UTM Computing Proceedings
Innovations in Computing Technology and Applications
Volume: 6 | Year: 2024 | ISBN No. : 2976-2278

Identification of Important Risk Factors of Diabetes and Heart Disease using Feature Selection on Classification Algorithms

Nur Izzah Amira binti Ibrahim, Weng Howe Chan
Faculty of Computing
Universiti Teknologi Malaysia
Johor Bahru, Malaysia
nurizzahamira@graduate.utm.my, cwenghowe@utm.my

Abstract*—Diabetes mellitus, a chronic disease caused by high blood sugar levels, can lead to serious health problems like kidney failure and heart disease. Coexistence of diabetes and hypertension further increases the risk of heart disease, making it a leading cause of mortality and disability in individuals with diabetes worldwide. The study addresses the challenge of processing and mining knowledge from large and diverse medical data, which often contains irrelevant and redundant features. The selection of attributes and approaches for predicting illnesses among patients significantly influences the efficiency of data mining. Therefore, a robust framework for identifying significant risk factors for diabetes and heart disease was employed, where two types of feature selection methods were used: Filter methods (ANOVA and Chi-square) and Wrapper methods (Forward Feature Selection and Recursive Feature Elimination), along with classification algorithms Support Vector Machine (SVM), Random Forest (RF) and Decision Tree (DT). The dataset used in this research was obtained from the UCI Machine Learning Repository and consisted of two types of datasets: Diabetes Health Indicators dataset and Key Indicators of Heart Disease dataset. The results of the classification algorithms indicated that the SVM model with feature selection was the best-performing model with an accuracy of 83.6% for predicting diabetes cases. The best selected features that contributed to the performance of the model as the risk factors of diabetes were genetic health, mental health, physical health, BMI, income, and any health care. While the Decision Tree (DT) model with feature selection achieved the highest accuracy of 71.8% for heart disease cases, with the best features identified as risk factors for heart disease: stroke, mental health, physical health, age category, kidney disease, and genetic health. In conclusion, this research provided valuable insights into predicting diabetes and heart disease by identifying important risk factors. The study's findings could have implications in developing effective strategies for disease management and prevention.

Keywords-Diabetes, Heart disease, Feature selection.

I. INTRODUCTION

Diabetes mellitus is a long-term condition that can develop into serious health issues such as kidney failure, heart disease, and stroke. Diabetes increases the risk of heart disease by two to four times. Heart disease is the major cause of premature mortality. The risk of heart disease is greatly increased when hypertension and diabetes coexist. According to the World Health Organization (WHO), the number of individuals dying from heart disease would have increased to almost 30 million by 2040. As the prevalence of diabetes mellitus develops, the incidence of heart disease is expected to climb in the future years. As a result, massive amounts of clinical data are stored in biomedical devices and other hospital systems. However, identifying important risk factors can be a complex task due to the large number of potential factors that may influence disease outcomes. The selection of attributes and approaches for predicting illnesses among patients significantly influences the efficiency of data mining. The lack of appropriate identification and combination of important features for predictive models hinders the performance of forecasting models. According to Tang (2020), the redundant and unnecessary attributes will hinder data mining and analysis. The effectiveness of the classifiers is also reduced when irrelevant features are present according to Zulfiker (2021).

Therefore, the problem at hand was the presence of irrelevant and redundant features that impede the performance and efficiency of machine learning and data mining techniques. This can impede healthcare professionals' ability to effectively identify individuals at high risk of developing diabetes and heart disease or its associated complications, ultimately impacting timely interventions, treatment plans, and overall patient outcomes. The aim of this research is to identify which features (risk factor) have the most significant impact on predicting the presence or absence of diabetes and heart disease by employing selected feature selection methods and developing a classification model that can aid in early detection, prevention, and effective management of these conditions.

