International Journal on Advanced Science Engineering Information Technology

Vol.15 (2025) No. 3
ISSN: 2088-5334

A Review of Feature Selection Methods on Diabetes Mellitus Classification

Nur Farahaina Idris [a], Mohd Arfian Ismail [a,b,*], Shahreen Kasim [c], Rohayanti Hassan [d], Deshinta Arrova Dewi [e], Abdullah Munzir Mohd Fauzi [f], Rahmat Hidayat [g]

[a] *Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Pahang, Malaysia*
[b] *Center of Excellence for Artificial Intelligence & Data Science, Universiti Malaysia Pahang Al-Sultan Abdullah, Malaysia*
[c] *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Johor Malaysia*
[d] *Faculty of Computing, Universiti Teknologi Malaysia, Johor Malaysia*
[e] *INTI International University, Nilai, Negeri Sembilan, Malaysia*
[f] *MZR Global Sdn Bhd, Shah Alam, Selangor, Malaysia*
[g] *Department of Information Technology, Politeknik Negeri Padang, Sumatera Barat, Indonesia*
Corresponding author: [*]arfian@umpsa.edu.my

***Abstract*— Diabetes is a leading cause of death in the United States and leads to serious health complications. In recent decades, artificial intelligence technology and its subfield, machine learning, have been increasingly utilized to aid in disease diagnosis. Machine learning methods must be robust enough to handle the variability in diabetes datasets, which often encompass diverse patient demographics, clinical characteristics, and environmental factors. This motivates researchers to develop suitable feature selection methods that complement machine learning methods, thereby reducing time and complexity. However, feature selection may negatively impact classification accuracy by inadvertently removing essential features, or it may increase the time required due to repetitive processes during evaluation. Hence, thorough reviews of feature selection methods for diabetes classification are being conducted to evaluate their effectiveness. There are three primary categories of feature selection methods: embedded, wrapper, and filter methods. All the methods had distinct mechanisms and effects during the classification process. This study reviewed feature selection methods in each category, such as Random Forest from the embedded method, Chi-Square test from the filter method, and Recursive Feature Elimination from the wrapper method. The Chi-Square test is efficient only with categorical features, Random Forest is effective but causes high complexity and increased time due to its ensemble nature, and Recursive Feature Elimination produces the best performance but is not very suitable for data with high dimensionality. The findings indicate that Recursive Feature Elimination is more suitable for diabetes classification, as it is fast and yields good performance.**

***Keywords*—Feature selection; classification; random forest; recursive feature elimination; chi-square test; diabetes.**

Manuscript received 11 Aug. 2024; revised 21 Dec. 2024; accepted 24 Feb. 2025. Date of publication 30 Jun. 2025.

I. INTRODUCTION

Diabetes mellitus, commonly referred to as diabetes, is a prevalent chronic metabolic disorder affecting individuals worldwide. Historical evidence suggests that diabetes has been documented in ancient Egypt as far back as 3000 years ago, and it remains a significant global health challenge [1]. Consequently, extensive research has been conducted to address this issue. A significant milestone in diabetes research occurred in 1936 when a thorough investigation discovered a clear distinction between type 1 and type 2 diabetes [2]. Type 2 diabetes is primarily caused by insulin resistance and is commonly observed in older individuals. The previous studies identified that there is a high correlation between type 2 diabetes and ethnicity, family history, age, and obesity, despite the exact causes not yet being determined [3]. In contrast, type 1 diabetes results from insufficient insulin secretion and predominantly affects the younger population. Diabetes is not confined to a specific age group and often remains asymptomatic, earning its title as a 'silent' illness, with many individuals unaware of their condition [4]. Thus, the onset of diabetes-related complications can be sudden, making compliance with treatment and lifestyle changes challenging for affected individuals [5]. This disease affects individuals of all ages, leading to serious health complications

such as kidney disease, coronary heart disease, stroke, retinopathy, and even cancer [6]. It has been identified that diabetes becomes one of the highest causes of mortality and morbidity worldwide. Fig. 1 illustrates that diabetes was among the leading causes of death in the United States in 2022, based on data collected from the US Centers for Disease Control and Prevention (CDC). Due to its widespread prevalence and the severe health issues associated with it, there is a pressing need for effective methods to diagnose and manage diabetes.

In recent decades, various approaches within artificial intelligence have been utilized to address these problems, including machine learning methods [7]. These technologies are beneficial for analyzing complex datasets and facilitating rapid and accurate diagnosis [8]–[10]. Classification, a subfield of machine learning, has proven beneficial in the diagnostic process [11]. However, the effectiveness of classification can be significantly enhanced through feature selection methods, which streamline the dataset to focus on the most relevant features during the training process. Feature selection assists classifiers in producing better performance.

