International Research Journal of Multidisciplinary Scope (IRJMS), 2025; 6(2): 1121-1138 IRJMS

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2024.v06i02.03513

An Effective Diabetes Mellitus Classification Model for Diabetes Predictions and Diagnosing Using Machine Learning Techniques

Prag Jain[1]*, Nidhi Tyagi[1], Birendra Kumar Sharma[2]

[1]Shobhit Institute of Engineering and Technology (Deemed to be University), Meerut, India, [2]Ajay Kumar Garg Engineering College, Ghaziabad, India. *Corresponding Author's Email: pragjain@gmail.com

Abstract

This paper applies onerous machine learning methods to enhance the precision and potency of diabetes mellitus (DM) analysis and classification. Then, it presents HOMED, a novel hybrid online model using Adaptive Principal Component Analysis (APCA) and Incremental Support Vector Machine (ISVM) algorithms to overcome typical overfitting, missing data imputation, and computational inefficiency problems experienced by traditional models. The model was evaluated on 768 records from the Pima Indian Diabetes Dataset (PIDD), and on 1,099 records on the Diabetes Treatment Dataset (DTD). When compared to well-known models such as Decision Trees (DT), Random Forests (RF), Naïve Bayes (NB), and k-Nearest Neighbors (k-NN), HOMED yielded higher accuracy (80.34%), higher sensitivity (92.67%), and higher specificity (98.43%). The ability to deal with high dimensional data while preserving precision and data integrity provides evidence for its potential to be a robust data tool for real time medical diagnosis. HOMED takes advantage of machine learning by bringing it to the support system for making clinical decision, to make healthcare practices more scalable, and more adaptive to early disease detection and effective patient management. The findings can be key for policymakers and healthcare practitioners to understand the need for using the most advanced algorithms to support sustainable, data driven medical practices. This research speaks to the use of technology to improve care outcomes worldwide.

Keywords: Clinical Decision Support Systems, Diabetes Classification, Hybrid Models, Machine Learning, Sustainable Healthcare.

Introduction

The healthcare industry, a critical sector for economic growth and employment, has seen significant industrialisation in countries such as the United States, China, and the United Kingdom (1-3). Global healthcare expenditures are projected to double in the next five years (4). In India, public health spending is expected to rise significantly, from 267,000 crores in 2018 to 486,000 crores by 2022, representing a 45% increase over five years. However, a substantial portion of healthcare costs stems from non-value-added practices, including incorrect diagnoses, prescription errors, antibiotic misuse, readmissions, and fraud. Annually, approximately 5.2 million deaths in India are attributed to medical errors and related issues (5). To mitigate these challenges, Clinical Decision Support Systems (CDSS) have emerged as transformative tools, expediting the transition to value-based healthcare. CDSS leverage risk analysis for disease progression and treatment strategies, enhancing the quality of care for both patients and medical professionals (6). The delivery of healthcare has been greatly enhanced by CDSS, which has decreased drug errors and diagnostic inaccuracy. Smarter and more efficient clinical decision-making is made possible by modern systems that integrate cutting-edge platforms including wireless sensor networks, data analytics, the Internet of Medical Things (IoMT), artificial intelligence (AI), and machine learning (ML). The use of machine learning and data mining techniques is crucial in the present healthcare environment to transform enormous volumes of raw data into insights that can be used to inform therapeutic choices. The diagnosis and treatment of diabetes mellitus (DM) is a crucial area where machine learning (ML) can have a revolutionary

(Received 13th December 2024; Accepted 21st April 2025; Published 30th April 2025)

Impact. Diabetes mellitus, a deadly and chronic metabolic disease, is a major global health concern, especially in underdeveloped nations. Alarming patterns are revealed by recent epidemiology studies: Currently, 463 million people between the ages of 20 and 79 have diabetes; by 2030, that figure is predicted to increase to 552 million, and by 2045, it will reach 628 million (7-9). With one death every six seconds and serious complications from decreased immunity and co-occurring infections, diabetes is the fifth greatest cause of death worldwide. High rates of morbidity and mortality are the outcome of these variables, which also make managing the disease more difficult.

Managing diabetes mellitus is a difficult and resource-intensive endeavor for medical practitioners. Accurate and fast diagnostic data are essential for effective treatment. However, there are a lot of obstacles facing conventional ML-based diagnostic systems. These algorithms are susceptible to overfitting, which diminishes their efficacy, because high-dimensional datasets frequently contain redundant or unnecessary features. Learning and testing are the two phases of classification in traditional Support Vector Machine (SVM) models. Features are taken from training data during the learning phase, and they are used to categorize fresh data samples during the testing phase. SVM's capacity to effectively handle high-dimensional and nonlinear data makes it especially well-suited for use in medical applications.

Traditional algorithms have drawbacks such trapping in local optima and a lack of incremental classification skills, despite the fact that many Disease Management Systems (DMS) use machine learning (ML) for DM diagnosis. For new classifications, traditional supervised machine learning approaches frequently necessitate retraining the entire dataset, which increases computing overhead and inefficiency. It is stated that 537 million people worldwide have diabetes. According to statistics data, 529 million people worldwide had diabetes in 2021. By 2050, the population is expected to reach 1.31 billion (10). Although the percentage of people with diabetes varies slightly from country to country, diabetes remains one of the world's major causes of death and disability. This holds true regardless of national borders, gender roles, or age groups. The cost of treating diabetes-related issues places a significant and continuously increasing financial strain on healthcare systems. The number of people with diabetes is rising significantly worldwide, which is a serious health issue. The timely and precise identification of diabetic mellitus (DM) through clinical information is essential for both therapy selection and disease prevention. Machine learning (ML)-based detection models have historically been developed as offline, non-incremental methods that use pre-existing datasets to train. In order to successfully diagnose DM, this study suggests a novel hybrid online model called HOMED that incorporates cutting-edge algorithms. Addressing these gaps necessitates innovative approaches to improve diagnostic accuracy and adaptability.

Diabetes Mellitus

DM has been studied for centuries. The term "diabetes" originates from Aretaeus, a 5th-century physician, who described it as a "melting down of limbs and flesh into urine." The term "mellitus," meaning "honey" in Latin, was later added due to the sweet taste of urine observed by Indian physicians in 5th century BC. This characteristic attracted ants, leading to early diagnostic methods. DM is a chronic non-communicable disease characterised by partial or total insulin deficiency, resulting in disturbances in the metabolism of glucose, fats, proteins, and carbohydrates. It leads to severe complications, including reduced life expectancy, high morbidity from microvascular issues (e.g., neuropathy, and nephropathy), macrovascular complications (e.g., peripheral vascular disease, stroke, and heart disease), infections, and diminished quality of life. The disease is considered one of the most pressing public health challenges due to its prevalence and the frequent comorbidities that accompany it.

In India, the prevalence of diabetes has been rising rapidly, with 8–10% of the population affected, and urban areas showing higher incidence rates than rural regions (11). Alarmingly, studies indicate that 46% of Delhi's population is prediabetic, with DM prevalence at approximately 27%. Similar trends are observed in other metropolitan cities. The age group of 30 to 34 years old has the highest frequency. The rising prevalence of diabetes places a significant financial

burden on the healthcare system, making accurate and effective diagnostic technologies essential (12).

The COVID-19 pandemic significantly increased the workload of healthcare workers, particularly doctors and nurses, who faced a sudden and substantial rise in patient numbers. During such crises, artificial intelligence (AI) techniques have shown their potential to assist in identifying critical illnesses. Conditions such as diabetes mellitus (DM), high blood pressure, and heart disease, which are associated with higher mortality and hospitalisation risks in coronavirus patients, underscore the importance of early detection. Diabetes diagnosis and management are inherently data-intensive processes, making machine learning (ML) techniques an ideal choice for enhancing outcomes and adopting advanced approaches.

