DS Journal of Digital Science and Technology
Volume 4 Issue 3, 1-15, Jul - Sep 2025
ISSN: 2583-5416 / https://doi.org/10.59232/DST-V4I3P101

Original Article

A Machine Learning-Based Solution for Diabetes Diagnosis: Integrating Clinical Data and Vital Signs

Huy Huynh[1], Thanh Cao[1], Hai Tran[2*]

[1]*Saigon University (SGU), HCM, Vietnam*
[2]*Ho Chi Minh University of Education (HCMUE), HCM, Vietnam*

*haits@hcmue.edu.vn

Received: 18 July 2025; Revised: 10 August 2025; Accepted: 18 September 2025; Published: 28 September 2025;

Abstract - The diagnosis of diabetes mellitus utilizing clinical data and vital signs is essential for the early identification and efficient management of the condition, particularly as its global prevalence rises. This study examines general diagnostic methodologies utilizing diverse indicators, with a particular emphasis on diabetes diagnostics, subsequently identifying pertinent features to develop an effective, cohesive diagnostic solution. The suggested solution uses both structured and unstructured features. Structured features include vital signs, demographics, and lab tests, while unstructured features include chief complaints and medical notes. This solution is based on the MIMIC-IV dataset, which is a rich and varied source of medical data. The research suggests creating a prototype diagnostic system that uses modern machine learning models like Logistic Regression, Support Vector Machine (SVM), Gradient Boosting, Random Forest, and XGBoost. This system is built to automatically process and combine information from different types of data, such as medical text and biological indicators, to improve the prediction process and help doctors make decisions. This solution not only offers a thorough and efficient method for diagnosing diabetes, but it also shows how it could be used in real-world healthcare systems. The prototype system can be used to help doctors make quick and correct clinical decisions. It can also be used as a starting point for future research and applications in smart healthcare.

Keywords - Diabetes Diagnosis, Machine Learning, Clinical Information, Vital Signs, MIMIC-IV Dataset.

1. Introduction

Artificial Intelligence (AI) has made significant strides in medical diagnosis over the past few years, helping to identify diseases such as diabetes, kidney disease, and cardiovascular disease. Machine Learning (ML) models have shown great promise in analyzing medical data, ranging from structured data, such as vital signs and demographics, to unstructured data, including medical notes. This improves the accuracy of diagnoses and speeds up the decision-making process [1]. Diabetes mellitus is one of the greatest global health challenges of the 21st century. It is a long-term illness that causes high blood sugar levels all the time because of a lack of insulin (type 1) and resistance to insulin combined with relative deficiency (type 2). If not well controlled, the disease leads to a series of dangerous complications such as heart failure, kidney failure, blindness, and peripheral nerve damage [2]. Traditional diagnostic procedures are often based on the comprehensive assessment of risk factors such as Body Mass Index (BMI), blood pressure, fasting blood glucose levels, HbA1c, and family history.

© 2025 Dream Science. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Many studies have achieved significant results in applying AI to solve this problem. Traditional machine learning algorithms, such as Support Vector Machine (SVM), Random Forest (RF), and Decision Tree, have proven highly effective in classifying disease risk based on standard datasets [3]. Recently, the development of Deep Learning with architectures such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) has enabled the exploitation of complex, non-linear relationships between health variables, achieving superior accuracy compared to classical statistical methods [4]. Despite encouraging results, diabetes classification using machine learning models still has some limitations [4-6]: Most current models are trained on small, old, or poorly diverse datasets in terms of race and underlying pathology (e.g., the dataset of PIMA Indian Diabetes or small, scattered datasets on Kaggle).

This leads to poor generalization when deploying the model to real-world clinical settings that have many different patient demographics. Deep learning models (DNN, Transformer), despite achieving high accuracy, often operate as a "black-box", making it hard for doctors to understand the reasoning behind the predictions. In contrast, interpretable models such as Logistic Regression often have lower performance and are prone to overfitting if not rigorously externally validated. A large amount of important information about patient conditions resides in clinical notes in the form of free text. However, most current studies only focus on tabular data without effectively exploiting this treasure trove of text data, reducing accuracy in complex cases with multiple comorbidities [7].

To overcome the above limitations, this study proposes a comprehensive diabetes diagnosis solution based on advanced classification models, using the MIMIC-IV dataset - a large-scale, diverse, and realistic clinical database.

The novelty of the study is the integration of natural language processing to extract information from medical notes (unstructured data), combined with vital signs and demographics (structured data). At the same time, an application of the results was developed to demonstrate the decision support capabilities, with the aim of improving the quality of healthcare and demonstrating the feasibility of implementation in a hospital environment.

2. Related Work

This study will conduct a study of some articles related to clinical diagnosis using machine learning models and delve into diabetes. Then, based on these syntheses, the study will select features and models suitable for the diabetes diagnosis problem. Finally, based on this model, the study will propose a solution for a diabetes diagnosis application for doctors or nurses in hospitals. The research framework is modeled as follows:

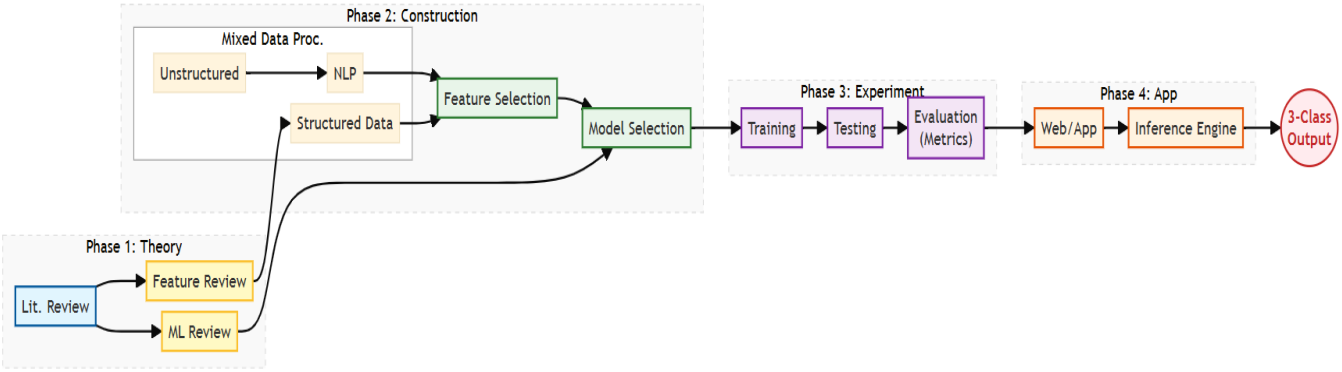

Fig. 1 Framework of research

2.1. Models for Diagnosing Diseases that Use Clinical Data and Vital Signs

The use of Artificial Intelligence (AI) and Machine Learning (ML) in medical diagnosis has led to major improvements, especially in finding diseases early and keeping an eye on patients' health. Numerous studies have concentrated on utilizing clinical information, vital signs, and demographic data to construct predictive models.

