Diagnosis and Classification of Diabetes Mellitus Type 1 and Type 2 Through Super Learner

Nisha A (✉ rushinishu@gmail.com)

BS Abdur Rahman University: B S Abdur Rahman Crescent Institute of Science & Technology

Kavitha G

BS Abdur Rahman Crescent Engineering College: B S Abdur Rahman Crescent Institute of Science & Technology

Research Article

Keywords: Diabetes Mellitus, Machine Learning, Super Learner, Predictive Models

Posted Date: July 27th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-718966/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Diagnosis and Classification of Diabetes Mellitus Type1 and Type2 through Super Learner

Nisha.A^{1*}, Kavitha .G²

¹Department of Computer Applications, B.S Abdur Rahman Crescent Institute of science and Technology, Chennai.

²Department of Information Technology, B.S Abdur Rahman Crescent Institute of science and Technology, Chennai.

*Corresponding Author: rushinishu@gmail.com

Abstract

Diabetes Mellitus (DM) plays a significant role in increasing the associated health problems worldwide by acting as a Comorbid condition. Moreover, it is a progressive illness without severe external symptoms leading to a fatal impact on the human body if left unnoticed or untreated. This research work aims to associate an individual's lifestyle and ethnic background in assessing the risk of Diabetes acting as a comorbid condition. A detailed assessment of lockdown impact with rapid modification in individual's lifestyle due to the pandemic gives specific insight into individuals becoming susceptible to Diabetes Mellitus. An ensemble of ML algorithms is utilized in predicting the risk of individuals turning Diabetic. The ensemble of the ML model is trained on the Pima Indian dataset and Vanderbilt biostatistics diabetes dataset providing the impact of Type 1 diabetes mellitus. The proposed super learner model provides the highest classification accuracy of T1DM & T2DM with 97% compared to an ensemble of algorithms in identifying and classifying the individuals as being susceptible to DM due to the lifestyle and ethnic background.

Keywords: Diabetes Mellitus, Machine Learning, Super Learner, Predictive Models.

1. Introduction

Diabetes mellitus (DM), aka diabetes, is a fatal syndrome and transmitted through genetic traits based on lifestyle, age, and ethnic structure. DM is characterized by high blood glucose due to insufficient insulin secretion resulting in retinopathy, nephropathy, and eye lesions. One in two adults with diabetes are undiagnosed (232 million), and one in five people with diabetes are above 65 years old (136 million) worldwide. One in six live births (20 million) is affected by hyperglycemia in pregnancy, 84% of which have gestational diabetes worldwide.

Over 1.1 million children and adolescents <20 years have T1DM worldwide. One in 11 adults (20-79) years has diabetes (463 million) worldwide.

DM is categorized into T1DM and T2DM with different variations like gestational and pre-diabetic conditions. T1DM affects individuals <30 years due to criteria like smoking, physical dormancy, and hereditary. T2DM affects elderly and middle-aged individuals due to associated illnesses such as A1C, hypertension, high cholesterol. The key indicator of T2DM is Oral Glucose tolerance (OGT) of 75g of glucose administered in a fasted state with high fasting glucose levels of >200mg/dL where average blood glucose level is >126mg/dL. Abnormalities like dyslipidemia, insulin resistance, and hyperinsulinemia complicate the diagnosis and treatment of T2DM. There is no hard rule for determining the appropriate analytical method, and the best algorithm is determined by the relationship between the data and variance. A key analytical challenge that stems out is identifying the relationship between the metabolomics dataset elements with high dimensions and various levels of quantitative relationship.

Recent years have seen a rise in the number of patients with diabetes mellitus at a young age. Therefore, to ensure Diabetes mellitus is understood adequately, it is studied in further detail. The recent decade has seen significant advances in detecting and identifying the characteristics of Diabetes Mellitus. This paper contributes through a diabetes prediction model that can predict and determine the type of diabetes. The experiment results show the improved accuracy in predicting diabetes through practical training using the Pima and Vanderbilt dataset. This study examines health care data that plays a crucial role in many systems such as disease prediction, prevention methods, medical guidance, and urgent medical decision-making to improve health care standards, reduce costs and increase efficiency.

