Diabetes Prediction Using Machine Learning: Methods, Challenges, and Insights from a Systematic Literature Review (SLR)

Mustafa M. Abd Zaid
College of Technical Engineering, Islamic University, Najaf, Iraq
mustafamajeed2014@gmail.com
Orcid.org/0000-0001-8116-5170

Ahmed Abed Mohammed
College of Computer Science and Information Technology, University of Al-Qadisiyah, Diwaniyah, Iraq
a.alsherbe@qu.edu.iq
Orcid.org/0009-0002-2485-4230

DOI: http://dx.doi.org/10.31642/JoKMC/2018/120204
Received Nov.13, 2024. Accepted for publication Apr. 2, 2025

***Abstract*—** *This study systematically reviews and critically evaluates the current state of machine learning models for diabetes prediction, addressing key methodologies, challenges, and insights. Despite the growing body of research, a comprehensive systematic literature review (SLR) on this topic has been lacking. By consulting five major scientific databases (IEEE, ScienceDirect, Springer, Scopus, and ACM), this paper offers an in-depth analysis of existing studies' strengths, limitations, and research gaps. Key challenges discussed include data collection and preprocessing, handling missing values, feature importance assessment, standardization, and addressing class imbalance in datasets. Additionally, the review identifies underexplored areas and highlights opportunities for future research, such as developing standardized preprocessing frameworks and exploring advanced hybrid models. This SLR aims to guide researchers by summarizing existing evidence, resolving conflicts in the literature, and providing actionable directions for advancing diabetes prediction through machine learning.*

***Keywords*— Diabetes Prediction; Machine Learning; SLR; Research Gaps; Data Preprocessing.**

I. INTRODUCTION

Diabetes is a chronic and serious health condition that affects millions of people worldwide. According to the World Health Organization (WHO), over 463 million individuals were diagnosed with diabetes as of 2019, a number projected to rise significantly to 10.2% of the global population by 2030 and 10.9% by 2045 [1].

Diabetes mellitus (DM) is characterized by high blood glucose levels (hyperglycemia), which, if left untreated, can result in severe complications such as cardiovascular disease, kidney failure, and neuropathy [2]. While genetic predisposition and environmental factors are known contributors to the disease, their complexity requires robust diagnostic and management strategies [3].

In recent years, machine learning (ML) has emerged as a promising tool for diabetes prediction and diagnosis [4]. These methods enable researchers and clinicians to analyze large datasets, uncover hidden patterns, and build predictive models with high accuracy [5]. For example, several studies have employed ML algorithms such as Decision Trees, Naïve Bayes, and ensemble techniques like AdaBoost to enhance predictive performance in diabetes detection [6], [7], [8]. However, the literature on this topic is extensive and often fragmented, with studies varying in their methodologies, datasets, and evaluation metrics.

Despite the wealth of research in this domain, a comprehensive systematic literature review (SLR) that synthesizes findings, identifies gaps, and evaluates the state-of-the-art machine-learning approaches for diabetes prediction has not yet been conducted. An SLR is a structured methodology that aggregates and critically evaluates existing studies to provide a high-level understanding of the research landscape. This paper addresses this gap by systematically reviewing studies on diabetes prediction using machine learning published between 2011 and 2024, sourced from five major databases: IEEE, ScienceDirect, Springer, Scopus, and ACM.

The objectives of this review are threefold:

- To summarize the current state of machine learning techniques applied to diabetes prediction.
- To identify key challenges, including data preprocessing, feature selection, class imbalance, and algorithm performance.
- To highlight research gaps and propose directions for future studies in this domain.

By synthesizing the existing body of work, this review aims to provide researchers with a comprehensive understanding of the strengths and limitations of current methodologies while laying a foundation for advancing diabetes prediction through machine learning. Figure 1 explains the overview of our study.

