HYBRID MACHINE LEARNING MODELS FOR DIABETES PREDICTION: A SYSTEMATIC LITERATURE REVIEW

Mukhriddin Arabboev
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan
Shohruh Begmatov
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan

Abstract. Diabetes mellitus is a chronic disease requiring timely and accurate diagnosis for effective management. Hybrid machine learning models have emerged as a promising approach for enhancing predictive accuracy. This systematic literature review categorizes existing hybrid models, highlights key trends, and evaluates their effectiveness in diabetes prediction. A taxonomy is proposed to classify hybrid approaches, and an in-depth analysis of feature selection methods, performance metrics, and datasets is provided. Finally, challenges and future research directions are discussed.

Keywords: diabetes prediction, hybrid machine learning models, feature selection, ensemble learning, Pima Indians Diabetes Dataset, deep learning, medical diagnosis, classification accuracy, systematic literature review.

Аннотация. Сахарный диабет — хроническое заболевание, требующее своевременной и точной диагностики для эффективного лечения. Гибридные модели машинного обучения стали перспективным подходом для повышения точности прогнозирования. Этот систематический обзор литературы классифицирует существующие гибридные модели, выделяет ключевые тенденции и оценивает их эффективность в прогнозировании диабета. Предлагается таксономия для классификации гибридных подходов и дается углубленный анализ методов выбора признаков, показателей производительности и наборов данных. Наконец, обсуждаются проблемы и будущие направления исследований.

Ключевые слова: прогнозирование диабета, гибридные модели машинного обучения, отбор признаков, ансамблевое обучение, Pima Indians Diabetes Dataset, глубокое обучение, медицинская диагностика, точность классификации, систематический обзор литературы.

Annotatsyia. Qandli diabet-bu samarali davolanish uchun o'z vaqtida va aniq tashxis qo'yishni talab qiladigan surunkali kasallik. Mashinani o'rganishning gibrid modellari bashorat qilish aniqligini oshirish uchun istiqbolli yondashuvga aylandi. Ushbu tizimli adabiyotlarni ko'rib chiqish mavjud gibrid modellarni tasniflaydi, asosiy tendentsiyalarni ta'kidlaydi va ularning diabetni bashorat qilishda

ISSN 3030-3907

samaradorligini baholaydi. Gibrid yondashuvlarni tasniflash uchun taksonomiya taklif etiladi va xususiyatlar, ishlash ko'rsatkichlari va ma'lumotlar to'plamlarini tanlash usullarini chuqur tahlil qiladi. Va nihoyat, muammolar va kelajakdagi tadqiqot yo'nalishlari muhokama qilinadi.

Kalit so'zlar: diabetni bashorat qilish, mashinani o'rganishning gibrid modellari, xususiyatlarni tanlash, ansamblni o'rganish, Pima Indians Diabetes Dataset, chuqur o'rganish, tibbiy diagnostika, tasniflash aniqligi, adabiyotlarni tizimli ko'rib chiqish.

1. Introduction

Diabetes is a major public health concern affecting millions worldwide. Traditional diagnostic methods rely on clinical tests, which can be expensive and time-consuming. Machine learning models, particularly hybrid approaches, offer an efficient alternative by utilizing patient data for early detection. Hybrid machine learning models integrate multiple algorithms to enhance accuracy, robustness, and generalizability. This paper systematically reviews existing literature on hybrid models for diabetes prediction, incorporating insights from 20 peer-reviewed journal articles retrieved exclusively from the Scopus database to provide a comprehensive taxonomy and discuss key findings, challenges, and future directions.

2. Methodology

A systematic review was conducted using Scopus, ensuring high-quality peer-reviewed journal articles. The search strategy included the following keywords: "hybrid methods" AND "diabetes" AND "Pima", with filters applied to select English-language, peer-reviewed, open-access journal articles published between 2010-2024.

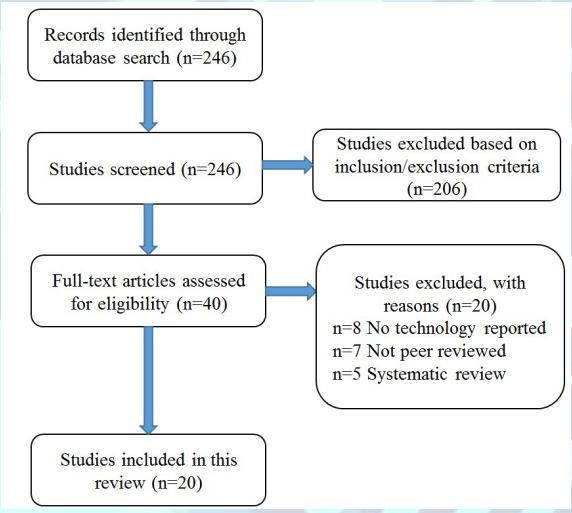


Figure 1. Literature screening and inclusion process for the systematic review.

ISSN 3030-3907

Figure 1 illustrates the step-by-step process of study selection conducted for the systematic review. A total of 246 records were initially identified through a database search. After applying inclusion and exclusion criteria, 206 studies were excluded. Of the remaining 40 full-text articles assessed for eligibility, 20 were excluded due to lack of technological focus (n=8), absence of peer review (n=7), or being systematic reviews themselves (n=5). Ultimately, 20 studies were included in the final review.