2. Related work

Health care lies in the intersection of science, technology, and human wellbeing, contributing to a well-developed society [2]. Artificial intelligence (AI) contributes to the ease and automates prediction of diseases occurring through genetic traits. The critical drawback of the existing methods is the absence of a well-defined solution to calculating carbohydrates' consumption, mostly manually performed by individual users. It is vulnerable to a mistake that can significantly affect predictive efficiency [1]. Besides, a universal method has not been established to estimate and measure the estimated effect on the BG level of physical

activities, tension, and infections [3]. Besides, a little has been done concerning model portability that can capture patient inter and intra variability. The impact of time lags is not clear between the CGM readings, and BG's actual levels tend [4]. In general, they predict that these developments will facilitate the development of next-generation BG prediction algorithms, which will make a meaningful contribution to the long-awaited, so-called artificial pancreas[5] [6]. The Primary findings included segregation metrics, calibration, and grouping. Relevant factors, level of validity, and intended use of models were used as secondary effects. A meta-analysis of c-indices, subgroup comparisons, meta-regression, reporting bias tests, and sensitivity analyses were carried out. Twelve studies demonstrated their models' possible applicability to Screening for T2DM with a high pooled c-index provided by meta-analysis (0.812).

The studies report a direct link between the diet and increase in the BG from integrating them to achieve better prediction [1]. The problem becomes even more problematic due to the complexity of BG dynamics. The limiting issue presents itself as the BG prediction Algorithm does not consider the effect of uncontrollable parameters. The factors to be considered are age group, BG regions, Time of day, living conditions, and generalizability for developing a universal approach.

Studies of subgroups and meta-regression established heterogeneity causes. Issues of methodological consistency and monitoring were noted. In the group, we find proof of the strong performance of ML models for T2DM prediction. Before they can be used on a scale, changes to methodology, documentation, and validation are required. [7] The fieldwork is based on a review of the Diabetes dataset and how it would make a diabetes forecast using various machine learning algorithms.

This study suggests a smart home health monitoring device to analyze the patient's blood pressure and glucose readings. In the case of any detected abnormality, the healthcare provider is alerted at home. To forecast hypertension and diabetes status, a mixture of conditional decision-making and machine-learning techniques was used. Utilizing the patient's blood pressure and glucose readings, the goal is to forecast hypertension and diabetes status. A framework learned to predict diabetes and hypertension state of a patient employing supervised machine learning classification algorithms. This proposed research introduces a home health monitoring system program with a user-friendly, easy-to-use interactive user interface to diagnose patients' blood pressure and diabetes condition and

submit classified reminders and real-time updates from home to their registered doctor or clinic [8].

By practicing its features in a five-fold and ten-fold cross-validation fashion, we suggest a technique for diagnosing diabetes using a deep neural network. From the UCI machine learning library database, the Pima Indian Diabetes (PID) data collection is extracted. The PID dataset findings reveal that an in-depth learning methodology constructs an auspicious diabetes prediction method with a prediction accuracy of 98.35 percent, and F1 score of 98, and a five-fold cross-validation MCC of 97. In comparison, for ten-fold cross-validation, precision is 97.11 percent, sensitivity is 96.25 percent, and specificity is 98.80 percent. The experimental findings indicate that the proposed system produces positive results in a five-fold cross-validation [4].

We use the support vector machine (SVM), and the K-nearest neighbor algorithms on the dataset gathered from a longitudinal analysis called the Framingham Heart Study to create the prediction models. The algorithm for the Synthetic Minority Oversampling Technique first balanced the dataset. The SVM algorithm model to correctly predict CVD's average precision prevalence due to T2DL as 96.5 percent. The recall rate was at an average of 89.8 percent after adjusting the parameters and practicing for 1000 times. Similarly, the KNN algorithm was also used to train the dataset, and the recall rate even hit 92.9%. The model's benefits are that it can forecast both the probability of T2DL and CVD growth concurrently with high accuracy. Model forecasts without the costly and cumbersome examination of oral glucose tolerance. Following testing on the Framingham Heart Study dataset, the model yielded high-performance outcomes [5].