Some important studies that have come out recently are: Garcés-Jiménez et al. suggested a way to predict infections (ARI, UTI, SSTI) in nursing home residents early on. They used XGBoost with vital signs (TEMP, BPM, SpO2) and demographic data. They got an AUC-ROC of 0.857, but the study had some problems, such as using data that was not public and being hard to do in real time [1]. In the same year, López-Izquierdo et al. created clinical phenotypes (Alpha, Beta, Gamma) to sort people by risk and predict short-term death from EMS data in Spain. They performed effectively with a Gaussian mixture model and PCA; however, applying the same results to different contexts is challenging [8].

Soares et al. employed Continuous-Wave (CW) radar to assess vital signs and personal attributes; their Random Forest model achieved 83% accuracy using data from 92 subjects, although the restricted scale and controlled experimental conditions presented considerable limitations [9]. Fascia et al. investigated machine learning algorithms, including Random Forest, Logistic Regression, CNN, and LSTM, for medical prognosis, proposing future directions such as Explainable AI and transfer learning, although the research was primarily theoretical [10]. In Orangi-Fard's study, COPD exacerbations were predicted from MIMIC-III data using NLP and SVM. Because it only used data from one source, its AUC of 0.82 was constrained [7].

Verhoeven et al. used SVM and Logistic Regression to predict Necrotizing Enterocolitis (NEC) in premature infants, achieving an F1-score of 0.82; however, the small sample size (267 infants) and lack of nutritional data were major limitations [11]. The DBICP deep learning model (DNN), which Kim et al. developed to predict critical events in ICU patients, achieved 95% accuracy on a sizable dataset (110,000 samples); however, it has not yet been validated in a real-world setting and requires additional optimization [12]. Rassam et al. created the HATCN-AD model (TCN with an attention mechanism) to find problems with the vital signs of older people. It was 99.15% accurate, but it could not be used on other data sets or in other situations [13].

Park et al. used XGBoost to combine vital signs, blood tests, and ICD-10 codes to predict In-Hospital Cardiac Arrest (IHCA) in 2025. They got an AUROC of 0.934, but only one South Korean hospital gave them the data [14]. Finally, Awad et al. made a remote monitoring system that uses an ANN and a LabVIEW interface to divide cardiovascular health into three groups with 98.8% accuracy. However, the dataset was small (750 patients), and the system has not been tested in the real world [15]. The results of these studies are presented in **Table 1** to facilitate evaluation and tracking:

Table 1. Summary of recent studies on disease prediction using clinical data and vital signs

Author(s) and Year	Article	Method	Dataset	Advantages	Disadvantages	Results
Garcés-Jiménez et al., 2024 [1]	Predictive health monitoring: Leveraging artificial intelligence for early detection of infectious diseases in nursing home residents through discontinuous vital signs analysis	XGBoost	5,004 samples from 60 people living in three nursing homes in Spain and the Dominican Republic. Vital signs: TEMP, BPM, SpO2, EDA, SYS/DIA.	An effective way to clean up data. Four days ahead of time prediction	- Discontinuous data. - Data imbalance. - Not applicable in real-time.	AUC-ROC = 0.857, which is better than kNN and Logistic Regression.
López-Izquierdo et al., 2024 [8]	Clinical phenotypes and short-term outcomes based on prehospital point-of-care testing and on-scene vital signs	Gaussian Mixture with PCA	7,909 patients from EMS Spain (2020–2023). Data: Vital signs, point-of-care blood tests, and the death	Stratification of risk that works. Examination of the correlation between phenotype and clinical	- The data is only for Spain. - POCT is not very common yet.	The alpha phenotype has the highest chance of death (33% after 30 days).

			rate after 2, 7, and 30 days	outcomes.		
Soares et al., 2024 [9]	Impact and Classification of Body Stature and Physiological Variability in the Acquisition of Vital Signs Using Continuous Wave Radar	Random Forest SVM KNN	92 volunteers (46 men and 46 women) between the ages of 18 and 50 gave us continuous wave (CW) radar data	No contact. Random Forest had the best performance	- Not enough people in the sample. - The best place to do experiments, but it is hard to use in real life.	RF attained 83% accuracy (cross-validation) in gender and CWP classification.
Fascia et al., 2024 [10]	Machine learning applications in medical prognostics: a comprehensive review	Random Forest Logistic Regression CNN LSTM	Synthesis from 87 studies (PubMed, IEEE Xplore, Google Scholar).	Full picture. Future direction: AI that can be explained and that learns all the time.	Only a theory, not yet tested in real life.	RF and CNN are both highly rated for working with unstructured and multidimensional data.
Orangi-Fard et al., 2024 [7]	Prediction of COPD Using Machine Learning and Clinical Summary Notes	SVM AdaBoost QDA	Information from MIMIC-III (31,667 records from clinical notes and 10,489 records from vital signs).	Combination of NLP and ML. SVM does better than other algorithms.	- Only uses information from MIMIC-III. - Does not combine lab results with medical imaging.	The AUC for SVM was 0.82 for clinical notes and 0.79 for vital signs.
Verhoeven et al., 2024 [11]	Using Vital Signs for the Early Prediction of Necrotizing Enterocolitis in Preterm Neonates with Machine Learning	Logistic Regression SVM XGBoost	267 premature infants in the NICU (2018-2022). Heart rate, respiratory rate, cerebral oxygenation, and splanchnic oxygenation are all examples of data.	Early prediction of NEC before it starts. SVM and LR did the best job.	- The sample size is small. - Data that is missing, especially splanchnic oxygenation.	F1-score = 0.82 (SVM, LR), AUC-PR = 0.83.
Kim et al., 2024 [12]	Deep Learning Model for Predicting Critical Patient Condition	DNN-based Intensive Care Prediction (DBICP)	110,000 records from the ICU in South Korea. Information: vital signs, underlying conditions, prescribed drugs, and ICD-10 codes.	High accuracy. The best DNN structure keeps it from overfitting.	- Not yet tried out in real life. - Very dependent on data from the ICU.	Accuracy 95.4%, AUC = 0.98.
Rassam et al., 2024 [13]	Monitoring Critical Health Conditions in the Elderly: A Deep Learning-Based Abnormal Vital Sign Detection Model	HATCN-AD (Hierarchical Attention-based Temporal CNN)	Two sets of data from MIMIC-II: Subject 330 and Subject 441. Heart rate, blood pressure, temperature, respiratory rate, and SpO2.	Includes a way to focus on important features. Better performance than CNN and LSTM.	- Only tried out on two small sets of data. - Needs to grow and try things out in real life.	The accuracy is 99.15% for Subject 330 and 98.96% for Subject 441.
Park et al., 2025 [14]	A Machine learning approach for predicting in-hospital cardiac arrest using single-	XGBoost	62,061 patients from CDW in South Korea. Data: Blood tests, vital signs,	Putting together data from different sources.	- Data from one center. - Not tested in real life.	AUROC = 0.934 (FS3: Vital signs + lab tests + ICD-10).