II. RELATED WORKS

Computational analysis has emerged as a valuable tool in the study of diabetes and heart disease risk factors, offering

insights into the complex interplay of various factors contributing to the development and progression of these conditions. By leveraging advanced techniques from the fields of data mining, machine learning, and statistical modelling, computational analysis enables researchers to uncover patterns, relationships, and predictive models that aid in understanding and predicting the risk factors associated with diabetes and heart disease. The application of computational analysis involves the integration and analysis of large and diverse datasets, including clinical records, genetic information, lifestyle factors, and medical imaging data. These datasets provide a wealth of information that can be utilized to identify and analyses the risk factors associated with diabetes and heart disease, paving the way for more accurate risk assessment and early detection. By employing sophisticated machine learning algorithms, computational analysis enables researchers to extract meaningful insights and build predictive models that can aid in risk assessment.

A. Machine Learning

Machine learning techniques have become increasingly utilized in the case study of diabetes and heart disease risk factors, offering valuable insights and predictive models to enhance our understanding and management of these conditions. By leveraging advanced algorithms and data analysis methods, machine learning enables the identification and classification of key risk factors, facilitating risk assessment, early detection, and personalized interventions. Supervised learning algorithms, such as support vector machines, decision trees, and neural networks, can be trained on labelled datasets to predict the risk of developing diabetes or heart disease based on specific risk factors. These algorithms can effectively analyze the relationships between various risk factors and outcomes, enabling the identification of important predictors and their impact on disease development.

Previous studies have demonstrated the potential of machine learning techniques in estimating the likelihood of an individual developing diabetes and heart disease based on various risk factors. In 2012 (Xue-Hui Meng, 2012), a study was conducted to predict diabetes using common risk factors. The study employed various classification techniques, including decision trees, Neural Networks, and logistic regression, to analyses their performance. Among these techniques, the logistic regression model demonstrated superior accuracy compared to the others. The study focused on common attributes such as family history, characteristics, and lifestyle risks as predictors for diabetes. A study in 2023 (S. V. Evangelin Soniaa, 2023), various classification techniques were evaluated to predict the future presence of cardiovascular disease (CVD) in diabetes patients over 10 years. Methods such as Decision Trees (DTs), K-Nearest Neighbor (KNN), Logistic Regression (LR), Artificial Neural Networks (ANNs), and Random Forest (RF) were compared. The results showed that LR had the highest accuracy of 84.4%, making it the optimal method for classification in this dataset. A multiple logistic regression classifier was used in a study from (Karolina Drożdż, 2022) to identify people who were most at risk of developing cardiovascular disease (CVD).

The classifier operated on patient parameters that were chosen based on their ability to discriminate through univariate feature ranking or extracted using principal component analysis (PCA). The study included 191 patients with metabolic-associated fatty liver disease (MAFLD), of which 25% had a history of CVD. Important clinical variables for predicting CVD risk included hypercholesterolemia, plaque scores, and duration of diabetes. In a 2019 study, machine learning techniques were used to examine different risk factors associated with diabetes. Four common machine learning algorithms, namely Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN), and C4.5 Decision Tree (DT), were employed to predict diabetes in adult population data. The results indicated that the C4.5 decision tree algorithm achieved higher accuracy compared to the other techniques.

B. Feature Selection

Feature selection is a critical step in machine learning for risk factor analysis. With numerous potential variables, feature selection techniques assist in identifying the most informative risk factors. These methods employ statistical approaches, information theory, or genetic algorithms to select the key predictors that significantly contribute to the risk of developing diabetes and heart disease. By focusing on these important features, machine learning models can improve their accuracy, interpretability, and predictive power. The goal is to reduce the dimensionality of the data by selecting a subset of features that capture the essential information while discarding irrelevant or redundant variables. By selecting the most relevant features, researchers and healthcare professionals can gain insights into the key variables that contribute to the development and progression of these diseases, leading to better prevention, early detection, and management strategies.

The risk factors for diabetes and heart disease can be ascertained by using a variety of feature selection techniques. The wrapper and filter approaches are two of the often-employed feature selection techniques. These methods employ different approaches to select the most relevant features for prediction or classification tasks. The wrapper technique evaluates the performance of a specific machine learning model by considering subsets of features. It involves iteratively selecting different subsets of features and training a model on each subset to assess its performance. The evaluation criterion, such as accuracy or cross-validation score, is used to determine the subset of features that yields the best model performance. This process is typically computationally intensive, as it requires training in multiple models, but it provides an accurate assessment of feature relevance. In the context of diabetes and heart disease risk factors, the wrapper technique would involve selecting subsets of features and evaluating the performance of a chosen machine learning model, such as logistic regression or random forest, on each subset.

