International Journal of Advanced Engineering and Management Research

Vol. 10, No. 06; 2025

ISSN: 2456-3676

Application of Machine Learning Techniques for Effective Diabetes Management

Seun Mayowa Sunday
ADEY Innovations Limited,
Stone House, GL10 3EZ, Gloucester, United Kingdom

doi.org/10.51505/ijaemr.2025.1517 URL: http://dx.doi.org/10.51505/ijaemr.2025.1517

Received: Nov 11, 2025 Accepted: Nov 20, 2025 Online Published: Dec 08, 2025

Abstract
The study employs machine learning techniques to predict and manage type 2 diabetes (T2DM). Various machine learning models, including Decision Tree, Support Vector Machine (SVM), Artificial Neural Network (ANN), and Gradient Boosting Machine (GBM), were assessed. The research involved extensive data preparation, feature selection using mutual information, and hyper parameter tuning via Grid Search CV. The GBM and SVM models demonstrated superior performance, achieving high accuracy and AUC values. Feature importance analysis identified glucose, BMI, and diabetes pedigree function as critical predictors. This study highlights the potential of machine learning to enhance diabetes management and advocates for future research to explore diverse datasets and advanced algorithms. The practical implications suggest significant improvements in personalized treatment plans and early intervention strategies for T2DM patients.

Keywords: Type 2 diabetes; machine learning; grid search; gradient boosting; feature selection

1. Introduction

Cardiovascular diseases (CVDs) impose a significant burden on international health, accounting for an estimated 17.9 million fatalities each year (figure 1.1), according to (World Heart Federation, 2023). These figures highlight the need for new cardiovascular disease identification, treatment, and management methods. The convergence of cardiovascular disease (CVD) and type 2 diabetes (T2DM) becomes an important topic of interest because of the intricate relationships and mutual exacerbates of these conditions.

Figure 1.1 illustrates the trends in the global number of deaths due to cardiovascular diseases from 1990 to 2019. The graph includes three lines: the black line represents the total number of CVD deaths globally, the red line represents the number of CVD deaths among females, and the grey line represents the number of CVD deaths among males. This figure shows that while the number of deaths has increased for both genders over the years, the total number of CVD deaths remains significantly higher for males compared to females.

Vol. 10, No. 06; 2025

ISSN: 2456-3676

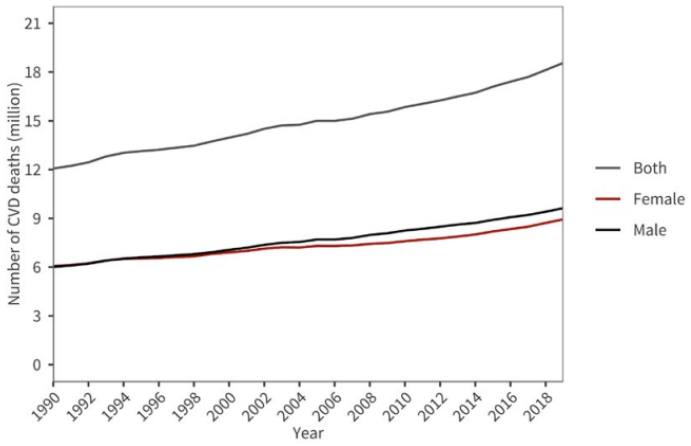


Figure 1.1 Trends in the global number of deaths due to cardiovascular diseases, 1990-2019 (source: World Heart Federation, 2023)

Type 2 diabetes increases the risk of coronary heart disease (CHD), a primary cause of myocardial infarction (MI), to levels comparable to those of non-diabetics who have had MI. The significance of confronting cardiovascular risk factors in diabetic patients with the same level of vigour as in those with a previous heart disease (Haffner, et al., 1998)

Beyond myocardial infarction, T2DM links to a wide range of cardiovascular effects, including peripheral artery disease and heart failure. The wide range of potential complications underscores the need for a holistic approach to managing cardiovascular disease in people with type 2 diabetes, thereby emphasising the significance of multifaceted care strategies (Shah, et al., 2015)

Genetic and epigenetic factors further reveal the relationship between T2DM and CVD, suggesting a shared pathophysiology. Hyperglycemia and prolonged hyperinsulinemia cause diabetic cardiomyopathy and atherosclerotic cardiovascular disease. Chronic pressure overload, myocardial infarction, and other processes can cause heart failure (De Rosa, et al., 2018)

This research aims to develop the ML model, which will help in the early detection of type 2 diabetes.

1.1 Background to the Study

The emergence of machine learning (ML) and artificial intelligence (AI) technologies marks a transformative era in healthcare, particularly in the management and early detection of type 2 diabetes mellitus (T2DM). While AI and ML algorithms have been very good at finding conditions like myocardial infarction (MI) through electrocardiograms (ECGs) and chest X-rays, using them to find T2DM presents its own opportunities and challenges (Liu, et al., 2021).

Vol. 10, No. 06; 2025

ISSN: 2456-3676

Despite the promising advances in AI and ML for healthcare, significant research gaps remain, particularly in the context of T2DM. These gaps include challenges like overfitting, where models capture noise instead of the underlying pattern, leading to misleading results and potentially harmful healthcare interventions (Park & Ho, 2020), (Ying, 2019). Furthermore, algorithmic biases pose ethical concerns, potentially exacerbating healthcare disparities, especially in T2DM management across diverse populations.

Incorporating AI and ML into T2DM research and treatment could yield substantial benefits, such as enhanced efficiency in clinical trials and personalized patient care. However, obstacles related to data standardization, interoperability, and regulatory validation of real-world evidence (RWE) persist. These challenges highlight the importance of utilizing practical data from electronic health records, wearable technology, and mobile health applications to inform patient outcomes and treatment efficacy while navigating regulatory and privacy concerns (Okolo, et al., 2024)

1.2 Aims and Objectives of the Study

To investigate and analyze Type 2 diabetes datasets using machine learning and data analytics techniques to improve the management and prediction of diabetes-related health outcomes.

1. Identify and select a suitable dataset pertaining to T2DM for the purpose of analysis.
2. Undertake initial data processing and exploratory data analysis (EDA) to gain insights into the understanding and connections of risk factors in the dataset, as well as assessing the impact of different feature selection techniques, including feature engineering, on the performance of models.
3. Evaluate the effectiveness of four distinct machine learning models in the prediction of T2DM.
4. Conduct an evaluation of the model's performance by using different measuring metrics such as the area under the curve (AUC), confusion matrix, classification accuracy, recall, precision, and F1-score.
5. Benchmarking the performance of the developed models against existing research in the field.

1.3 Research Questions

1. Which machine learning models are most effective in predicting the onset of type 2 diabetes based on patient data?
2. What patterns and features within diabetes datasets are most indicative of effective diabetes management strategies?
3. How can machine learning aid in the personalization of treatment plans for individuals with type 2 diabetes?

1.4 Research Rationale

Despite existing diagnostic methods and interventions, the incidence of T2DM continues to rise, highlighting the critical need for innovative approaches to identify at-risk populations early and

Vol. 10, No. 06; 2025

ISSN: 2456-3676

implement preventive measures more effectively. The integration of artificial intelligence (AI) and machine learning (ML) in analyzing complex patterns within large-scale health data presents a promising avenue to enhance our understanding of T2DM pathogenesis, risk factors, and progression. This study aims to leverage the transformative potential of AI and ML to develop more accurate and efficient tools for T2DM risk assessment, ultimately contributing to improved patient outcomes and reducing the disease's global impact.

1.5 The significance of the Research

The research holds considerable importance due to its potential to improve the accuracy of T2MD risk predictions. By employing machine learning algorithms, the goal of this research is to enhance the timely identification and control of T2DM. Furthermore, it addresses critical ethical concerns, including the protection of data privacy and the mitigation of biases, thereby showcasing a dedication to the ethical and responsible implementation of AI in the healthcare sector.

1.6 Scope of the Research

The goal of this research endeavor is to construct and assess four unique machine learning models: Support Vector Machines (SVM), Decision Trees (DT), Gradient Boosting Machines (GBM), and Artificial Neural Networks (NN). With this, it is expected that the data management strategy will be implemented. Although the integration of the best performing model is important into the healthcare end user device is important, however, in this study, the scope of will be the utilization of diverse performance metrics such as, recall, f1-score, accuracy, etc. to evaluate the precision with which these models predict T2DM risk.

1.7 Structure of the Research

The present study is methodically structured into six chapters, with each chapter fulfilling a vital function. The figure 1.1 summarizes the structure of this study and briefly documented subsequently.

Vol. 10, No. 06; 2025

ISSN: 2456-3676

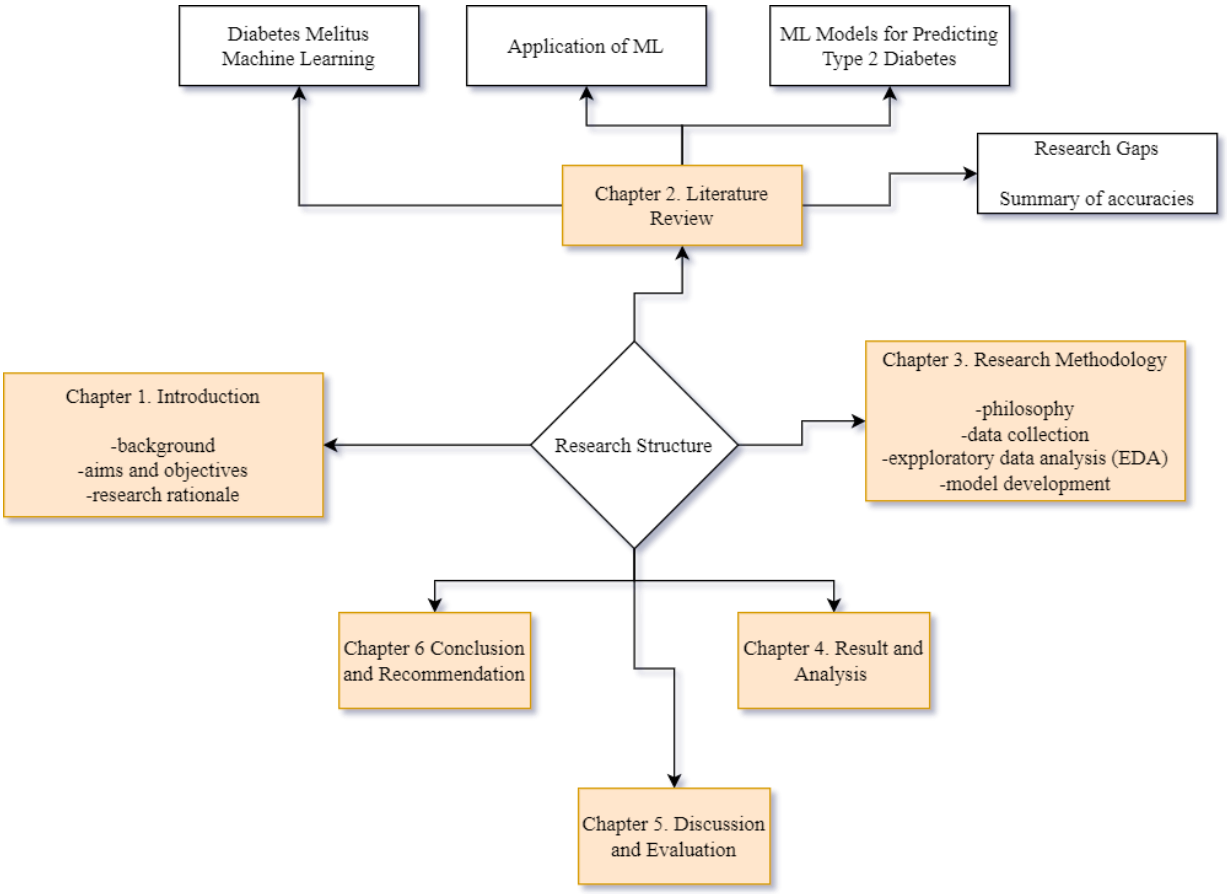


Figure 1.2 Research Structure (Source: Researcher Computation)

Chapter 1 introduces the study, thisalso documents aims, objectives. Chapter documents the conceptual, theoretical and empirical review. Chapter 3 outlines the research methodology, covering the selection of the dataset, data preprocessing steps, exploratory analyses, and the specific machine learning models employed. Chapter 4 carefully documents the results and analysis of the analyzed data. Chapter 5 presents an analysis of the findings in relation to chapter 4 of this study. This chapter 5 provides an analysis of the results, examining their significance, applicability, and inputs within the healthcare research. Chapter 6 documents the conclusion and the future research in this field.

2. Literature Review

This chapter discusses the detailed explanation of diabetes and the different machine learning techniques. Also, careful tabular summary of the analyzed research papers was also used to

Vol. 10, No. 06; 2025

ISSN: 2456-3676

provide quick summary of the accuracies of different works in this domain. The chapter concludes with the identification of the relevant research gaps and summary.

2.1 Definition of Type 2 Diabetes Mellitus

Type 2 Diabetes Mellitus (T2DM) is a prevalent chronic disease characterized by insulin resistance and impaired insulin secretion, leading to high blood glucose levels. It has a substantial effect on global health, leading to higher rates of illness and death. T2DM development is influenced by genetic, environmental, and lifestyle factors. This condition causes a progressive dysfunction of pancreatic β-cells and impairs the responsiveness of insulin-sensitive tissues to insulin. This malfunction results in persistent high blood sugar levels, a characteristic feature of the condition. If left untreated, it can give rise to many problems such as cardiovascular disease (CVD), kidney failure, damage to the retina, and nerve damage (Galicia-Garcia, 2020)

2.2 Introduction to Machine Learning

Machine learning (ML) is an essential part of artificial intelligence (AI) that enables computers to acquire knowledge from data and enhance their performance without the need for explicit programming for each specific task. It has become an essential component of multiple industries, particularly healthcare, where it is transforming the methods used to diagnose, treat, and control diseases.

2.2.1 Introduction to Machine Learning (ML)

Machine learning algorithms utilize statistical techniques to enable computers to 'learn' with data. The essence of ML lies in its ability to identify patterns and make decisions with minimal human intervention. Its applications range from everyday tasks like spam filtering and recommendation systems to complex problems such as autonomous driving and real-time fraud detection (Friedman & Labbi, 2020).

