Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 47 (2026), No. 1, pp. 371–398
DOI : 10.47974/JIOS-2182

Integrating natural therapeutics in machine learning models for diabetes mellitus optimization

Mona Adlakha ‡
M. Afshar Alam †
Department of Computer Science and Engineering
School of Engineering Sciences and Technology (SEST)
Jamia Hamdard
Delhi 110062
India

Sameena Naaz §
Department of Computer Science
University of Roehampton
London SW15 5PJ
U.K.

Sherin Zafar *
Department of Computer Science and Engineering
School of Engineering Sciences and Technology (SEST)
Jamia Hamdard
Delhi 110062
India

Abstract

Diabetes Mellitus is a rapidly growing global health concern that poses significant risks, if not detected and managed at an early stage. This study aims to develop a model that accurately predicts diabetes mellitus using machine learning (ML) models, namely hybrid and ensemble models. Various natural remedies that have traditionally been used to manage

‡ *ORCID-ID:* 0009-0009-6278-8271
† *ORCID-ID:* 0000-0002-4677-711X
§ *ORCID-ID:* 0000-0003-0080-5063
* *ORCID-ID:* 0000-0001-6656-1095

‡ *E-mail:* monaadlakham@gmail.com
† *E-mail:* aalam@jamiahamdard.ac.in
§ *E-mail:* sameena.naaz@roehampton.ac.uk
* *E-mail:* sherin.zafar@jamiahamdard.ac.in (Corresponding Author)

diabetes have also been studied. Standalone machine learning algorithms were combined as hybrid and ensemble models. Performance metrices were used to evaluate the models. This integrated approach is a complete package for diabetes management, reducing the overall cost of this life-long disease.

Subject Classification: *92B20 - Neural networks, artificial intelligence for/ in biological.*

Keywords: *Diabetes prediction, Ensemble model, Hybrid model, Machine learning (ML), Natural remedies for diabetes cure.*

1. Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels, due to body's impaired production of insulin, insulin action, or both. Over the years, this disease has become a significant global health concern due to its rising prevalence and link to severe complications like cardiovascular disease, kidney failure, neuropathy, and in severe cases death due to its complications.

A recent study, by Lacent, conducted the first global analysis of trends of both prevalence of diabetes and its treatment from 1990 to 2022, covered 200 countries and territories. The study showed that in 2022, an estimated 828 million adults (18 years and older) were diabetic, an increase of 630 million from 1990. The study concluded that low-income and middle-income countries exhibited the largest increase in diabetes prevalence, whereas high-income and industrialized nations and some emerging economies displayed the largest improvements in treatment. Table 1 shows the top 20 countries with the highest prevalence and treatment of diabetes along with that total diabetes cases, untreated cases and treated cases (in million) [1].

Table 1

Top 20 countries with prevalence and treatment of diabetes in 2022.

Country	Total Diabetes cases (in million)	Untreated Diabetes cases (in million)	Treated Diabetes cases (in million)
India	212	133	79
China	148	78	70
USA	42	13	29
Pakistan	36	24	12

Contd...

Indonesia	25	18	7
Brazil	22	10	12
Mexico	13	5	8
Egypt	18	8	10
Russia	12	4	8
Bangladesh	18	13	5
Japan	12	5	7
Turkey	11	5	6
Iran	9	6	3
Nigeria	10	6	4
Philippines	9	6	3
Germany	8	8	0
Thailand	9	5	4
Vietnam	8	5	3
UK	6	6	0
South Korea	6	3	3

The Figure 1 depicts the country-wise treated and untreated diabetes cases in the top 10 countries where diabetes is prevalent [1].

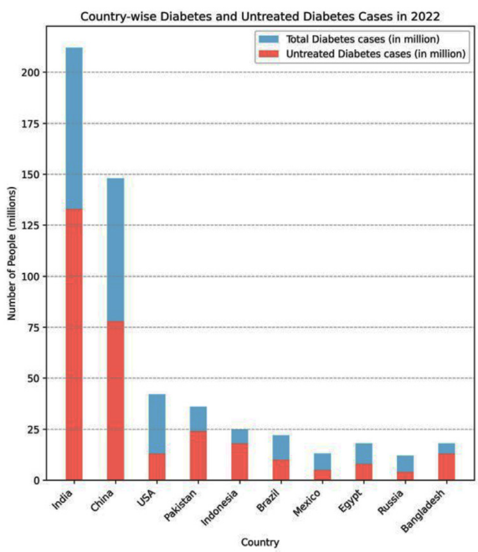

Figure 1

Country-wise treated and untreated diabetes cases (in the top 10 countries).

Diabetes mellitus is divided into three categories: type 1, type 2, and gestational diabetes. Type 1 diabetes is an autoimmune condition where the body's immune system destroys insulin-producing β-cells in the pancreas, leading to absolute deficiency of insulin. Typically developed in childhood and adolescence, it is generally caused by genetic factors and the body's autoimmune response. There is no cure for type 1 diabetes. It requires regular monitoring, along with lifelong insulin therapy. Type 2 diabetes is a metabolic disorder caused due to insulin resistance in peripheral tissues, combined with dysfunction of β-cells, resulting in insulin deficiency. This disease usually develops in adults and is generally caused due to lifestyle factors - poor diet, obesity, and low physical activity. In majority cases, type 2 diabetes is non-reversable and has to be managed through changes in lifestyle, appropriate diet, exercise, medication, and in some cases insulin injections. Gestational diabetes occurs during pregnancy and is generally resolved after childbirth. The hormonal imbalance during pregnancy, causes the insulin response. The disease has to be managed through diet, medication, exercise and sometimes insulin injections during pregnancy.

Type 2 diabetes is the most prevalent type and is primarily caused by insulin resistance, sometimes causing several serious complications, such as diabetic foot, neuropathy, bone disease, and a weakened immune system, which increases vulnerability to bacterial and viral infections. To avoid complications and lessen the burden of disease on people and healthcare systems around the world, extensive research has been conducted for early diagnosis, and efficient management through medication, lifestyle changes, and routine monitoring.

There are several biochemical pathways that contribute to the pathogenesis of diabetes and its complications. Numerous research studies have been conducted to understand these pathways and develop targeted therapies that mitigate the complications associated with diabetes. Currently, several synthetic drug classes are being used to manage diabetes. However, over the past two decades, there has been increasing interest in utilizing natural remedies for managing chronic diseases, including type 2 diabetes [2].

For several decades, computational techniques have been used to solve problems in various fields of study. The last decade has seen a steep rise in research that uses various machine learning, hybrid, ensemble and deep learning models for medical diagnosis and treatment. Some of these models have been integrated with clinical practice, to aid in timely detection and treatment of disease.

In traditional medical practice, there are centuries old traditions of using locally grown herbs and plants for mitigation and cure of diseases, including diabetes mellitus. Research has confirmed the significance of medicinal plant therapy for the mitigation and treatment of diabetes mellitus. Some herbs have shown antidiabetic activities through various mechanisms, by reducing the oxidative stress and inflammation, increasing the insulin sensitivity and glucose absorption, and thus regulate insulin induced signalling in different tissues. Numerous types of herbs are freely available globally. These are low in cost, low in toxicity, and rich phytochemical contents [2].