The subset that achieves the highest performance metric would be considered the optimal set of risk factors. The filter technique, on the other hand, selects features based on their intrinsic properties without considering the performance of a specific machine learning model. It relies on statistical measures or predefined criteria to rank and select features that

are most relevant to the target variable. This technique is computationally less demanding than the wrapper method, making it suitable for high-dimensional datasets. In the case of diabetes and heart disease risk factors, the filter technique would involve applying statistical measures, such as correlation, information gain, chi-square, or mutual information, to evaluate the relationship between each feature and the target variable. Features that exhibit strong correlations or high information gain with the target variable would be selected as relevant risk factors.

In a study conducted by (Channabasavaraju & Vinayakamurthy, 2020), various approaches were developed for predicting diabetes and heart disease. The researchers utilized the Pima Indians Diabetes dataset and the heart disease dataset from UCI datasets to create a new combined dataset. The feature selection process involved using Recursive Feature Elimination (RFE) methods to validate the results. The dataset was divided into 70% for training and 30% for testing. The performance of RFE approaches with Artificial Neural Network (ANN), Fuzzy, and Support Vector Machine (SVM) models was evaluated using accuracy, sensitivity, specificity, and precision. The study focused on determining if diabetic patients would experience a heart attack or not. The RFE approach demonstrated a higher accuracy of 83.49% for the heart disease dataset, while the current method achieved an accuracy of 83%. A study by (Bagherzadeh-Khiabani et al., 2016) a clinical dataset of 803 pre-diabetic females with 55 characteristics was analyzed. The researchers explored different feature selection techniques, including wrapper and filter approaches, to predict diabetes mellitus (DM). The findings indicated that wrapper techniques yielded the best overall results. Among the filtering techniques tested, symmetrical uncertainty demonstrated the highest prediction accuracy. In the study conducted (Georga et al., 2015) various features, including Random Forest (RF) and ReliefF, were analyzed to predict transient subcutaneous glucose levels. Meanwhile, (Spencer et al., 2020) evaluated Principal Component Analysis, Chi-squared testing, ReliefF, and symmetrical uncertainty on four heart disease datasets. The authors found that different feature selection techniques had varying benefits depending on the machine learning method used for analyzing the cardiac datasets. One of the most accurate models achieved 85.0% accuracy, 84.73% precision, and 85.56% recall when combining Chi- squared feature selection with the BayesNet classifier.

III. METHODOLOGY

To ensure this study was conducted systematically, there are several phases that were done. The research framework has five phases which first consists of a literature review and problem identification activities. Second, data preparation which was data collection and data pre-processing. The third phase was featuring selection to execute feature selection methods and forth phase was classification algorithm. Finally, the last phase of this research framework was performance measures.

A. Phase 1: Research Planning

Activity 1: Literature review. Prior knowledge of how the investigation was carried out was crucial in any research. The literature review and problem identification, data collection and data pre-processing were the three activities that must be completed in phase 1. The basic concept of disease categorization and the procedures used to diagnose disease were discussed in the literature review. By evaluating prior related works done by other researchers, detailed information from previous studies was gathered to get a complete picture of the domain of the problem.

Activity 2: Problem identification. In this initial phase of the study, a thorough examination of the domain, algorithms, methodologies, and tools employed in previous research has been conducted. The purpose of this activity was to ensure a comprehensive understanding of the unsolved problem that the study aims to address. By analyzing existing literature and research, the specific problem to be tackled in this study has been identified. This step was crucial to ensure that the study was conducted with a clear grasp of the problem at hand, allowing for the development of appropriate solutions and contributions to the field.