2.2.2 Categories of ML techniques

ML techniques are broadly categorized into three types based on the nature of the learning signal or feedback available to the learning system: supervised, unsupervised, and reinforcement learning (figure 2.1).

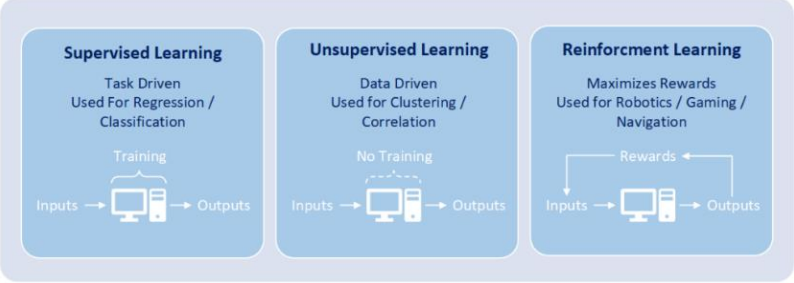


Figure 2.1 Types of Machine Learning (Source: Getz, 2019)

Vol. 10, No. 06; 2025

ISSN: 2456-3676

2.2.3 Role of ML in Healthcare

By enhancing diagnostics, modifying treatment, optimizing operational efficiency, and forecasting outbreaks, the implementation of machine learning in healthcare is reshaping the industry. ML algorithms identify trends in huge data sets that are impossible for humans to identify, thereby facilitating the early detection of diseases, genomics, and treatment efficacy insights. As an illustration, ML models have the capability of predicting patient risks, facilitate radiology imaging, and contribute to drug discovery through the analysis of drug molecular structures and the prediction of their interactions (Rani, et al., 2023).

The potential of ML in the healthcare sector is substantial, as it presents prospects for improving patient outcomes, decreasing expenses, and cultivating novel treatments. Nevertheless, the integration of AI tools into clinical workflows, concerns regarding data privacy, and the requirement for robust and interpretable models are all challenges that it presents. Notwithstanding these challenges, the indisputable advantages of machine learning in healthcare render it a pivotal domain for ongoing investigation and implementation.

2.3 Application of ML in Diabetes Management

In recent years, the application of machine learning (ML) in managing type 2 diabetes mellitus (T2DM) has witnessed significant evolution, shifting from basic data analysis to complex predictive modeling. This represents a paradigm shift in how healthcare professionals approach the disease, emphasizing personalized treatment plans and proactive management strategies (Afsaneh, Sharifdini, Ghazzaghi, & Ghobadi, 2022).

2.3.1 Overview of ML Techniques Used in Diabetes Prediction

Neural networks (NNs) have emerged as a cornerstone in the ML landscape for T2DM, primarily due to their ability to model complex, non-linear relationships between a multitude of variables. The depth and flexibility of NNs make them particularly effective in sifting through large datasets to uncover patterns and insights that elude traditional statistical methods. Their application ranges from identifying early markers of diabetes to predicting the risk of complications, thereby facilitating timely and targeted interventions. For instance, a study conducted by (Valchev, et al., 2023) demonstrated the use of NNs in identifying potential candidates for diabetes at its nascent stages, showcasing the model's predictive prowess in healthcare settings.

Support Vector Machines (SVMs) represent another critical ML technique in diabetes management. Known for their robustness and accuracy in classification tasks, SVMs have been adeptly used to categorize patients based on risk factors and predict disease outcomes. Their kernel-based approach allows for the handling of high-dimensional data, making SVMs invaluable in scenarios where the relationships between disease markers are complex and intertwined. Research, such as that by (Anton, et al., 2021), illustrates the application of SVMs in diagnosing diabetic retinopathy, underscoring the potential of ML in enhancing diagnostic precision.

Vol. 10, No. 06; 2025

ISSN: 2456-3676

2.4 ML Models for Predicting Type 2 Diabetes

To fully grasp the complex efficacy of machine learning (ML) models in predicting type 2 diabetes mellitus (T2DM), it is imperative to conduct an exhaustive examination of recent studies and methodologies. The purpose of this extensive analysis is to evaluate the application and performance of support vector machines (SVMs), decision trees (DTs), gradient boosting machines (GBMs), and neural networks (NNs) in various research projects, with a particular emphasis on contrasting their limitations and effectiveness.

2.4.1 Neural Networks (NNs) in T2DM Prediction

Recent developments in neural networks have exhibited considerable promise in enhancing the precision of prognostication for Type 2 Diabetes Mellitus (T2DM). An investigation that utilized convolutional neural networks (CNNs) in conjunction with Raman spectroscopy to detect diabetes is one example. The investigation revealed that neural networks exhibit potential in the realm of non-invasive disease detection, as evidenced by a classification accuracy of 95 (Yang & Zeng, 2023). In a recent investigation carried out by (Srinivasu, et al., 2022), deep neural networks (DNNs) were employed to predict Type 2 Diabetes Mellitus (T2DM) through the utilization of genomic and tabular data. The research emphasized the remarkable sensitivity and specificity of the model, thereby showcasing the capacity of neural networks to effectively handle complex biological data.

Although NNs exhibit remarkable performance, they are subject to criticism because of their opaque character, which gives rise to questions over interpretability, particularly in clinical contexts where comprehending the decision-making process is essential. For instance, (Chen, et al., 2020) highlighted that typical deep learning models are black-boxes, making the prediction outcomes difficult to interpret (Chen, et al., 2020). Similarly, (Zhang, et al., 2020) discussed the importance of interpretability for gaining trust in neural networks, especially in critical fields like healthcare (Zhang, et al., 2020). In addition, the need for extensive datasets and significant computational resources further restricts their usefulness in contexts with limited resources.

2.4.2 Support Vector Machines (SVMs) in T2DM Datasets

SVMs have been extensively explored for T2DM prediction due to their robustness in dealing with high-dimensional spaces. A study by (Caixeta, et al., 2023)used SVMs to screen for T2DM using salivary ATR-FTIR spectroscopy. They got a sensitivity of 93.3% and a specificity of 74%, which shows how useful SVMs are for clinical diagnostics. Another study utilizing SVMs for constructing a ten-gene biomarker prediction model for T2DM diagnosis achieved remarkable accuracy, underscoring SVMs' capability in biomarker identification and disease prediction (Li, et al., 2022)While SVMs offer considerable advantages in classification accuracy and the handling of complex datasets, their performance heavily relies on the choice of kernel function and the tuning of hyperparameters, which can be a challenging and time-consuming process.

Vol. 10, No. 06; 2025

ISSN: 2456-3676

2.4.3 Gradient Boosting Machines (GBMs) for Improving T2DM Prediction Accuracy

GBMs, particularly XGBoost and LightGBM, have emerged as powerful tools for T2DM prediction. A study that used XGBoost to predict the risk of T2DM found that it worked better than common machine learning models like SVMs, Random Forests, and K-Nearest Neighbors (Wang, et al., 2020), with an average accuracy of 89.09%. Another study into the use of LightGBM and adaptive boosting for predicting type-2 diabetes demonstrated the ensemble model's superior performance, with an accuracy of 94%, showing the potential of GBMs in enhancing predictive accuracy (Sai, et al., 2023).

Despite their efficacy, GBMs require careful tuning of parameters and can be computationally intensive, which may limit their use in settings with constrained computational resources.

2.5 Personalization of Diabetes Treatment Through Machine Learning

The personalization of diabetes treatment through machine learning (ML) signifies a transformative step in managing type 2 diabetes mellitus (T2DM), enabling tailored treatment strategies that cater to the individualized needs of patients. This personalized approach leverages ML's ability to analyses vast datasets, uncover patterns, and predict outcomes, thereby enhancing treatment efficacy and patient quality of life.

2.5.1 Case Studies on the Personalization of Treatment Using ML

Recent studies illustrate ML's capacity to personalize diabetes treatment. For instance, (Liao, et al., 2022) developed and validated ML prediction models for gestational diabetes treatment modality, demonstrating how ML can aid in deciding between medication and lifestyle changes based on patient-specific factors. Another innovative application by (Zhang, et al., 2023) utilized ML for post-acute pancreatitis, diabetes mellitus prediction, and personalized treatment recommendations, showcasing the potential of ML in addressing diabetes resulting from other health complications.

2.5.2 Impact of ML on Treatment Outcome Predictability

The impact of ML on treatment outcome predictability is profound. (Oikonomou & Khera, 2023) highlighted how ML improves precision in diabetes care and cardiovascular risk prediction, allowing for earlier interventions that could prevent severe complications. In a similar vein, (Hendawi, et al., 2023) developed XAI4Diabetes, a mobile app leveraging explainable AI to make diabetes predictions more interpretable to healthcare providers, thereby enhancing decision-making confidence.

Table 2.1 shows the dataset, model(s) and the summary of the accuracy gotten from different authors in their developed model.

Vol. 10, No. 06; 2025

ISSN: 2456-3676

Table 2.1 Summary of Accuracies of Research papers

Authors	Dataset	Model	Accuracy
(Hasan, et al., 2020)	PIMA Indians Diabetes (PID)	k-nearest Neighbor (k-NN), Decision Trees (DT), Random Forest (RF), Naive Bayes (NB), AdaBoost (AB) and XGBoost (XB)	K-NN (92.6), DT (91.2), RF (93.9), AB (94.1), NB (87.9), XB (94.6)
(Yahyaoui, et al., 2019))	PIMA Indians Diabetes (PID)	Support Vector Machine (SVM), Random Forest (RF), and Convolutional Neural Network (CNN)	SVM (65.38%), RF (83.67%), CNN (76.81%)
(Sonar & JayaMalini, 2019)	Not specified	Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM),	DT (85%), 77% (NB), 77.3% (SVM)
(Khanam & Foo, 2021)	Pima Indian diabetes (PID)	Decision Tree (DT), K Nearest Neighbor (KNN), Random Forest (RF), Naive Bayes (NB), Adaboost (AB), Logistic Regression (LR), Support Vector Machine (SVM), and Neural Network (NN)	DT (74.24%), RF (77.14%), NB (78.28%), LR (78.85%), KNN (79.42%), AB (79.42%), SVM (77.71%) and NN (88.6%)

Table 2.1 Research Paper Accuracy Summary

2.6 Gaps in Literature

The literature highlights significant advancements in the use of machine learning (ML) for managing type 2 diabetes mellitus (T2DM), emphasizing prediction, treatment personalization, and monitoring. One noted gap is the limited exploration of specific ML models' effectiveness across diverse diabetes datasets. While studies have showcased the potential of neural networks, support vector machines, decision trees, and gradient boosting machines in diabetes prediction

Vol. 10, No. 06; 2025

ISSN: 2456-3676

and management, it can be noticed that the dataset publicly available in Kaggle are commonly used as their benchmark which is simply good for learning data science (PyCoach, 2022). This gap is filled in the present study which utilize the dataset available at IEEE-Data (Singh, 2024). Although the accuracies of the developed models as shown in table 2.1 are high, the results present room for more accurate models. The present study intends to also fill this gap by ensuring that the developed models in this study present higher accuracy.

3. Research Methodology

The research methodology section details the techniques and procedures used to assess the risk prediction of type 2 diabetes (T2DM). This methodology ensures the robustness, reliability, and validity of the findings. By integrating a curated dataset, feature engineering, strategic feature selection, and machine learning models, the study aims to refine the precision of predictive capabilities for the risk of developing T2DM.

3.1 Research Philosophy

The philosophy that informs this research illuminates the path for the methodological approach and analytical perspective. It is the philosophical stance that not only influences the selection of methods, but also profoundly impacts the interpretation of data and the conclusions that emerge. The study used the positivist approach to actualize the research aim and the objectives (Mkansi & Mkansi, 2023).

3.1.1 The Role of Statistical Validation

A fundamental aspect of the positivist approach is the focus on statistical validation and the ability to replicate findings. The analytical tools and techniques chosen are prized for their robustness and reliability. Through rigorous statistical testing and validation, we ensure that the results are not only applicable to the specific dataset analyzed but also hold relevance in wider contexts. This dedication to replicability and generalizability is a defining feature of our research philosophy.

3.1.2 Justification of Methodological Choices

To ensure methodological soundness, each analytical decision in this study was grounded in established best practices. The selection of SVM, DT, ANN, and GBM was informed by their consistent performance in clinical prediction tasks reported in contemporary literature. GridSearchCV with 5-fold cross-validation was employed to minimize overfitting and provide reliable hyper parameter optimization. StandardScaler was adopted because several algorithms—particularly SVMs and ANNs—are highly sensitive to differences in feature magnitude. RandomOverSampler was chosen instead of synthetic sampling methods such as SMOTE, as the dataset includes zero-inflated variables (e.g., insulin) where synthetic interpolation may distort clinical meaning. The 80/20 train-test split reflects common practice for medium-sized clinical datasets and ensures a balanced approach between training depth and validation reliability.

Vol. 10, No. 06; 2025

ISSN: 2456-3676

3.2 Data Source and Collection

The study utilized a dataset developed by (Singh, 2024), which is vital for understanding the complexities associated with the risk of type 2 diabetes in the healthcare industry. To load this dataset, the use of the Pandas library was indispensable. Pandas, a cornerstone of the Python programming ecosystem, holds great esteem in data science and research due to its powerful capabilities in managing large datasets, efficiently organizing them, and conducting preliminary analyses. Pandas provides an extensive set of tools that simplify data manipulation, from restructuring datasets to visualizing them in various forms.

Appendix A displays the dataset in a tabular format, showing the features and their definitions.

3.3 Data Cleaning

Data cleaning is an essential step in data-driven research, crucial for ensuring analyses are performed on data that is not only clean but also relevant and of high quality.

The initial stage of this process involved a detailed examination of the dataset's structure. This analysis was necessary to understand the types of objects in the dataset and identify any missing or null values.

To address missing data, Pandas was used, however, the Seaborn library was utilized for proper visualization, which helps to quickly visualize if there is any null or missing value. As noted by (Li, et al., 2022), a heatmap indicates the presence of missing values and shows a clear view when there are none.

3.4 Feature Transformation

During the initial stage of machine learning processing, it is essential to verify that the data is correctly structured. The dataset exclusively consists of numerical features, represented as either integers or floating-point numbers, to meet the specific criteria of the selected machine learning algorithms: Decision Trees (DT), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Gradient Boost (GBM). According to (Brownlee, 2020), states that the process of converting data into numerical form not only simplifies calculations but also enhances the performance of models by making relationships clearer. In this dataset, all the entries are in numerical form from the onset, hence, no further step to ensure conversion to numerical data.

