SJCS Scientific Journal of Computer Science

futuristech PT. Teknologi Futuristik Indonesia

Date of publication March 29, 2026, date of current version March 29, 2026.
Digital Object Identifier 10.64539/sjcs.v2i2.2026.439

Article

Machine Learning-Based Diabetes Classification Using Vital Signs and Clinical Information from the MIMIC-IV Dataset

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

[1] Faculty of Information Technology, Saigon University (SGU), HCM, Vietnam; e-mail: hnkhuy@sgu.edu.vn (H. Huynh), thanh.cao@sgu.edu.vn (T. Cao).
[2] NTT Institute of International Education (NIIE), Nguyen Tat Thanh University, HCM, Vietnam; e-mail: tshai@ntt.edu.vn (H. Tran)
* Correspondence

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

Abstract: Diagnosing diabetes based on clinical data is very important because the number of people with diabetes is growing around the world. The main focus of this study is on using machine learning models to figure out what kind of sickness someone has from a variety of clinical data. The MIMIC-IV dataset was used, which has both structured and unstructured data. The structured data includes vital signs, demographics, and lab tests. The unstructured data includes medical notes, major complaints, and a list of medications. Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, and XGBoost were some of the models that were tested. Accuracy, Precision, Recall, F1-score, and AUC-ROC were used to measure how well the models worked. When random text data was added to the experiments, the results showed a big improvement in performance**:** the accuracy increased from approximately 68% to up to 87% across models**.** The best-performing models achieved AUC-ROC values above 0.95, with Random Forest and XGBoost showing the strongest performance**.** This shows that semantic mining from clinical notes is a key part of making medical decision support systems more reliable.

Keywords: Diabetes Classification; MIMIC-IV; Natural Language Processing; Ensemble Learning; Clinical Decision Support; Machine Learning; Vital Signs.

1. Introduction

Among noncommunicable diseases, diabetes mellitus (DM) has become one of the biggest problems in the world, with projected global prevalence exceeding 700 million adults by 2045 [1]. The disease comprises two main classifications with distinctly different etiologies: Type 1, characterized by insulin deficiency due to an autoimmune mechanism, and Type 2, due to progressive insulin resistance combined with insufficient insulin secretion. Both classifications share a common pathway of chronic hyperglycemia which, if left untreated or misdiagnosed at screening, leads to life-threatening complications including diabetic ketoacidosis, hyperosmolar hyperglycemia, end-stage renal disease, and cardiovascular events. Therefore, timely and accurate classification is essential not only for prognostic purposes but also to initiate the correct treatment pathway from the first contact within the healthcare system [2].

Artificial intelligence and machine learning have been used in clinical evaluation a lot more in the last ten years. Modern systems demonstrate considerable capabilities in analyzing diverse data modalities—including structured biomarkers such as vital signs and laboratory values to unstructured textual records such as discharge summaries, medication diaries, and principal symptom narratives. In the fields of endocrinology and emergency medicine, these tools are particularly promising: they can assist overwhelmed clinicians in risk stratification, accelerate differential diagnosis, and potentially reduce the rate of adverse events due to misclassification of disease type.

Despite these advances, a specific and persistent challenge remains unresolved. In the acute care setting, the physiological manifestations of type 1 and type 2 diabetes converge significantly: both groups may exhibit tachycardia, high blood glucose, and similar respiratory profiles upon admission. Therefore, conventional machine

learning processes trained exclusively on structured numerical traits struggle to establish reliable classification boundaries, yielding accuracy of only about 68% in recent reviews. Two structural shortcomings explain this limitation: first, the inefficient use of unstructured clinical texts encoding medication history, symptom descriptions, and diagnostic reasoning [3]; second, the scarcity of sufficiently large and demographically diverse datasets needed to fully capture the differences in real-world cases. Publicly available standard datasets such as the Pima Human Diabetes Dataset or similar repositories from a single organization are often too small and too homogeneous to support robust multiclass models for large-scale deployment [4].

This study directly addresses these limitations by conducting a systematic experimental comparison of five classical machine learning classifiers under two feature configurations, using a balanced 6,000-encounter cohort derived from the MIMIC-IV and MIMIC-IV-ED repositories. The core contribution is a multimodal pipeline that augments structured vital-sign and laboratory features with TF-IDF representations of chief complaints, BART-model-summarised discharge narratives, and medication-reconciliation lists processed through natural language processing. By contrasting performance in the text-absent and text-present conditions, the study furnishes controlled empirical evidence of the incremental value of clinical text for three-class diabetes triage. Additionally, this work introduces one-vs-rest macro AUC estimation to ensure statistically valid discrimination measurement in the multiclass context, addresses label-assignment transparency through standardised ICD-9/10 coding, and discusses data-filtering decisions and sampling methodology to support reproducibility.

The rest of this essay is organized as follows. Additionally, Section 2 looks at related work on using machine learning to diagnose diabetes. It includes both traditional and deep learning techniques. Including dataset creation, feature engineering, preprocessing pipelines, and classifier configuration, Section 3 explains the suggested way. There is a detailed discussion of findings, limitations, and observed error patterns in Section 4. It shows the experimental evaluation framework and the comparative results achieved under both feature conditions. The paper concludes in Section 5 with a synthesis of key insights and a delineation of future research directions.

2. Related works

The intersection of machine learning and diabetes research has generated a rich body of literature spanning traditional biomarker-based approaches, large-scale intensive-care data mining, deep learning architectures, and real-time monitoring platforms. A selective review of recent advances reveals both the progress achieved and the gaps that motivated the present work.

Investigations focused on traditional clinical biomarkers have produced encouraging initial results. Morgan-Benita et al. [5] constructed a Random Forest model from lipid and blood-pressure indicators across a cohort of 1,726 patients, attaining an accuracy of 88.2%; however, the authors acknowledged that external validation across heterogeneous populations remained outstanding. In parallel, Gudiño-Ochoa et al. [6] pursued a non-invasive strategy by coupling an electronic nose with breath analysis, reporting 94% accuracy when Random Forest, XGBoost classifiers were applied to exhaled volatile compounds, albeit without integration of standard vital-sign covariates.

Leveraging large-scale intensive-care repositories has become increasingly tractable with the public release of MIMIC derivatives. Hu et al. [7] mined MIMIC-III records from more than 14,000 patients to predict 30-day mortality and ICU readmission among Type 2 diabetes cases, with AdaBoost achieving an AUROC of 0.7952 for mortality. The authors noted that the single metropolitan health-system origin of MIMIC constrained generalisability to community or rural hospital settings. Deep learning architectures have meanwhile pushed accuracy metrics substantially higher: Rustam et al. [8] combined convolutional and long short-term memory layers for feature extraction before applying a Random Forest classifier, claiming 99% accuracy across three public benchmarks. Nevertheless, the substantial computational overhead of this approach limits its deployment in resource-constrained environments, and the reliance on curated public datasets leaves real-world robustness unverified.

Real-time glycaemic monitoring represents another active research strand. Muller et al. [9] constructed a MANET-based platform incorporating recurrent gated units, reporting 95.7% accuracy in blood-glucose variability prediction, though the system's dependence on wireless mesh infrastructure and unresolved data-security considerations impede routine clinical adoption. Alkalifah et al. [10] conducted a regression benchmark across 14,733 continuous glucose monitoring records, finding that Gaussian Process Regression yielded the lowest mean squared error of 1.64 mg/dL; the study, however, confronted pronounced class-level data imbalance. Collectively, the above body of work illustrates that accuracy gains are consistently observed when richer feature representations and larger, more diverse datasets are employed, yet methodological transparency regarding data selection and multiclass evaluation protocols remains uneven.