B. Phase 2: Data Preparation

Activity 3: Data Collection. In data gathering, this research focuses on which datasets were to be applied. There are two datasets: one for diabetes and one for heart disease. The Diabetes Health Indicators dataset is used in this study. The CDC conducts an annual telephone survey on health-related topics called the Behavioral Risk Factor Surveillance System (BRFSS). Over 400,000 Americans participate in the survey each year, providing information on risky behaviors, chronic health issues, and usage of preventative treatments. Only 253,680 survey responses from the cleaned BRFSS 2015 dataset, which are 253,680 refers as rows and 22 features refers as columns, were included due to diabetes disease research regarding factors influencing diabetes disease and other chronic health conditions. The Key Indicators of Heart Disease dataset utilized in this study traces all the way from the Centers for Disease Control and Prevention (CDC). The CDC reports that heart disease is one of the leading causes of death for persons of most races in the US, including African Americans, American Indians, and Alaska Natives, as well as white people. The latest recent dataset (as of February 15, 2022) contained data from 2020. It has 279 columns and 401,958 rows, but only about 18 columns and 319,795 rows were given to it.

Activity 4: Data Pre-processing. Pre-processing the dataset was required for an accurate depiction of data quality. To clean up the data, pre-processing techniques including StandardScaler(SS) and MinMaxScaler were used to remove the missing characteristics (Ayon et al., 2020). A data preparation technique called missing value handling was used to build a smooth dataset. As a result, the initial step was to check for missing values in the dataset. Missing values may be ignored, substituted with any numeric value, substituted with the value that occurs the most frequently (the mode) for that feature or substituted with the mean value of the attribute.

C. Feature Selection

Activity 5: Employ feature selection. Finding the most important risk factors of diabetes and heart disease was the aim of feature selection. In this research, the behavior of numerous feature selection algorithms was evaluated across two primary categories (filter and wrapper). As shown in Figure 3.1, initial datasets were subjected to feature selection techniques from

two different categories. The creation of a subset was the first step in feature selection methods, but the type of approach determines how that subset was created (Dissanayake & Johar, 2021). The feature selection methods utilized are shown in Figure 1.

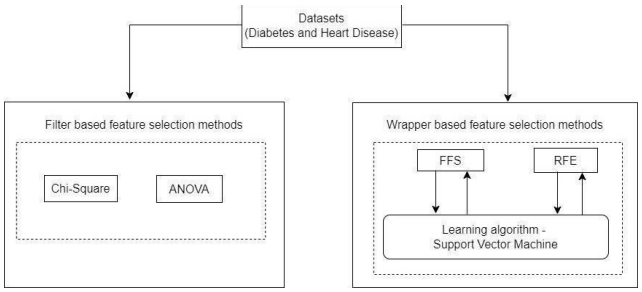

Figure 1 Feature selection methods

Feature Selection using Filter Methods: In this research, the Chi-square and ANOVA methods were chosen as the specific filter-based feature selection techniques. The Chi-square method is particularly useful for categorical data, as it measures the independence between two categorical variables and helps identify features that may have a significant relationship with the target variable. On the other hand, the ANOVA method is suitable for continuous data and assesses the variance between different groups, helping to identify features that show significant variations across various categories or classes.

Feature Selection using Wrapper Methods: In this research, two specific searching strategies were used with the wrapper method: forward feature selection and recursive feature elimination. Forward Feature Selection: This strategy starts with an empty subset and iteratively adds one feature at a time based on the algorithm's performance. At each step, the feature that contributes the most to the model's performance was selected and added to the subset. This process continues until a stopping criterion was met or no further improvement in performance was achieved. Recursive Feature Elimination (RFE): In RFE, the process begins with all features included in the subset. At each iteration, the least important feature was eliminated, and the model's performance was reevaluated. This process continues until the desired number of features is reached or the performance no longer improves. The wrapper methods are particularly useful for datasets with many features and have the advantage of considering feature interactions, which can lead to more accurate and robust models.

D. Phase 4: Classification

Activity 6: Build classification model. On both structured and unstructured data, classification is a way of categorizing data sets into various classes. Classification predictive modelling makes some attempts to map discrete input pieces to discrete output variables. There are various categorization methods accessible, however it is impossible to determine one is superior to the others. This is dependent on the issue domain and the nature of the dataset (Senan et al., 2021). The support vector machine and decision tree were supervised learning algorithms that were utilized in this study.