3.5 Exploratory Data Analysis (EDA)

Every dataset is distinctive, and the one being examined here is no exception. Before constructing models, it is essential to understand the core structure of the dataset and detect any anomalies (Brownlee, 2020). A range of visualization methods were utilized to highlight these features. Upon careful examination of the 'label' attribute, it was observed that there is an uneven distribution, which is a common issue that can cause models to favor the most common class, thereby impacting the accuracy of predictions (Venkatesh & Anuradha, 2019)

Vol. 10, No. 06; 2025

ISSN: 2456-3676

3.6 Correlation Matrix

A correlation matrix is a statistical technique employed to ascertain the association between two variables within a dataset. The information is displayed in a tabular manner, with each cell containing a correlation coefficient (figure 3.1). A coefficient of 1 implies a high degree of connection, whereas a coefficient of 0 suggests no association, and a coefficient of -1 suggests a low degree of correlation (Wagavkar, 2023). In this research, after following the mutual information (MI) algorithm (refer to Section 3.7), the features were visualized using Matplotlib to examine their correlation matrix.

3.7 Feature Importance and Selection

This segment of the document highlights the use of the mutual information algorithm for identifying the most effective features remaining in the dataset.

3.7.1 Mutual Information on Feature Importance

Not all features have an equal impact on a model's predictive abilities. In order to differentiate between features that have a greater or lesser impact, the mutual information metric was computed for each feature in relation to the target variable (Brownlee, 2020). Mutual information measures the extent to which the target variable relies on each independent variable. The interdependence was visually depicted using a bar chart, effectively illustrating the relative significance of each characteristic.

3.7.2 Deciding on Features and Their Importance

Initial analyses led to the adoption of a 0.02 threshold for feature selection. This threshold was based on statistical significance and the need for mutual information. The mutual information expects that after checking the importance of all the independent variables against the dependent variables, the features which are not contributing effectively (based on their score) are dropped (Beraha, et al., 2019). Features falling below this threshold were considered less critical and consequently excluded from the dataset. This decision was informed by the goal of enhancing model performance by reducing complexity and focusing on the most influential variables.

3.8 Model Development and Training

The process of transitioning from raw data to a functional predictive model involves a sequence of crucial steps, each of which is essential to the overall process. The construction of machine learning models is based on these fundamental techniques.

3.8.1 Data Splitting

Before commencing training, it is crucial to clearly define the dataset to be used for training and validation. This decision is crucial. This study employed an 80-20 split, which is a widely used approach in machine learning research. Assigning 80% of the data to the training set resulted in a significant number of data points that could be used for identifying patterns. The remaining 20%, which is set aside for testing, holds similar significance. This division guarantees an impartial

Vol. 10, No. 06; 2025

ISSN: 2456-3676

assessment of the model's effectiveness under unexpected conditions (Khanam & Foo, 2021). This approach has been widely adopted in several studies. For instance, (Huč, et al., 2021) used an 80-20 split in their analysis of machine learning algorithms for anomaly detection on edge devices, achieving high accuracy and robust validation results (Huč, et al., 2021). Similarly, (Kahloot & Ekler, 2021) employed an 80-20 split to prepare sub-dataset splits for machine learning models, ensuring balanced and representative training and validation sets.

3.8.2 Data Scaling

Datasets sometimes exhibit variability in the scales of their features, with certain features covering wide ranges while others are more limited. If these disparities are not addressed, machine learning algorithms may develop biases, perhaps misinterpreting the significance of a feature based on its magnitude (Kumar, et al., 2020). To mitigate this, the standard scaler was employed.

The standard scaler is a widely used method in data preprocessing that standardizes features by removing the mean and scaling to unit variance. This process transforms the data into a distribution with a mean of 0 and a standard deviation of 1, which makes it suitable for many machine learning algorithms that rely on the assumption that the input features are normally distributed. According to (Ahsan, et al., 2021), the use of standard scaling significantly improves the performance of machine learning models, particularly in medical datasets where feature scales can vary widely.

3.8.3 Oversampling

Achieving the data balance is an important aspect of machine learning (ML). Imbalances, particularly in class distribution, introduce significant challenges (Kumar, et al., 2020). Models trained on unbalanced data tend to be biased towards the predominant class, often neglecting the nuances of the minority class (Qaddoura, et al., 2021). The dataset used in this study exhibited such an imbalance. To rectify this, the Random Over Sampler technique was implemented. To augment the minority class rather than reduce the majority class avoided the loss of valuable data (Gadelrab, et al., 2018). This approach ensured a balanced representation of both classes during training, safeguarding the model against inherent biases and fostering a understanding of the data.

The combined application of data scaling, class balancing, and feature selection ensured that each machine-learning model was trained on standardized and statistically meaningful data. This structured pipeline, implemented consistently across all models, enhances comparability and reduces the risk of model bias, thereby improving the validity of the results presented in Chapter 4.

3.9 Model Implementation

The process of transforming unprocessed data into practical insights requires careful examination and deliberate decision-making. To exploit the dataset's depth and peculiarities, we employed

Vol. 10, No. 06; 2025

ISSN: 2456-3676

four separate machine learning algorithms: decision tree (DT), support vector machine (SVM), artificial neural network (ANN), and gradient boost (GBM).

3.9.1 Decision Tree

The Decision Tree classifier (Charbuty & Abdulazeez, 2021), was meticulously calibrated to enhance the predictive analysis of Type 2 diabetes (T2DM). A parameter grid was established, specifying 'max_depth' with options [None, 10, 20, 30] to determine the tree's growth limit, and 'min_samples_split' with values [2, 10, 20] to define the minimum number of samples required to split an internal node.

A GridSearchCV (Alibrahim & Ludwig, 2021), was instantiated for the Decision Tree, which incorporates the parameter grid and a 5-fold cross-validation scheme. The objective was to optimize the model for accuracy, ensuring that the most predictive features of T2DM were utilized. The grid search was executed on the training data, systematically exploring the parameter space to identify the optimal tree structure.

3.9.2 Support vector machine (SVM)

In the pursuit of improving the predictive analytics of Type 2 diabetes (T2DM), the study also employed Support Vector Machine (SVM) as one of the pivotal machine learning models. The SVM's capability to handle non-linear boundaries and its robustness in high-dimensional spaces make it an ideal candidate for such complex classification (Abdullah & Abdulazeez, 2021)

The methodology began with the definition of a parameter grid, which is essential for optimizing the SVM. The parameters included the regularization parameter 'C', with values [0.1, 1, 10], to control the trade-off between achieving a low training error and a low testing error. The 'gamma' parameter, set to ['scale', 'auto'], determined the influence of a single training example, and the 'kernel' parameter, with options ['rbf', 'linear'], specified the kernel type to be used in the algorithm, thus dictating the decision boundary's shape.

A GridSearchCV object was instantiated with the SVM estimator, the defined parameter grid, and a 5-fold cross-validation strategy (Alibrahim & Ludwig, 2021). The scoring metric was set to 'accuracy', ensuring that the model's performance was evaluated based on its ability to correctly predict T2DM instances. The return_train_score parameter was set to True to retain the scores on the training sets during the cross-validation process, providing insight into the model's generalization capabilities.

3.9.3 Artificial neural network (ANN)

The Artificial Neural Network (ANN), was leveraged to enhance the predictive analysis of Type 2 diabetes (T2DM). The ANN's architecture, inspired by biological neural networks, offers a robust framework for capturing complex patterns in data (Ahmed, et al., 2022).

Vol. 10, No. 06; 2025

ISSN: 2456-3676

A parameter grid was defined to optimize the ANN. This grid included `hidden_layer_sizes` with configurations [(50,), (100,), (50, 50)] to determine the network's capacity, `activation` functions ['tanh', 'relu'] to introduce non-linearity, `solver` options ['sgd', 'adam'] for optimization, `alpha` values [0.0001, 0.001, 0.01] as regularization terms, and `learning_rate` strategies ['constant', 'adaptive'] to modulate the learning process.

A GridSearchCV object was created, encapsulating the MLPClassifier with a maximum iteration limit of 1000, the parameter grid, and a 5-fold cross-validation setup. The scoring metric was set to 'accuracy', aligning the optimization process with the study's goal of accurate T2DM risk prediction.

3.9.4 Gradient boost

In this study, the Gradient Boosting model was optimized. This model, renowned for its predictive prowess, was subjected to a hyperparameter tuning process to ensure its efficacy in identifying T2DM risk factors (Kopitar, et al., 2020).

The parameter grid was carefully constructed to explore a range of values for 'n_estimators', 'learning_rate', 'max_depth', and 'min_samples_split'. These parameters were chosen to fine-tune the model's complexity and learning capacity. Specifically, 'n_estimators' were varied from 100 to 300 to determine the optimal number of trees, while 'learning_rate' values of 0.01, 0.1, and 0.2 were tested to control the contribution of each tree. The 'max_depth' was adjusted between 3 and 7 to manage the depth of each tree, and 'min_samples_split' was set from 2 to 6 to dictate the minimum samples required for a node split.

3.10 Model Evaluation

Assessing the performance of created and trained models is a critical stage in establishing their effectiveness in producing precise predictions. This section examines the many metrics employed to evaluate the performance of a model and considers their significance.

3.10.1 Confusion Matrix

The confusion matrix (figure 3.1) is a fundamental tool for model evaluation, providing a visual representation of the performance of classification models (Luque, et al., 2019). It is a table with two dimensions: “Actual” and “Predicted,” and each cell in this matrix represents the count of instances for the combination of the predicted and actual values. The key parameters of confusion matrix are provided below:

- True Positives (TP): Correctly predicted positive observations.
- True Negatives (TN): Correctly predicted negative observations.
- False Positives (FP): Incorrectly predicted positive observations (Type I error).
- False Negatives (FN): Incorrectly predicted negative observations (Type II error).

Vol. 10, No. 06; 2025

ISSN: 2456-3676

The confusion matrix allows researchers to calculate various performance metrics, such as accuracy, precision, recall, and F1-score. Accuracy, calculated as ((TP + TN) / (TP + TN + FP + FN)), gives the overall correctness of the model but may not be sufficient for imbalanced datasets.

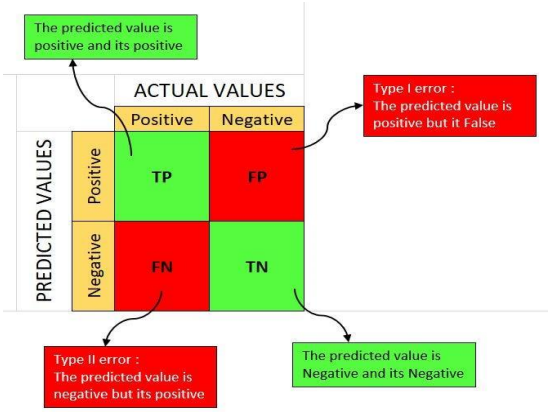


Figure 3.1 The diagrammatic representation of Confusion Matrix (Source: Research Computation)

3.10.2 Classification Report

The classification report expands on the confusion matrix by providing a text report showing the main classification metrics (Luque, et al., 2019). The below are included in the classification report:

- Precision: The ratio of correctly predicted positive observations to the total predicted positives. It is given by (Precision = TP / (TP + FP)).
- Recall (Sensitivity): The ratio of correctly predicted positive observations to all observations in the actual class. It is given by (Recall = TP / (TP + FN)).
- F1-Score: The weighted average of Precision and Recall. It is calculated as (F1 = 2 \times (Precision \times Recall) / (Precision + Recall)).
- Support: The number of actual occurrences of the class in the specified dataset.

3.10.3 ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) are pivotal in evaluating the performance of machine learning models, particularly in the context of binary classification tasks like the present study (Nahm, 2022). These metrics provide a measure of a model's ability to distinguish between the two classes (Hoo, et al., 2017).

In this study, a comparative analysis of four distinct models—Support Vector Machine (SVM), Decision Tree, Gradient Boosting Machine (GBM), and Artificial Neural Network (ANN)—was conducted using the ROC curve and AUC. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings, providing insight into the trade-

Vol. 10, No. 06; 2025

ISSN: 2456-3676

off between sensitivity and specificity. The AUC, a scalar value, quantifies the overall ability of the model to rank positive instances higher than negative ones.

The models were assessed using a function plot combined_roc_curve, which iterated through each model's best estimator obtained from grid search optimization. For models with a predict proba method, the probabilities of the positive class were used, while for those with a decision function, decision scores were normalized to a 0-1 range to serve as probabilities. Any model yielding NaN values was flagged for investigation, as this indicates potential issues in probability estimation or decision function computation.

The ROC curves for each model were plotted, and their respective AUCs were calculated. The AUC values, ranging from 0.5 (no discrimination) to 1 (perfect discrimination), provided a clear metric for comparison. Models with higher AUC values demonstrated a greater capacity to differentiate between patients at risk of T2DM and those not at risk.

3.11 Ethical Considerations in Machine Learning and Data Analysis

In the domain of machine learning and data analysis, especially when working with datasets that include potentially sensitive personal information, ethical considerations are of utmost importance. During the development and evaluation of models using the Type 2 diabetes dataset developed by (Singh, 2024), several ethical concerns were rigorously addressed:

Protection of Data and Ensuring Security: The dataset containing health information pertinent to Type 2 diabetes, could include sensitive information. It was critical to ensure that personal identities or specific details remained confidential, protecting individual privacy. Although as documented earlier, the dataset used in this study is secondary data which is from (Singh, 2024), however, based on the actual source of the data, it is obvious that the data is well protected.

Potential Bias and Upholding Fairness: Machine learning models can inherently reflect biases present in the training data. A biased model may yield unequal outcomes for different Type 2 diabetes scenarios, potentially leading to incorrect predictions or overlooked diagnoses (Torrie & Payette, 2020). It was imperative to maintain fairness in model outcomes, not only for accuracy but also for the ethical application of these models.

3.12 Conclusion

This chapter has carefully documented the detailed research methodology. From the data cleaning to the model development. The next chapter discusses the result of this chapter.

4. Method and Results

4.1 Method and Results

This chapter provides a detailed explanation of the methodologies and strategies employed in the execution of the research, followed by a documentation of the findings in response to the stated

Vol. 10, No. 06; 2025

ISSN: 2456-3676

research questions. This chapter started the documentation from the data preprocessing, feature selection, then concluded at the model development, confusion matrix and the general summary.