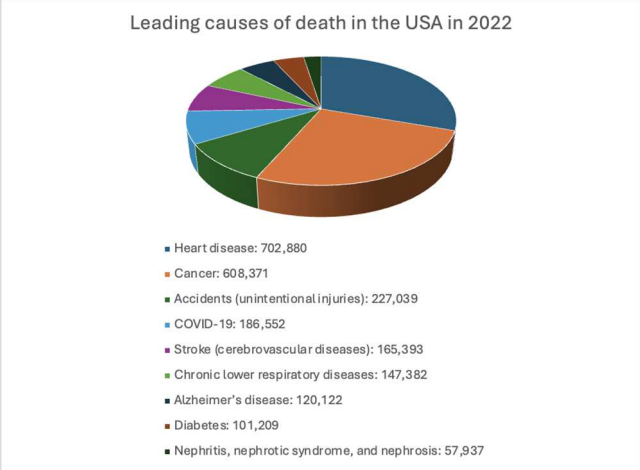

Fig. 1 Leading causes of death in the USA in 2022

The specific problem addressed in this study is to identify the best feature selection methods for diabetes classification in order to achieve better classification performance. Feature selection aims to improve classification performance by identifying and using only the most relevant features, thus reducing data complexity and potentially increasing accuracy [12]. This study reviews several more well-known feature selection methods, including Recursive Feature Elimination (RFE), Chi-Square test, and Random Forest, in order to determine which method works the best on diabetes classification [13]–[16].

Previous research has explored the application of feature selection classification with varied outcomes [17]. While these studies have demonstrated that feature selection can improve classification performance, the results have been inconsistent [18], [19]. Hence, some of the feature selection methods may not be suitable for the diabetes data and certain classifiers. This study builds upon these efforts by critically evaluating these feature selection methods to identify the most effective approach for diabetes classification.

II. Materials and Method

Machine learning is the induction of rules from observed data, but it can only be advantageous when the data features are informative [20], [21]. As it involves hundreds or thousands of features, the large number of features is not necessarily informative, as they are either irrelevant or redundant in relation to the class concept [22]. Feature selection is the most common technique to remove irrelevant features. The effects of redundant features are that they can extensively increase the computational and learning time, as a large number of features will clutter the learning process [23]. Typically, the feature selection method not only reduces training time but also reduces the dimensionality of the training data by removing irrelevant features and those with lower predictive power [24]. Various feature selections can be categorized into three categories: filter, wrapper, and embedded [25]. Overall, the crucial aspect of the feature selection method is to identify the level or rate the efficiency of the feature subset and use only the optimal features [26]. This study decides to uncover the effectiveness of all the categories of feature selection methods in diabetes by reviewing the most famous techniques in each category. The method involves RFE, which uses a wrapper method; the Random Forest, which uses an embedded method; and a Chi-Square test, which uses a filter method.

A. Recursive Feature Elimination (Wrapper Method)

The wrapper method searches for the optimal model by evaluating the model's performance for every conceivable combination of available features, like finding a search problem [27]. The objective of the wrapper is to identify the model with the highest performance [28]. One of the most well-known techniques from the wrapper method is RFE, which has been used in multiple studies related to diabetes [12], [29]. The first step for RFE is to train a machine-learning model. During the training of the model, rank the essential features. Secondly, it required the elimination of the least important feature. Thirdly, rebuild the model using the remaining features. Processes 1 to 3 would be repeated and continued until the desired number of features is achieved. The RFE process is depicted in Fig. 2. One of the unique characteristics of RFE is that it considers the interactions between the features in the dataset, unlike some other methods like the Chi-Square test and Analysis of Variance (ANOVA) from the filter method. Hence, it enables the redundant correlated features to be removed. RFE is more model-dependent, leading to higher flexibility and adaptability when used in conjunction with various machine learning methods, including linear methods such as Linear Support Vector Machine and Logistic Regression, as well as non-linear methods like Decision Tree and K-Nearest Neighbor.

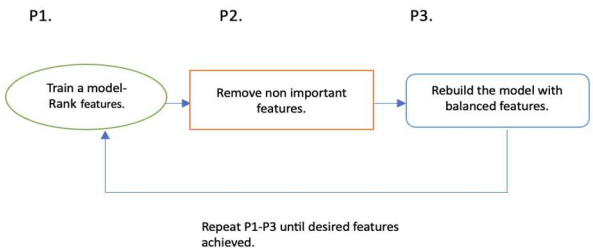

Fig. 2 RFE Process

Based on the study by [29], the Gated Recurrent Unit (GRU) with RFE managed to score an accuracy of 90.7% and an F1-score of 90.5% with the PIMA Indian Diabetes dataset, outperforming the GRU without RFE, which only acquired 87.65% accuracy and 87.61% F1-score. Then, it was identified that using RFE with Random Forest achieved an accuracy of 99% with the Iraqi Society Diabetes dataset, outperforming Random Forest with Genetic Algorithm, which only attained an accuracy of 94% [12]. A newer introduced method known as SRFEI that consists of RFE, stacking, and Isolation Forest acquired 79.077% test accuracy which is better compared to traditional stacking with only 76.363% using the PIMA Indian Diabetes dataset while the method that applied RFE also gained a better score of 97.466% accuracy in contrast to traditional stacking with 96.689% using diabetes prediction dataset [7]. The utilization of RFE in the diabetes domain simultaneously reduces the complexity of the method and improves the performance of the machine learning model, as it effectively eliminates unnecessary features [30]. The RFE method has been extensively researched for its straightforwardness, accessibility, and versatility. However, the utilization of RFE with high-dimensional data would be impractical as it would take more training time. Also, the existing studies identified that RFE utilization with high-dimensional data would make the model less robust [14].