Support Vector Machine (SVM). SVM works by transforming the input data into a higher-dimensional feature space using a kernel function. In this transformed space, the algorithm finds an optimal hyperplane that maximizes the margin between the classes. The choice of kernel function determines the type of decision boundary that SVM can create, allowing it to capture complex relationships in the data. One of the key advantages of SVM is its ability to handle high-dimensional data and find non- linear decision boundaries. SVM is also less affected by overfitting compared to other algorithms, as it seeks to maximize the margin rather than fitting the training data exactly. Additionally, SVM can handle datasets with small ample sizes.

Decision Tree (DT). The decision tree algorithm makes decisions by following a flowchart-like structure, where each internal node evaluates a feature and determines the next node to visit based on the feature's value. This process continues until a leaf node is reached, which provides the final prediction or decision. Decision trees have several advantages. They are easy to understand and interpret, as the decision-making process is transparent and can be visualized. Decision trees can handle both numerical and categorical features, and they can capture non-linear relationships and interactions between variables. They can also handle missing values in the data. Overall, decision trees are widely used in various domains due to their simplicity, interpretability, and ability to manage both classification and regression tasks. They provide a powerful tool for decision-making and understanding complex relationships in data.

Random Forest (RF). Random Forest is an ensemble method that combines multiple decision trees to improve performance and reduce overfitting. Random Forest can handle both classification and regression problems effectively. It performs well with high-dimensional datasets and can handle many features without overfitting. Random Forest can capture complex relationships and interactions between features. It provides a measure of feature importance, allowing for feature selection and interpretation.

E. Phase 5: Evaluation Measures

Activity 7: Evaluate performance measures. Several assessment measures, including accuracy, sensitivity, and specificity, were utilized to examine the efficacy of the classification algorithm in this study using items from the confusion matrix. A classification result's performance rate was evaluated in Table 1. Table 1 shows the confusion matrix that has been used to calculate the performance of classifier, where TP: the number of instances correctly predicted as positive. FP: the number of instances incorrectly predicted as positive. FN: the number of instances incorrectly predicted as negative. TN: the number of instances correctly predicted as negative.

CONFUSION MATRIX TO BE USED FOR THE CLASSIFIERS PERFORMANCE.

		ACTUAL	
		Positive	**Negative**
PREDICTED	**Positive**	TP	FP
	Negative	FN	TN

Therefore, the classification accuracy has been calculated using equations as follows:

Sensitivity = TP * 100 / (TP + FN) (3.1)

Specificity = TN * 100 / (TN + FP) (3.2)

Accuracy = (TP + TP) / (TP + FP + TN + FN) * 100 (3.3)

IV. RESULT AND ANALYSIS

Three methods, Chi-Square, ANOVA, and RFE, utilized a fixed number of features, k=6 —were selected based on their ability to achieve higher accuracy compared to k=8 and k=10. As a result of the feature selection process, a subset of the original set of features has been obtained, which consists of the selected features. These selected features were considered to have the most significant impact on predicting diabetes. The subsequent step involved evaluating the performance of the model using these selected features to determine which combination of features resulted in the highest model accuracy. Table 2 presents the results of this evaluation, demonstrating the accuracy achieved by each feature selection method. The RFE method stood out by producing the highest accuracy when compared to the other methods. This indicates that RFE is effective in identifying the best selected features, which in turn represent the most important risk factors for diabetes. Overall, the study's findings in Table 2 demonstrate that by using RFE, the study successfully pinpointed the subset of features that have the most significant impact on predicting diabetes. These best selected features, (BMI, AnyHealthcare, GenHlth, MentHlth, PhysHlth, and Income) capture essential information and patterns related to the disease, making them valuable in enhancing our comprehension of diabetes risk factors. Understanding these risk factors can be instrumental in guiding healthcare professionals in identifying high-risk individuals, implementing early interventions, and designing personalized treatment plans.

Table II THE SELECTED FEATURES FOR DIABETES DATASET.