4.1.1 Data Pre-processing

The analysis of the dataset reveals several key insights regarding the features and their distributions. Initially, the top five entries (Figure 4.1) offer a quick look of the dataset, including columns such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome. Each entry illustrates the varied nature of the data points, highlighting differences in glucose levels, BMI, and other attributes among individuals.

The descriptive statistics (Figure 4.2) provide a summary of the dataset, showing measures such as the count, mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum values for each feature. For instance, the mean glucose level is approximately 121.18, with a standard deviation of 32.07, indicating a moderate spread around the mean. The median values, included in the statistical summary (Figure 4.3), further clarify the central tendency of the data, emphasizing the skewness in some features, such as Insulin and Skin Thickness, where the means are significantly higher than the medians.

To ensure the dataset's completeness, a check for null values was performed (Figure 4.4). The heatmap confirms the absence of missing data, ensuring the integrity of subsequent analyses. This step is crucial for maintaining data quality and reliability in predictive modeling and other analytical procedures (Albahri, et al., 2023)

The imbalance in the dataset is evident in the distribution of the Outcome variable (Figure 4.5). With 1,316 instances of the outcome '0' and only 684 instances of '1', there is a clear imbalance that could affect the performance of machine learning models. This disparity validates the need for potential resampling techniques, or the application of algorithms designed to handle imbalanced data to avoid biased predictions (Mohammed, et al., 2020).

Lastly, the distribution plots of each feature (Figure 4.6) reveal the underlying patterns and tendencies within the dataset. For example, the distributions of Pregnancies and Age are right-skewed, indicating that most observations fall within the lower ranges. Conversely, the distributions of BMI and Glucose exhibit a more normal distribution with some skewness, suggesting varied health profiles among the subjects. The Insulin levels show a significant skewness towards lower values, with many instances of zero insulin, likely due to missing or unmeasured data (Mohammed, et al., 2020).

Vol. 10, No. 06; 2025

ISSN: 2456-3676

Out[4]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	62	35	0	33.6	0.127	47	1
1	0	84	82	31	125	38.2	0.233	23	0
2	0	145	0	0	0	44.2	0.630	31	1
3	0	135	68	42	250	42.3	0.365	24	1
4	1	139	62	41	480	40.7	0.536	21	0

Figure 4.1 Top 5 Entries of dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               2000 non-null   int64
 1   Glucose                   2000 non-null   int64
 2   BloodPressure             2000 non-null   int64
 3   SkinThickness             2000 non-null   int64
 4   Insulin                   2000 non-null   int64
 5   BMI                       2000 non-null   float64
 6   DiabetesPedigreeFunction  2000 non-null   float64
 7   Age                       2000 non-null   int64
 8   Outcome                   2000 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 140.8 KB

(None,
        Pregnancies      Glucose  BloodPressure  SkinThickness      Insulin  \
 count  2000.000000  2000.000000    2000.000000    2000.000000  2000.000000
 mean      3.703500   121.182500      69.145500      20.935000    80.254000
 std       3.306063    32.068636      19.188315      16.103243   111.180534
 min       0.000000     0.000000       0.000000       0.000000     0.000000
 25%       1.000000    99.000000      63.500000       0.000000     0.000000
 50%       3.000000   117.000000      72.000000      23.000000    40.000000
 75%       6.000000   141.000000      80.000000      32.000000   130.000000
 max      17.000000   199.000000     122.000000     110.000000   744.000000

               BMI  DiabetesPedigreeFunction          Age      Outcome
 count  2000.000000               2000.000000  2000.000000  2000.000000
 mean     32.193000                  0.470930    33.090500     0.342000
 std       8.149901                  0.323553    11.786423     0.474498
 min       0.000000                  0.078000    21.000000     0.000000
 25%      27.375000                  0.244000    24.000000     0.000000
 50%      32.300000                  0.376000    29.000000     0.000000
 75%      36.800000                  0.624000    40.000000     1.000000
 max      80.600000                  2.420000    81.000000     1.000000  )
```

Figure 4.2 Descriptive Statistics

Vol. 10, No. 06; 2025

ISSN: 2456-3676

```
# Calculate the statistical summaries including mean and median for each column
statistical_summary = diabetes_data.describe()
statistical_summary.loc[['mean', '50%']]  # '50%' represents the median
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
mean	3.7035	121.1825	69.1455	20.935	80.254	32.193	0.47093	33.0905	0.342
50%	3.0000	117.0000	72.0000	23.000	40.000	32.300	0.37600	29.0000	0.000

Figure 4.3 Statistical Summary

```
plt.figure(figsize=(15,3))
plt.title('Dataset With/Without Null Value')
sns.heatmap(diabetes_data.isnull()[:-1],yticklabels=False,cbar=False,cmap='PuBuGn_r')
```

```
<Axes: title={'center': 'Dataset With/Without Null Value'}>
```

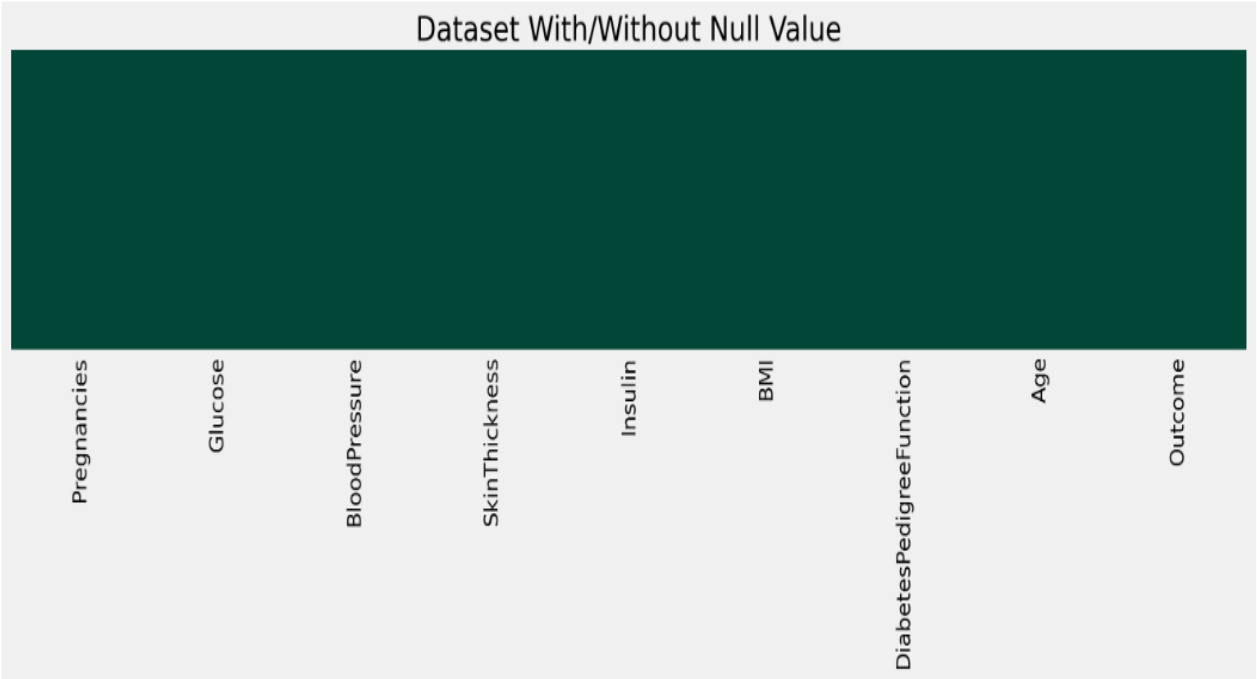


Figure 4.4 Check if the Dataset is Null or Not

Vol. 10, No. 06; 2025

ISSN: 2456-3676

```
sns.set_style('whitegrid')
sns.countplot(x='Outcome',data=new_data,palette='gist_ncar_r')

Outcome
0    1316
1     684
Name: count, dtype: int64

<Axes: title={'center': 'Imbalanced Dataset'}, xlabel='Outcome', ylabel='count'>
```

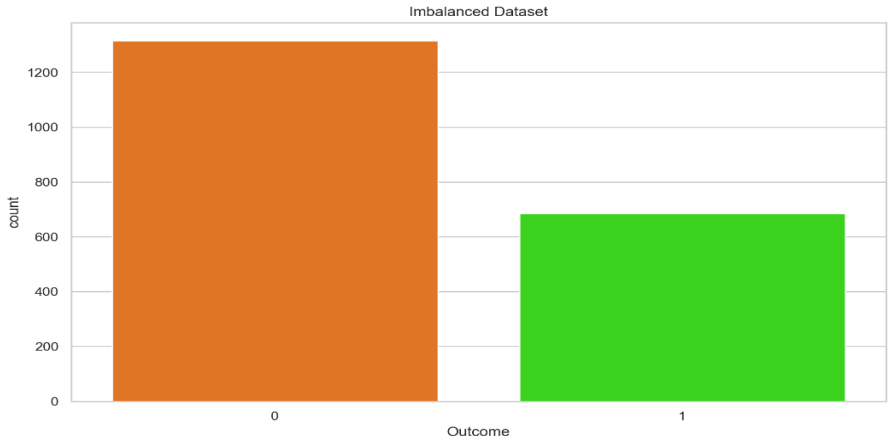


Figure 4.5 Balanced and Imbalanced Nature of the Dataset

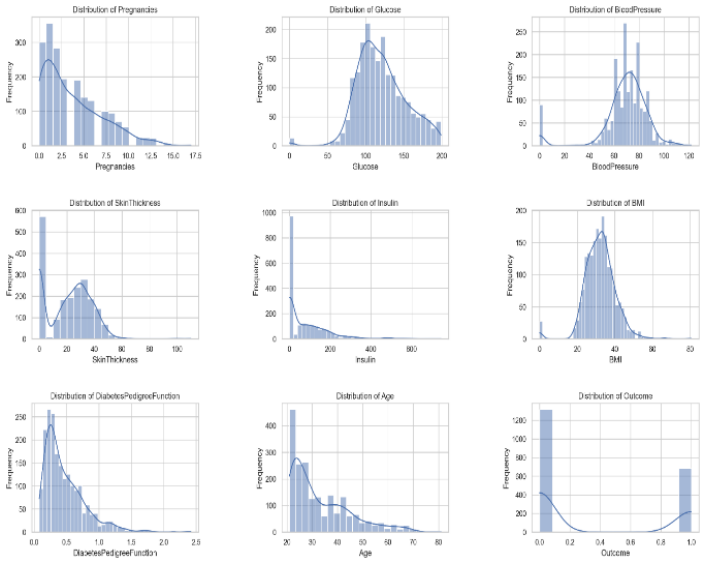


Figure 4.6 Distribution of features

4.1.2 Feature Selection

The feature selection process using the mutual information algorithm provides critical insights into the most influential variables in predicting diabetes outcomes. Mutual information measures the interdependence between variables, helping to identify features that provide the most significant information about the target variable. This method ensures that the selected features

Vol. 10, No. 06; 2025

ISSN: 2456-3676

are both relevant and minimally redundant, enhancing the model's predictive accuracy (Beraha, et al., 2019).

The mutual information algorithm, which is calculated from the scikit-learn library (Scikit-Learn, 2019), shows that the bar chart (Figure 4.7) clearly illustrates that glucose, BMI, and diabetes pedigree function are the most significant predictors, with glucose having the highest importance score. This aligns with existing medical knowledge, as elevated glucose levels are a primary indicator of diabetes. BMI and diabetes pedigree function also play crucial roles, emphasizing the importance of body mass and genetic predisposition in the development of type 2 diabetes. Studies have shown that a higher BMI is associated with an increased risk of developing T2DM, and a family history of diabetes, captured by the Diabetes Pedigree Function, further elevates this risk (Shah, et al., 2020)

Post feature selection, the dataset was refined to include only the most relevant features (Figure 4.8). This step is essential to enhance model performance by reducing dimensionality and focusing on the variables that contribute the most to predictive accuracy. The selected features—Pregnancies, Glucose, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age—are retained for subsequent analysis. The snapshot of the refined dataset maintains a representation of the diverse risk factors associated with diabetes while eliminating less informative variables.