Plants like chicory (Cichorium intybus), gurmar (Gymnema sylvestre), Bengal quince (Aegle marmelos), garlic (Allium sativum), Indian lilac (Azadirachta indica), and many others have been proven to reduce the fasting glucose and Haemoglobin A1C (HbA1c) in type 2 diabetes, have anti-hyperglycemic, insulin-releasing properties [3] [4] [5] [6].

Type 2 diabetes being a lifelong disease, a patient can be comforted and can lead a normal life if the symptoms are managed and the cure causes minimal side effects. Early and accurate disease diagnosis can increase the efficiency of the chosen treatment and give the patient a longer and more comfortable life.

The structure of this paper starts with Section 2 that discusses the related work on the worldwide diabetes trends, gives a range of machine learning algorithms being used for disease prediction, and summarizes some natural remedies being used for managing diabetes mellitus. The Section 3 discusses the methodology of the proposed model, describes the dataset use, gives a brief description of data analysis and pre-processing techniques, along with the algorithm employed for the proposed model and the performance metrices used. Section 4 outlines the experimental work done along with the result analysis. Lastly, Section 5, presents the conclusion drawn from this study and outlines roadmap for future research.

2. Related work

The recent study on 200 countries and territories, on the worldwide trends in the diabetes prevalence and its treatment from 1990 to 2022, included 108 population representative studies with measurements of glycemic biomarkers. The study included 141 million adult participants, 18 years and older. It estimated that, in 2022, 828 million adults (18 years and older) were diabetic, which was an increase of 630 million adults from 1990. Definition of diabetes, according to the study, is - fasting plasma glucose (FPG) greater than equal to 7.0 mmol/L and HbA1c greater than equal to, or use of diabetes medication. It defined treatment as the use of oral hypoglycemic drugs or insulin. In the population studied, the diabetes prevalence, in year 2022, had more than doubled (since 1990) in 145 countries, with 66 countries showing this increase in women and 79 countries showing this increase in men. Of the top 5 countries with the prevalence of diabetes, 4 are in Asia [1].

Numerous computational techniques have been used for detection of various diseases. A recent study devised a technique for identifying lung disease from chest X-rays and categorised pneumonia and Covid-19 cases. Their suggested framework combined machine learning, deep learning, and soft computing. This novel model for detecting Covid-19, uses a picture improvement technique, normalization, and a ROI-based feature extraction, and subsequently takes into account the lung-specific features. This maximizes the performance and computing efficiency of the classification [7].

Another study used machine learning to investigate the patterns derived from physical activity monitoring (PAM), through an at-home screening machine, and used it to identify individuals that were at risk of Type 2 diabetes mellitus. HbA1c and other data (7532 subjects) from the National Health and Nutrition Examination Survey (NHANES) (for 2011–2014), were studied, focusing on the PAM component. Using age, and the features resulting from PAM, as inputs for an Extreme Gradient Boosting (XGBoost) model, SHapley additive exPlanations (SHAP) was used to enhance the interpretability of the models. The XGBoost model that utilized the PAM features with age, showed a better AUC of 0.80, as compared the model which solely included PAM features [8].

Recent research, on the 'Indicators of Heart Disease (2022 UPDATE)' dataset, which originated from CDC and constitutes a fundamental component of the Behavioral Risk Factor Surveillance System (BRFSS), USA; used various machine learning models for cardiovascular disease prediction. To improve the accuracy of predictions, features associated with herbal medicines and their effects on heart health were selected. Few ensemble and hybrid models were used for disease prediction. A hybrid model that combined random forest (RF) for interpretability and Long Short-Term Memory (LSTM) networks for time series analysis achieved an accuracy of 100%. The study suggested including natural remedies into machine learning models, thereby offering a promising approach to enhancing the prediction and early prevention of heart disease [9].

Recent studies, on the Pima Indians Diabetes dataset, used Support Vector Machine (SVM) and deep learning models for diabetes prediction. The studies implemented Artificial Neural Networks (ANN) with 1, 2 and 3 hidden layers, with multiple neurons in each layer, using RMSprop and Adam optimizers. Models were trained during epochs. The maximum accuracy of 84.42% was obtained with ANN with three hidden layers with Adam optimizer [10] [11].

In a study, on the Pima Indians Diabetes dataset, the data was pre-processed, before feeding this data into the training models. ML techniques: random forest classifiers with gridsearchCV, Natural Gradient Boosting (NGBoost), Extreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), bagging and Light Gradient Boosting Machine (LightGBM), were employed as base classifiers, for training. Feature engineering was used to create new features. The proposed stacked ensemble model showed an accuracy of 92.91%. SHAP was used, which facilitated the interpretation of the machine learning models [12].

A research proposed a robust approach based on the stacked ensemble method, using several ML methods, for predicting diabetes, on the Pima Indian Diabetes dataset. The base models SVM, K-Nearest Neighbor (KNN), Random Forest and Naïve Bayes were used. The final predictions were made using Logistic Regression. Linear and stratified sampling was used to ensure data consistency. K-fold cross-validation endured that the model did not show overfitting. The proposed stacked ensemble model outperformed the base classifier, and gave an accuracy of 94.17% [13].

A latest study experimented with the effectiveness of a stacked ensemble ML model, by incorporating a meta-model approach, on the Pima Diabetes dataset. In the first step of modeling (Level 1), two base models: random forest and SVM were trained using distinct feature selection, and the model was tuned and optimization strategies were applied. In the second stage, to combine the predictions from these base models, the ensemble methods, (voting and stacking classifiers), were employed The final prediction was made by a meta-model gradient boosting classifier (GBC). This system gave a real-time assessment of diabetes risk. The Meta-Model (GBC) gave an accuracy of 99.13%, and a precision of 100% [14].

A study focused on developing a hybrid model that combined LightGBM and CatBoost to minimize overfitting and improve accuracy by reducing variance. This model was compared with the performance of

LightGBM, XGBoost, CatBoost, Ada-Boost, Decision Tree, Random Forest, and GBM algorithms. The Bayesian hyperparameter optimization was used, to find the best hyperparameters to use with the under-lying learners. The regularization parameter was fine-tuned, which reduced the variance (overfitting) of the hybrid model. The hybrid model demonstrated the best F1-score of 99.37% and accuracy of 99.37% [15].

Recent research used a dataset from Sylhet Diabetes Hospital. This dataset included symptom and demographic information of the patients. The research used ensemble feature selection for identifying critical predictor features and Generative Adversarial Networks (GANs) to address class imbalance, by generating synthetic data. A hybrid loss function was implemented to improve classification. Results showed that the attention-based DBN model achieved an Area Under the ROC Curve (AUC) of 1.00 with high values, F1-score, precision and recall and outperformed several baseline models [16].

Research recently used machine learning, ensemble and deep learning techniques for diabetes prediction. The study used two datasets, the Pima Indians Diabetes dataset (768 records, 9 features) and another diabetes dataset (2000 records, attributes). Machine algorithms: K-Nearest Neighbors, Naive Bayes, decision tree, random forest, SVM, and logistic regression; ensemble deep learning techniques: TabNet, XGBoost, and multi-layer perceptron, were used for prediction. Their proposed hybrid deep learning ensemble approach showed an accuracy of 96%, a much higher efficiency over common ML algorithms [17].