Machine learning has been employed in diabetes diagnosis systems (DDS) to handle various data-related tasks. Beyond healthcare, ML has proven effective in fields like image processing (13), web searches, speech and picture recognition, spam filtering, fraud detection, and credit ratings. These capabilities highlight the versatility and transformative potential of ML across diverse applications. Traditional supervised machine learning techniques often require retraining the entire dataset for new classifications, increasing computational overhead and inefficiency. These steps reduce the risk of morbidity by promptly identifying and predicting DM. By training a large number of real-world databases, the DDS's data classification allows physicians and caregivers to explore unknown datasets. Diabetes is one of the major global health issues. Therefore, it is critical to recognize the importance of hyperglycemia and to promptly detect, examine, and treat it.

Hybrid Models for Diabetes Diagnosis

For Type 2 Diabetes Mellitus (T2DM), a hybrid DDS model technique that combines K-means clustering with Random Forest (RF) and XGBoost algorithms effectively found hidden features in databases, producing accurate results (14). The Pima Indian Diabetes Dataset (PIDD) assessed the performance of RF, K-means clustering, and Artificial Neural Networks (ANN) for the categorization of diabetes mellitus. The results showed that ANN had an accuracy rate of 75.7%, demonstrating its usefulness in supporting clinical decision-making (15). T2DM classification in rural populations using ANN and Multivariate Logistic Regression (MLR) models, with a remarkable 89.1% accuracy rate (16). DNN and SVM-based ML techniques have also been used in DM diagnosis. A system for classifying diabetes using real-world datasets with 14 attributes that combines ANN, Decision Tree (DT), and RF, with an 80.84% classification accuracy (17).

SVM-Based Approaches

The SVM-based classifier model achieved classification accuracies of 83.33% for microaneurysms and optic discs and 91.67% for blood vessels and exudates (18). A hybrid approach integrating K-means clustering and SVM, using longitudinal datasets to enhance DM classification accuracy (19). The combined convolutional neural networks (CNN) with SVM to extract high-level features and classify DM with improved precision (20).

Decision Tree and Ensemble Methods

A DT-based classification algorithm for diabetes diagnosis, achieved an accuracy of 75.8% (21). By leveraging World Health Organization (WHO) classification standards, the study demonstrated that each feature effectively distinguished between diabetic and non-diabetic cases (22).

Advanced Architectures for Diabetes Detection

A three-tier healthcare system using ANN to monitor vital signs utilised robust servers for database management and social welfare operations, offering a comprehensive approach to disease diagnosis (23). NB to be the most accurate as compared SVM, DT, and NB classifiers for diabetes prediction using PIDD with K-10 cross-validation (24).

Machine learning methods, assessing the performance of RF, NB, and neural networks (NN) using Matthew's correlation coefficient (25). Ensemble methods combining NB, RF, and Logistic Regression (LR), achieving a classification accuracy of 79% (26). A strong interpretability tool is Shapley Additive explanations (SHAPs), which are based on cooperative game theory and precisely the Shapley values that quantify each player's involvement to the game (or prediction

model) as a whole (27). SHAP values are essential for determining how each dataset feature affects the prediction result in diabetes detection, giving precise information about the variables. Because SHAP values provide each feature a "contribution" to the prediction, they offer a means of comprehending why a machine learning model generates a specific prediction. Since SHAP is model-agnostic, it may be used to clarify the predictions of a number of ML models that involve those that are based on trees, such as Random Forest and XGBoost.

In medical applications, Local Interpretable Model-agnostic Explanations (LIME) assist physicians comprehend and trust AI predictions, improving patient outcomes by elucidating the methods by which these models arrive at certain diagnostic or prognostic conclusions. For each prediction model, explanation is provided at the individual level using LIME. By using the LIME technique, the prediction probability for each risk factor was displayed at the individual level. This could help medical professionals make individualized judgments and address public health issues. The predictions of machine learning classifiers can be interpretably explained by LIME (28).

Feature Selection and Preprocessing Techniques

Feed-forward multilayer neural networks for diabetes classification, normalising dataset values between 0 and 1 to enhance numerical stability (29). Authors compared SVM and NB classifiers using K-fold cross-validation, with SVM outperforming NB (30). In order to classify PIDD, PCA and linear discriminant analysis have been investigated; k-NN and radial basis kernel approaches have been found to achieve excellent accuracy (31). Using classifiers like decision trees, bootstrap, and adaptive boosting, and integrating characteristics like blood pressure, triglycerides, and BMI. Recent developments in hybrid models, which employ an ensemble or stacking technique to incorporate many classifiers. Key research in the subject of type 2 diabetes prediction are compiled in Table 1, which also details the solution approaches used, performance metrics attained, and study constraints.

Table 1: Summary of Existing Studies

Work	Method	Accuracy (%)	Limitations	References
Application of Supervised ML for Type 2 DM Prediction	Random Forest	93.8	Needs more intricate interpretation and greater capacity for processing.	(32)
Diabetes Detection and Classification	Random Forest	96.75	High computing costs as a result of random forest complexity, component evaluation, and kernel entropy	(33)
Diabetes Prediction Using a Healthcare Framework	AdaBoost	80	Moderate applicability and accuracy.	(34)
Diabetes Detection Using Artificial Neural Network (ANN)	ANN	80.79	Needs a substantial quantity of data to train; lesser datasets may cause overfitting.	(35)
E-Diagnosis System for Diabetes Using ML Models	Naïve Bayes classifier, Random Forest classifier	79.57	Without model fine-tuning, the decision-making process can be challenging to understand.	(36)
Diabetes Prediction Using Deep Dense	DDLNN	84.42	K-fold cross-validation and hyperparameter	(37)

Layer Neural Network (DDLNN)			adjustment are necessary; more intricate datasets or features would be advantageous.	
Diabetes Prediction	Random Forest	83	Restricted to the chosen features.	(38)
Diabetes Prediction Using Deep Learning	Convolutional Neural Network	92.31	Needs a balanced dataset, is computationally demanding.	(39)

Methodology

The main problems with traditional ML-based diabetes classification algorithms are missing values in the input dataset, loss of privacy, trapping into a local optimum solution, and inadequacy of incremental classification. In contrast to conventional machine learning (ML) models like SVM or Random Forest, which treat data as flat labels and lack this essential interpretability, HOMED models provide visible, explainable reasoning, which makes them advantageous in the management of diabetes. The goal of HOMED is to create a secure diabetes categorization model that safeguards sensitive user data. With the use of Kronecker product-based CSA, which maximizes the Secrecy Usage parameter, this study creates a secure DDS to safeguard user information. The accurate and prompt detection of DM using clinical datasets is crucial for disease prevention and therapy selection. Traditionally, machine learning (ML)-based detection models have been designed as non-incremental, offline techniques that learn from predefined datasets. This study proposes a novel hybrid online model, named HOMED, which integrates advanced algorithms to identify DM effectively. The system employs two primary methods such as Enhanced Incremental Support Vector Machines (ISVM) for incremental classification and Adaptive Principal Component Analysis (APCA) for attribute selection, data clustering, and handling missing data.

The Pima Indian Diabetes Dataset (PIDD) serves as the primary dataset for performance assessment, evaluated using metrics. The hybrid framework addresses the limitations of traditional offline classifiers by significantly enhancing efficiency and reducing processing costs, enabling healthcare professionals to make timely decisions for DM management.