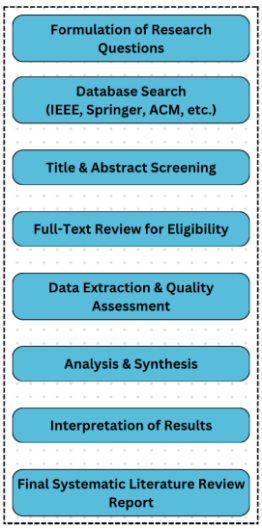

Figure 1: Systematic Literature Review Research Process

A. Problem Statement

Diabetes prediction is a critical area of research, given the global prevalence and severe health consequences of the disease [9]. Machine learning (ML) has emerged as a powerful tool for early detection and risk assessment, enabling more accurate and data-driven predictions [10]. Over the past decade, a significant number of studies have explored various ML models for diabetes prediction, utilizing diverse datasets, preprocessing techniques, and evaluation metrics. However, the rapid expansion of this research has resulted in a fragmented body of literature, making it increasingly challenging to synthesize and compare existing findings comprehensively.

This study addresses this challenge by systematically reviewing the current literature on diabetes prediction using ML models. It aims to consolidate and analyze key methodologies, datasets and model performance metrics while identifying trends and best practices. Additionally, this review highlights critical challenges such as data preprocessing, feature selection, model interpretability, and class imbalance issues that significantly impact the effectiveness of predictive models. By evaluating research gaps and unresolved issues, this SLR provides a structured foundation for future research, supporting both experienced researchers and those new to the field in advancing ML-driven diabetes prediction.

B. Research Contributions

Significant research has been conducted in the field of diabetes detection and prediction. However, there remains a lack of a comprehensive systematic literature review (SLR) that systematically organizes and synthesizes existing evidence, identifies key challenges, and addresses unresolved gaps in the field. This SLR aims to fill this gap by reviewing studies published between January 2011 and 2024.

The primary contributions of this review are as follows:

- RQ1: What are the key characteristics and formulations in the Pima Indians Diabetes dataset?
- RQ2: What techniques can be used to effectively balance sample distributions in diabetes prediction datasets?
- RQ3: Which machine learning models are most used for diabetes prediction, and how do they compare in terms of performance?

II. RELATED WORKS

In this Systematic Literature Review (SLR), we conducted a structured search across five major scientific databases: IEEE, ACM, Springer, Scopus, and ScienceDirect. These databases were selected because they comprehensively cover high-quality research in machine learning and healthcare applications.

To ensure a thorough and systematic search, we formulated a well-defined search string incorporating key terms related to diabetes prediction, machine learning techniques, and data preprocessing challenges. The final search string used was:

(Diagnostic OR Characteristic OR Distinguishing OR Indicative OR Symptomatic) AND (Classification OR Order OR Division OR Sorting OR Grouping OR Kind) AND (Prediction OR Conjecture OR Foretelling OR Prognostication OR Hunch) AND (Disease OR Illness OR Malady OR Sickliness) AND (SVM OR SVMs OR Support Vector Machine) AND (KNN OR k-NN OR k-nearest neighbors) AND (Balance OR Counterbalance OR Equivalence OR Evenness OR Parity) AND (Correlation OR Interrelationship OR Relationship OR Interconnection).

This search query was carefully designed to capture a broad yet relevant set of studies, ensuring the inclusion of diverse methodologies and perspectives on diabetes prediction using machine learning. The retrieved studies were then systematically reviewed based on relevance, publication year, and methodological rigor to construct a comprehensive analysis of the existing literature.

After conducting our systematic search, we identified numerous studies on diabetes prediction using machine learning, primarily spanning the period from 2012 to 2022. This review examines key aspects of these studies, including datasets used, preprocessing techniques, and machine learning models applied. While some studies utilized well-structured datasets that required minimal preprocessing, others incorporated essential steps such as handling missing values, examining feature correlations, and applying normalization—particularly when using algorithms like Support Vector Machines (SVM). Although these preprocessing techniques enhance model accuracy, they may also increase computational complexity and execution time [11].

