Type 2 diabetes mellitus", *Nature Reviews Disease Primers*, ISSN: 2056676X, pp. 1-22, Vol. 1, No. 1, 23rd July 2015, Published by Nature Publishing Group, DOI: 10.1038/nrdp.2015.19, Available: https://www.nature.com/articles/nrdp201519.

[3] Jasmine F Plows, Joanna L Stanley, Philip N Baker, Clare M Reynolds and Mark H Vickers, "The pathophysiology of gestational diabetes mellitus", *International Journal of Molecular Sciences,* Print ISSN: 14220067, Online ISSN: 16616596, pp. 3342, Vol. 19, No. 11, 26th October 2018, Published by MDPI Multidisciplinary Digital Publishing Institute, DOI: 10.3390/ijms19113342, Available: https://www.mdpi.com/1422-0067/19/11/3342.

[4] Victor Chang, Jozeene Bailey, Qianwen Ariel Xu and Zhili Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms", *Neural Computing and Applications*, Print ISSN: 09410643, Online ISSN: 14333058, pp. 1-17, 24 March 2022, Published by Springer, DOI: 10.1007/s00521-022-07049-z, Available: https://link.springer.com/article/10.1007/s00521-022-07049-z.

[5] AK Mohiuddin, "Diabetes fact: Bangladesh perspective", *International Journal of Diabetes Research*, ISSN: 2414-2409, Vol. 2, No. 1 pp. 14-20, 24th February 2019, DOI: 10.17554/j.issn.2414-2409.2019.02.12, Available: http://96.126.98.199/index.php/ijdr/article/view/2457/2835.

[6] Raja Krishnamoorthi, Shubham Joshi, Hatim Z. Almarzouki, Piyush Kumar Shukla, Ali Rizwan *et al.*, "A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques", *Journal of Healthcare Engineering*, Print ISSN: 20402295, Online ISSN: 20402309, 11th January 2022, Published by Hindawi Limited, DOI: 10.1155/2022/1684017, Available: https://www.hindawi.com/journals/jhe/2022/1684017.

[7] Rashmi Srivastava and Rajendra Kumar Dwivedi, "A Survey on Diabetes Mellitus Prediction Using Machine Learning Algorithms", in *Lecture Notes in Networks and Systems (LNCS), ICT Systems and Sustainability*, Vol. 321, Print ISBN: 978-981-16-5986-7, Online ISBN: 978-981-16-5987-4, DOI: 10.1007/978-981-16-5987-4_48, pp. 473-480, 2022, Published by Springer, Singapore, Available: https://link.springer.com/chapter/10.1007/978-981-16-5987-4_48.

[8] Saloni Kumari, Deepika Kumar and Mamta Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier", *International Journal of Cognitive Computing in Engineering*, ISSN: 2666-3074, Vol. 2 pp. 40-46, June 1 2021, Published by KeAi Publisher, DOI: 10.1016/j.ijcce.2021.01.001, Available: https://www.sciencedirect.com/science/article/pii/S2666307421000048.

[9] Prachi Ahlawat, "DCPM: An effective and robust approach for diabetes classification and prediction", *International Journal of Information Technology*, Print ISSN: 2511-2104, Electronic ISSN: 2511-2112, Vol. 13, No. 3, pp. 1079-1088, 18 April 2021, Published by Springer, DOI: 10.1007/s41870-021-00656-4, Available: https://link.springer.com/article/10.1007/s41870-021-00656-4.

[10] Kannadasan K, Damodar Reddy Edla and Venkatanareshbabu Kuppili, "Type 2 diabetes data classification using stacked autoencoders in deep neural networks", *Clinical Epidemiology and Global Health*, ISSN: 2213-3984, Vol. 7, No. 4, pp. 530-535, December 2019, Published by Elsevier, DOI: 10.1016/j.cegh.2018.12.004, Available: https://www.sciencedirect.com/science/article/abs/pii/S221339841830277X.

[11] Namrata Singh and Pradeep Singh, "Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus", *Biocybernetics and Biomedical Engineering,* Vol. 40, No. 1, pp. 1-22, 2020, DOI: 10.1016/j.bbe.2019.10.001, Available: https://www.sciencedirect.com/science/article/abs/pii/S020852161930467X.

[12] Hang Lai, Huaxiong Huang, Karim Keshavjee, Aziz Guergachi and Xin Gao, "Predictive models for diabetes mellitus using machine learning techniques", *BMC Endocrine Disorders*, ISSN: 14726823, Vol. 19, No. 1, pp. 1-9, 15 October 2021, Published by BioMed Central Ltd., DOI: 10.1186/s12902-019-0436-6, Available: https://link.springer.com/article/10.1186/s12902-019-0436-6.

[13] Ambika Choudhury and Deepak Gupta, "A survey on medical diagnosis of diabetes using machine learning techniques", *In Advances in Intelligent Systems and Computing: Recent Developments in Machine Learning and Data Analytics*, Singapore: Springer, 2022, Vol. 740, pp. 67-78, Print ISBN: 978-981-13-1279-3, Online ISBN: 978-981-13-1280-9, DOI: 10.1007/978-981-13-1280-9_6, Available: https://link.springer.com/chapter/10.1007/978-981-13-1280-9_6.

[14] Umair Muneer Butt, Sukumar Letchmunan, Mubashir Ali, Fadratul Hafinaz Hassan, Anees Baqir *et al.*, "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications", *Journal of Healthcare Engineering*, Print ISSN: 20402295, Online ISSN: 20402309, 29th September 2021, Published by Hindawi Limited, DOI: 10.1155/2021/9930985, Available: https://www.hindawi.com/journals/jhe/2021/9930985.