B. Random Forest (Embedded Method)

Embedded algorithms have the feature selection process within the classifier training procedure, explicitly refining the feature set to attain the best performance [23]. As a result, they address the challenge of minimal optimality. The embedded methods are built-in feature selection methods. The advantage of using the embedded method as feature selection is that it is highly accurate and can be generalized better. Methods like Decision Trees and Random Forests are embedded methods that can be used as classification or additional steps for feature selection processes. This study will review the Random Forest as it is commonly studied in the diabetes domain and shows relatively good classification performance in most existing studies [31], [32]. Implementing a Random Forest is defined when only a random subset of features is selected when building each decision tree from the total features available. The randomness in the method (due to random subset features used) assists in decorrelating the individual trees and ensures diversity in the ensemble method. The construction process involves recursively splitting the data based on the Gini Index or Information Gain (splitting criteria). Fig. 3 illustrates the architecture of the Random Forest, which consists of multiple decision trees.

The top-ranked features in Random Forest are typically determined based on the importance scores of splitting criteria, which are calculated during the training phase. The algorithm assigns a critical value to each feature, ranking them accordingly. Higher-ranking features would be used as they are deemed to make more accurate predictions [33]. This selection process serves as an embedded feature selection mechanism, an integral part of the Random Forest method's training procedure.

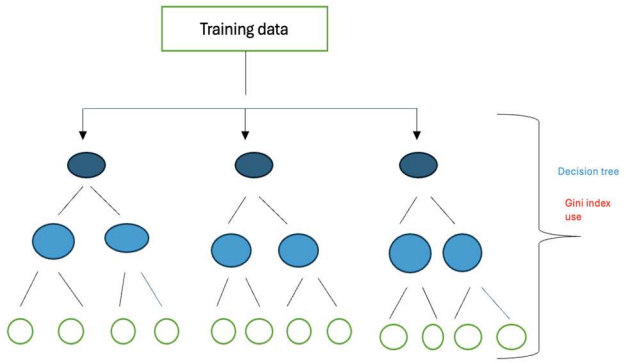

Fig. 3 Random Forest Architecture

Fig. 4 illustrates the importance levels of features as determined by Random Forest feature selection. The features in the root split are considered the most crucial as it is responsible for the first division of data that separates the data into different class groups. Features in level 1 splits of the decision tree are considered moderately important. As the level of the tree increases (deeper tree), the importance of the features decreases. The features in leaf nodes are commonly more specialized and granular. Features at the leaf nodes are considered the least important and influential in the hierarchy of feature importance derived by the Random Forest.

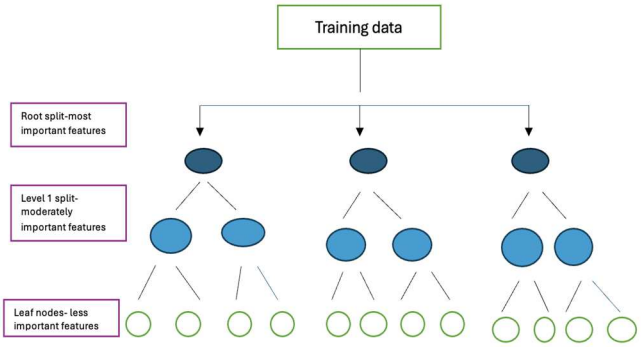

Fig. 4 Importance levels of features by Random Forest

The method proposed by [34] integrates fine-tuned K-Nearest Neighbor (FKNN) and Random Forest for feature selection. The method scored an accuracy of 90% and an F1 measure of 83.4% using the PIMA Indian Diabetes dataset, outperforming the Support Vector Machine, which obtained an accuracy of 83% and an F1 measure of 79%. According to a study by [35] that utilized data from physical examinations for diabetes (2010-2011), the use of only the top nine features subset, determined by Random Forest, managed to achieve a better AUC result of 0.828 compared to the use of 28 features with an AUC of 0.728. Not only that, the best nine features subsets consist of fasting blood glucose, triglycerides, age, urea nitrogen, low-density lipoprotein cholesterol, creatinine, alanine aminotransferase, high-density lipoprotein cholesterol, and mean platelet volume, which closely aligns with current medical studies. These attributes are considered some of the most critical factors in diagnosing diabetes. Thus, it shows the credibility of Random Forest, capable of optimally determining the best feature subset. This study further reviews another paper that applies Random Forest as

a feature selection method, using a dataset collected from diabetic hospital data in Sylhet, Bangladesh, which consists of 520 instances. However, using the Random Forest as a feature selection with Gradient Boosting classifier, the method acquired an accuracy of 97.69% and a time taken of 0.07563s when using nine features while with a total of 16 features had a 96.82% accuracy and time taken of 0.08278s [36]. Despite the count of features used in the classification of almost half, the time taken is not much different due to the high computational complexity of Random Forest during feature selection processing.

