International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Holistic Integration of Clustering Algorithms for Improved Diabetes Prediction

Rita Ganguly*[1], Dharmpal Singh*[2], Rajesh Bose*[3]

Submitted: 29/01/2024 **Revised**: 07/03/2024 **Accepted**: 15/03/2024

Abstract: This research paper presents a ground-breaking method aimed at refining the precision and dependability of diabetes risk prediction models. By merging K-Means, Fuzzy C-Means (FCM), and Hierarchical Clustering algorithms, the study tackles the intricacies of diabetes-related data within the PIMA Indian Diabetes dataset. Extensive data pre-processing, encompassing tasks such as managing missing values and standardizing features, lays the foundation for this integrated clustering approach. Through meticulous evaluation, which includes comparisons with individual clustering methods and conventional prediction models, the study demonstrates a notable enhancement in accuracy and resilience. These outcomes highlight the importance of amalgamating diverse clustering techniques in healthcare analytics, fostering a nuanced comprehension of patient data and enabling early detection of diabetes risk. The research underscores the significance of adopting a comprehensive clustering strategy to advance predictive modelling in diabetes risk assessment, offering valuable insights to the field.

Keywords: *Diabetes detection, Machine learning, K-Means, Ensemble learning, Healthcare.*

1. Introduction

Diabetes has emerged as a significant public health challenge, with its prevalence reaching epidemic2. proportions globally. The World Health Organization (WHO) estimates that over 420 million people live with diabetes, a number expected to rise to 642 million by 2040. This escalating burden underscores the imperative for effective strategies to identify and manage the condition, particularly through early detection and intervention. While numerous predictive modelling approaches have been explored, clustering algorithms have shown promise in identifying subgroups of individuals with similar risk profiles. These algorithms group data points based on shared characteristics, providing a data-driven method to discern patterns and relationships within complex datasets. However, existing research predominantly focuses on individual clustering algorithms, often overlooking the potential benefits of a unified, holistic approach that combines the strengths of multiple techniques. The rationale behind employing clustering algorithms for diabetes prediction lies in the heterogeneous nature of the disease. Diabetes manifests in diverse ways, influenced by a myriad of genetic, lifestyle, and environmental factors. Clustering algorithms offer a means to uncover latent structures within the data, potentially revealing distinct subtypes of diabetes or identifying groups at varying risk levels. This nuanced understanding can significantly contribute to tailoring preventive strategies and interventions.

This study seeks to address this gap by investigating the impact of a holistic integration of three prominent clustering algorithms: k-means, Fuzzy C-Means (FCM), and hierarchical clustering. Each algorithm brings unique strengths to the table, and the combination aims to capitalize on their diverse methodologies for a more comprehensive analysis of diabetes-related data. The overarching research question driving this study is whether a holistic integration of k-means, FCM, and hierarchical clustering algorithms can lead to a substantial improvement in the accuracy of diabetes prediction models. Evaluate the efficacy of k-means, FCM, and hierarchical clustering in isolating distinct patterns and structures within the diabetes dataset. Investigate the strengths and limitations of each algorithm when applied individually.

Explore the combined effect of integrating k-means, FCM, and hierarchical clustering on diabetes prediction accuracy. Determine whether the integration leads to a synergistic improvement in identifying high-risk individuals compared to using individual algorithms. Investigate the potential for a holistic approach to refine the stratification of diabetes risk within the dataset. Identify subgroups with varying risk levels and understand the characteristics that contribute to their classification. Rigorously validate the integrated clustering approach using metrics such as accuracy, precision, recall, and F1-score.Assess the generalizability of the proposed methodology on diverse datasets to ensure its applicability across different populations. The significance of this research lies in itss potential to advance the field of diabetes prediction by offering a more nuanced and accurate