Machine learning and deep learning algorithms have been used for a wide range of applications. A recent study used pre-processed images as inputs for garbage detection classification models - K-Nearest Neighbour (KNN), Naive Bayes (NB), Logistic Regression (LR), and SVM. The proposed ConvNet model, with an accuracy of 79%, was suitable for classifying garbage in open space, while the LR model with an ROC of 1.00, was suitable for the garbage detection problem [18]. Other researchers have designed hybrid deep learning models that combine GRU and LSTM networks for real-time, DDoS detection and build robust models that can be utilized in critical applications like healthcare IoT systems [19].

A comparatively new area of study is the use of mobile applications in the field of health care (mHealth apps). A recent study accessed the behavioural elements that affected Indian customers in accepting and using mHealth apps [20]. Since healthcare datasets are large, have varied data types, and may contain hidden structure, researches proposed a model that combines concepts from differential geometry and tools from manifold learning. When these tools were incorporated in deep neural network, guiding the model towards more reliable disease predictions [21].

Recent research used the database, Medical Information Mart for Intensive Care IV (MIMIC-IV), in their study. They proposed a model (MXLPred) for prediction of mortality and hospital stay length. The model used FastText embedding on the clinical discharge summary notes. The proposed MXLPred mortality prediction model used the ensemble XLNet-GRU architecture and displayed a significant improvement in the accuracy and AUC, as compared to available state of the art techniques for ICU mortality risk prediction, on the MIMIC-IV dataset [22].

Historically, in many countries, there are centuries old traditions of using locally grown herbs and plants to cure various diseases. Many clinical trials and experiments have been conducted, that have proven the efficacy of these herbs and plants. Not just for managing diabetes, most of these plants have multiple health benefits. Numerous studies have been conducted to understand the biochemical pathways that contribute to the pathogenesis of diabetes mellitus and its complications. Scientists devise targeted therapies that moderate the complications associated with diabetes. Numerous synthetic drug classes are available, that manage diabetes. The past two decades have seen a growing interest in utilizing natural remedies for managing chronic diseases, including type 2 diabetes [2].

A recent study, focused on the comprehensive role of Gymnema sylvestre (GS) and other herbs for their ayurvedic impact in controlling blood sugar, by releasing insulin from the pancreas (in diabetic type 2). GS supplement also leaves an impact on various cardiovascular risk factors, improving the lipid profile, blood pressure and controls glycemia. Extracts of Gymnema have been shown to have an inhibitory impact on triglyceride accumulation in muscles and the liver, while reducing the buildup of fatty acids in the bloodstream, thus also reducing HbA1c parameters [4].

A recent literature reviewed herbal and natural remedies in the traditional Indian medical system and how these are used for diabetes treatment. The study emphasized on how India, with its traditional medical heritage, has an edge globally, for providing supplement treatment for diabetes. The paper discussed numerous natural remedies that gives cure in diabetes and other ailments. One such plant, gurmar, contains phytochemicals that have insulin-releasing properties, is an anti-hyperglycemic, and reduces HbA1c and fasting glucose in type 2 diabetes [5].

Table 2 lists plants (in alphabetical order of scientific name) that have been traditionally used as natural remedies for diabetes management, and their medicinal benefits.

Table 2

Traditionally used therapeutics for diabetes mellitus

Scientific Name (Common Name)	Medicinal Benefits
Azadirachta indica (Margosa/ Indian lilac/ neem)	Neem leaf extracts significantly reduces the fasting glucose in type 2 diabetics [5].
Aegle marmelos (Bengal quince/ Bael)	Enhances β-cell function and insulin sensitivity [6]
Allium sativum (Garlic)	Regulates insulin secretion and lowers glucose [6].
Berberis vulgaris (Berberine)	Improves glucose uptake, reduces insulin resistance, activates AMPK*. Additionally, it is also antihyperlipidemic, anti-inflammatory, anticancer, antioxidant and hepatoprotective [2].
Cichorium intybus (Chicory)	Has shown anti-diabetic properties [3].
Cinnamomum verum (Cinnamon)	Enhances insulin function and regulates glucose levels [6].
Clerodendrum glandulosum (East Indian Glory Bower)	Leaves are used to manage diabetes and hypertension [23].
Coriandrum sativum (Coriander)	Increases insulin production and is an antioxidant [6].
Costus pictus D. Don (Insulin Plant)	Consumed as a tea or raw leaves are chewed to reduce blood sugar levels [24].
Curcuma longa (Turmeric)	Turmeric is an antioxidant, anti-inflammatory substance that enhances insulin receptor activity; is antihyperglycemic, anticancer, neuroprotective, antiapoptotic, antimicrobial, and has cardioprotective properties [2]. It is used in various forms (powder, decoction) to manage blood sugar and inflammation [24].
Grifola frondosa (Maitake Mushroom)	Consumed to enhance insulin sensitivity and lower blood glucose. [25]

Contd...

Gymnema sylvestre (Gurmar)	Leaves are used to reduces blood glucose, regenerates β-cells, suppresses sweet taste, improves insulin sensitivity [4]. Phytochemicals in this plant have anti-hyperglycemic, insulin-releasing properties, reducing fasting glucose and HbA1c* in type 2 diabetes [5].
Momordica charantia (Bitter Melon)	Bitter melon has shown slight anti-hyperglycemic efficacy in diabetes. [5]. Improves glucose tolerance and insulin sensitivity [6]. Fruit and leaf extracts are used to lower blood glucose levels [26].
Moringa oleifera (Moringa)	Moringa is a hypoglycemic, neuroprotective, hepatoprotective, hypolipidemic, and antiviral agent. It also improves β-cell function, regulates lipid and carbohydrate metabolism. [2]
Murraya koenigii (Curry Leaf)	Curry leaves improve insulin sensitivity and postprandial glucose. [6]
Nigella sativa (Black cumin/ kalongi)	Black cumin is an anti-inflammatory agent that enhances insulin sensitivity and regenerates β-cells [2]. The seeds and oil are used to reduce blood sugar and improve insulin sensitivity [27].
Ocimum tenuiflorum (Holy Basil/ Tulsi)	Holy Basil improves the glycemic control and lipid profile [5]. It lowers blood glucose and boosts insulin [6].
Phyllanthus emblica (Indian Gooseberry/ amla)	Indian Gooseberries stimulate insulin, lowers lipids and glucose [6].
Portulaca oleracea (Purslane)	Purslane has anti-inflammatory, antidiabetic, anticancer, analgesic, immunostimulant, antimicrobial, and antiviral properties. It also stimulates insulin secretion, is an antioxidant and improves glucose uptake [2].
Punica granatum (Pomegranate)	Pomegranate enhances insulin sensitivity and inhibits glucose absorption enzymes [2].
Salacia reticulata Wight. (Ekanayakam (in hindi))	The bark and root extracts are used to inhibit carbohydrate digesting enzymes [28].
Syzygium cumini (Java Plum)	Java plum enhances insulin secretion and lowers LDL* cholesterol [6].

Contd...

