

Optimized stacking ensemble for early-stage diabetes mellitus prediction

Aman, Rajender Singh Chhillar

Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak, India

Article Info

Article history:

Received May 23, 2023

Revised Jul 9, 2023

Accepted Jul 17, 2023

Keywords:

Artificial neural network

Diabetes mellitus

Feature normalization

Random forest

Stacking

ABSTRACT

This paper presents an optimized stacking-based hybrid machine learning approach for predicting early-stage diabetes mellitus (DM) using the PIMA Indian diabetes (PID) dataset and early-stage diabetes risk prediction (ESDRP) dataset. The methodology involves handling missing values through mean imputation, balancing the dataset using the synthetic minority over-sampling technique (SMOTE), normalizing features, and employing a stratified train-test split. Logistic regression (LR), naïve Bayes (NB), AdaBoost with support vector machines (AdaBoost+SVM), artificial neural network (ANN), and k-nearest neighbors (k-NN) are used as base learners (level 0), while random forest (RF) meta-classifier serves as the level 1 model to combine their predictions. The proposed model achieves impressive accuracy rates of 99.7222% for the ESDRP dataset and 94.2085% for the PID dataset, surpassing existing literature by absolute differences ranging from 10.2085% to 16.7222%. The stacking-based hybrid model offers advantages for early-stage DM prediction by leveraging multiple base learners and a meta-classifier. SMOTE addresses class imbalance, while feature normalization ensures fair treatment of features during training. The findings suggest that the proposed approach holds promise for early-stage DM prediction, enabling timely interventions and preventive measures.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Aman

Department of Computer Science and Applications, Maharshi Dayanand University

Rohtak, India

Email: sei@live.in

1. INTRODUCTION

Diabetes Mellitus is a persistent metabolic ailment characterized by elevated blood glucose levels. As per the International Diabetes Federation, approximately 463 million adults were affected by diabetes in 2019, with projections indicating a surge to 700 million individuals by 2045 [1]. In the Indian context, the incidence of diabetes is substantial, with an approximate population of 77 million adult individuals affected by the condition in the year 2019 [2]. Furthermore, there is a significant association between type 2 diabetes mellitus (T2DM) and depression among individuals in India. In a study conducted in Hue City, Vietnam, it was observed that approximately 23.2% of patients diagnosed with T2DM experienced symptoms of depression [3].

Diabetes is often regarded as a systemic disease with far-reaching consequences, as it exerts detrimental effects on vital organs. Individuals diagnosed with diabetes face an increased susceptibility to various complications, including but not limited to miscarriage, renal failure, myocardial infarction, vision impairment, and other chronic and potentially life-threatening conditions [4]. Therefore, it is essential to diagnose diabetes mellitus (DM) faster to prevent or delay the onset of these complications. Machine learning

(ML) algorithms (such as support vector machine (SVM) [5], k-nearest neighbors (k-NN) [6], random forest (RF) [7], and artificial neural network (ANN) [8]) can help in the early diagnosis and accurate prediction of DM by analyzing various health indicators such as plasma glucose concentration, serum insulin resistance, and blood pressure [9]–[11]. Timely identification and precise prognostication of DM hold paramount importance in facilitating efficacious interventions and optimal disease management. Leveraging the capabilities of ML algorithms, which possess the ability to process vast volumes of data, enables the detection of intricate patterns that might elude human experts [12]–[14]. However, there is a need for further research and improved methodologies to enhance diagnostic accuracy and personalized treatment strategies.

To address this problem, this study proposes a novel stacking-based hybrid ML approach for the prediction of early-stage DM. By integrating multiple base classifiers through a stacked classifier, the proposed approach can capture complex relationships and patterns within the data, leading to improved predictive performance. The use of ML algorithms offers a comprehensive understanding of DM and aids in disease management, including the identification of individuals at risk of developing complications [15]–[18]. This approach has the potential to improve health outcomes and enhance the quality of life for individuals with DM [19], [20].