The box plots of the selected features (Figure 4.9) offer a detailed view of the data distribution and potential outliers. For instance, the Glucose and Insulin plots reveal significant outliers, indicating variability in these measures among individuals. Such outliers are critical to consider, as they may represent extreme cases of diabetes or pre-diabetes conditions. The box plots for BMI and Diabetes Pedigree Function also show a relatively widespread, highlighting the diversity in body composition and genetic predisposition within the dataset (Alghushairy, et al., 2020)

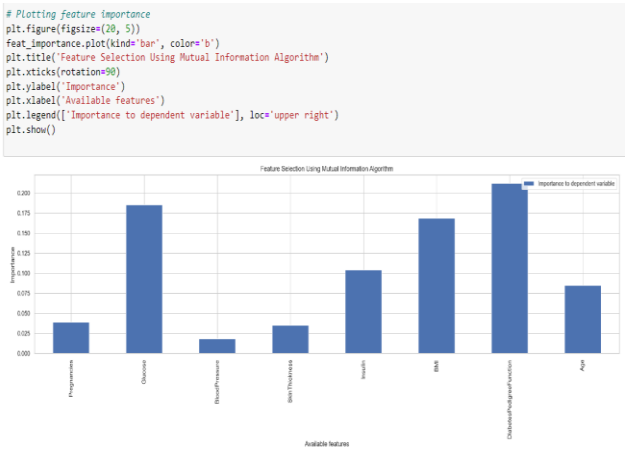


Figure 4.7 Feature selection using mutual information

Vol. 10, No. 06; 2025

ISSN: 2456-3676

```
threshold = 0.025
selected_features = feat_importance[feat_importance >= threshold].index.tolist()

new_data = diabetes_data[selected_features + ['Outcome']]
```

```
new_data.head()
```

	Pregnancies	Glucose	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	35	0	33.6	0.127	47	1
1	0	84	31	125	38.2	0.233	23	0
2	0	145	0	0	44.2	0.630	31	1
3	0	135	42	250	42.3	0.365	24	1
4	1	139	41	480	40.7	0.536	21	0

Figure 4.8 Selected features after feature selection

```
for i, col in enumerate(diabetes_data.columns):
    sns.boxplot(y=diabetes_data[col], ax=axes[i], color='skyblue')
    axes[i].set_title(f'Boxplot of {col}')
    axes[i].set_ylabel('')

plt.show()
```

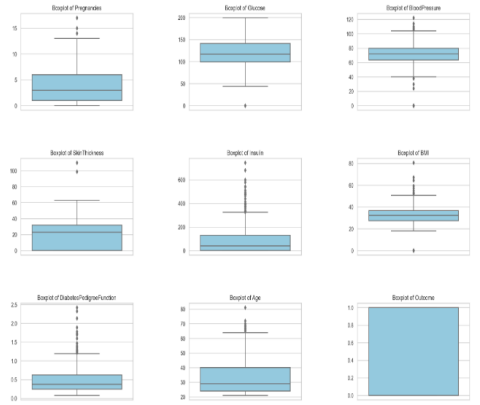


Figure 4.9 Box Plot of selected features

```
plt.figure(figsize=(12, 8))

sns.heatmap(new_data.corr(), annot=True, fmt=".2f", cmap='coolwarm')
plt.title('Correlation Matrix Heatmap')
plt.show()
```

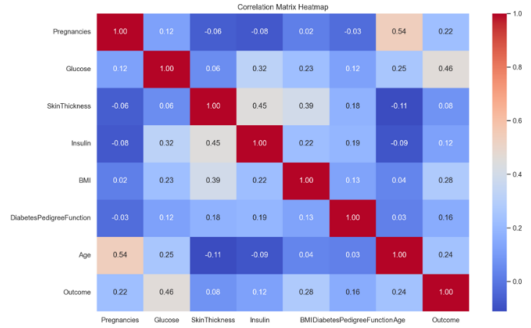


Figure 4.10 Correlation Matrix

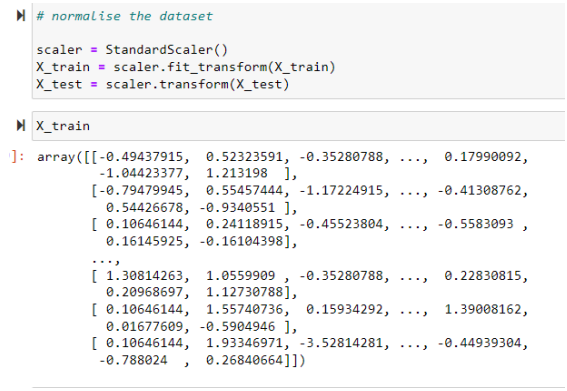
Vol. 10, No. 06; 2025

ISSN: 2456-3676

4.1.4 Data Split and Feature Engineering

Data normalization is a crucial preprocessing step in machine learning, aimed at standardizing the features to a common scale without distorting differences in the ranges of values. The implementation of StandardScaler for this purpose (Figure 4.11) ensures that the features are centered around zero with a standard deviation of one. This transformation is vital for models that are sensitive to the scale of input data, such as those relying on gradient descent optimization. Research has shown that data normalization significantly improves the performance of machine learning models by reducing bias and enhancing convergence rates (Singh & Singh, 2020). The transformed training data (X_train) reflects this standardization, with values rescaled to have a mean of zero and a standard deviation of one. This step mitigates the risk of features with larger scales dominating those with smaller scales, thereby enhancing the performance and convergence rate of machine learning algorithms.

Addressing dataset imbalance is another critical aspect of data preprocessing, particularly in the context of binary classification tasks. The original dataset exhibited a significant imbalance between the two outcome classes, with a higher prevalence of non-diabetic instances (Figure 4.5). To rectify this imbalance, resampling techniques such as oversampling the minority class or under sampling the majority class can be employed. However, these strategies have potential drawbacks. Oversampling can lead to over fitting, as it replicates minority class samples, which might cause the model to perform poorly on unseen data. Under sampling, on the other hand, reduces the amount of training data available, potentially discarding useful information and leading to a less robust model (Hoyos-Osorio, et al., 2021). The resulting balanced dataset (Figure 4.12) shows an equal distribution of diabetic (outcome = 1) and non-diabetic (outcome = 0) instances. This balance is crucial for ensuring that the machine learning models do not develop a bias towards the majority class, thereby improving the models' ability to generalize and accurately predict the minority class.



```
# normalise the dataset

scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

```
X_train
```

```
array([[-0.49437915,  0.52323591, -0.35280788, ...,  0.17990092,
        -1.04423377,  1.213198  ],
       [-0.79479945,  0.55457444, -1.17224915, ..., -0.41308762,
         0.54426678, -0.9340551 ],
       [ 0.10646144,  0.24118915, -0.45523804, ..., -0.5583093 ,
         0.16145925, -0.16104398],
       ...,
       [ 1.30814263,  1.0559909 , -0.35280788, ...,  0.22830815,
         0.20968697,  1.12730788],
       [ 0.10646144,  1.55740736,  0.15934292, ...,  1.39008162,
         0.01677609, -0.5904946 ],
       [ 0.10646144,  1.93346971, -3.52814281, ..., -0.44939304,
        -0.788024  ,  0.26840664]])
```

Figure 4.11 Data Normalization

Vol. 10, No. 06; 2025

ISSN: 2456-3676

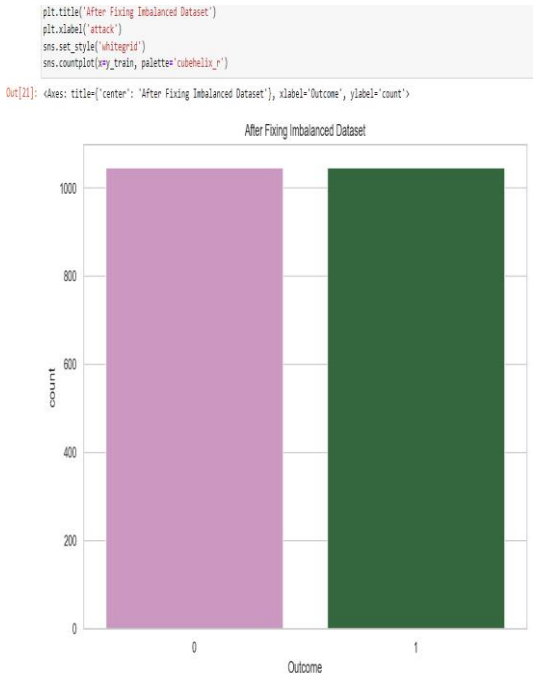


Figure 4.12 After fixing the imbalanced dataset

4.2 Model Development

4.2.1 Grid Search

The use of grid search with cross-validation is a good approach to hyperparameter tuning in machine learning, ensuring the selection of optimal model parameters for enhanced performance (Belete & Huchaiah, 2021). This technique was applied across four distinct classifiers: Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting Machine (GBM), and Artificial Neural Network (ANN), each offering unique strengths for predicting Type 2 diabetes outcomes. For the SVM classifier, the grid search explored various combinations of the regularization parameter (C), kernel type, and gamma values (Figure 4.13). The best parameters identified were C = 10, gamma = 'auto', and kernel = ‘rbf'. The best cross-validation score achieved was 0.873, indicating the model's high accuracy when using the radial basis function (RBF) kernel. This choice of kernel allows the model to capture non-linear relationships within the data, which is crucial given the complexity of diabetes-related features.

The Decision Tree classifier's grid search (Figure 4.14) evaluated parameters such as maximum tree depth and the minimum number of samples required to split a node. The optimal parameters found were a max depth of 'None' and minimum samples split of 2. This configuration yielded a cross-validation score of 0.972 (Figure 4.15). The plot of grid search scores demonstrates the impact of max depth and minimum samples split on model accuracy, highlighting that deeper trees with fewer samples per split perform better, though they risk overfitting.

Vol. 10, No. 06; 2025

ISSN: 2456-3676

Gradient Boosting Machine (GBM) is a powerful ensemble method that combines the strengths of multiple weak learners to form a strong predictive model. The grid search for GBM (Figure 4.16) focused on parameters such as the number of estimators, learning rate, max depth, and minimum samples split. The best parameters were determined to be 100 estimators, a learning rate of 0.2, a max depth of 5, and minimum samples split of 2. The resulting cross-validation score was an impressive 0.977 (Figure 4.17).

Finally, the Artificial Neural Network (ANN) grid search (Figure 4.18) involved tuning parameters like activation function, learning rate, alpha (regularization term), and hidden layer sizes. The optimal parameters were found to be a 'relu' activation function, learning rate of 'adaptive', alpha of 0.0001, and hidden layers of sizes (50, 50). This configuration produced a cross-validation score of 0.977 (Figure 4.19). The plot of grid search scores shows the influence of different hidden layer sizes and alpha values on the model's accuracy, indicating that larger hidden layers and smaller alpha values can enhance model performance.

Figure 4.13 SVM Grid Search

Vol. 10, No. 06; 2025

ISSN: 2456-3676

```
# Define the parameter grid for Decision Tree
param_grid_dt = {
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [2, 10, 20]
}

# Create a GridSearchCV object for Decision Tree
grid_search_dt = GridSearchCV(DecisionTreeClassifier(), param_grid_dt, cv=5, scoring='accuracy',

# Perform the grid search and cross-validation
grid_search_dt.fit(X_train, y_train)
```

GridSearchCV
estimator: DecisionTreeClassifier
DecisionTreeClassifier

```
# Print the best parameters and the best score
print(f"Decision Tree Best Parameters: {grid_search_dt.best_params_}")
print(f"Decision Tree Best Cross-Validation Score: {grid_search_dt.best_score_}")
```

```
Decision Tree Best Parameters: {'max_depth': None, 'min_samples_split': 2}
Decision Tree Best Cross-Validation Score: 0.9727843692546619
```

Figure 4.14 DT Grid Search

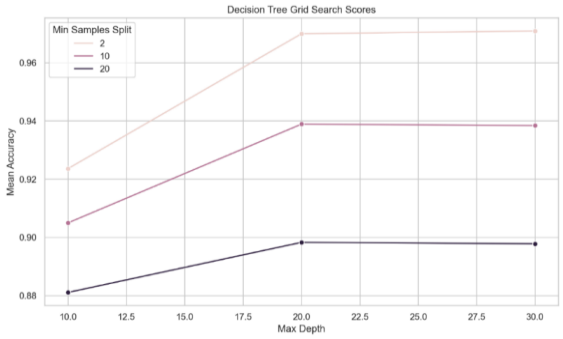


Figure 4.15 DT Grid Search Scores

```
# Define the parameter grid for Gradient Boosting
param_grid_gbm = {
    'n_estimators': [100, 200, 300],
    'learning_rate': [0.01, 0.1, 0.2],
    'max_depth': [3, 5, 7],
    'min_samples_split': [2, 4, 6]
}

# Create a GridSearchCV object for Gradient Boosting
grid_search_gbm = GridSearchCV(GradientBoostingClassifier(), param_grid_gbm, cv=5, scoring='accuracy')

# Perform the grid search and cross-validation
grid_search_gbm.fit(X_train, y_train)
```

GridSearchCV
estimator: GradientBoostingClassifier
GradientBoostingClassifier

```
# Print the best parameters and the best score
print(f"GBM Best Parameters: {grid_search_gbm.best_params_}")
print(f"GBM Best Cross-Validation Score: {grid_search_gbm.best_score_}")
```