The effectiveness of the proposed method is evaluated by comparing its performance to that of the other recognized classification strategies. HOMED significantly reduces processing overhead and improves forecast accuracy as compared to earlier offline techniques. Several performance indicators are used to evaluate the effectiveness of this classifier. Machine learning techniques to improve the accuracy and effectiveness of DM categorization and analysis. Adaptive Principal Component Analysis (APCA) and Incremental Support Vector Machine (ISVM) methods are used in HOMED, a unique hybrid online model, to address common issues with overfitting, missing data imputation, and computational inefficiency that traditional models face. HOMED leverages machine learning by integrating it into clinical decision-making support systems, increasing the scalability and adaptability of healthcare practices to early disease detection and efficient patient care. Policymakers and healthcare professionals may find the findings crucial in comprehending the necessity of employing the most cutting-edge algorithms to promote data-driven, sustainable medical practices. The primary issues with conventional ML-based diabetes classification algorithms include insufficient incremental categorization, privacy loss, trapping in a local optimum solution, and missing values in the input dataset. HOMED models are useful in the treatment of diabetes because they offer visible, explicable reasoning, in contrast to traditional ML models like SVM or Random Forest, which treat data as flat labels and lack this crucial interpretability.

Hybrid Online Model for Early Detection of Diabetes

Managing diabetes mellitus is a complex, resource-intensive challenge for medical professionals. High-dimensional datasets with redundant features often lead to overfitting, making traditional ML techniques inefficient. HOMED introduces a hybrid incremental framework that leverages the strengths of both ISVM and APCA to address these challenges. APCA uses the Expectation-Maximization (EM) technique to combine attribute selection, data clustering, and missing data imputation. Retaining important data patterns while minimizing computing complexity is the main objective. Managing diabetes mellitus is a difficult and resource-intensive endeavor for medical practitioners. Accurate and fast diagnostic data are essential for effective treatment. However, there are a lot of obstacles facing conventional ML-based diagnostic systems. These algorithms are susceptible to overfitting, which diminishes their efficacy, because high-dimensional datasets frequently contain redundant or unnecessary features. Traditional algorithms have drawbacks such trapping in local optima and a lack of incremental classification skills, despite the fact that many Disease Management Systems (DMS) use machine learning (ML) for DM diagnosis. For new classifications, traditional supervised machine learning approaches frequently necessitate retraining the entire dataset, which increases computing overhead and inefficiency. DM is identified and predicted in a timely manner using these measures to lower the risk of morbidity. The DDS's data classification enables doctors and caregivers to investigate unknown datasets by training a significant number of real-world databases. One of the biggest global health concerns is diabetes. The significance of hyperglycemia and its prompt detection, examination, and treatment are therefore vital.

System Architecture

The architecture of HOMED is illustrated in Figure 1. The system combines APCA for dimensionality reduction and ISVM for efficient classification of incremental datasets. The proposed APCA assists DDS models in removing inappropriate characteristics, which reduces learning time and expense while simultaneously boosting DDS enactment. In order to accomplish data clustering, the APCA uses the Gaussian mixture model, and the model variables are calculated using the expectation-maximization technique. The technique of attribute extraction is essential for enhancing classifier performance, particularly when working with high-dimensional datasets. Removing superfluous and inappropriate characteristics is an important preprocessing technique. In order to resolve the association issues, which present difficulties for disease detection techniques seeking to determine the correlation between the data, the original set of attributes is converted using a dimensionality reduction technique based on principal component analysis.

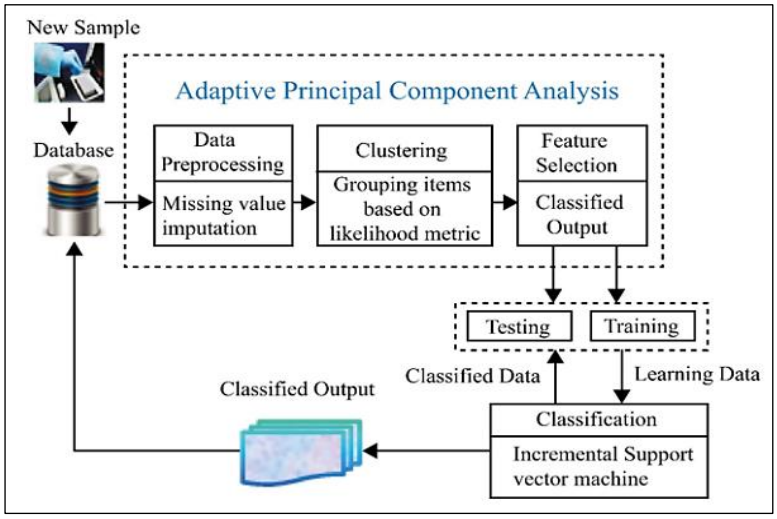

Figure 1: System Architecture of HOMED

Steps in the Hybrid Model

Data Preprocessing: Handles missing values using APCA.

Dimensionality Reduction: APCA reduces the feature space while preserving key variations.

Incremental Classification: ISVM continuously updates the classifier with new data, eliminating the need for full retraining.

Adaptive Principal Component Analysis (APCA)

APCA combines attribute selection, data clustering, and missing data imputation using the Expectation-Maximisation (EM) algorithm. The primary goal is to minimise computational complexity while retaining key data patterns.

Steps in APCA:

Weighted Adaptive Imputation: Missing values are imputed using weighted averages of neighbouring data points.

Clustering via EM Algorithm: Data samples are grouped into clusters based on Gaussian Mixture Models (GMM).

Dimensionality Reduction: The feature space is reduced by identifying the most relevant attributes using PCA.

SVM-Based Diabetes Diagnosis System

The classification process in HOMED employs a hybrid Crow Search Algorithm (CSA) and Binary Grey Wolf Optimisation (BGWO)-based SVM classifier. This section outlines the mathematical foundation of the SVM classifier and its optimisation process.

SVM Classifier

Assume $(x\text{i,i})_{i=1}$ is the set of learning data.

Here, represents the input datasets, and$y \in \{+1, -1\}$ is the class label. Then the hyperplane isdefined inEquation [1].

$$w^T.x+b=0 \quad [1]$$

The hyperplane is perpendicular to a coefficient vector (w) in Equation (1). The term b indicates the distance in the database between the origin and the point.

Finding the values of b and w is the SVM's main objective. To create an optimal hyperplane, $\|w\|^2$should be reduced under theconstraintof$y$i ($w^T.\ x$xi + b) ≥ 1 as given in Figure 2. SVM classification seeks to maximize the margin between data points the distance between the hyperplane and the nearest data points of each class, in order to identify the optimal hyperplane for classifying data points. In order to classify data, the supervised machine learning method SVM finds the best line or hyperplane in an N-dimensional space that optimizes the distance between each class. Therefore, the optimization problem is model ledas Lagrange multipliers are used by SVM to solve linear problems. The data points that are located on the judgement margin are the support vectors in this approach. Consequently, the value of w can be calculated as given in Equation [2].

$$w=\sum_{i=1}^{n} \alpha i y i x i \quad [2]$$

where αi signifies the Lagrange multipliers and n denotes the number of SVs.Once w is calculated, the value of b can be computed using Equation [3].

$$y\text{i}(w^T.x\text{i}+b)-1=0 \quad [3]$$

The linear discriminating function can be defined as givenin Equation [4].

$$y = sgn\left(\sum_{i=1}^{n} \alpha i y i x^T x\text{i}+b\right) \quad [4]$$

The kernel trick is applied by SVM to resolve non-linear problems. Equation [5] then defines the choice function.

$$y = sgn\left(\sum_{i=1}^{n} \alpha i y i K(x\text{i},x)+b\right) \quad [5]$$

In this case, y represents the kernelized label for the unclassified input x, and the sgn function indicates whether positive or negative classification results are expected. Typically, any positive definite kernel functions including Gaussian function $(x\text{i}, x) = \exp(-\gamma\|x - x\text{i}\|^2)$ and the polynomial function $K(x\text{i}, x) = (x^T x\text{i} + 1)^d$ that satisfy Mercer's limitation.