[15] Akm Ashiquzzaman, Abdul Kawsar Tushar, Md Islam, Dongkoo Shon, Kichang Im *et al.*, "Reduction of overfitting in diabetes prediction using deep learning neural network", in *Lecture Notes in Electrical Engineering, IT*

Convergence and Security (LNEE), Vol. 449, Print ISBN: 978-981-10-6450-0, Online ISBN: 978-981-10-6451-7, pp. 35-43, 31st August 2017, DOI: 10.1007/978-981-10-6451-7_5, Available: https://link.springer.com/chapter/10.1007/978-981-10-6451-7_5.

[16] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas *et al.*, "Machine learning and data mining methods in diabetes research", *Computational and Structural Biotechnology Journal*, ISSN: 20010370, Vol. 15, pp. 104-116, January 2017, Research Network of Computational and Structural Biotechnology, DOI: 10.1016/j.csbj.2016.12.005, Available: https://www.sciencedirect.com/science/article/pii/S2001037016300733.

[17] Han Wu, Shengqi Yang, Zhangqin Huang, Jian He and Xiaoyi Wang, "Type 2 diabetes mellitus prediction model based on data mining", *Informatics in Medicine Unlocked*, ISSN: 23529148, Vol. 10, pp. 100-107, January 2018, DOI: 10.1016/j.imu.2017.12.006, Available: https://www.sciencedirect.com/science/article/pii/S2352914817301405.

[18] Ayman Mir and Sudhir N. Dhage, "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare", In *Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (IEEE ICCUBEA '18)*, 16 August 2018, Pimpri Chinchwad College of Engineering, Pune, Maharastra, India, Print ISBN: 978-1-5386-5258-9, E-ISBN: 978-1-5386-5257-2, DOI: 10.1109/ICCUBEA.2018.8697439, pp. 1-6, Published by IEEE, Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8697439.

[19] Norma Latif Fitriyani, Muhammad Syafrudin, Ganjar Alfian, Agung Fatwanto, Syifa Latif Qolbiyani *et al.*, "Prediction Model for Type 2 Diabetes using Stacked Ensemble Classifiers", In *Proceedings of the 2020 International Conference on Decision Aid Sciences and Application (DASA)*, Print ISBN: 978-1-7281-9678-7, Online ISBN: 978-1-7281-9677-0, pp. 399-402, 8 November 2020, Published by IEEE, DOI: 10.1109/DASA51403.2020.9317090, Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9317090.

[20] Hariprasath Manoharan, Sulaima Lebbe Abdul Haleem, S. Shitharth, Pravin R. Kshirsagar, *et al.*, "A machine learning algorithm for classification of mental tasks", *Computers and Electrical Engineering*, ISSN: 0045-7906, Vol. 99, pp. 107785, 1 April 2022, Published by Elsevier, DOI: 10.1016/j.compeleceng.2022.107785, Available: https://www.sciencedirect.com/science/article/abs/pii/S0045790622000854.

[21] Pravin R. Kshirsagar, Hariprasath Manoharan, S. Shitharth, Abdulrhman M. Alshareef, Nabeel Albishry *et al.*, "Deep Learning Approaches for Prognosis of Automated Skin Disease", *Life*, Vol. 12, No. 3, pp. 426, 15 March 2022, Published by MDPI Multidisciplinary Digital Publishing Institute, DOI: 10.3390/life12030426, Available: https://www.mdpi.com/2075-1729/12/3/426.

[22] Ravinder Ahuja, Subhash C. Sharma and Maaruf Ali, "A Diabetic Disease Prediction Model Based on Classification Algorithms", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 44-52, Vol. 3, No. 3, 1st July 2019, Published by International Association of Educators and Researchers (IAER), DOI: 10.33166/AETiC.2019.03.005, Available: http://aetic.theiaer.org/archive/v3/v3n3/p5.html.

[23] Yingquan Wu, Krassimir Ianakiev and Venu Govindaraju, "Improved k-nearest neighbor classification", *Pattern Recognition*, ISSN: 313203, Vol. 35, No. 10, pp. 2311-2318, October 2002, DOI: 10.1016/S0031-3203(01)00132-7, Available: https://www.sciencedirect.com/science/article/abs/pii/S0031320301001327.

[24] Yong Shi, "Support Vector Machine Classification", *Advances in Big Data Analytics*, Print ISSN: 978-981-16-3606-6, Online ISSN: 978-981-16-3607-3, pp. 97-246, 13 January 2022, Published by Springer, DOI: 10.1007/978-981-16-3607-3_3, Available: https://link.springer.com/chapter/10.1007/978-981-16-3607-3_3.

[25] Bo-Suk Yang, Xiao Di and Tian Han, "Random forests classifier for machine fault diagnosis", *Journal of Mechanical Science and Technology*, Print ISSN: 1738494X, Online ISSN: 19763824, pp. 1716-1725, Vol. 22, No. 9, September 2008, Published by Korean Society of Mechanical Engineers, DOI: 10.1007/s12206-008-0603-6. Available: https://link.springer.com/article/10.1007/s12206-008-0603-6.

[26] Jae-Cheol Park and Jea-Young Lee, "How to build nomogram for type 2 diabetes using a naïve Bayesian classifier technique", *Journal of Applied* Statistics, Print ISSN: 02664763, Online ISSN: 13600532, Vol. 45, No. 16, pp. 2999-3011, 10 December 2018, Published by Routledge, DOI: 10.1080/02664763.2018.1450366, Available: https://www.tandfonline.com/doi/full/10.1080/02664763.2018.1450366.