International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Enhanced Diagnosis and Classification of Type 1 and Type 2 Diabetes Mellitus with Super Learner

Nisha A.[1*], Kavitha G.[2]

Submitted: 15/01/2024 **Revised**: 23/02/2024 **Accepted**: 01/03/2024

Abstract: Diabetes Mellitus (DM) plays a huge part in expanding the related medical issues overall by going about as a Comorbid condition. Besides, it is an ever-evolving disease without serious outer side effects prompting a deadly effect on the human body whenever left inconspicuous or untreated. This study aims to assess the risk of diabetes occurring as a comorbid condition by relating an individual's lifestyle and ethnic background. A detailed analysis of the lockdown's impact on people's rapid lifestyle changes brought on by the epidemic provides clear insight into how persons with diabetes mellitus become powerless. The chance of someone developing diabetes is predicted using a collection of machine learning computations. The Pima Indian dataset and the Vanderbilt biostatistics diabetes dataset, which show the effects of Type 1 diabetes mellitus, are used to create the ML model. The suggested super learner model produces the most remarkable classification accuracy of 97% for T1DM and T2DM when compared to an ensemble of algorithms in identifying and categorising people as being susceptible to DM because of their ethnic heritage and way of life.

Keywords: *Diabetes Mellitus; Machine Learning; Super Learner; Predictive Models*

1. Introduction

As The no free lunch theorem, which states that no single algorithm is optimal for all problems, is a well-known result. The explanation for this, intuitively, is that machine-learning algorithms learn in a certain way. Machine Learning is ultimately about minimizing a loss function in a given context. We have implicitly constrained the kind of patterns an algorithm should learn until we have decided what it can learn. Because of this diversity of learning, various algorithms would be able to catch different facets of the signal for a given problem. They often catch the same signal dynamics, but more often than not, their predictive power complements each other. In other words, if one learning algorithm fails to find a successful signal, another does. Ensembles allow us to take advantage of this diversity in disease prediction communicated through hereditary characteristics in light of way of life, age and ethnic construction.

High blood glucose brought on by inadequate insulin secretion, which causes retinopathy, nephropathy, and ocular diseases, is the hallmark of diabetes mellitus (DM). Worldwide, 136 million adults over 65 have diabetes, while another 232 million adults with the disease are undiagnosed. Twenty million live births, or one in six cases, have hyperglycemia during pregnancy; of these, 84% develop gestational diabetes globally. Globally, T1DM affects more than 1.1 million children and teenagers under the age of 20. Worldwide, 463 million adults (20–79 years old) have diabetes, or one in eleven.

Diabetes mellitus is classified as either T1DM or T2DM, with distinct variations such as gestational and pre-diabetic situations. People under 30 years old are affected by T1DM because of factors include physical inactivity, smoking, and genetics.

. A1C, hypertension, and high cholesterol are among the related disorders that cause type 2 diabetes (T2DM) in middle-aged and older adults. The main indicator of type 2 diabetes is oral glucose tolerance (OGT), which is determined by giving 75g of glucose during a fast with high fasting glucose levels of >200 mg/dL when the normal blood glucose level is >126 mg/dL.

Diagnosing and treating type 2 diabetes is made more difficult by abnormalities such as dyslipidaemia, insulin resistance, and hyperinsulinemia. The link between the data and variance determines the optimal algorithm, and there is no hard rule for choosing the right analytical approach. Deciding the connection between the things in the metabolomics dataset with high aspects and various degrees of quantitative affiliation is a critical scientific trouble that emerges.

The number of young individuals with diabetes mellitus has increased in recent years. Therefore, more research is done on diabetes mellitus to make sure it is well understood. Significant progress has been made in the last ten years in detecting and characterising the symptoms of diabetes mellitus.

*[1] Research Scholar Department of Computer Applications, B.S Abdur Rahman Crescent Institute of science and Technology, Chennai, Tamilnadu - 600048, India. Email: rushinishu@gmail.com**
[2] K Associate Professor Departrrent of lnformation Technology, B.S Abdur Rahman Crescent Institute of science and Technology, Chennai, Tamilnadu – 600048, India. Email: gkavitha.78@crescent.education.

The first reason for this work is to determine the forecast of two models by taking the average of these forecasts and constructing an ensemble. Better still, use linear regression to figure to see what the optimal (linear) mix of the two assumptions will be. Our initial estimators, which were fitted to the input data, are referred to as base learners, and the algorithm used to find the best combination is referred to as the meta learner. Of necessity, nothing stops us from discovering non-linear combinations, and these ensembles often outperform linear combinations in terms of predictive ability.

In this work with the use of a diabetes prediction model sexperiment's outcomes demonstrate how practical training using the Pima and Vanderbilt datasets increased the accuracy of diabetes prediction. The goal of this project is to better understand health care data, which is essential to various systems including disease prediction, preventative strategies, medical advice, and emergency medical decision-making**.**

2. Related Work

A well-developed civilization is facilitated by the convergence of research, technology, and health care. [2]. Genetic trait-based disease forecasting is made easier and more automated by artificial intelligence (AI). The absence of a well-defined methodology for calculating carbohydrate consumption—which is usually done manually by individual users and prone to error, which can have a significant impact on predictive efficiency—is the primary flaw in the current approaches [1].Furthermore, there is currently no standard technique for estimating and quantifying the estimated effects of stress, illnesses, and physical activity on the BG level [3]. Additionally, not much research has been done on model portability that can capture intra- and inter-patient diversity in patients. The effect of time gaps between the real levels of BG and the CGM measurements is unclear [4]. They generally anticipate that these advancements will speed up the creation of next-generation BG prediction algorithms, which will significantly advance the long-awaited "artificial pancreas" [5].[6].Secondary impacts included the level of validity, the relevance of the components, and the models' intended application. Subgroup comparisons, reporting bias tests, meta-regression, c-indices, and sensitivity analyses were all done. Based on a meta-analysis, twelve studies showed that their models may be used for T2DM screening, with a high pooled c-index of 0.812.