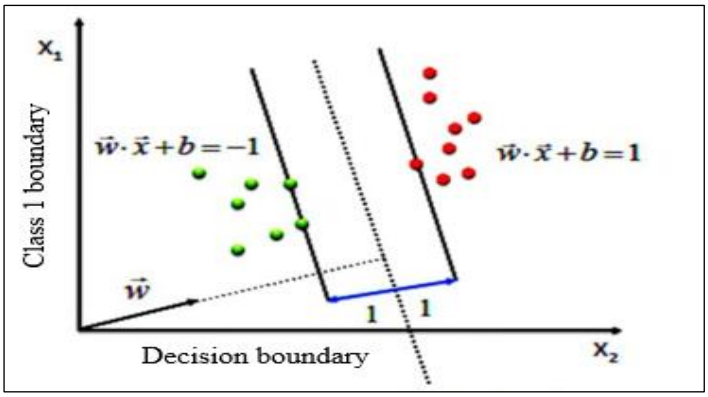

Figure 2: SVM Classification Process

Optimisation with CSA and BGWO

The hybrid optimiser integrates CSA and BGWO to enhance SVM performance.

Crow Search Algorithm (CSA): Simulates crows' food caching behaviour to identify the optimal solution.

Binary Grey Wolf Optimisation (BGWO): Models grey wolf hunting strategies for refining solutions.

The optimisation process is summarised as follows:

- Initialise crow and wolf populations.
- Use CSA to find initial hyperplane parameters.
- Refine parameters with BGWO to maximize classification accuracy.

Experimental Dataset

The empirical analysis uses the following datasets.

Pima Indian Diabetes Dataset (PIDD): Contains 768 samples with attributes like plasma glucose, BMI, and age as shown in Table 2.

Diabetes Treatment Dataset (DTD): Includes 1,099 samples, focusing on HbA1c levels and fasting glucose.

Dataset Features:

PIDD: Attributes include plasma glucose, serum insulin, BMI, age, and more.

DTD: Attributes include HbA1c, fasting glucose, and postprandial glucose levels.

Table 2: PIDD Dataset Statistics

Feature	Mean	Min	Max
Plasma Glucose (mg/dL)	120.9	0	199
BMI (kg/m^2)	32.0	0	67.1
Age (years)	33.2	21	81

Results

This section outlines the application and PIDD-based empirical analysis of the suggested model. A disease diagnosis model's accuracy and classification performance are critical factors that must be demonstrated before it can be put to use in a real-world setting. In order to assess HOMED's efficacy, its performance is compared to five other cutting-edge online classification techniques. The aforementioned writers have adjusted the original SVM, RF, NB, and k-NN classification algorithms to efficiently categories online data samples. Several evaluation metrics, including accuracy, precision, specificity, sensitivity, PPV, and NPV, are used in the assessment process of online diabetes classifiers.

The results of many categorization techniques, including the suggested model, are shown in Table 3. The findings show that whereas Naïve Bayesian (NB) can properly identify 68.92% of data samples, the KNN-based online classification model can only categories 64.43% of data samples. The results of the DT classifier were 75.54%, random forest 77.95%, and SVM 79.33%, in that order. Furthermore, as Figure 3 illustrates, the suggested HOMED classifier proved to be the most accurate classification model with an accuracy of 80.34%. The accuracy of all the Classification techniques is shown in Figure 3. It has been observed from the Figure 4 that the maximum accuracy is noticed from HOMED.

Table 3: The Results of Categorization Methods

Classification Technique	Correctly Classified	Incorrectly classified	Accuracy (%)
Naïve Bayesian	132	62	68.92
KNN	47	87	64.43
Decision Tree	565	208	75.54
Random Forest	582	192	77.95
SVM	598	174	79.33
HOMED	605	170	80.34

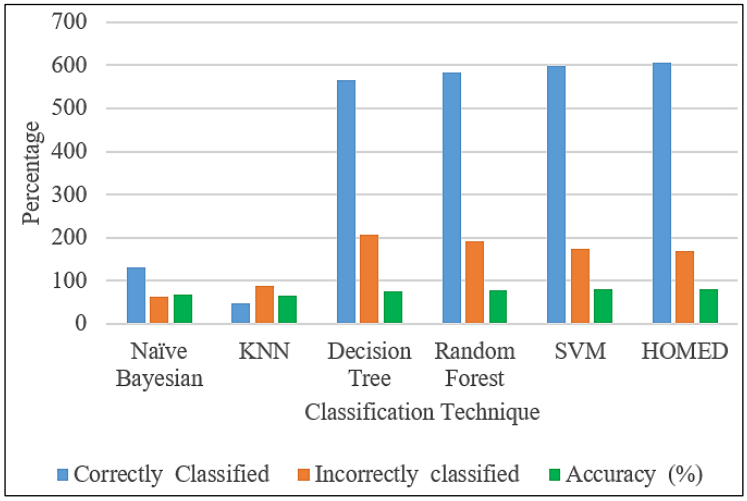

Figure 3: Comparing Different Categorization Methods

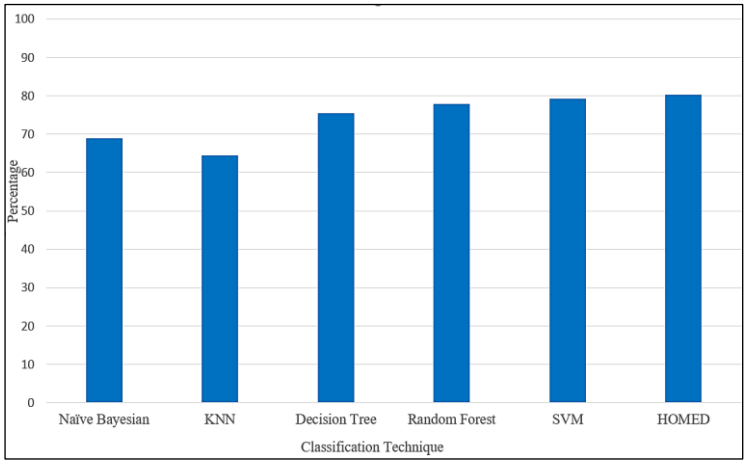

Figure 4: The Accuracy of Classification Techniques

The sum of diagonals on the matrix indicates how many occurrences were successfully identified; all other examples were misclassified. The degree to which a new test value resembles an anticipated value is called accuracy. It is expressed as the percentage of cases in the collection that are correctly classified relative to all cases. If, at least on the training set, every instance was correctly classified and every mistake was 0, it is obvious that the algorithm is a perfect classifier. In fact, though, such is not the case. Thus, we may acknowledge that the method with the highest number of correctly classified instances or the lowest number of mistakenly categorized instances is the best classifier. According to a number of assessment metrics, the performance of the constructed classifiers is contrasted with that of other classification methods, as shown in Table 4. Based on the data shown in this table, it can be observed that the suggested HOMED model outperforms the other online classification methods in terms of precision (83.54%). In

contrast to the KNN (76.43%), NB (79.23%), DT (82.94%), RF (81.54%), and SVM (79.83%) algorithms, the suggested approaches for classification, clustering, and missing value imputation help to improve the proposed classifier's precision. As indicated in Table 4, HOMED achieves 98.43% specificity, 93.68% sensitivity, 81.89% negative predicted value, and 90.54% positive predicted value.