2.1. Inclusion and Exclusion Criteria

The inclusion and exclusion criteria were established to ensure the relevance, quality, and focus of the studies included in this systematic review. Only peer-reviewed journal articles published in English and available as open-access were considered, ensuring both academic rigor and accessibility. The review specifically targeted studies that applied hybrid machine learning models for diabetes prediction using real-world medical datasets sourced from the Scopus database. Studies were excluded if they were conference papers, book chapters, non-English, subscription-based, or focused on non-hybrid or standalone machine learning approaches. These criteria helped maintain consistency in the reviewed literature and aligned with the study's objective of evaluating hybrid model effectiveness in clinical prediction settings.

Table 1

Eligibility criteria applied in the systematic review

Criteria	Inclusion	Exclusion
Language	English only	Non-English papers
Publication Type	Peer-reviewed journal articles (research)	Conference papers, book chapters
Availability	Open-access articles	Subscription-based papers
Topic	Hybrid machine learning models for diabetes	Standalone ML models, non-hybrid approaches
Data Source	Scopus database (real medical datasets)	Non-Scopus databases

3. Taxonomy of Hybrid Machine Learning Models for Diabetes Prediction

To systematically analyze the diverse approaches in hybrid machine learning for diabetes prediction, this section introduces a structured taxonomy. The classification categorizes hybrid models based on their underlying methodologies, including combinations of neural networks with evolutionary algorithms, decision tree-based ensembles, stacked generalization techniques, and deep learning integrations. Additionally, models emphasizing feature selection, imputation, and classification are presented as a distinct category due to their relevance in handling

ISSN 3030-3907

real-world medical data challenges. This taxonomy provides a comprehensive framework for comparing models and understanding the methodological trends shaping hybrid AI solutions in diabetes diagnosis and risk prediction.

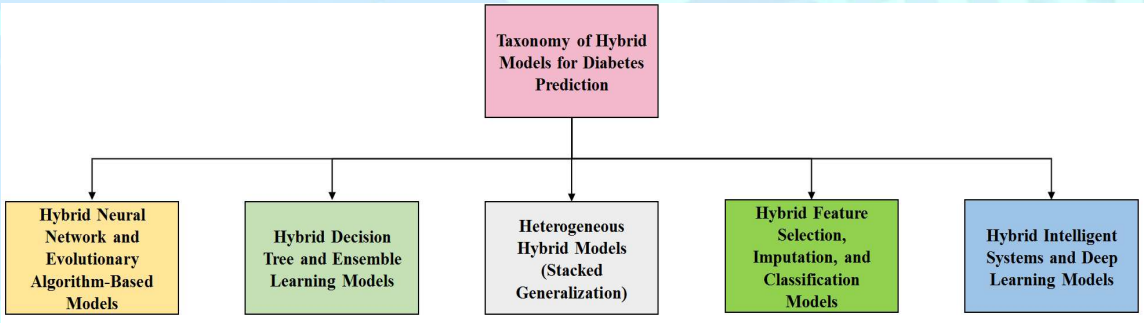


Figure 2. Taxonomy of hybrid approaches for diabetes prediction used in this work.

Figure 2 illustrates the taxonomy of hybrid machine learning approaches analyzed in this systematic review. The models are categorized into five main groups: (1) Hybrid Neural Network and Evolutionary Algorithm-Based Models, (2) Hybrid Decision Tree and Ensemble Learning Models, (3) Heterogeneous Hybrid Models using Stacked Generalization, (4) Hybrid Feature Selection, Imputation, and Classification Models, and (5) Hybrid Intelligent Systems and Deep Learning Models. This classification provides a structured framework for understanding the diverse techniques applied in diabetes prediction research.

3.1. Hybrid Neural Network and Evolutionary Algorithm-Based Models

This category of hybrid models integrates neural networks with evolutionary algorithms to enhance feature selection, data imputation, and classification accuracy. The synergy between adaptive learning and optimization techniques allows these models to handle noisy, incomplete, or high-dimensional datasets more effectively. This section explores various models, including Hopfield networks, genetic algorithms, and multilayer perceptrons, which have shown remarkable improvements in predictive performance.

In [1], the authors propose a hybrid Evolutionary Correlated Gravitational Search Algorithm (ECGS) with Genetic Optimized Hopfield Neural Network (GHNN) for diabetes prediction. The study utilizes the Pima Indians Diabetes dataset, which undergoes data preprocessing using mean mode normalization to handle noise. Following this, feature selection and dimensionality reduction are performed by computing correlation, feature mass, and direction change to retain the most relevant diabetic abnormality features. The selected features are then classified using GHNN, effectively identifying diabetes-related abnormalities. The performance of the proposed ECGS-GHNN model is evaluated through experimental validation, demonstrating improved accuracy in diabetes classification. In [2], a hybrid prediction model with missing value imputation

ISSN 3030-3907

(HPM-MI) is introduced to handle missing data in medical datasets. The model integrates K-means clustering with Multilayer Perceptron (MLP) for imputation and classification, ensuring enhanced data quality. K-means clustering validates class labels, removing incorrectly classified instances before applying the classifier. The model achieves 99.82% accuracy, 100% sensitivity, and 99.74% specificity on the Pima Indians Diabetes dataset, outperforming existing methods. A comparative analysis of eleven imputation techniques confirms the effectiveness of HPM-MI for predictive accuracy in medical diagnosis.