Diabetes is emerging at an unprecedented pace every day across the world and can cause certain long-term complications such as affecting the skin, heart, liver, brain, feet, nerves. Before it becomes one of the biggest human challenges, it is crucial to find an efficient means of forecasting diabetes. If we take care of it properly, It is possible to take steps at an early stage to take care of diabetes. In this study, 340 cases with 26 characteristics of patients that have already been affected by diabetes with different symptoms, classified by two categories, namely Typical and Non-Typical Symptoms, were obtained. This research aims to correctly classify the type of Diabetes Mellitus using the Random Forest algorithm, an Ensemble Computer. We also achieved 98.24 percent accuracy for seeds 2 and 97.94 percent accuracy for sources 1 and 3 [9].

To build a high-precision model for the estimation of T2DM at various onsets. It is possible to integrate clustering and classification approaches better to diagnose diabetes at an earlier level, without removing findings with missed values at the previous stage and minimizing irrelevant features to get the most associated features during data collection. Using K-means clustering accompanied by running the noise reduction process, we apply a noise reduction-based technique. To remove the unknown secret parts of the dataset and for more detailed performance, random forest and XGBoost classifiers. Through benchmarking prediction accuracy of the model is compared against up-to-date predictive models and standard classification algorithms. Our T2ML model, with an accuracy of 97.53 percent by 10-fold cross-validation, compared to other studies recorded by other researchers in the literature and over several traditional classification algorithms [10].

A fast prediction algorithm for constructing a super learner uses candidate learners with V-fold cross-validation using selected weights [super learner 2007]. The author demonstrates various data generating distributions adaptivity with their proposed super learner. The proposed learner defines any parameter that can generalize as a minimizer of loss function. In this section of the paper, we have discussed a stacked ensemble approach utilizing three benchmark datasets.

3. Super learner Methodology

Super learning is a loss based supervised learning method utilizing a cross validation-based learning approach. The Meta learner trains on the results of the base learners to predict the output of the ensemble.

Fig 1: Flow Diagram of Super learner [refer super learner 2007]

The observations in the training set are randomly split into $X_i \rightarrow$ feature values and $Y_i \rightarrow$ class value. The learners are indexed as 'k' index and represented as a function in equation 1,

$$\hat{\Psi}(P_n) \equiv \hat{\Psi}_{K(P_n)}(P_n) \quad (1)$$

Where $P_n \rightarrow$ probability distributions, $\Psi \rightarrow$ parameter space, $K_n \rightarrow$ collection of learners.

The learners B_L predicted values are cross-validate the n^{th} learner,

$$\hat{K}(P_n) \equiv \arg \min_k E_{B_n} \sum_{i, B_n(i)=1} (Y_i - \hat{\Psi}_k(P_{n, B_n}^0)(X_i))^2 \quad (2)$$

Where $B_n \in \{0,1\} \rightarrow$ random binary vector, $\hat{K}(P_n) \rightarrow$ Cross-validation selector.

The cross-validated risk criteria is defined as in equation 3,

$$R_{CV}(\beta) \equiv \sum_{i=1}^n (Y_i - m(Z_i|\beta))^2 \quad (3)$$

3.1 Base Learners

The basic base learners are Logistic Regression, Decision Tree Classifier, SVC, Gaussian NB, K Neighbors Classifier, AdaBoost Classifier, Bagging Classifier, Random Forest Classifier, Extra Trees Classifier, and Super Learner.

DELM deep extreme learning machine-learned through statistical measures

$$Misrate = \frac{(O_1/T_0 + O_0/T_1)}{(T_0 + T_1)} \quad (4)$$

$$Accuracy = \frac{(O_0/T_0 + O_0/T_1)}{(T_0 + T_1)} \quad (5)$$

$$Positive Prediction value = \frac{O_0/T_1}{(O_0/T_0 + O_0/T_1)} \quad (6)$$

$$Negative Prediction value = \frac{O_0/T_0}{(O_0/T_0 + O_1/T_0)} \quad (7)$$

$$Specificity = \frac{O_0/T_0}{(O_0/T_0 + O_0/T_1)} \quad (8)$$

4. Experiments Result

This section provides the overall view of the experimental setup that we used. The choice was the python programming language with its vast library for implementing the super learner.