The research that is now available shows a clear correlation between diet and an increase in BG from combining the two to improve prediction [1]. The intricacy of BG dynamics makes the situation considerably more difficult. The BG prediction algorithm's failure to account for the impact of uncontrollable parameters is the limiting factor. Subgroup analyses and meta-regression studies determined the sources of heterogeneity. Methodological consistency and monitoring issues were identified. Look for evidence in the group demonstrating the ML models' excellent performance in T2DM prediction. Methodology, documentation, and validation improvements must be made before they can be implemented on a large scale. [7] The examination of the Diabetes dataset and the potential applications of several machine learning techniques for diabetes prediction form the basis of the fieldwork.

An intelligent home health monitoring system to evaluate the patient's glucose and blood pressure values. The healthcare provider is notified at home in the event that any irregularity is found. A combination of machine-learning and conditional decision-making techniques was utilised to predict the status of diabetes and hypertension. The objective is to predict the patient's condition for hypertension and diabetes by using their blood pressure and glucose values. A system that uses supervised machine learning classification algorithms to forecast a patient's status of diabetes and hypertension.With a user-friendly, interactive user interface, the author presents a programme for home health monitoring that allows patients to diagnose their blood pressure and diabetes and send real-time information and classified reminders to their registered physician or clinic from the comfort of their own home [8].

Using five- and ten-fold cross-validation, the author proposes a deep neural network-based method for diabetes diagnosis. The Pima Indian Diabetes (PID) dataset is taken from the UCI machine learning library database. An auspicious diabetes prediction system is constructed using an in-depth learning methodology, according to the PID dataset findings. By contrast, the accuracy, sensitivity, and specificity of ten-fold cross-validation are 97.11 percent, 96.25 percent, and 98.80 percent, respectively. According to the experimental data, the suggested system performs well in five-fold cross-validation [4].

Miao et al., (2020) discusses 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 dataset was first balanced by the Synthetic Minority Oversampling Technique programme. The model developed by the SVM technique was able to predict the average precision prevalence of CVD attributable to T2DL as 96.5 percent and the average recall rate as 89.8 percent after changing the parameters and training 1000 times. The model's advantages include its high accuracy capacity to predict the likelihood of concomitant increases in CVD and T2DL.Following testing on the Framingham Heart Review dataset, the model yielded superior presentation results [5].

Diabetes can lead to long-term issues that damage the skin, heart, liver, brain, foot, and nerves. It is spreading around the world at an unprecedented rate every day. It is imperative to develop an effective method of predicting diabetes before it becomes one of the greatest human challenges. Diabetes is something that can be managed early on if it is properly taken care of. In this study, 340 cases with 26 features of individuals with diabetes who had previously experienced various symptoms were categorised into two groups: Typical and Non-Typical Symptoms. The Random Forest algorithm, an Ensemble Computer, is used to categorise the kind of Diabetes Mellitus. Islam et al., (2020) achieved 98.24 % accuracy for seeds 2 and 97.94 percent accuracy for sources 1 and

3 [9].

A high-precision model was developed by Albahli (2020) to estimate T2DM at different onsets. By reducing superfluous variables to acquire the most related features during data collection or eliminating findings with missing values from a previous stage, a better integration of clustering and classification approaches can result in an earlier diagnosis of diabetes. utilising a noise reduction-based approach and K-means clustering in conjunction with the noise reduction procedure. XG Boost and random forest classifiers are used for more detailed performance and to eliminate the dataset's unknown hidden sections. By benchmarking, the model's prediction accuracy is evaluated against the most recent predictive models and accepted categorization techniques. T2ML model, which achieved a 97.53 percent accuracy rate by 10-fold cross-validation, outperformed various conventional classification algorithms as well as other experiments documented by other researchers in the literature. Using V-fold cross-validation with chosen weights, a quick prediction technique is used to create super learners [super learner 2007]. The author uses their suggested super learner to adaptively generate different data generating distributions [10].

3. Methodology

The suggested learner allows for the definition of any parameter that may be defined as a minimizer of the loss function. We have covered a stacked ensemble method using three benchmark datasets in this part.

Let X = {x^((1)),x^((2)),…,x^((n)) define the set of 1046 observations of some input data with associated output y = { y^((1)),y^((2)),…,y^((n))}. For each observation of 746 features, x^i = {x^1,x^2,…,x^n} we predict the expected output of y^i.

Suppose the observations are fitted into the set of models l_1,l_2,l_3,….,l_k. the ensemble of the models are combined into a library of base learners L={l_1,l_2,…,l_k }. Observations are given as such in the library of the base learners x^((i)), is used to determine the set of predictions. Each of the base learners l_j ∈L outputs the prediction, p_j^((i))=l_j (x^((i)))thereby stacking the predictions into a vector of predictions and the proposed workflow is given in figure 1.

$$\boldsymbol{p}^{(i)} = \left(p_1^{(i)}, p_2^{(i)}, \dots, p_k^{(i)}\right) \tag{1}$$

It's worth noting that p^i and x^iboth describe a collection of features associated with an output y^i. The distinction is that the features in p^i are each base learner's predictions. It's now simple to train a meta learner g on a series of predictions rather than the original results.

$$\hat{y}^{(i)} = g\left(\boldsymbol{p}^{(i)}\right) \tag{2}$$

Assume that g is a linear regression model, g(p)=wp, to get a sense of what g is learning. We must find the coefficients that minimize the number of squared errors over the training set T to suit the meta learner.

$$\boldsymbol{w}^{*} = \arg\min_{\boldsymbol{w}} \sum_{i \in \mathcal{T}} \left(y^{(i)} - \boldsymbol{w}^{\top}\boldsymbol{p}^{(i)}\right)^2 \tag{3}$$

For our super learner model, ∑_jw_j =1 is a reasonable restriction. The ensemble serves as a model averaging method in this scenario. Although this restriction is not recommended in practice because g can still learn here to demonstrate the learning dynamics. The cost term can be written as in the equation for any given observation i in equation (2).

$$y^{(i)} - \boldsymbol{w}^{\top}\boldsymbol{p}^{(i)} = y^{(i)} \sum_{j=1}^{k} w_j - \sum_{j=1}^{k} w_j p_j^{(i)} = \sum_{j=1}^{k} w_j \left(y^{(i)} - p_j^{(i)}\right) \tag{4}$$

Now, y^((i))-p_j^((i))is the j-th base learner's prediction error, so g is learning to compensate for each learner's errors by selecting a linearly optimal combination of predictions, as we saw earlier in equation 3.Assume g is a non-linear construct in general and G learns to correct mistakes locally in this situation. That is it learns the best combination of predictions in each subdivision of the prediction space. This can be very useful since it only takes one or two models to have predictive power over the whole function space. Instead, one model can catch one series of signals and another can capture a different kind of signal. The meta learner's job is then to figure out which series of models it can trust in which region.