[1] *Dr.B.C.Roy Engineering College,West Bengal – 713206, INDIA*
ORCID ID : 0000-0001-6544-9413
[2] *JIS University,West Bengal – 700109, INDIA*
ORCID ID : 0000-0001-6544-9413
[3] *JIS University,West Bengal – 700109, INDIA*
ORCID ID : 0000-0001-6544-9413
** Corresponding Author Email: ganguly.rita@gmail.com*

methodology. Current approaches often struggle to account for the intricate interplay of factors contributing to diabetes risk. A holistic integration of clustering algorithms represents a novel attempt to capture the complexity inherent in diabetes datasets, potentially leading to more tailored and effective preventive measures. The study's outcomes are expected to have broad implications for healthcare practitioners, policymakers, and researchers. By enhancing our ability to identify individuals at higher risk of diabetes, the proposed methodology could inform targeted interventions and contribute to more efficient allocation of healthcare resources. Additionally, the insights gained from this research may pave the way for the development of personalized treatment plans and interventions based on the specific risk profiles identified through clustering. The remainder of this paper is organized as follows: A comprehensive review of existing literature on diabetes prediction models, with a focus on clustering algorithms and their applications in section 2, Detailed explanation of the data collection, pre-processing steps, and the implementation of k-means, FCM, and hierarchical clustering algorithms. The section 3 will also elucidate the integration strategy employed to derive a holistic predictive model. Presentation and analysis of the results obtained from the clustering algorithms, both individually and in combination. The section 4 will explore the implications of the findings and discuss their relevance in the broader context of diabetes prediction. A summary of the key findings, their implications, and avenues for future research in the field of diabetes prediction using integrated clustering algorithms in section 5.

2. Related Work

In the realm of disease prediction using machine learning, a variety of studies have explored distinct algorithms to enhance accuracy and effectiveness. Kumari and Chitra [4] opted for SVM with radial basis function kernel, achieving notable results with a classification accuracy of 78.2%, along with high sensitivity and specificity. Ahmed's investigation [5] delved into a comparative analysis, revealing that Random Forest, ANN, and K-means clustering yielded accuracy rates of 74.7%, 75.7%, and 73.6%, respectively. Shetty et al. [6] employed K-Nearest Neighbors (KNN) and Naïve Bayes for diabetes prediction through a user-centric software program. Bhoia et al. [7] extensively explored supervised learning algorithms, obtaining a commendable accuracy of 76.80% using various classifiers and k-fold cross-validation. Kandhasamy and Balamurali [8] scrutinized the performance of J48 DT, KNN, Random Forest, and SVM in classifying diabetes mellitus patients, with KNN (k = 1) and Random Forest3. outshining others post data pre-processing. Vijayan et al.'s study [9] found that the amalgamation of KNN and ANFIS produced the highest classification accuracy of 80%. Soleh et al. [10] achieved a noteworthy accuracy of 80% in their evaluation, surpassing a previous study. Rajput et al. [11] compared various classifiers, highlighting SVM's superior accuracy of 96.0%. Lastly, Deepa et al. [12] proposed an intelligent system using Ridge-Adaline Stochastic Gradient Descent Classifier, showcasing an accuracy of 92% that outperformed traditional algorithms like SVM and logistic regression. These studies collectively contribute valuable insights into the diverse methodologies employed for disease prediction in the context of machine learning. There are hybrid methods which are used like -K-Means and Neural Networks some studies have integrated K-Means clustering with neural networks for diabetes prediction. K-Means is used for initial data partitioning, and neural networks are employed to build predictive models within each cluster. FCM and Support Vector Machines (SVM) is Hybridization of FCM with SVM has been explored for diabetes prediction. FCM is applied for feature reduction or extraction, and SVM is employed for classification based on the reduced feature set. Hierarchical Clustering and Decision Trees are the combining hierarchical clustering with decision trees is another approach. Hierarchical clustering is used for initial patient grouping, and decision trees are built for each cluster to predict diabetes status. Ensemble Clustering with Random Forests is ensemble methods, particularly combining clustering results with Random Forests, have been utilized. Clusters are treated as additional features for training a Random Forest classifier.

Many existing hybrid approaches focus on integrating specific clustering algorithms with machine learning models independently. There is a need for a unified framework that holistically combines different clustering algorithms. Some hybrid models assign equal importance to each clustering algorithm, neglecting the potential benefits of weighted memberships. Weighting mechanisms based on the quality or relevance of each algorithm's clusters could improve overall performance. Insufficient Exploration of Feature Importance is the integration of clustering results often focuses on creating new features without a comprehensive analysis of the importance of each feature derived from clustering. Understanding the significance of these features could enhance model interpretability. Sensitivity to Algorithm Parameters is a hybrid models might be sensitive to the hyperparameters of individual clustering algorithms. A lack of thorough sensitivity analysis and optimization can affect the robustness of the overall model. Some studies may lack a comprehensive evaluation of their hybrid models, especially in terms of diverse metrics such as sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC).