4.1 Data set

This study's advantage includes having data from a large repository to forecast whether a patient has diabetes or not. The Pima Indian diabetes dataset (PIDD) used in this study is obtained from the UCI repository supported by Diabetes, Digestive, and Kidney Disorders National Institute and the other from the biostatistics program (Vanderbilt dataset). The PIDD contains 768 observations on 9 variables, all patients being females at least 21 years old with one variable target outcome and predictor variables like pregnancies, glucose, blood pressure, insulin, BMI, diabetes pedigree function, and age. The Biostatistics study program comprised 1046 subjects, of which 403 subjects are selected based on the age group with BG levels and physical activity. The dataset consists of nineteen variables out of 1046. Only 403 subjects

were interviewed to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors. The second set contains 60 out of 390 African-American Patients with A1c > 6.5 were classified as diabetic in the Vanderbilt dataset. The objective is to predict diabetes using the demographic variables. T2DM is also associated with hypertension; they may both be part of the syndrome x; the 403 subjects were the ones who screened for diabetes glycosylated hemoglobin > 7.0 is usually taken as a positive diagnosis of diabetes. The objective is to predict the onset of diabetes based on diagnostic measures. The schematic overview of the dataset PIDD is shown in figure 1 below

Fig 2: Overview of PIDD dataset

A schematic diagram representing the relationship between the various attributes of the PIDD dataset in Figure 2 & 3 as follows,

Fig 3: Plot of Pima Indian Dataset

Fig 4: Plot of Pima Indian Dataset

As seen in Figures 2 & 3, the correlation between the features plotted for the Pima Indian Dataset. Figure 4 illustrates the feature importance of the T2DM dataset with that of permutation importance of Test cases.

Fig 5: Feature importance

4.2 Result and Discussion

An ensemble of classification models, in combination with Super learner, is utilized for assessing the given dataset. We have compared all the data we received, and the best performance of super learner is based on the results. The statistical data analysis package “scikit-learn” in python is utilized for Exploratory Data Science (EDS). Studies of the data demonstrate a general pattern of increased risk for diabetes in age groups with limited physical activity.

The super learner model allows for the selection and fit algorithms based on their performance and make predictions. The analysis process steps are that the base models provide input to the Meta model, which predicts the output. The 768 samples of the training dataset are then provided as the Meta model input as 768 rows and 19 columns. The base learners’ prediction is concatenated into vectors and given as input to the metamodel shown as confusion matrix in figure 6.

Fig 6: Confusion matrix of predicted vs. actual values

The super learner classifies the dataset as class labels or class probabilities with linear and nonlinear classification models. The stochastic algorithms predict with high accuracy, as shown in figure 7.

Fig 7: Cross-entropy loss

4.3 Evaluation

Statistical methods are applied sensitivity, specificity, and accuracy to measure the performance of the proposed model. The measures of the classifier's evaluation are as shown in equations (1) to (3)

$$accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (9)$$

$$sensitivity = \frac{TP}{(TP + FN)} \quad (10)$$

$$Specificity = \frac{TN}{(TN + TP)} \quad (11)$$

Where TP = True Positive, TN = True Negative, FN = False Negative, FP = False Positive

The data is imbalanced as the datasets are combined, and specific fields are extracted and utilized. The model displays an overall accuracy of 79.17%.

4.4 Principle findings

The proposed model stacks base learners in layers propagating to the next layer and improves the predictions through ensemble combinations. The proposed model is trained on three different datasets displaying varying sensitivity, specificity, accuracy, and AUC in table 1. As the Area provides the best prediction aggregate under the ROC curve (AUC), which is scale-invariant where the model's probability of positive examples rather than negative examples.

Fig 8: Glucose vs. Blood pressure

Figure 8 shows the relationship between glucose and blood pressure for the diabetic and healthy individuals.

Figure 9 shows the relationship between the glucose and BMI of the healthy individuals and the diabetic individuals.

Fig 9: Glucose vs. BMI

Figure 10 shows the T2DM patients with the insulin levels for the healthy and diabetic patients as the secretion of insulin can vary as the age factor increases unless there occurs a genetic trait of the individual patient.

Fig 10: Insulin for diabetic and healthy

Figure 11 shows the chart of distribution of the individuals with glucose levels ≤ 105 and having BMI ≤ 30 for the diabetic and healthy individuals. The PIDD dataset patients are

classified in to 268 as diabetic and 500 as healthy with proper management within the required levels.

