Abstract

Diabetes is a chronic endocrine disorder characterized by elevated blood glucose levels, leading to severe complications. Its global prevalence has risen sharply over the past decade, with over 570 million cases projected by 2025. Concurrently, advancements in machine learning have enabled new possibilities for disease prediction and classification.

This study aims to develop a machine learning model capable of predicting both the presence and type of diabetes using an integrated dataset.

To address the limitations of existing models—particularly their inability to differentiate between diabetes types due to limited data—a compounded dataset was created by merging two publicly available sources : one a real-world datset, and one a synthetic dataset. The model performs binary classification to detect diabetes presence and multiclass classification to identify type, including Prediabetes, Type 1, Type 2, and Pancreatogenic Diabetes. Algorithms implemented include Naive Bayes, K-Nearest Neighbours, Logistic Regression, Random Forest, and XGBoost.

XGBoost achieved the highest performance among all models. Its feature importance scores were analyzed to validate predictive accuracy and to identify key factors distinguishing diabetes types.

The proposed model demonstrates potential for effective diabetes screening and classification using routinely collected clinical data, supporting early diagnosis and personalized treatment planning.

1 Introduction

Diabetes mellitus, commonly known as diabetes, is an umbrella term used to refer to a group of endocrine diseases.

It is a chronic condition where the body does not produce enough insulin, or cannot effectively use the insulin it produces, leading to high blood sugar levels [1]. There are multiple factors that can cause diabetes, such as pancreatic cancer, pancreatitis, genetic defects, and surgery. Apart from these medical factors, unhealthy dietary patterns, socio-economic development, and sedentary lifestyles have been identified as determinants that are driving an increase in prevelance of diabetes [2].

It is estimated that 537 million people in the age range of 20 to 79 are affected by diabetes, and this number will grow to 643 million by 2045. A study by Kumar et al. (2021) concluded that incidence of diabetes in India would grow from 9.6% in 2021 to 10.9% in 2045 [3].

Resistance to, or an insufficient amount of insulin causes suboptimal conversion of food to energy, resulting in increased hunger levels, called as polyphagia. Further, the kidney utilises more water to filter the excessive glucose in the bloodstream causing abnormal urination frequency (polyuria) and excessive thirst (polydipsia).

The World Health Organization classifies diabetes into 6 categories. Among these are Type 1 Diabetes and Type 2 Diabetes have the highest prevelance. A Prediabetic state has been identified, as an early stage of diabetes. Further, diabetic states that occur as a consequence of pancreatic diseases have been grouped under Pancreatogenic Diabetes, also termed as Type 3c diabetes.

Type 1 Diabetes is typically a consequence of autoimmune destruction, causing the pancreatic β cells to stop producing insulin. Due to its autoimmune nature and by extension, genetic predisposition, it is likely to occur at a younger age compared to other types of diabetes. People with this type of diabetes are at a higher risk of developing other autoimmune disorders. They require external doses of insulin for survival [4].

On the other hand, Type 2 Diabetes is characterised by insulin resistance and a progressive lack of insulin. Initially, the pancreas compensates by producing more insulin, but over time, it becomes unable to keep up with the demand, causing the blood sugar to rise. Although some people are more genetically prone, it also heavily depends on lifestyle factors, like lack of exercise and obesity. It accounts for nearly 90% of all diabetes cases. Type 2 Diabetes may lead to severe complications, such as cardiovascular diseases, kidney damage, nerve damage and retinopathy.

Prediabetes is a considered a precursor to Type 2 Diabetes, where the blood sugar level is high, but not high enough to be classified into the latter. People in this stage show many of the common symptoms of diabetes like polyuria and polyphagia. A study by Schlesinger et al. showed that prediabetes markedly increased the risk for incidence of cardiovascular, renal, hepatic failurea, as well as the risk of cancer and dementia [5]

Pancreatogenic diabetes is the most common after Type 2 Diabetes. Patients with acute pancreatitis are at a 34.5% risk of developing diabetes. [6] A study by Shivaprasad et al. concluded that the mortality and morbidity is higher for Type 3c Diabetes than Type 2 Diabetes [7].