C. Chi-Square Test (Filter Method)

The filter method reviewed in this paper is the Chi-Square test, as it is commonly applied in the diabetes study [37]. The Chi-Square feature selection technique demonstrates efficacy in handling multiclass data by assessing the correlation strength of each feature through the Chi-Square distribution [38]. This method is grounded in hypothesis testing, wherein the initial step involves scrutinizing the deviation between actual and theoretical values. Subsequently, the analysis of this deviation reveals the correlation between the feature and the class. Each feature is assigned a Chi value during the classification process, and the algorithm starts with the assumption that each feature is independent of the class. The significance of the correlation between a feature and the class is reflected in the magnitude of the Chi value, with higher values indicating stronger correlations. Consequently, selecting the most important feature is determined by ordering Chi values from largest to smallest, with the feature exhibiting the highest Chi value considered the most influential [15]. Below is Equation 1 which depicts Pearson's Chi-Square test in which $\chi 2$ refers to Chi-square statistic or values, O refers to observed frequency in the table while E refers to expected frequency which is calculated under the assumption of independence between the variables and i is the specific index of the cell in the table. The expected frequency is shown using Equation 2 which consists of row total, column total and grand total elements. Overall, the steps start by calculating the expected frequencies, E for each category, then compute the differences between the observed and expected frequencies, square the differences, and divide by E. Finally, sum these values for all categories, and compare the results of Chi-Square values, $\chi 2$ to identify the best feature in the dataset.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{1}$$

$$E_i = \frac{(\text{row total} \times \text{column total})}{\text{grand total}} \tag{2}$$

The utilization of Random Forest and AdaBoost with the Chi-Square feature selection yielded an accuracy of 81% and 79%, respectively, while without the feature selection, it achieved an accuracy of 73% for both methods when tested with the PIMA Indian Diabetes dataset [39]. This shows the proficiency of the Chi-Square test across machine learning methods. Other findings collectively suggest that the Chi-square test for feature selection outperforms Information Gain, delivering higher accuracy while requiring less implementation time across both SVM and KNN scenarios, while also using the PIMA Indian diabetes [40]. The experimental results indicate that utilizing the Chi-square test for feature selection in conjunction with Support Vector Machine (SVM) yielded an accuracy of 88%, accomplished in a swift implementation time of 0.02 seconds. Additionally, employing the same Chi-square test with k-Nearest Neighbors (KNN) resulted in an accuracy of 84%, achieved within a short implementation time of 0.03 seconds. In comparison, when Information Gain was used in tandem with SVM, the accuracy achieved was 87%. The implementation time remained at 0.02 seconds, while KNN coupled with Information Gain exhibited an accuracy of 82%, and the implementation process took 0.02 seconds. The advantage of using this method is that it is fast and performs well. Nevertheless, it is suitable only with categorical features [41].

III. Results and Discussion

This study conducts a comprehensive comparative analysis based on the existing studies to identify the most suitable feature selection method for the diabetes domain. Various factors have been considered to know the most reliable methods to be integrated with machine learning methods (comparisons are being made using the performance with diabetes data), including the time taken, the effect on classification accuracy based on multiple existing diabetes studies, and suitability with large data. Table 1 presents a summary of the feature selection methods (RFE, Random Forest, and Chi-Square test). Table 2 presents the advantages of the feature selection methods, while Table 3 presents the disadvantages of the feature selection methods.

TABLE I
BRIEF SUMMARY OF THE FEATURE SELECTION METHODS

Method	Disadvantage
RFE	-Wrapper type -Model-dependent (handling feature interactions) -Moderate interpretability
Random Forest	-Embedded type -Model-dependent (handling feature interactions) -Lower interpretability (black box model)
Chi-Square test	-Filter type -Model-independent (not handling feature interactions) -High interpretability due to straightforward statistical computation

TABLE II
ADVANTAGES OF THE FEATURE SELECTION METHODS.

Method	Advantage
RFE	-Suitable to be applied to any classifier including ANN and ensembles [42]. -Superior performance in determining the best subset features in the diabetes data (outperforming SelectKBest) [43]. -Faster compared to the Chi-Square test [44].
Random Forest	-High accuracy and better generalization. Works well in determining the best features that should be used for classification [36]. -Suitable for small datasets and high dimensional data [45].
Chi-Square test	-Works well with large datasets [46]. -Simple and works well with multiclass data [38], [47]. -Lower computational complexity and fast processing.