Fig 11: Distribution by the outcome

Figure 12 provides a detailed view of diabetic and healthy individuals' insulin levels. The percentage of abnormal blood glucose levels needs to be regulated, resulting in long-term and short-term health complications. Figure 13 provides a distribution of the healthy and diabetic individuals' insulin levels as 17.5% and 8.4%, respectively.

Fig 12: Distribution by Insulin

Fig 13: Distribution by Insulin target

Figure 14 provides a detailed view of diabetic and healthy individuals' blood pressure where the percentage of abnormal blood glucose levels needs to be regulated, resulting in long-term and short-term health complications.

Fig 14: Distribution by blood pressure

Figure 15 provides a distribution of the blood pressure between the healthy and diabetic individuals as 33.6% and 23%, respectively.

N10 distribution by target
BloodPressure < 80

Fig 15: Distribution by blood pressure

Fig 16: Correlation matrix of PIDD dataset

The table 1 shows the comparison of classifier accuracy through the proposed model.

Fig 17: Model Accuracy and Loss for predicting DM

Table 1: Classifier accuracy for diabetes datasets

Classifier	PIDD	T1DM	T2DM
Logistic Regression	93.600	94.460	96.560
DecisionTreeClassifier	72.200	71.760	76.040
SVC	95.800	94.300	97.320
GaussianNB	96.200	94.500	97.680
KNeighborsClassifier	91.000	87.300	93.560
AdaBoostClassifier	89.600	88.640	92.280
BaggingClassifier	84.200	82.340	86.760
RandomForestClassifier	83.400	82.120	86.640
ExtraTreesClassifier	81.200	81.460	85.600
Super Learner	96.400	94.660	97.720

Fig 18: Comparison of an ensemble of classifiers algorithms

The following table shows the RMSE for scores for the base learner and the super learner.

Table 2: Comparison of RMSE score for super learner

	score-m	score-s	ft-m	ft-s	pt-m	pt-s
layer-1 AdaBoostClassifier	0.91	0.02	0.85	0.05	0.05	0.01
layer-1 BaggingClassifier	0.84	0.04	0.30	0.01	0.01	0.00
layer-1 DecisionTreeClassifier	0.74	0.04	0.04	0.01	0.00	0.00
layer-1 ExtraTreesClassifier	0.83	0.06	0.17	0.06	0.01	0.00
layer-1 GaussianNB	0.96	0.02	0.02	0.01	0.00	0.00
layer-1 KNeighborsClassifier	0.92	0.02	0.01	0.00	0.03	0.01
layer-1 logistic regression	0.96	0.02	0.01	0.01	0.00	0.00
layer-1 RandomForestClassifier	0.84	0.03	0.14	0.01	0.01	0.00
layer-1 svc	0.97	0.02	0.12	0.00	0.00	0.00
Super Learner: 97.400						
	score-m	score-s	ft-m	ft-s	pt-m	pt-s
layer-1 AdaBoostClassifier	0.93	0.02	1.89	0.02	0.08	0.02
layer-1 BaggingClassifier	0.87	0.02	1.80	0.08	0.03	0.05
layer-1 DecisionTreeClassifier	0.76	0.02	0.28	0.02	0.00	0.00
layer-1 ExtraTreesClassifier	0.87	0.02	0.31	0.25	0.01	0.01
layer-1 GaussianNB	0.98	0.01	0.05	0.04	0.02	0.03
layer-1 KNeighborsClassifier	0.94	0.01	0.08	0.03	0.28	0.06
layer-1 Logistic regression	0.96	0.01	1.32	0.26	0.01	0.01
layer-1 RandomForestClassifier	0.87	0.02	0.22	0.02	0.00	0.00
layer-1 svc	0.98	0.01	3.64	0.06	0.02	0.00