The recent rise in artificial intelligence has had an impact on the healthcare field, with analytical and machine learning models processing the available health records in order to predict patient outcomes and diagnose patients more efficiently. It has accelerated the efficiency and accuracy of diagnosis by enabling data-driven decisions with greater confidence. AI systems can analyze vast amounts of medical data—such as electronic health records (EHRs), medical imaging, and genomics, allowing for early detection of conditions like cancer, heart disease, and neurological disorders, often at stages when treatments are more effective.

However, the healthcare field poses its own set of challenges to Artificial Intelligence. Availability of credible datasets, especially in economically challenged areas is scarce. This might inadvertently lead to a model that is biased due the prevalence of data from

socio-ethnic groups with higher accesibility to healthcare facilities. Data privacy is yet another concern, and the abstract nature of many machine learning models raises questions about their internal working and their reliability.

Various machine learning models have been employed for diabetes prediction. K-Nearest Neighbours (KNN) has been used due to its effectiveness in handling non-linear data, achieving moderate accuracy in binary classification tasks. Naive Bayes uses a probabilistic approach, but it often struggles with feature dependencies. XGBoost has emerged as the leading algorithm, utilizing gradient-boosted trees to achieve high accuracy and robust performance on imbalanced datasets. Random Forest is yet another important algorithm, providing interpretability through feature importance. Despite these advancements, most models focus on binary classification (diabetes vs. non-diabetes) and fail to differentiate between diabetes types, such as Type 1, Type 2, Prediabetes, and Pancreatogenic Diabetes. This limitation underscores the need for more comprehensive approaches, such as the one proposed in this study.

This limitation stems from the lack of datasets that include sufficient data points for type-specific classification. To bridge this gap, this study proposes a novel two-stage approach : An initial binary classification using the Pima Indians dataset, and a subsequent multiclass classification using a secondary, programmatically generated dataset to distinguish between Type 1 Diabetes, Type 2 Diabetes, Prediabetes, and Pancreatogenic Diabetes. By developing the model in this manner, we address the critical limitation of existing approaches, which lack the granularity to differentiate between diabetes subtypes, thereby enabling more precise predictions.

2 Literature Review

Many attempts have been made to successfully integrate ML techniques for detection of diabetes, and several different techniques have been used for the same [8].

KNN predicts the class of a data point by looking at the classes of its nearest neighbours in the feature space. It is a simple, non-parametric method that relies totally on the data's structure. While it gives a straightforward interpretation, it is sensitive to the choice of distance metric and the number of neighbors. Logistic Regression predicts the probability of a binary outcome using a sigmoid function. It learns the relationship between input features and the log-odds of the target variable, making linear classifier. However, it assumes a linear relationship between features and the target, which can limit its performance on complex datasets. Decision Trees split the data into regions based on feature values, aiming to maximize the purity of each region. They are intuitive and easy to interpret, but they tend to overfit the training data, especially when grown too deep. To counter this, a Random Forest classifier can be used, that combines multiple decision trees, each trained on a random subset of the data and features. The final prediction is made by averaging or taking a majority vote from all trees. This ensemble approach makes Random Forest robust and reliable for classification tasks.

Abdulhadi and Al-Mousa obtained an accuracy of 82% using the Random Forest Classifier [9]. Ghosh et.al obtained an accuracy of 99.35% using Random Forest Classifier [10]. Thotad et.al also concluded that Random Forest outperformed other models in detection on Indian demographic dataset [11].