Within the classical supervised learning paradigm, several well-established algorithm families have been widely adopted for clinical classification tasks. Logistic Regression serves as an interpretable linear baseline that estimates class probabilities through a sigmoid transformation of a weighted feature sum; its principal limitation is an inability to capture non-linear decision boundaries without explicit feature engineering [11]. SVMs address

this shortcoming by identifying the maximum-margin hyperplane that best separates class instances, using kernel functions, like the radial basis function, to project data into higher-dimensional areas where linear separation is possible [11]. Random Forest constructs an ensemble of decision trees through bootstrap aggregation, averaging their outputs to reduce variance and improve generalisation on high-dimensional medical data [11]. The Gradient Boosting family adopts an additive strategy: a sequence of shallow trees is fitted iteratively, each correcting the cumulative residual error of the preceding ensemble [12]. XGBoost [13] refines this approach through parallelised tree construction, column sub-sampling, and L1/L2 regularisation, achieving state-of-the-art performance on tabular biomedical datasets. LightGBM [14] further accelerates training via a histogram-based, leaf-wise growth strategy that is especially advantageous on feature-rich records.

Artificial neural networks extended to many hidden layers—commonly termed Deep Neural Networks (DNNs) are a flexible, data-driven way to learn hierarchical representations from raw inputs without having to create features by hand [15]. In the field of biomedicine, DNN depth lets you find abstract clinical trends that shallow models might miss; however, this expressive capacity comes at the cost of substantial computational demands, a large labelled-sample requirement, and reduced transparency compared with tree-based methods. Convolutional Neural Networks (CNNs) introduce spatial inductive biases through learnable filter banks and pooling operations, making them well-suited to structured physiological time-series and medical imaging modalities [16]. In prior diabetes research, CNN-derived features have been transferred to downstream classifiers to boost predictive accuracy, though the resulting pipelines typically require GPU acceleration and impose significant engineering overhead that can be prohibitive in resource-constrained clinical settings.

Ensemble and hybrid architectures that combine the predictions of multiple heterogeneous base learners have demonstrated consistent gains in discriminative accuracy and probability calibration relative to any single constituent model [17]. In multi-class clinical settings, such approaches reduce the risk of high-variance predictions on minority classes—a particularly important property for rare-but-serious conditions such as Type 1 diabetes in adult emergency cohorts. Beyond simple voting or averaging, stacked generalisation introduces a meta-learner that learns an optimal weighting of base-model outputs, typically yielding further improvements in precision and recall balance. The present study evaluates five classifiers under a unified protocol, laying the empirical groundwork for future stacked architectures that may further narrow the performance gap between structured-only and multimodal configurations.

3. Methodology

3.1. Dataset Selection

This study draws its clinical data from two complementary sub-repositories of the MIMIC (Medical Information Mart for Intensive Care) project: MIMIC-IV and MIMIC-IV-ED [18], [19]. MIMIC-IV is a longitudinal, de-identified electronic health-record (EHR) database assembled from inpatient admissions at Beth Israel Deaconess Medical Center (BIDMC) spanning several decades. Its modular schema separates data by care setting and information type, enabling researchers to selectively join only the tables relevant to a given research question while maintaining referential integrity across admissions, laboratory results, prescriptions, and clinical notes. De-identification was performed in compliance with HIPAA Safe Harbor provisions, removing all eighteen Protected Health Information identifiers and replacing dates with shifted temporal offsets that preserve within-patient chronology. The database encompasses records for more than 40,000 ICU patients, providing the demographic breadth and clinical richness required to support robust multi-class modelling of diabetes subtypes [18].

Within MIMIC-IV, the hospitalisation (hosp) module aggregates data sourced from the institution-wide EHR system and constitutes the backbone of the structured feature set used in this study. Its constituent tables collectively document [18]:

- Demographics and admission logistics: patient demographics (patients), hospitalisation records (admissions), and ward-level transfer events (transfers)
- Quantitative laboratory findings: time-stamped test results (labevents) and their corresponding item catalogue (d_labitems)
- Microbiological investigation records: culture specimens and sensitivity results (microbiologyevents)
- Clinical order data: physician-entered orders and associated detail annotations (poe, poe_detail)
- Drug administration records: electronic medication administration records and line-item detail (emar, emar_detail)
- Pharmacological prescriptions: discharge prescriptions and pharmacy fulfilment data (prescriptions, pharmacy)
- Billing and diagnosis coding: ICD-coded diagnoses and procedures, HCPCS event codes, and DRG assignments (diagnoses_icd, d_icd_diagnoses, procedures_icd, d_icd_procedures, hcpcsevents, d_hcpcs, drgcodes)
- Care-team routing: records of the clinical service responsible for each patient episode (services)

The emergency-department extension, MIMIC-IV-ED [19], complements the inpatient module by capturing approximately 425,000 ED visits at BIDMC between 2011

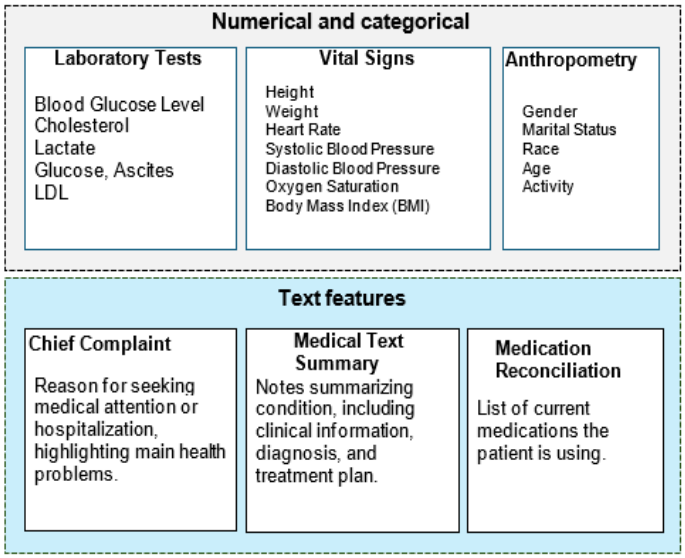


Figure 1. Overview of Numerical, Categorical, and Textual Features Used for Model Training.

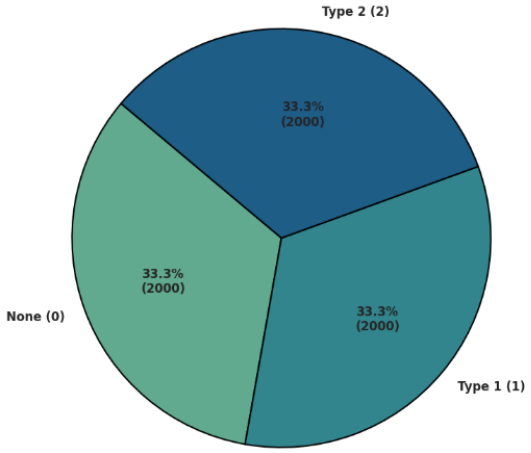


Figure 2. Class distribution in the balanced 6,000-encounter cohort (2,000 samples per class).

and 2019. Crucially for this study, it contributes triage-point vital signs recorded at first clinical contact, chief-complaint free-text fields entered by nursing staff, medication-reconciliation lists compiled at admission, and final discharge diagnoses. Because these observations precede definitive workup, they represent the information realistically available for early triage classification—an ecological validity advantage over datasets assembled from post-hoc inpatient summaries alone. Figure 1 summarises the overall input space used for model development, including numerical, categorical, and textual variables. Figure 2 shows the class distribution of the final balanced cohort, with 2,000 encounters in each class.