Super Learner: 98.160							
	score-m	score-s	ft-m	ft-s	pt-m	pt-s	
layer-1	AdaBoostClassifier	0.93	0.01	3.99	0.05	0.09	0.03
layer-1	BaggingClassifier	0.87	0.01	4.57	0.14	0.13	0.24
layer-1	DecisionTreeClassifier	0.76	0.02	0.68	0.04	0.00	0.00
layer-1	ExtraTreesClassifier	0.88	0.01	0.63	0.55	0.04	0.05
layer-1	GaussianNB	0.98	0.01	0.08	0.10	0.03	0.07
layer-1	KNeighborsClassifier	0.94	0.01	0.20	0.07	1.28	0.22
layer-1	Logistic regression	0.97	0.01	0.52	0.18	0.00	0.00
layer-1	RandomForestClassifier	0.88	0.02	0.49	0.03	0.00	0.00
layer-1	svc	0.98	0.01	15.66	0.18	0.08	0.01
Super Learner: 97.640							

Table 3: Performance of Proposed Techniques

Datasets	Sensitivity	Specificity	Accuracy	AUC
PIDD	89.90	77.5	79.17	0.875
T1DM	93.12	55.67	97.57	0.787
T2DM	90.65	93.65	68.52	0.973

We have used the ensemble-based technique in PIDD datasets with 10-fold cross-validation. They obtained 89% sensitivity, 71.5% specificity, 79.17% accuracy, and 0.875 AUC. Among the classifiers, super learners obtained the highest classification accuracy, with 97% on the datasets. The measured and predicted values are plotted for logistic regression providing a visual aspect of the proposed model's identified data. For the T1DM datasets with 10-fold cross-validation, they obtained 89% sensitivity, 71.5% specificity, 79.17% accuracy, and 0.875 AUC. Among the classifiers, super learners obtained the highest classification accuracy, with 98% on the datasets. The measured and predicted values are plotted for logistic regression providing a visual aspect of the identified data by the proposed model.

Fig 19: ROC curve analysis for PIDD dataset

For the T2DM datasets with 10-fold cross-validation, they obtained 89% sensitivity, 71.5% specificity, 79.17% accuracy, and 0.875 AUC. Among the classifiers, super learners obtained the highest classification accuracy, with 97% on the datasets. The measured and predicted values are plotted for logistic regression providing a visual aspect of the identified data by the proposed model.

Fig 20 : Measured and Predicted values

5. Conclusion

ML's essential task is to identify patterns from the available data with limited error and high prediction accuracy. Determining the best algorithm for the specified task depends on the researcher selecting them. This paper provides a thorough study of the blood glucose level through automatic diabetes diagnosis and detection methods. A combination of PIDD, Kaggle, and CGM publicly available datasets is utilized for DM detection. The emphasis of this study is to determine the type of DM based on the combination of Datasets. The super learner is used to identifying the optimal algorithm for a specific classification task in classifying Diabetes as T1DM and T2DM. The proposed model presents analysis and predicting of DM type I and II with high accuracy. The performance measure for the proposed approach reflects better accuracy. Compared with past practices, the proposed method encapsulates both the statistical and ML, providing much more benefits balancing the imbalanced data through the class weight method. Furthermore, a thorough analysis of datasets is made in research areas that need improvement possible detection of genetic ancestry, incorporation of intelligent diagnosis and analysis, and personalized DM disease pathology. Therefore, in the future, the study may be extended to address the limitations.

Declarations

Funding: No Funding support.

Conflict of Interests/Competing Interests: No conflict to declare by the authors.

Availability of Data and Material: Not Applicable.

Code Availability: Not Applicable.

Author's Contributions: Nisha. A designed the Methodology, Implemented Classification of Diabetes and written the manuscript, Kavitha.G - Resources Validation and verification presented in the manuscript.

References

- [1] A. Z. Woldaregay *et al.*, "Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes," *Artif. Intell. Med.*, vol. 98, pp. 109–134, 2019.
- [2] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Comput. Struct.*

Biotechnol. J., vol. 15, pp. 104–116, 2017.