In [3], a customized decision-making algorithm is introduced for diabetes prediction using the Pima Indians Diabetes Dataset, addressing key limitations of previous studies that rely on general-purpose packages. The study identifies that missing values are often misrepresented as zeros, requiring a specialized handling approach to improve data accuracy. A novel AI-driven feature prioritization technique is developed, ranking blood sugar levels, BMI, age, and pregnancy count as the most significant factors, aligning with medical intuition. Despite working with incomplete data, the model achieves 80%-81% accuracy, closely matching real-world medical diagnoses. The algorithm favors positive values over negative ones, ensuring a more reliable prediction process. In [4], a hybrid attribute optimization model called Enhanced and Adaptive Genetic Algorithm with Multilayer Perceptron (EAGA-MLP) is proposed for Type 2 diabetes prediction. The model applies EAGA for attribute optimization, reducing irrelevant information and improving dataset reliability for classification. The optimized dataset is used by MLP to classify diabetes presence, enhancing predictive accuracy. The EAGA-MLP model is validated on seven different disease datasets, achieving a maximum accuracy of 97.76% with an execution time of 1.12 seconds. The model outperforms existing classifiers, with a mean accuracy of 94.7%, precision of 91%, recall of 89.8%, and F-score of 90.4%. These results confirm that EAGA-MLP effectively aids medical experts in diagnosing Type 2 diabetes and other diseases by optimizing feature selection and classification accuracy.

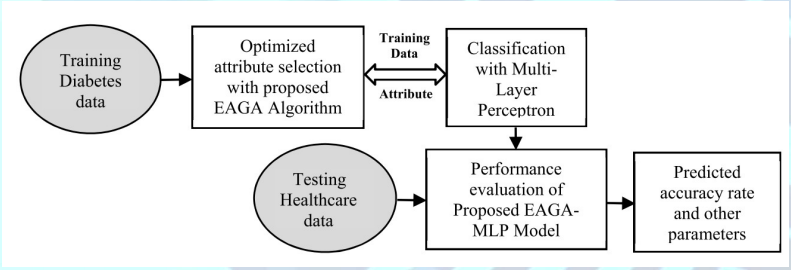


Figure 3. Hybrid EAGA-MLP model for diabetes risk classification [4]

3.2. Hybrid Decision Tree and Ensemble Learning Models

ISSN 3030-3907

Hybrid models based on decision trees and ensemble methods combine interpretability with high classification accuracy. These systems often utilize techniques like random forests, boosting algorithms, and fuzzy logic to balance predictive strength and clinical usability. This section examines how integrating decision trees with ensemble learners can offer robust solutions for diabetes prediction, especially in early diagnosis scenarios.

In [5], a hybrid intelligent system combining Fuzzy Min–Max Neural Network (FMMNN), Classification and Regression Tree (CART), and Random Forest (RF) is proposed for medical data classification. The model leverages FMMNN for incremental learning, CART for explainability, and RF for high classification accuracy. It is evaluated on three benchmark medical datasets (Breast Cancer Wisconsin, Pima Indians Diabetes, and Liver Disorders) using accuracy, sensitivity, specificity, and AUC metrics. Experimental results confirm its effectiveness in medical data classification, outperforming existing methods. In [6], a hybrid classification model combining Genetic Algorithm (GA) for feature selection and J48graft Decision Tree (J48graft DT) for classification is proposed for diabetes diagnosis. The two-stage approach first applies GA to reduce data dimensionality, improving storage efficiency and computational cost. The selected features are then classified using J48graft DT, enhancing predictive accuracy. Evaluated on the Pima Indians Diabetes Dataset, the model achieves 74.78% accuracy and a ROC value of 0.786, demonstrating its effectiveness for early and cost-effective diabetes diagnosis.

In [7], a super ensemble learning model is proposed for early diabetes diagnosis, leveraging a cross-validation-based Super Learner approach that integrates multiple machine learning algorithms. The model consists of four base learners (Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting) and a Support Vector Machine (SVM) as the meta-learner. Feature selection is optimized using Chi-square, and hyperparameters are fine-tuned using GridSearch. The model is evaluated on three datasets (Early-Stage Diabetes Risk, PIMA, and Diabetes 130-US Hospitals), achieving 99.6%, 92%, and 98% accuracy, respectively. The results confirm that Super Learner ensembles enhance predictive accuracy, making them a robust and efficient approach for diabetes detection. In [8], a hybrid ensemble model is proposed for early prediction of Type 2 Diabetes Mellitus, aiming to alert patients in advance and reduce associated health risks. The model integrates supervised classification algorithms (Naïve Bayes) with ensemble learning techniques, including bagging with Random Forest and Adaboost for Decision Tree, to improve accuracy and precision. By aggregating multiple classifiers, the approach enhances predictive performance and

ISSN 3030-3907

helps physicians efficiently diagnose and manage diabetes. The system also features real-time patient notifications, allowing individuals to take immediate action based on risk predictions.