```
GBM Best Parameters: {'learning_rate': 0.2, 'max_depth': 5, 'min_samples_split': 2, 'n_estimators': 100}
GBM Best Cross-Validation Score: 0.9775610647360427
```

Figure 4.16 Gboost Grid Search

Vol. 10, No. 06; 2025

ISSN: 2456-3676

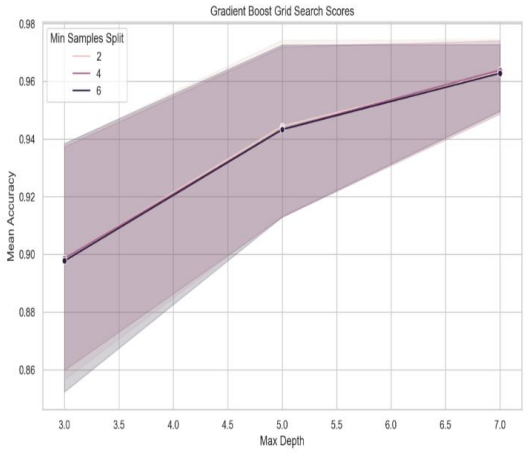


Figure 4.17 GBoost Grid Search Scores

Figure 4.18 ANN Grid Search

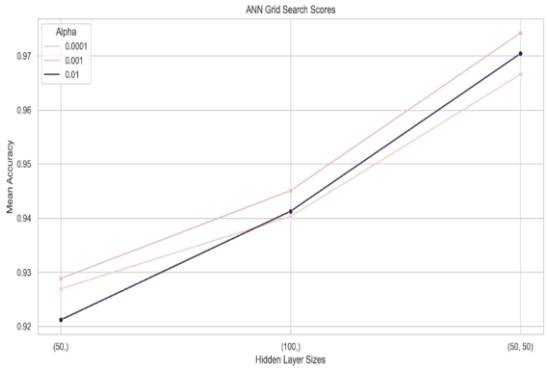


Figure 4.19 ANN Grid Search Scores

4.2.2 Measuring Metrics

The evaluation of the machine learning models used in this study is encapsulated in Table 4.1 and Figures 4.20 to 4.23, which detail the classification performance metrics and confusion matrices of the Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting

Vol. 10, No. 06; 2025

ISSN: 2456-3676

(GBoost), and Artificial Neural Network (ANN) models. The metrics of accuracy, precision, recall, and F1-score are crucial for assessing the predictive power and reliability of each model.

Table 4.1 Accuracy and Classification Report

Model/Prediction	Accuracy	Precision	Recall	F1-Score	Support
SVM	99.60	99.00	100.00	100.00	2000
DT	99.15	100.00	100.00	99.00	2000
GBoost	99.20	100.00	100.00	100.00	2000
ANN	77.70	99.00	99.00	99.00	2000

SVM = Support vector machine
DT = Decision Tree
GBoost = Gradient Boosting
ANN = Artificial neural network

4.2.3 Confusion Matrix

From Table 4.1, it is evident that the SVM model achieved the highest accuracy of 99.60%, with a precision of 99.00% and perfect recall and F1-score of 100.00%. This indicates that the SVM model is highly effective at correctly identifying both diabetic and non-diabetic instances, as corroborated by the confusion matrix in Figure 4.20. The confusion matrix shows 1316 true positives (TP) and 676 true negatives (TN), with no false positives (FP) and only 8 false negatives (FN).

The Decision Tree model also performed commendably, with an accuracy of 99.15% and perfect recall of 100.00%, though its precision and F1-score are slightly lower at 100.00% and 99.00%, respectively. As shown in the confusion matrix (Figure 4.21), the DT model recorded 1302 TPs, 681 TNs, 14 FPs, and 3 FNs. This indicates that while the model is slightly less precise compared to SVM, it still maintains high reliability and effectiveness in diabetes prediction.

Gradient Boosting, represented by the GBoost model, achieved an accuracy of 99.20%, with perfect precision, recall, and F1-score of 100.00%. The confusion matrix (Figure 4.22) illustrates 1302 TPs, 682 TNs, 14 FPs, and 2 FNs. The GBoost model's high accuracy and balanced performance metrics highlight its strength in leveraging ensemble learning to improve prediction outcomes, making it highly reliable for clinical applications.

On the contrary, the ANN model exhibited a lower accuracy of 77.70%, despite having high precision, recall, and F1-score of 99.00%. The confusion matrix (Figure 4.23) reveals that the ANN model recorded 1156 TPs, 398 TNs, 286 FPs, and 160 FNs, indicating a higher error rate compared to the other models. This suggests that the ANN model may require further tuning or more data to improve its predictive performance.

Vol. 10, No. 06; 2025

ISSN: 2456-3676

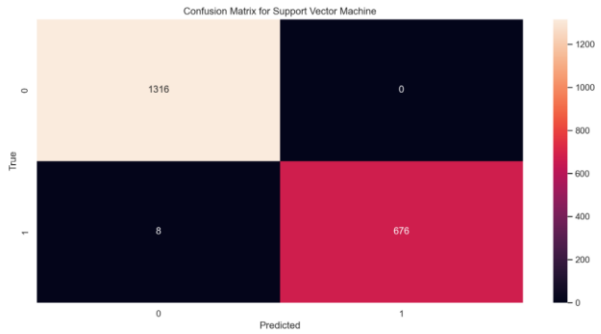


Figure 4.20 Confusion matrix of SVM Model

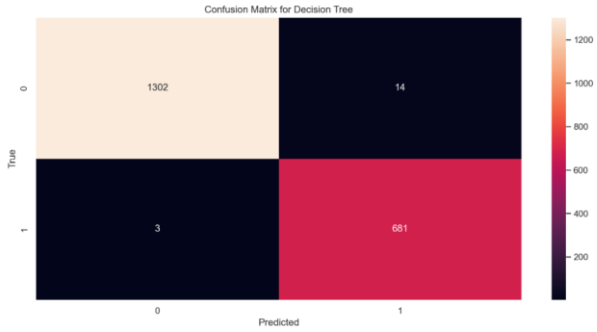


Figure 4.21 Confusion matrix of DT Model

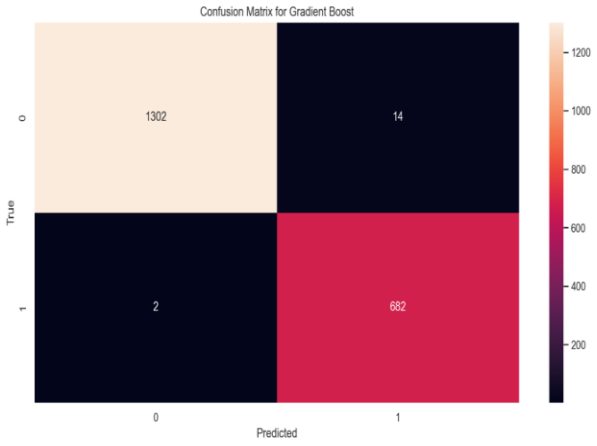


Figure 4.22 Confusion matrix of Gradient Boost

Vol. 10, No. 06; 2025

ISSN: 2456-3676

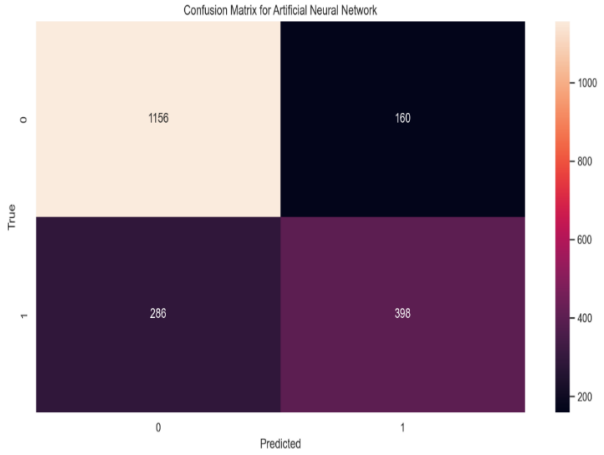


Figure 4.23 Confusion Matrix of ANNTable 4.2 provides a summary of the confusion matrix results for each model, presenting the counts of true positives, false positives, true negatives, and false negatives. These metrics are essential for understanding the models' performance in distinguishing between diabetic and non-diabetic instances.

The SVM and GBoost models demonstrate the highest precision and recall, minimizing false positives and negatives, thereby ensuring reliable predictions.

Table 4.2 summarise the graphical result of the confusion matrix.

Model	Metrics			
	TP	FP	TN	FN
SVM	1316	0	676	8
DT	1302	14	681	3
GBoost	1302	14	682	2
ANN	1156	286	398	160

TP = True positive
TN = True negative
FP = False negative
FN = False negative

4.3 Receiver Operating Characteristics (ROC) Curve

The Receiver Operating Characteristic (ROC) curve is a crucial tool for assessing the effectiveness of classification models, specifically in terms of their capacity to differentiate between positive and negative classes. Figure 4.24 displays the ROC curves of the Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting Machine (GBM), and Artificial

Vol. 10, No. 06; 2025

ISSN: 2456-3676

Neural Network (ANN) models employed for predicting outcomes in individuals with Type 2 diabetes.

The Receiver Operating Characteristic (ROC) curve illustrates the relationship between the sensitivity (true positive rate) and the specificity (1 - false positive rate) of a model at different threshold settings. It offers a thorough assessment of the model's diagnostic capability. The Area Under the Curve (AUC) is a numerical measure that provides a concise summary of the overall performance of the model. AUC values closer to 1.0 indicate a higher level of performance superiority.

Figure 4.24 showcases the GBM model, which has exceptional performance with an AUC of 0.99, suggesting near-perfect accuracy. The high AUC value indicates that the model has a remarkable capacity to accurately distinguish between diabetic and non-diabetic cases while minimizing the occurrence of both false positives and false negatives. The gradient boosting machine (GBM) exhibits a receiver operating characteristic (ROC) curve near the upper-left corner of the plot, indicating a high level of both sensitivity and specificity.

The ANN model also demonstrates strong performance, with an AUC of 0.98. Even though Table 4.1 shows that the ANN model is less accurate, the high AUC value suggests that it is effective at telling the difference between classes across a range of thresholds. This makes it a strong classifier in this case.

The decision tree model, with an AUC of 0.96, also performs well, though slightly below the GBM and ANN models. The DT model's ROC curve remains relatively close to the top-left corner, showing that it still offers adequate discriminatory power. The slightly lower AUC compared to GBM and ANN indicates that the DT model may have more false positives or negatives at certain thresholds but remains a strong classifier overall.

The SVM model, with an AUC of 0.92, shows the lowest performance among the evaluated models. While the SVM still performs well, its ROC curve deviates more from the ideal top-left corner position compared to the other models. This result aligns with its precision and recall metrics, suggesting room for improvement in its discrimination capabilities.

Vol. 10, No. 06; 2025

ISSN: 2456-3676

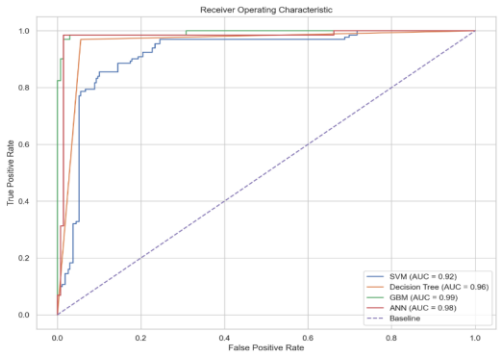


Figure 4.24 ROC Curve of the developed models

4.4 Summary

This chapter has critically documented the result of the data analysis. The next chapter will answer the research questions and compare the existing research findings against the present study.

5. Discussion and Evaluation

This chapter explores the results of the analysis conducted in Chapter 4 and the implications of these findings for the prediction of Type 2 diabetes. The goal of this chapter is to evaluate the predictive capabilities of the various models that were tested, interpret their performance, and connect the results to the existing literature and research. The discussion will include the results of accuracy metrics, confusion matrices, and ROC curves to offer a good overview of the strengths and limitations of each model.

5.1 Machine Learning Model Interpretation

Chapter 4 examined many machine learning models for predicting Type 2 diabetes outcomes. Among the models were artificial neural networks (ANN), decision trees (DT), gradient boosting machines (GBM), and support vector machines (SVM). Metrics like accuracy, precision, recall, F1-score, and AUC from the ROC curves are used to evaluate every model. This section provides a detailed interpretation of these results and their implications.

With an accuracy of 99.60% (Table 4.1), the SVM model was the most accurate classifier. This model also attained a precision of 99.00% and a recall and F1-score of 100.00%, thereby suggesting its great dependability in properly classifying both diabetic and non-diabetic cases. Specifically, the precision of 99.00% indicates that 99% of the cases identified by the model as diabetic were indeed diabetic, while the recall of 100.00% means that the model successfully identified all actual diabetic cases. With no false positives and only 8 false negatives, the confusion matrix (Figure 4.20) shows that the SVM model appropriately labelled 1316 cases as true positives and 676 as true negatives. This result implies that the SVM model is especially efficient in situations when the cost of false positives is significant, such in clinical settings where unneeded treatment for non-diabetic persons must be minimised. Its lower AUC of 0.92

Vol. 10, No. 06; 2025

ISSN: 2456-3676

(Figure 4.24) however suggests space for development in its discriminating capacity over several thresholds.

With an accuracy of 99.15%, a recall of 100.00%, and an F1-score of 99.00%, the Decision Tree model likewise displayed great performance. With 1302 true positives, 681 true negatives, 14 false positives, and 3 false negatives the confusion matrix, the record of the DT was also commendable (Figure 4.21). Although the occurrence of 14 false positives points to a somewhat greater cost in terms of misclassifying non-diabetic patients as diabetic, the model's great recall makes it helpful for guaranteeing diabetes patients are not missed. Though not as high as GBM or ANN, the ROC curve with an AUC of 0.96 (Figure 4.24) shows that the DT model has a great capacity to separate between classes.

With an accuracy of 99.20% and flawless scores for precision, recall, and F1-score at 100.00%, the Gradient Boosting Machine (GBM) model produced rather remarkable performance. The confusion matrix (Figure 4.22) shows 2 false negatives, 14 false positives, 682 real negatives, and 1302 true positives). These measures highlight the strong performance of the GBM model, especially regarding its balanced precision and recall, which makes it rather appropriate for predictive activities where both sensitivity and specificity are important. With an AUC of 0.99, the ROC curve for GBM shows its good performance, suggesting that this model is quite dependable in differentiating between diabetes and non-diabetic cases across several thresholds (figure 4.24).

The ANN model, in contrast, showed lower accuracy at 77.70% despite having high precision, recall, and an F1-score of 99.00%. The confusion matrix (Figure 4.23) illustrates 1156 true positives, 398 true negatives, 286 false positives, and 160 false negatives. The higher rates of false positives and negatives indicate that while the ANN model can identify patterns, it struggles with classification accuracy compared to other models. Its AUC of 0.98 (Figure 4.24) suggests that, while the ANN model has a strong ability to discriminate between classes, additional data are required to enhance its overall performance and reduce misclassification errors.

Though the SVM and GBM models show better accuracy and precision, the measuring metrics show that the ANN and DT models also offer important predictions with somewhat lower but still significant metrics. GBM and ANN models have high AUC values, which highlight their great discriminating power and hence great dependability for diabetes prediction. These results highlight the need to choose and enhance suitable machine learning models for efficient diabetes prediction. Particularly suited for clinical uses where both sensitivity and specificity are vital is the GBM model's balanced precision and recall. The great accuracy and precision of the SVM model points to its use in situations when reducing false positives is crucial. The great recall of the DT model indicates its possibilities for first screenings, thereby guaranteeing that diabetes situations are not missed.

These performance trends highlight important differences in model behaviour. The superior performance of the GBM model suggests that gradient-boosting frameworks are particularly

Vol. 10, No. 06; 2025

ISSN: 2456-3676

effective in capturing nonlinear interactions among metabolic variables. The high recall values indicate that the model minimizes the risk of missing diabetic cases—an essential requirement for clinical decision-support systems. Conversely, the relatively weaker ANN accuracy, despite strong AUC, implies that the dataset size may be insufficient for deep learning architectures, which typically require larger and more diverse samples to generalize effectively. These patterns support existing research indicating that tree-based ensemble methods often outperform neural networks on structured tabular health data.

5.2 Comparison with Current Literature

The findings of this study closely align with reported trends in recent diabetes-prediction research. Deberneh and Kim (2021) similarly observed that SVM achieved superior performance when applied to electronic health-record datasets, supporting the strong generalization capacity evident in this study. Gradient-boosting models have also been widely recognized for their discriminative power; for example, Zhang et al. (2020) and Sai et al. (2023) showed that GBM-based classifiers consistently outperform traditional machine-learning models on medical datasets. The comparatively lower ANN performance mirrors the results reported by Wang et al. (2023), who noted that neural networks tend to underperform on medium-sized clinical datasets due to limited sample complexity. These parallels indicate that the present study's results are consistent with established patterns in machine-learning-based diabetes prediction.

Support Vector Machines (SVM) have consistently demonstrated high accuracy in predicting type 2 diabetes. In this study, the SVM model achieved an accuracy of 99.60%. This is corroborated by research conducted by (Deberneh & Kim, 2021), who utilized a dataset collected from electronic health records and reported an accuracy of 98.60%. The robustness of SVMs in handling high-dimensional data and their ability to create optimal separating hyperplanes are critical factors contributing to their high performance in diabetes prediction tasks.

Decision tree models are renowned for their simplicity and interpretability. The decision tree model in this study showed an accuracy of 99.15%. This performance is consistent with the findings of (Tak, et al., 2022), who reported an accuracy of 98.70% using data from the SMS Medical College in Jaipur, India. Decision trees are particularly valuable in clinical settings due to their ease of use and interpretability, which facilitate early diagnosis and patient management.