Trigonella foenum-graecum L. (Fenugreek)	Fenugreek seeds are consumed to improve lipid profile and controls glycemic index and improves lipid markers [5] [29].
Withania somnifera (Ashwagandha)	Improves fasting glucose and HbA1c* [5].
Zingiber officinale (Ginger)	Improves glycemic index and antioxidant status. [6]

*AMPK: Adenosine Monophosphate-activated protein kinase, HbA1c: Hemoglobin A1C, LDL: low-density lipoprotein

3. Materials and Methods

3.1 *Methodology*

The methodology used in our study, focused on developing an optimized machine learning based diabetes prediction model, with emphasis on ensemble techniques, is shown in Figure 2. The process begins with data being retrieved from the Pima Indians Diabetes dataset for relevant analysis [25]. This data undergoes data pre-processing, data processing, wrangling, feature scaling and extraction of relevant features. Following the pre-processing, predictive models are constructed. These models used machine learning algorithms and ensemble models. An ensemble model combines multiple base learners to improve the model's

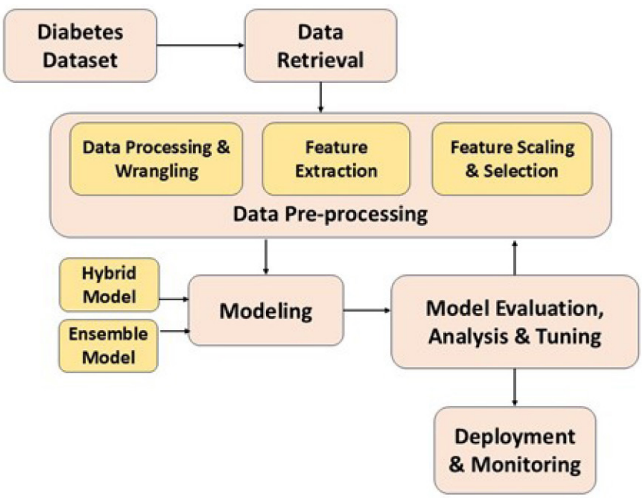

Figure 2

Methodology of the proposed model.

accuracy and enhance its robustness. Each model is further evaluated, analysed and fine turned; based on its performance. The model's performance is assessed using metrics such as accuracy, precision, recall, AUC and Receiver Operating Characteristic (ROC) curve. Once a satisfactory model is achieved, the model is deployed into a real-life practical application for continuous use. Monitoring ensures that the model maintains performance over a period of time. This structured is vital in providing clinical support in decision making for management of diabetes.

3.2 *Dataset Source and Description*

This study was conducted using the Pima Indians Diabetes dataset. This dataset, originally collected from the National Institute of Diabetes and Digestive and Kidney Diseases (USA), has been widely used to predict the onset of diabetes mellitus in 21 year or older women, of Pima Indian descent [30]. Based on medical diagnostic measurements, the eight input predictor features aid in determining the output target feature, Outcome,

Table 3

Description of Pima Indians diabetes dataset.

Feature Name (P/ T) [P: Predictor variable T: Target variable]	**Description**
Pregnancies (P)	Number of pregnancies
Glucose (P)	2-hour plasma glucose concentration a in an oral glucose tolerance test (OGTT)
BloodPressure (P)	Diastolic blood pressure (mm Hg)
SkinThickness (P)	Triceps skinfold thickness (mm)
Insulin (P)	2-hour serum insulin (mu U/ml)
BMI (P)	Body Mass Index (kg/ m^2) (patient's weight in kg/ height in m^2)
DiabetesPedigreeFunction (P)	Diabetes pedigree function estimates the possibility of an individual developing diabetes
Age (P)	Age (years)
Outcome (T)	Class variable 1: Presence of diabetes 0: Absence of diabetes

with 0 indicating absence of diabetes and 1 indicating the presence of diabetes. This dataset comprises of 768 instances, with 268 (34.9%) diabetic and 500 (65.1%) non-diabetic records. Table 3 gives a detail description of the features.

3.3 *Data Analysis and Pre-processing*

The analysis of the dataset reveals the number of null or missing values and outliers. Correlation of the features in the Pima Indians Diabetes dataset show how strongly features are related. Skewness of the features in a dataset helps to understand the distribution of data of each feature.

3.4 *Standardization*

Machine learning models give biased predictions when some features in the dataset have skewed distributions. To ensure unbiased predictions, data preprocessing techniques of normalization and standardization are applied. These techniques rescale the data, so as to ensure that the prediction model robust.

A feature scaling technique standardizes the values of numeric features. 'StandardScaler', follows a standard normal distribution. This transforms data, so that the standard deviation (s) becomes 1, mean (u) becomes 0. Given sample x, if u is the mean of the training samples and s is the standard deviation of the training samples; we obtain a scaled value z, by using the Equation 1:

$$z = (x - u)/s \tag{1}$$

3.5 *Machine Learning Algorithms*

Once the data is prepared, various classification-based machine learning models are developed on the training set. To implement the models, open-source python-based machine learning frameworks are used. The test set is used for prediction of the 'Outcome'. The Pseudocode below gives the steps of each of the proposed machine learning models:

1. Load Dataset: Load Pima Indians diabetes dataset.
2. Data Preprocessing: Feature analysis, and standardize data.
3. Data Train-Test Split: 80-20 split for training and testing.
4. Train multiple models using cross-validation to find the best parameters.

5. Model Selection: Implement multiple models (standalone models, hybrid model's, Ensemble Models).
6. Model Training & Evaluation: Confusion Matrix, classification report, accuracy, precision, recall, F1-Score, ROC Curve & AUC, cross-validation accuracy
7. Graphical Visualization: ROC Curves are plotted for comparison.

3.6 *Performance Metrices and Model Evaluation*

Once the model is trained on a training set, the model's performance is assessed with the testing set, to predict the 'Outcome' (whether the patient is diabetic or not), using performance metrices. The performance of the model is assessed and through the confusion matrix, that gives the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) counts.

Accuracy, the proportion of correct predictions, i.e. the percentage of instances that are correctly classified (TP + TN), as given in Equation 2.

$$accuracy = (TP + TN)/\ (TP + TN + FP + FN) \tag{2}$$

Precision gives the proportion of true positive instances out of all predicted instances that are positive (TP and FP), given in Equation 3:

$$precision = TP/(TP + FP) \tag{3}$$

Recall or the true positive rate or sensitivity, is the proportion of true positive instances out of all actual positive instances (TP and FN), as given in Equation 4:

$$recall = TP/(TP + FN) \tag{4}$$

F1 score measures the test accuracy, as it is the harmonic mean of precision and recall, as in Equation 5:

$$F1\ score = 2*((precision * recall)/(precision + recall)) \tag{5}$$

Specificity (true negative rate), is the proportion of true negative instances out of all actual negative instances (TN and FP), as given in Equation 4:

$$specificity = TN/(TN + FP) \tag{6}$$

A ROC curve is a graphical representation of the True Positive Rate (TPR) (on the y-axis) vs the False Positive Rate (FPR) (on the x-axis). Area Under the ROC Curve (AUC) is a quantitative measure of the overall ability of the given model to distinguish between the positive and negative classes. AUC value lies between 0 and 1. A good AUC score of 0.7 or higher being considered acceptable, and a score of 0.8 or higher is considered excellent. In medical research, a minimum score of 0.8 may be considered good enough for a model to be considered effective.