	day vital signs, laboratory test results, and International Classification of Diseases-10 block for diagnosis		and ICD-10 codes.			
Awad et al., 2025 [15]	An encoding-based machine learning approach for health status classification and remote monitoring of cardiac patients	ANN	There are 750 patients from two to three hospitals in Mosul, Iraq. Information: SPO2, HR, RR, blood pressure, GC, chest pain, and trouble breathing.	Bringing together TinyML. A cheap way to keep an eye on things from afar.	- Small number of samples. - Not yet put to the test in real life.	The accuracy was 98.8% (RS + VS encoding), and the ANN got 98.4%.

These studies have created a strong foundation for using AI and ML in medicine, especially for predicting diseases and helping doctors make decisions. However, there are still problems with the size of the datasets, their ability to be generalized, and their usefulness in real life that need to be solved in future research.

2.2. Models for Diagnosing Diabetes

The study found that AI and ML could be useful for finding and treating diabetes. They say that we need to use machine learning models to find and treat this illness early and well. Gudiño-Ochoa et al. built an e-nose system with TinyML to look at breath in 2024. It was 94% accurate at finding diabetes and had an R^2 value of 0.86 for predicting blood sugar levels. The study solely depends on breath analysis and has not yet integrated supplementary vital signs [16].

Hu et al. used the MIMIC-III dataset that same year to make models that could guess how many people with type 2 diabetes would die or have to go back to the hospital. Adaboost's AUROC was 0.7952, and MLP's AUROC was 0.8487. The data only came from big-city hospitals, so it does not work as well in rural areas [17]. Khalifa et al. arranged the function of AI in diabetes management, emphasizing individualized treatment and risk assessment. But their research was largely theoretical and did not include any real-world testing [3]. Muller et al. developed a monitoring system utilizing a MANET network and RNN-GRU, attaining 95.7% accuracy in predicting hyperglycemia and hypoglycemia; however, it is significantly reliant on network infrastructure and necessitates enhancements in security [18].

Alkalifah et al. assessed eight regression algorithms for forecasting blood glucose levels, finding that GPR yielded the lowest mean squared error (1.64 mg/dL) and BSTE exhibited the highest accuracy (92.58%); however, the study was constrained by data imbalance [19]. Rustam et al. created a CNN-LSTM method to classify patients using the Random Forest method. It was 99% accurate on three sets of data, but it needs a lot of computing power, which makes it hard to use in medical settings with limited resources [20]. Finally, Morgan-Benita et al. used biomarkers such as triglycerides, SBP, and DBP to find type 2 diabetes early. Their Random Forest model was 88.2% accurate, but the study needs to get more data and make sure it works for people from different backgrounds [21].

These studies have changed a lot about how diabetes is classified. They have made it possible to create non-invasive monitoring systems and advanced models that use deep learning and machine learning. But there are still a lot of problems to solve, like making sure that technology works, making models that work in different situations, and keeping data safe. These problems need to be fixed so that real-world applications can work. The table below, Table 2, shows a summary of the studies above:

Table 2. Summary of research on diabetes diagnosis

Author(s) and Year	Article	Method	Dataset	Advantages	Disadvantages	Results
Gudiño-Ochoa et al., 2024 [16]	Enhanced diabetes detection and blood glucose prediction using TinyML-integrated E-nose and breath analysis: A novel approach combining synthetic and real-world data	Random Forest XGBoost LightGBM DNN	58 people took part (29 were healthy, 12 had type 1 diabetes, and 17 had type 2 diabetes). Data on VOCs and breath (acetone, alcohol, and CO).	- An e-nose system that does not hurt. - CTGAN for adding more synthetic data. - TinyML integration for analysis in real time.	- Needs to breathe, but does not have any other vital signs (like heart rate or blood pressure). - The sample size is small; there are only 58 participants	94% accuracy (Random Forest, XGBoost). R^2 = 0.86 (LightGBM for blood glucose prediction).
Hu et al., 2024 [17]	Machine learning-based predictions of mortality and readmission in type 2 diabetes patients in the ICU	Adaboost Bagging GaussianNB Logistic Regression MLP SVC	14,222 people with type 2 diabetes from MIMIC-III. Data: Vital signs, lab tests, length of stay.	- Adaboost and Bagging work well. - Use of SMOTE to balance the data.	- Most of the data came from urban hospitals, so it is hard to make generalizations. - Not yet tried on other datasets.	Adaboost got 92.49% accuracy, while AUROC = 0.8487 (MLP for predicting readmission).
Khalifa et al., 2024 [3]	Artificial intelligence for diabetes: Enhancing prevention, diagnosis, and effective management	Review of ML and AI	43 studies from PubMed, Embase, and Google Scholar (2019–2023).	- A thorough look at how AI is used to manage diabetes. - Advises on Explainable AI and continuous learning.	- An overview of the theory, but no real-world testing.	AI helps with personalized treatment, predicting complications, and managing a healthy lifestyle.
Muller et al., 2024 [18]	Improving Diabetes Diagnosis in Instantaneous Situations with MANET and Data Mining	RNN-GRU k-Medoids Clustering	1,000 data samples from glucose meters and wearable devices.	- MANET lets you watch things in real time. - RNN-GRU gets very accurate results. - Treatment is based on the patient's group.	- Based on the MANET infrastructure. - Needs to be tested on a larger scale.	95.7% accuracy (RNN-GRU), which is better than Decision Tree and Random Forest.
Alkalifah et al., 2024 [19]	Evaluation of machine learning-based regression techniques for the prediction of diabetes level fluctuations	GPR BDT BSTE ANN SVM LR LRSGD SW	14,733 records from a continuous glucose monitoring (CGM) system.	GPR, BDT, and BSTE achieve high performance in continuous prediction and classification.	- Data that is not balanced (not many records of hyperglycemia). - The dataset size is small and not fully representative.	GPR got an MSE of 1.64 mg/dL, and BDT got an accuracy of 92.58%.
Rustam et al., 2024 [20]	Enhanced detection of diabetes mellitus using a novel ensemble feature engineering approach and machine learning model	CNN-LSTM combined with Random Forest	There are 3 public datasets: Aravindpcoder, Mathchi, and Ishandutta. Each has 27 attributes and 3 main classification classes.	- CNN-LSTM works better (99% of the time). - High generalization because it combines many datasets.	- Not enough data; it needs to be increased. - CNN-LSTM needs a lot of computing power.	99% accuracy (CNN-LSTM + RF), better than earlier studies (ANN, DNN).
Morgan-Benita et al.	Setting Ranges in Potential	Logistic Regression	1,726 people in Mexico have T2DM	- Analyzing by gender helps	- Small number of samples.	RF achieved 88.2% accuracy.