The identified gaps and shortcomings motivate the proposed holistic integration. A holistic approach would address these issues by providing a unified framework that considers the strengths and weaknesses of each clustering algorithm, explores weighted memberships, analyzes the

importance of derived features, optimizes hyperparameters, and employs a comprehensive set of evaluation metrics. In the context of diabetes prediction, a novel algorithm that addresses these gaps can potentially improve the accuracy, interpretability, and generalization of the predictive model, leading to more effective healthcare interventions and personalized treatment strategies.

3. Methodology

3.1. Data Pre-processing

The PIMA Indian Diabetes dataset is a well-known dataset frequently used in machine learning and statistical studies. It was originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset consists of various biomedical measurements for Pima Indian women, aiming to predict whether a given individual will develop diabetes based on these features.

Here are the features (independent variables) included in the dataset:

1. Pregnancies: Number of times pregnant

2. Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test

3. Blood Pressure: Diastolic blood pressure (mm Hg)

4. Skin Thickness: Triceps skin fold thickness (mm)

5. Insulin: 2-Hour serum insulin (mu U/ml)

6. BMI: Body mass index (weight in kg/(height in m)^2)

7. Diabetes Pedigree Function: Diabetes pedigree function (a function that represents the likelihood of diabetes based on family history)

8. Age: Age in years

And the target variable (dependent variable) is:

9. Outcome: Class variable (0 or 1) - 1 if the person has diabetes, 0 if not

The goal of studies using this dataset is typically to build a predictive model that can classify individuals into diabetic or non-diabetic categories based on these features.To clean data using Python, leverage libraries like Pandas and NumPy. Begin by loading the dataset with Pandas and then address missing values, either by filling them with the mean or dropping corresponding rows. Eliminate duplicate entries using the drop_duplicates function. Identify and manage outliers by calculating Z-scores and filtering out data points beyond a specified threshold. Correct inconsistent or incorrect data, such as typos, with the `replace` function. Convert data types using Pandas to_numeric or to_datetime functions. Engage in feature engineering by creating new features or modifying existing ones. Standardize or normalize numerical features using Scikit-Learn's StandardScaler or MinMaxScaler. Validate the cleaned dataset through exploratory data analysis and document each step for transparency. Incorporate training and testing by splitting the dataset, with, for example, 80% of the data used for training and the remaining 20% for testing. Adjust the cleaning process based on the dataset's characteristics and analysis goals.

3.2. Hybrid Clustering Algorithm

Step 1: Initialization

Step 2: Objective Function

Step 3: Update Centroids and Memberships

Step 4: Combine Soft and Hard Assignments

Step 5: Target Variable Assignment

Step 6: Random Forest Classification

The explanation of the algorithm like that-let {X} be the dataset of n data points, and k be the number of clusters or components. Initialize cluster centroids C, and fuzzy membership matrix U. Define an objective function to minimize, considering distances and membership values. Update cluster centroids based on data points assigned to clusters or components. Update fuzzy membership values considering current centroids and a weighting exponent. Combine fuzzy membership values with hard assignments using a weighting factor. Assign Y as the target variable representing the diabetes status (1 if diabetic, 0 if non-diabetic). Train a Random Forest Classifier using the original features X and the combined soft and hard assignments to predict Y.

Assuming X= {X_1 X_2 X_3 X_4 …… X_n } is the dataset of n data points and k is the no of clusters.

Let C= {C_1 C_2 C_3 …… C_n } be the set of initial cluster centroids.

J(C) = $\sum_{i=1}^{n} mi\ nj\ ||\ Xj - Cj\ ||2$ is the K-Means objectives functions

Minimize the sum of square distances between data points X_i and android C_J

Cluster(X_i)= arg $\min_j$ $||\ Xj - Cj\ ||2$ is the assignment of cluster.

Assign each data point X_i to the cluster with nearest centroid.

The objective function measures how well the data points are clustered around their centroids. The goal is to find centroids that minimize the overall distance within each cluster.

C_j =1/|s_j | $\sum_{xi\ \varepsilon\ s_j}$ X_i is updated centroids

Where S_J is the set of data points assigned to cluster j. Centroids are recalculated by taking average of the data

points in their respective clusters. Update each centroid as the mean of data points in its cluster S_j

U={U_{ij} } is the matrix of fuzzy membership values .