4. Results and Discussion

4.1 *Results of Data Analysis*

This study on the Pima Indians Diabetes dataset, implemented various standalone ML algorithms, along with hybrid and ensemble models. After loading the dataset, the data was analysed. As shown in Figure 3, the generated descriptive statistics provided key metrics, and exhibited no null values in the data.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768	768	768	768	768	768	768	768	768
mean	0.23	0.61	0.57	0.21	0.09	0.48	0.17	0.2	0.35
std	0.2	0.16	0.16	0.16	0.14	0.12	0.14	0.2	0.48
min	0	0	0	0	0	0	0	0	0
25%	0.06	0.5	0.51	0	0	0.41	0.07	0.05	0
50%	0.18	0.59	0.59	0.23	0.04	0.48	0.13	0.13	0
75%	0.35	0.7	0.66	0.32	0.15	0.55	0.23	0.33	1
max	1	1	1	1	1	1	1	1	1

Figure 3

Descriptive statistics of Pima Indians diabetes dataset.

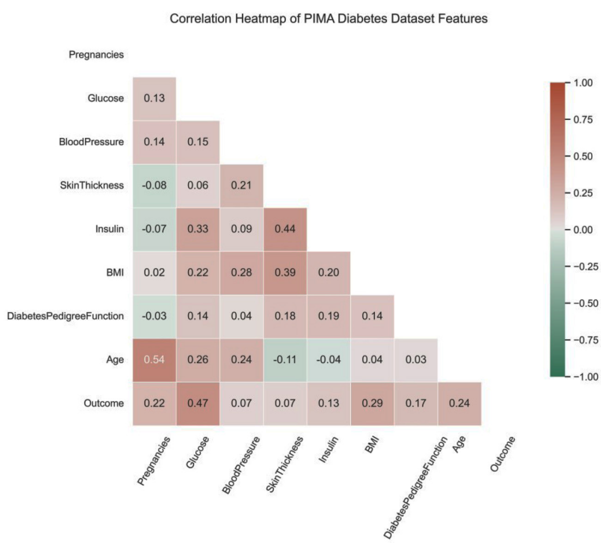

Figure 4

Correlation heatmap of Pima Indians diabetes dataset.

On further analysis of the data, the correlation of the features was studied. The resultant heatmap is given in the Figure 4. The heatmap shows the correlation lying between -0.08 ('SkinThickness' and 'Pregnancies') and 0.54 ('Age' and 'Pregnancies'), indicating weak or moderate correlation between the features of this dataset. 'Insulin' and 'Glucose'; 'Insulin' and 'SkinThickness'; 'BMI' and 'SkinThickness'; 'Age' and 'Pregnancies' show moderate positive correlation.

Additionally, the skewness of the dataset features was studied. As seen in Figure 5, it was observed that the features 'Insulin', 'DiabetesPedigreeFunction', 'Pregnancies' and 'SkinThickness' showed a skewed distribution (with skewness of 2.17, 1.92, 0.89 respectively). To avoid machine learning models giving biased predictions using the skewed distributions of certain features, feature scaling method, standardization, was applied. By rescaling the data, these techniques make model prediction less sensitive to the skewed distributions, aiding in building a robust diabetes prediction model.

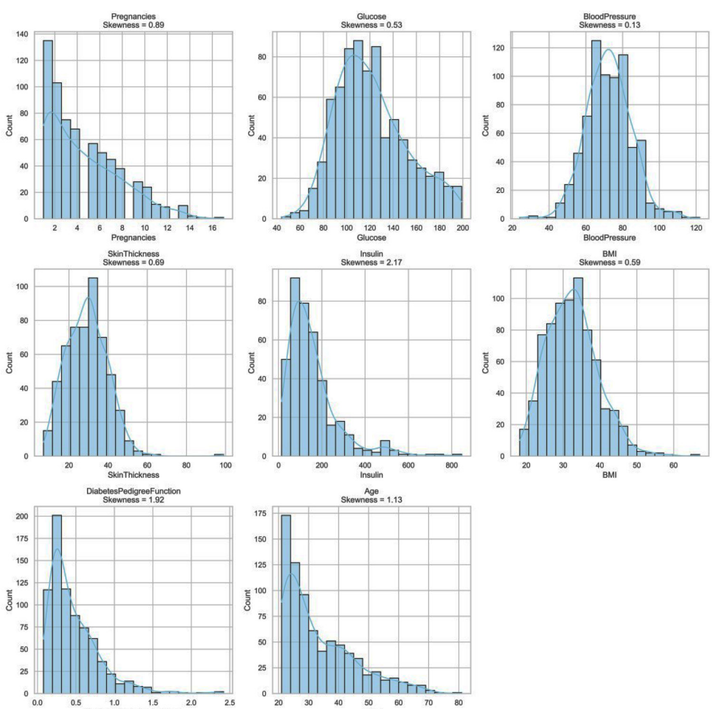

Figure 5

Skewness of parameters in Pima Indians diabetes dataset.

4.2 *Natural Therapeutics for Diabetes Mellitus*

As discussed in Section 2, and enumerated in Table 2, numerous natural remedies have been globally used in traditional medical practice, for mitigation and cure of diabetes mellitus. Numerous experiments and clinical trials that have been conducted, and have proven the efficacy of these herbs and plants. The advantage of these remedies is that, the plants are generally locally available, not expensive and usually do not have much side effects. Most of these plants and their extracts have multiple health benefits, not just for managing diabetes.

4.3 *Results of Machine Learning Models*

This study implemented four standalone machine learning models, five hybrid models and six ensemble models for diabetes prediction (whether the patient has diabetes or not). All models used the Pima Indians Diabetes Dataset. K-fold cross-validation was employed to prevent over fitting.

Evaluating the results of the models implementing standalone algorithms - gaussian naïve bayes, Support Vector Machine, Logistic Regression and K-Nearest Neighbor, it was observed that the model using the gaussian naïve bayes gave the best accuracy of 76.62%. Table 4 shows the results of the models, in the decreasing order of the accuracy. K-Nearest Neighbor demonstrated the lowest accuracy of 70.78%.

Table 4

Results of standalone machine learning models.

Model	Accuracy	Precision	Recall	F1 Score	AUC Score	5-fold Cross Validation Accuracy
Gaussian Naïve Bayes	0.7662	0.6610	0.7091	0.6842	0.7535	0.7525 (+/- 0.0272)
Support Vector Machine	0.7597	0.6667	0.6545	0.6606	0.7364	0.7622 (+/- 0.0247)
Logistic Regression	0.7532	0.6491	0.6727	0.6607	0.7354	0.7606 (+/- 0.0299)
K-Nearest Neighbor (neighbors=3)	0.7078	0.6087	0.5091	0.5545	0.6636	0.7443 (+/- 0.0271)

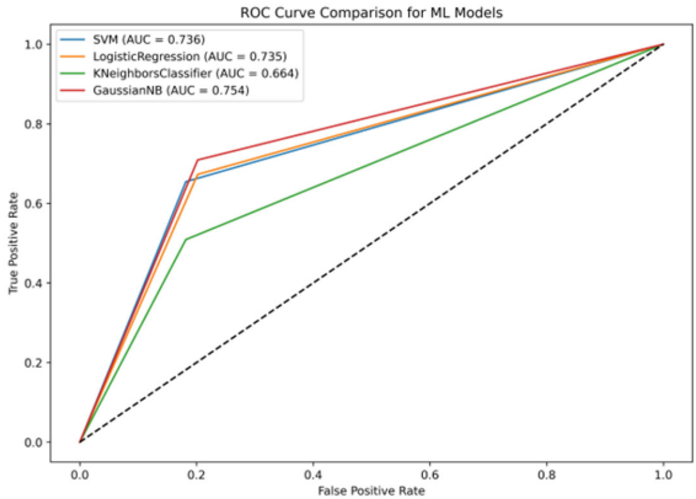

Figure 6

Comparative ROC curves for standalone machine learning models.