It has been shown that ensemble techniques perform better than logistic regression, due to the fact that LR sufferrs from assuming linear relationships and is much more sensitive to outliers. One exception is the study by Khaleel and Al-Bakry, where Logistic Regression showed an accuracy of 94%, above that of Naive Bayes 79% and KNN 69% [12]. Rajendra and Latifi concluded that the ensemble techinques Max Voting and Stacking fare better than Logistic Regression, with the latter having an accuracy 93.04% of to the former's 74.03 when trained without feature selection [13]. Seto et al. provided evidence that Gradient Boosted Decision Trees (GBDT) and logistic regression models perform somewhat equally upto training data size of 10^4, GBDT shows a significant increase after this number, while LR the model's metrics become saturated [14].

Salem et al. utilised the basic K - nearest neighbours and two of its variants which allow partial membership in multiple classes, Fuzzy KNN nand TFKNN. TFKNN turned out to be the best performer with 94.13% specific accuracy [15]

While instance-based algorithms like KNN and its fuzzy variants excel in modeling local data relationships, they tend to struggle with high-dimensional datasets and are computationally expensive during inference. These limitations have led researchers to increasingly explore ensemble methods and boosting techniques, which offer a better balance between predictive power and computational efficiency. In particular, Gradient Boosting frameworks such as XGBoost have emerged as state-of-the-art due to their ability to model complex, non-linear interactions while mitigating overfitting through regularization.

Maulana et.al compared siz other popular machine learning models with XGBoost, and concluded its superiority in diabetes predictive capabilities, and notes that it achieves the achieves the highest performance in accuracy, precision, sensitivity, and F1-score [16].

Despite the abundance of research in binary diabetes classification, the task of differentiating between diabetes subtypes remains largely underexplored. The majority of publicly available datasets, including the widely-used Pima Indians dataset, lack subtype annotations, making it challenging for researchers to develop models that can distinguish between conditions like Type 1 Diabetes, Type 2 Diabetes, Prediabetes, and Pancreatogenic Diabetes. As a result, current models often provide a binary outcome that fails to capture the nuanced differences between these clini-

cally distinct subtypes, limiting their real-world applicability in personalized treatment and early intervention strategies.

Uddin et.al notes the necessity for a balanced dataset to aid in the reduction of false-negative detections [17].

3 Methodology

3.1 Dataset

A targeted set of physiologically relevant features was selected based on clinical significance and exploratory data analysis. The second dataset utilised is synthetic in nature and was used only to provide double verification to the machine learning models. These attributes were chosen to ensure effectiveness in both diabetes detection and subtype classification. The selection balances medical relevance with predictive utility.

Age plays a critical role in subtype differentiation. Type 1 Diabetes is commonly diagnosed in youth due to autoimmune β-cell destruction. Type 2 Diabetes correlates with age-related insulin resistance and β-cell decline. Type 3c (Pancreatogenic) Diabetes often follows chronic pancreatic diseases in older adults. Age thus reflects both pathophysiology and onset window. It helps the model separate autoimmune, metabolic, and secondary causes.

BMI (Body Mass Index) quantifies body fat based on height and weight. High BMI is directly linked to obesity-induced insulin resistance, a hallmark of Type 2 Diabetes. Excess adipose tissue disrupts insulin signaling and increases systemic inflammation. Obesity also elevates free fatty acids, impairing glucose uptake in muscles and liver. Higher BMI correlates with increased diabetes onset risk and disease progression.

Waist Circumference measures visceral fat, which is more metabolically active than subcutaneous fat. Visceral adiposity is strongly linked to insulin resistance and chronic inflammation. Unlike BMI, it specifically captures central obesity, a better predictor of Type 2 Diabetes. Excess abdominal fat increases hepatic glucose production and impairs insulin action. Studies show waist circumference independently correlates with diabetes risk.

Cholesterol Levels, particularly HDL and LDL, are often disrupted in diabetic patients. Type 2 Diabetes commonly presents with dyslipidemia: high LDL, low HDL, and elevated triglycerides. This lipid imbalance contributes to insulin resistance and vascular complications. Cholesterol abnormalities are also part of the metabolic syndrome cluster. They signal cardiovascular risk, which is elevated in all diabetes types.