Table 4: Comparative Evaluation of Categorization Techniques Using Different Metrics

Algorithm	Positive Predicted	Negative Predicted	Sensitivity	Specificity	Precision
Naïve Bayesian	64.54	81.43	63.09	82.54	79.23
KNN	48.84	73.65	49.47	74.18	76.43
Decision Tree	90.73	73.19	28.04	98.22	82.94
Random Forest	90.43	74.54	35.76	98.91	81.54
SVM	75.43	82.54	60.34	89.43	79.83
HOMED	90.54	81.89	92.67	98.43	83.54

In comparison to KNN, NB, DT, RF, and SVM, the suggested HOMED offers better specificity, sensitivity, positive predictive value, and negative predictive value. The HOMED classification algorithm's performance metrics are improved by the suggested techniques, which include clustering, missing value imputation, and classification. The efficiency of HOMED in classifying DM is seen in Figure 5 and HOMED performance in terms of accuracy is shown in Figure 6. The results demonstrate that, in comparison to previous offline methods, HOMED considerably increases forecast accuracy and lowers processing overhead. A number of performance metrics are used to assess this classifier's efficacy. The performance metrics of the suggested strategy are compared to those of the other five classification techniques currently in use, namely SVM, RF, NB, DT, and K-NN-based classifiers, in order to assess its efficacy. DM is a severe chronic endocrine disease characterized by elevated plasma sugar levels in affected persons. This endeavour aims to improve the lives of people by categorizing patient laboratory test findings and preventing the early repercussions of DM through prognostic research and better DDS implementation. The chronic non-communicable disease known as diabetes mellitus is characterized by partial or complete insulin insufficiency, which leads to abnormalities in the metabolism of proteins, carbohydrates, lipids, and glucose. Data categorization in the DDS assists doctors and carers in examining unknown datasets by training a large number of real-world databases. Diabetes is a serious worldwide health issue. It is therefore essential for the significance of hyperglycemia and for the prompt detection, examination, and treatment of it. This work uses an SVM classifier based on a hybrid optimizer to develop an efficient model to recognize DM. Because of the disease's high prevalence and common comorbidity, it is regarded as one of the most urgent public health issues. Artificial intelligence (AI) approaches have demonstrated their promise to help diagnose serious illnesses during such crises. Early identification is crucial since conditions including DM, high blood pressure, and heart disease are linked to increased mortality and hospitalization risks in coronavirus patients. Because diabetes diagnosis and treatment are data-intensive procedures by nature, machine learning (ML) techniques are a great option for improving results and implementing cutting-edge strategies. The thorough empirical research demonstrates that, in comparison to offline classifiers, HOMED dramatically increases DDS's performance and lowers its processing overhead. This model can

assist medical professionals in making the best choices regarding the condition and course of treatment. Digital therapeutics has emerged as a prominent field of intervention for the therapy of illness. Both medical professionals and patients are seeing benefits from CDSS. The CDSS is a system that enhances care quality by calculating the risk of sickness progression and medical therapy. It helps physicians and patients come to the best clinical decisions. It has greatly enhanced clinical therapy by reducing drug errors and misdiagnoses.

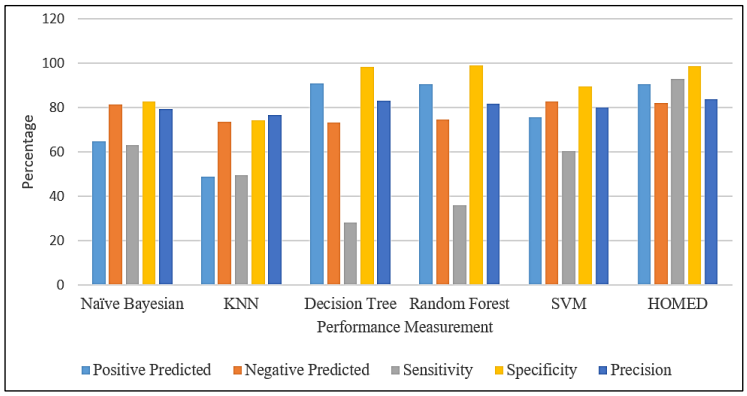

Figure 5: Comparing Different Classifiers According to Performance Metrics

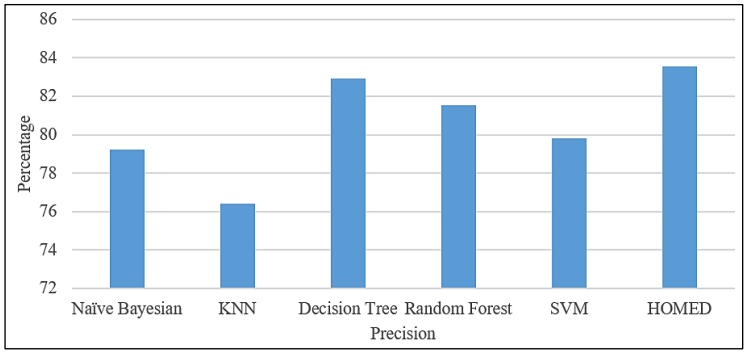

Figure 6: HOMED Performance in Terms of Accuracy

Additionally, because medical databases require ongoing evaluation of the data, it is crucial to update the trained models gradually. This makes the categorization process less complicated to process and more efficient in terms of storage needs. Therefore, this study adopts the HOMED model to improve the classification performance and reduce the computing complexity of the classification process. Using several assessment indicators, HOMED's performance is assessed on PIDD. The experimental findings demonstrate that, in comparison to previous offline methods, HOMED considerably lowers processing overhead and increases forecast accuracy. This classifier's efficacy is assessed using a number of performance metrics. By contrasting the suggested approach's performance with the other five accepted categorization techniques, its efficacy is assessed. Diabetes has an impact on the distribution of healthcare services, but it also causes serious problems for diabetics because of autoimmune illnesses that reduce their immunity to pathogens. This reverses the challenges and obstacles associated with treating the illness as well as the possibility of grave repercussions that increase morbidity and mortality. Managing diabetes is also an expensive, complex, and difficult task for medical personnel. To help doctors make the best treatment decisions and extend patients' lives, a great deal of vital data on patients and ailments needs to be maintained on file. A high-dimensional database's abundance of redundant and unnecessary characteristics makes the algorithm more likely to overfit, which undermines the effectiveness of conventional machine learning-based diagnostic methods. Diabetes diagnosis systems (DDS) have used machine learning to manage a variety of data-related activities. Beyond the healthcare industry, machine learning has shown promise in domains such as web searches,

image processing, speech and image identification, spam filtering, fraud detection, and credit ratings. These features demonstrate the adaptability and revolutionary potential of machine learning in a variety of applications.

Two datasets are used by the suggested classifier: the PIDD from the UCI repository and the DTD from the Data World repository. The National Institute of Diabetes and Digestive and Kidney Diseases provides patient records for PIDD. This dataset comprises 768 clinical records of Pima Indian heritage, people who are surviving. This database uses eight characteristics: blood sugar level, two-hour serum insulin, diastolic blood pressure, skinfold thickness in the triceps, number of pregnancies, diabetes function, nutrition, body mass index, and age. The statistical analysis of every data sample in this database is shown in Table 5. The range of binary parameters is confined to "1 or 0". The output parameter 0" denotes a negative result (i.e., non-diabetic) and 1" denotes a positive result for DM (i.e., diabetic).

The Data World repository is where DTD obtained the clinical data samples (40). This dataset comprises 1099 data samples with 8 features (e.g., glycated hemoglobin (HbA1c), age, class, and type), as well as blood glucose test results that can be taken at any time and plasma sugar test results that are typically taken in the morning or eight hours after a meal. The statistical analysis of every sample in this database is shown in Table 6.