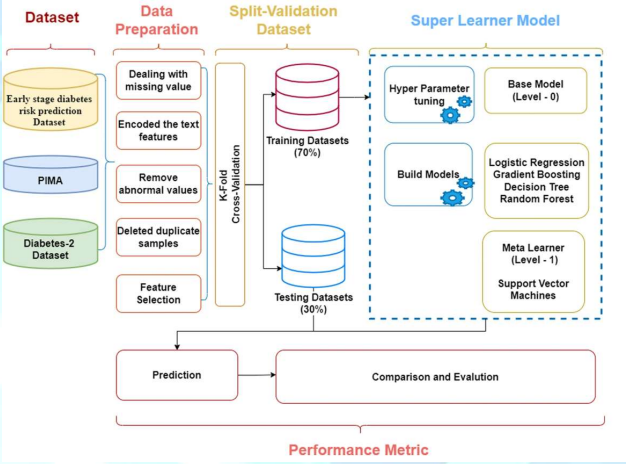


Figure 4. Early diabetes prediction using a hybrid super learner ensemble model [7]

3.3. Heterogeneous Hybrid Models (Stacked Generalization)

Heterogeneous hybrid models leverage stacked generalization to combine diverse learning algorithms through meta-learning strategies. These models outperform conventional approaches by integrating strengths from multiple classifiers, often using optimization algorithms to refine ensemble architecture. This section discusses various implementations of stacked generalization and their effectiveness in capturing complex patterns in diabetes datasets.

In [9], a comparative evaluation of diabetes prediction models is conducted, analyzing classical machine learning models (MLP, KNN, SVM, NB), homogenous hybrid models (RF, AdaBoost, XGBoost, Extra Trees, GBT), and a heterogeneous hybrid model using Stacked Generalization (SG). The SG algorithm is employed to select the optimal base classifiers, identifying AdaBoost, XGBoost, and GBT, with GBT serving as the meta-classifier. Experimental results demonstrate that the SG model outperforms both classical and homogenous hybrid models, achieving 83.9% accuracy on the Pima dataset and 98.5% accuracy on an early-stage diabetes dataset. The findings suggest that stacked generalization is an effective approach for diabetes prediction, assisting clinicians in early diagnosis and decision-making. In [10], a hybrid meta-classifier approach is proposed for Type 2 diabetes prediction, integrating Fuzzy C-means Clustering (FCM) and Logistic Regression (LR). The model follows a two-level ensemble learning framework, where the base learner consists of six machine learning algorithms, and the hybrid meta-learner combines FCM with LR to refine the final prediction. The FCM algorithm generates highly significant clusters of predictions, which are then processed by LR for improved classification accuracy. The model is evaluated on

ISSN 3030-3907

two publicly available datasets (Pima Indians Diabetes Database and Schorling Diabetes Dataset), achieving 99.00% and 95.20% accuracy, respectively, outperforming existing approaches. The results confirm that hybrid meta-classifiers effectively enhance diabetes prediction accuracy, making them valuable tools for medical diagnosis.

In [11], a stacking-based integrated Kernel Extreme Learning Machine (KELM) model is proposed for Type-II diabetes risk prediction. The model utilizes the Pima Indians Diabetes Dataset and a Diabetic Research Center dataset, applying min-max normalization for data preprocessing. The Hybrid Particle Swarm Optimization-Artificial Fish Swarm Optimization (HAFPSO) algorithm is employed to optimize classification accuracy and reduce kernel complexity. A KELM meta-classifier integrates predictions from twenty base learners to enhance accuracy. The model achieves 98.5% accuracy and is validated using multiple performance metrics, including sensitivity, specificity, Matthews Correlation Coefficient, and Kappa Statistics. The results confirm that KELM-HAFPSO is a promising and effective tool for Type-II diabetes prediction, aiding clinicians in early risk assessment.

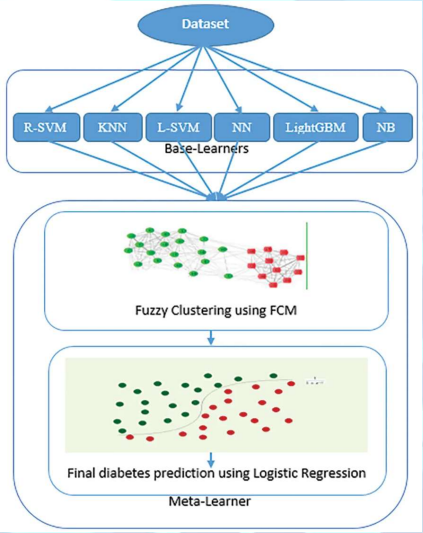


Figure 5. Hybrid meta-learner model (Fuzzy C-Means + LR) [10]

3.4. Hybrid Feature Selection, Imputation, and Classification Models

Accurate diabetes prediction depends heavily on the quality of input features and the handling of missing or imbalanced data. This section focuses on hybrid models that combine feature selection techniques, such as F-score ranking and recursive elimination, with imputation strategies and robust classifiers. These models aim to improve classification accuracy while addressing the limitations of real-world clinical datasets.