TABLE III
DISADVANTAGES OF THE FEATURE SELECTION METHODS.

Method	Disadvantage
RFE	-Not suitable to use with the high dimensional data [14]. -Can cause overfitting [24].
Random Forest	-Method has high complexity that can lead to longer training time [45].
Chi-Square test	-Only suitable for categorical data [48]. -Not able to handle feature interactions, leading to the possibility of retaining redundant features [49].

Feature selection plays a crucial role in machine learning, aiding in the identification of relevant features that contribute to model performance while reducing computational complexity and training time, and also overcoming overfitting issues. While RFE, Random Forest, and the Chi-Square test stand out as popular choices, all of them have pros and cons, making the impacts of feature selection on each dataset may be considerably different.

RFE is a versatile feature selection method known for its ability to be applied to any classifier, including complex models like Random Forest and XGBoost ensemble methods [50], [51] One of its key advantages is its superior performance in determining the best subset of features for certain datasets, often outperforming simpler methods like SelectKBest. Additionally, RFE tends to be faster compared to alternatives like the Chi-Square test, making it suitable for various applications where computational efficiency is paramount. It is also suitable for use with any classifier and exhibits superior performance. However, RFE also has its drawbacks. Despite its versatility, RFE requires careful tuning to prevent the selection of an excessive number of features, which can degrade model performance on unseen data. It is also not well-suited for high-dimensional data, as the iterative process of feature elimination can become computationally expensive and prone to overfitting. However, it may not be an issue with diabetes classification, as the data typically does not have high dimensionality compared to domains such as gene expression.

Random Forest is a powerful ensemble learning method known for its high accuracy and robust generalization capabilities. In the context of feature selection, Random Forest excels at identifying the most informative features for classification tasks, making it a popular choice for both small datasets and high-dimensional data. Its ability to handle a large number of features while mitigating the risk of overfitting is particularly advantageous in complex modeling scenarios. Nevertheless, Random Forest comes with its own set of limitations. The method's high complexity can result in longer training times and is very computationally expensive, especially for large datasets with numerous decision trees. Additionally, while Random Forest performs well in many scenarios, its performance can be sensitive to hyperparameter settings, requiring careful optimization to achieve optimal results. Random Forest is the preferred choice when prioritizing accuracy over computational efficiency and time.

The Chi-Square test is a simple yet effective feature selection method, particularly well-suited for categorical data. It works well with large datasets, providing a straightforward means of identifying relevant features based on their association with the target variable. Moreover, the Chi-Square test is adept at handling multiclass data, making it a versatile option for a wide range of classification tasks. However, the Chi-Square test has its limitations, primarily stemming from its categorical nature. It is only applicable to categorical data and may not be suitable for continuous or mixed-type datasets, which are commonly used in the diabetes domain. While it offers simplicity and ease of interpretation, it may not capture more complex relationships between features and the target variable, limiting its utility in certain scenarios.

IV. CONCLUSION

This study presents an overview of prominent feature selection methods to determine which method yields the best accuracy and time efficiency for diabetes classification. The reviewed methods include the Chi-Square test, Random Forest, and RFE. In the analysis, the accuracy of the methods, their effects on training time, and their suitability for large datasets are critical components to consider when selecting a feature selection method. An in-depth study of these methods is crucial to ensure improvements in classification performance and reliability. This paper thoroughly discusses these feature selection methods, covering their advantages and disadvantages, particularly in the context of benchmark diabetes datasets. Based on the analysis, it is challenging to definitively determine which feature selection method is consistently the best for diabetes domain because each method has its own set of strengths and weaknesses.

Random Forest is highly accurate and suitable for most datasets but has a long training time and high complexity due to its ensemble nature. Aside from that, it is also a black-box model. In contrast, the Chi-Square test is more straightforward than the Random Forest. It works well with multi-class data but only supports categorical features, making it less relevant for continuous data commonly used in the diabetes domain. Lastly, RFE is suitable for most classifiers, works faster than the Chi-Square test based on some existing studies, and performs well in identifying the best subset of features. RFE is particularly suitable for benchmark diabetes datasets, such as the Pima Indians Diabetes dataset, or any real-world diabetes data that do not have an extremely high number of features (i.e., very high dimensionality).

Thus, this study identifies RFE as the best method overall, as it generally works quite fast based on existing studies and is deemed suitable for most diabetes data, which typically do not have an extreme number of features. RFE also supports the use of mixed data types and continuous numeric values commonly found in diabetes data. However, while RFE is selected as the best method among the three, the choice ultimately depends on the specific problem, situation, features, and data, as no technique performs excellently in every scenario Future research should explore the combination of feature selection methods with hyperparameter optimization, as tuning hyperparameters is essential for maximizing the performance of feature selection methods [52]. Thus, evaluating the optimal performance of these methods requires effective hyperparameter optimization [53]. The success in diabetes classification relies on both selecting the right feature selection method and optimizing hyperparameters effectively.