Feature Selection Methods	**Selected features (k=6)**	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**
Chi-Square	HeartDiseaseorAttack NoDocbcCost GenHlth MentHlth PhysHlth DiffWalk	80.9	73.0	93.1
ANOVA	BMI GenHlth MentHlth PhysHlth DiffWalk Income	83.0	73.0	93.1
FFS	PhysHlth MentHlth GenHlth Veggies DiffWalk Income BMI Fruits NoDocbcCost AnyHealthcare HvyAlcoholConsump CholCheck Stroke	82.9	73.0	93.1
RFE	BMI AnyHealthcare GenHlth MentHlth PhysHlth Income (Most important risk factors)	83.6	73.0	93.1

For the heart disease dataset, three methods also, Chi-Square, ANOVA and RFE, use a fixed number of features, k=6 —selected based on their ability to achieve higher accuracy compared to k=8 and k=10. Table 3 displays the evaluation results, indicating the accuracy achieved by each feature selection method. Surprisingly, the accuracy for each method appears to be identical. Despite this, the RFE method has been selected as the most notable approach due to its consistent performance across different numbers of selected features when compared to other methods. This observation highlights the effectiveness of RFE in identifying the best features, which are important risk factors for heart disease. The stability of RFE's performance regardless of the number of features selected reinforces its reliability in identifying important characteristics that contribute significantly to heart disease prediction. In conclusion, the study's findings from Table 3 highlight the success of utilizing RFE in identifying a subset of features with the most significant impact on predicting heart disease. These selected features, including Stroke, PhysicalHealth, MentalHealth, AgeCategory, KidneyDisease, and GenHealth, contain crucial information and patterns related to the disease, enhancing our understanding of heart disease risk factors. By leveraging these important risk factors, this research contributes

valuable insights that can lead to improved heart disease management and prevention strategies.

TABLE III THE SELECTED FEATURES FOR HEART DISEASE DATASET.

Feature Selection Methods	**Selected features (k=6)**	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**
Chi-Square	Stroke PhysicalHealth MentalHealth DiffWalking GenHealth	69.1	54.7	88.6
ANOVA	Stroke PhysicalHealth MentalHealth DiffWalking Diabetic GenHealth	68.6	54.7	88.6
FFS	Stroke PhysicalHealth MentalHealth Asthma PhysicalActivity SleepTime GenHealth	69.4	54.7	88.6
RFE	Stroke PhysicalHealth MentalHealth AgeCategory KidneyDisease GenHealth (Most important risk factors)	69.7	54.7	88.6

To assess the effectiveness of the best selected features, various evaluation metrics, including accuracy, sensitivity, and specificity, were utilized. The model classifiers were tested using the best features to determine how well they could predict the disease. The results obtained in Table 4 served as a crucial validation of the model's effectiveness in predicting disease based on the best selected features. Based on the accuracy results for diabetes dataset provided in Table 4, the Support Vector Machine (SVM) model with feature selection achieved the highest accuracy of 83.6%, compared to 83.2% for the Decision Tree (DT) model, 83.0% for the Random Forest (RF), and 82.7% for the SVM model without feature selection. The sensitivity for the SVM model with feature selection was 73.2%, slightly higher than the SVM without feature selection. This implies that the selected features aided in correctly identifying a higher percentage of diabetes cases. The specificity for the SVM model with feature selection was 94.8%, significantly higher than the SVM without feature selection. This suggests that the selected features contributed to better distinguishing non-diabetes cases. Therefore, the SVM model with feature selection was the best-performing model in terms of accuracy for predicting diabetes cases in this study. It outperformed both the DT model and the SVM model without feature selection, making it the preferred choice for diabetes prediction based on the provided evaluation metrics. Based on the accuracy results provided for the heart disease dataset, the Decision Tree (DT) model with feature selection achieved slightly higher accuracy of 71.8% compared to 71.7% for Random Forest (RF). This shows that the Decision Tree model, when given the relevant features, was more effective at predicting heart disease cases compared to the SVM models. However, the sensitivity of all models was relatively low. Sensitivity represents the model's ability to accurately identify positive heart disease cases. The low sensitivity values (ranging from 54.1% to 57.9%) indicate that the models have some difficulty in correctly classifying heart disease cases, leading to a relatively high number of false negatives. On the other hand, the specificity values (ranging from 78.5% to 88.6%) indicate that the models perform better in accurately identifying negative cases (non-heart disease cases). Overall, the Decision Tree model was the best performer in terms of accuracy for predicting heart disease cases.