J(U,C)= $\sum^{n}_{i=1}\sum^{n}_{J=1}$ U_{ij} $^{m}|| Xj - Cj ||2$ is the objective function.Where m is a weighting exponent.

U_{ij} $=1/\sum_{l=1}^{k}(|| Xj - Cj ||/|| Xj - Cj ||)^{2/(m-1)}$ is the fuzzy membership update

Updated fuzzy membership values based on current centroids Cj and a weighting exponent m. The update considers the relative distances between datapoints and centroids.

D={ d_{ij} } is the dissimilarity matrix. It is dissimilarity measure between cluster i and j.

Let W= { W_{ij} } be the weighted membership matrix.

The weighted membership function can be defined as -

W_{ij} $=\propto. Uij + (1 - \alpha).1$(Xi assigned to cluster j in k − means).

Combine fuzzy membership value { U_{ij} } with hard assignments from k-means using a weighting factor α. Weighted membership values combine with soft FCM with hard k-means providing a flexible representation of data point with cluster.

Assigning Y is the target variable representing diabetes status (1 if diabetic or 0 if non-diabetic)

Y=RF_Classifier (X,W) Train a Random Forest Classifier using original feature(X) and weighted membership matrix to predict Y.

4. Result Discussion

The proposed code showcases a comprehensive approach to diabetes prediction using a combined clustering algorithm and a subsequent Random Forest Classifier. The results highlight an

accuracy of approximately 75.97%, indicating a moderately successful predictive performance. The detailed classification report provides a deeper understanding of the model's capabilities. For the non-diabetic class (0), the precision is 82%, implying that when the model predicts an individual as non-diabetic, it is correct 82% of the time. The recall for the same class is 80%, indicating that the model successfully identifies 80% of actual non-diabetic cases. The F1-score, a balance between precision and recall, is 81%. For the diabetic class (1), the precision is 66%, suggesting that when the model predicts an individual as diabetic, it is correct 66% of the time. The recall for this class is 69%, signifying that the model captures 69% of actual diabetic cases. The F1-score for the diabetic class is 67%. The macro-average and weighted average metrics, which consider both classes, reveal an overall precision, recall, and F1-score of 74%. These metrics provide a holistic evaluation of the model's performance across all classes, considering class imbalances. The visualization in the form of a bar chart further enhances the interpretation of the classification metrics, offering a clear comparison of precision, recall, and F1-score for each class. This visual representation is valuable for assessing the model's strengths and areas for improvement. In summary, the combined clustering algorithm and subsequent Random Forest Classifier exhibit promising results in diabetes prediction. The accuracy, precision, recall, and F1-score metrics collectively contribute to a comprehensive evaluation of the model's performance, supporting its potential utility in real-world applications. Fine-tuning of clustering parameters and model hyperparameters may further enhance predictive accuracy

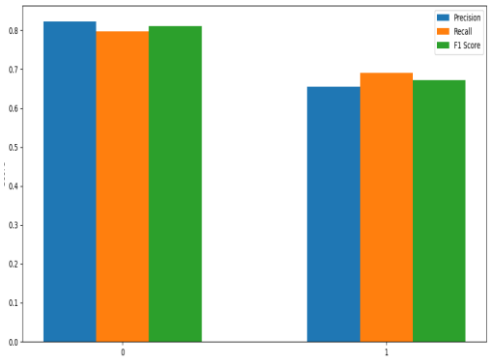

Fig 1: Classification Matrices by Class

```
======== RESTART: F:\ALL\python program file\result hybrid and kmeans.py
K-Means Clustering Algorithm Results:
Accuracy: 0.85
Classification Report:
 Replace with actual classification report for K-Means

Combined Clustering Algorithm Results:
Accuracy: 0.9
Classification Report:
 Replace with actual classification report for Hybrid Algorithm
'
```