In recent years, there has been a growing emphasis on the early detection and prediction of DM, prompting extensive research in this field. Doğru *et al.* [21] introduced a hybrid super ensemble learning model that integrated multiple algorithms, yielding remarkable accuracy rates of 99.6%, 92%, and 98% for the prediction of early-stage DM across diverse datasets. Krishnamoorthi *et al.* [22] devised an innovative healthcare disease prediction framework for DM, leveraging RF and SVM models, attaining an accuracy of 86% in DM prediction. In the study [23], a stacked-based model was employed to predict the presence of DM in individuals. Compared to other existing models such as LR, NB, and linear discriminant analysis (LDA), the stacked-based model predicted blood sugar disease with 93.1% accuracy. This demonstrates the effectiveness of the stacked ensemble method for enhancing DM prediction results. In addition, Chakravarthy and Rajaguru [24] proposed a voting-based approach for the early diagnosis of DM. To enhance DM prediction, they applied a mixture of three ML algorithms: LR, RF, and XGBoost classifiers. After evaluating the efficacy of each algorithm separately, it was found that the ensemble method with weighted voting provided the best results for binary classification in terms of accuracy, precision, and F1-score. Mushtaq *et al.* [25] studied the effectiveness of voting-based models and hyperparameter-tuned ML algorithms for predicting DM. The study focused on addressing the challenge of imbalanced datasets through the implementation of methods such as Tomek and synthetic minority over-sampling technique (SMOTE). A two-stage model selection approach was employed, where LR, SVM, k-NN, gradient boost, NB, and RF algorithms were evaluated. RF emerged as the top-performing algorithm, achieving an accuracy of 80.7% after dataset balancing using SMOTE. Subsequently, a voting algorithm was applied to combine three superior models, resulting in an accuracy of 82.0%. These studies have demonstrated the effectiveness of various ML algorithms in enhancing DM prediction results.

The specific aim of this work is to develop a novel stacking-based hybrid ML approach for the prediction of early-stage DM. The integration of a stacked classifier in this research enables the combination of multiple base classifiers, leveraging their collective decision-making capabilities to enhance prediction accuracy. By employing the stacked classifier, the proposed approach can effectively capture complex relationships and patterns within the data, leading to improved predictive performance for early-stage DM detection. Through a comprehensive review of existing literature, we compare our proposed approach with the currently available models and methodologies. Section 2 describes the detailed methodology of the proposed model containing data pre-processing, handling missing values, balancing the dataset, normalizing features, and hyperparameter tuning of base and meta classifiers. Section 3 presents the results of the experiments conducted, including accuracy rates and performance metrics of the stacking-based hybrid model. Finally, section 4 compares the results and performance of the proposed stacking-based hybrid model with existing literature on early-stage DM prediction, and highlights the improvements achieved by the proposed approach.

2. MATERIAL AND METHOD

The methodology employed in this study aims to develop a precise and reliable model for predicting early-stage DM. The process, as illustrated in Figure 1, follows a systematic approach involving data acquisition, data preprocessing, model formulation, model training, and model evaluation. The initial step involves gathering relevant datasets for analysis. Subsequently, the collected data undergoes preprocessing to ensure its quality and eliminate inconsistencies. Once the data is appropriately prepared, the model formulation stage entails selecting suitable algorithms and techniques for constructing the predictive model. The model is then trained using the preprocessed data, optimizing its parameters and refining its performance. Finally, the model's predictive accuracy and performance are evaluated using appropriate evaluation metrics.

To reproduce the results obtained in this study, all experiments were conducted on the Waikato environment for knowledge analysis (WEKA) platform [26]. The experimental setup utilized a system equipped with a 3.2 GHz Intel Core i5 CPU and 16 GB RAM. WEKA is a comprehensive and user-friendly environment for data analysis and machine learning tasks, providing a diverse range of algorithms and evaluation methods. By utilizing the WEKA platform, researchers can replicate the procedures described in this study and achieve comparable outcomes.