Our review indicates that Support Vector Machines (SVM) are among the most frequently employed algorithms in diabetes prediction research. One of the key advantages of SVM is its robustness against outliers, making it particularly useful in disease prediction, where unreliable data points can significantly impact model performance [12],[13],[14]. Similarly, the k-nearest neighbors (KNN) algorithm is widely used, though selecting an appropriate value for K requires

domain expertise. In many cases, setting K = 5 has demonstrated high classification accuracy, particularly when combined with k-means clustering for feature selection and data segmentation [15].

Despite the growing application of machine learning in healthcare, our analysis reveals a critical gap: many studies do not adopt systematic data mining methodologies, such as CRISP-DM or Fayyad's model, which are designed to provide structured and reproducible data analysis processes. The absence of such frameworks represents an opportunity for future research to standardize diabetes prediction methodologies, improving interpretability and comparability across studies [16].

Furthermore, our review highlights specific challenges related to gestational diabetes mellitus (GDM) prediction. One study noted that populations examined before 24 weeks of gestation were often excluded from analysis, limiting the generalizability of findings. Additionally, diabetes prevalence results were not consistently categorized into multiple age groups, further complicating global trend analysis. A recent study on the global prevalence of hyperglycemia in pregnancy (HIP) projected that assuming current trends persist, HIP prevalence will stabilize between 15.8% and 16.0% from 2019 to 2045. However, these results are often confounded by data heterogeneity, inconsistencies in diagnostic criteria, and variations in GDM screening protocols across different regions. Establishing a uniform screening and diagnosis framework for GDM, like standardized approaches in diabetes prediction, would significantly improve research consistency and clinical applicability [17].

Ahmed et al. proposed a hybrid approach consisting of k-means and principal component analysis (PCA) to predict and check future diabetes using nine machine learning algorithms; this study achieved a good accuracy of 95% using random forest [18].

III. METHODOLOGY

This systematic literature review (SLR) covers a wide range of studies related to diabetes detection and prediction. The primary objective is identifying the best machine-learning model for diabetes prediction. Although variations in data types can be crucial in SLRs, this study focuses on a single dataset: the Pima Indians Diabetes dataset. The goal is to highlight potential challenges and gaps in the methodologies proposed in the literature, aiming to improve the effectiveness of diabetes identification and prediction algorithms.

As the literature defines, an SLR involves organizing, evaluating, and summarizing available research relevant to a specific area of interest. The motivation behind conducting this review is to identify the most effective methods, highlight challenges, and uncover research gaps, thereby guiding future work in this domain.

A. Research Questions

The following steps outline the systematic process of conducting this SLR:

- Define the research questions.
- Identify relevant studies and perform a pilot study.
- Search for relevant information across major databases (IEEE, Springer, ACM, ScienceDirect).
- Document the search strategy.
- Evaluate and select studies for inclusion.
- Analyze and present the results.
- Discuss the generalized conclusions and limitations of the review.
- Provide recommendations for future research.

The questions that need to be addressed in this SLR are:

- RQ1: What are the definitions and formulations of the characteristics in the Pima Indians Diabetes dataset (PIMA)?
- RQ1a: What techniques are used to identify and relate characteristics to each other in PIMA?
- RQ1b: How are features that negatively affect the label removed in PIMA?
- RQ2: How can the number of samples be balanced in the PIMA dataset?
- RQ2a: What techniques are used for data balancing in PIMA?
- RQ2b: What criterion determines that the data is unbalanced in PIMA?
- RQ2c: How does unbalanced data affect the accuracy of machine learning models?
- RQ3: What machine learning models are the most widely used for predicting patient diabetes?
- RQ3a: How is the most efficient machine learning model determined?
- RQ3b: What criterion determines the model is most appropriate for the PIMA dataset?