Equation (2) also reveals that g can learn to compensate for the particular kind of prediction error embedded in the training set T. Let's say using the same data to match the base learners first, then the meta learners. In this case, g learns to adjust for training errors, but it will still correct for test errors during testing. Models usually do better on the training set than on the test set; in this situation, the gap would be exacerbated by the mistaken belief that the base learners are more reliable than they are. Worse, it will come to depend on the base learner who overfits the most, rather than the one who extrapolates the most.

- Let $X_L^{(train)}$ denote the training set of the base learners.
- Let $X_g^{(train)}$ denote the training set of the meta learners.
- Fitting the base learners to predict $X_g^{(train)}$ thereby using these predictions to train the meta leaner as the sets are disjoint the models will fit on all the training data and base learner models in testing the predictions.

$$X_L^{(train)} \cap X_g^{(train)} = \emptyset$$

The training set's observations are divided at random into Y_i→ class value and X_i→ feature values. Equation 1 represents the learners as a function, indexed as 'k' index:

Ψ ̃(P_n)≡ Ψ^_(K ̃(P_n)) (P_n (5)

where Ψ→ parameter space, K_n→ collection of learners, and P_n→ probability distributions. The nth learner is cross validated using the learners B_L predicted values.

$$\widehat{K}(P_n) \equiv \arg \underset{k}{min}\, E_{B_n} \sum_{i,B_n(i)=1} \left(Y_i - \widehat{\Psi}_k\left(P^0_{n,B_n}\right)(X_i)\right)^2$$

(6)

Where K ̃(P_n)→ Cross-validation selector, and B_n∈{0,1}→random binary vector.In equation 3, the cross-validated risk criteria are defined.

$$R_{CV}(\beta) \equiv \sum_{i=1}^{n}(Y_i - m(Z_i|\beta))^2 \qquad (7)$$

Where R_CV→risk criteria,Z_i→sample set of observations

Blending is a term used to describe this process as shown in equation 3. The rationale for this is that only have predictions for a subset of the training data, and the predictions used to train the meta learners are made by base learners which only saw a portion of the training data. Unless the data set is large, the meta learner's forecasts are hard to achieve the behaviour of the base learners at test time. Subsample is a class of ensemble that can be a very efficient way for Stacking. These ensembles segment the training data into J partitions and allocate K-fold cross-validation to each partition to fit the base learners. The base learner is fitted on the predictions made in each partition, as each partition functions as its mini-ensemble. The number of features for the meta learner is k × J because there are k foundation learners and J partitions. Super learning is a supervised learning technique that relies on loss and employs a cross-validation learning strategy.

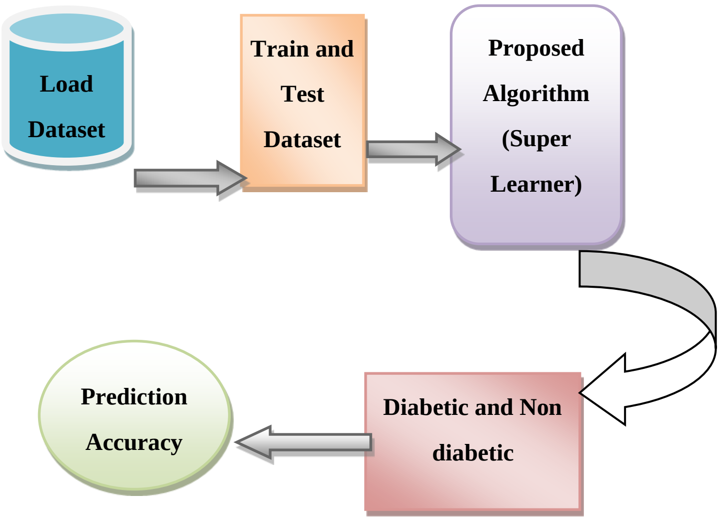

Fig. 1. Proposed Workflow

A learning algorithm can be thought of as a search through a space H of theories to find the right one. When the sum of training data relative to the size of the hypothesis space is insufficient, a statistical issue occurs. Without enough data, the learning algorithm will generate a large number of hypotheses in H, each of which has the same accuracy on the training data. The algorithm will minimize the probability of selecting the incorrect classifier by making an ensemble out of all of these accurate classifiers and averaging their votes. The steps involved can be summed up as follows:

1. Choose a k-fold split between the Vanderbilt and Pima Indian diabetes datasets (PIDD and VDD).
2. Choose m model configurations or base models.
3. For every foundational model:

a. Apply k-fold cross-validation for evaluation.
b. Keep track of every out-of-fold prediction.
c. Utilising the entire training dataset, fit the model and save it.
4. Apply a meta-model to the predictions that are out-of-fold.
5. Test the model to forecasts or assess it on a holdout dataset.

The present meta-model's foundation predicts the training dataset's aim by using predictions from base models as input and predictions for the training dataset as an output. A vector of 50 values, each of which represents a projection from a base model for a single training dataset sample, would constitute one input sample if we had eight base models. The input data for the meta-model would be 1,000 rows and 50 columns if the training dataset contained 1,000 occurrences (rows) and 50 models. Logistic regression, decision tree classifiers, SVC, Gaussian NB, K neighbours classifiers, Ada Boost classifiers, bagging classifiers, random forest classifiers, extra trees classifiers, and super learners are among the base learners used in the suggested technique.

4. Experiments

The general overview of our experimental setup is given in this section. Python was selected because of its extensive library for developing the super learner.