3.2. Feature Selection and Label Assignment

Outcome labels were assigned at the admission level using ICD billing codes. Under ICD-10, code prefix E10 designates Type 1 diabetes mellitus and E11 designates Type 2; under ICD-9, codes of the form 250.x1 or 250.x3 correspond to Type 1, while 250.x0 or 250.x2 correspond to Type 2. Admissions with no qualifying diabetes ICD entry were assigned the label None. The use of standardised international coding ensures label consistency and facilitates future comparisons against other MIMIC-derived studies.

Two mutually exclusive feature sets were constructed to enable a controlled ablation comparison:

- **Structured features (no-text condition):** heart rate, blood pressure (systolic and diastolic), body temperature, peripheral oxygen saturation (SpO_2), respiratory rate, body mass index (BMI), blood glucose, lactate, cholesterol, age, and sex.
- **Multimodal features (with-text condition):** all structured features above, plus TF-IDF representations of the chief-complaint field, BART-summarized discharge text, and a concatenated medication-reconciliation string (see Section 3.3 for preprocessing details).

3.3. Evaluation Metrics

Performance was quantified using a comprehensive suite of metrics computed on the held-out test set (20% of the cohort). The dataset was partitioned as follows:

- Training set (80%): used to fit classifier parameters and pipeline transformers.
- Test set (20%): reserved exclusively for unbiased evaluation; no transformation statistics from this partition were used during training.

The following metrics were computed for each classifier [1], [20]:

- Precision: the proportion of predicted positive instances that are truly positive — TP / (TP + FP).
- Recall (Sensitivity): the proportion of actual positive instances correctly identified — TP / (TP + FN).
- F1-Score: the harmonic mean of Precision and Recall — 2 × (Precision × Recall) / (Precision + Recall), providing a balanced summary when class support is equal.
- Accuracy: the fraction of all predictions that are correct — (TP + TN) / (TP + TN + FP + FN).

For multi-class discrimination, AUC-ROC was estimated using a one-vs-rest (OvR) strategy in which a separate binary ROC curve is computed for each class against the union of all remaining classes; the three resulting AUC values are then averaged (macro averaging) to yield a single scalar summary. This approach is appropriate for three-class problems and avoids the binary-classification assumption that would render a single ROC curve misleading. All per-class and macro-averaged AUC values are reported in the tables below.

Confusion matrices for all classifiers under both experimental conditions were additionally generated and are presented as heatmap visualisations to provide a granular view of per-class prediction errors. These matrices

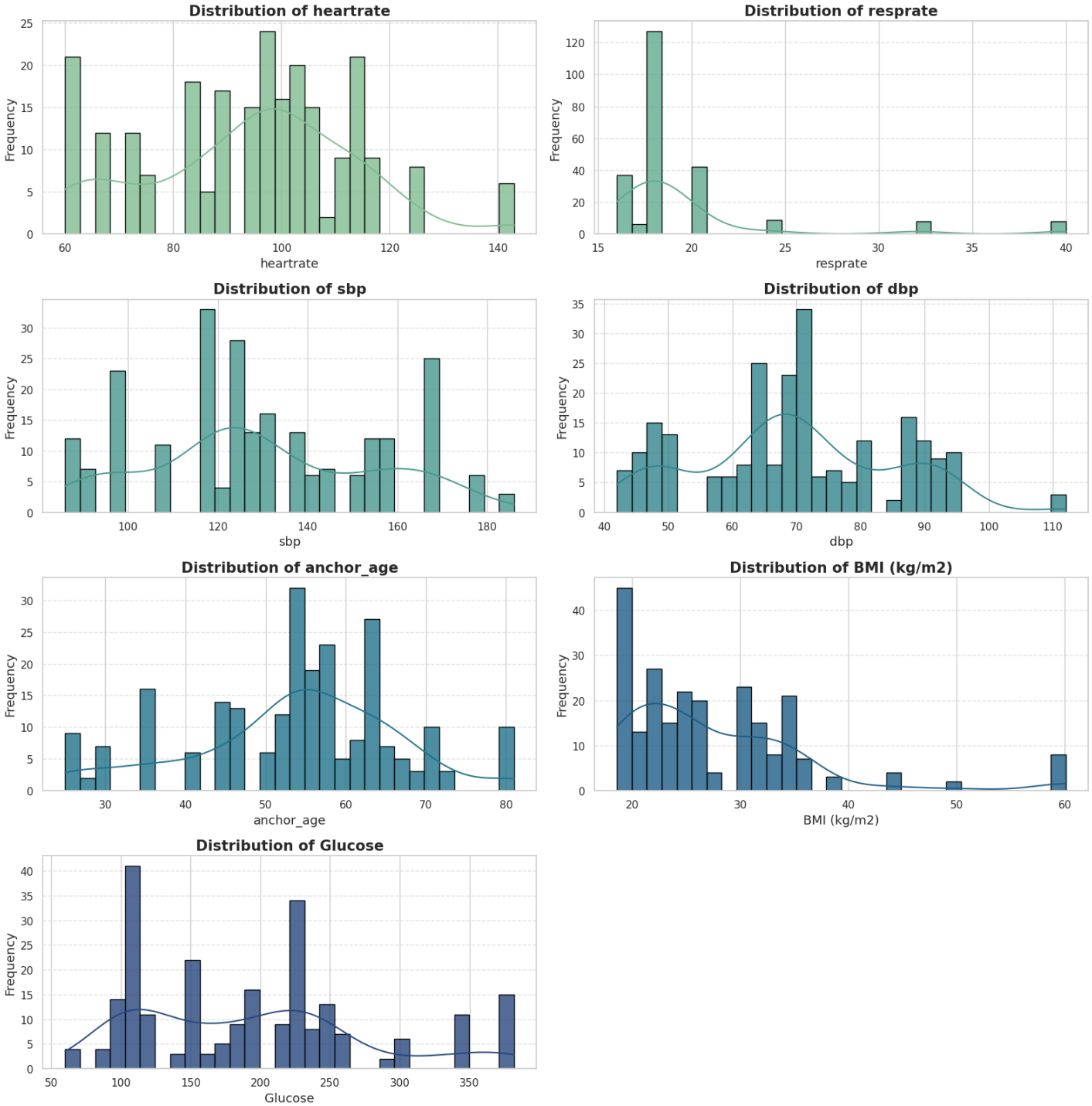


Figure 3. Frequency distributions of all seven structured numerical variables (heart rate, systolic BP, temperature, SpO_2, respiratory rate, BMI, blood glucose).

reveal, for instance, the degree of confusion between Type 1 and Type 2 at the feature-only baseline and the extent to which text integration resolves that ambiguity Figure 3 presents the frequency distributions of the structured numerical variables, providing an overview of their empirical ranges before model training. Figure 4 illustrates the Pearson correlation structure among the structured numerical variables.