Table 5: Statistical Analysis of PIDD (41)

Index	Attribute	Mean	Min/Max
F_1	Plasmasugarlevel	120.9	0/199
F_2	Twohour seruminsulin(muU/ml)	79.8	0/846
F_3	Diastolicbloodpressure(mmHg)	69.1	0/122
F_4	Tricepsskinfoldthickness (mm)	20.5	0/99
F_5	Numberoftimespregnant	3.8	0/17
F_6	Functionof diabetes nutrition	0.5	0.078/2.42
F_7	Age(years)	33.2	21/81
F_8	Bodymassindex(kg/m^2)	32	0/67.1

Table 6: Statistical Analysisof DTD (42)

Label	Feature	Mean	Min/Max
F_1	Blood sugar test taken at any time	10.73	7.9/13.1
F_2	Blood sugar test usually taken in the morning or 8 hours after a meal	6.14	3.9/9.1
F_3	Plasma glucose while fasting	12.57	0/54
F_4	Plasma glucose 90 minutes after a meal	6.65	4.2/8.8
F_5	HbA1c	43.48	28/66
F_6	Type	Normal, Type 1,Type2	
F_7	Class	0,1	
F_8	Age	33.39	21/81

Choosing the appropriate kernel width (γ) and penalty factor (C) is crucial when using SVM models on real-world datasets. The classifier performs poorly when the value of C is high. In the

learning phase, accuracy is higher when the value of C is very high, but it decreases during the testing phase. The algorithm is unfeasible if C is less because the classification accuracy becomes unacceptable. Because of its effect on how the classification process is implemented, the value of γ has a greater influence on classification performance than C. Prediction of diabetes readmission following the aforementioned pre-processing, dimensional medical data from 127,386 diabetes datasets were acquired. The diabetes database offers a lot more information and numerous features than the heart failure database. Table 7 displays the outcomes of applying the same algorithm to the diabetic data set for experimental comparison.

Table 7: Diabetes Prediction Experiment Results

Method	Training Set Accuracy/%	Test Set Accuracy/%	Training Time/s
Logistic Regression	75.67	73.15	1.46
Random Tree	89.27	88.74	2.45
Deep Neural Networks (DNNs)	93.58	91.47	93.68
Convolutional Neural Networks (CNNs)	92.98	93.82	102.67
Wide Neural Network (BLS)	93.56	92.94	50.90
DAE-BLS	94.69	93.65	52.26

Figure 7 presents the findings of the comparison. It has been observed from Figure 7 that the accuracy of predication method for DAE-BLS is 93.65% and the accuracy of BLS, CNN, DNN, and Random Tree and Logistic regression is 92.94%, 93.82%, 91.47%, 88.74% and 73.15% respectively.

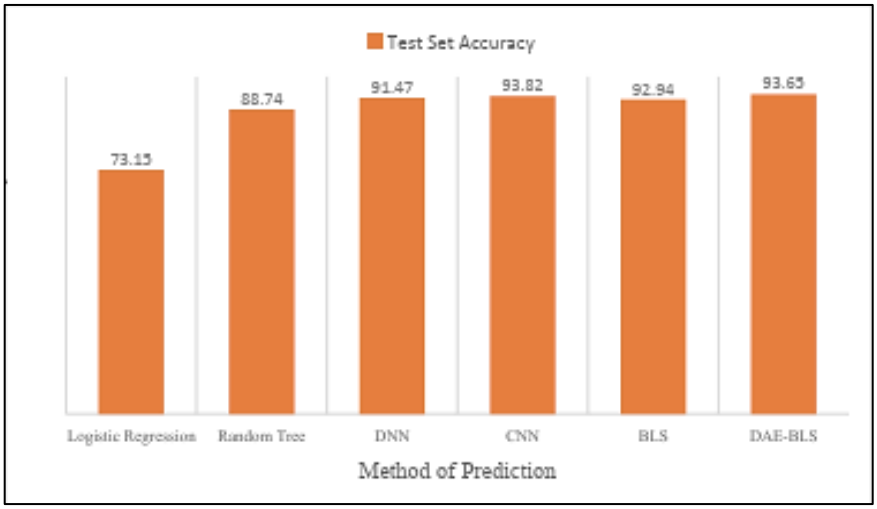

Figure 7: Comparison of Diabetes Experiment Results

In order to find suitable treatments to represent us the condition, diabetes has emerged as a critical problem in the medical field. It is critical to act to diagnose diseases early on and to prevent them in addition to treating them. Prognostic modelling, or the use of data and knowledge to compute future outcomes from past data, is achieved by the techniques used in data mining, machine learning, and other artificial intelligence domains. In order to reduce the risk of morbidity, these strategies are used for the timely prediction and identification of DM. Through training a sizable number of real-world databases, data classification in the DDS helps physicians and carers examine unknown datasets. Diabetes is a major global health concern. As a result, it is crucial for the relevance of hyperglycemia and its early diagnosis, investigation, and management. This work creates an effective model to identify DM using an SVM classifier based on a hybrid optimizer. To make use of SVM's full capacity in the DDS, the established optimization integrates the CSA and BGWO

optimization methods. The suggested cohesive optimizer efficiently delivers good candidate solutions and attains global optimum outcomes by combining the advantages of both the CSA and BGWO algorithms. In this work, the CSA designates an enhanced initial population, and BGWO is then utilized to assess the exploratory agent's locations in the unique exploratory space in order to obtain the best outcomes with the best classification performance. Using real-time datasets, the effectiveness of this well-established cohesive optimizer-based SVM classifier is evaluated in detail. The enactment of the suggested CS-BGWO-SVM classifier is compared with a number of cutting-edge SVM-based diabetes mellitus classifiers in order to determine how effective it is. The experimental investigation shows that the suggested CS-BGWO-SVM is a fortunate classification algorithm with remarkable performance metrics.

Discussion

The conclusion section of this study highlights the findings and their implications concerning the application of machine learning techniques for diabetes classification and prediction. The proposed HOMED model demonstrates superior performance metrics, surpassing existing classifiers, and offers significant contributions to clinical decision-making and patient management. By addressing key challenges such as overfitting, missing data imputation, and real-time classification, the model aligns with the goals of enhancing healthcare delivery.

The results of the study reveal that HOMED outperformed other state-of-the-art classification models in terms of accuracy, sensitivity, specificity, precision, and efficiency. The model's hybrid architecture, combining Adaptive Principal Component Analysis (APCA) and Incremental Support Vector Machines (ISVM), has proven effective for handling high-dimensional datasets and incremental data processing. These findings are consistent with prior research that emphasised the importance of hybrid machine learning techniques in improving predictive accuracy and reducing computational overhead.

In healthcare environments, practical applications bridge the gap between theory and practice by converting theoretical knowledge into practical abilities through techniques including nursing rounds, bedside clinics, and demonstrations. Through the collection and centralization of patient-related data, clinical applications also assist with healthcare planning, delivery, management, and research. These programs compile all patient-related data collected during different patient meetings into a single, all-inclusive data file.

Healthcare providers should concentrate on data-driven decision-making, individualized treatment plans, and using technology for remote monitoring and patient education in order to successfully integrate a new paradigm into the current diabetes medical workflows. It should also make sure that patients are engaged and follow their treatment plans. Managing diabetes mellitus presents a number of challenges and restrictions that impact the application of customized medicine in clinical settings. To improve diabetes care and maximize the potential of customized medicine in successfully managing this complex chronic condition, several issues must be overcome.

Diabetes is a prevalent condition that affects people all over the world. Among other chronic conditions, diabetes raises the risk of renal failure and heart disease. If this disease is identified early, people may live longer and be healthier. The primary diagnosis of diabetes can be aided by a number of supervised machine learning models that have been trained on suitable datasets. It can be difficult for medical experts to diagnose diabetes mellitus accurately and early. People can use machine learning and artificial intelligence approaches as a guide to learn more about this illness and adjust their workload accordingly.