In [12], a hybrid imputation model (FPCA–SVM–FCM) is proposed to address missing data challenges in medical datasets, particularly for diabetes

ISSN 3030-3907

prediction. The model integrates Fuzzy Principal Component Analysis (FPCA) for feature selection, Fuzzy C-Means (FCM) for missing data imputation, and Support Vector Machines (SVM) for classification, ensuring improved accuracy and efficiency. Unlike traditional imputation methods, FPCA considers outliers, reducing the impact of irrelevant features on prediction accuracy. The approach is tested on the Pima Indians Diabetes Dataset, demonstrating superior performance compared to existing imputation techniques. The model achieves higher accuracy, lower RMSE, and MAE, and is further validated using Wilcoxon rank sum and Theil's U tests, confirming its effectiveness as a reliable alternative for handling missing data in medical diagnosis. In [13], a hybrid preprocessing method called GASMOTEPSO_ENN is proposed to address class imbalance in medical datasets. This approach combines SMOTE (Synthetic Minority Oversampling Technique) and ENN (Edited Nearest Neighbour) with Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) to optimize the balance and quality of training data. The method is tested on CKD, CSP, and PIMA Indian Diabetes (PID) datasets using classifiers such as Logistic Regression, XGBoost, and SVM. It achieves strong performance with an MCC of 1.00 (CKD), 0.94 (CSP), and 0.87 (PID), showing superior accuracy across all datasets. The results confirm that GASMOTEPSO_ENN significantly improves classification performance in imbalanced medical datasets compared to existing methods. In [14], a hybrid prediction model with F-score feature selection is proposed to enhance the diagnostic accuracy of diabetes using the Pima Indians Diabetes Dataset. The method ranks features using the F-score metric, and selects the optimal subset that minimizes clustering error. These selected features are used for further classification with a Support Vector Machine (SVM), improving performance across key metrics. The model achieves 98.94% accuracy, along with high sensitivity, specificity, and AUC, outperforming existing models in the literature. This study confirms that informative feature selection significantly boosts classification performance for diabetes prediction.

In [15], a hybrid feature selection algorithm combining Recursive Feature Elimination (RFE) with Ridge Regression (L2 regularization) is proposed to address the issue of overfitting in small medical datasets. The approach identifies the most relevant features from the Pima Indians Diabetes Dataset, and the selected features are classified using a Random Forest (RF) classifier. The model is designed to enhance generalization by reducing model variance through L2 regularization. It achieves 100% accuracy and 97% AUC, clearly outperforming existing feature selection and classification methods. The results confirm the method's robustness and effectiveness in diabetes prediction. In [16], a novel

ISSN 3030-3907

hybrid optimization strategy called BGW-CSO-SVM is proposed to improve the accuracy and generalization of diabetes prediction models. The approach integrates Binary Grey Wolf Optimization (BGWO) with Crow Search Optimization (CSO) for enhanced feature selection, addressing the challenges of high-dimensional data. After selecting optimal features from two Pima Indians Diabetes datasets, a Support Vector Machine (SVM) is used for classification. The BGW-CSO technique optimizes the SVM's structure, including the number of hidden neurons, to improve performance. The proposed method achieves 96.62% accuracy, outperforming conventional classifiers such as Soft Voting, Random Forest, and Bootstrap Aggregation, with an average improvement of around 16%. This confirms its strong potential for real-world diabetes risk prediction in healthcare applications. In [17], an improved diabetes prediction algorithm (IDPA) is proposed by combining hierarchical clustering with the Naïve Bayes classification algorithm for early detection of Type-II diabetes. The method first groups patients into diabetic and non-diabetic categories using hierarchical clustering and then filters and classifies them using Naïve Bayes based on group-class value comparison. Several hybrid models were explored (e.g., hierarchical clustering with SVM, Decision Tree, Logistic Regression, and K-means), achieving accuracies between 0.90 and 0.946, while the IDPA model achieved a higher accuracy of 0.96. This demonstrates the method's effectiveness over traditional approaches. The study confirms that IDPA offers an efficient early prediction system, reducing the risk of diabetes-related complications.

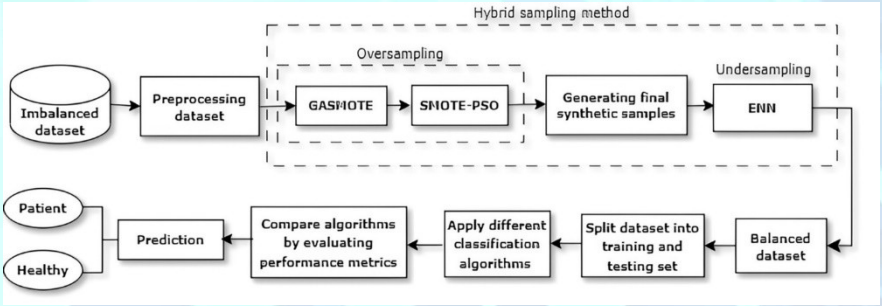


Figure 6. The hybrid heuristic sampling model [13]

3.5. Hybrid Intelligent Systems and Deep Learning Models

The growing availability of computational resources has facilitated the integration of deep learning and intelligent systems into hybrid machine learning frameworks. These advanced models blend conventional algorithms with deep neural architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), often enhanced by meta-heuristic optimization techniques. This section explores the role of such hybrid intelligent systems in improving prediction accuracy, scalability, and real-time applicability in diabetes risk assessment and diagnosis.

