Predictive Modeling for Diabetes and Risk of Associated Chronic Heart Disease: A Comparative Evaluation of Machine Learning Techniques

Golam Md Mohiuddin1, a) Md Shohel Sayeed2, b) and Yoon Thiri Ko3, c)

1Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia.

2Centre for Intelligent Cloud Computing, CoE for Advanced Cloud, Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia.

3Faculty of Science, King Mongkut's Institute of Technology Ladkrabang (KMITL), Bangkok 10520, Thailand.

b) Corresponding author: shohel.sayeed@mmu.edu.my

a)GOLAM.MD.MOHIUDDIN@student.mmu.edu.my

*c) 67056004@kmitl.ac.th*

**Abstract.** Diabetes Mellitus is a significant global health concern, contributing to the development of chronic diseases such as heart disease. Individuals with diabetes face a significantly higher risk of developing heart disease, with studies showing they are approximately 2 to 3 times more likely to experience cardiovascular conditions compared to non-diabetic individuals and nearly one in three diabetic patients develop heart disease. However, traditional diagnostic procedures frequently struggle to provide timely and accurate risk predictions, necessitating machine learning based solutions for early detection. This study leverages machine learning models to predict the risk of heart disease in diabetes patients. Machine learning approaches have been shown to be highly effective at improving early detection and risk assessment. This study aims to predict and analyze the risk of heart disease in diabetes patients using machine learning algorithms. The method uniquely demonstrates how machine learning can be used in order to predict risk in addition provide insightful information on personalized healthcare management by integrating data-driven feature selection with advanced model tuning in an innovative approach to improve the accuracy of predictions. Logistic Regression, Random Forest, Support Vector Machines, K-Nearest Neighbours (KNN), XGBoost, LightGBM and CatBoost were the seven predictive models used for training and evaluation. Among them, CatBoost and LightGBM consistently delivered the most robust performance, achieving F1 scores up to 0.74 in diabetes prediction and 0.69 in heart disease risk assessment, demonstrating their strong potential for supporting early diagnosis and personalized healthcare interventions. This study implements a two-stage hierarchical classification system based on two publicly available and separate datasets diabetes and cardiovascular. While the two datasets are not connected at the individual level, their common elements and the clinical catheterization path identify them as appropriate for sequential modelling as typical comorbidities patterns.

# Introduction

Diabetes mellitus represents a significant global health concern, substantially reducing life expectancy and heightening the risk of developing serious chronic conditions, including heart disease, stroke, renal failure, and various forms of cancer. As reported by the World Health Organization (WHO), diabetes is among the primary contributors to global morbidity and mortality, largely due to its association with severe complications such as vision loss, kidney dysfunction, and heart-related disorders [1]. The global surge in Type 2 Diabetes Mellitus (T2DM) cases is largely attributed to unhealthy diets, sedentary lifestyles, and genetic predisposition [2]. Among its many complications, heart disease is the most prevalent, with diabetic individuals being two to four times more likely to develop cardiovascular conditions compared to non-diabetics [3]. Mechanisms such as endothelial dysfunction, chronic inflammation, and metabolic abnormalities drive the connection between diabetes and heart disease [4]. Therefore, conventional diagnostic approaches often lack the sensitivity and timing required for early risk detection, leading to delayed interventions.

To overcome such constraints, machine learning (ML) promises the early prediction of diabetes complications. ML algorithms like Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and Neural Networks have been shown to be proficient in processing large-scale health data and identifying hidden risk patterns [5]. Empirical confirmation exists to indicate that these models are more accurate and early in their identification compared to conventional statistical methods. Supervised learning algorithms such as Decision Trees and K-Nearest Neighbors (KNN) can successfully classify patients based on key risk factors including Body Mass Index (BMI), blood pressure, insulin status, and genetic predispositions [6]. In addition, ML helps create individual patient risk profiles so that doctors can apply preventive care strategies more specifically [7]. Still, there are concerns about understanding the models, data quality, applying it on a large scale, and ethical questions about data confidentiality. Despite these concerns, ML has tremendous potential in diabetic patient risk assessment improvement and advancing personalized medicine. Previous studies have modeled prediction of heart disease and diabetes separately. However, few have established the risk of heart disease as predicated on diabetes risk. This research addresses that gap by proposing a two-stage machine learning model. This method reflects clinical realities in which diabetes risk increases the chance of heart problems.