3.4. Data preprocessing and model construction

The training cohort comprises 6,000 encounters assembled from MIMIC-IV and MIMIC-IV-ED through a multi-step filtering and balancing procedure. Starting from the full MIMIC-IV patient population (>40,000 records), admissions were selected according to the following eligibility criteria: (a) the presence of at least one diabetes-relevant ICD-9 or ICD-10 discharge code or an explicit None designation confirmed by the absence of any endocrine-chapter diagnosis; (b) availability of at least 70% of the required structured vital-sign fields; and (c) the existence of a legible chief-complaint entry in MIMIC-IV-ED. After applying these filters, surviving records were stratified by class label and randomly downsampled to 2,000 encounters per stratum, yielding a perfectly balanced

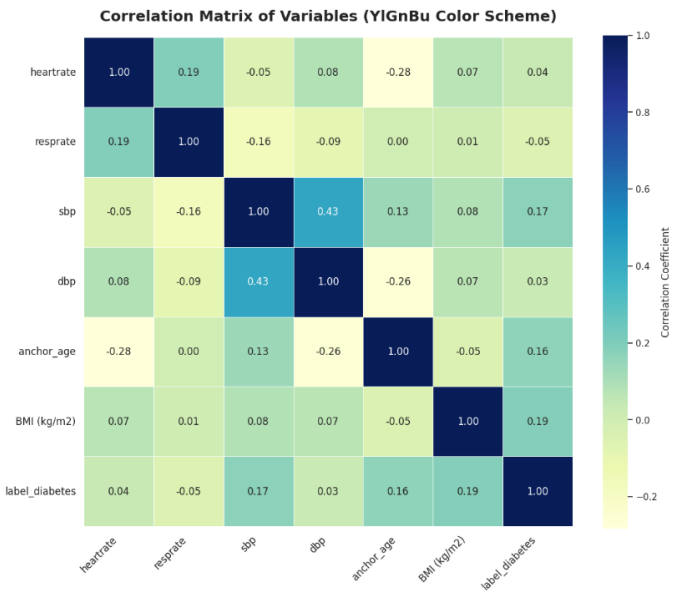


Figure 4. Pearson correlation matrix of structured numerical variables.

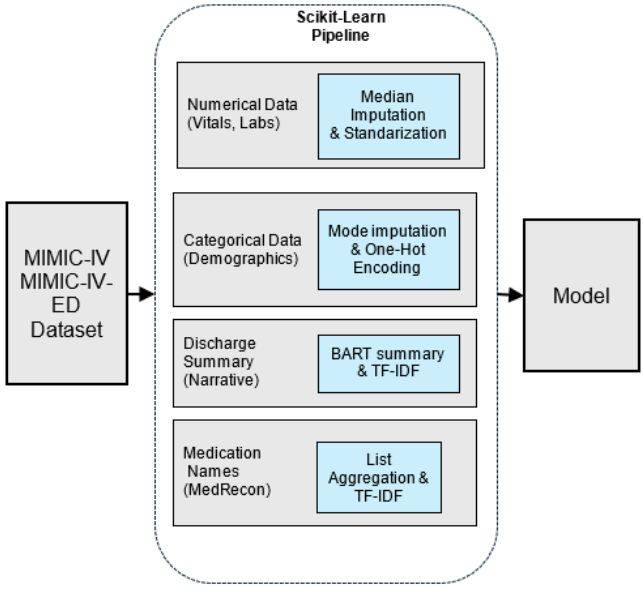


Figure 5. End-to-end data pre-processing pipeline.

cohort. This design eliminates majority-class bias and provides equal support for each target label during classifier training.

The preprocessing pipeline applied the following transformations sequentially (Figure 5):

a) Numerical imputation and normalisation: missing values in continuous fields were imputed with the column median; categorical missingness was resolved using the mode. All continuous features were subsequently scaled to unit variance using a StandardScaler fitted exclusively on the training partition [21], [22].

b) Categorical encoding: nominal variables such as sex and admission type were one-hot encoded, expanding each category into a binary indicator column [23].

c) Text vectorisation: free-text fields (chief complaint, medication list) were tokenised, lowercased, stripped of domain-specific stop words, and transformed into TF-IDF feature vectors with a maximum vocabulary size of 5,000 terms per field [24].

d) Discharge-note summarisation: raw discharge narratives are verbose and contain extensive boilerplate. To distil clinically relevant content while controlling dimensionality, each note was condensed to a 128-token summary using the facebook/bart-large-cnn model [25] before TF-IDF vectorisation was applied.

e) Medication-list concatenation: all drug names associated with a single admission were concatenated into a single string entry, preserving polypharmacy patterns that are clinically informative for diabetes subtype identification.

f) Label-leakage mitigation: A potential concern is that text fields may contain terms such as insulin or metformin that are strongly predictive of subtype labels. In clinical settings, however, information on a patient's current medication regimen is routinely available at the point of triage and forms part of standard medication reconciliation procedures. The model therefore leverages clinically relevant features rather than artificially restricted inputs. Future work will include systematic ablation studies to evaluate the contribution of individual text features and to better distinguish between informative clinical context and potential label leakage.

All transformations were encapsulated within a scikit-learn Pipeline object fitted solely on the training split to prevent any information leakage between training and test partitions [26]. The 80:20 train–test split was applied with stratification to preserve class proportions across both subsets.

4. Experimental results and discussion

Model training and evaluation were executed on Google Colab, leveraging an NVIDIA A100 GPU to accelerate matrix operations during feature construction and cross-validation. All classical classifiers (SVM, Logistic Regression, Random Forest, Gradient Boosting, and XGBoost) were implemented via the scikit-learn and XGBoost Python libraries. TensorFlow and PyTorch were available in the environment for potential deep-learning extensions but were not employed for the classical classifiers evaluated in this study. Experiments were conducted under two feature configurations—structured features only (no_text) and the full multimodal feature set (with_text)—to enable a controlled, pairwise comparison of text's incremental diagnostic value (Figure 6).

Table 1. Classification report — structured features only (no_text).

Classifier	None Prec	None Rec	None F1	T1 Prec	T1 Rec	T1 F1	T2 Prec	T2 Rec	T2 F1
SVM	0.578	0.631	0.603	0.471	0.426	0.447	0.605	0.590	0.598
Random Forest	0.703	0.746	0.724	0.714	0.479	0.573	0.655	0.779	0.712
Logistic Regression	0.510	0.592	0.548	0.402	0.351	0.375	0.540	0.500	0.519
Gradient Boosting	0.721	0.715	0.718	0.613	0.521	0.563	0.635	0.713	0.672
XGBoost	0.725	0.731	0.728	0.667	0.532	0.592	0.650	0.746	0.695

Table 2. Classification report — multimodal features (with_text).