ACKNOWLEDGMENT

This study was supported by a Fundamental Research Grant (FRGS) with FRGS/1/2022/ICT02/UMP/02/2 (RDU220134) from the Ministry of Higher Education Malaysia. This work was supported by the Ministry of Higher Education, Malaysia (MOHE) under the Fundamental Research Grant Scheme (FRGS), FRGS/1/2023/ICT02/UTM/02/8.

REFERENCES

[1] W. Animaw and Y. Seyoum, "Increasing prevalence of diabetes mellitus in a developing country and its related factors," *PLoS One*, vol. 12, no. 11, pp. 1–11, 2017, doi:10.1371/journal.pone.0187670.

[2] A. B. Olokoba, O. A. Obateru, and L. B. Olokoba, "Type 2 diabetes: A review of current trends," *J. Clin. Med.*, vol. 7, no. 18, pp. 61–66, 2015, doi: 10.5001/omj.2012.68.

[3] A. P. Lovic, A. Piperidou, I. Zografou, and H. Grassos, "The growing epidemic of diabetes mellitus," *Curr. Vasc. Pharmacol.*, vol. 18, no. 2, 2020, doi: 10.2174/1570161117666190405165911.

[4] The Lancet Diabetes & Endocrinology, "Undiagnosed type 2 diabetes: An invisible risk factor," *Lancet Diabetes Endocrinol.*, vol. 12, no. 4, p. 215, 2024, doi: 10.1016/S2213-8587(24)00072-X.

[5] J. A. da Silva *et al*., "Diagnosis of diabetes mellitus and living with a chronic condition: Participatory study," *BMC Public Health*, vol. 18, no. 699, pp. 1–8, 2018, doi: 10.1186/s12889-018-5637-9.

[6] D. Tomic, J. E. Shaw, and D. J. Magliano, "The burden and risks of emerging complications of diabetes mellitus," *Nat. Rev. Endocrinol.*, vol. 18, no. 9, pp. 525–539, 2022, doi: 10.1038/s41574-022-00690-7.

[7] N. F. Idris *et al*., "Stacking with recursive feature elimination-isolation forest for classification of diabetes mellitus," *PLoS One*, vol. 19, no. 5, pp. 1–18, 2024, doi: 10.1371/journal.pone.0302595.

[8] K. Devasena and J. Shana, "Building machine learning model for predicting breast cancer using different regression techniques," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 1166, no. 1, Art. no. 012029, Jul. 2021, doi: 10.1088/1757-899X/1166/1/012029.

[9] S. Jebapriya, S. David, J. W. Kathrine, and N. Sundar, "Support vector machine for classification of autism spectrum disorder based on abnormal structure of corpus callosum," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 9, pp. 489–493, 2019, doi:10.14569/ijacsa.2019.0100965.

[10] D. Lavanya and K. U. Rani, "Performance evaluation of decision tree classifiers on medical datasets," *Int. J. Comput. Appl.*, vol. 26, no. 4, pp. 1–4, 2011, doi: 10.5120/3095-4247.

[11] V. O. Khilwani *et al*., "Diabetes prediction using stacking classifier," in *Proc. 2021 1st IEEE Int. Conf. Artif. Intell. Mach. Vis. (AIMV)*, 2021, pp. 1–6, doi: 10.1109/aimv53313.2021.9670920.

[12] X. Li, M. Curiger, R. Dornberger, and T. Hanne, "Optimized computational diabetes prediction with feature selection algorithms," *ACM Int. Conf. Proc. Ser.*, no. ML, pp. 36–43, 2023, doi:10.1145/3596947.3596948.

[13] Md. Maniruzzaman *et al*., "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm," *Comput. Methods Programs Biomed.*, vol. 152, pp. 23–34, Dec. 2017, doi:10.1016/j.cmpb.2017.09.004.

[14] B. F. Darst, K. C. Malecki, and C. D. Engelman, "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data," *BMC Genet.*, vol. 19, no. Suppl. 1, pp. 1–6, 2018, doi: 10.1186/s12863-018-0633-8.

[15] L. J. Cai, S. Lv, and K. B. Shi, "Application of an improved CHI feature selection algorithm," *Discret. Dyn. Nat. Soc.*, vol. 2021, 2021, doi: 10.1155/2021/9963382.

[16] M. A. M. Hasan, M. Nasser, S. Ahmad, and K. I. Molla, "Feature selection for intrusion detection using random forest," *J. Inf. Secur.*, vol. 7, no. 3, pp. 129–140, 2016, doi: 10.4236/jis.2016.73009.