Through the analysis and use of diabetic data, machine learning approaches can identify suggested pathogenic variables, leading to a diagnosis of diabetes that is very accurate and stable. Thus, novel approaches to diabetes screening and diagnosis are made possible by machine learning techniques that can identify physiological indicators and realistic threshold risk variables. Diabetes is a highly dangerous condition that can cause life-threatening consequences, including death, if it is not appropriately and promptly treated. Diabetes is therefore a top priority for medical science research, which produces a vast amount of data.

Key Findings

Performance Metrics: HOMED achieved the highest accuracy (80.34%), sensitivity (92.67%), and specificity (98.43%) among the evaluated models. These metrics are critical for reliable diagnosis and early intervention, aligning with clinical needs for precision and dependability.

Improved Computational Efficiency: By incorporating APCA and ISVM, HOMED addressed challenges related to overfitting and computational inefficiency, often associated with high-dimensional medical datasets.

Practical Applicability: HOMED's ability to process real-time incremental datasets positions it as a practical solution for clinical settings, where data is continuously updated.

Implications

The findings of this study have significant practical implications for the development of clinical decision support systems (CDSS), healthcare policies, and machine learning applications in medical diagnostics:

Enhanced Clinical Decision-Making: The HOMED model provides healthcare professionals with a reliable tool for early diagnosis and management of diabetes, aiding in personalised treatment plans and improving patient outcomes.

Data-Driven Healthcare: By effectively handling high-dimensional and incomplete datasets, HOMED underscores the potential of machine learning to transform clinical data into actionable insights, paving the way for data-driven healthcare systems.

Scalability and Adaptability: The model's hybrid architecture makes it scalable for various medical conditions beyond diabetes, indicating broader applicability in disease prediction and management.

Integration with Digital Health Platforms: HOMED's real-time capabilities can be integrated with telemedicine platforms and electronic health records (EHRs), facilitating remote monitoring and diagnosis.

Advancements in Healthcare Informatics: This study highlights the need for continuous innovation in machine learning algorithms to address evolving challenges in medical data analysis and decision support systems.

Limitations and Scope for Future Research

While the study offers valuable insights, certain limitations should be acknowledged to guide future research:

Dataset Limitations: The study utilised the Pima Indian Diabetes Dataset (PIDD) and Diabetes Treatment Dataset (DTD), which may not fully represent diverse populations. Future studies should include larger and more diverse datasets for generalisability.

Cross-Sectional Analysis: The cross-sectional nature of the analysis limits the ability to track changes in model performance over time. Longitudinal studies are needed to evaluate the model's efficacy in real-world scenarios.

Limited Variables: The current study focused on specific features of the datasets. Exploring additional clinical, behavioural, and genetic variables could enhance the model's predictive power.

Lack of Qualitative Insights: While the study provides robust quantitative results, incorporating qualitative methods such as clinician feedback and patient interviews could offer deeper insights into the model's practical utility.

Integration with Real-World Systems: Future research should focus on implementing HOMED in real-world healthcare systems to evaluate its operational feasibility and effectiveness.

Future Directions

To address these limitations and build upon the current findings, the following future research directions are proposed:

Multimodal Data Integration: Integrate diverse data types, including imaging, genomic, and wearable sensor data, to improve prediction accuracy and expand the model's applicability.

Real-Time Deployment: Develop cloud-based or edge-computing solutions for real-time implementation of the HOMED model in clinical environments.

Explainability and Interpretability: Focus on enhancing the interpretability of the model's predictions to facilitate clinician trust and adoption.

Customisation for Local Populations: Adapt the model to specific population characteristics,

considering regional variations in diabetes prevalence and risk factors.

Ethical and Privacy Considerations: Address ethical concerns related to patient data privacy and ensure compliance with healthcare regulations when deploying machine learning models in clinical settings.

Conclusion

The study demonstrates that the HOMED model, powered by APCA and ISVM, is a significant advancement in diabetes classification and prediction. Its superior performance metrics, scalability, and adaptability position it as a promising tool for enhancing healthcare delivery. However, addressing the outlined limitations and pursuing the proposed research directions will be essential for maximising the model's potential and ensuring its successful integration into real-world medical practice.

By leveraging innovative machine learning techniques, HOMED sets a benchmark for future research and development in medical diagnostics, contributing to the broader goal of improving global healthcare outcomes.

Abbreviations

APCA: Adaptive Principal Component Analysis, BGWO: Binary Grey Wolf Optimisation, CDSS: Clinical Decision Support Systems, CSA: Crow Search Algorithm, DM: Diabetes Mellitus, DTD: Diabetes Treatment Dataset, HOMED: Hybrid Online Model for Early Detection, ISVM: Incremental Support Vector Machine, ML: Machine Learning, NPV: Negative Predictive Value, PCA: Principal Component Analysis, PIDD: Pima Indian Diabetes Dataset, PPV: Positive Predictive Value, RF: Random Forest, SVM: Support Vector Machine.

Acknowledgement

The authors would like to extend their gratitude to the faculty and staff of Shobhit Institute of Engineering and Technology, Meerut, for their unwavering support and guidance throughout this research. The authors are also grateful to Ajay Kumar Garg Engineering College, Ghaziabad, for providing the necessary resources and a conducive environment for carrying out this study.

Author Contributions

Prag Jain: Conceptualisation, methodology design, data collection, analysis, manuscript drafting, Nidhi Tyagi: Supervision, validation, critical review, editing of the manuscript, Birendra Kumar Sharma: Technical guidance, review of computational models, data interpretation.

Conflict of Interest

The authors declare no conflict of interest.

Ethics Approval

This study does not involve human or animal participants. All data used were anonymised and sourced from publicly available datasets. Ethics approval was not required for this research.

Funding

This research received no external funding.

References

1. Raghupathi V, Raghupathi W. Healthcare expenditure and economic performance: insights from the United States data. Frontiers in public health. 2020 May 13;8:156.
2. Zhou L, Xu X, Antwi HA, Wang L. Towards an equitable healthcare in China: evaluating the productive efficiency of community health centers in Jiangsu Province. International journal for equity in health. 2017 Dec;16:1-0.
3. Altman IL, Holkham C, Gosrani R. A mixed method service evaluation of Electronic Prescribing and Medicines Administration (EPMA) within 10 community hospitals in England. International Journal of Pharmacy Practice. 2024 Nov;32(Supplement_2):ii32-3.
4. Barlow P, van Schalkwyk MC, McKee M, Labonté R, Stuckler D. COVID-19 and the collapse of global trade: building an effective public health response. The Lancet Planetary Health. 2021 Feb 1;5(2):e102-7.
5. Patel S, Patel, A Patel V, Solanki N. Study of medication error in hospitalized patients in tertiary care hospital. Indian Journal of Pharmacy Practice. 2018; 11(1): 32-36.
6. Sheikh A, Anderson M, Albala S, Casadei B, Franklin BD, Richards M, Taylor D, Tibble H, Mossialos E. Health information technology and digital innovation for national learning health and care systems. The Lancet Digital Health. 2021 Jun 1;3(6):e383-96.
7. Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, Colagiuri S, Guariguata L, Motala AA, Ogurtsova K, Shaw JE. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas.