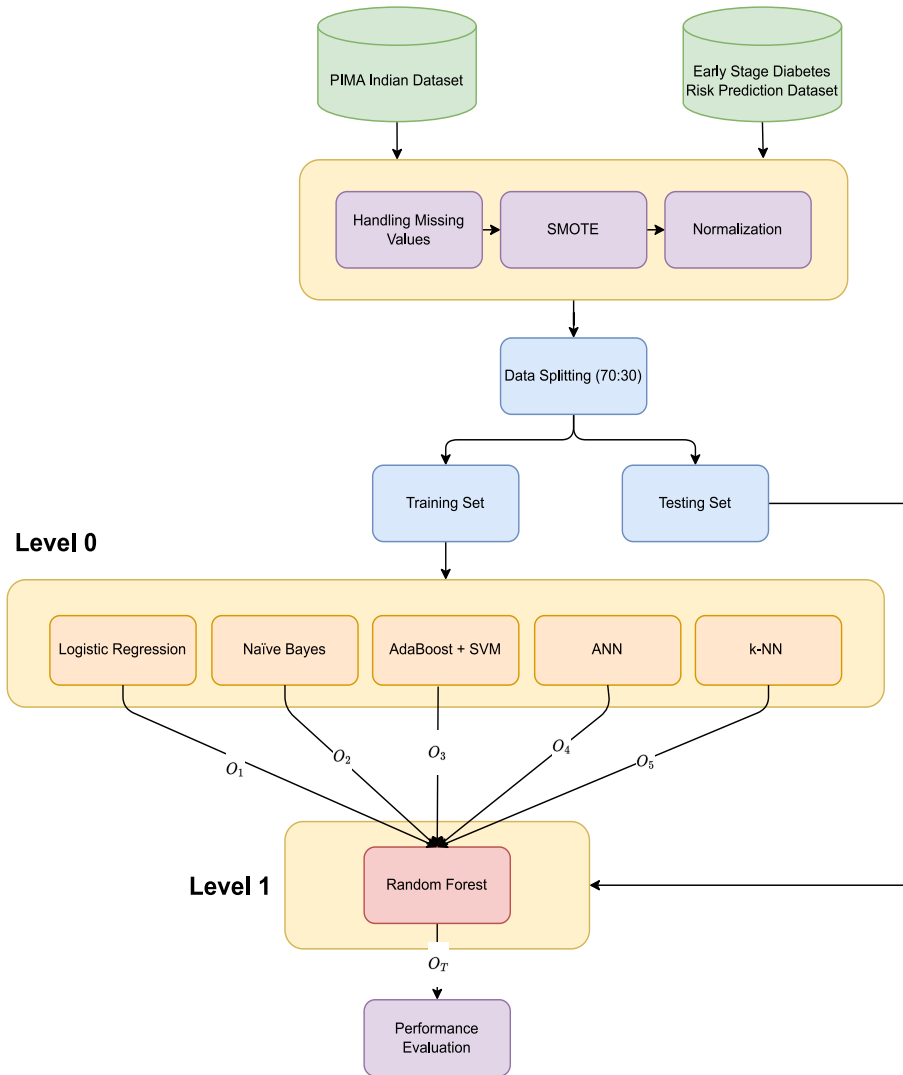


Figure 1. Workflow of the proposed stacking-based hybrid model for early-stage DM prediction

2.1. Data pre-processing

The data pre-processing phase involves handling missing values through mean imputation, addressing class imbalance through SMOTE, and normalizing the datasets. These steps prepare the datasets for subsequent modelling and analysis, ensuring a reliable and standardized foundation for accurate early-stage DM prediction. In this study, both the PID dataset [27] and the ESDRP dataset [28] were collected from the OpenML website, a reliable platform for sharing datasets and machine learning experiments. The PID dataset initially consists of 768 instances and 9 attributes, while the ESDRP dataset comprises 520 instances and 17 attributes as shown in Table 1. To handle missing values in the datasets, the mean imputation technique is employed. This can be implemented using the “ReplaceMissingValues” filter available in the WEKA platform (*WEKA.filters.unsupervised.attribute.ReplaceMissingValues*). This filter replaces the missing values with the mean value of the corresponding attribute, ensuring that the datasets are complete and ready for analysis.

There are 320 occurrences labelled “Positive” and 200 examples labelled “Negative” in the DM prediction dataset. Before data preprocessing, there are 268 cases labelled “Positive” and 500 instances labelled “Negative” in the PID dataset. The SMOTE algorithm is utilized to address class imbalance issue by generating synthetic examples of the minority class. This can be achieved using the “SMOTE” filter in WEKA (*WEKA.filters.supervised.instance.SMOTE*), with specific parameters such as the number of instances to generate (P), the number of nearest neighbors (K), and the class index (C). By applying this filter, the class imbalance is alleviated, allowing for more balanced datasets. Following the application of SMOTE, the number of cases labelled “Positive” rises to 320, equaling the number of instances labelled “Negative” at 400. Similarly, the number of “Positive” examples in the PID dataset climbs to 536 and the number of “Negative” instances increases to 500.

Normalization is performed on the preprocessed datasets to ensure consistency and comparability. The “Normalize” filter in WEKA (*WEKA.filters.unsupervised.attribute.Normalize*) can be applied to scale all attribute values between 0 and 1. This step eliminates any potential bias introduced by varying attribute scales, enhancing the accuracy and interpretability of the subsequent modeling and analysis.