[21]	Biomarkers for Type 2 Diabetes Mellitus Patients Early Detection By Sex—An Approach with Machine Learning Algorithms	ANN SVM RF	(855 men and 871 women). Data: 21 factors, including triglycerides, cholesterol, and blood pressure.	make treatment more personal. - It sets the value ranges for important biomarkers.	- Needs to be tested on different groups of people.	The most important features are triglycerides and DBP. .

2.3. Commonly Used Models for Diagnosing Diseases

Researchers have used Machine Learning (ML) and Deep Learning (DL) models a lot to classify and predict diabetes. These models have been used for many tasks, including classifying patients, predicting blood glucose levels, and finding complications. Some of the most common algorithms for Machine Learning (ML) classification include Logistic Regression, Support Vector Machine (SVM), Random Forest, LightGBM, XGBoost, and Gradient Boosting. Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs) are two kinds of neural networks that can also handle data that is hard to understand and not organized:

- Logistic Regression (LR): LR is a linear algorithm that can be used to classify things into two groups. It uses the logistic (sigmoid) function to figure out how likely it is that something will happen. LR is easy to understand and works well on linearly separable data, but it does not work as well when relationships are not linear [22].
- Support Vector Machine (SVM): SVM is a strong tool for sorting data that finds the best hyperplane to divide data points into groups. The "kernel trick" moves data that is not linear to a space with more dimensions so that it can be used. SVM is a great tool for dealing with complicated datasets, even when there aren't many samples [22].
- Random Forest (RF): RF is an ensemble learning algorithm that prevents overfitting and improves prediction accuracy by utilizing multiple Decision Trees. The final outcome is based on either majority voting (for classification) or averaging (for regression), with each tree learning from a random subset of the data. RF works well with non-linear and multidimensional data [22].
- LightGBM (Light Gradient Boosting Machine): LightGBM is a gradient boosting framework that works best with big datasets and fast training times. LightGBM is faster and more efficient than other algorithms because it grows trees leaf-wise instead of level-wise. This is especially true for datasets with a lot of features [25].
- Gradient Boosting is a type of ensemble learning in which weak learners, like decision trees, are added one at a time to fix the mistakes of the ones that came before them. Gradient Boosting is suitable for non-linear or unbalanced data. Some of the most common types are XGBoost, LightGBM, and CatBoost [23].
- Extreme Gradient Boosting, or XGBoost, is a better and more scalable version of gradient boosting that makes calculations faster and more accurate. It adds new decision trees one at a time to reduce mistakes, and it uses regularization (L1 and L2) to stop overfitting. People know that it can handle big, uneven datasets [24].
- Deep Neural Networks (DNN): A DNN is a type of neural network that has a lot of hidden layers between the input and output layers. This helps it learn from data that is hard to understand and has many levels. Deep Neural Networks (DNNs) work well with large, high-dimensional datasets, but they need a lot of computing power and can overfit if they aren't regularized properly [26].
- Convolutional Neural Networks (CNN): A type of neural network that is designed to work with grid-like data, such as images. They automatically and adaptively learn spatial hierarchies of features using convolutional layers, and then they use pooling layers to lower the number of dimensions. A lot of medical signal analysis and image recognition work is done with CNNs [27].

These models work together to help doctors make informed decisions, improve classification performance, and make more accurate predictions. They also give people with diabetes personalized ways to manage their condition. Below is a list of some important studies in this area.

3. Proposed Method for Diabetes Diagnosis

3.1. Introduction to the MIMIC-IV Dataset

This study employs the MIMIC-IV and MIMIC-IV-ED datasets, which contain comprehensive intensive care data for more than 40,000 ICU patients at the Beth Israel Deaconess Medical Center (BIDMC) [28]. The MIMIC data has been stripped of any identifying information in accordance with HIPAA rules. It is very important for research in clinical informatics, epidemiology, and machine learning. MIMIC-IV is a better version of MIMIC-III that makes it easier to do medical and healthcare research by putting data into modules in a way that makes sense.

The hospital module gets its data from the hospital's Electronic Health Record (EHR) system, which mostly has information about patients who are currently in the hospital. Some tables also have data from places other than the hospital. The labevents table, for instance, has information about lab tests that were done on an outpatient basis. This module shows everything that happens to a patient while they are in the hospital, such as all of the clinical activities and information. The MIMIC-IV-ED dataset is a big, free set of data that has information about visits to the Emergency Department (ED) at the Beth Israel Deaconess Medical Center from 2011 to 2019. The dataset contains around 425,000 ED visits [29].

To follow the Safe Harbor rules of the Health Information Portability and Accountability Act (HIPAA), all data has been de-identified. This study will suggest a way to classify diabetes by going through the steps of feature extraction, data preprocessing, choosing a classification model, and making a prototype application interface.

3.2. Feature Selection

This research will utilize previous studies to select and integrate various feature types for the development of the classification model. The integration of text-based features constitutes a significant innovation of this study. This method tries to see how much better classification accuracy gets compared to older methods that only use numbers. Table 3 shows some of the most common features:

Table 3. Numerical feature

Biological	Demographics
• Height • Weight • Heart rate • Systolic blood pressure • Diastolic blood pressure • Oxygen saturation • Body Mass Index	• Gender • Marital status • Race • Age • Physical activity

This study uses unstructured, text-based features to get information from natural language, which is different from traditional methods that only use structured, numerical features. This is a new feature that lets the model use more complicated and information-rich data, especially in the medical field.