Blood Glucose Levels are the primary diagnostic indicator for diabetes. Fasting glucose ¿126 mg/dL or postprandial ¿200 mg/dL confirms diabetic status. Chronic hyperglycemia results from impaired insulin production or action. It causes cellular damage, inflammation, and glucose toxicity across tissues. Prediabetes, Type 1, and Type 2 each show distinct glucose patterns. The model relies on glucose levels for both detection and subtype separation.

Insulin Levels indicate the pancreas's ability to produce insulin and the body's response to it. Low insulin suggests Type 1 Diabetes due to β-cell destruction. High insulin with hyperglycemia points to insulin resistance, typical of Type 2 and Prediabetes. Insulin levels help differentiate between deficiency-driven and resistance-driven diabetes. They also aid in identifying compensation failure in later Type 2 stages.

Pancreatic Health Indicators reflect structural or functional pancreatic abnormalities. Conditions like chronic pancreatitis or pancreatic tumors impair insulin secretion. Such damage leads to Pancreatogenic (Type 3c) Diabetes, distinct from Types 1 and 2. Markers may include pancreatic enzyme levels, history of surgery, or imaging findings. These features are rare in standard datasets but critical for detecting Type 3c. Their inclusion enables the model to capture secondary diabetes from exocrine dysfunction.

Figure 1 presents a Pearson correlation matrix for eight clinical features. Blood Glucose Levels exhibit the strongest positive correlation (0.86), indicating a strong linear association and potential predictive utility. Insulin Levels show a high correlation with the Target (0.59), further supporting the validity of the dataset with medical observation. In contrast, Pancreatic Health demonstrates a negligible negative correlation with the Target (r = -0.06), implying minimal linear dependence. Age, BMI, and Waist Circumference show moderate to strong inter-correlations (r = 0.63 to 0.68), mostly reflecting the wide range of people affected by diabetes with respect to there indices. BMI correlates most strongly with Age (r = 0.68), capturing age-related changes in body mass. Cholesterol Levels are moderately correlated with Age (r = 0.61) and BMI (r = 0.64), early the influence of metabolic factors across age and body composition. Insulin Levels correlate moderately with both Blood Glucose (r = 0.57) and Waist Circumference (r = 0.57), consistent with known metabolic relationships. Blood Glucose also shows moderate correlations with Insulin (r = 0.57), BMI (r = 0.48), and Waist Circumference (r = 0.56), further underscoring its central role in the metabolic profile.