Table 1. Characteristics of diabetes datasets

	PID dataset	ESDRP dataset
Description	This dataset was collected from the PID population in Arizona, USA. It consists of data from females of PID heritage and includes attributes such as age, body mass index (BMI), number of pregnancies, glucose levels, insulin levels	This dataset was collected from patients who were suspected of having early-stage DM. The data includes various biomedical attributes such as age, BMI, glucose levels, and insulin levels
Size	768×9	520×17
Target variable	Outcome (Positive/Negative)	Class (Positive/Negative)
Missing values and unbalanced	Yes	Yes

2.2. Classifiers and hyperparameter tuning

This section encompasses the process of learner selection and hyperparameter tuning for the stacking-based hybrid ML approach. In this approach, base learners are utilized as level 0 models within the stacking ensemble, while hyperparameter tuning is employed to enhance their performance. A meta-classifier is employed as the level 1 model in the stacking ensemble, which combines the predictions from the base learners and generates the final prediction. This integration of diverse outputs from the base learners enables improved overall predictive performance [29]. In this study, the meta-classifier used is RF. To optimize the performance of the base learners, hyperparameter tuning was conducted. Table 2 provides an overview of the hyperparameter settings for each classifier used in this study, specifically for the PID dataset and the ESDRP dataset. The hyperparameters were fine-tuned using cross-validation with the CVParameterSelection technique. This approach enabled the selection of the most suitable hyperparameter values for each classifier, ensuring optimal performance within the stacking ensemble.

Table 2. Hyperparameters used for classifier tuning with CVParameterSelection

Classifier	PID dataset	ESDRP dataset
NB	No hyperparameters	No hyperparameters
LR	R 0.01 (Ridge Parameter), M 4 (Max Iterations)	R 0.01, M 4
RF	Bagsizepercent 100, Batch size 100, Max Depth 100, Num Features 8, Num Iterations 200	Bagsizepercent 100, Batch size 100, Max Depth 100, Num Features 8, Num Iterations 200
k-NN	F (Weight by 1-Distance), k 5 (Number of Neighbors)	F (Weight by 1-Distance), k 1
ANN	Number of hidden layers (a): 8, Batch size: 1000, Epochs: 560	Number of hidden layers (a): 16, Batch size: 1000, Epochs: 560
AdaBoost+SVM	P 10 (Weight Threshold), I 1 (Number of Iterations)	P 100, I 1

Naïve Bayes (NB) is a probabilistic classifier that assumes independence between features and calculates the probability of a sample belonging to a specific class using Bayes' theorem. It is computationally efficient and works well with high-dimensional data but may oversimplify complex relationships between features [30]. To apply naïve Bayes in WEKA, researchers can utilize the naïve Bayes classifier available in the platform's library.

Logistic regression (LR) is a linear classifier that models the connection between the input features and the sample's likelihood of belonging to a certain class. It is interpretable, and supports categorical and continuous data, but presupposes a linear relationship between features and the target variable's log odds [31]. In WEKA, the Logistic classifier can be used to apply logistic regression.

k-nearest neighbors (k-NN) classifier is a non-parametric algorithm that assigns a sample to the predominant class based on the class labels of its k closest neighbors in the feature space. It is easy to comprehend and does not require model training, but it can be sensitive to the selection of k and may be subject to the curse of dimensionality for high-dimensional data [32]. In WEKA, the instance-based learning with k-NN (IBk) classifier can be used for k-NN classification.

Artificial neural networks (ANNs) are a class of mathematical models that simulate the behavior of biological neural networks. Composed of interconnected nodes called neurons, organized in layered architectures, ANNs aim to unravel intricate relationships between input features and target variables. ANN can capture non-linear relationships, but require careful architecture design, and training data, and can be computationally intensive [33], [34]. In WEKA, researchers can utilize the MultilayerPerceptronClassifier to implement ANNs.

AdaBoost with support vector machine is a boosting-based classifier that combines multiple weak classifiers to create a strong classifier. In this case, SVMs are used as the weak classifiers. AdaBoost iteratively adjusts the weights of misclassified samples to focus on difficult-to-classify instances. SVMs provide robust classification boundaries but may be sensitive to the choice of kernel and hyperparameters [35]. In WEKA, researchers can apply AdaBoost with SVM using the AdaBoostM1 classifier.

Random forest (RF) is a classifier that uses a bagging approach to aggregate the predictions of many DTs. It utilizes bootstrapping and feature randomization to reduce overfitting and improve generalization. It handles both categorical and continuous features and provides feature importance measures, but can be computationally expensive for large datasets [36]. In WEKA, the random forest classifier can be used to apply random forest classification.