- Chief Complaint: A short explanation of why the patient came in or was admitted, including their main health issues.
- Medical Text Summary: Clinical notes that give a brief overview of the patient's condition, including clinical data, diagnoses, and treatment plans.

The medical text is often a long record with extra characters. Therefore, the Facebook/bart-large-cnn model [30] was used to preprocess this text, converting it into concise, medically relevant information. This is also a new point compared with previous studies. Table 4 shows an example of a data sample from the MIMIC-IV dataset:

Table 4. An example data sample from the MIMIC-IV dataset

biological_metrics	demographics	chief_complaint	medical_text_summary
"height_cm": 170, "weight_kg": 75, "heart_rate_bpm": 85, "systolic_bp_mmHg": 130, "diastolic_bp_mmHg": 80, "oxygen_saturation_%": 98, "bmi": 25.95	"gender": "Male", "marital_status": "Married", "race": "Asian", "age": 45, "physical_activity": "Moderate"	Going to the bathroom a lot and being very thirsty	The patient has had polyuria and polydipsia for the past month. Recent lab results show that fasting blood glucose levels are higher than normal. A family history of type 2 diabetes. There is no significant history of other chronic conditions reported.

The diabetes classification model is meant to predict and put patients into three main groups:

- None
- Type 1
- Type 2

The model's output labels are based on the ICD_CODE column, which uses the International Classification of Diseases (ICD) system. ICD-9 and ICD-10 codes can both be used to find diabetes mellitus. For example, codes E10.x for Type 1 diabetes and E11.x for Type 2 diabetes (ICD-10) or codes 250.x1/250.x3 for Type 1 and 250.x0/250.x2 for Type 2 (ICD-9). This labeling method ensures everything is the same, works with international medical databases, and is compatible. Data from different schemas in the MIMIC-IV and MIMIC-IV-ED datasets is then extracted and merged into a single comprehensive dataset for model training and testing.

3.3. Data Preprocessing

Before the model gets the data, the data preprocessing workflow is set up for each feature type to make sure it is consistent and works well. For numerical features, the median is used to fill in missing values. This method works well even when there are outliers [31]. After that, StandardScaler is used to standardize the data so that all of the features are on the same scale [32]. For categorical features, missing values are filled in with the mode (the most common value) [33]. Then, One-Hot Encoding is used to turn these features into numbers. This makes a new binary column for each unique category [34]. To process text features, the chief complaint and text_Summary columns are first combined. Then, Term Frequency-Inverse Document Frequency (TF-IDF) is used to extract features, but only the top 5000 features [35]. All of these steps are put together in one scikit-learn Pipeline. This method automates the workflow, makes sure that processing is always the same, and most importantly, stops data from leaking between the training and test sets [36].

3.4. The Prototype of the Diagnostic Solution for Diabetes Application

The diabetes diagnosis application is built using modern software engineering tools to ensure ease of use and good performance. The front-end is designed to make it easy for doctors and healthcare professionals to enter data and clearly view the diagnosis results. The back-end uses classification results from a set of ML models, such as Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, and XGBoost. We chose these

models because of the characteristics of the problem and their performance in previous studies. The system is made to quickly and accurately analyze complicated medical data and make predictions. This gives doctors the tools they need to make better decisions about patient care. The application shows that smart healthcare systems could use it in many different ways. It also shows how modern software development and machine learning can work together to make healthcare better.

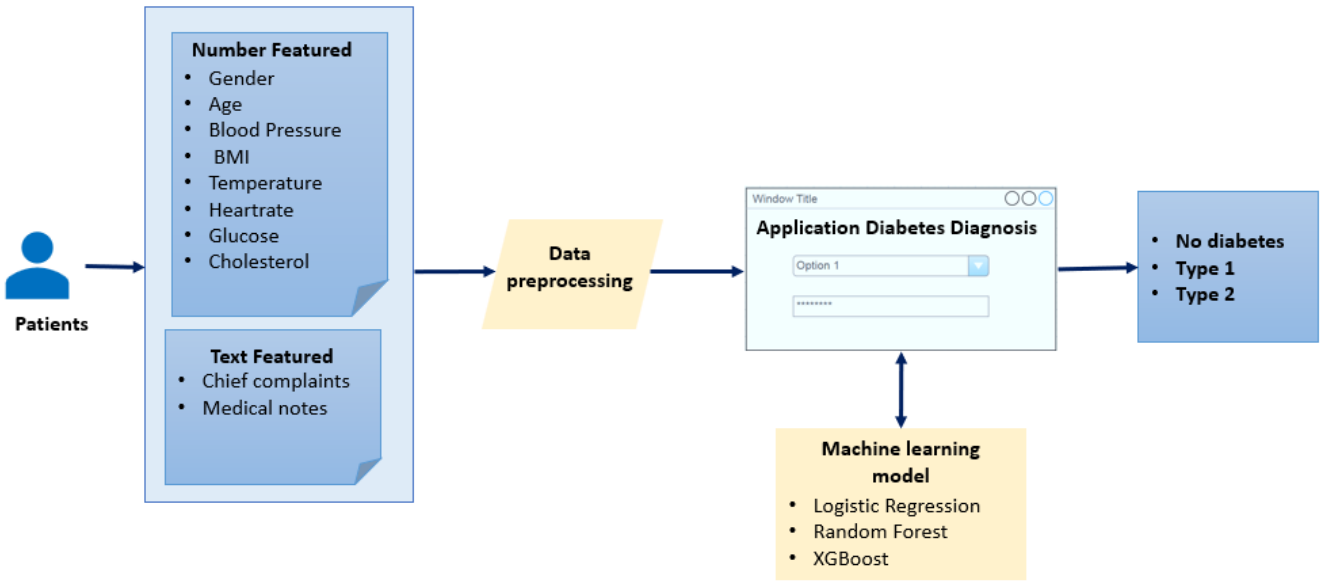

Fig. 2 Proposed method for diabetes classification using integrated data

4. Implementation and Experiments

Modern front-end technology is used to create interface demos for the published solution. Based on a thorough review of the literature, the XGBoost model was chosen to train and test on the MIMIC-IV dataset. The model was trained on Google Colab using an NVIDIA A100 GPU. To speed up the process and improve performance, we used a combination of libraries such as Scikit-learn, TensorFlow, and PyTorch. This study uses standard classification metrics such as Precision, Recall, F1-Score, and Accuracy [1] to evaluate the performance of the model. A confusion matrix is then used to visualize the results. This matrix details the number of correct and incorrect predictions made by the model on the test set.