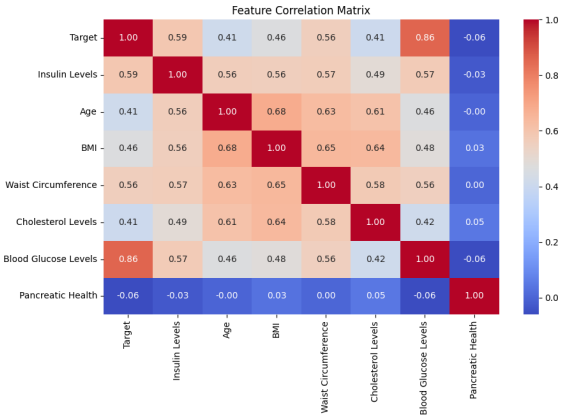

Figure 1: Correlation matrix of input features

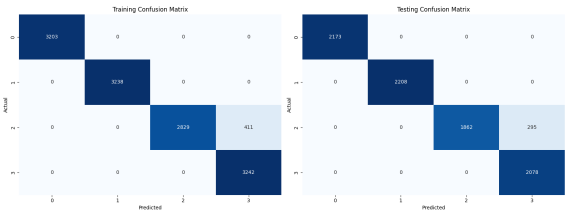

Figure 2: Random Forest confusion matrix

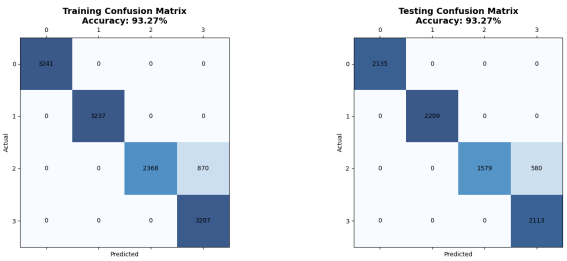

Figure 3: Decision Tree confusion matrix

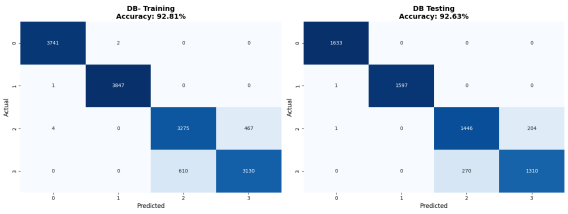

Figure 4: Discrete Bayes confusion matrix

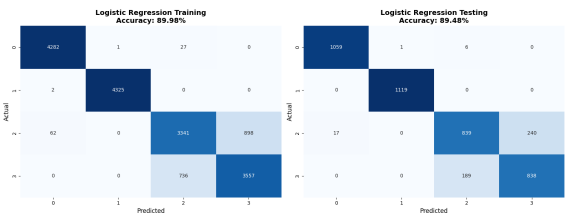

Figure 5: Logistic Regression confusion matrix

3.2 Machine Learning Models

Five supervised learning algorithms were employed to build and evaluate the predictive models: Naive Bayes Classifier, K-Nearest Neighbours (KNN), Logistic Regression, Random Forest, and XGBoost. These models were selected to represent a range of learning strategies — from simple probabilistic methods to advanced ensemble techniques. Each model was trained and tested in both stages: a binary classification to detect the presence of diabetes, and a multiclass classification to distinguish between subtypes. Among these, XGBoost consistently outperformed the others, showing higher accuracy and better handling of class imbalance, making it the most promising candidate for final deployment and interpretation.

4 Results and Conclusion

4.1 Results

In this section, we present the results of our machine learning experiments for diabetes detection using various supervised learning algorithms. We interpret the performance of each model based on relevant metrics such as accuracy, precision, recall, F1-score, specificity. The findings are visualized and compared comprehensively to determine the most effective classifier for the problem domain. Performance of the Decision Tree classifier was optimized using cost-complexity pruning with cross-validation. The optimal ccp alpha value was de- termined through a 10-fold Stratified K-Fold cross-validation process, balancing generalization and model simplicity. The final pruned tree achieved a testing accuracy of approximately 80 without overfitting the training data. The classification report demonstrated fairly balanced precision and recall values across the target classes, with the F1-score reflecting this balance.

The confusion matrices further revealed that the Decision Tree classifier was moderately effective in correctly classifying samples across all categories, although minor misclassifications between adjacent or overlapping conditions (e.g., class 2 vs. class 3) were observed. The feature importance plot provided insights into which attributes most influenced the model's decisions. In this case, features like Blood Glucose Level, Insulin, and BMI were identified as strong contributors to the model's decisions.

In contrast, the Naive Bayes classifier, optimized using GridSearchCV over the var smoothing parameter, showed slightly lower performance compared to the Decision Tree. The best model achieved a test accuracy of approximately 78 but slightly lower recall. This indicates that while the model was confident in its positive predictions, it had difficulty in identifying all relevant instances across classes. Given that Naive Bayes assumes independence between features, the slightly diminished performance is expected, especially in a dataset where attributes like BMI and insulin levels may be correlated. Nonetheless, the confusion matrix for Naive Bayes revealed that the classifier

managed to achieve reasonably good separation across the five target classes, although misclassifications were more prominent when compared to the Decision Tree. Particularly, class overlaps and false positives in adjacent classes highlighted the limitation of the strong independence assumption of the model. A consolidated table comparing all the models implemented (including additional models such as Lo- gistic Regression, KNearest Neighbors, Random Forest, XGBoost, and Voting Classifier) demonstrates that ensemble-based approaches consistently outperformed individual classifiers. XGBoost achieved the highest testing accuracy of 88. These models also recorded the highest F1-scores and AUC values, signifying strong discrimination capabilities. Decision Tree and KNN offered a balance between interpretability and accuracy, while Naive Bayes and Logistic Regres- sion, though slightly less accurate, maintained good precision.