The AUC scores, of these standalone models ranged from 0.6636 to 0.7535, which are considered a poor (less than 0.7) to good (0.7 and greater) AUC score. Figure 6 gives the comparative ROC curves for these standalone ML models.

For hybrid models, multiple algorithms were used, one for feature selection, and another for model fit. For convenience purpose, the five models are named Hybrid1, Hybrid2, Hybrid3, Hybrid4, and Hybrid5. Models Hybrid1 and Hybrid2, used SVM for feature selection and Gaussian Naïve Bayes and ANN for model fit respectively. Hybrid3 used random forest and logistic regression for feature selection and model fit respectively. In Hybrid4, feature selection was done using LSTM and model was fit with random forest. Hybrid5 used gradient boosting and KNN for feature selection and model fit respectively.

Most of the hybrid models showed similar accuracy to the standalone models, with all models showing a higher AUC score. These models combined the strengths of the multiple algorithms. Table 5, displays the experimental results of the hybrid models, in the decreasing order of the accuracy. On evaluating the performance of the five hybrid models, it was observed that the model Hybrid1, that used SVM for feature selection and the gaussian naïve bayes for model fit, gave an excellent AUC score of 0.8253 and maximum accuracy (among hybrid models) of 76.62%. The lowest accuracy of 68.83% was observed in the Hybrid5, while it gave a good AUC score of 0.7643. Hybrid5's accuracy was even lower than the standalone models.

Table 5

Results of hybrid models

Model	Accuracy	Precision	Recall	F1 Score	AUC Score	5-fold Cross Validation Accuracy
Hybrid1 (feature selection: SVM; model fit: GNB)*	0.7662	0.6610	0.7091	0.6842	0.8253	0.7525 (+/- 0.0272)
Hybrid2 (feature selection: SVM; model fit: ANN)*	0.7597	0.6667	0.6545	0.6606	0.8154	-
Hybrid3 (feature selection: RF; model fit: LR)*	0.7532	0.6491	0.6727	0.6607	0.8147	0.7606 (+/- 0.0299)
Hybrid4 (feature selection: LSTM; model fit: RF)*	0.7208	0.6071	0.6182	0.6126	0.7911	-
Hybrid5 (feature selection: GB; model fit: KNN)*	0.6883	0.5745	0.4909	0.5294	0.7643	0.7395 (+/- 0.0283)

*Algorithms used in the models: ANN: Artificial Neural Network, GNB: Gaussian Naïve Bayes, GB: Gradient Boosting, KNN: K-Nearest Neighbor, LR: Logistic Regression, LSTM: Long-Short term Memory, RF: Random Forest, SVM: Support Vector Machine

The AUC scores, of these hybrid models ranged from 0.7643 to 0.8253, which is considered a good to an excellent score. Figure 7 gives the comparative ROC curves for these hybrid models.

The ensemble models used a combination of multiple algorithms. For the purpose of convenience, the six models are named Ensemble1, Ensemble2, Ensemble3, Ensemble4, Ensemble5, and Ensemble6. Ensemble1 model is a voting model. Ensemble2, Ensemble3 are stacking models that use linear regression for the final-estimator meta model. Boosting algorithms of AdaBoost and gradient boosting are implemented

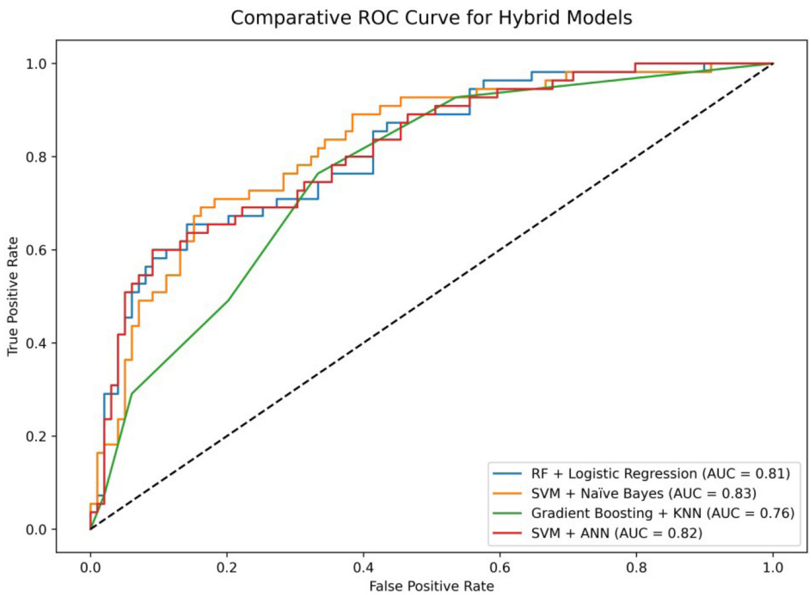

Figure 7

Comparative ROC curves for hybrid models.

in Ensemble4 and Ensemble5 models. Ensemble6 implements the random forest algorithm.

The ensemble models showed better accuracy than standalone models, with all models showing a higher AUC score, greater than 0.8. The combination of multiple algorithms makes these models robust. Table 6, displays the experimental results of the ensemble models, in the decreasing order of the accuracy. On evaluating the performance of these ensemble models, it was observed that Ensemble1, that used voting, gave the maximum accuracy of 76.62% (among the ensemble models), with an excellent AUC score of 0.8213. The lowest accuracy of 72.08% was observed in Ensemble6, while it gave an excellent AUC score of 0.8125.

The AUC scores, of these ensemble models ranged from 0.8044 to 0.8239, which are considered as excellent AUC score. Figure 8 gives the comparative ROC curves for these ensemble models.

A comparison of the performance metrices of the fifteen models (standalone algorithms: four, hybrid: five, ensemble: six), was done. The accuracy of prediction of the Outcome (no diabetes or diabetes) on the Pima Indians Diabetes dataset, gave a maximum value of 76.62% for three algorithms: Gaussian Naïve Bayes algorithm, Hybrid 1 and Ensemble1. Further analysis showed a lower AUC score of 0.7535 Gaussian Naïve

Table 6
Results of ensemble models.

MODEL	Accuracy	Precision	Recall	F1 Score	AUC Score	5-fold Cross Validation Accuracy
Ensemble1: Voting	0.7662	0.6792	0.6545	0.6667	0.8213	0.7606 (+/- 0.0269)
Ensemble2: Stacking (estimator: base model: LR, RF, GB; final_estimaor: Metamodel: LR)*	0.7597	0.6607	0.6727	0.6667	0.8239	-
Ensemble3: Stacking (estimator: base model: RF, GB, KNN; final_estimator: Meta-model: LR)*	0.7468	0.6429	0.6545	0.6486	0.8228	-
Ensemble4: Gradient Boosting	0.7468	0.6379	0.6727	0.6549	0.8072	0.7704 (+/- 0.0278)
Ensemble5: AdaBoost	0.7403	0.6415	0.6182	0.6296	0.8044	0.7606 (+/- 0.0196)
Ensemble6: Random Forest	0.7208	0.6071	0.6182	0.6126	0.8125	0.7721 (+/-0.0317)

*Algorithms used in the models: GB: Gradient Boosting, KNN: K-Nearest Neighbor, LR: Logistic Regression, RF: Random Forest, SVM: Support Vector Machine

Bayes algorithm. On the other hand, Hybrid1 model and Ensemble1 demonstrated a similar AUC score of 0.8253 and 0.8213 respectively, which are excellent scores for medical data. The fractional difference in the AUC scores make both Hybrid1 that uses and Ensemble1, candidates for selecting our proposed model. Table 7 lists the two Proposed Models for diabetes prediction.