3. RESULTS AND DISCUSSION

After the evaluation in WEKA, the performance metrics for each hyper-tuned model were obtained. These metrics typically include accuracy, precision, recall, F-measure, mean absolute error (MAE), and area under the curve (AUC). For the ESDRP dataset, the hyper-tuned models were individually optimized using the CVParameterSelection technique in WEKA to maximize their performance before being combined in the proposed stacking-based hybrid model. The evaluation results showcased promising outcomes, with each model demonstrating its strengths and areas of improvement. Table 3 presents a comprehensive performance analysis, highlighting the accuracy achieved by each hyper-tuned model. Notably, the proposed stacking-based hybrid model stands out with an impressive accuracy of 99.7222%. This result surpasses the other hyper-tuned models, including NB (92.5926%), LR (93.5185%), RF (99.0741%), k-NN (98.6111%), ANN (96.2963%), and AdaBoost with SVM (93.9815%). These findings emphasize the effectiveness of the hyper-tuning process and the potential of the stacking ensemble approach.

Moving on to the PID dataset, similar efforts were made to optimize the hyper-tuned models using the CVParameterSelection technique in WEKA. The performance analysis in Table 4 provides valuable insights into the performance of each model on this specific dataset. The proposed stacking-based hybrid model demonstrates remarkable accuracy, achieving a score of 94.2085%. This accuracy outperforms other hyper-tuned models such as ANN (87.4598%), NB (77.8135%), LR (82.9582%), RF (90.9968%), k-NN (81.0289%), and AdaBoost with SVM (81.672%). These results underscore the importance of the hyperparameter tuning process and its impact on the models' performance.

The outstanding performance of the proposed model can be attributed to its stacking-based hybrid approach, which combines the predictions of the hyper-tuned base learners, effectively leveraging their strengths. It is important to note that each base learner was hyper-tuned independently to optimize its performance, ensuring that the model benefits from the best possible configurations for each classifier. The use of the RF meta-classifier in the stacking ensemble further enhances the model's predictive capability.

Table 3. Comparative analysis of models performance on the ESDRP dataset

Model	Accuracy (%)	MAE	Precision	Recall	F-measure	AUC
ANN	96.2963	0.0418	0.963	0.963	0.963	0.974
NB	92.5926	0.0867	0.927	0.937	0.926	0.967
LR	93.5185	0.0803	0.936	0.935	0.935	0.958
RF	99.0741	0.0348	0.991	0.991	0.991	0.999
k-NN	98.6111	0.015	0.986	0.986	0.986	0.999
AdaBoost + SVM	93.9815	0.0602	0.940	0.940	0.940	0.937
Proposed	99.7222	0.002	0.997	0.997	0.997	1

Table 4. Comparative analysis of models performance on the PID dataset

Model	Accuracy (%)	MAE	Precision	Recall	F-measure	AUC
ANN	87.4598	0.1438	0.875	0.875	0.875	0.912
NB	77.8135	0.2717	0.780	0.778	0.778	0.841
LR	82.9582	0.2526	0.830	0.830	0.830	0.868
RF	90.9968	0.1534	0.910	0.910	0.910	0.965
k-NN	81.0289	0.241	0.817	0.810	0.808	0.884
AdaBoost + SVM	81.672	0.1833	0.817	0.817	0.817	0.816
Proposed	94.2085	0.0967	0.945	0.942	0.943	0.984

4. COMPARISON WITH THE EXISTING LITERATURE

The comparison with existing literature reveals a diverse range of models and methodologies proposed for the prediction of early-stage DM. However, the proposed model, which utilizes a stacking ensemble approach, demonstrates superior performance in DM prediction when compared to the existing models. This highlights the effectiveness and potential of the novel approach in enhancing the accuracy and reliability of early-stage DM prediction. According to Table 5, the proposed model achieves an accuracy of 94.2085% on the dataset and 99.7222% on the ESDRP dataset. This reflects a substantial improvement over the existing models, with the proposed model outperforming them by absolute differences ranging from 10.2085% to 16.7222% in terms of accuracy.