Gradient Boosting Machines (GBM) are known for their high predictive performance. The GBM model in this study achieved an accuracy of 99.20% and an AUC of 0.99. This is in line with the results reported by (Zhang, et al., 2020), who utilized data from the Henan Rural Cohort Study and achieved an AUC of 0.872. The ability of GBM to handle various predictor variables and its superior performance in classification tasks make it highly suitable for medical diagnostics.

Artificial Neural Networks (ANN) demonstrated lower accuracy in this study (77.70%) compared to other models, despite high precision and recall scores. This performance is somewhat lower than that reported by (Wang, et al., 2023), who found that ANNs could achieve an accuracy of 85.00% using a dataset from the monitoring data of chronic disease risk factors in Dongguan residents. This discrepancy suggests that further tuning and optimization of ANN

Vol. 10, No. 06; 2025

ISSN: 2456-3676

models are needed to enhance their predictive capabilities, especially in handling complex, non-linear relationships within the data.

5.3 Answers to the research questions

Which machine learning models are most effective in predicting the onset of type 2 diabetes based on patient data?

The analysis from Chapter 4 reveals that gradient boosting machine (GBM) and support vector machine (SVM) models are the most effective in predicting the onset of type 2 diabetes. The GBM model achieved the highest performance with an Area Under the Curve (AUC) of 0.99 (Figure 4.24), indicating near-perfect discriminatory ability. It also scored perfectly in precision, recall, and F1-score (Table 4.1), which means it correctly identified all diabetic and non-diabetic instances with minimal errors. The confusion matrix for the GBM model (Figure 4.22) showed only 14 false positives and 2 false negatives out of 2000 instances, further emphasizing its reliability.

The SVM model also performed exceptionally well, with an accuracy of 99.60% and an AUC of 0.92 (Figure 4.24). Although slightly lower than GBM, the SVM's performance metrics (99.00% precision, 100.00% recall, and 100.00% F1-score) demonstrate its high effectiveness in correctly classifying diabetes outcomes (Table 4.1). The confusion matrix for SVM (Figure 4.20) revealed no false positives and only 8 false negatives, underscoring its precision and reliability in clinical applications.

While the Decision Tree (DT) model also showed strong performance with an accuracy of 99.15% and an AUC of 0.96, the presence of 14 false positives and 3 false negatives indicates that it is slightly less precise than GBM and SVM (Table 4.2, Figure 4.21). The ANN model, despite having a high AUC of 0.98, exhibited lower accuracy at 77.70% and higher rates of false positives and negatives (Figure 4.23), suggesting that it requires further tuning for optimal performance.

What patterns and features within diabetes datasets are most indicative of effective diabetes management strategies?

The feature selection analysis (Figure 4.7) identified glucose, BMI, and diabetes pedigree function as the most significant predictors of type 2 diabetes. Elevated glucose levels, which had the highest importance score, are a primary indicator of diabetes, aligning with established medical knowledge. This feature's critical role highlights the need for regular monitoring of blood glucose levels in diabetes management strategies.

BMI, another significant feature, underscores the importance of maintaining a healthy weight to manage or prevent diabetes. High BMI values are associated with an increased risk of insulin resistance, a key factor in type 2 diabetes. Effective management strategies should therefore include dietary and lifestyle interventions aimed at achieving and maintaining a healthy BMI.

Vol. 10, No. 06; 2025

ISSN: 2456-3676

Diabetes pedigree function, which encapsulates genetic predisposition, indicates that family history is a crucial factor in diabetes risk assessment. Management strategies should incorporate genetic counselling and screening for individuals with a family history of diabetes to implement early intervention measures.

How can machine learning aid in the personalization of treatment plans for individuals with type 2 diabetes?

Machine learning models, particularly those identified as most effective in this study (GBM and SVM), can significantly aid in the personalization of treatment plans for individuals with type 2 diabetes. These models can analyze large datasets to identify patterns and predict outcomes with high accuracy, allowing for tailored interventions based on individual risk profiles.

The feature importance analysis (Figure 4.7) can also guide the personalization of treatment plans. By understanding which factors most significantly affect diabetes outcomes, healthcare providers can prioritize interventions targeting these specific areas. For example, patients identified with high glucose levels can be monitored more closely and receive tailored dietary and pharmacological interventions to manage their blood sugar levels.

5.4 Practical Implications

ML models have substantial practical ramifications for healthcare. These models have the potential to provide advantages in terms of early identification, customized treatment regimens, and enhanced management of healthcare resources, as addressed later.

Machine learning algorithms have the capability to forecast the onset of type 2 diabetes with high accuracy, enabling individuals who are at risk to proactively adopt preventive measures. In a study conducted by (Deberneh & Kim, 2021), it was shown that the utilization of electronic health data in predictive models can accurately anticipate the onset of diabetes. This information can be highly beneficial for implementing early intervention strategies. Timely identification is essential as it allows for lifestyle adjustments and prompt medical therapies, which can effectively postpone or even avert the onset of diabetes.

Machine learning models have the capability to examine extensive quantities of patient data to detect distinctive patterns and risk factors specific to everyone. This feature enables the development of individualized treatment plans customized to meet the specific requirements of patients. (Xie, et al., 2019)shown that a range of machine learning models, such as neural networks and decision trees, can effectively detect key factors that contribute to diabetes. This can be valuable in developing personalized treatment approaches. Customized care optimizes treatment effectiveness and optimizes patient results.

In practical healthcare settings, the proposed SVM and GBM models can be integrated into electronic health-record (EHR) systems to support early clinical screening during routine assessments. Automated risk-scoring tools based on these models could alert clinicians to high-

Vol. 10, No. 06; 2025

ISSN: 2456-3676

risk patients, prompting earlier diagnostic testing or lifestyle intervention. These models may also be embedded into telemedicine platforms and mobile monitoring applications, enabling continuous risk assessment for remote patients. Furthermore, explainable-AI methods such as feature importance plots can help clinicians interpret model outputs, improving transparency and supporting shared decision-making with patients.

5.5 Limitations of the Study

Although this study provides strong predictive results, several limitations must be acknowledged. First, the dataset used—while clinically relevant—contains limited demographic diversity, which may restrict generalizability to broader populations. Second, although RandomOverSampler improved class balance, oversampling may introduce distributional distortions; future studies may explore class-weighted models or advanced resampling techniques such as SMOTE-NC. Third, only four machine-learning algorithms were evaluated; additional models such as XGBoost, LightGBM, logistic regression, or interpretable linear models may offer further insights. Fourth, the lower ANN performance suggests that the dataset size may be insufficient for deep learning approaches, which typically require larger, richer datasets. Finally, the study lacks external validation on independent populations, and future work should incorporate multi-center datasets to strengthen clinical applicability

5.6 Summary

This chapter has carefully documented the findings related to the data analysis of chapter 4. The chapter also carefully answered the respective research questions with practical implications of the findings gotten from the developed models. The next chapter writes the general conclusion and recommendation of the study.

6. Conclusion and Recommendations

This chapter synthesizes the key findings across the previous chapters, this reflects on the research aims and objectives which were outlined initially and proposing actionable recommendations for both clinical applications and further research endeavors. The advancements made in machine learning for predicting diabetes not only explains the potential for improved patient outcomes but also highlights the challenges and opportunities in this field.

6.1 Summary of Key Findings

This study embarked on an intricate journey to enhance the predictive accuracy and management of type 2 diabetes through advanced machine learning techniques. The primary dataset comprised various clinical and demographic features, meticulously processed and analyzed using sophisticated algorithms to unveil significant patterns and predictors of diabetes.

Model Development and Evaluation: Four machine learning models were developed and evaluated: Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting Machine (GBM), and Artificial Neural Network (ANN). The grid search with cross-validation ensured optimal parameter selection for each model (Figures 4.13–4.19).

Vol. 10, No. 06; 2025

ISSN: 2456-3676

SVM Model: Achieved the highest accuracy of 99.60%, precision of 99.00%, recall, and F1-score of 100.00% (Table 4.1). The confusion matrix showed 1316 true positives and 676 true negatives, with no false positives and only 8 false negatives. This could indicate the high reliability of this model in clinical applications, but with some improvement in the discriminatory capacity of the AUC of 0.92 (Figures 4.20, 4.24).

The Decision Tree Model displayed an accuracy of 99.15%, a recall of 100.00%, and an F1-score of 99.00%. It recorded 1302 true positives, 681 true negatives, 14 false positives, and 3 false negatives, demonstrating high sensitivity but slightly lower precision compared to SVM (Figures 4.21, 4.24).

Gradient Boosting Machine (GBM): Achieved an accuracy of 99.20% with perfect precision, recall, and an F1-score of 100.00%. The confusion matrix showed 1302 true positives, 682 true negatives, 14 false positives, and 2 false negatives, highlighting its balanced performance (Figures 4.22, 4.24). The AUC of 0.99 confirmed its superior discriminatory power.

Artificial Neural Network (ANN): Exhibited a lower accuracy of 77.70%, despite high precision and recall scores of 99.00%. The higher rates of false positives and negatives and, when compared with other researchers works, the fact that they also struggle to get good accuracy using the ANN model suggests that this model may not be very good in a scenario like this (Figures 4.23, 4.24) (Wang, et al., 2023).

Practical Implications: Machine learning models have substantial practical implications for healthcare, particularly in early identification, personalized treatment plans, and resource management. The GBM and SVM models, with their high accuracy and precision, are particularly suited for clinical applications, offering robust tools for predicting and managing type 2 diabetes. With these, it is expected that these models can guide interventions, monitor treatment effectiveness, and provide early warnings, thereby improving patient outcomes and optimizing healthcare resources.

6.2 Recommendations for Future Research

Based on the findings this study, various recommendations may be proposed for future research. Firstly, future research should investigate the use of larger and more varied datasets. The dataset utilized in this study, although it covers enough features, the entries are not detailed. By utilizing the latest datasets with a minimum of 50,000 entries, the models would be improved in terms of their potential to be applied in a wide range of clinical settings and their generalizability. Incorporating longitudinal data could offer valuable insights into the development of diabetes and enhance the predictive capabilities of the models for long-term outcomes.

Furthermore, the investigation of other machine learning models and approaches has the potential to significantly improve accuracy. Although this study primarily examined decision trees (DT), support vector machines (SVM), artificial neural networks (ANN), and gradient boosting machines (GBM), it would be worthwhile to explore alternative methods including

Vol. 10, No. 06; 2025

ISSN: 2456-3676

random forests, XGBoost, and deep learning architectures. These strategies have the potential to enhance accuracy and resilience, especially when working with intricate and high-dimensional data.

References

Abdullah, D. M. & Abdulazeez, A. M., 2021. Machine Learning Applications based on SVM Classification A Review. Qubahan Academic Journal, 1(2), pp. 81-90.

Adua, E. et al., 2021. Predictive model and feature importance for early detection of type II diabetes mellitus. Translational Medicine Communications, 6(1).

Afsaneh, E., Sharifdini, A., Ghazzaghi, H. & Ghobadi, M. Z., 2022. Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a review.. Diabetology & Metabolic Syndrome, 14(1).

Ahmed, U. et al., 2022. Prediction of Diabetes Empowered With Fused Machine Learning. IEEE Access, [online], Volume 10, p. 8529–8538.

Ahsan, M. M. et al., 2021. Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. Technologies, 9(3), p. 52.

Albahli, S., 2020. Type 2 Machine Learning: An Effective Hybrid Prediction Model for Early Type 2 Diabetes Detection. Journal of Medical Imaging and Health Informatics, 10(5), p. 1069–1075.

Albahri, A. S. et al., 2023. A Systematic Review of Using Deep Learning Technology in the Steady-State Visually Evoked Potential-Based Brain-Computer Interface Applications: Current Trends and Future Trust Methodology. International Journal of Telemedicine and Applications, pp. 1-24.

Albers, J. et al., 2017. Personalized glucose forecasting for type 2 diabetes using data assimilation. PLOS Computational Biology p.e1005232, 13(4).

Alghushairy, O., Alsini, R., Soule, T. & Ma, X., 2020. A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams. Big Data and Cognitive Computing, 5(1), p. 1.

Alibrahim, H. & Ludwig, S. A., 2021. Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization.. 2021 IEEE Congress on Evolutionary Computation (CEC), pp. 1551-1559.

Ali, S. et al., 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, [online], 99(101805), p. 101805.