B. Search Strategy

A well-planned search strategy is essential in an SLR to extract relevant research studies that address the research questions. A meticulous search was conducted to gather the necessary literature to ensure comprehensive coverage. The following steps were taken to design the search terms for this SLR:

- Determine relevant search terms: The key terms were derived from the research questions by identifying the primary elements of interest: population, intervention, outcome, and context.
- List keywords: Keywords were gathered from relevant papers on diabetes prediction and machine learning models.

- Identify alternative spellings and synonyms: A dictionary was used to identify alternative spellings and synonyms for the search terms, ensuring comprehensive search coverage.
- 4 Use Boolean AND: Boolean operators like AND were used to concatenate search terms, narrowing the scope to ensure that the retrieved studies were directly relevant to the research questions.
- Use Boolean OR: The OR operator was applied to expand the search by including terms with similar meanings, capturing a broader range of relevant studies.

C. Search String

The resulting search strings used for this systematic literature review are as follows:

- Diagnostic: "Characteristic" OR "Distinguishing" OR "Indicative" OR "Symptomatic".
- Classification: "Order" OR "Division" OR "Sorting" OR "Grouping" OR "Kind".
- Prediction: "Conjecture" OR "Foretelling" OR "Prognostication" OR "Hunch".
- Disease: "Illness" OR "Malady" OR "Sickliness".
- SVM: "SVMs" OR "Support Vector Machine".
- KNN: "k-NN" OR "k-nearest neighbors".
- Balance: "Counterbalance" OR "Equivalence" OR "Evenness" OR "Parity".
- Correlation: "Interrelationship" OR "Relationship" OR "Interconnection".

These search strings were designed to identify relevant papers in the literature related to diabetes prediction. While these terms were included to maximize the reliability of the search results, studies that do not specifically address diabetes detection and prediction were excluded from the selection process.

D. String Refinement

Once the search string is established, validating the search results returned by the defined search engines is crucial. Potential papers for primary review should be visible in the results [19]. If no relevant papers appear or very few are returned, the search string must be refined [20].

To enhance the search string, we need to:

- Refine synonyms: Reassess the synonyms used in the search string to ensure comprehensive coverage of relevant literature.
- Adjust search criteria: Modify the search criteria in each search engine as needed, considering factors such as inclusion and exclusion of synonyms, publication type, year constraints, language, research area, and specific journals.
- Evaluate impact: Systematically examine the impact of these adjustments on the search results until satisfactory outcomes are achieved.

The process of refining the search string for this SLR involves iterating through these steps to ensure that the most relevant studies are captured.

E. Study Selection

The literature search yielded 72 research papers. However, several papers were excluded based on specific criteria for the machine learning models used for prediction, data balancing methods, and the relationship between characteristics.

In the first stage of the selection process, 17 papers were excluded based on their relevance to this review's objectives. This left us with 49 papers, which were thoroughly analyzed through their introductions and conclusions.

Subsequently, a more detailed analysis was conducted, resulting in the exclusion of an additional 26 papers. Ultimately, 21 papers were identified as suitable for inclusion in this systematic literature review; Figure 2 explains all study selections.