TABLE IV RESULTS OF MODEL CLASSIFIERS FOR BOTH DATASETS.

Diabetes Health Indicators dataset				
Performance Measure	SVM (Without feature selection)	SVM	Decision Tree (DT)	Random Forest (RF)
Accuracy (%)	82.7	83.6	83.2	83.0
Sensitivity (%)	73.0	73.2	73.9	73.1
Specificity (%)	93.1	94.8	93.1	91.4
Heart Disease Personal Key dataset				
Performance Measure	SVM (Without feature selection)	SVM	Decision Tree (DT)	Random Forest (RF)
Accuracy (%)	68.3	69.7	71.8	71.7

Sensitivity (%)	57.9	54.1	54.7	55.1
Specificity (%)	78.5	85.0	88.6	88.0

V. DISCUSSION

The best features that contribute to the performance of models as the risk factor of diabetes are genetic health, mental health, physical health, BMI, income, and any health care. Genetic factors can play a significant role in diabetes risk (Genetic of Diabetes, 2023). Certain genes and family history can increase the likelihood of developing diabetes. Individuals with a family history of diabetes may have a higher genetic predisposition to the disease. Mental health can indirectly affect diabetes risk as stated in study (Tiziana Leone, 2012). Chronic stress, depression, and anxiety can influence behaviors such as unhealthy eating habits, lack of physical activity, and poor sleep patterns, which are all associated with an increased risk of developing diabetes. Physical health is strongly linked to diabetes risk. Regular physical activity and maintaining a healthy weight can help prevent or manage diabetes. Lack of exercise and being overweight, high BMI especially or obese are significant risk factors for type 2 diabetes. Socioeconomic factors, including income, can influence diabetes risk in study by (Selena E. Richards, 2022). Lower income individuals may have limited access to healthy food options, education about diabetes prevention, and healthcare resources. These factors can contribute to an increased risk of developing diabetes. Access to healthcare is crucial for managing and preventing diabetes. Regular check-ups, screenings, and diabetes management resources provided by healthcare professionals can help individuals control their blood sugar levels and prevent complications. It can be concluded that this is the most important risk factor in determining diabetes.

The best features that were identified as risk factors for heart disease: stroke, mental health, physical health, age category, kidney disease, and genetic health. Stroke is a serious medical condition that occurs when there is a disruption of blood flow to the brain. It is often caused by a clot or rupture of a blood vessel. Individuals who have had a stroke may be at a higher risk of heart disease because both conditions share common risk factors, such diabetes (Diabetes and Your Heart, 2022). Additionally, a history of stroke may indicate underlying cardiovascular issues that increase the likelihood of heart disease. Mental health plays a significant role in overall well-being, and there is a growing body of evidence linking mental health conditions, such as depression and anxiety, to heart disease (heart disease and Mental Health Disorders, 2020). Moreover, mental health issues can lead to physiological changes, including inflammation and hormonal imbalances, that may adversely affect the cardiovascular system. Good physical health is essential for heart disease prevention. Regular exercise and maintaining a healthy weight can help lower the risk of heart disease (Prevention Coronary Heart Disease, 2020). Physical activity improves cardiovascular fitness, reduces blood pressure, and lowers cholesterol levels, which are all crucial factors in heart health. Age is a well-known risk factor for heart disease. As people age, the risk of developing heart-related issues increases. This may be due to the natural aging process leading to changes in blood vessels and heart muscle, as well as the cumulative impact of other risk factors over time. The kidneys play a critical role in maintaining overall health, including heart health. Chronic kidney disease is associated with an increased risk of developing heart disease (Chronic Kidney Disease Initiative, 2022). Kidney dysfunction can lead to imbalances in electrolytes and fluid retention, which can strain the heart and contribute to the development of cardiovascular problems. Family history and genetics can influence an individual's risk of heart disease. If there is a family history of heart disease, individuals may have inherited certain genetic factors that predispose them to cardiovascular issues. It's important for individuals with a family history of heart disease to be vigilant about managing other risk factors and adopting a heart-healthy lifestyle.