5. Conclusion

This study used the Pima Indians Diabetes dataset to implement various machine learning models. We studied the performance metrices

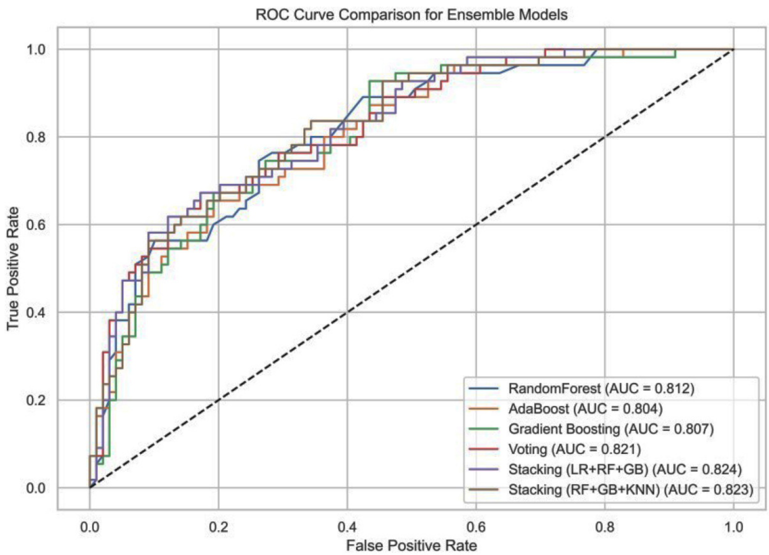

Figure 8

Comparative ROC Curves for Ensemble Models.

Table 7

Proposed model(s) for diabetes prediction.

Candidates For Proposed Model	AUC Score
Hybrid1 (feature selection: SVM; model fit: GNB)	0.8253
Ensemble1: Voting	0.8213

of the four standalone machine learning models, the five hybrid models and the six ensemble models for the prediction of diabetes (diabetes or no diabetes) was evaluated. Three of the experimental models, Gaussian Naïve Bayes algorithm, Hybrid1 and Ensemble1 demonstrated the maximum accuracy of 76.62%.

Hybrid1 model used SVM algorithm for feature selection and the Gaussian Naïve Bayes algorithm for model fit. The other performance metrices of Hybrid1 are a precision of 0.6610, recall of 0.7091 and an F1 score of 0.6842. The Ensemble1 model implemented Voting. The other performance metrices of Ensemble1 are: a precision of 0.6792, recall: 0.6545 and an F1 score of 0.6667. Further, Gaussian Naïve Bayes showed a lower AUC score of 0.7535. On the other hand, Hybrid1 and Ensemble1 demonstrated higher AUC scores of 0.8253 and 0.8213 respectively, which are excellent scores (AUC greater than 0.8). There is a fractional difference in these AUC scores, which make both Hybrid1 and Ensemble1 as candidates for our proposed model, and any one may be chosen.

The lowest accuracy of 68.83% was observed in the Hybrid5 model, despite the fact that it gave a good AUC score of 0.7643. This accuracy was even lower than all the standalone models implemented in this study. Hybrid5 implemented gradient boosting for feature selection and the model was fit using KNN.

The AUC scores of these standalone models were lower than hybrid and ensemble models, with AUC score values that are considered 'poor' or 'good' (0.7 and greater). The standalone algorithm, K-Nearest Neighbor (neighbors=3) showed the lowest AUC score of 0.6636. This is considered a poor score, as it is less than 0.7.

Thus, as seen in our study though the standalone models gave a similar accuracy, the hybrid and ensemble methods out performed in terms of the AUC score. This is because these models combine the strengths of multiple algorithms, and hence improve prediction accuracy.

As discussed in our study, traditional medical practices have used numerous plants as natural remedies for prevention and cure of diabetes mellitus. The experiments and clinical trials carried out by researchers, have proven the efficacy of these plants. These remedies give a cheaper alternative for mitigation of diabetes and in some cases, its cure. These plants are generally locally available, and generally do not have much side effects, though more research is required. Most of these plants and their extracts have multiple health benefits, not just for managing diabetes.

It may be noted that accuracy achieved in our experiments is not a benchmark, as the Pima Indians Diabetes dataset is a smaller dataset. The experiments are demonstrative of the models that gave higher performance than others. In future, the authors would implement feature engineering and use variations of hybrid and ensemble algorithms on larger diabetes datasets. The ultimate goal is to integrate an optimized prediction model with clinical practice, and provide an alternate cost-effective diabetes mellitus management program that utilizes natural remedies.

Acknowledgments

The authors would like to acknowledge DST-FIST (Department of Computer Science & Engineering, Jamia Hamdard) No. SR/FST/ET-11/2019/313 (C).

References

[1] B. Zhou *et al.*, "Worldwide trends in diabetes prevalence and treatment from 1990 to 2022: a pooled analysis of 1108 population-representative studies with 141 million participants," *The Lancet*, vol. 404, no. 10467, pp. 2077–2093 (Nov. 2024), doi: 10.1016/S0140-6736(24)02317-1.

[2] M. Balbaa, M. El-Zeftawy, and S. A. Abdulmalek, "Therapeutic screening of herbal remedies for the management of diabetes," Nov. 02 (2021), *MDPI*. doi: 10.3390/molecules26226836.

[3] A. H. Afzal, O. Alam, S. Zafar, A. Alam, and J. Khan, "Exploring the Anti-Diabetic Potential of Cichorium intybus through Integrated Network Pharmacology Analysis and Molecular Docking Validation," *Journal of Natural Remedies*, vol. 24, no. 10, pp. 2253–2261 (Oct. 2024), doi: 10.18311/jnr/2024/44047.

[4] S. M. Bhatt and N. Sharma, "A Comprehensive Review on the Role of <i>Gymenma sylvestre</i> in Combating Diabetes-2," *Journal of Natural Remedies*, pp. 2143–2155 (Nov. 2024), doi: 10.18311/jnr/2024/44121.

[5] K. N. Sharma and A. Kashyap, "Review of herbal and natural remedies for diabetes in traditional Indian medicine," (2023). [Online]. Available: https://www.virtued.in/s/pages/screji

[6] Z. Sitobo, L. Tinotenda Navhaya, and X. Henry Makhoba, "Medicinal plants as a source of natural remedies in the management of diabetes," *INNOSC Theranostics and Pharmacological Sciences*, vol. 7, no. 3, p. 1885 (Jul. 2024), doi: 10.36922/itps.1885.

[7] N. Vinothkumar *et al.*, "Deep learning and probabilistic neural networks-based identification and classification of Pneumonia-related lung diseases," *Journal of Information and Optimization Sciences*, vol. 46, no. 1, pp. 213–222 (2025), doi: 10.47974/jios-1865.