4.2 Conclusion

In contrast, the Naive Bayes classifier, optimized using GridSearchCV over the var smoothing parameter, showed slightly lower performance compared to the Decision Tree. The best model achieved a test accuracy of approximately 78sion but slightly lower recall. This indicates that while the model was confident in its positive predictions, it had difficulty in identifying all relevant instances across classes. Given that Naive Bayes assumes independence between features, the slightly diminished performance is expected, especially in a dataset where attributes like BMI and insulin levels may be correlated. Nonetheless, the confusion matrix for Naive Bayes revealed that the classifier managed to achieve reasonably good separation across the five target classes, al- though misclassifications were more prominent when compared to the Decision Tree. Particularly, class over- laps and false positives in adjacent classes highlighted the limitation of the strong independence assumption of the model. A consolidated table comparing all the models implemented (including additional models such as Logistic Regression, KNearest Neighbors, Random Forest, XGBoost, and Voting Classifier) demonstrates that ensemble-based approaches consistently outperformed individual classifiers. XGBoost achieved the highest testing accuracy of 88ing Classifier and Random Forest. These models also recorded the highest F1-scores and AUC values, signifying strong discrimination capabilities. Decision Tree and KNN offered a balance between interpretability and accuracy, while Naive Bayes and Logistic Regression, though slightly less accurate, maintained good precision In contrast, the Naive Bayes classifier, optimized using GridSearchCV over the var smoothing parameter, showed slightly lower performance compared to the Decision Tree. The best model achieved a test accuracy of approximately 78sion but slightly lower recall. This indicates that while the model was confident in its positive predictions, it had difficulty in identifying all relevant instances across classes. Given that Naive Bayes assumes independence between features, the slightly diminished performance is expected, especially in a dataset where attributes like BMI and insulin levels may be correlated. Nonetheless, the confusion matrix for Naive Bayes revealed that the classifier managed to achieve reasonably good separation across the five target classes, although misclassifications were more prominent when compared to the Decision Tree. Particularly, class overlaps and false positives in adjacent classes highlighted the limitation of the strong independence assumption of the model. A consolidated table comparing all the models implemented (including additional models such as Logistic Regression, KNearest Neighbors, Random Forest, XGBoost, and Voting Classifier) demonstrates that ensemble-based approaches consistently outperformed individual classifiers. XGBoost achieved the highest testing accuracy of 88ing Classifier and Random Forest. These models also recorded the highest F1-scores and AUC values, signifying strong discrimination capabilities. Decision Tree and KNN offered a balance between interpretability and accuracy, while Naive Bayes and Logistic Regression, though slightly less accurate, maintained good precision.

References

[1] Gojka Roglic. Who global report on diabetes: A summary. *International Journal of Noncommunicable Diseases*, 1(1):3–8, 2016.

[2] Alfredo Caturano, Margherita D'Angelo, Andrea Mormone, Vincenzo Russo, Maria Pina Mollica, Teresa Salvatore, Raffaele Galiero, Luca Rinaldi, Erica Vetrano, Raffaele Marfella, et al. Oxidative stress in type 2 diabetes: impacts from pathogenesis to lifestyle modifications. *Current Issues in Molecular Biology*, 45(8):6651–6666, 2023.

[3] Arvind Kumar, Ruby Gangwar, Abrar Ahmad Zargar, Ranjeet Kumar, and Amit Sharma. Prevalence of diabetes in india: A review of idf diabetes atlas 10th edition. *Current diabetes reviews*, 20(1):105–114, 2024.