Classifier	None Prec	None Rec	None F1	T1 Prec	T1 Rec	T1 F1	T2 Prec	T2 Rec	T2 F1
SVM	0.832	0.839	0.835	0.742	0.702	0.721	0.738	0.762	0.750
Random Forest	0.911	0.946	0.928	0.807	0.798	0.802	0.848	0.820	0.833
Logistic Regression	0.870	0.923	0.896	0.772	0.755	0.763	0.802	0.762	0.782
Gradient Boosting	0.912	0.954	0.932	0.817	0.809	0.813	0.829	0.795	0.812
XGBoost	0.939	0.946	0.943	0.843	0.798	0.820	0.825	0.853	0.839

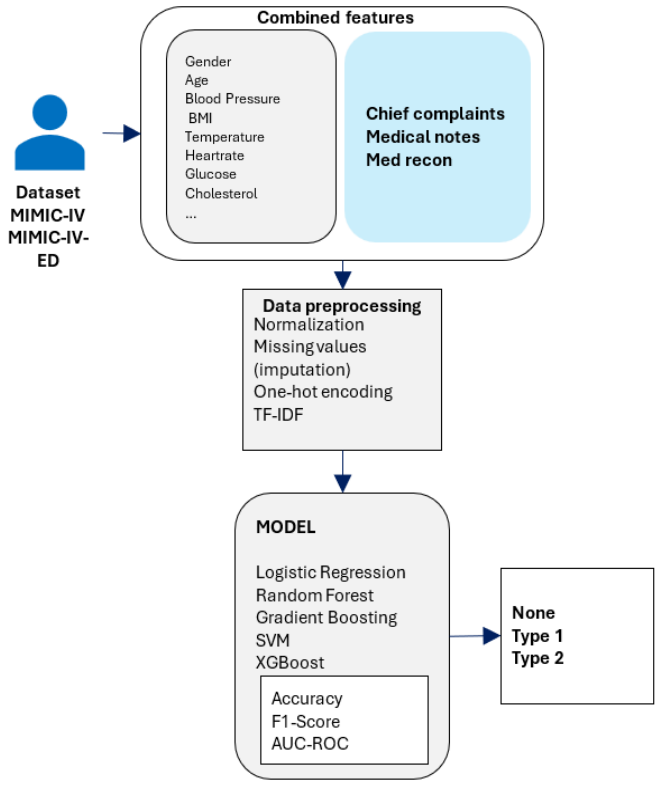


Figure 6. Experimental workflow: classifier training and evaluation under no-text and with-text configurations.

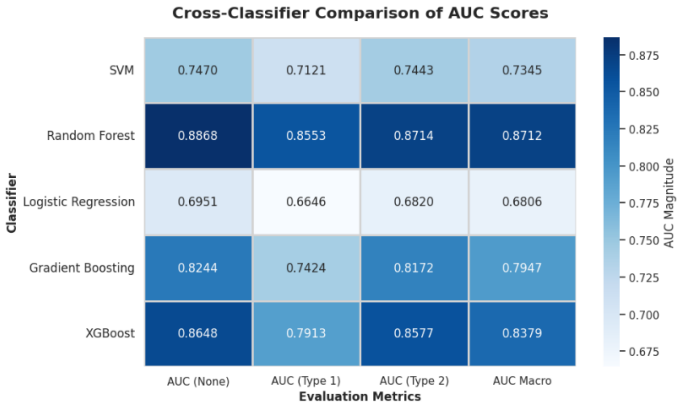


Figure 7. AUC-ROC (one-vs-rest, macro) — structured features only (no_text).

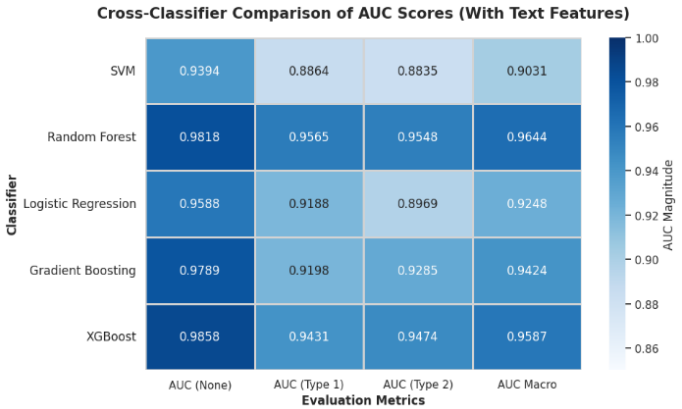


Figure 8. AUC-ROC (one-vs-rest, macro) — multimodal features (with_text).

4.1. Results

The experiments were conducted with two different cases: Structured Features Only and Multimodal Features. The results are presented below:

a) **Structured Features Only (no_text condition)**
When classifiers were restricted to structured vital-sign, demographic, and laboratory inputs, overall accuracy remained moderate across all models (Table 1 and Figure 7). This reflects the clinical reality that physiological measurements alone carry insufficient discriminative power to differentiate between diabetes subtypes at triage, particularly given the substantial overlap in vital-sign profiles between Type 1 and Type 2 presentations.

b) **Multimodal Features (with_text condition)**
Incorporating TF-IDF representations of the three text fields produced substantial and consistent performance improvements across all classifiers as reported in Table 2. The corresponding macro-AUC improvements are shown in Figure 8.

Table 3. Comparative summary: no_text vs. with_text across all classifiers (macro average).

Classifier	Acc No Txt	Acc W/Txt	Prec No Txt	Prec W/Txt	Rec No Txt	Rec W/Txt	F1 No Txt	F1 W/Txt
SVM	0.5607	0.7746	0.5510	0.7706	0.5488	0.7676	0.5491	0.7689
Random Forest	0.6850	0.8613	0.6908	0.8550	0.6679	0.8546	0.6696	0.8546
Logistic Regression	0.4942	0.8208	0.4841	0.8143	0.4811	0.8136	0.4807	0.8135
Gradient Boosting	0.6618	0.8584	0.6562	0.8527	0.6499	0.8525	0.6511	0.8523
XGBoost	0.6821	0.8728	0.6806	0.8690	0.6695	0.8655	0.6714	0.8670

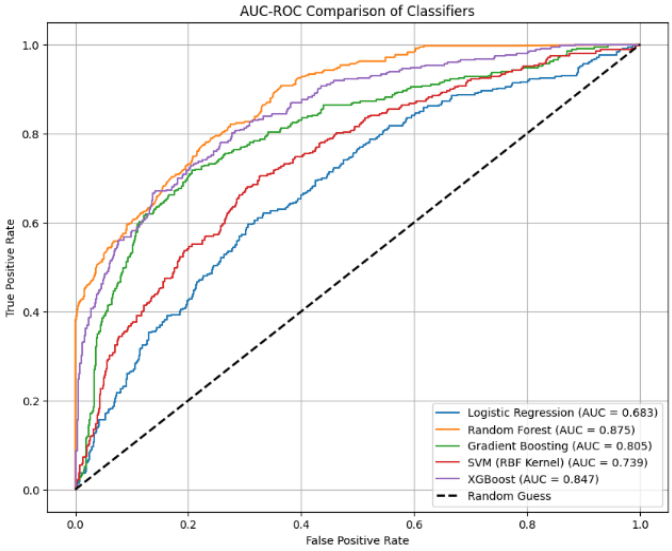


Figure 9. AUC-ROC curves (one-vs-rest, macro) — no_text condition.