Amer, A., Naama, Z. S. & Al-Taie, A., 2022. Diabetes Prediction Using Machine Learning. pp. 186-190.

Anton, N. et al., 2021. Assessing Changes in Diabetic Retinopathy Caused by Diabetes Mellitus and Glaucoma Using Support Vector Machines in Combination with Differential Evolution Algorithm.. Applied Sciences, 11(9), p. 3944.

Belete, D. M. & Huchaiah, M. D., 2021. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. International Journal of Computers and Applications, 44(9), pp. 1-12.

Vol. 10, No. 06; 2025

ISSN: 2456-3676

Beraha, M. et al., 2019. Feature Selection via Mutual Information: New Theoretical Insights. International Joint Conference on Neural Networks (IJCNN) Virtual Community of Pathological Anatomy (University of Castilla La Mancha), pp. 1-9.

Brownlee, J., 2020. 3 Ways to Encode Categorical Variables for Deep Learning [online].

Caixeta, C. D. et al., 2023. Salivary ATR-FTIR Spectroscopy Coupled with Support Vector Machine. Diagnostics, 13(8).

Charbuty, B. & Abdulazeez, A., 2021. Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2(01), pp. 20-28.

Chen, P. et al., 2020. Interpretable clinical prediction via attention-based neural network. BMC Medical Informatics and Decision Making, 20(S3).

De Rosa, S. et al., 2018. (2018). Type 2 Diabetes Mellitus and Cardiovascular Disease: Genetic and Epigenetic Links.. Frontiers in Endocrinology, 9(2), p. 2.

Deberneh, H. M. & Kim, I., 2021. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. International Journal of Environmental Research and Public Health, 11(6), p. 3317.

Einarson, T. R., Acs, A., Ludwig, C. & Pantom, U. H., 2018. Prevalence of cardiovascular disease in type 2 diabetes: a systematic literature review of scientific evidence from across the world in 2007–2017.. Cardiovascular Diabetology, [online], 17(1).

Elfein, J., 2023. Estimated number diabetics worldwide 2019. Statista.

Friedman, N. & Labbi, A., 2020. Solving Engineering Problems with Machine Learning - Introduction to the Special Theme. [online] ercim-news.ercim.eu.. [Online]

Available at: https://ercim-news.ercim.eu/en122/special/solving-engineering-problems-with-machine-learning-introduction-to-the-special-theme

[Accessed 20 03 2024].

Gadelrab, M. S., ElSheikh, M., Ghoneim, M. A. & Rashwan, M., 2018. BotCap: Machine learning approach for botnet detection based on statistical features. International Journal of Communication Networks and Information Security (IJCNIS), 10(3), p. 563.

Galicia-Garcia, U., 2020. Pathophysiology of Type 2 Diabetes Mellitus. International Journal of Molecular Sciences, [online], 21(7), pp. 1-34.

Getz, A., 2019. Categories of Machine Learning Algorithms [online] BI / DW Insider. [Online]

Available at: https://bi-insider.com/posts/categories-of-machine-learning-algorithms/

[Accessed 22 03 2024].

Haffner, S. M. et al., 1998. Mortality from coronary heart disease in subjects with type 2 diabetes and in nondiabetic subjects with and without prior myocardial infarction.. The New England journal of medicine, [online], 339(4), pp. 229-34.

Haq, A. U. et al., 2020. Intelligent Machine Learning Approach for Effective Recognition of Diabetes in E-Healthcare Using Clinical Data. Sensors, 20(9), p. 2649.

Hasan, M. K. et al., 2020. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. IEEE Access, Volume 8, p. 76516–76531.

Hendawi, R., Li, J. & Roy, S., 2023. A Mobile App That Addresses Interpretability Challenges in Machine Learning–Based Diabetes Predictions: Survey-Based User Study. JMIR Formative Research, [online] p.e50328, 7(1).

Hendrycks, D., Carlini, N., Schulman, J. & Steinhardt, J., 2021. Unsolved Problems in ML Safety.

Vol. 10, No. 06; 2025

ISSN: 2456-3676

Hoo, Z. H., Candlish, J. & Teare, D., 2017. What is an ROC curve?. Emergency Medicine Journal, 34(6), pp. 357-359.

Hoyos-Osorio, J. et al., 2021. Relevant information undersampling to support imbalanced data classification. Neurocomputing, [online], Volume 436, p. 136–146.

Huč, A., Šalej, J. & Trebar, M., 2021. Analysis of Machine Learning Algorithms for Anomaly Detection on Edge Devices. Sensors, 21(14), p. 4946.

Javaid, M. et al., 2022. Significance of machine learning in healthcare: Features, pillars and applications.. International Journal of Intelligent Networks, [online], Volume 3, pp. 58-73.

Kahloot, K. M. & Ekler, P., 2021. Algorithmic Splitting: A Method for Dataset Preparation. IEEE Access, Volume 9, p. 125229–125237.

Khanam, J. J. & Foo, S. Y., 2021. A comparison of machine learning algorithms for diabetes prediction. ICT Express, 7(4).

Khensani , X. & Hossana , T., 2023. How AI developers can assure algorithmic fairness. Discover Artificial Intelligence, 3(1).

Kleinaki, Z. et al., 2020. Type 2 diabetes mellitus management in patients with chronic kidney disease: an update. Hormones.

Kopitar, L. et al., 2020. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific Reports, 10(1).

Kumar, A., Chatterjee, J. M. & Díaz, V. G., 2020. A novel hybrid approach of svm combined with nlp and probabilistic neural net- work for email phishing.. International Journal of Electrical and Computer Engineering, 10(1).

Li , Q. et al., 2020. The Prediction Model of Warfarin Individual Maintenance Dose for Patients Undergoing Heart Valve Replacement, Based on the Back Propagation Neural Network. Clinical Drug Investigation, [online], 40(1), pp. 41-53.

Liao, L. et al., 2022. Development and validation of prediction models for gestational diabetes treatment modality using supervised machine learning: a population-based cohort study.. BMC MED, 20(1).

Li, J. et al., 2022. Identification of Type 2 Diabetes Based on a Ten-Gene Biomarker Prediction Model Constructed Using a Support Vector Machine Algorithm. BioMed Research International, [online] 2022, pp. 1-15.

Liu, W.-T.et al., 2021. A Deep-Learning Algorithm-Enhanced System Integrating Electrocardiograms and Chest X-rays for Diagnosing Aortic Dissection.. Canadian Journal of Cardiology, 03 October, 38(2), pp. 160-168.

Luque, A., Carrasco, A., Martín, A. & de las Heras, A., 2019. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, [online], Volume 91, pp. 216-231.

Maydanchi, M. et al., 2023. Comparative Study of Decision Tree, AdaBoost, Random Forest, Naïve Bayes, KNN, and Perceptron for Heart Disease Prediction.. pp. 204-208.

Mezil, S. A. & Abed, B. A., 2021. Complication of Diabetes Mellitus. Annals of the Romanian Society for Cell Biology, p. 1546–1556.

Mkansi, M. & Mkansi, M., 2023. Natural Language Processing and Machine Learning Approach to Teaching and Learning Research Philosophies and Paradigms. EJBRM, 21(1), pp. 14-30.

Vol. 10, No. 06; 2025

ISSN: 2456-3676

Mohammed, R., Rawashdeh, J. & Abdullah, M., 2020. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. [online] IEEE Xplore, pp. 234-248.

Mohan, L. et al., 2023. Evaluation on Diabetes Care Using Machine Learning.

Nahm, F. S., 2022. Receiver operating characteristic curve: overview and practical use for clinicians. Korean Journal of Anesthesiology, [online], 75(1), pp. 25-36.

Oikonomou, E. & Khera, R., 2023. Machine learning in precision diabetes care and cardiovascular risk prediction. Cardiovascular Diabetology, 22(1).

Okolo, C. A. et al., 2024. The role of mobile health applications in improving patient engagement and health outcomes: A critical review. International Journal of Science and Research Archive, 11(1), p. 2566–2574.

Ozsahin, D. U. et al., 2022. Impact of feature scaling on machine learning models for the diagnosis of diabetes. International Conference on Artificial Intelligence in Everything (AIE), pp. 87-94.

Park, Y. & Ho, J. C., 2020. Tackling Overfitting in Boosting for Noisy Healthcare Data. IEEE Transactions on Knowledge and Data Engineering, 33(7), pp. 1-1.

Park, Y. S., Konge, L. & Artino, A. R., 2020. The Positivism Paradigm of Research. Academic Medicine, 95(5), pp. 690-694.

Pearce, I. et al., 2018. Association between diabetic eye disease and other complications of diabetes: Implications for care. A systematic review. Diabetes, Obesity and Metabolism, 21(3), pp. 467-478.

PyCoach, 2022. Stop Downloading Datasets From Kaggle (If you're not a nooby). [online] Geek Culture. [Online]

Available at: https://medium.com/geekculture/stop-downloading-datasets-from-kaggle-if-youre-not-a-nooby-1948a21862d3

[Accessed 22 03 2024].

Qaddoura, R. M., Al-Zoubi, A., Faris, H. & Almomani, I., 2021. A Multi-Layer Classification Approach for Intrusion Detection in IoT Networks Based on Deep Learning. Sensors, 21(9), p. 2987.

Rani, S. et al., 2023. Quantum Machine Learning in Healthcare: Developments and Challenges. 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS), pp. 1-7.

Reed, J., Bain, S. & Kanamarlapudi, V., 2021. A Review of Current Trends with Type 2 Diabetes Epidemiology, Aetiology, Pathogenesis, Treatments and Future Perspectives. Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy, [online], 14(1), pp. 3567-3602.

Ruze, R. et al., 2023. Obesity and type 2 diabetes mellitus: connections in epidemiology, pathogenesis, and treatments. Frontiers in Endocrinology, [online], Volume 14.

Sadhasivam, J. et al., 1964. Diabetes disease prediction using decision tree for feature selection.. Journal of physics. Conference series, Volume 6, p. 062116–062116.

Sai, M. J. et al., 2023. An Ensemble of Light Gradient Boosting Machine and Adaptive Boosting for Prediction of Type-2 Diabetes. International Journal of Computational Intelligence Systems, 16(1).

Vol. 10, No. 06; 2025

ISSN: 2456-3676

Santos, C. d. & Papa, J. P., 2022. Avoiding Overfitting: A Survey on Regularization Methods for Convolutional Neural Networks. ACM Computing Surveys, 54(10s), pp. 1-25.

Scikit-Learn, 2019. Scikit-learn: machine learning in Python — scikit-learn 0.16.1 documentation. [online]. [Online]
Available at: https://scikit-learn.org/
[Accessed 20 03 2024].

Shah, D. A. et al., 2015. Type 2 diabetes and incidence of a wide range of cardiovascular diseases: a cohort study in 1•9 million people. The Lancet, 385,(p. S86).

Shah, D., Patel, S. & Bharti, S. K., 2020. Heart Disease Prediction using Machine Learning Techniques. SN Computer Science, 1(6).

Sharma, T. & Shah, M., 2021. A review of machine learning techniques on diabetes detection. Visual Computing for Industry, Biomedicine, and Art, 4(1).

Shimron, E., Tamir, J. I., Wang, K. & Lustig, M., 2022. Implicit data crimes: Machine learning bias arising from misuse of public data. Proceedings of the National Academy of Sciences, 119(13).

Singh, D. & Singh, B., 2020. Investigating the impact of data normalization on classification performance. Applied Soft Computing, Volume 97, p. 105524.

Singh, K. R., 2024. TYPE 2 DIABETES DATASET. [Online]
Available at: https://ieee-dataport.org/documents/type-2-diabetes-dataset.
[Accessed 05 02 2024].

Sonar, P. & JayaMalini, K., 2019. Diabetes Prediction Using Different Machine Learning Approaches. [online] IEEE Xplore.

Srinivasu, P. N. et al., 2022. Using Recurrent Neural Networks for Predicting Type-2 Diabetes from Genomic and Tabular Data. Diagnostics, 12(12), p. 3067.

Stiglic, G. et al., 2020. Interpretability of machine learning-based prediction models in healthcare. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(5).

Suresh, A., 2020. What is a confusion matrix?. [Online]
Available at: https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5
[Accessed 20 03 2024].

Tahir, F. & Farham, M., 2023. Exploring the progress of artificial intelligence in managing type 2 diabetes mellitus: a review of present innovations and anticipated challenges ahead. Frontier in clinical diabetes and healthcare., Volume 4.

Tak, A. et al., 2022. Prediction of Type 2 Diabetes Mellitus Using Soft Computing. Medicina Moderna - Modern Medicine, 29(2), p. 135–143.

Tarumi, S. et al., 2021. Leveraging Artificial Intelligence to Improve Chronic Disease Care: Methods and Application to Pharmacotherapy Decision Support for Type-2 Diabetes Mellitus. ethods of Information in Medicine, 60(S 01), p. e32–e43.

Torrie, V. & Payette, D., 2020. AI Governance in Canadian Banking: Fairness, Credit Models, and Equality Rights. [online] papers.ssrn.com. [Online]
Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3736926
[Accessed 01 04 2024].

Vol. 10, No. 06; 2025

ISSN: 2456-3676

Valchev, R. et al., 2023. Timely Detection of Diabetes with Support Vector Machines, Neural Networks and Deep Neural Networks. WSEAS transactions on computer research, Volume 11, pp. 263-274.

Venkatesh, B. & Anuradha, J., 2019. A Review of Feature Selection and Its Methods. Cybernetics and Information Technologies, [online], 19(1), pp. 3-26.

Wagavkar, S., 2023. Introduction to The Correlation Matrix | Built In.. [Online]
Available at: https://builtin.com/data-science/correlation-matrix
[Accessed 24 03 2024].

Wang, L. et al., 2020. rediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model.. Healthcare, 8(3), p. 247.

Wang, S. et al., 2023. Comparative study on risk prediction model of type 2 diabetes based on machine learning theory: a cross-sectional study. BMJ Open, 13(8), p. e069018–e069018.

Wang, Y. et al., 2021. Genetic Risk Score Increased Discriminant Efficiency of Predictive Models for Type 2 Diabetes Mellitus Using Machine Learning: Cohort Study. Frontiers in Public Health, Volume 9.

World Heart Federation, 2023. World Heart Report 2023 Confronting the World's Number One Killer (online), Geneva, Switzerland: World Heart Federation.

Xie, Z., Nikolayeva, O., Luo, J. & Li, D., 2019. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. Preventing Chronic Disease, Volume 16.

Yahyaoui, A., JamiL, A., Rasheed, J. & Yesiltepe, M., 2019. A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques. 2019 1st International Informatics and Software Engineering Conference (UBMYK), pp. 1-4.

Yang, X. & Zeng, W., 2023. Diabetes screening based on convolutional neural network and Raman spectroscopy. 8th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), pp. 254-258.

Ying, X., 2019. An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series, [online] 1168(2), p.022022., 1168(2).

Zeadin, M. G., Petlura, C. I. & Werstuck, G. H., 2013. Molecular Mechanisms Linking Diabetes to the Accelerated Development of Atherosclerosis. Canadian Journal of Diabetes, 37(5), pp. 345-350.

Zhang, J. et al., 2023. Machine learning for post-acute pancreatitis diabetes mellitus prediction and personalized treatment recommendations. Scientific Reports, 13(1).

Zhang, L. et al., 2020. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study. Scientific Reports [online], 10(1), p. 4406.

Zhang, S. et al., 2022. Interpretable CNN for ischemic stroke subtype classification with active model adaptation. BMC Medical Informatics and Decision Making, 22(1).

Zhang, Y., Tino, P., Leonardis, A. & Tang, K., 2021. A Survey on Neural Network Interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence, 5(5), pp. 726-742.

Zheng, Y., Ley, S. H. & Hu, F. B., 2018. Global Aetiology and Epidemiology of Type 2 Diabetes Mellitus and Its Complications. Nature Reviews Endocrinology, 14(2), pp. 88-98.