4.1. Dataset

One of the advantages of this study is its ability to predict whether or not a patient has diabetes by trimming data from a vast library. The Diabetes, Digestive, and Kidney Disorders National Institute-supported UCI repository provided the Pima Indian diabetes dataset (PIDD) for this work, whereas the Vanderbilt dataset was sourced via the biostatistics programme.. The PIDD has 768 observations on nine variables. All patients were female and at least 21 years old, and the predictor variables included age, glucose, blood pressure, insulin, BMI, and diabetes pedigree function. The goal variable outcome was one of the variables. Of the 1046 participants in the biostatistics study programme, 403 were chosen based on their age group, BG levels, and degree of physical activity. The research study aimed to determine the prevalence of obesity, diabetes, and other cardiovascular risk factors. Of the 1046 respondents in the dataset, 19 variables were interviewed; the summary information is displayed in Table 1. Out of the 390 African-American patients with an A1c more than 6.5 who were identified as diabetics in the Vanderbilt dataset, 60 are included in the second set. The goal is to use the demographic data to predict diabetes. T2DM is also linked to hypertension; both conditions may be part of the syndrome x. Of the 403 participants who underwent diabetes screening, a glycosylated haemoglobin level of greater than 7.0 is often considered a positive diagnosis of diabetes. The goal is to anticipate Diabetes utilizing the segment factors. T2DM is likewise connected with hypertension the two of them might be essential for the condition x the 403 subjects were the ones who evaluated for Diabetes glycosylated hemoglobin> 7.0 is normally taken as a positive finding of Diabetes. The goal is to use diagnostic measurements to forecast when diabetes will manifest.

The PIMA Indian diabetes was classified using hybrid classification methods, and the results are shown in Table 2. The dataset was divided into two halves for this experiment: 20% for the test dataset and 80% for the training dataset. Table 2 shows the classification accuracy of the various classifiers. Feature value scores can be used to aid in data interpretation, but they can also be used to rate and pick the most useful features for a predictive model. There are 1,046 samples in the dataset, each with ten input variables, five of which are redundant and five of which are critical to the result. To get started, we can separate the training dataset into test and train sets, use the training dataset to train a model, use the test set to generate predictions, and evaluate the outcomes using classification accuracy.

Table 1 Summary details of PIDD

Feature	**Description**
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index (weight in kg/(height in m)^2)
Diabetes Pedigree Function	Diabetes pedigree function
Age	Age (years)
Class	Class variable (0 or 1) indicating presence or absence of diabetes

4.2. Result and Discussion

In this work a Super learner joined with classification models for evaluating the given dataset. We have looked at every one of the information got, and the best exhibition of super learners depends on the outcomes. For exploratory data science (EDS), the statistical data analysis software "scikit-sklearn" in Python is used. Analyses of the data show a broad trend of age groups with low levels of physical activity having an elevated risk for diabetes.

•First, generate out-of-fold predictions using k-fold cross-validation; these predictions will be used to train the meta-model, also known as the "super learner." To do this, the data must first be divided into k folds, of which 10 can be used.

• On the testing side of the break, the model is matched, and on the testing portion, the out-of-fold predictions. This is performed for each model, and all predictions that are out-of-fold are preserved.

• The meta-model input will be a column for each out-of-fold projection. For one-fold of the results, columns are collected from each algorithm, and the rows are stacked horizontally.

•A fresh sample is predicted by feeding it into each base model first, which produces a prediction.The base-model projections are then concatenated into a vector and fed into the meta-model as data. After that, the final forecast for the data row will be done by meta-model.

The following is a summary of this approach:

1. Use a sample that the models have not yet seen during

testing.

2. For each base model, write the following:

a. Based on the sample, make a guess.

b. Estimation of the store.

3. Combine sub model projections into a single vector.

4. To make a final forecast, feed the vector into the meta-model.

The super learner model generates predictions and enables the fitting and selection of algorithms according on their performance. The stage in the analysis process involves feeding the meta model's output prediction with data from the base models.

4.3. Evaluation

The evaluation metrics like Sensitivity, specificity, and accuracy are statistical measures used to assess the performance of the suggested model; the key characteristics are displayed in figure 2. The proportions of the classifier's assessment are as displayed in equation (8) to (10)

$$\text{accuracy} = \frac{(\text{TP+TN})}{(\text{TP+FP+TN+FN})} \tag{8}$$

$$\text{sensitivity} = \frac{\text{TP}}{(\text{TP+FN})} \tag{9}$$

$$Specificity = \frac{TN}{(TN+TP)} \tag{10}$$

where FN = False Negative, FP = False Positive, TP = True Positive, and TN = True Negative As the datasets are joined and particular fields are taken out and used, the data becomes unbalanced. The overall accuracy of the model is 79.17%.

4.4. Principle findings

Through the use of an ensemble of permutations, the proposed models improve predictions by stacking base learners in layers that propagate to the next layer. Three different datasets with varying AUC, sensitivity, specificity, and accuracy are used to train the suggested model. Table 2 provides demographic information, and Figure 3 displays the properties of the separate datasets. Since the Area under the ROC curve (AUC), which is scale-invariant and indicates the model's likelihood of positive cases rather than negative examples, provides the best prediction aggregate.