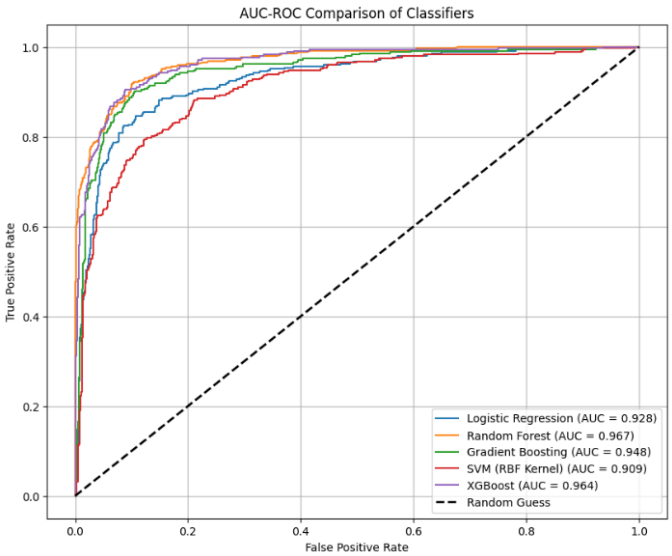


Figure 10. AUC-ROC curves (one-vs-rest, macro) — with_text condition.

Accuracy rose into the 77–88% range, and macro-AUC values for all models exceeded 0.90.

4.2. Discussion

As shown in Table 3, the integration of unstructured clinical text led to consistent performance improvements across all evaluated classifiers. Accuracy increased substantially for every model, with XGBoost achieving the highest score (0.8728), followed by Random Forest and Gradient Boosting, both exceeding 0.85 in the multimodal configuration. Logistic Regression exhibited the largest relative gain, improving from 0.4942 to 0.8208, indicating that textual features provide strong linear separability when combined with structured variables.

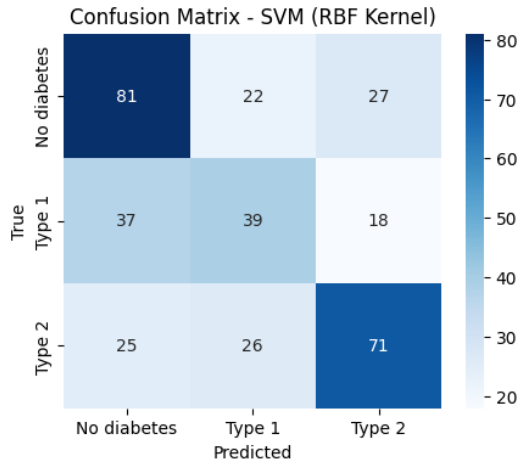
The AUC–ROC curves in Figure 9 and Figure 10 further confirm the enhanced discriminative capacity of the multimodal feature set. While class separation remained moderate under the structured-only condition, the inclusion of TF-IDF representations shifted all ROC curves closer to the optimal region, yielding macro-averaged AUC values above 0.90 for all models. XGBoost achieved a strong macro-averaged AUC of 0.9587, while Random Forest attained the highest overall macro AUC at 0.9644.

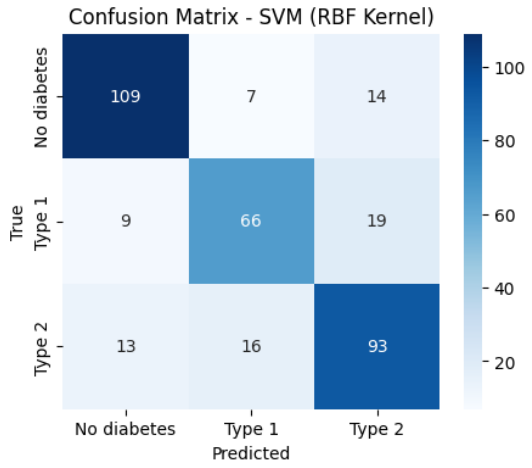
Confusion matrix (Figure 11) reveal that the primary source of error in the structured-only setting was the misclassification between Type 1 and Type 2 diabetes. This confusion was markedly reduced when textual information was incorporated, suggesting that medication history and symptom descriptions provide clinically specific cues that are not captured by vital signs and laboratory measurements alone. The root cause of the structured-feature performance ceiling is the physiological overlap between Type 1 and Type 2 presentations in the emergency setting: both groups can exhibit comparable heart-rate elevations, blood-glucose ranges, and respiratory profiles at triage, rendering numerical decision boundaries ambiguous. Textual fields resolve this ambiguity primarily through two mechanisms: (i) the chief-complaint field contains patient- and nurse-reported symptom descriptors (e.g., DKA, polydipsia) that are diagnostically specific; and (ii) the medication-reconciliation list reveals subtype-defining treatment histories, with insulin-only regimens characteristic of Type 1 and oral agents such as metformin characteristic of Type 2.

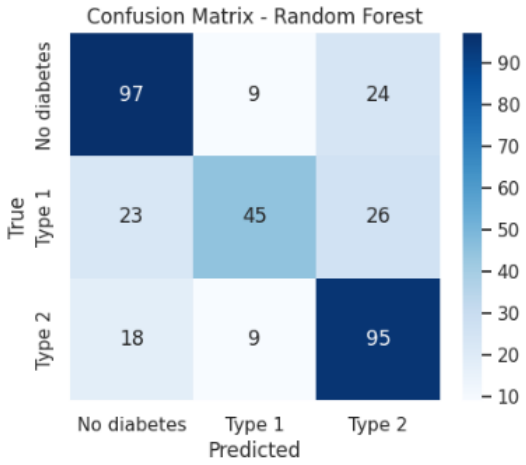
Overall, these findings demonstrate that the integration of unstructured clinical narratives significantly enhances multi-class diabetes classification performance, independent of the underlying classifier architecture.

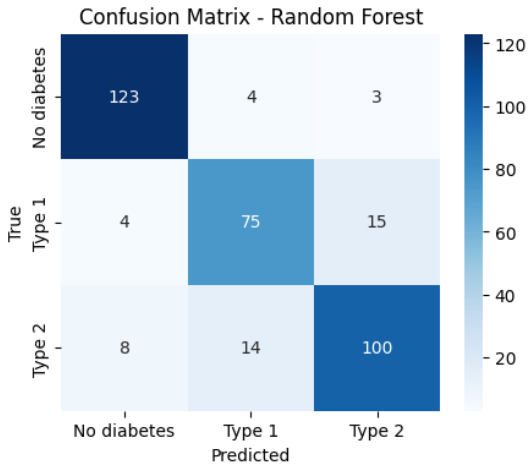
5. Conclusion and future works

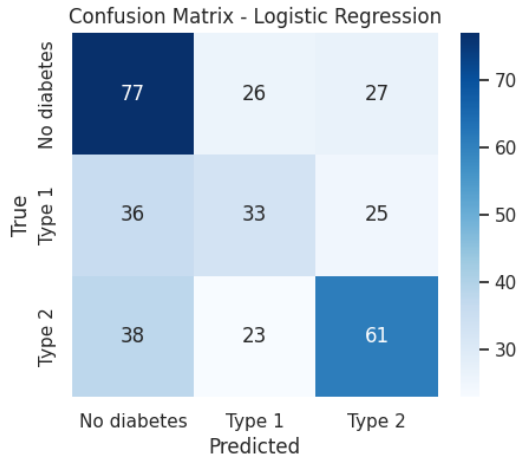
This research conducted a controlled empirical comparison of five classical machine learning classifiers for three-way diabetes triage (None / Type 1 / Type 2) applied