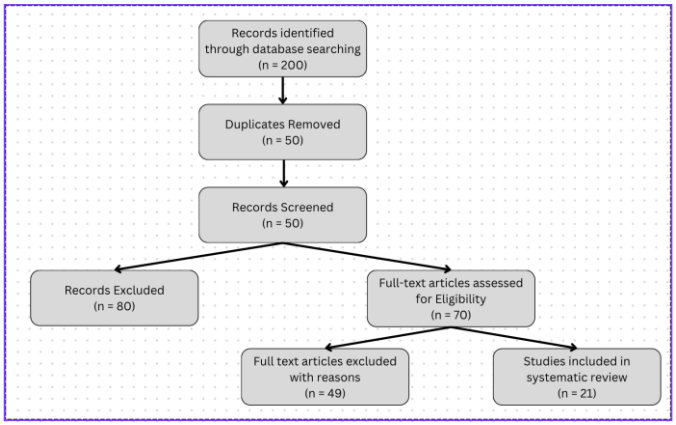

Figure 2. Flowchart for Study Selection in the SLR

IV. RESULTS AND DISCUSSION

A. Techniques Used to Identify and Relate Characteristics in PIMA (RQ1a)

- To analyze the relationships between characteristics in the PIMA dataset, various techniques and models were applied. These include:
- Figures: Visual representations such as correlation matrices and scatter plots help in understanding the relationships between attributes [21].
- Tables: Summary statistics presented in tables highlight key information about the dataset, such as means, variances, and distributions [22].
- Data Head(): Provides a preview of the dataset, showing the first few rows for an initial exploration of the characteristics [23].

- Data Describe(): Summarizes statistical metrics like count, mean, and standard deviation for each feature, helping identify potential relationships or anomalies [24].
- Data.info(): This website displays data types and missing values, giving insight into the dataset's structure [25].

Several machine learning models were utilized in this review to relate characteristics, including Naive Bayes, which is useful when the data lacks correlation. Naive Bayes assumes independence between features, which can be beneficial and limiting [26]. For smaller datasets, this independence may be advantageous, but for larger datasets, ignoring relationships between attributes may reduce accuracy. On the other hand, decision trees were found to be highly effective, especially when properly tuned. Decision trees can handle both correlated and uncorrelated data, making them versatile. The hybrid models have proven effective in diabetes prediction [27]. The diversity of models chosen in the reviewed studies was based on achieving the highest possible accuracy, with models selected according to the dataset's specific characteristics.

B. Features that Negatively Affect the Label Removed in PIMA (RQ1b)

To remove features that negatively impact the prediction label in the PIMA dataset, the following methods were commonly applied:

- Heatmap: A heatmap of the correlation matrix visualizes the relationships between features and the target label. Features with low or negative correlation to the label can be identified for potential removal [28].
- Feature Importance: Machine learning models like Random Forest or Gradient Boosting calculate feature importance scores, which rank features based on their contribution to the model's prediction accuracy [29]. Features with low importance scores are often removed to improve model performance.
- Statistical Measures: Various statistical tests, such as ANOVA or Chi-squared tests, assess the significance of each feature [30]. Features that are not statistically significant in predicting the label can be eliminated from the model.

These techniques ensure that only the most relevant features are retained, enhancing the overall accuracy of the machine learning models.

C. Techniques Used in Data Balancing in PIMA (RQ2a)

To address the class imbalance in the PIMA dataset, various techniques have been employed:

- Use the Right Evaluation Metrics: Precision, recall, F1-score, and area under the ROC curve are used to evaluate models trained on imbalanced data rather than relying solely on accuracy.
- Resample the Training Set: Resampling techniques, such as oversampling the minority class or undersampling the majority class, are applied to balance the dataset.
- Use K-fold Cross-Validation in the Right Way: K-fold cross-validation helps reduce bias when training on imbalanced data by ensuring each fold contains a similar distribution of classes.
- Ensemble Different Resampled Datasets: Ensemble methods, such as Bagging or Boosting, are used on different resampled versions of the training set to improve prediction performance.
- Resample with Different Ratios: Various resampling ratios are tested to find the best balance between the minority and majority classes.
- Cluster the Abundant Class: Clustering techniques can be used in the majority class to group similar instances, reducing redundancy and balancing class distribution.
- 7Design Your Own Models: To address class imbalance, custom models specifically designed to handle imbalanced datasets can be created.

D. The criterion that Determines the Data is Unbalanced in PIMA (RQ2b)

A common criterion for determining data imbalance is the imbalance ratio (IR). The dataset is considered imbalanced if the IR is greater than 3 (i.e., IR > 3).