- [3] B. P. Nguyen *et al.*, “Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records,” *Comput. Methods Programs Biomed.*, vol. 182, p. 105055, 2019.
- [4] S. I. Ayon, M. Islam, and others, “Diabetes Prediction: A Deep Learning Approach,” *Int. J. Inf. Eng. Electron. Bus.*, vol. 11, no. 2, 2019.
- [5] L. Miao, X. Guo, H. T. Abbas, K. A. Qaraqe, and Q. H. Abbasi, “Using Machine Learning to Predict the Future Development of Disease,” in *2020 International Conference on UK-China Emerging Technologies (UCET)*, 2020, pp. 1–4.
- [6] W. K. Lee, A. Forbes, R. T. Demmer, C. Barton, J. Enticott, and K. De Silva, “Use and performance of machine learning models for type 2 diabetes prediction in community settings: a systematic review and meta-analysis,” *Int. J. Med. Inform.*, p. 104268, 2020.
- [7] L. Beqiri, A. Velinov, B. Fetaji, L. Loku, A. Buçuku, and Z. Zdravev, “Analysis of Diabetes Dataset,” in *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, 2020, pp. 309–314.
- [8] S. P. Chatrati *et al.*, “Smart home health monitoring system for predicting type 2 diabetes and hypertension,” *J. King Saud Univ. Inf. Sci.*, 2020.
- [9] M. T. Islam, M. Raihan, F. Farzana, N. Aktar, P. Ghosh, and S. Kabiraj, “Typical and Non-Typical Diabetes Disease Prediction using Random Forest Algorithm,” in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, pp. 1–6.
- [10] S. Albahli, “Type 2 Machine Learning: An Effective Hybrid Prediction Model for Early Type 2 Diabetes Detection,” *J. Med. Imaging Heal. Informatics*, vol. 10, no. 5, pp. 1069–1075, 2020.
- [11] T. R. Gadekallu *et al.*, “Early detection of diabetic retinopathy using PCA-firefly based deep learning model,” *Electronics*, vol. 9, no. 2, p. 274, 2020.
- [12] A. Rghioui, J. Lloret, S. Sendra, and A. Oumnad, “A Smart Architecture for Diabetic Patient Monitoring Using Machine Learning Algorithms,” in *Healthcare*, 2020, vol. 8, no. 3, p. 348.

- [13] A. Yaganteeswarudu, "Multi Disease Prediction Model by using Machine Learning and Flask API," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 2020, pp. 1242–1246.
- [14] M. F. Faruque, A. Asaduzzaman, S. M. M. Hossain, M. H. Furhad, and I. H. Sarker, "Predicting Diabetes Mellitus and Analysing Risk-Factors Correlation," *EAI Endorsed Trans. Pervasive Heal. Technol.*, vol. 5, no. 20, p. e7, 2020.
- [15] M. T. Islam, M. Raihan, N. Aktar, M. S. Alam, R. R. Ema, and T. Islam, "Diabetes Mellitus Prediction using Different Ensemble Machine Learning Approaches," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, pp. 1–7.
- [16] K. S. Ryu, S. W. Lee, E. Batbaatar, J. W. Lee, K. S. Choi, and H. S. Cha, "A Deep Learning Model for Estimation of Patients with Undiagnosed Diabetes," *Appl. Sci.*, vol. 10, no. 1, p. 421, 2020.
- [17] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," *Heal. Inf. Sci. Syst.*, vol. 8, no. 1, p. 7, 2020.
- [18] J. Chaki, S. T. Ganesh, S. K. Cidham, and S. A. Theertan, "Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review," *J. King Saud Univ. Inf. Sci.*, 2020.
- [19] D. R. Nemade and R. K. Gupta, "Diabetes Prediction using BPSO and Decision Tree Classifier," in *2nd International Conference on Data, Engineering and Applications (IDEA)*, 2020, pp. 1–5.
- [20] G. Alfian *et al.*, "Deep Neural Network for Predicting Diabetic Retinopathy from Risk Factors," *Mathematics*, vol. 8, no. 9, p. 1620, 2020.
- [21] J. D. Bodapati *et al.*, "Blended Multi-Modal Deep ConvNet Features for Diabetic Retinopathy Severity Prediction," *Electronics*, vol. 9, no. 6, p. 914, 2020.
- [22] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," *Procedia Comput. Sci.*, vol. 167, pp. 706–716, 2020.
- [23] T. Daghistani and R. Alshammari, "Comparison of Statistical Logistic Regression and RandomForest Machine Learning Techniques in Predicting Diabetes," *J. Adv. Inf.*

Technol. Vol, vol. 11, no. 2, 2020.