ISSN 3030-3907

In [18], a real-time hybrid deep learning model (CNN-BiLSTM) is proposed for the detection and prediction of Type 2 Diabetes Mellitus, using the PIMA Indian Diabetes Dataset. The study first conducts a comparative analysis of various deep learning architectures, then merges Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) to form a superior hybrid model. The proposed CNN-BiLSTM model achieves 98% accuracy, 97% sensitivity, and 98% specificity, outperforming other state-of-the-art approaches by 1.1%. It supports real-time monitoring and provides clinicians with vital, timely information for diabetes management, demonstrating high practical utility in healthcare settings. In [19], an intelligent hybrid deep learning approach is proposed for diabetes detection and blood glucose level prediction using the PIMA and UCI datasets. The model incorporates a meta-heuristic-based CNN architecture with two convolutional and pooling layers, enhanced by a novel Moth Flame-based Crow Search Algorithm (MF-CSA) for multi-objective optimization. MF-CSA optimizes the number of hidden neurons and minimizes feature correlation to enhance feature extraction. A modified fuzzy classifier is used to classify glucose levels as high or low, followed by an enhanced RNN for predicting glucose ranges. Experimental results demonstrate that the MF-CSA hybrid model outperforms existing algorithms, proving effective in improving the accuracy of glucose and diabetes prediction systems. In [20], a hybrid meta-heuristic-enhanced Deep Neural Network (DNN) model is proposed to improve the prediction accuracy of medical datasets, including the Pima Indians Diabetes Dataset. The approach combines two types of meta-heuristic algorithms—population-based (Genetic Algorithm and Differential Evolution) and trajectory-based (Simulated Annealing and Tabu Search)—to balance exploration and exploitation during model training. This hybrid strategy enhances the DNN's ability to search for global optima and avoid overfitting. The model is evaluated on three datasets: anemia, liver disorders, and diabetes, achieving notable performance improvements compared to individual meta-heuristics. The results demonstrate the model's effectiveness in boosting DNN classification accuracy in complex medical prediction tasks.

ISSN 3030-3907

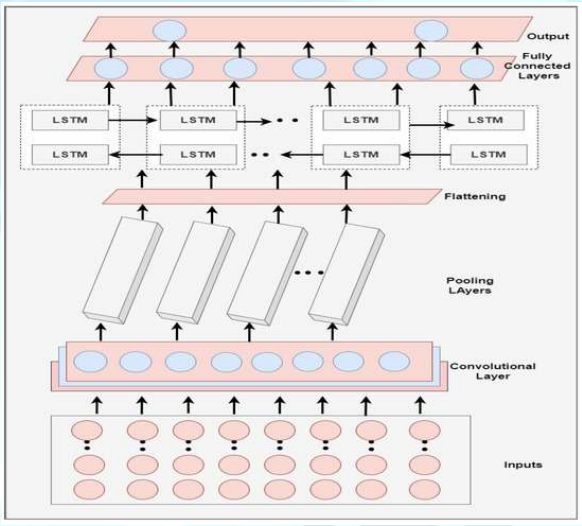


Figure 7. CNN-Bi-LSTM hybrid architecture for diabetes prediction [18]

4. Performance Evaluation

This section presents a comparative analysis of the hybrid models based on key performance metrics, including accuracy, sensitivity, precision, specificity, F1-score, RMSE, and area under the curve (AUC). The evaluation is based on the benchmark Pima Indians Diabetes Dataset, emphasizing both the predictive accuracy of the models and their potential for real-world clinical application. By examining each model's strengths and limitations, this performance analysis identifies the most effective approaches and provides insights to guide the development of more robust, scalable, and clinically relevant diabetes prediction systems.

Table 2

Performance metrics across reviewed hybrid ML models

Reference	Model	Metrics/Accuracy
[1]	ECGS-GHNN	Acc: 99.85 Sens: 99.77%
[2]	HPM-MI	Acc: 99.82% Sensitivity: 100%
[3]	customized hybrid model of artificial neural network (ANN) and genetic algorithm	Acc: 80-81 %
[4]	EAGA-MLP	Acc: 97.76% F1-score: 86.8%
[5]	FMM-CART-RF	AUC: 0.732
[6]	GA-J48graft DT	Acc: 76.52% AUC: 0.765
[7]	Hybrid super ensemble learning model	Acc: 92%
[8]	Hybrid Ensemble Machine Learning	Acc: 96.56 %

ISSN 3030-3907

		Precision:96.12 %
[9]	Stacked Generalization with optimized base learners (AdaBoost, XGBoost, GBT)	Acc: 83.9% Precision: 83.7 %
[10]	Hybrid Meta-Learner Model (Fuzzy C-Means + Logistic Regression)	Acc: 99 %
[11]	KELM-HAFPSO	Acc: 99.92 %
[12]	FPCA–SVM–FCM Model (Feature Selection, Imputation, and Classification)	RMSE: 0.049
[13]	GASMOTEPSO_ENN	Acc: 94 % AUC: 0.94
[14]	Hybrid Prediction Model with F-Score Feature Selection (HPM-FS)	Acc: 98.93 % AUC: 0.9997
[15]	Hybrid Recursive Feature Elimination with Ridge Regression	Acc: 100 % AUC: 0.97
[16]	BGW-CSO-SVM	Acc: 84.63% Recall: 93.8%
[17]	IDPA	Acc: 96 %
[18]	CNN-Bi- LSTM	Acc: 98.85%
[19]	MF-CSA	Acc: 97.4 %
[20]	DNN-DESA	Acc: 84.89 %