E. How Unbalanced Data Affects Model Accuracy (RQ2c)

Unbalanced datasets can significantly reduce model accuracy, especially for the minority class. Models trained on unbalanced data often perform poorly when generalizing to new data. This happens because the model tends to predict the majority class, leading to misleadingly high accuracy but poor performance on the minority class. This phenomenon is known as the Accuracy Paradox, where accuracy may appear high while the model is not learning to distinguish between classes effectively.

F. Most Efficient Machine Learning Models in Predicting Diabetes (RQ3a)

Based on the literature, the most used machine learning models in predicting diabetes are:

- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Decision Tree (DT)
- Logistic Regression (LR)
- Neural Networks (NN)
- Naive Bayes (NB)

G. Criteria to Determine the Most Efficient Machine Learning Model (RQ3b)

The efficiency of a machine learning model is determined based on:

- Accuracy: The overall correctness of the model in predicting diabetes.
- Build System: How well the model integrates into the system and its scalability.
- Run Time: The model's time to train and make predictions is particularly important for large datasets or real-time predictions.

V. ANALYSIS

The essential objective of this SLR is to display and comprehensively display existing procedures for diabetes forecasting. Specifically, it points to reply to the defined inquiry about questions by completely looking into the chosen articles, which were filtered utilizing the consideration, prohibition, and quality (RQ1) It aims to identify the data format and its properties for predicting diabetes mellitus,(RQ2) It aims to balance the data and how to deal with it, for predicting diabetes mellitus,(RQ3) It aims to select the most efficient model for predicting diabetes mellitus, it is clear that one of the most important factors affecting diabetic patients is glucose, so it always has a strong relationship with the adsorbent and is not omitted from the features because it is an important feature in predicting diabetes patients [28], utilized the data from a longitudinal clinical think about, the San Antonio Heart Ponder. Eight ML techniques were utilized to anticipate whether a person would create sort two diabetes within the following 7-8 years. An exactness of 95.94% was fulfilled by the gathering Naïve Bayes (NB). Irregular Woodland (RF) and Support Vector Machine (SVM) models were assessed as well for this dataset, coming about in a worthy precision but with low sensitivity (58.5% for RF and 51.9% for SVM). All that came about was utilized to foresee the rate of diabetes, which was computed based on a 10-crease cross-validation. Wang et al. [29], to assist in assessing the diverse calculations, Cruel Outright Mistake (MAE), Root Cruel Square Blunder (RMSE), and Log Loss (LL) were calculated for the MRMR highlights, see Table 5. The RF calculation had the lowest MSE score, 0.649%, for all highlights. Scores for SVM (9.74%), AB (12.34%), and GB (2.59%) were all impressively higher. RMSE for AB and SVM were very tall (35.13% and 31.21%). The Log Misfortune score of RF was exceptionally moo (22.45%) with LL scores for SVM, AB, and GB of 336.42%, 426.14%, and 89.71% individually [30]; one of the most important elements that affect the accuracy of the model is the correct data, sometimes medical data is close to health, but this affects because diseases are sensitive and need more accuracy. Also, the process of hiding people's information affects the dissemination of data and the need for more freely accessible datasets. So, an analyst can contribute to the information on wrongdoing forecast by pre-sending a novel dataset. This SLR concludes by edifying the unwavering quality, precision, and opportuneness issues of wrongdoing datasets that can influence the general execution and efficiency of wrongdoing forecast calculations.

VI. CONCLUSION AND FUTURE STUDY

This paper systematically examined the various factors influencing the prediction of diabetes patients. Through an extensive review of relevant literature, we did not find any prior research that conducted a systematic literature review (SLR) specifically on diabetes prediction models. Papers published between 2011 and 2024 were analyzed, with a particular focus on data handling, the selection of appropriate machine learning models, and the preprocessing steps involved before feeding data into these models. This review highlights the rapid development of diabetes prediction methods over the past decade. Several gaps in existing research were identified, including the lack of focus on effective preprocessing techniques, such as feature selection, and the absence of standardized data mining methods that would make the research process easier to follow and interpret. The findings of this SLR underline that accurate detection and prediction of diabetes patients rely heavily on precise medical data and expertise in data processing. The importance of choosing the right machine learning models and identifying the most influential features can- not be overstated. This review also emphasizes the need for laboratory professionals to provide accurate test results, as early detection of diabetes is critical for effective intervention and patient care.