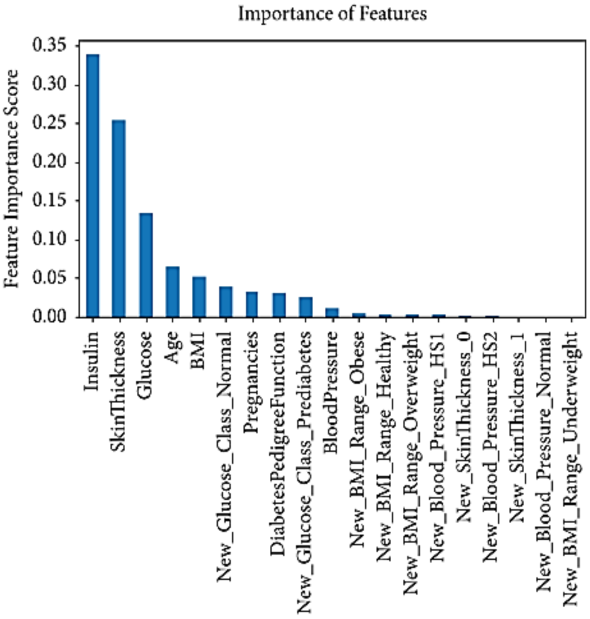

Fig. 2. Features and its Range

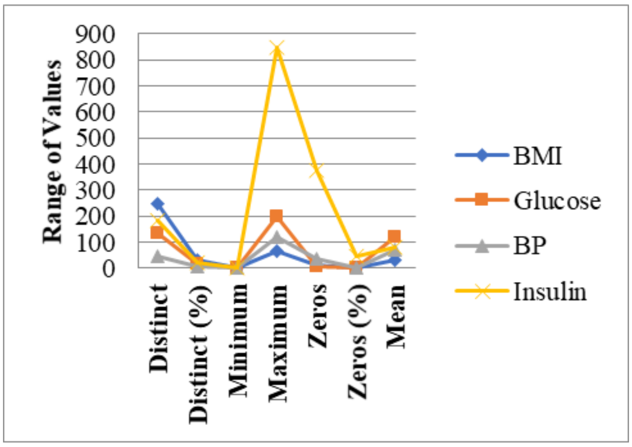

Fig. 3. Different Features of the Dataset

Unless there is a hereditary characteristic unique to each patient, the amounts of insulin secreted by healthy and diabetic patients can change based on the age.

Table 2 Demographic details of PIDD

S/N	**Pregnancies**	**Glucose**	**BP**	**Insulin**	**BMI**	**DPF**
Unit	Number of times	mg/dL	mmHg	muU/ml	kg/m2	
Distinct count	17	136	47	186	248	517
Unique (%)	2.20%	17.70%	6.10%	24.20%	32.30%	67.30%
Mean	3.8451	120.89	69.105	79.799	31.993	0.47188

Range	0–17	0–199	0–122	0–846	0–67.1	0.078–2.42
Zeros (%)	14.50%	0.70%	4.60%	48.70%	1.40%	0.00%
5-th percentile	0	79	38.7	0	21.8	0.14035
Q1	1	99	62	0	27.3	0.24375
Median	3	117	72	30.5	32	0.3725
Q3	6	140.25	80	127.25	36.6	0.62625
Range	17	199	122	846	67.1	2.342
IQR	5	41.25	18	127.25	9.3	0.3825
Co-eff of variative	0.876	0.264	0.28	1.444	0.246	0.702
MAD	2.772	25.182	12.639	84.505	5.842	0.247
Skewness	0.902	0.174	-1.844	2.272	-0.429	1.92
Sum	2953	92847	53073	61286	24570	362.4
Variance	11.354	1022.2	374.65	13281	62.16	0.10978
Memory size	6.1 KB	6.1 KB	6.1 KB	6.1 KB	6.1 KB	6.1 KB

Table 3 Rules to be followed for the dataset.

Rules	Rule description
R1	Diabetes is considered negative if BMI is less than 23.4 and the diabetes pedigree function is less than 0.647.
R2	Diabetes is positive if blood pressure is 70 mm Hg and glucose is 100 mm Hg.
R3	Diabetes is positive if the skin thickness is less than 22 and BMI is less than 25.8.

Table 3 lists the guidelines followed in PIDD. Of the patients in the PIDD dataset, 268 are classified as diabetics and 500 as healthy with appropriate control within the necessary levels.The way the body uses glucose is affected by the type of diabetes T1DM or T2DM which determine the treatment options. Glucose builds up in the blood if insulin isn't working properly leads to a disease known as hyperglycemia. Hypoglycemia is the medical term for a low blood sugar level. Both T1DM and T2DM have the potential to cause major problems the figure shows the comparison of healthy vs diabetic individual's insulin count.

When the pancreas cells cease releasing insulin, type 1 diabetes emerges. Glucose cannot reach muscle cells for energy without the presence of insulin. Instead, glucose levels in the blood increase, making a person sick. If insulin is not substituted, type 1 diabetes will lead to death. For the remainder of their life, people with type 1 diabetes would continue to administer insulin. Type 1 diabetes is most common in children and young adults under the age of 30, but it can strike someone at any age. Insulin levels are distributed as follows: 17.5% of healthy people and 8.4% of people with diabetes, respectively.

The percentage of aberrant blood glucose levels that need to be controlled in both diabetics and healthy individuals' blood pressure can lead to both short- and long-term health issues. Type 1 diabetes symptoms frequently occur unexpectedly, prompting blood sugar testing. The American Diabetes Association (ADA) has established screening guidelines since certain types of diabetes and prediabetes have slower developing or less obvious symptoms. According to the ADA, the following people should get tested for diabetes: Anybody over 25 who has additional risk factors, such as a history of polycystic ovarian syndrome, high blood pressure, excessive cholesterol, or a sedentary lifestyle.

The blood pressure values of those with diabetes and those in good health were 23% and 33.6%, respectively. High blood pressure is twice as likely to occur in diabetics as in non-diabetics. High blood pressure will cause heart failure and stroke if it is not treated. An individual with diabetes and high blood pressure is four times more likely to experience heart disease than someone who does not have this disorder. Around two-thirds of diabetic adults have blood pressure that is higher than 130/80 mm Hg or takes hypertension drugs.

Table 4 compares the accuracy of the classifiers using the suggested model. After analyzing various classification algorithms with different characteristics, it was discovered that the combination of different algorithms gives better accuracy when the three features of insulin, glucose, and age are combined. 71.48 percent accuracy was attained. For the sequence of five features glucose, blood pressure, pedigree, and BMW, the Naive Bayes Classifier would offer an outstanding accuracy value of 77.73 percent. Insulin, Glucose, BMI, Pedigree, Blood, Pressure, and Age are all combined in the Random Forest Classifier and achieve an increased accuracy of 76.43 percent. The best accuracy value of each algorithm with no features is listed with a graphical representation in figure 18.