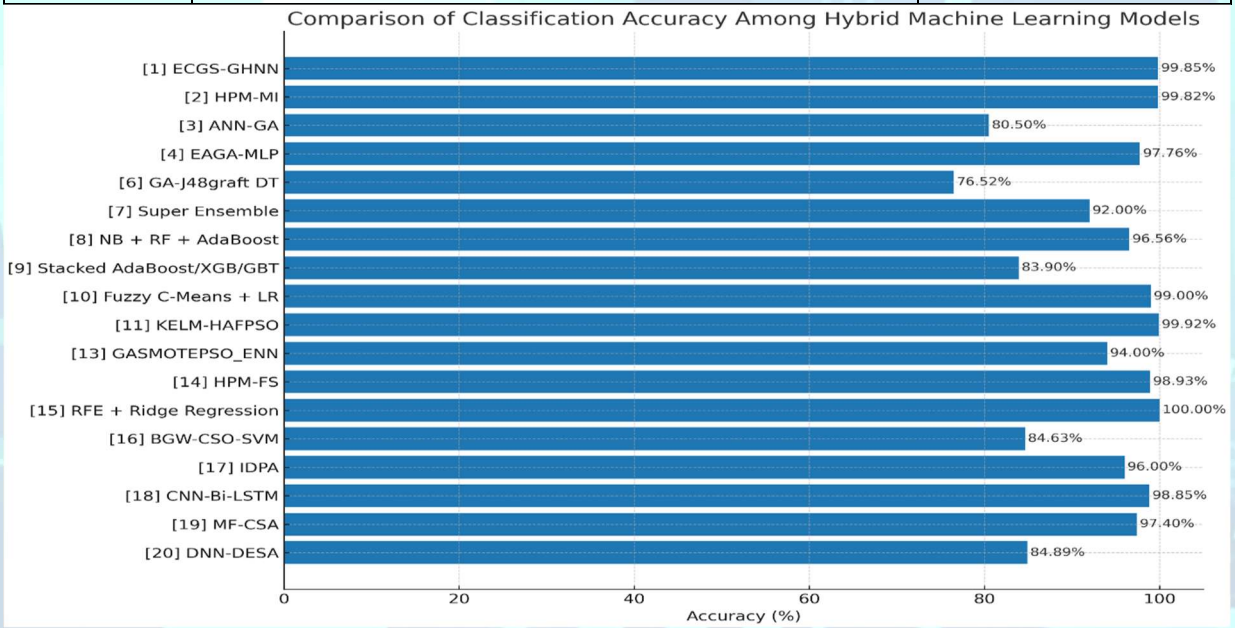


Figure 8. Accuracy performance of reviewed hybrid models for diabetes prediction.

Figure 8 provides a comparative overview of the classification accuracy achieved by various hybrid machine learning models for diabetes prediction. Among the twenty models analyzed, the RFE with Ridge Regression model stands

ISSN 3030-3907

out with a perfect accuracy of 100%, followed closely by KELM-HAFPSO (99.92%), ECGS-GHNN (99.85%), and HPM-MI (99.82%), indicating the effectiveness of combining feature selection, optimization algorithms, and ensemble classifiers. Deep learning-based hybrids such as CNN-BiLSTM and MF-CSA also show high performance, demonstrating the growing potential of deep architectures in medical diagnosis. In contrast, models like GA-J48graft DT (76.52%) and ANN-GA (80.50%) perform less competitively, suggesting room for improvement in traditional hybrid approaches. Overall, Figure 8 underscores the superiority of optimized and ensemble-based hybrid models in achieving high predictive accuracy, reinforcing their relevance in clinical decision-making for diabetes management.

5. Challenges and Future Research Directions

Despite the promising capabilities of hybrid machine learning models in diabetes prediction, several critical challenges hinder their clinical adoption. A primary concern is data quality and availability, as widely used datasets like PIMA often suffer from missing values, class imbalance, and limited diversity. Additionally, many models lack generalizability, having been trained and validated on narrowly scoped benchmark datasets that may not reflect broader patient populations. Another persistent issue is model interpretability—complex hybrid or deep learning systems frequently function as “black boxes,” limiting clinician trust and understanding.

Further, the computational complexity of some hybrid models, especially those employing ensemble or meta-heuristic algorithms, makes them difficult to deploy in resource-constrained environments. Lastly, integration into clinical workflows, including compatibility with electronic health records (EHRs) and real-time decision support systems, is often overlooked in research.

To address these challenges, future work should focus on developing explainable AI (XAI) techniques specific to hybrid models, improving multi-modal learning by combining structured and unstructured clinical data, and incorporating real-time data from wearable devices. Additionally, transfer learning and federated learning approaches can enhance generalization while maintaining data privacy. Emphasis should also be placed on designing lightweight models suitable for edge devices and rural clinics. These advancements are essential for ensuring that hybrid models transition effectively from research to real-world healthcare applications.

6. Conclusion

This review presents a detailed examination of hybrid machine learning models for diabetes prediction, with a focus on studies using the Pima Indians Diabetes Dataset. By integrating methods such as neural networks, decision trees,

ISSN 3030-3907

optimization algorithms, and deep learning, these models demonstrate improved accuracy and robustness. The proposed taxonomy categorizes hybrid approaches to offer a clear understanding of current methodologies.