# RELATED WORK

The use of machine learning has also enhanced disease prediction, specifically diabetes and related complications like heart disease. Conventional methods like statistical modeling and clinical screening are generally not precise or timely enough for proper risk assessment. In contrast, machine learning is based on big data to uncover hidden patterns, which allows for earlier and more precise identification. The literature has spoken of a number of ML models, including Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), and XGBoost, and their performance in predicting diabetes and cardiovascular disease development. Ensemble methods RF and XGBoost have been especially outstanding, with over 85% accuracy, due to their capability in handling non-linear relationships and variable interaction [8]. Combining clinical variables such as insulin resistance, BMI, and blood glucose with machine learning algorithms has enhanced individualized risk prediction [9].

Data preprocessing and feature extraction have become paramount in optimizing the predictive accuracy. Principal Component Analysis (PCA) has been used to reduce dataset dimensionality without affecting model performance, and research concentrates on the aspect of structured data that is normalized yielding a far better model accuracy compared to raw datasets [10]. Deep learning techniques have also gained much popularity. For instance, Convolutional Neural Networks (CNNs) reported an AUC of 0.92 in diabetic retinopathy diagnosis from retina images, performing better than traditional ML classifiers [3]. Similarly, Deep Neural Networks (DNNs) in electronic health records (EHRs) accurately predicted the onset of diabetes with applicability for clinical decision-making support [11]. However, deep learning models are generally claimed to lack proper interpretability, which hinders clinical acceptance and use. Comparative research indicates RF and SVM possess higher sensitivity and specificity than LR for heart disease prediction in diabetic patients, with RF having an AUC of 0.87 [5]. Long Short-Term Memory (LSTM) networks, capable of analyzing temporal health records, also demonstrated high predictive power, though they demand significant computational resources, limiting their everyday use [12].

Despite promising advancements, ML-based models face challenges including interpretability, generalizability, and ethical concerns. Explainable AI is essential to increase transparency and foster trust among healthcare professionals. Additionally, models trained on homogeneous datasets often underperform on external populations, stressing the importance of using diverse and representative data [7]. Privacy and fairness must also be prioritized to ensure responsible use of patient data. However, while ML shows strong potential for early detection and personalized risk segmentation of diabetes related heart disease, future research must focus on model transparency, scalability, and ethical robustness for successful clinical implementation. Emerging research in comorbidity-aware machine learning has emphasized the importance of modeling health conditions sequentially. However, most existing studies lack an integrated architecture for conditional disease risk modeling. Our study extends this work by introducing a hierarchical classification pipeline using diabetes status as an upstream predictor for heart disease risk.

# Methodology

This section outlines the systematic steps undertaken to build and evaluate predictive models for diabetes and associated chronic heart disease risk. The methodology consists of several key stages: data collection, preprocessing, feature engineering, model training, evaluation, and integration of prediction models. The system process flowchart we have used in this paper are as shown (see Figure 1).

## Data Collection

This research utilized two publicly available datasets obtained from Kaggle to develop and evaluate predictive models for diabetes and heart disease risk: (i) the Diabetes Health Indicators Dataset and (ii) the Cardiovascular Disease Dataset, both widely adopted in machine learning research due to their clinical relevance, feature diversity, and real-world applicability. The diabetes dataset, diabetes\_binary\_5050split\_health\_indicators\_BRFSS2015.csv, is derived from the U.S. CDC’s BRFSS 2015 survey and comprises 70,692 samples evenly split between diabetic and non-diabetic cases, with 21 features including demographics, lifestyle, and clinical metrics ideal for balanced classification tasks [13]. The heart disease dataset includes 70,000 structured medical records with 11 features such as age, cholesterol, glucose, smoking status, and BMI, targeting binary cardiovascular disease outcomes [14]. These datasets were used both independently and in a hybrid model to evaluate the compounded risk of heart disease in diabetic individuals. Their size, accessibility, and clinical precision offer a robust foundation for predictive modeling and comparative algorithm assessment in healthcare applications. It is important to note that the diabetes and heart disease datasets were not originally collected from the same population. Thus, the study does not attempt a literal patient-level fusion. Instead, a logical integration is performed, using common features (e.g., age, BMI, cholesterol, activity levels) to simulate a clinical inference process: first assess diabetes risk, then conditionally assess heart disease risk. This abstraction aligns with clinical practice and maintains the statistical integrity of both datasets.