Table 5. Comparison of models for early-stage DM prediction in existing literature

Reference	Model	Methodology	Accuracy in % (Dataset)
Proposed	Level 0 (ANN, NB, LR, AdaBoost + SVM, k-NN), Level 1 (RF)	Stacking ensemble	94.2085% (PID dataset), and 99.7222% (ESDRP dataset)
[21]	Level 0 (LR, DT, RF, gradient boosting), Level 1 (SVM)	Stacked ensemble	92% (PID dataset), 99.6% (ESDRP dataset), and 98% (Diabetes 130-US hospitals dataset)
[22]	LR, RF, SVM, and KNN with grid search	Parameter optimization using grid search	83% (PID dataset)
[23]	Level 0 (RF, DT, gradient boost, Gaussian NB, SVM KNN), Level 1 (LR)	Stacking ensemble	93% (PID dataset)
[24]	LR, RF, extreme gradient boosting	Voting ensemble	92.21% (PID dataset)
[25]	LR, SVM, k-NN, gradient boost, NB, and RF with majority voting	Voting ensemble	82% (PID dataset)

5. CONCLUSION

In conclusion, this research work introduces a novel stacking-based hybrid machine learning approach for accurately predicting early-stage DM. The approach combines multiple base learners at level 0 and utilizes an RF meta-classifier at level 1 to effectively aggregate their predictions. The obtained high accuracy rates on the early-stage DM and PID datasets highlight the effectiveness of the proposed model in predicting early-stage DM. The proposed approach demonstrates significant improvements over existing literature, outperforming them by absolute differences ranging from 10.2085% to 16.7222% in terms of accuracy. This substantial enhancement in accuracy showcases the superiority of the proposed model. The findings of this study have several implications. Firstly, the high accuracy rates achieved by the proposed model indicate its potential as a valuable tool in aiding early intervention and prevention strategies for DM. Timely identification of individuals at risk can enable proactive healthcare management and improve patient outcomes. Secondly, the stacking-based hybrid approach proves to be an effective methodology for integrating the predictions of multiple base learners, leveraging their strengths and enhancing overall performance. Future research endeavors should focus on further validating the generalizability of the proposed model on larger and more diverse datasets. Additionally, exploring feature importance analysis techniques can enhance the interpretability of the model, enabling better insights into the factors influencing early-stage DM.

REFERENCES





- [1] F. Alanazi and V. Gay, "e-health for diabetes self-management in Saudi Arabia: barriers and solutions," *Proceedings of the 36th International Business Information Management Association Conference (IBIMA)*, pp. 4-5, Feb. 2020, doi: 10.2196/preprints.18085.
- [2] M. S. Paulo, N. M. Abdo, R. Bettencourt-Silva, and R. H. Al-Rifai, "Gestational diabetes mellitus in europe: a systematic review and meta-analysis of prevalence studies," *Frontiers in Endocrinology*, vol. 12, Dec. 2021, doi: 10.3389/fendo.2021.691033.
- [3] T. L. Haraldsdottir *et al.*, "Diabetes mellitus prevalence in tuberculosis patients and the background population in Guinea-Bissau: a disease burden study from the capital Bissau," *Transactions of The Royal Society of Tropical Medicine and Hygiene*, vol. 109,