Hybrid models generally outperform traditional approaches across key performance metrics. However, challenges related to data quality, model transparency, and clinical integration persist. Addressing these gaps is essential for transitioning models into real-world healthcare systems. With ongoing advances in explainable AI, real-time data integration, and computational efficiency, hybrid models are well-positioned to enhance early diabetes diagnosis and support personalized treatment strategies.

REFERENCES

[1] J. Jayashree and S. Ananda Kumar, "Evolutionary correlated gravitational search algorithm (ECGS) with genetic optimized Hopfield neural network (GHNN) – A hybrid expert system for diagnosis of diabetes," *Meas. J. Int. Meas. Confed.*, vol. 145, pp. 551–558, 2019.

[2] A. Purwar and S. K. Singh, "Hybrid prediction model with missing value imputation for medical data," *Expert Syst. Appl.*, vol. 42, no. 13, pp. 5621–5631, 2015.

[3] A. Rajagopal, S. Jha, R. Alagarsamy, S. G. Quek, and G. Selvachandran, "A novel hybrid machine learning framework for the prediction of diabetes with context-customized regularization and prediction procedures," *Math. Comput. Simul.*, vol. 198, pp. 388–406, 2022.

[4] S. Mishra, H. K. Tripathy, and P. K. Mallick, "EAGA-MLP—An Enhanced and Adaptive Hybrid Classification Model for Diabetes Diagnosis," *Sensors (Basel).*, vol. 20, no. 14, pp. 1–31, 2020.

[5] M. Seera and C. P. Lim, "A hybrid intelligent system for medical data classification," *Expert Syst. Appl.*, vol. 41, no. 5, pp. 2239–2249, 2014.

[6] D. K. Choubey and S. Paul, "GA_J48graft DT: A hybrid intelligent system for diabetes disease diagnosis," *Int. J. Intell. Syst. Appl.*, vol. 8, no. 1, pp. 49–59, 2016.

[7] A. Doğru, S. Buyrukoğlu, and M. Arı, "A hybrid super ensemble learning model for the early-stage prediction of diabetes risk," *Med. Biol. Eng. Comput.*, vol. 61, no. 3, pp. 785–797, 2023.

[8] G. Geetha and K. M. Prasad, "An Hybrid Ensemble Machine Learning Approach to Predict Type 2 Diabetes Mellitus," *Webology*, vol. 18, no. SpecialIss, pp. 311–331, 2021.

[9] K. S. M. Anbananthen, M. B. M. A. Busst, R. Kannan, and S. Kannan, "A Comparative Performance Analysis of Hybrid and Classical Machine Learning Method in Predicting Diabetes," *Emerg. Sci. J.*, vol. 7, no. 1, pp. 102–115, 2023.

ISSN 3030-3907

[10] A. A. Taha and S. J. Malebary, "A Hybrid Meta-Classifier of Fuzzy Clustering and Logistic Regression for Diabetes Prediction," *Comput. Mater. Contin.*, vol. 71, no. 2, pp. 6089–6105, 2022.
[11] N. Kanimozhi and G. Singaravel, "Hybrid artificial fish particle swarm optimizer and kernel extreme learning machine for type-II diabetes predictive model," *Med. Biol. Eng. Comput.*, vol. 59, no. 4, pp. 841–867, 2021.
[12] M. F. Dzulkalnine and R. Sallehuddin, "Missing data imputation with fuzzy feature selection for diabetes dataset," *SN Appl. Sci.*, vol. 1, no. 4, pp. 1–12, 2019.
[13] H. Nizam-Ozogur and Z. Orman, "A heuristic-based hybrid sampling method using a combination of SMOTE and ENN for imbalanced health data," *Expert Syst.*, vol. 41, no. 8, 2024.
[14] B. S. Ilango and N. Ramaraj, "A hybrid prediction model with F-score feature selection for type II Diabetes databases," *Proc. 1st Amrita ACM-W Celebr. Women Comput. India, A2CWiC'10*, 2010.
[15] M. Mravik, T. Vetriselvi, K. Venkatachalam, M. Sarac, N. Bacanin, and S. Adamovic, "Diabetes prediction algorithm using recursive ridge regression l2," *Comput. Mater. Contin.*, vol. 71, no. 1, pp. 457–471, 2022.
[16] A. L. Sravanthi, S. Al-Ashmawy, C. Kaur, M. S. A. Ansari, K. A. Saravanan, and V. A. Vuyyuru, "Utilizing Multimodal Medical Data and a Hybrid Optimization Model to Improve Diabetes Prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 11, pp. 754–764, 2023.
[17] S. P. Nimmagadda, S. Yeruva, and R. Siempu, "Improved diabetes prediction model for predicting type-II diabetes," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 12, pp. 230–235, 2019.
[18] P. Madan *et al.*, "An Optimization-Based Diabetes Prediction Model Using CNN and Bi-Directional LSTM in Real-Time Environment," *Appl. Sci.*, vol. 12, no. 8, 2022.
[19] S. Naveena and A. Bharathi, "A new design of diabetes detection and glucose level prediction using moth flame-based crow search deep learning," *Biomed. Signal Process. Control*, vol. 77, no. March, p. 103748, 2022.
[20] F. M. Aswad, K. M. Saffer, and I. Salman, "Evolutionary trajectory hybridization for improving deep learning accuracy in medical data prediction system," *Mater. Today Proc.*, vol. 61, pp. 653–659, 2022.

ISSN 3030-3907