Future research directions include developing improved preprocessing methodologies and feature selection techniques that can further enhance model accuracy. Additionally, the creation of standardized data mining frameworks for medical data analysis could greatly assist researchers in building more effective predictive systems, not only for diabetes but also for other diseases.

REFERENCES

[1] H. Sun et al., "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," Diabetes Research and Clinical Practice, vol. 183, p. 109119, Jan. 2022, doi: 10.1016/j.diabres.2021.109119.

[2] S. Alam, Md. K. Hasan, S. Neaz, N. Hussain, Md. F. Hossain, and T. Rahman, "Diabetes Mellitus: Insights from Epidemiology, Biochemistry, Risk Factors, Diagnosis, Complications and Comprehensive Management," Diabetology, vol. 2, no. 2, pp. 36–50, Apr. 2021, doi: 10.3390/diabetology2020004.

[3] D. J. Hunter, "Gene–environment interactions in human diseases," Nature Reviews Genetics, vol. 6, no. 4, pp. 287–298, Apr. 2005, doi: 10.1038/nrg1578.

[4] E. Afsaneh, A. Sharifdini, H. Ghazzaghi, and M. Z. Ghobadi, "Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review," Diabetology & Metabolic Syndrome, vol. 14, no. 1, p. 196, Dec. 2022, doi: 10.1186/s13098-022-00969-9.

[5] I. D. Dinov, "Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data," GigaScience, vol. 5, no. 1, p. 12, Dec. 2016, doi: 10.1186/s13742-016-0117-6.

[6] A. F. Fadhlullah and T. Widiyaningtyas, "Comparative Analysis of Decision Tree and Random Forest Algorithms for Diabetes Prediction," JTAM (Jurnal Teori dan

Aplikasi Matematika), vol. 8, no. 4, p. 1121, Oct. 2024, doi: 10.31764/jtam.v8i4.24388.

[7] A. Z. Arrayyan, H. Setiawan, and K. T. Putra, “Naive Bayes for Diabetes Prediction: Developing a Classification Model for Risk Identification in Specific Populations,” Semesta Teknika, vol. 27, no. 1, pp. 28–36, Apr. 2024, doi: 10.18196/st.v27i1.21008.

[8] B. Thuraka, V. Pasupuleti, C. S. Kodete, R. S. Chigurupati, N. S. K. M. K. Tirumanadham, and V. Shariff, “Enhancing Diabetes Prediction using Hybrid Feature Selection and Ensemble Learning with AdaBoost,” in 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), IEEE, Oct. 2024, pp. 1132–1139. doi: 10.1109/I-SMAC61858.2024.10714776.

[9] M. A. B. Khan, M. J. Hashim, J. K. King, R. D. Govender, H. Mustafa, and J. al Kaabi, “Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends,” Journal of Epidemiology and Global Health, vol. 10, no. 1, p. 107, 2019, doi: 10.2991/jegh.k.191028.001.

[10] W. Strielkowski, A. Vlasov, K. Selivanov, K. Muraviev, and V. Shakhnov, “Prospects and Challenges of the Machine Learning and Data-Driven Methods for the Predictive Analysis of Power Systems: A Review,” Energies, vol. 16, no. 10, p. 4025, May 2023, doi: 10.3390/en16104025.

[11] M. Bangar and P. Chaudhary, "A novel approach for the classification of diabetic maculopathy using discrete wavelet transforms and a support vector machine," *AIMS Electronics & Electrical Engineering*, vol. 7, no. 1, 2023.