**FIGURE 1.** System Process flowchart

## Data Cleaning & Preprocessing

Prior to model development, both datasets underwent systematic data cleaning and preprocessing to enhance data quality and ensure consistency for machine learning algorithms. Observations with missing or null values were removed using listwise deletion dropna(), maintaining the integrity of the observed relationships and minimizing imputation bias. Duplicate entries were checked and removed.

In the data preprocessing phase, numerical variables were standardized using z-score normalization via Scikit-learn’s StandardScaler to ensure consistent scaling across continuous features, which is particularly important for distance-based algorithms such as K-Nearest Neighbors and gradient-based models like Logistic Regression. Categorical variables including classifications of blood pressure, BMI, and age were transformed through one-hot encoding, with the first category omitted to avoid multicollinearity. To mitigate class imbalance within the training dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. Although the diabetes dataset appeared balanced, SMOTE was retained to enhance generalization and reduce overfitting, particularly for tree-based classifiers [15]. SMOTE's effectiveness in improving classification performance in imbalanced datasets is well documented [16]. A unified preprocessing pipeline was applied to both datasets to ensure methodological consistency across models, minimizing potential biases in performance comparison.

## Feature Engineering

Feature engineering was essential to harmonize the diabetes and heart disease datasets, which, despite sharing risk factors like age, BMI, blood pressure, and cholesterol, differed in feature representation. To ensure consistency and enable reliable comparative modeling, features were systematically aligned and transformed. Age in the heart disease dataset, initially recorded in days, was converted to years and categorized into three demographic groups: Young (<30), Middle-aged (30–60), and Senior (>60), a stratification commonly applied in epidemiological studies. BMI, while continuous in both datasets, was computed from weight and height in the heart disease data using the standard formula as depicted in Equation (1).

(1)

Then the calculated BMI is categorized as "Underweight," "Normal," "Overweight," and "Obese" based on WHO guidelines. Blood pressure required harmonization, with the heart disease dataset offering continuous systolic (ap\_hi) and diastolic (ap\_lo) readings, while the diabetes dataset used a binary HighBP variable. These were categorized using AHA guidelines into: Normal (<120/<80), Elevated (120–129/<80), Hypertension Stage 1 (130–139 or 80–89), and Stage 2 (≥140 or ≥90). The diabetes variable was mapped accordingly: diabetes\_df['BP\_Category'] = diabetes\_df['HighBP']. Map ({0: "Normal", 1: "Hypertension Stage 1"}). Cholesterol levels, categorized as 1 = Normal, 2 = Borderline High, and 3 = High in the heart disease dataset, and as 0 = Normal, 1 = High in the diabetes dataset, were simplified into two unified categories: "Normal" and "High". These transformations ensured consistent feature representation and structure across both datasets, enhancing the comparability and interpretability of the models. To reduce potential bias introduced by manual categorization (e.g., BMI and BP categories), all threshold values were derived from established clinical guidelines. Nonetheless, future versions of this work may use quantile-based binning or data-driven categorization to further minimize threshold bias.

Table 1 summarizes the transformations applied to datasets, comparing before and after features engineering, including categorization and new variables.