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$f1 - score = \frac{2 * precision * recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Additionally, the AUC-ROC metric is also used to evaluate the performance of the classification model. The higher the AUC (closer to 1), the better the model is at discriminating between classes. AUC-ROC is often used in imbalanced datasets because it not only measures the overall accuracy but also considers the performance for each

class [1]. After completing the data preprocessing, this study conducted experiments on diabetes classification models on various classification algorithms. The performance of the models was evaluated based on the previously mentioned metrics. The detailed results of the experiments are presented in Table 5, and to visualize the experimental results, this study also evaluates them using the AUC-ROC measure; the results are shown in Figure 3.

Table 5. Detailed experimental results

Classifier	AUC-ROC	Accuracy	Precision (Macro Avg)	Recall (Macro Avg)	F1-Score (Macro Avg)
Logistic Regression	0.882	0.730	0.724	0.723	0.723
Random Forest	0.850	0.730	0.724	0.726	0.723
Gradient Boosting	0.887	0.730	0.717	0.716	0.709
SVM (RBF Kernel)	0.797	0.604	0.604	0.592	0.593
XGBoost	0.893	0.757	0.741	0.739	0.735

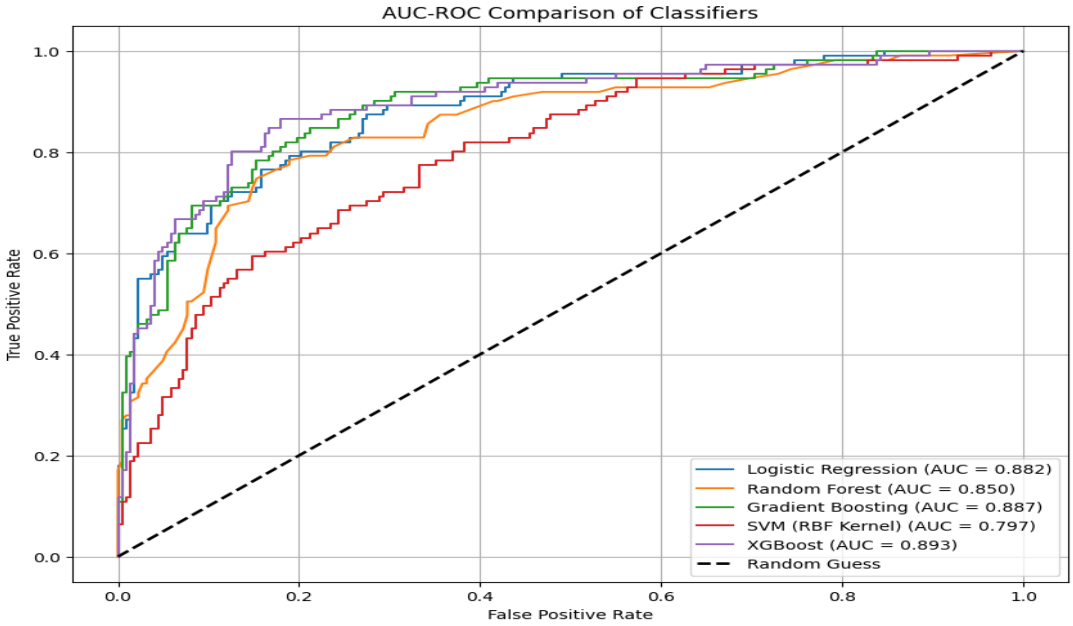

Fig. 3 Performance evaluation by AUC-ROC

Based on the results above, XGBoost is the best model for the application solution because it works the same for all labels and is especially good at telling the difference between "None" and "Type 2." In some cases, like when you want to focus on the "None" label, Gradient Boosting and Logistic Regression can be used. Logistic Regression is a good example of a simple model that works well for the "Type 1" label. You should not use SVM because it does not work well. You can improve XGBoost even more by tweaking its hyperparameters, using methods to deal with data imbalance if the label distribution isn't even, or combining it with other strong models, like an ensemble of XGBoost and Random Forest, to get the best of both worlds. These changes could make the diabetes classification problem even better.

Next, an application interface prototype is developed to illustrate the proposed solution, including full numeric and text feature information as stated. Based on the input information, the application will process through the machine learning model in the back-end to make output predictions. For this research classification model, this application will use XGBoost with the optimal parameter set after fine-tuning: learning_rate=0.05, max_depth=6,

n_estimators=500, and subsample=0.8. Modern interface technologies are used to design an illustrative interface for the solution. The application prototype is illustrated below:

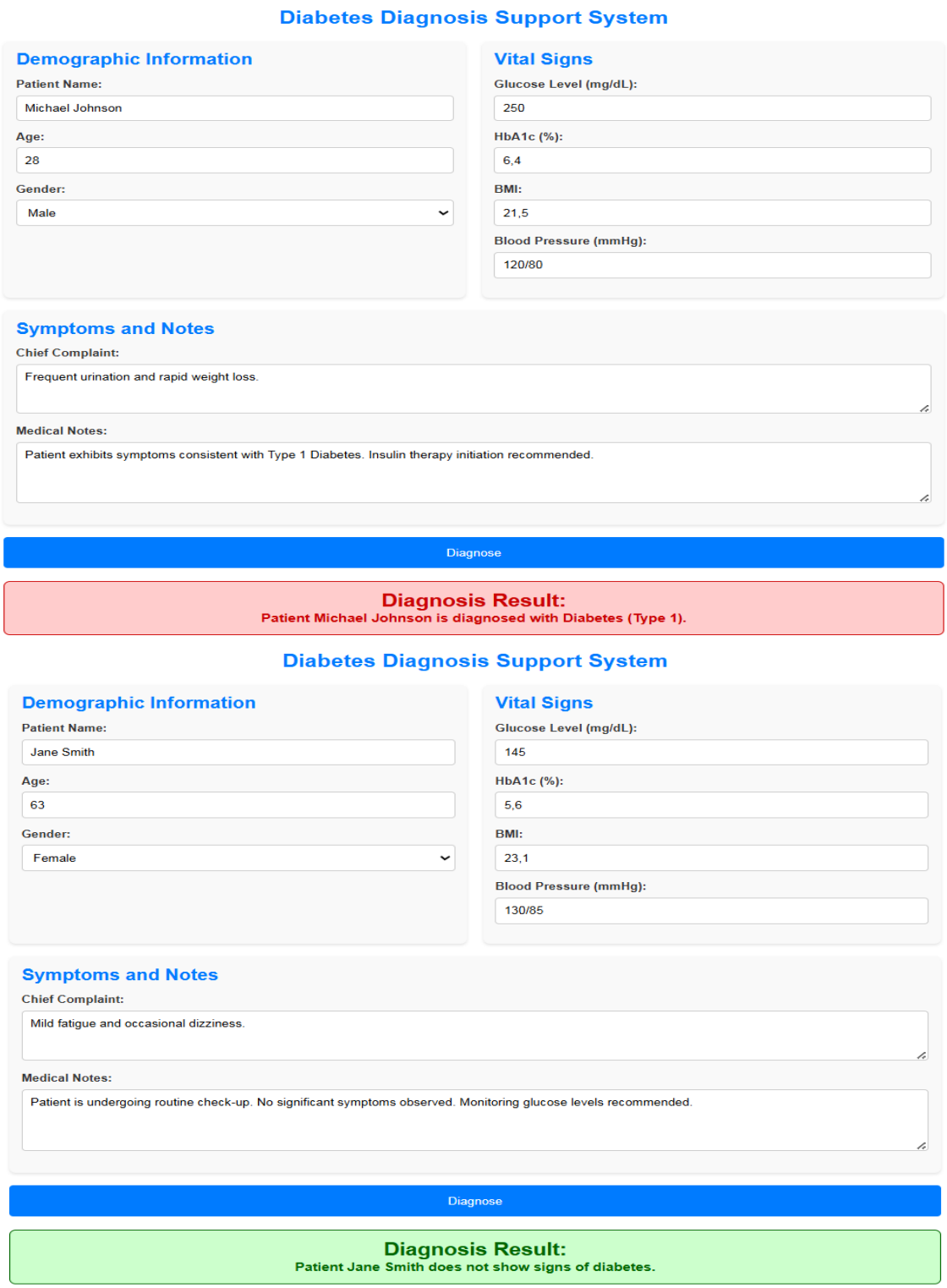

Fig. 4 Prototype application interface to demonstrate the proposed solution

The interface test shows that the app works well and gives users all the information they need, just like the proposed solution said it would. The interface provides clear and useful information that helps doctors figure out if someone has diabetes. The interface is also stable and mature, which means it has a lot of potential for real-world use. This means it could be used in modern clinical support systems.

5. Conclusion

This study analyzed sophisticated machine learning models for disease diagnosis, focusing on the application of clinical data and vital signs from the MIMIC-IV dataset. The proposed classification model for diabetes mellitus utilizing XGBoost demonstrated superior performance with high accuracy, indicating its potential applicability in real-world healthcare systems. The model has become a useful tool for making decisions because it uses the rich features of MIMIC-IV and advanced data processing methods. This has made diagnoses better and made better use of healthcare resources.

A prototype of a diagnostic interface was created to demonstrate the practical application of the model. You can enter basic clinical information like age, gender, glucose, HbA1c, BMI, and blood pressure into this interface, and it will show you a visual classification result (Type 1 Diabetes, Type 2 Diabetes, or No Diabetes). This is a big step toward turning the model from research into clinical practice, which will help doctors make quick and accurate decisions. Even though the results are promising, the study still has some problems. First, the model was only trained and tested on the MIMIC-IV dataset. It was not cross-validated on other, more varied data sources, which could make it less useful when used with other types of patients. Second, the study has not made any major changes to the model architecture or the specific loss function, and it has not fully used advanced hyperparameter optimization methods, so the current performance might not be the best it can be.

In the future, we will continue to improve the model by adding more medical features such as laboratory tests (HbA1c, glucose), medical images (X-rays, CT scans), and physiological signals (ECG, EEG). Deep learning models such as CNNs and RNNs will also be able to work with text, images, and multidimensional time series data. Methods such as oversampling and ensemble learning will also be used for more accurate classification. The model is expected to achieve enhanced accuracy and to expand its applicability within actual healthcare systems, tailored to the unique demographic and epidemiological characteristics of diverse regions.

Acknowledgments

This research is supported by Sai Gon University under grant number CSB.2025.065.

References

[1] Alberto Garcés-Jiménez et al., "Predictive Health Monitoring: Leveraging Artificial Intelligence for Early Detection of Infectious Diseases in Nursing Home Residents through Discontinuous Vital Signs Analysis," *Computers in Biology and Medicine*, vol. 174, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[2] Leila Ismail, Huned Materwala, and Juma Al Kaabi, "Association of Risk Factors with Type 2 Diabetes: A Systematic Review," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 1759-1785, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Mohamed Khalifa, and Mona Albadawy, "Artificial Intelligence for Diabetes: Enhancing Prevention, Diagnosis, and Effective Management," *Computer Methods and Programs in Biomedicine Update*, vol. 5, pp. 1-14, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Luis Fregoso-Aparicio et al., "Machine Learning and Deep Learning Predictive Models for Type 2 Diabetes: A Systematic Review," *Diabetology & Metabolic Syndrome*, vol. 13, no. 1, pp. 1-22, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Toshita Sharma, and Manan Shah, "A Comprehensive Review of Machine Learning Techniques on Diabetes Detection," *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, no. 1, pp. 1-16, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Kuo Ren Tan et al., "Evaluation of Machine Learning Methods Developed for Prediction of Diabetes Complications: A Systematic Review," *Journal of Diabetes Science and Technology*, vol. 17, no. 2, pp. 474-489, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] Negar Orangi-Fard, "Prediction of COPD Using Machine Learning, Clinical Summary Notes, and Vital Signs," *arXiv Preprint*, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Raúl López-Izquierdo et al., "Clinical Phenotypes and Short-Term Outcomes based on Prehospital Point-of-Care Testing and on-Scene Vital Signs," *npj Digital Medicine*, vol. 7, no, 1, pp. 1-8, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Beatriz Soares et al., "Impact and Classification of Body Stature and Physiological Variability in the Acquisition of Vital Signs Using Continuous Wave Radar," *Applied Sciences*, vol. 14, no. 2, pp. 1-20, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Michael Fascia, "Machine Learning Applications in Medical Prognostics: A Comprehensive Review," *arXiv preprint*, pp. 1-30, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Rosa Verhoeven et al., "Using Vital Signs for the Early Prediction of Necrotizing Enterocolitis in Preterm Neonates with Machine Learning," *Children*, vol. 11, no. 12, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Gayoung Kim, "Deep Learning Model for Predicting Critical Patient Conditions," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 24, no. 3, pp. 287-294, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Murad A. Rassam, and Amal A. Al-Shargabi, "Monitoring Critical Health Conditions in the Elderly: A Deep Learning-Based Abnormal Vital Sign Detection Model," *Technologies*, vol. 12, no. 12, pp. -23, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Haeil Park, and Chan Seok Park, "A Machine Learning Approach for Predicting in-Hospital Cardiac Arrest using Single-Day Vital Signs, Laboratory Test Results, and International Classification of Disease-10 Block for Diagnosis," *Annals of Laboratory Medicine*, vol. 45, no. 2, pp. 209-217, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Sohaib R. Awad, and Faris S. Alghareb, "Encoding-based Machine Learning Approach for Health Status Classification and Remote Monitoring of Cardiac Patients," *Algorithms*, vol. 18, no. 2, pp. 1-24, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[16] Alberto Gudiño-Ochoa et al., "Enhanced Diabetes Detection and Blood Glucose Prediction using Tinyml-Integrated E-Nose and Breath Analysis: A Novel Approach Combining Synthetic and Real-World Data," *Bioengineering*, vol. 11, no. 11, pp. 1-26, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Tung-Lai Hu et al., "Machine Learning-Based Predictions of Mortality and Readmission in Type 2 Diabetes Patients in the ICU," *Applied Sciences*, vol. 14, no. 18, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Priya Shirley Muller et al., "Improving Diabetes Diagnosis in Instantaneous Situations with MANET and Data Mining," *Journal of Environmental Protection and Ecology*, vol. 25, no. 4, pp. 1330-1343, 2024. [Google Scholar]