[4] Fatima Z Syed. Type 1 diabetes mellitus. *Annals of internal medicine*, 175(3):ITC33–ITC48, 2022.

[5] Sabrina Schlesinger, Manuela Neuenschwander, Janett Barbaresko, Alexander Lang, Haifa Maalmi, Wolfgang Rathmann, Michael Roden, and Christian Herder. Prediabetes and risk of mortality, diabetes-related complications and comorbidities: umbrella review of meta-analyses of prospective studies. *Diabetologia*, pages 1–11, 2022.

[6] Diego García-Compeán, Alan R Jiménez-Rodríguez, Juan M Muñoz-Ayala, José A González-González, Héctor J Maldonado-Garza, and Jesús Z Villarreal-Pé rez. Post-acute pancreatitis diabetes: A complication waiting for more recognition and understanding. *World Journal of Gastroenterology*, 29(28):4405, 2023.

[7] Channabasappa Shivaprasad, Yalamanchi Aiswarya, Shah Kejal, Atluri Sridevi, Biswas Anupam, Barure Ramdas, Kolla Gautham, and Premchander Aarudhra. Comparison of cgm-derived measures of glycemic variability between pancreatogenic diabetes and type 2 diabetes mellitus. *Journal of diabetes science and technology*, 15(1):134–140, 2021.

[8] Toshita Sharma and Manan Shah. A comprehensive review of machine learning techniques on diabetes detection. *Visual Computing for Industry, Biomedicine, and Art*, 4(1):30, 2021.

[9] Nour Abdulhadi and Amjed Al-Mousa. Diabetes detection using machine learning classification methods. In *2021 international conference on information technology (ICIT)*, pages 350–354. IEEE, 2021.

[10] Pronab Ghosh, Sami Azam, Asif Karim, Mehedi Hassan, Kuber Roy, and Mirjam Jonkman. A comparative study of different machine learning tools in detecting diabetes. *Procedia Computer Science*, 192:467–477, 2021.

[11] Puneeth N Thotad, Geeta R Bharamagoudar, and Basavaraj S Anami. Diabetes disease detection and classification on indian demographic and health survey data using machine learning methods. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 17(1):102690, 2023.

[12] Fayroza Alaa Khaleel and Abbas M Al-Bakry. Diagnosis of diabetes using machine learning algorithms. *Materials Today: Proceedings*, 80:3200–3203, 2023.

[13] Priyanka Rajendra and Shahram Latifi. Prediction of diabetes using logistic regression and ensemble techniques. *Computer Methods and Programs in Biomedicine Update*, 1:100032, 2021.

[14] Hiroe Seto, Asuka Oyama, Shuji Kitora, Hiroshi Toki, Ryohei Yamamoto, Jun'ichi Kotoku, Akihiro Haga, Maki Shinzawa, Miyae Yamakawa, Sakiko Fukui, et al. Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. *Scientific reports*, 12(1):15889, 2022.

[15] Hanaa Salem, Mahmoud Y Shams, Omar M Elzeki, Mohamed Abd Elfattah, Jehad F. Al-Amri, and Shaima Elnazer. Fine-tuning fuzzy knn classifier based on uncertainty membership for the medical diagnosis of diabetes. *Applied Sciences*, 12(3):950, 2022.

[16] Aga Maulana, Farassa Rani Faisal, Teuku Rizky Noviandy, Tatsa Rizkia, Ghazi Mauer Idroes, Trina Ekawati Tallei, Mohamed El-Shazly, and Rinaldi Idroes. Machine learning approach for diabetes detection using fine-tuned xgboost algorithm. *Infolitika Journal of Data Science*, 1(1):1–7, 2023.

[17] Md Ashraf Uddin, Md Manowarul Islam, Md Alamin Talukder, Md Al Amin Hossain, Arnisha Akhter, Sunil Aryal, and Maisha Muntaha. Machine learning based diabetes detection model for false negative reduction. *Biomedical Materials & Devices*, 2(1):427–443, 2024.