- no. 6, pp. 400–407, Jun. 2015, doi: 10.1093/trstmh/trv030.
- [4] P. Fasching, “The new ways of preventing and treating diabetes mellitus,” in *Practical Issues in Geriatrics*, Springer International Publishing, 2019, pp. 71–81.
 - [5] A. M. Elshewey, M. Y. Shams, N. El-Rashidy, A. M. Elhady, S. M. Shohieb, and Z. Tarek, “Bayesian optimization with support vector machine model for parkinson disease classification,” *Sensors*, vol. 23, no. 4, Feb. 2023, doi: 10.3390/s23042085.
 - [6] S. Lahmiri, “Integrating convolutional neural networks, kNN, and Bayesian optimization for efficient diagnosis of Alzheimer’s disease in magnetic resonance images,” *Biomedical Signal Processing and Control*, vol. 80, Feb. 2023, doi: 10.1016/j.bspc.2022.104375.
 - [7] Q. Abbas, A. Hussain, and A. R. Baig, “CAD-ALZ: a blockwise fine-tuning strategy on convolutional model and random forest classifier for recognition of multistage alzheimer’s disease,” *Diagnostics*, vol. 13, no. 1, Jan. 2023, doi: 10.3390/diagnostics13010167.
 - [8] R. Sawhney, A. Malik, S. Sharma, and V. Narayan, “A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease,” *Decision Analytics Journal*, vol. 6, Mar. 2023, doi: 10.1016/j.dajour.2023.100169.
 - [9] G. Annuzzi *et al.*, “Impact of nutritional factors in blood glucose prediction in type 1 diabetes through machine learning,” *IEEE Access*, vol. 11, pp. 17104–17115, 2023, doi: 10.1109/ACCESS.2023.3244712.
 - [10] K. Liu *et al.*, “Machine learning models for blood glucose prediction in patients with Diabetes Mellitus: a systematic review and network meta-analysis,” *SSRN Electronic Journal*, 2023, doi: 10.2139/ssrn.4401684.
 - [11] D. Srivastava, H. Pandey, and A. K. Agarwal, “Complex predictive analysis for health care: a comprehensive review,” *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 12, no. 1, pp. 521–531, Feb. 2023, doi: 10.11591/eei.v12i1.4373.
 - [12] H. Pallathadka, M. Mustafa, D. T. Sanchez, G. S. Sajja, S. Gour, and M. Naved, “Impact of machine learning on management, healthcare and agriculture,” *Materials Today: Proceedings*, vol. 80, pp. 2803–2806, 2023, doi: 10.1016/j.matpr.2021.07.042.
 - [13] S. Yang, P. Varghese, E. Stephenson, K. Tu, and J. Gronsbell, “Machine learning approaches for electronic health records phenotyping: a methodical review,” *Journal of the American Medical Informatics Association*, vol. 30, no. 2, pp. 367–381, Jan. 2023, doi: 10.1093/jamia/ocac216.
 - [14] T. A. Assegie, T. Karpagam, S. Subramanian, S. M. Janakiraman, J. Arumugam, and D. O. Ahmed, “Prediction of patient survival from heart failure using a cox-based model,” *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 27, no. 3, pp. 1550–1556, Sep. 2022, doi: 10.11591/ijeecs.v27.i3.pp1550-1556.
 - [15] Aman and R. S. Chhillar, “Analyzing predictive algorithms in data mining for cardiovascular disease using WEKA tool,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, pp. 144–150, 2021, doi: 10.14569/IJACSA.2021.0120817.
 - [16] A. Darolia and R. S. Chhillar, “Analyzing three predictive algorithms for diabetes mellitus against the pima indians dataset,” *ECS Transactions*, vol. 107, no. 1, pp. 2697–2704, Apr. 2022, doi: 10.1149/10701.2697ecst.
 - [17] M. Atif, F. Anwer, F. Talib, R. Alam, and F. Masood, “Analysis of machine learning classifiers for predicting diabetes mellitus in the preliminary stage,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 12, no. 3, pp. 1302–1311, Sep. 2023, doi: 10.11591/ijai.v12.i3.pp1302-1311.
 - [18] K. Yothapakdee, S. Charoenkhum, and T. Boonnuk, “Improving the efficiency of machine learning models for predicting blood glucose levels and diabetes risk,” *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 27, no. 1, Jul. 2022, doi: 10.11591/ijeecs.v27.i1.pp555-562.
 - [19] G. Mathur, A. Pandey, and S. Goyal, “Applications of machine learning in healthcare,” in *The Internet of Medical Things (IoMT) and Telemedicine Frameworks and Applications*, 2022, pp. 177–195.
 - [20] S. A. Abdulkareem, H. Y. Radhi, Y. A. Fadil, and H. Mahdi, “Soft computing techniques for early diabetes prediction,” *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 25, no. 2, pp. 1167–1176, Feb. 2022, doi: 10.11591/ijeecs.v25.i2.pp1167-1176.
 - [21] A. Doğru, S. Buyrukoglu, and M. Ari, “A hybrid super ensemble learning model for the early-stage prediction of diabetes risk,” *Medical, Biological Engineering, and Computing*, vol. 61, no. 3, pp. 785–797, Mar. 2023, doi: 10.1007/s11517-022-02749-z.
 - [22] R. Krishnamoorthi *et al.*, “A novel diabetes healthcare disease prediction framework using machine learning techniques,” *Journal of Healthcare Engineering*, vol. 2022, pp. 1–10, Jan. 2022, doi: 10.1155/2022/1684017.
 - [23] S. R., S. M., M. K. Hasan, R. A. Saeed, S. A. Alsubhany, and S. Abdel-Khalek, “An empirical model to predict the diabetic positive using stacked ensemble approach,” *Frontiers in Public Health*, vol. 9, Jan. 2022, doi: 10.3389/fpubh.2021.792124.
 - [24] S. R. S. Chakravarthy and H. Rajaguru, “Ensemble-based weighted voting approach for the early diagnosis of diabetes mellitus,” in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 93, Springer Nature Singapore, 2022, pp. 451–460.
 - [25] Z. Mushtaq, M. F. Ramzan, S. Ali, S. Baseer, A. Samad, and M. Husnain, “Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques,” *Mobile Information Systems*, vol. 2022, pp. 1–16, Mar. 2022, doi: 10.1155/2022/6521532.
 - [26] Weka Wiki, “Weka 3 - data mining with open source machine learning software in Java.” University of Waikato, <https://www.cs.waikato.ac.nz/ml/weka/> (accessed May 22, 2023).
 - [27] D. Carrion, “Pima-Indians-Diabetes,” OpenML, 2022. Accessed: May 22, 2023. [Online], Available: <https://www.openml.org/search?type=data&status=active&id=43582&sort=runs>
 - [28] D. Carrion, “Early-stage-diabetes-risk-prediction-dataset,” OpenML, 2022. Accessed May 22, 2023. [Online], Available: <https://www.openml.org/search?type=data&status=active&id=43643&sort=runs>
 - [29] T. Yoon and D. Kang, “Multi-modal stacking ensemble for the diagnosis of cardiovascular diseases,” *Journal of Personalized Medicine*, vol. 13, no. 2, Feb. 2023, doi: 10.3390/jpm13020373.
 - [30] R. Rajni and A. Amandeep, “RB-Bayes algorithm for the prediction of diabetic in Pima Indian dataset,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 6, pp. 4866–4872, Dec. 2019, doi: 10.11591/ijece.v9i6.pp4866-4872.
 - [31] Z. Li, S. Pang, H. Qu, and W. Lian, “Logistic regression prediction models and key influencing factors analysis of diabetes based on algorithm design,” *Neural Computing and Applications*, Mar. 2023, doi: 10.1007/s00521-023-08447-7.
 - [32] S. Avinash, H. N. N. Kumar, M. S. G. Prasad, R. M. Naik, and G. Parveen, “Early detection of malignant tumor in lungs using feed-forward neural network and k-nearest neighbor classifier,” *SN Computer Science*, vol. 4, no. 2, Feb. 2023, doi: 10.1007/s42979-022-01606-y.
 - [33] S. P. Shankar, M. S. Supriya, D. Varadam, M. Kumar, H. Gupta, and R. Saha, “A comprehensive study on algorithms and applications of artificial intelligence in diagnosis and prognosis: AI for healthcare,” in *Digital Twins and Healthcare: Trends, Techniques, and Challenges*, 2022, pp. 35–54.