In this proposed work, we used various machine learning classification algorithms to predict the form of Diabetes Occurrence using different characteristics of Diabetes Mellitus. It has been observed that as the number of features increases, the algorithm's accuracy improves. In comparison to other algorithms, Naive Bayes, SVM, and Logistic Regression are more effective. It is analyzed, and the result obtained is better, but it also needs to be changed.

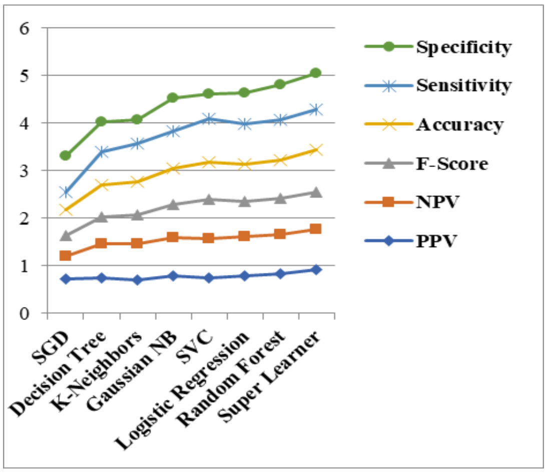

Fig. 4. Before Feature Extraction Classification results

Table 4 Accuracy of the classifier for Diabetes datasets

Classifier	PIDD	T1DM	T2DM
Logistic Regression	93.610	94.450	96.550
Decision Tree Classifier	72.210	71.750	76.030
SVC	95.810	94.300	97.310
Gaussian NB	96.100	94.400	97.670
K Neighbors Classifier	91.200	87.310	93.550
Ada Boost Classifier	89.700	88.630	92.260
Bagging Classifier	84.100	82.330	86.750
Random Forest Classifier	83.300	82.110	86.650
Extra Tree	81.100	81.450	85.600
Super Learner	**96.300**	**94.650**	**97.730**

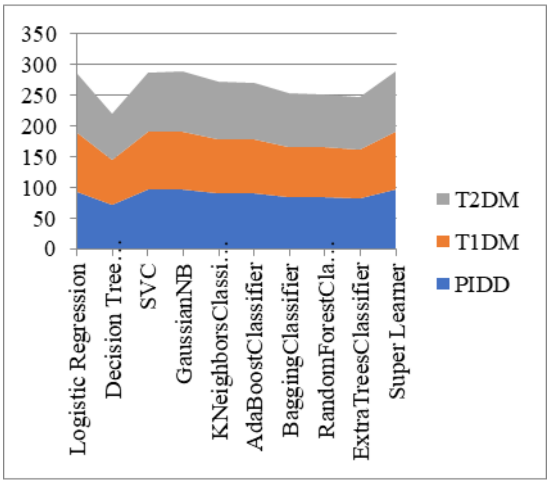

Fig. 5. Comparison of an ensemble of classifiers algorithms

The following table 5 shows the RMSE for a score for the base learner and the super learner. The true score for the ith data point is denoted by yi from the equation, and the expected value is denoted by yi. The Euclidean interval between the vector of true scores and the vector of expected scores, averaged by n, where n is the number of data points, is one way to consider this calculation. From the table, we can see that adjusted R2 is displaying the right trend even if the model is penalised for more variables, even though we are not introducing any new information from case 1 to case 2. In this instance, Adjusted R2 performs better than RMSE, which is only capable of comparing projected and actual values. Moreover, the model's quality cannot be determined by the RMSE's absolute value. It is limited to cross-model comparisons, whereas Adjusted R2 makes this comparison with ease. A model is considered subpar if its adjusted R2 value is 0.05.