- [24] M. Rahman, D. Islam, R. J. Mukti, and I. Saha, "A deep learning approach based on convolutional LSTM for detecting diabetes," *Comput. Biol. Chem.*, vol. 88, p. 107329, 2020.
- [25] L. Wang, X. Wang, A. Chen, X. Jin, and H. Che, "Prediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model," in *Healthcare*, 2020, vol. 8, no. 3, p. 247.
- [26] D. J. Reddy *et al.*, "Predictive machine learning model for early detection and analysis of diabetes," *Mater. Today Proc.*, 2020.
- [27] K. Vidhya and R. Shanmugalakshmi, "Deep learning based big medical data analytic model for diabetes complication prediction," *J. Ambient Intell. Humaniz. Comput.*, pp. 1–12, 2020.
- [28] M. J. Mehdi, N. Srinivasrao, and A. Sireesha, "Detection and prognosis of diabetes based on data science techniques," *Mater. Today Proc.*, 2020.
- [29] F. Nusrat, B. Uzba\cs, and Ö. K. Baykan, "Prediction of Diabetes Mellitus by using Gradient Boosting Classification," *Avrupa Bilim ve Teknol. Derg.*, pp. 268–272.
- [30] K. Driss, W. Boulila, A. Batool, and J. Ahmad, "A Novel Approach for Classifying Diabetes' Patients Based on Imputation and Machine Learning," in *2020 International Conference on UK-China Emerging Technologies (UCET)*, 2020, pp. 1–4.
- [31] N. Pradhan, G. Rani, V. S. Dhaka, and R. C. Poonia, "Diabetes prediction using artificial neural network," in *Deep Learning Techniques for Biomedical and Health Informatics*, Elsevier, 2020, pp. 327–339.
- [32] K. Saiti, M. Macaš, L. Lhotská, K. Štechová, and P. Pithová, "Ensemble methods in combination with compartment models for blood glucose level prediction in type 1 diabetes mellitus," *Comput. Methods Programs Biomed.*, vol. 196, p. 105628, 2020.
- [33] G. Mainenti, L. Campanile, F. Marulli, C. Ricciardi, and A. S. Valente, "Machine Learning Approaches for Diabetes Classification: Perspectives to Artificial Intelligence Methods Updating," in *IoTBDs*, 2020, pp. 533–540.
- [34] A. U. Haq *et al.*, "Intelligent Machine Learning Approach for Effective Recognition of

- Diabetes in E-Healthcare Using Clinical Data,” *Sensors*, vol. 20, no. 9, p. 2649, 2020.
- [35] B. Ljubic *et al.*, “Predicting complications of diabetes mellitus using advanced machine learning algorithms,” *J. Am. Med. Informatics Assoc.*, vol. 27, no. 9, pp. 1343–1351, 2020.
 - [36] R. Unnikrishnan and A. Misra, “Infections and diabetes: Risks and mitigation with reference to India,” *Diabetes Metab. Syndr. Clin. Res. Rev.*, 2020.
 - [37] F. Hassan and M. E. Shaheen, “Predicting Diabetes from Health-based Streaming Data using Social Media, Machine Learning and Stream Processing Technologies.”
 - [38] R. Alshammari, N. Atiyah, T. Daghistani, and A. Alshammari, “Improving Accuracy for Diabetes Mellitus Prediction by Using Deepnet,” *Online J. Public Health Inform.*, vol. 12, no. 1, 2020.
 - [39] L. Zhang *et al.*, “Predicting the Development of Type 2 Diabetes in a Large Australian Cohort Using Machine-Learning Techniques: Longitudinal Survey Study,” *JMIR Med. informatics*, vol. 8, no. 7, p. e16850, 2020.
 - [40] G. Emmanuel, G. G. Hungilo, and A. W. R. Emanuel, “Performance Evaluation of Machine Learning Classification Techniques for Diabetes Disease.”
 - [41] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, “Logistic Regression based Feature Selection and Classification of Diabetes Disease using Machine Learning Paradigm.”
 - [42] P. A. Senior *et al.*, “Pharmacologic Glycemic Management of Type 2 Diabetes in Adults: 2020 Update--The User’s Guide,” *Can. J. Diabetes*, vol. 44, no. 7, pp. 592–596, 2020.