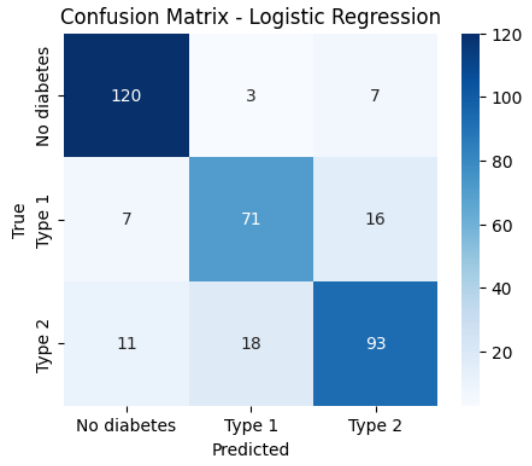

Confusion Matrix - SVM (RBF Kernel)
No diabetes
Type 1
Type 2
True
Predicted
81
22
27
37
39
18
25
26
71

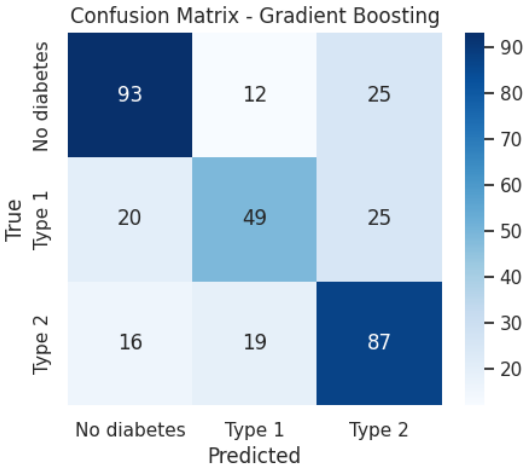

Confusion Matrix - SVM (RBF Kernel)
No diabetes
Type 1
Type 2
True
Predicted
109
7
14
9
66
19
13
16
93

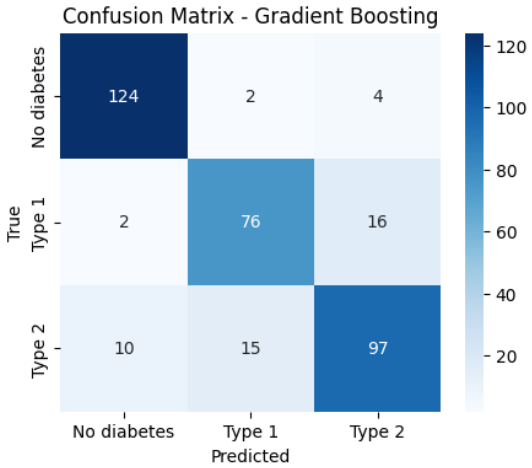

Confusion Matrix - Random Forest
No diabetes
Type 1
Type 2
True
Predicted
97
9
24
23
45
26
18
9
95


Confusion Matrix - Random Forest
No diabetes
Type 1
Type 2
True
Predicted
123
4
3
4
75
15
8
14
100


Confusion Matrix - Logistic Regression
No diabetes
Type 1
Type 2
True
Predicted
77
26
27
36
33
25
38
23
61


Confusion Matrix - Logistic Regression
No diabetes
Type 1
Type 2
True
Predicted
120
3
7
7
71
16
11
18
93


Confusion Matrix - Gradient Boosting
No diabetes
Type 1
Type 2
True
Predicted
93
12
25
20
49
25
16
19
87


Confusion Matrix - Gradient Boosting
No diabetes
Type 1
Type 2
True
Predicted
124
2
4
2
76
16
10
15
97

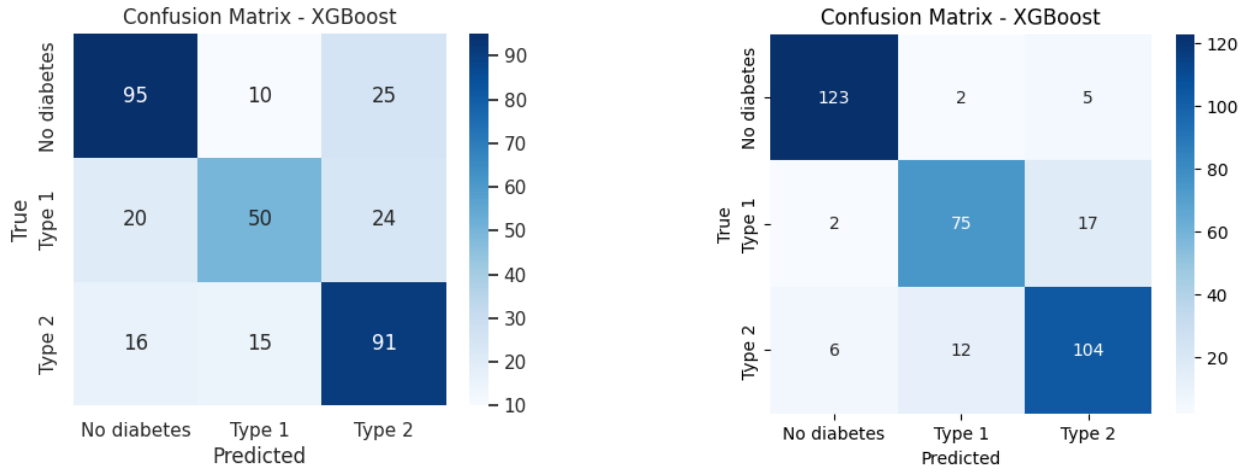


Figure 11. Confusion matrix under two feature settings: left, structured-only (no_text); right, multimodal (with_text). The multimodal setting reduces confusion between Type 1 and Type 2 diabetes.

to a balanced 6,000-encounter cohort derived from the MIMIC-IV inpatient and MIMIC-IV-ED emergency-department repositories. The experimental design deliberately isolated the incremental effect of clinical text by evaluating each classifier under two strictly comparable feature configurations: one restricted to structured vital-sign, demographic, and laboratory inputs, and one augmented with TF-IDF representations of chief complaints, BART-summarised discharge narratives, and medication-reconciliation strings.

Several limitations temper the generalisability of these findings. First, the cohort originates exclusively from a single academic medical centre (BIDMC), and performance on community hospitals, rural clinics, or non-English-language record systems is unknown. Second, the balanced sampling strategy, while beneficial for training, does not reflect the natural prevalence ratio of diabetes subtypes in the emergency population, and prospective deployment would require recalibration of decision thresholds. Third, the possible contribution of label-revealing terminology in medication text warrants further investigation through systematic per-field ablation experiments. Fourth, external validation on temporally or geographically held-out MIMIC partitions or on independent EHR systems is needed before clinical deployment can be responsibly considered.

Future research should pursue four principal directions. First, a rigorous feature-level ablation study should be conducted to isolate the contribution of each text field—chief complaint, discharge summary, and medication list—thereby quantifying the extent of genuine semantic inference versus incidental label correlation. Second, the feature set should be expanded to encompass laboratory markers with established pathophysiological relevance (HbA1c, C-peptide, GAD antibodies) and structured imaging outputs (e.g., retinal photographs for diabetic retinopathy staging). Third, the ensemble paradigm should be extended through stacked generalisation, combining the five classifiers evaluated here within a meta-learner, and through the inclusion of LightGBM—whose efficiency and accuracy on large, wide feature matrices make it a natural candidate for this task. Fourth, the model should be validated across multi-institutional EHR systems, including non-English records via multilingual NLP pipelines, to verify its universality and calibrate its suitability for diverse population characteristics. Achieving these goals will bring NLP-augmented, multi-class diabetes triage systems meaningfully closer to routine clinical deployment.