- [34] G. Alfian, Y. M. Saputra, L. Subekti, A. D. Rahmawati, F. T. D. Atmaji, and J. Rhee, "Utilizing deep neural network for web-based blood glucose level prediction system," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 30, no. 3, pp. 1829–1837, Jun. 2023, doi: 10.11591/ijeecs.v30.i3.pp1829-1837.
- [35] A. Belghit, M. Lazri, F. Ouallouche, K. Labadi, and S. Ameer, "Optimization of one versus all-SVM using AdaBoost algorithm for rainfall classification and estimation from multispectral MSG data," *Advances in Space Research*, vol. 71, no. 1, pp. 946–963, Jan. 2023, doi: 10.1016/j.asr.2022.08.075.
- [36] R. R. K. AL-Taie, B. J. Saleh, A. Y. Falih Saedi, and L. A. Salman, "Analysis of WEKA data mining algorithms Bayes net, random forest, MLP and SMO for heart disease prediction system: a case study in Iraq," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 6, pp. 5229–5239, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5229-5239.

BIOGRAPHIES OF AUTHORS



Aman     is a researcher in the field of Data Mining at the Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak, India. His research focuses on data science, healthcare, metaheuristic algorithms, and cyber security. He has published his work on prestigious platforms such as Scopus, web of science (WoS), and respected conferences. Additionally, he has authored a book on MATLAB. He is a member of the International Association of Engineers (IAENG). He can be contacted at email: sei@live.in.



Rajender Singh Chhillar     is a professor and former head of the Department of Computer Science at Maharshi Dayanand University, Rohtak, India. He obtained his Ph.D. in Computer Science from Maharshi Dayanand University, Rohtak, India, and a master's degree from Kurukshetra University, Kurukshetra, India. Additionally, he holds a Master of Business Administration (MBA) degree from Sikkim Manipal University, Sikkim, India. His research interests include software engineering, software testing, software metrics, web metrics, bio metrics, data mining, computer networking, and software design. He has published over 100 journal papers and 65 conference papers in recent years. Rajender has also written two books in the fields of software engineering and information technology. He is a director of the CMAI Asia Association in New Delhi and a senior member of the IACSIT in Singapore. Furthermore, he is a member of the Computer Society of India. He can be contacted at email: chhillar02@gmail.com.