**TABLE 1.** Summary of feature representations before and after feature engineering

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **Heart Disease Dataset (Before)** | **Heart Disease Dataset (After)** | **Diabetes Dataset (Before)** | **Diabetes Dataset (After)** |
| Age | Age in days | Age in years, Age Group (Young, Middle-aged, Senior) | Age in years | Age in years, Age Group (Young, Middle-aged, Senior) |
| Blood Pressure | Systolic and Diastolic (continuous) | BP Category (Normal, Elevated, Hypertension) | HighBP (binary) | BP Category (Normal, Hypertension) |
| BMI | Continuous BMI value | BMI Category (Underweight, Normal, Overweight, Obese) | BMI (continuous) | BMI Category (Underweight, Normal, Overweight, Obese) |
| Cholesterol | Categorical (1, 2, 3) | Chol Category (Normal, Borderline High, High) | HighChol (binary) | Chol Category (Normal, High) |
| Age Group | Not present | Age Group (Young, Middle-aged, Senior) | Not present | Age Group (Young, Middle-aged, Senior) |
| BP Category | Not present | BP Category (Normal, Hypertension) | Not present | BP Category (Normal, Hypertension) |
| BMI Category | Not present | BMI Category (Underweight, Normal, Overweight, Obese) | Not present | BMI Category (Underweight, Normal, Overweight, Obese) |
| Chol Category | Not present | Chol Category (Normal, High) | Not present | Chol Category (Normal, High) |

## Data Splitting

To ensure effective model evaluation, the dataset was divided into 80% training and 20% testing sets using stratified sampling to preserve class distribution and address class imbalance. This approach maintains sufficient training data while ensuring an unbiased test subset. Stratified k-fold cross-validation was applied to reduce variance and improve metric reliability. Models were trained on the training set and independently evaluated on the testing set. Additionally, feature standardization using StandardScaler was conducted to optimize the performance of scaling-sensitive models like logistic regression and k-nearest neighbors, ensuring consistent feature ranges and enhancing convergence and accuracy.

## Model Selection & Training

In this study seven machine learning models are used, including Random Forest, Gradient Boosting, Logistic Regression, K-Nearest Neighbors, XGBoost, LightGBM and CatBoost to predict the diabetes and heart disease outcomes. The models were selected to provide an even comparison of performance on healthcare prediction tasks. Random Forest was chosen due to robustness and adaptation to highly non-linear medical data, as well as its effectiveness in avoiding overfitting. And finally, the high-accuracy Gradient Boosting for further predictions fine-tuning with error correction. To avoid overfitting, Logistic Regression is chosen due to its simplicity and interpretability in clinical use K-Nearest Neighbors performed well in smaller sets based on similarity-based classification. XGBoost was chosen because it is scalable and fast in structured clinical problems, LightGBM being the fastest, most accurate, and most efficient in handling large data sets. Finally, CatBoost was highlighted for its effective capability to process categorical features. The study aims to identify models that are well-balanced in predictive power, interpretability, and computationally efficient.

## Integrated Diabetes and Heart Disease Risk Prediction System

This system does not make the assumption of shared identity between individuals of the two datasets but uses a two-stage design inspired by disease progression and comorbid research. There is the first model to predict diabetes status; the second model is used for heart disease risk estimation with features harmonized across both data sets, thereby creating a conditional inference pathway. If an individual is predicted as diabetic, the system proceeds to the second step, where the model estimates the risk of heart disease based on cardiovascular and diabetic risk factors. This approach aligns with clinical findings that diabetes significantly increases the risk of heart disease. By addressing the two diseases in a sequential manner, the system presents a broad picture of risk that can enable early treatment and preventive care.

## Evaluation of Models

To check if the model is performing well, we used Accuracy, Precision, Recall, and F1 Score. These are standard metrics for healthcare classification issues. Accuracy tells us how accurate the predictions are overall, but it becomes confusing with imbalanced data. Precision attempts to minimize false positives, which avoids unnecessary treatment and stress for patients. Recall aims for correct identification of true positives, which is crucial to avoid false negatives. F1 Score is a blend of Precision and Recall, giving an even way to measure how well a model performs. Together, they make an overall testing system for evaluating how accurate health prediction models are [17],[18],[19]. AUC-ROC and PR-AUC were also calculated to deal with class imbalance and the significant influence of misclassification in health situations. AUC-ROC makes it clear the trade-off between true positives and false positives. PR-AUC is more informative for the diabetic group since there's a smaller number of positive cases.

# Result and Discussion

The performance of the predictive models in detecting diabetes and assessing heart disease risk was analyzed to use Accuracy, Precision, Recall, and F1-Score. Even with balanced class distributions, comprehensive evaluation was used to reveal the strengths and weaknesses of each model to ensure false positives and negatives were accounted for in clinical practice.