[12] M. Khanna, L. K. Singh, and H. Garg, "A novel approach for human diseases prediction using nature inspired computing & machine learning approach," *Multimedia Tools and Applications*, vol. 83, no. 6, pp. 17773–17809, 2024.

[13] A. I. Saleh, F. M. Talaat, and L. M. Labib, "A hybrid intrusion detection system (HIDS) based on prioritized k-nearest neighbors and optimized SVM classifiers," *Artificial Intelligence Review*, vol. 51, pp. 403–443, 2019.

[14] V. Krishna B, et al., "A novel application of K-means cluster prediction model for diabetes early identification using dimensionality reduction techniques," *The Open Bioinformatics Journal*, vol. 16, no. 1, 2023.

[15] B. V. V. S. Prasad, et al., "Predicting diabetes with multivariate analysis: an innovative KNN-based classifier approach," *Preventive Medicine*, vol. 174, p. 107619, 2023.

[16] F. S. Butt, et al., "Application of CRISP-DM and DMME to a case study of condition monitoring of lens coating machines," in *2023 IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4.0&IoT)*, IEEE, 2023.

[17] T. Tiftik, et al., "The paradoxical impact of diabetes mellitus on osteoporosis and sarcopenia: The ParaDOS study," 2023.

[18] A. A. Mohammed, P. Sumari, and K. Attabi, “Hybrid K-means and Principal Component Analysis (PCA) for Diabetes Prediction:,” International Journal of Computing and Digital Systems, vol. 15, no. 1, pp. 1719–1728, Jun. 2024, doi: 10.12785/ijcds/1501121.

[19] S. S. Panesar, et al., "How safe is primary care? A systematic review," BMJ Quality & Safety, vol. 25, no. 7, pp. 544-553, 2016. doi: 10.1136/bmjqs-2015-004178.

[20] M. Gusenbauer and N. R. Haddaway, "Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources," Research Synthesis Methods, vol. 11, no. 2, pp. 181-217, 2020. doi: 10.1002/jrsm.1378.

[21] A. Sarikaya and M. Gleicher, "Scatterplots: Tasks, data, and designs," IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 1, pp. 402-412, 2017. doi: 10.1109/TVCG.2017.2744184.

[22] R. M. Heiberger and B. Holland, Statistical Analysis and Data Display: An Intermediate Course with Examples in R, 2nd ed. Springer, 2015. doi: 10.1007/978-1-4939-2122-5.

[23] A. A. Patel, Hands-On Unsupervised Learning Using Python: How to Build Applied Machine Learning Solutions from Unlabeled Data, 1st ed. O'Reilly Media, 2019.

[24] Muzaferija, Ibrahim, and Assist Prof Dr Zerina Mašetić. "Cloud Computing Anomaly and Threat Detection Using Big Data Analytics and Machine Learning."

[25] X. Zhang, "Machine learning insights into digital payment behaviors and fraud prediction," Applied and Computational Engineering, vol. 67, pp. 61-67, 2024.

[26] N. A. Zaidi, J. Cerquides, G. I. Webb, and K. M. Ting, "Alleviating naive Bayes attribute independence assumption by attribute weighting," Journal of Machine Learning Research, vol. 14, no. Jul, pp. 1947-1988, 2013.

[27] M. M. Abd Zaid and A. A. Mohammed, "Hybrid models in diabetes prediction: A review of techniques, performance, and potential," Journal of Al-Qadisiyah for Computer Science and Mathematics, vol. 16, no. 4, pp. 298-308, 2024.

[28] Paleczek, Anna, and Artur Rydosz. "Review of the algorithms used in exhaled breath analysis for the detection of diabetes." Journal of breath research 16.2 (2022): 026003.

[29] Patel, R. "Diabetes Prediction using Data Mining Techniques-A Comparative Study." (2023): 818-824.

[30] Pushpo, Mahzebin. Predicting diabetes using machine learning: a comparative study of supervised classification models. Diss. Brac University, 2023.