[19] Badriah Alkalifah et al., "Evaluation of Machine Learning-Based Regression Techniques for Prediction of Diabetes Levels Fluctuations," *Heliyon*, vol. 11, no. 1, 1-12, 2025. [Google Scholar] [Publisher Link]

[20] Furqan Rustam et al., "Enhanced Detection of Diabetes Mellitus using Novel Ensemble Feature Engineering Approach and Machine Learning Model," *Scientific Reports*, vol. 14, no. 1, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] Jorge A. Morgan-Benita et al., "Setting Ranges in Potential Biomarkers for Type 2 Diabetes Mellitus Patients Early Detection By Sex-An Approach with Machine Learning Algorithms," *Diagnostics*, vol. 14, no. 15, pp. 1-43, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[22] Paidipati Dinesh, A.S. Vickram, and P. Kalyanasundaram, "Medical Image Prediction for Diagnosis of Breast Cancer Disease Comparing the Machine Learning Algorithms: SVM, KNN, Logistic Regression, Random Forest and Decision Tree to Measure Accuracy," *AIP Conference Proceedings*, vol. 2853, no. 1, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[23] Surajit Das et al., "Machine Learning in Healthcare Analytics: A State-of-the-Art Review," *Archives of Computational Methods in Engineering*, vol. 31, no. 7, pp. 3923-3962, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[24] Shikha Prasher, and Leema Nelson, "Early Prediction of Obesity Risk in Older Adults using XGBoost Classifier," *2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT)*, Kollam, India, pp. 1599-1603, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[25] Prince Jain et al., "Enhanced Cardiovascular Diagnostics using Wearable ECG and Bioimpedance Monitoring with LightGBM Classifier," *Biosensors and Bioelectronics: X*, vol. 24, pp. 1-7, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[26] Abdulaziz Aldaej, Tariq Ahamed Ahanger, and Imdad Ullah, "Deep Neural Network-Based Secure Healthcare Framework," *Neural Computing and Applications*, vol. 36, no. 28, pp. 17467-17482, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[27] Naif Al Mudawi et al., "Innovative Healthcare Solutions: Robust Hand Gesture Recognition of Daily Life Routines using 1D CNN," *Frontiers in Bioengineering and Biotechnology*, vol. 12, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[28] Alistair Johnson et al., "MIMIC-IV," *PhysioNet*, RRID:SCR_007345, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[29] Alistair Johnson et al., "MIMIC-IV-ED (version 2.2)," *PhysioNet*, RRID:SCR_007345, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[30] Nilesh Kumar Sahu et al., "Leveraging Language Models for Summarizing Mental State Examinations: A Comprehensive Evaluation and Dataset Release," *Proceedings of the 31st International Conference on Computational Linguistics*, Abu Dhabi, UAE, pp. 2658-2682, 2025. [Google Scholar] [Publisher Link]

[31] Luke Oluwaseye Joel, Wesley Doorsamy, and Babu Sena Paul, "On the Performance of Imputation Techniques for Missing Values on Healthcare Datasets," *arXiv Preprint*, pp. 1-20, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[32] Zulfikar Setyo Priyambudi, and Yusuf Sulistyo Nugroho, "Which Algorithm is better? An Implementation of Normalization to Predict Student Performance," *AIP Conference Proceedings*, vol. 2926, no. 1, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[33] JiaHang Li et al., "Comparison of the Effects of Imputation Methods for Missing Data in Predictive Modelling of Cohort Study Datasets," *BMC Medical Research Methodology*, vol. 24, no. 1, pp. 1-9, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[34] Yi Sun et al., "Modifying the One-Hot Encoding Technique Can Enhance the Adversarial Robustness of the Visual Model for Symbol Recognition," *Expert Systems with Applications*, vol. 250, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[35] Zakia Labd et al., "Text Classification Supervised Algorithms with Term Frequency Inverse Document Frequency and Global Vectors for Word Representation: A Comparative Study," *International Journal of Electrical & Computer Engineering*, vol. 14, no. 1, pp. 589-599, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[36] Jiang Wu et al., "Data Pipeline Training: Integrating Automl to Optimize the Data Flow of Machine Learning Models," *arXiv preprint*, pp. 1-5, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[37] S. Sathyanarayanan, and B. Roopashri Tantri, "Confusion Matrix-Based Performance Evaluation Metrics," *African Journal of Biomedical Research*, vol. 27, no. 4S, pp. 4023-4031, 2024. [CrossRef] [Google Scholar] [Publisher Link]