Table 5 Comparison of RMSE score for super learner

```
score-m  score-s  ft-m  ft-s  pt-m  pt-s
layer-1  Ada Boost Classifier      0.91     0.02  0.85  0.05
0.05  0.01
layer-1  Bagging Classifier        0.84     0.04  0.30  0.01
0.01  0.00
layer-1  Decision Tree Classifier  0.74     0.04  0.04  0.01
0.00  0.00
layer-1  Extra Trees Classifier    0.83     0.06  0.17  0.06
0.01  0.00
layer-1  Gaussian NB               0.96     0.02  0.02  0.01
0.00  0.00
layer-1  K Neighbour's Classifier       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  Random Forest Classifier     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  Ada Boost Classifier       0.93     0.02  1.89  0.02
0.08  0.02
layer-1  Bagging Classifier         0.87     0.02  1.80  0.08
0.03  0.05
layer-1  Decision Tree Classifier      0.76      0.02  0.28
0.02  0.00  0.00
layer-1  Extra Trees Classifier      0.87     0.02  0.31  0.25
0.01  0.01
layer-1  Gaussian NB                0.98     0.01  0.05  0.04
0.02  0.03
layer-1  K Neighbour's Classifier  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  Random Forest Classifier      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  Ada Boost Classifier      0.93     0.01   3.99  0.05
0.09  0.03
layer-1  Bagging Classifier       0.87     0.01   4.57  0.14
0.13  0.24
layer-1  Decision Tree Classifier  0.76     0.02   0.68  0.04
0.00  0.00
layer-1  Extra Trees Classifier    0.88     0.01   0.63  0.55
0.04  0.05
layer-1  Gaussian NB              0.98     0.01   0.08  0.10
0.03  0.07
layer-1  K Neighbour's Classifier     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  Random Forest Classifier   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 6 Performance of Suggested Methods

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

With the ensemble-based approach, they were able to attain 89% sensitivity, 71.5% specificity, 79.17% accuracy, and 0.875 AUC in PIDD datasets using 10-fold cross-validation. Super learners achieved the greatest classification accuracy of all the classifiers, achieving 97% on the datasets displayed in figure 5. Plotting the measured and expected values for logistic regression gives a visual representation of the data that the suggested model has found. They achieved 89% sensitivity, 71.5% specificity, 79.17% accuracy, and 0.875 AUC for the T1DM datasets using 10-fold cross-validation. Super learners achieved the best classification accuracy of any classifier, scoring 98% on the datasets. Table 6 shows the plotted measured and predicted values for logistic regression, which gives a visual representation of the data that the suggested model discovered.

Graphical

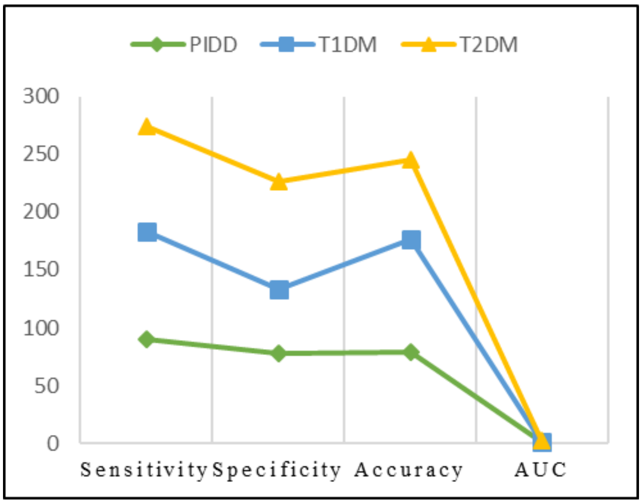

Fig. 6. Analysis of using various datasets

The 10-fold cross-validation is used for the T2DM datasets, they got 89.1% sensitivity, 71.53% specificity, 79.16% accuracy, and 0.876 AUC. Among the classifiers, super learners acquired the most elevated classification exactness with 97% on the datasets is displayed in figure 6. Plotting the deliberate and anticipated values for logistic regression gives a visual portrayal of the information that the recommended model has found...In contrast to other classification algorithms, the maximum accuracy of 97% was attained by spitting data into 30% testing and 70% preparation for diabetes prediction using hybrid classifiers for the updated PIMA Indian diabetes dataset. One of the most intriguing findings of the proposed model is that although specificity was previously discussed in terms of whether or not a patient has diabetes, the authors here concentrate on not only diabetes but also its classification.

5. Conclusion

Finding patterns in the available data with little error and excellent prediction accuracy is the main goal of machine learning. The researcher must choose the algorithms in order to determine which is optimal for the given task. This work offers an in-depth analysis of blood glucose levels using automated techniques for detecting and diagnosing diabetes. For DM identification, publically available datasets from Kaggle, CGM, and PIDD are combined. Finding the combination of Datasets based pn the type of DM is the main focus of this work. The super learner is employed to choose the best method for a particular classification problem, such as differentiating between Type 1 and Type 2 diabetes. With great accuracy, the suggested model analyses and predicts DM types I and I. Better accuracy is reflected in the performance metric for the suggested strategy. In contrast to previous methods, by using the class weight approach to balance the unbalanced data, the suggested approach combines the strengths of machine learning and statistical modelling, offering significantly more advantages.

Additionally, a comprehensive examination of datasets is conducted in research areas that are expected to benefit from improvement, such as personalised DM disease pathology, intelligent diagnosis and analysis, and genetic ancestry recognition. As a result, the study might be expanded in the future to solve the drawbacks.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] S. Albahli, "Type 2 Machine Learning: An Effective Hybrid Prediction Model for Early Type 2 Diabetes Detection," *Journal of Medical Imaging & Health Informatics*, vol. 10, no. 5, pp. 1069–1075, 2020. [Online]. Available: https://doi.org/10.1166/jmihi.2020.3000.

[2] G. Alfian et al., "Deep Neural Network for Predicting Diabetic Retinopathy from Risk Factors," *Mathematics*, vol. 8, no. 9, p. 1620, 2020. [Online]. Available: https://doi.org/10.3390/math8091620.

[3] R. Alshammari et al., "Improving Accuracy for Diabetes Mellitus Prediction by Using Deepnet," *Online Journal of Public Health Informatics*, vol. 12, no. 1, 2020. [Online]. Available: https://doi.org/10.5210/ojphi.v12i1.10611.

[4] S. I. Ayon et al., "Diabetes Prediction: A Deep Learning Approach," *Int. J. Inf. Eng. Electron. Bus.*, vol. 11, no. 2, 2019.

[5] L. Beqiri et al., "Analysis of Diabetes Dataset," in *43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, 2020, pp. 309–314. [Online]. Available: https://doi.org/10.23919/MIPRO48935.2020.9245318.

[6] J. D. Bodapati et al., "Blended MultitModal Deep ConvNet Features for Diabetic Retinopathy Severity Prediction," *Electronics*, vol. 9, no. 6, p. 914, 2020. [Online]. Available: https://doi.org/10.3390/electronics9060914.

[7] S. P. Chatrati et al., "Smart Home Health Monitoring System for Predicting Type 2 Diabetes and Hypertension," *J. King Saud Univ. Informing Science*, 2020.

[8] T. Daghistani and R. Alshammari, "Comparison of Statistical Logistic Regression and RandomForest Machine Learning Techniques in Predicting Diabetes," *Journal of Advances in Information Technology*, vol. 11, no. 2, pp. 78–83, 2020. [Online]. Available: https://doi.org/10.12720/jait.11.2.78-83.