[17] H. M. Farghaly and T. Abd El-Hafeez, "A high-quality feature selection method based on frequent and correlated items for text classification," *Soft Comput.*, vol. 27, no. 16, pp. 11259–11274, 2023, doi: 10.1007/s00500-023-08587-x.

[18] M. E. Cintra and H. A. Camargo, "Feature subset selection for fuzzy classification methods," in *Inf. Process. Manag. Uncertain. Knowl.-Based Syst.*, vol. 80, pt. 1, pp. 318–327, 2010, doi: 10.1007/978-3-642-14055-6_33.

[19] M. R. Mahmood, "Two feature selection methods comparison Chi-square and Relief-F for facial expression recognition," in *J. Phys.: Conf. Ser.*, vol. 1804, no. 1, Art. no. 012056, 2021, doi: 10.1088/1742-6596/1804/1/012056.

[20] H. Habehh and S. Gohel, "Machine learning in healthcare," *Curr. Genomics*, vol. 22, no. 4, pp. 291–300, 2021, doi:10.2174/1389202922666210705124359.

[21] M. Phongying and S. Hiriote, "Diabetes classification using machine learning techniques," *Computation*, vol. 11, no. 5, 2023, doi:10.3390/computation11050096.

[22] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, 2004, doi: 10.5555/1005332.1044700.

[23] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Front. Bioinform.*, vol. 2, no. June, pp. 1–17, 2022, doi: 10.3389/fbinf.2022.927312.

[24] N. M. Abdelwahed, G. S. El-Tawel, and M. A. Makhlouf, "Effective hybrid feature selection using different bootstrap enhances cancers classification performance," *BioData Min.*, vol. 15, no. 1, pp. 1–54, 2022, doi: 10.1186/s13040-022-00304-y.

[25] Y. Chen and Y. Zhong, "Improved filter method for feature selection," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 569, no. 5, Art. no. 052008, Aug. 2019, doi: 10.1088/1757-899X/569/5/052008.

[26] S. E. Awan, M. Bennamoun, F. Sohel, F. M. Sanfilippo, B. J. Chow, and G. Dwivedi, "Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death," *PLoS One*, vol. 14, no. 6, pp. 1–13, 2018, doi: 10.1371/journal.pone.0218760.

[27] S. Xia and Y. Yang, "A model-free feature selection technique of feature screening and random forest-based recursive feature elimination," *Int. J. Intell. Syst.*, vol. 2023, 2023, doi:10.1155/2023/2400194.

[28] E. Sreehari and L. D. D. Babu, "Critical factor analysis for prediction of diabetes mellitus using an inclusive feature selection strategy," *Appl. Artif. Intell.*, vol. 38, no. 1, 2024, doi:10.1080/08839514.2024.2331919.

[29] M. Y. Shams, Z. Tarek, and A. M. Elshewey, "A novel RFE-GRU model for diabetes classification using PIMA Indian dataset," *Sci. Rep.*, vol. 15, no. 1, pp. 1–22, 2025, doi: 10.1038/s41598-024-82420-9.

[30] R. K. Sachdeva, P. Bathla, P. Rani, V. Kukreja, and R. Ahuja, "A systematic method for breast cancer classification using RFE feature selection," in *Proc. 2022 2nd Int. Conf. Adv. Comput. Innov. Technol. Eng. (ICACITE)*, 2022, pp. 1673–1676, doi:10.1109/icacite53722.2022.9823464.

[31] S. Raghavendra and S. K. J, "Performance evaluation of random forest with feature selection methods in prediction of diabetes," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 1, pp. 353–359, 2020, doi:10.11591/ijece.v10i1.pp353-359.

[32] A. A. Alhussan *et al*., "Classification of diabetes using feature selection and hybrid Al-Biruni Earth Radius and Dipper Throated optimization," *Diagnostics*, vol. 13, no. 12, pp. 1–40, 2023, doi:10.3390/diagnostics13122038.

[33] R. Natras, B. Soja, and M. Schmidt, "Ensemble machine learning of random forest, AdaBoost and XGBoost for vertical total electron content forecasting," *Remote Sens.*, vol. 14, no. 15, pp. 1–34, Aug. 2022, doi: 10.3390/rs14153547.

[34] S. Ramya, T. Vijayaraghavan, and D. Kalaivani, "Diabetic prediction using feature selection-based random forest and fine-tuned K-nearest neighbor classifier algorithm—A design thinking approach," in *Proc. 2023 4th Int. Conf. Electron. Sustain. Commun. Syst. (ICESC)*, 2023, pp. 1303–1309, doi: 10.1109/icesc57686.2023.10193333.

[35] S. Lin, W. Ji, and J. Pei, "A method for selecting diabetes features based on random forest," *J. Phys.: Conf. Ser.*, vol. 1237, no. 2, Art. no. 022123, 2019, doi: 10.1088/1742-6596/1237/2/022123.

[36] S. Gündoğdu, "Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique," *Multimed. Tools Appl.*, vol. 82, no. 22, pp. 34163–34181, 2023, doi:10.1007/s11042-023-15165-8.