## Diabetes Prediction Model Performance

The models trained for diabetes prediction yielded generally consistent performance across all metrics. The highest accuracy was achieved by the CatBoost model, recording 72.9%, followed closely by LightGBM and Gradient Boosting. In terms of F1-score, which balances precision and recall, CatBoost again emerged as the best performing model with a score of 0.742, indicating a well-rounded ability to identify both diabetic and non-diabetic individuals.

Table 2 results show that tree-based ensemble methods excel at capturing non-linear patterns, while Logistic Regression performed well, indicating the dataset's suitability for linear classification.

**TABLE 2.** Diabetes prediction results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| Random Forest | 0.725 | 0.704 | 0.777 | 0.739 |
| Gradient Boosting | 0.728 | 0.709 | 0.773 | 0.740 |
| Logistic Regression | 0.726 | 0.714 | 0.754 | 0.734 |
| K-Nearest Neighbors | 0.700 | 0.717 | 0.661 | 0.688 |
| XGBoost | 0.725 | 0.708 | 0.766 | 0.736 |
| LightGBM | 0.729 | 0.710 | 0.775 | 0.739 |
| CatBoost | 0.729 | 0.710 | 0.777 | 0.742 |

## Heart Disease Risk Prediction Model Performance

Heart disease prediction models were evaluated using the same metrics. The best performing model in terms of F1-score was LightGBM, scoring 0.691, closely followed by Gradient Boosting and CatBoost. Although in Table 3 the absolute performance was slightly lower compared to the diabetes models, the models still demonstrated reasonable classification power.

**TABLE 3.** Heart disease risk prediction results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| Random Forest | 0.702 | 0.721 | 0.658 | 0.688 |
| Gradient Boosting | 0.704 | 0.723 | 0.661 | 0.690 |
| Logistic Regression | 0.701 | 0.728 | 0.641 | 0.682 |
| K-Nearest Neighbors | 0.678 | 0.705 | 0.613 | 0.656 |
| XGBoost | 0.701 | 0.715 | 0.667 | 0.690 |
| LightGBM | 0.702 | 0.716 | 0.668 | 0.691 |
| CatBoost | 0.702 | 0.720 | 0.661 | 0.689 |

While the predictive performance for heart disease was slightly lower, this is consistent with the increased complexity of heart risk prediction, which is influenced by a wider range of interacting physiological and behavioral factors. Nonetheless, the models achieved respectable recall scores, suggesting that they are generally sensitive to high-risk patients.

## Interpretation of Evaluation Metrics

Multiple evaluation metrics were essential in this study, as accuracy alone does not capture the impact of misclassifications in healthcare. Precision, recall, and F1-score were used to assess the models' ability to minimize false positives and false negatives, critical in medical applications. Tree-based ensemble methods, such as CatBoost, LightGBM, and XGBoost, showed strong performance across tasks, excelling in handling non-linear relationships and categorical data. While no model exceeded 80% performance, the results highlight the potential of machine learning for early diagnosis and risk assessment. Future improvements could involve incorporating longitudinal health data, feature expansion, and ensemble stacking to enhance performance.

## Limitations of Dataset Integration

A key limitation of this study is the use of two distinct datasets with no patient-level correspondence. While this prevents direct comorbidity modeling at the individual level, the two-stage classification framework reflects a plausible clinical inference process, where diabetes screening informs cardiovascular risk management. The results should be interpreted in this conceptual context. Future work will aim to validate this model using longitudinal, linked patient records.

# CONCLUSION

This study introduced a two-stage Integrated Diabetes and Heart Disease Risk Prediction System employing machine learning models to evaluate an individual's risk of developing diabetes and, if applicable, heart disease. By leveraging the known correlation between these two diseases, the system provides a more holistic health risk assessment. Models like CatBoost and LightGBM were used due to their efficiency with tabular clinical data, producing F1-scores between 68% and 74%. While the results show moderate predictive accuracy, they are valuable in the absence of detailed patient data.