[8] S. L. Cichosz, C. Bender, and O. Hejlesen, "Explainable Machine Learning-Based Approach to Identify People at Risk of Diabetes Using Physical Activity Monitoring," *BioMedInformatics*, vol. 5, no. 1 (Mar. 2025), doi: 10.3390/biomedinformatics5010001.

[9] L. M. S. Akoosh, F. Siddiqui, S. Zafar, S. Naaz, and M. A. Alam, "Enhancing Heart Health Prediction with Natural Remedies Through Integration of Hybrid Deep Learning Models," *Journal of Natural*

Remedies, vol. 24, no. 11, pp. 2481–2491 (Nov. 2024), doi: 10.18311/jnr/2024/44045.

[10] M. Adlakha, S. Naaz, S. Zafar, and M. A. Alam, "Deep Learning Approach for Accurate Prediction of diabetes," in *Proceedings of the 5th International Conference on Information Management & Machine Intelligence*, New York, NY, USA: ACM, pp. 1–6 (Nov. 2023). doi: 10.1145/3647444.3647832.

[11] M. Adlakha, M. A. Alam, S. Zafar, and S. Naaz, "Prediction Optimization for Type 2 Diabetes Mellitus Using Artificial Neural Networks," in *Proceedings of the 6th International Conference on Communication and Intelligent Systems*, Springer Nature (2025). doi: 10.1007/978-981-96-5723-0_2.

[12] K. Oliullah, M. H. Rasel, M. M. Islam, M. R. Islam, M. A. H. Wadud, and M. Whaiduzzaman, "A stacked ensemble machine learning approach for the prediction of diabetes," *J Diabetes Metab Disord*, vol. 23, no. 1, pp. 603–617 (Jun. 2024), doi: 10.1007/s40200-023-01321-2.

[13] M. A. Rahim, M. A. Hossain, M. N. Hossain, J. Shin, and K. S. Yun, "Stacked Ensemble-Based Type-2 Diabetes Prediction Using Machine Learning Techniques," *Annals of Emerging Technologies in Computing*, vol. 7, no. 1, pp. 30–39 (2023), doi: 10.33166/AETiC.2023.01.003.

[14] Md. Rashed, Md. I. Hossain, A. Mahdi, and G. Mustofa, "Hybrid Machine Learning Models for Accurate Type 2 Diabetes Mellitus Prediction Using a Stacking Classifier and a Meta-Model Approach," *Cureus Journal of Computer Science* (Mar. 2025), doi: 10.7759/s44389-025-03135-0.

[15] M. Nagassou, R. W. Mwangi, and E. Nyarige, "A Hybrid Ensemble Learning Approach Utilizing Light Gradient Boosting Machine and Category Boosting Model for Lifestyle-Based Prediction of Type-II Diabetes Mellitus," *Journal of Data Analysis and Information Processing*, vol. 11, no. 04, pp. 480–511 (2023), doi: 10.4236/jdaip.2023.114025.

[16] O. Olabanjo, A. Wusu, O. Olabanjo, M. Asokere, O. Afisi, and B. Akinnuwesi, "A novel deep learning model for early diabetes risk prediction using attention-enhanced deep belief networks with highly imbalanced data," *International Journal of Information Technology (Singapore)*, vol. 17, no. 4, pp. 1933–1955 (May 2025), doi: 10.1007/s41870-025-02459-3.

[17] P. Chowdhury, P. Barua, and M. N. Uddin, "Diabetes Prediction Using Machine Learning and Hybrid Deep Learning Ensemble Technique,"

in *2024 IEEE Conference on Computing Applications and Systems, COMPAS 2024*, Institute of Electrical and Electronics Engineers Inc. (2024). doi: 10.1109/COMPAS60761.2024.10796486.

[18] R. K. Sharma and M. Jailia, "Machine learning and IoT-based garbage detection system for smart cities," *Journal of Information and Optimization Sciences*, vol. 44, no. 3, pp. 393–406 (2023), doi: 10.47974/jios-1349.

[19] P. R. Anisha and K. Anjali, "CyberShield-H: A hybrid sequential deep learning model for robust DDoS mitigation in critical healthcare IoT systems," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 28, no. 7, pp. 2889–2898 (Oct. 2025), doi: 10.47974/JDMSC-2559.

[20] S. Nimbagal, A. Sarangmath, M. Unakal, and A. Gongade, "Factors influencing the acceptance of mobile health apps among Indians : A behavioral perspective," *Journal of Statistics and Management Systems*, vol. 28, no. 8, pp. 1433–1442 (2025), doi: 10.47974/JSMS-1506.

[21] A. Kumar, K. M. Bhagat, P. Dadheech, S. R. Dogiwal, and R. Sharma, "Manifold learning and differential geometry for disease detection and risk prediction," *Journal of Interdisciplinary Mathematics*, vol. 28, no. 6, pp. 2059–2069 (2025), doi: 10.47974/jim-2347.

[22] D. Samuel and M. Anita Rajam, "MXLPred: A Natural Language Processing Deep Learning Model for ICU Patients' Data," *The International Arab Journal of Information Technology*, vol. 22, no. 4 (2025), doi: 10.34028/iajit/22/4/12.

[23] "Clerodendrum glandulosum." Accessed: Sep. 25 (2025). [Online]. Available: https://en.wikipedia.org/wiki/Clerodendrum_glandulosum

[24] S. Alam *et al.*, "Antidiabetic Phytochemicals From Medicinal Plants: Prospective Candidates for New Drug Discovery and Development," Feb. 24, , *Frontiers Media S.A* (2022). doi: 10.3389/fendo.2022.800714.

[25] Brittany Lubeck, "Everything You Need to Know About Maitake Mushroom." Accessed: Sep. 25 (2025). [Online]. Available: https://www.verywellhealth.com/maitake-mushroom-benefits-nutrition-8606730

[26] M. A. S. Hairani, F. A. Abdul Majid, N. H. Zakaria, D. Hudiyanti, A. Fadhlina, and H. I. Sheikh, "Anti-diabetic properties of traditional herbal concoction containing Eleutherine palmifolia (L.) Merr., Momordica charantia L., and Syzygium polyanthum (Wight.):

a bibliometric analysis," Dec. 01 (2023), *BioMed Central Ltd*. doi: 10.1186/s43014-023-00172-x.

[27] Brittany Lubeck, "Nigella Sativa (Black Seed) and What It Can Do For Your Health." Accessed: Sep. 25 (2025). [Online]. Available: https://www.verywellhealth.com/everything-you-should-know-about-nigella-sativa-7775388

[28] A. Charkravarty and G. Ramasamy, "Profiling Of Volatiles In Tissues Of Salacia Reticulata Wight. With Anti-Diabetic Potential Using GC-MS And Molecular Docking,"

[29] S. Deyno, K. Eneyew, S. Seyfe, and E. Wondim, "Efficacy, safety and phytochemistry of medicinal plants used for the management of diabetes mellitus in Ethiopia: a systematic review," *Clinical Phytoscience*, vol. 7, no. 1 (Dec. 2021), doi: 10.1186/s40816-021-00251-x.

[30] "Pima Indians Diabetes Database," Kaggle. Accessed: Jun. 20 (2025). [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

Received August, 2025