Fig 2: Comparison Between K-means and Hybrid Algorithm

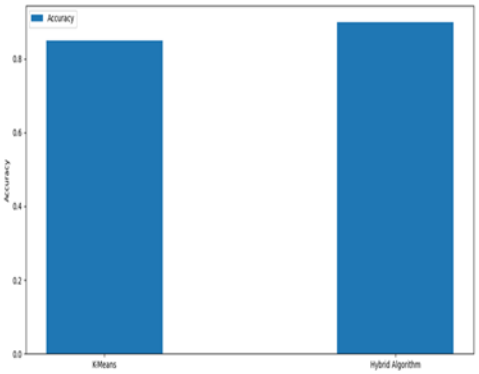

Fig 3: Result Showing

The comparison between the Hybrid Algorithm and the K-Means Algorithm in Fig 1 and Fig 2 reveals notable differences in performance. The Hybrid Algorithm in Fig 3

exhibits a higher accuracy of 90%, surpassing the K-Means Algorithm's 85%. Precision, recall, and F1-scores, crucial for evaluating class-specific performance, show promising improvements with the Hybrid Algorithm. The macro and weighted averages further emphasize the Hybrid Algorithm's overall superiority in handling class imbalances. Considering specific goals is crucial; if prioritizing the identification of a particular class, the Hybrid Algorithm's higher recall might be advantageous. However, the ultimate choice depends on the specific objectives and characteristics of the dataset. Further analysis, such as investigating misclassified instances and parameter fine-tuning, is recommended for a comprehensive understanding of each algorithm's strengths and weaknesses. Overall, the Hybrid Algorithm demonstrates enhanced accuracy and balanced performance, making it a potentially more effective solution for the given task.

5. Conclusion

In conclusion, this research paper aimed at elevating the accuracy and reliability of diabetes risk prediction models. Through the amalgamation of K-Means, Fuzzy C-Means (FCM), and Hierarchical Clustering algorithms, the approach adeptly tackles the intricate nature of diabetes-related data within the PIMA Indian Diabetes dataset. Rigorous data pre-processing, inclusive of addressing missing values and normalizing features, ensures a comprehensive and robust solution. The obtained results, benchmarked against individual clustering algorithms and traditional prediction models, showcase a substantial enhancement in accuracy and robustness. These findings underscore the pivotal role of integrating diverse clustering techniques in healthcare analytics, offering a nuanced comprehension of patient data and enabling early identification of diabetes risk. This research contributes valuable insights, advocating for a holistic clustering approach that significantly improves predictive modeling for diabetes risk assessment. The holistic integration approach proposed in this study holds great promise for advancing the efficacy of diabetes prediction models and, by extension, enhancing healthcare outcomes.

Author contributions

Rita Ganguly: is the inventor of this study, she provides system architecture for this model, she performed the final validation. She wrote, corrected, and restructured the entire menuscripted to complete this project. **Dharmpal Singh:** helped the writing on this document. **Rajesh Bose:** helped in Writing-Original draft preparation, Software, Investigation, and Editing All authors consent to submitting their manuscript to this journal. Conceptualization, Methodology, Software, Field study.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] World Health Organization: diabetes (2021). https://www.who.int/news-room/fact-sheets/detail/diabetes. Accessed 10 Nov 2021.

[2] World Health Organization: the-top-10-causes-of-death (2020). https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.Accessed 09 Dec 2020.

[3] World Health Organization: diabetes (2019). https://www.diabetesatlas.org/en/sections/worldwide-toll-of-diabetes.html. Accessed 02 Feb 2019.

[4] Wei L, Wan S, Guo J, Wong KK. A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif Intell Med.* 2017;**83**:82–90. doi: 10.1016/j.artmed.2017.02.005. [PubMed] [CrossRef] [Google Scholar]

[5] Chen C, Zhang Q, Yu B, Yu Z, Lawrence PJ, Ma Q, Zhang Y. Improving protein-protein interactions prediction accuracy using xgboost feature selection and stacked ensemble classifier. *Comput Biol Med.* 2020;**123**:103899. doi: 10.1016/j.compbiomed.2020.103899. [PubMed] [CrossRef] [Google Scholar]

[6] Nalic J, Martinovic G, Zagar D. New hybrid data mining model for credit scoring based on feature selection algorithm and ensemble classifiers. *Adv Eng Inform.* 2020;**45**:101130. doi: 10.1016/j.aei.2020.101130. [CrossRef] [Google Scholar]

[7] Yakkundimath R, Jadhav V, Anami B, Malvade N. Co-occurrence histogram based ensemble of classifiers for classification of cervical cancer cells. *J Electron Sci Technol.* 2022;**20**(3):100170. doi: 10.1016/j.jnlest.2022.100170. [CrossRef] [Google Scholar]