However, the study has limitations, including restricted feature sets that do not consider factors like genetics, medications, or diet, and reliance on predefined thresholds for feature engineering. Additionally, the system lacks external validation on diverse clinical datasets, limiting its generalizability. Future work should aim to enhance the dataset by including more comprehensive clinical variables, such as lab results, family history, and lifestyle factors. Further directions include exploring advanced models like deep learning and enhancing model interpretability for medical AI applications. Moreover, validating the system with external datasets from various population groups is crucial for broader applicability. This integrated system represents a promising approach to personalized preventive care, with potential applications in early diagnosis and chronic disease management.

# References

1. H. Sang, H. Lee, M. Lee, J. Park, S. Kim, H. G. Woo, M. Rahmati, A. Koyanagi, L. Smith, S. Lee, Y.-C. Hwang, T. S. Park, H. Lim, D. K. Yon, and S. Y. Rhee, “Prediction model for cardiovascular disease in patients with diabetes using machine learning derived and validated in two independent Korean cohorts,” Scientific Reports **14**(1), 2024.
2. World Health Day 2024. (2024). Who.int. https://www.who.int/campaigns/world-health-day/2024
3. J. Chaki, S. Thillai Ganesh, S.K. Cidham, and S. Ananda Theertan, “Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review,” Journal of King Saud University - Computer and Information Sciences **34**(6), 3204–3225 (2022).
4. H. Zhang, and H. Shi, “Construction of a prediction model for coronary heart disease in type 2 diabetes mellitus: A cross-sectional study,” Scientific Reports **15**(1), 7003 (2025).
5. 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,” Diabetol Metab Syndr **13**(1), 148 (2021).
6. E.K. Oikonomou, and R. Khera, “Machine learning in precision diabetes care and cardiovascular risk prediction,” Cardiovasc Diabetol **22**(1), 259 (2023).
7. F. Y. Osisanwo et al., “Supervised machine learning algorithms: Classification and comparison,” International Journal of Computer Trends and Technology **48**(3), 128–138 (2017).
8. B. Singh, S. Indu, and S. Majumdar, “Comparison of machine learning algorithms for classification of Big Data sets,” Theoretical Computer Science **1024**, 114938 (2025).
9. A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, “A data-driven approach to predicting diabetes and cardiovascular disease with machine learning,” BMC Medical Informatics and Decision Making **19**(1), 2019.
10. R. Singh, et al., “Data-driven approaches for diabetes risk stratification: A review of machine learning techniques,” Journal of Biomedical Informatics **141**, 104385 (2024).
11. H. Pang, L. Zhou, Y. Dong, P. Chen, D. Gu, T. Lyu, and H. Zhang, “Electronic Health Records-Based Data-Driven Diabetes Knowledge Unveiling and Risk Prognosis,” IECE Transactions on Intelligent Systematics **2**(1), 1–13 (2024).
12. T.K. Revathi, S. Balasubramaniam, V. Sureshkumar, and S. Dhanasekaran, “An Improved Long Short-Term Memory Algorithm for Cardiovascular Disease Prediction,” Diagnostics **14**(3), 239 (2024).
13. Teboul, A. (2018). *Diabetes Health Indicators Dataset*. Kaggle.com. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
14. Ulianova, S. (2019). *Cardiovascular Disease dataset*. Kaggle.com. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset
15. M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, “Handling imbalanced medical datasets: review of a decade of research,” Artif Intell Rev **57**(10), 273 (2024).
16. J. Kamwele Mutinda, A. Langat, R.K. Marcel Djaha, J. Ndoto Munyao, L. Whitaker, and M. Auma Omondi, “Enhancing Obesity Detection Through SMOTE -based Classification Models: A comparative Study,” Interdiscip J Epidemiol Pub Health **7**(1), (2024).
17. L. Ferrer. Analysis and comparison of classification metrics. *arXiv preprint arXiv:2209.05355* (2002).
18. C. M. Chituru, S.-B. Ho, and I. Chai, “Diabetes risk prediction using Shapley additive explanations for feature engineering,” *Journal of Informatics and Web Engineering* **4**(2), 18–35 (2025), doi:10.33093/jiwe.2025.4.2.2.
19. J. Jayapradha, Y. Kulkarni, L. Vadhanie G, P. Naveen, and E. Abdulwahab Anaam, “Treatment recommendation using BERT personalization,” *Journal of Informatics and Web Engineering* **3**(3), 41–62 (2024), doi:10.33093/jiwe.2024.3.3.3.