6. Declarations

6.1. Author Contributions

Huy Huynh: Conceptualization, Methodology, Formal analysis, Writing - Original Draft; **Thanh Cao:** Validation, Investigation, Writing - Review & Editing; **Hai Tran:** Supervision, Writing - Review & Editing.

6.2. Institutional Review Board Statement

Not applicable.

6.3. Informed Consent Statement

Not applicable.

6.4. Data Availability Statement

The data analyzed in this study was obtained from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database. Access to this database is restricted to credentialed researchers who complete required human subjects training and sign a data use agreement via PhysioNet.

6.5. Acknowledgment

The authors would like to acknowledge Saigon University (SGU) for the financial support and resource facilitation that made this research possible. Furthermore, we extend our gratitude to the Laboratory for Computational Physiology at the Massachusetts Institute of Technology (MIT) for developing and maintaining the MIMIC-IV database.

6.6. Conflicts of Interest

The author declares no conflicts of interest.

7. References

[1] L. Ismail, H. Materwala, and J. Al Kaabi, "Association of risk factors with type 2 diabetes: A systematic review," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 1759–1785, 2021. https://doi.org/10.1016/j.csbj.2021.03.003.

[2] M. Khalifa and M. Albadawy, "Artificial intelligence for diabetes: Enhancing prevention, diagnosis, and effective management," *Computer Methods and Programs in Biomedicine Update*, vol. 5, Art. no. 100141, 2024. https://doi.org/10.1016/j.cmpbup.2024.100141.

[3] N. Orangi-Fard, "Prediction of COPD using machine learning, clinical summary notes, and vital signs," *arXiv preprint arXiv:2408.13958*, 2024. https://doi.org/10.48550/arXiv.2408.13958.

[4] L. Fregoso-Aparicio, J. Noguez, L. Montesinos, and J. A. García-García, "Machine learning and deep learning predictive models for type 2 diabetes: A systematic review," *Diabetology & Metabolic Syndrome*, vol. 13, no. 1, 2021. https://doi.org/10.1186/s13098-021-00767-9.

[5] J. 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, Art. no. 1623, 2024. https://doi.org/10.3390/diagnostics14151623.

[6] A. 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, Art. no. 1065, 2024. https://doi.org/10.3390/bioengineering11111065.

[7] T. L. Hu, C. M. Chao, C. C. Wu, T. N. Chien, and C. Li, "Machine learning-based predictions of mortality and readmission in type 2 diabetes patients in the ICU," *Applied Sciences*, vol. 14, no. 18, Art. no. 8443, 2024. https://doi.org/10.3390/app14188443.

[8] F. Rustam *et al.*, "Enhanced detection of diabetes mellitus using novel ensemble feature engineering approach and machine learning model," *Scientific Reports*, vol. 14, no. 1, Art. no. 23274, 2024. https://doi.org/10.1038/s41598-024-74357-w.

[9] P. S. 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. https://www.researchgate.net/publication/382500931.

[10] B. Alkalifah, M. T. Shaheen, J. Alotibi, T. Alsubait, and H. Alhakami, "Evaluation of machine learning-based regression techniques for prediction of diabetes level fluctuations," *Heliyon*, vol. 11, no. 1, 2025. https://doi.org/10.1016/j.heliyon.2024.e41199.

[11] P. 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," in *AIP Conference Proceedings*, vol. 2853, no. 1, Art. no. 020140, May 2024. https://doi.org/10.1063/5.0203746.

[12] S. Das, S. P. Nayak, B. Sahoo, and S. C. Nayak, "Machine learning in healthcare analytics: A state-of-the-art review," *Archives of Computational Methods in Engineering*, vol. 31, pp. 3923–3962, 2024. https://doi.org/10.1007/s11831-024-10098-3.

[13] S. Prasher and L. Nelson, "Early prediction of obesity risk in older adults using XGBoost classifier," in *Proc. 2024 7th Int. Conf. Circuit Power and Computing Technologies (ICCPCT)*, 2024, pp. 1599–1603. https://doi.org/10.1109/ICCPCT61902.2024.10673336.

[14] P. Jain, R. Gupta, A. Joshi, and A. Kuzmin, "Enhanced cardiovascular diagnostics using wearable ECG and bioimpedance monitoring with LightGBM classifier," *Biosensors and Bioelectronics: X*, vol. 24, Art. no. 100617, 2025. https://doi.org/10.1016/j.biosx.2025.100617.

[15] A. Aldaej, T. A. Ahanger, and I. Ullah, "Deep neural network-based secure healthcare framework," *Neural Computing and Applications*, vol. 36, no. 28, pp. 17467–17482, 2024. https://doi.org/10.1007/s00521-024-10039-y.

[16] N. 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, Art. no. 1401803, 2024. https://doi.org/10.3389/fbioe.2024.1401803.

[17] M. Carletti *et al.*, "Multimodal AI correlates of glucose spikes in people with normal glucose regulation, pre-diabetes and type 2 diabetes," *Nature Medicine*, vol. 31, no. 9, pp. 3121–3127, 2025. https://doi.org/10.1038/s41591-025-03849-7.

[18] A. Johnson *et al.*, "MIMIC-IV (version 1.0)," *PhysioNet*, 2021. https://doi.org/10.13026/s6n6-xd98.

[19] A. Johnson *et al.*, "MIMIC-IV-ED (version 2.2)," *PhysioNet*, 2023. https://doi.org/10.13026/5ntk-km72.

[20] S. Sathyanarayanan and B. R. Tantri, "Confusion matrix-based performance evaluation metrics," *African Journal of Biomedical Research*, vol. 27, no. 4S, pp. 4023–4031, 2024. https://doi.org/10.53555/AJBR.v27i4S.4345.

[21] L. O. Joel, W. Doorsamy, and B. S. Paul, "On the performance of imputation techniques for missing values on healthcare datasets," *arXiv preprint arXiv:2403.14687*, 2024. https://doi.org/10.48550/arXiv.2403.14687.

[22] Z. S. Priyambudi and Y. S. Nugroho, "Which algorithm is better? An implementation of normalization to predict student performance," in *AIP Conference Proceedings*, vol. 2926, no. 1, Art. no. 020110, Jan. 2024. https://doi.org/10.1063/5.0182879.

[23] Y. Sun *et al.*, "Modifying the one-hot encoding technique can enhance the adversarial robustness of visual models for symbol recognition," *Expert Systems with Applications*, vol. 250, Art. no. 123751, 2024. https://doi.org/10.1016/j.eswa.2024.123751.

[24] Z. Labd, S. Bahassine, K. Housni, F. Z. A. H. Aadi, and K. Benabbes, "Text classification supervised algorithms with term frequency–inverse document frequency and global vectors for word representation: A comparative study," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 1, 2024. https://doi.org/10.11591/ijece.v14i1.pp589-599.

[25] N. K. Sahu *et al.*, "Leveraging language models for summarizing mental state examinations: A comprehensive evaluation and dataset release," in *Proc. 31st Int. Conf. Computational Linguistics*, 2025, pp. 2658–2682. https://aclanthology.org/2025.coling-main.182.pdf.

[26] J. Wu, H. Wang, C. Ni, C. Zhang, and W. Lu, "Data pipeline training: Integrating AutoML to optimize the data flow of machine learning models," *arXiv preprint arXiv:2402.12916*, 2024. https://doi.org/10.48550/arXiv.2402.12916.