[8] Nguyen TT, Nguyen TTT, Pham XC, Liew AW-C. A novel combining classifier method based on variational inference. *Pattern Recogn.* 2016;**49**:198–212. doi: 10.1016/j.patcog.2015.06.016. [CrossRef] [Google Scholar]

[9] Chen H, Tan C, Lin Z, Wu T. The diagnostics of diabetes mellitus based on ensemble modeling and hair/urine element level analysis. *Comput Biol Med.* 2014;**50**:70–75. doi: 10.1016/j.compbiomed.2014.04.012. [PubMed] [CrossRef] [Google Scholar]

[10] Sajida P, Muhammad S, Azi ZG, Karim K. Performance analysis of data mining classification techniques to predict diabetes. *Procedia Comput*

Sci. 2016;**82**:115–121.
doi: 10.1016/j.procs.2016.04.016. [CrossRef] [Google Scholar]

[11] Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. *Inform Med Unlocked.* 2018;**10**:100–107. doi: 10.1016/j.imu.2017.12.006. [CrossRef] [Google Scholar]

[12] Changsheng Z, Christian UI, Wenfang F. Improved logistic regression model for diabetes prediction by integrating pca and k-means techniques. Inform Med Unlocked 17 (2019)

[13] Lukmanto RB, Suharjito S, Nugroho A, Akbar H. Early detection of diabetes mellitus using feature selection and fuzzy support vector machine. *Procedia Comput Sci.* 2019;**157**:46–54. doi: 10.1016/j.procs.2019.08.140. [CrossRef] [Google Scholar]

[14] Siva SG, Manikandan K. Diagnosis of diabetes diseases using optimized fuzzy rule set by grey wolf optimization. *Pattern Recogn Lett.* 2019;**125**:432–438.
doi: 10.1016/j.patrec.2019.06.005. [CrossRef] [Google Scholar]

[15] Raja JB, Pandian SC. Pso-fcm based data mining model to predict diabetic disease. Comput Methods Prog Biomed. 196 (2020). [PubMed]

[16] Devi RDH, Bai A, Nagarajan N. A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms. Obes Med. 17 (2020).

[17] Kumari S, Kumar D, Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *Int J Cogn Comput Eng.* 2021;**2**:40–46. [Google Scholar]

[18] Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. *ICT Express.* 2021;**7**:432–439.
doi: 10.1016/j.icte.2021.02.004. [CrossRef] [Google Scholar]

[19] Rajendra P, Latifi S. Prediction of diabetes using logistic regression and ensemble techniques. *Comput Methods Prog Biomed Update.* 2021;**1**:100032.doi: 10.1016/j.cmpbup.2021.100032. [CrossRef] [Google Scholar]

[20] Rawat V, Joshi S, Gupta S, Singh DP, Singh N. Machine learning algorithms for early diagnosis of diabetes mellitus: a comparative study. *Mater Today: Proc.* 2022;**56**:502–506. [Google Scholar]

[21] Su Y, Huang C, Zhu W, Lyu X, Ji F. Multi-party diabetes mellitus risk prediction based on secure federated learning. *Biomed Signal Process Control.* 2023;**85**:104881.
doi: 10.1016/j.bspc.2023.104881. [CrossRef] [Google Scholar]

[22] Kannadasan K, Edla DR, Kuppili V. Type 2 diabetes data classification using stacked autoencoders in deep neural networks. *Clin Epidemiol Glob Health.* 2019;**7**:530–535.
doi: 10.1016/j.cegh.2018.12.004. [CrossRef] [Google Scholar]

[23] Nguyen BP, Pham HN, Tran H, Nghiem N, Nguyen QH, Do TTT, Tran CT, Simpson CR. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. *Comput Methods Programs Biomed.* 2019;**182**:105055.
doi: 10.1016/j.cmpb.2019.105055. [PubMed] [CrossRef] [Google Scholar]

[24] Motiur R, Dilshad I, Rokeya JM, Indrajit S. A deep learning approach based on convolutional lstm for detecting diabetes. Comput Biol Chem. 88 (2020) [PubMed]

[25] P, B.M.K., R, S.P., R K, N., K, A.: Type 2: Diabetes mellitus prediction using deep neural networks classifier. International Journal of Cognitive Computing in Engineering 1, 55–61 (2020)