[37] P. Rajendra and S. Latifi, "Prediction of diabetes using logistic regression and ensemble techniques," *Comput. Methods Programs Biomed. Updat.*, vol. 1, p. 100032, 2021, doi:10.1016/j.cmpbup.2021.100032.

[38] I. S. Thaseen and C. A. Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi-class SVM," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 29, no. 4, pp. 462–472, 2017, doi:10.1016/j.jksuci.2015.12.004.

[39] V. Rupapara, F. Rustam, A. Ishaq, E. Lee, and I. Ashraf, "Chi-square and PCA-based feature selection for diabetes detection with ensemble

classifier," *Intell. Autom. Soft Comput.*, vol. 36, no. 2, pp. 1931–1949, 2023, doi: 10.32604/iasc.2023.028257.

[40] A. S. Jaddoa and Z. T. M. Al-Ta'i, "Diagnosis of diabetes mellitus using (chi square-information gain) selectors and (SVM and KNN) classifiers," in *Proc. 1st Int. & 4th Local Conf. Pure Sci. (ICPS)*, 2023, doi: 10.1063/5.0102761.

[41] L. A. S. Cardona, H. D. Vargas-Cardona, P. N. González, D. A. C. Peña, and Á. Á. O. Gutiérrez, "Classification of categorical data based on the chi-square dissimilarity and t-SNE," *Computation*, vol. 8, no. 4, pp. 1–15, 2020, doi: 10.3390/computation8040104.

[42] A. B. Pillay, D. Pathmanathan, A. Abu, and H. Omar, "RFE-based feature selection to improve classification accuracy for morphometric analysis of craniodental characters of house rats," *Sains Malaysiana*, vol. 52, no. 7, pp. 1901–1914, 2023, doi: 10.17576/jsm-2023-5207-01.

[43] S. Srivatsan and T. Santhanam, "Early onset detection of diabetes using feature selection and boosting techniques," *ICTACT J. Soft Comput.*, vol. 12, no. 1, pp. 2474–2485, 2021, doi:10.21917/ijsc.2021.0344.

[44] Alifah, T. Siswantining, D. Sarwinda, and A. Bustamam, "RFE and chi-square based feature selection approach for detection of diabetic retinopathy," in *Proc. Int. Joint Conf. Sci. Eng. (IJCSE 2020)*, 2020, no. Feb. 2021, doi: 10.2991/aer.k.201124.069.

[45] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J. M. Lagniez, and P. Marquis, "Trading complexity for sparsity in random forest explanations," in *Proc. 36th AAAI Conf. Artif. Intell. (AAAI)*, vol. 36, 2022, pp. 5461–5469, doi: 10.1609/aaai.v36i5.20484.

[46] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, "Feature selection using an improved Chi-square for Arabic text classification," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 32, no. 2, pp. 225–231, Feb. 2020, doi: 10.1016/j.jksuci.2018.05.010.

[47] W. H. Nugroho, S. Handoyo, Y. J. Akri, and A. D. Sulistyono, "Building multiclass classification model of logistic regression and decision tree using the Chi-square test for variable selection method," *J. Hunan Univ. Nat. Sci.*, vol. 49, no. 4, pp. 172–181, 2022, doi:10.55463/issn.1674-2974.49.4.17.

[48] Vikas and P. Kaur, "Lung cancer detection using Chi-square feature selection and support vector machine algorithm," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 10, no. 3, pp. 2050–2060, Jun. 2021, doi:10.30534/ijatcse/2021/801032021.

[49] M. L. Mchugh, "The Chi-square test of independence," *Biochem. Medica*, vol. 23, no. 2, pp. 143–149, 2013, doi:10.11613/bm.2013.018.

[50] W. Li *et al.*, "Predictive model and risk analysis for diabetic retinopathy using machine learning: A retrospective cohort study in China," *BMJ Open*, vol. 11, no. 11, pp. 1–11, 2021, doi:10.1136/bmjopen-2021-050989.

[51] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Assessing feature selection method performance with class imbalance data," *Mach. Learn. Appl.*, vol. 6, p. 100170, 2021, doi:10.1016/j.mlwa.2021.100170.

[52] M. H. Rizky, M. R. Faisal, I. Budiman, D. Kartini, and F. Abadi, "Effect of hyperparameter tuning using random search on tree-based classification algorithm for software defect prediction," *IJCCS (Indones. J. Comput. Cybern. Syst.)*, vol. 18, no. 1, pp. 95–104, 2024, doi: 10.22146/ijccs.90437.

[53] B. H. Shekar and G. Dagnew, "Grid search-based hyperparameter tuning and classification of microarray cancer data," in *Proc. 2nd Int. Conf. Adv. Comput. Commun. Paradigms (ICACCP)*, 2019, doi:10.1109/icaccp.2019.8882943.