[9] K. Driss et al., "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.

[10] G. Emmanuel et al., "Performance Evaluation of Machine Learning Classification Techniques for Diabetes Disease," *IOP Conference Series: Materials Science & Engineering*, vol. 1098, no. 5, 2021. [Online]. Available: https://doi.org/10.1088/1757-899X/1098/5/052082.

[11] M. F. Faruque et al., "Predicting Diabetes Mellitus and Analysing Risk-Factors Correlation," *Earth/ Endorsed*, vol. 5, no. 20, p. e7, 2020. [Online]. Available: https://doi.org/10.4108/eai.13-7-2018.164173.

[12] 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. [Online]. Available: https://doi.org/10.3390/s20092649.

[13] F. Hassan and M. E. Shaheen, "Predicting Diabetes from Health-Based Streaming Data Using Social Media, Machine Learning and Stream Processing Technologies," *International Journal of Engineering Research & Technology*, vol. 13, no. 8, 2020. [Online]. Available: https://doi.org/10.37624/IJERT/13.8.2020.1957-1967.

[14] M. T. Islam et al., "Typical and Non-typical Diabetes Disease Prediction Using Random Forest Algorithm," in *fifth International Conference on Computing, Communication and Networking Technologies ICCCNT*, 2020, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICCCNT49239.2020.9225430.

[15] J. Chaki et al., "Machine Learning and Artificial Intelligence Based Diabetes Mellitus Detection and Self-Management: A Systematic Review," *J. King Saud Univ. Inf. Sci.*, 2020.

[16] I. Kavakiotis et al., "Machine Learning and Data Mining Methods in Diabetes Research," *Comput. Struct. Computational & Structural Biotechnology Journal*, vol. 15, pp. 104–116, 2017. [Online]. Available: https://doi.org/10.1016/j.csbj.2016.12.005.

[17] W. K. Lee et al., "Use and Performance of Machine Learning Models for Type 2 Diabetes Prediction in Community Settings: A Systematic Review and Meta-analysis," *International Journal of Medical Informatics*, p. 104268, 2020.

[18] B. Ljubic et al., "Predicting Complications of Diabetes Mellitus Using Advanced Machine Learning Algorithms," *Journal of the American Medical Informatics Association*, vol. 27, no. 9, pp. 1343–1351, 2020. [Online]. Available: https://doi.org/10.1093/jamia/ocaa120.

[19] *Logistic Regression Based Feature Selection and Classification of Diabetes Disease Using Machine Learning Paradigm.*

[20] G. Mainenti et al., "Machine Learnng Approaches for Diabetes Classification: Perspectives to — Artificial Intelligence Methods Updating.," in *JoTBDS*, 2020, pp. 533–540.

[21] M. Maniruzzaman et al., "Classification and Prediction of Diabetes Disease Using Machine Learning Paradigm," *Health Information Science & Systems*, vol. 8, no. 1, p. 7, 2020. [Online]. Available: https://doi.org/10.1007/s13755-019-0095-z.

[22] M. J. Mehdi et al., "Detection and Prognosis of Diabetes Based on Data Science Techniques," *Materials Today: Proceedings*, 2020.

[23] L. Miao et al., "Using Machine Learning to Predict the Future Development of Disease," in *International Conference on UK-China Emerging Technologies (UCET)*, 2020, pp. 1–4. [Online]. Available: https://doi.org/10.1109/UCET51115.2020.9205373.

[24] 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. [Online]. Available: https://doi.org/10.1109/IDEA49133.2020.9170744.

[25] B. P. Nguyen et al., "Predicting the Onset of Type 2 Diabetes Using Wide and Deep Learning with Electronic Health Records," *Computer Methods & Programs in Biomedicine*, vol. 182, p. 105055, 2019. [Online]. Available: https://doi.org/10.1016/j.cmpb.2019.105055.

[26] F. Nusrat et al., "Prediction of Diabetes Mellitus by Using Gradient Boosting Classification," *Avrupa Bilim ve Teknol. Derg.*, pp. 268–272.

[27] N. Pradhan et al., "Diabetes Prediction Using Artificial Neural Network," in *Deep Learning Techniques for Biomedical & Health Informatics*, 2020, pp. 327–339. Elsevier.

[28] M. Rahman et al., "A Deep Learning Approach Based on Convolutional LSTM for Detecting Diabetes," *Computational Biology & Chemistry*, vol. 88, p. 107329, 2020. [Online]. Available: https://doi.org/10.1016/j.compbiolchem.2020.107329.

[29] D. J. Reddy et al., "Predictive Machine Learning Model for Early Detection and Analysis of Diabetes," *Materials Today: Proceedings*, 2020.

[30] A. Rghioui et al., "A Smart Architecture for Diabetic Patient Monitoring Using Machine Learning Algorithms," *Healthcare*, vol. 8, no. 3, p. 348, 2020. [Online]. Available: https://doi.org/10.3390/healthcare8030348.

[31] K. S. Ryu et al., "Cha, A Deep Learning Model for Estimation of Patients with Undiagnosed Diabetes," *Applied Sciences*, vol. 10, no. 1, p. 421, 2020.

[32] P. A. Senior et al., "Pharmacologic Glycemic Management of Type 2 Diabetes in Adults: 2020 Update—The User’s Guide," *Canadian Journal of Diabetes*, vol. 44, no. 7, pp. 592–596, 2020. [Online]. Available: https://doi.org/10.1016/j.jcjd.2020.08.002.

[33] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes Using Machine Learning Classification Methods," *Procedia Computer Science*, vol. 167, pp. 706–716, 2020. [Online]. Available: https://doi.org/10.1016/j.procs.2020.03.336.

[34] R. Unnikrishnan and A. Misra, "Infections and Diabetes: Risks and Mitigation with Reference to India," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 6, pp. 1889–1894, 2020. [Online]. Available: https://doi.org/10.1016/j.dsx.2020.09.022.

[35] K. Vidhya and R. Shanmugalakshmi, "Deep Learning Based Big Medical Data Analytic Model for Diabetes Complication Prediction," *Journal of Ambient Intelligence & Humanized Computing*, vol. 11, no. 11, pp. 5691–5702, 2020. [Online]. Available: https://doi.org/10.1007/s12652-020-01930-2.

[36] L. Wang et al., "Prediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model," *Healthcare*, vol. 8, no. 3, p. 247, 2020. [Online]. Available: https://doi.org/10.3390/healthcare8030247.

[37] A. Z. Woldaregay et al., "Data-Driven Modeling and Prediction of Blood Glucose Dynamics: Machine Learning Applications in Type 1 Diabetes," *Artificial Intelligence in Medicine*, pp. 109–134, 2019.

[38] A. Yaganteeswarudu, "Multi Disease Prediction Model by Using Machine Learning and Flask API," in *Sth International Conference on Communication and Electronics Systems (ICCES)*, 2020, pp. 1242–1246. [Online]. Available: https://doi.org/10.1109/ICCES48766.2020.9137896.

[39] L. Zhang et al., "Predicting the Development of Type 2 Diabetes in a Large Australian Cohort Using Machine-Learning Techniques: Longitudinal Survey Study," *Journal of Medicine/R Med. informatics*, vol. 8, no. 7, p. e16850, 2020. [Online]. Available: https://doi.org/10.2196/16850.