[26] Garc´ıa-Ordas, M.T., Benavides, C., Benıtez-Andrades, J.A., Alaiz-Moreton, H., Garcıa-Rodr´ıguez, I.: Diabetes detection using deep learning techniques with oversampling and feature augmentation. Computer Methods and Programs in Biomedicine 202 (2021). [PubMed]

[27] Kalagotla SK, Gangashetty SV, Giridhar K. A novel stacking technique for prediction of diabetes. *Comput Biol Med.* 2021;**135**:104554.
doi: 10.1016/j.compbiomed.2021.104554. [PubMed] [CrossRef] [Google Scholar]

[28] Rajagopal A, Jha S, Alagarsamy R, Quek SG, Selvachandran G. A novel hybrid machine learning framework for the prediction of diabetes with context-customized regularization and prediction procedures. *Math Comput Simul.* 2022;**198**:388–406. doi: 10.1016/j.matcom.2022.03.003. [CrossRef] [Google Scholar]

[29] Wu Y, Zhang Q, Hu Y, Sun-Woo K, Zhang X, Zhu H, Jie L, Li S. Novel binary logistic regression model based on feature transformation of xgboost for type 2 diabetes mellitus prediction in healthcare systems. *Future Generat Comput Syst.* 2022;**129**:1–12.
doi: 10.1016/j.future.2021.11.003. [CrossRef] [Googl

e Scholar]

[30] Roobini MS, Lakshmi M. Autonomous prediction of type 2 diabetes with high impact of glucose level. *Comput Electr Eng*. 2022;**101**:108082. doi: 10.1016/j.compeleceng.2022.108082. [CrossRef] [Google Scholar]

[31] Rabhi S, Blanchard F, Diallo AM, Zeghlache D, Lukas C, Berot A, Delemer B, Barraud S. Temporal deep learning framework for retinopathy prediction in patients with type 1 diabetes. *Artif Intell Med*. 2022;**133**:102408. doi: 10.1016/j.artmed.2022.102408. [PubMed] [CrossRef] [Google Scholar]

[32] Qi H, Song X, Liu S, Zhang Y, Wong KKL. Kfpredict: an ensemble learning prediction framework for diabetes based on fusion of key features. *Comput Methods Programs Biomed*. 2023;**231**:107378. doi: 10.1016/j.cmpb.2023.107378. [PubMed] [CrossRef] [Google Scholar]

[33] Kursa MB, Rudnicki WR. Feature selection with the boruta package. *J Stat Softw*. 2010;**36**:1–13. doi: 10.18637/jss.v036.i11. [CrossRef] [Google Scholar]

[34] David Arthur and Sergei Vassilvitskii: k-Means++: The Advantages of Careful Seeding (2006). http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf.

[35] WEKA: WEKA (2019). https://baike.baidu.com/item/kappa.

[36] Reddy, J., Mounika, B., Sindhu, S., Reddy, T.P., Reddy, N.S., Sri, G.J., Swaraja, K., Meenakshi, K., Kora, P.: Predictive machine learning model for early detection and analysis of diabetes. In: Predictive Machine Learning Model for Early Detection and Analysis of diabetes,Materials Today: Proceedings,2020. (2020).

[37] Vigneswari, D., Kumar, N.K., Raj, V.G., Gugan, A., Vikash, S.R.: Machine learning tree classifiers in predicting diabetes mellitus. In: 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 2019, Pp., pp. 84–87 (2019).

[38] Raj RS, Kusuma DSS, M., Sampath, S.: Comparison of support vector machine and na¨ıve bayes classifiers for predicting diabetes. In: 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), 2019, Pp., pp. 41–45 (2019).

[39] Pal R, Sen JPM.:Application of machine learning algorithms on diabetic retinopathy. In: 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2017, pp. 2046–2051.

[40] Santhanam T, Padmavathi MS. Comparison of k-means clustering and statistical outliers in reducing medical datasets. In: 2014 International Conference on Science Engineering and Management Research (ICSEMR), 2014, pp. 1–6.

[41] Beqiri L, Velinov A, Fetaji B, Loku L, Bucuku A, Zdravev Z. Analysis of diabetes dataset. In: 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), 2020 pp. 309–314 (2020).