Diabetes research and clinical practice. 2019 Nov 1;157:107843.
8. Martín-Carro B, Donate-Correa J, Fernández-Villabrille S, Martín-Vírgala J, Panizo S, Carrillo-López N, Martínez-Arias L, Navarro-González JF, Naves-Díaz M, Fernández-Martín JL, Alonso-Montes C. Experimental models to study diabetes mellitus and its complications: limitations and new opportunities. International journal of molecular sciences. 2023 Jun 18;24(12):10309.
9. Singla R, Singla A, Gupta Y, and Kalra S. Artificial intelligence/machine learning in diabetes care. Indian Journal of Endocrinology and Metabolism. 2019; 23(4): 495–497.
10. Ong KL, Stafford LK, McLaughlin SA, Boyko EJ, Vollset SE, Smith AE, Dalton BE, Duprey J, Cruz JA, Hagins H, Lindstedt PA. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. The Lancet. 2023 Jul 15;402(10397):203-34.
11. Koulali R, Zaidani H, Zaim M. Image classification approach using machine learning and an industrial Hadoop based data pipeline. Big Data Research. 2021 May 15;24:100184.
12. Singla R, Garg A, Singla S, Gupta Y. Temporal change in profile of association between diabetes, obesity, and age of onset in urban India: A brief report and review of literature. Indian journal of endocrinology and metabolism. 2018 May 1;22(3):429-32.
13. Albahli S. Type 2 machine learning: an effective hybrid prediction model for early type 2 diabetes detection. Journal of Medical Imaging and Health Informatics. 2020 May 1;10(5):1069-75.
14. Alam T. M, Iqbal M. A, Ali Y, Wahab A, Ijaz S, Baig T I, Hussain A, Malik M A, Raza M R, Ibrar S, Abbas Z. A model for early prediction of diabetes. Informatics in Medicine Unlocked. 2019; 16: 100204.
15. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics. 2018 Nov 6;9:515.
16. Sah P, Sarma KK. Bloodless technique to detect diabetes using a soft computational tool. In Ophthalmology: Breakthroughs in Research and Practice. 2018; 15(4): 34-52. https://doi.org/10.4018/978-1-5225-5736-5.ch004
17. Sarwar A, Ali M, Manhas J, Sharma V. Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. International Journal of Information Technology. 2020 Jun;12:419-28.
18. Qomariah DU, Tjandrasa H, Fatichah C. Classification of diabetic retinopathy and normal retinal images using CNN and SVM. In2019 12th International Conference on Information & Communication Technology and System (ICTS). IEEE. 2019 Jul 18;12(1): 152-157.
19. Gupta D, Choudhury A, Gupta U, Singh P, Prasad M. Computational approach to clinical diagnosis of diabetes disease: a comparative study. Multimedia Tools and Applications. 2021 Jan 1;80(6):1-26.
20. Sardu C, Gargiulo G, Esposito G, Paolisso G, Marfella R. Impact of diabetes mellitus on clinical outcomes in patients affected by Covid-19. Cardiovascular diabetology. 2020 Dec;19:1-4.
21. Rabie O, Alghazzawi D, Asghar J, Saddozai FK, Asghar MZ. A decision support system for diagnosing diabetes using deep neural network. Frontiers in public health. 2022 Mar 17;10:861062.
22. Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia computer science. 2018 Jan 1;132:1578-85.
23. Hussain A, Naaz S. Prediction of diabetes mellitus: comparative study of various machine learning models. InInternational Conference on Innovative Computing and Communications: Proceedings of ICICC 2020. Springer Singapore.2021;2:103-115 .
24. Kumari S, Kumar D, Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. International Journal of Cognitive Computing in Engineering. 2021 Jun 1;2:40-6.
25. Srivastava S, Sharma L, Sharma V, Kumar A, Darbari H. Prediction of diabetes using artificial neural network approach. InEngineering Vibration, Communication and Information Processing: ICoEVCI 2018, India. Springer Singapore. 2019:679-687. https://doi.org/10.1007/978-981-13-1642-5_59
26. Gupta S, Verma HK, Bhardwaj D. Classification of diabetes using Naive Bayes and support vector machine as a technique. InOperations Management and Systems Engineering: Select Proceedings of CPIE 2019. Springer Singapore. 2021:365-376. https://link.springer.com/chapter/10.1007/978-981-15-6017-0_24
27. Ndjaboue R, Ngueta G, Rochefort-Brihay C, Delorme S, Guay D, Ivers N, Shah BR, Straus SE, Yu C, Comeau S, Farhat I. Prediction models of diabetes complications: a scoping review. J Epidemiol Community Health. 2022 Oct 1;76(10):896-904.
28. Makroum MA, Adda M, Bouzouane A, Ibrahim H. Machine learning and smart devices for diabetes management: Systematic review. Sensors. 2022 Feb 25;22(5):1843.
29. Abnoosian K, Farnoosh R, Behzadi MH. Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. BMC bioinformatics. 2023 Sep 12;24(1):337.
30. Choubey DK, Kumar M, Shukla V, Tripathi S, Dhandhania VK. Comparative analysis of classification methods with PCA and LDA for diabetes. Current diabetes reviews. 2020 Nov 1;16(8):833-50.
31. Perveen S, Shahbaz M, Guergachi A, Keshavjee K. Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science. 2016 Jan 1;82:115-21.
32. Ebrahim OA, Derbew G. Application of supervised machine learning algorithms for classification and prediction of type-2 diabetes disease status in Afar regional state, Northeastern Ethiopia 2021. Scientific reports. 2023 May 13;13(1):7779.
33. Thotad PN, Bharamagoudar GR, Anami BS. Diabetes disease detection and classification on Indian demographic and health survey data using machine

learning methods. Diabetes & Metabolic Syndrome: Clinical Research & Reviews. 2023 Jan 1;17(1):102690.
34. AlZu'bi S, Elbes M, Mughaid A, Bdair N, Abualigah L, Forestiero A, Zitar RA. Diabetes monitoring system in smart health cities based on big data intelligence. Future Internet. 2023 Feb 20;15(2):85.
35. Ahamed BS, Arya MS, Sangeetha SK, Auxilia Osvin NV. Diabetes mellitus disease prediction and type classification involving predictive modeling using machine learning techniques and classifiers. Applied Computational Intelligence and Soft Computing. 2022;2022(1):7899364.
36. Chang V, Bailey J, Xu QA, Sun Z. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications. 2023 Aug;35(22):16157-73.
37. Gupta N, Kaushik B, Imam Rahmani MK, Lashari SA. Performance evaluation of deep dense layer neural network for diabetes prediction. Computers, Materials & Continua. 2023; 76(1): 1-20.
38. Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. Ict Express. 2021 Dec 1;7(4):432-9.
39. García-Ordás MT, Benavides C, Benítez-Andrades JA, Alaiz-Moretón H, García-Rodríguez I. Diabetes detection using deep learning techniques with oversampling and feature augmentation. Computer Methods and Programs in Biomedicine. 2021 Apr 1;202:105968.
40. Liaw ST, Guo JG, Ansari S, Jonnagaddala J, Godinho MA, Borelli Jr AJ, de Lusignan S, Capurro D, Liyanage H, Bhattal N, Bennett V. Quality assessment of real-world data repositories across the data life cycle: a literature review. Journal of the American Medical Informatics Association. 2021 Jul 1;28(7):1591-9. https://doi.org/10.1093/jamia/ocaa340
41. Khan MM, Arif RB, Siddique MA, Oishe MR. Study and observation of the variation of accuracies of KNN, SVM, LMNN, ENN algorithms on eleven different datasets from UCI machine learning repository. In2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT).IEEE. 2018 Sep 13:124-129. https://doi.org/10.48550/arXiv.1809.06186
42. Mousa A, Mustafa W, Marqas RB, Mohammed SH. A comparative study of diabetes detection using the Pima Indian diabetes database. Journal of Duhok University. 2023 Oct 12;26(2):277-88.