VI. CONCLUSION

The research objectives have been successfully achieved, leading to valuable insights into the identification of important risk factors for both diabetes and heart disease. Through meticulous analysis and feature selection methods, the study identified the key risk factors that have a substantial impact on the occurrence and progression of these diseases. For diabetes, the critical risk factors were found to be genetic health, mental health, physical health, BMI, income, and any health care indicators. On the other hand, for heart disease, the prominent risk factors were stroke, mental health, physical health, age category, kidney disease, and genetic health. The selected features provided essential insights into the specific risk factors that play a crucial role in disease prediction. Classification models using two different algorithms, Support Vector Machine (SVM), Random Forest (RF) and Decision Tree (DT) were built using the best selected features from the feature selection process. By utilizing these models, the study ensured the creation of accurate and reliable tools for predicting diabetes and heart disease. The results indicated that SVM with feature selection yielded the highest accuracy for the diabetes dataset, while Decision Tree with feature selection achieved the best accuracy for the heart disease dataset. Overall, the research outcomes have significant implications for both medical and public health fields. The identified risk factors provide critical insights into the underlying causes and drivers of diabetes and heart disease, aiding healthcare professionals in better understanding and managing these conditions. The developed classification models can serve as valuable tools in clinical settings, supporting early diagnosis and personalized treatment strategies.

Expand the set of classifiers used in the analysis. Consider a range of classifiers, including but not limited to decision trees, random forests, and support vector machines. Select classifiers that are known for their performance in handling imbalanced datasets. Analyze and compare the performance of classifiers both on the imbalanced and balanced datasets. Evaluate their performance in terms of accuracy, precision, recall, F1-score, and other relevant metrics. Identify the classifiers that demonstrate robust performance across different evaluation

metrics and determine their suitability for handling imbalanced datasets.

ACKNOWLEDGMENT

The Universiti Teknologi Malaysia (UTM) supported this work through UTM Encouragement Research Grant with Cost Centre No. Q.J130000.3828.42J11.

REFERENCES

[1] Aggrawal, R., & Pal, S. (2020). Sequential feature selection and machine learning algorithm-based patient's death events prediction and diagnosis in heart disease. SN Computer Science, 1(6), 1–16.

[2] Alić, B., Gurbeta, L., & Badnjević, A. (2017). Machine learning techniques for classification of diabetes and cardiovascular diseases. 2017 6th Mediterranean Conference on Embedded Computing (MECO), 1–4.

[3] Almansour, N. A., Syed, H. F., Khayat, N. R., Altheeb, R. K., Juri, R. E., Alhiyafi, J., Alrashed, S., & Olatunji, S. O. (2019). Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study.Computers in Biology and Medicine, 109, 101–111.

[4] Ayon, S. I., Islam, M. M., & Hossain, M. R. (2020). Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. IETE Journal of Research, 1–20.

[5] Bagherzadeh-Khiabani, F., Ramezankhani, A., Azizi, F., Hadaegh, F., Steyerberg, E. W., & Khalili, D. (2016). A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results.Journal of Clinical Epidemiology, 71, 76–85. https://doi.org/10.1016/j.jclinepi.2015.10.002

[6] Birjais, R., Mourya, A. K., Chauhan, R., & Kaur, H.(2019). Prediction and diagnosis of future diabetes risk: a machine learning approach. SN Applied Sciences, 1(9), 1–8.

[7] Channabasavaraju, B. D., & Vinayakamurthy, U. (2020). An analysis of heart disease for diabetic patients using recursive feature elimination with random forest. Journal of Computer Science, 16(1), 105–116. https://doi.org/10.3844/jcssp.2020.105.116

[8] Collaboration, E. R. F. (2010). Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. The Lancet, 375(9733), 2215–2222.

[9] Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Medical Informatics and Decision Making, 19(1), 1–15. https://doi.org/10.1186/s12911-019-0918-5

[10] Dissanayake, K., & Johar, M. G. M. (2021). Comparative study on heart disease prediction using feature selection techniques on classification algorithms. Applied Computational Intelligence and Soft Computing, 2021. https://doi.org/10.1155/2021/5581806