Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 6, 2026

A COMPARATIVE ANALYSIS OF MACHINE LEARNING AND DEEP LEARNING-BASED PREDICTIVE FRAMEWORKS FOR DIABETES CLASSIFICATION USING CLINICAL DATA

[1]Mian Farhan Shah, [*2]Shams Ul Arifeen, [3]Zia Ur Rahman, [4]Ikram Ullah, [5]Ahmad Saeed, [6]Waqas Ahmad, [7]Muhammad Naeem Ullah, [8]Naeem Jan, [9]Atta Ur Rahman

[1]Department of Computer Science, Bacha Khan University, Charsadda
[*2]Department of Computer Science, Abdul Wali Khan University, Mardan
[3]Department of Computer Science, Bacha Khan University, Charsadda
[4]Department of Computer Science, Abdul Wali Khan University, Mardan
[5]Department of Computer Science, Bacha Khan University, Charsadda
[6]Department of Computer Science, Bacha Khan University, Charsadda
[7]Department of Computer Science, Bacha Khan University, Charsadda
[8]Department of Computer Science, Bacha Khan University, Charsadda
[9]Department of Computer Science, Bacha Khan University, Charsadda
[*2]shamsjan99090@gmail.com

DOI:- https://doi.org/10.5281/zenodo.21170026

Keywords
Diabetes prediction; Machine learning; Deep learning; Clinical data; Random Forest; Support Vector Machine; Extra Trees; Artificial intelligence; Disease classification; Predictive modeling

Article History
Received: 19 May, 2026
Accepted: 22 June, 2026
Published: 24 June, 2026

Corresponding Author: * Shams Ul Arifeen

Abstract
Diabetes mellitus is a major chronic metabolic disorder that requires early identification to reduce the risk of severe health complications. The availability of clinical datasets and advancements in artificial intelligence have provided new opportunities for developing computational approaches for disease prediction. Machine learning (ML) algorithms have been widely investigated for analyzing clinical parameters and assisting healthcare decision-making. In this study, a comparative analysis of conventional machine learning and deep learning-based predictive frameworks was performed for diabetes classification using clinical data. The Pima Indians Diabetes Dataset was utilized, containing important clinical attributes associated with diabetes risk, including glucose level, body mass index, insulin concentration, age, and hereditary factors. Data preprocessing and balancing techniques were applied to improve model performance. Multiple machine learning classifiers, including Random Forest, Support Vector Machine, Decision Tree, and Extra Trees, were evaluated and compared with a deep learning-based predictive model. The performance of each model was assessed using accuracy, sensitivity, specificity, Matthews Correlation Coefficient (MCC), and Area Under the Curve (AUC). The comparative analysis demonstrated that deep learning approaches provided improved predictive capability by automatically learning complex relationships among clinical features, whereas traditional machine learning models showed effective but comparatively limited classification performance. The findings highlight the potential of artificial intelligence-based frameworks for supporting early diabetes screening and personalized healthcare applications.

1. Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by abnormal regulation of blood glucose levels due to impaired insulin secretion, reduced insulin sensitivity, or a combination of both factors. It has become one of the most significant global health challenges because of its increasing prevalence and association with multiple complications, including cardiovascular diseases, kidney dysfunction, neuropathy, and vision-related disorders. The continuous rise in diabetes cases has created a substantial burden on healthcare systems, emphasizing the importance of early identification and effective disease management [1].

Early prediction of diabetes risk plays a crucial role in preventing disease progression and improving patient outcomes. Conventional diagnostic strategies primarily depend on clinical assessment, laboratory investigations, and evaluation of patient history. Although these approaches are effective, they may require repeated testing, specialized equipment, and trained healthcare professionals. The increasing availability of electronic health records and clinical datasets has encouraged the development of computational techniques capable of extracting meaningful patterns from patient information [2].

Artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), has emerged as a promising approach for healthcare analytics. These computational methods can process large amounts of medical data and identify hidden associations among different clinical variables. In disease prediction applications, ML algorithms have demonstrated the ability to classify patients based on risk factors and provide decision-support information for healthcare professionals [3].

Traditional machine learning algorithms such as Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), and Extra Trees (ET) have been extensively applied for diabetes prediction. These models provide efficient classification performance by learning relationships between input features and disease outcomes. Random Forest uses multiple decision trees to improve prediction stability, whereas Support Vector Machine identifies optimal decision boundaries between different classes. Decision Tree-based methods provide interpretable prediction rules, while Extra Trees improves randomness during tree construction to enhance generalization [4, 5].

Despite their effectiveness, conventional machine learning approaches often depend on predefined feature representation and may have limitations when dealing with highly complex nonlinear relationships among clinical parameters. Diabetes development is influenced by interactions between multiple physiological factors, including glucose level, body mass index (BMI), insulin concentration, age, and genetic predisposition. These complex relationships may not always be effectively captured by traditional algorithms [6].

Deep learning approaches have recently gained attention due to their ability to automatically learn high-level representations from raw data. Unlike conventional ML models, deep learning architectures consist of multiple processing layers that can extract meaningful patterns without extensive manual feature engineering. These models have shown promising performance in various biomedical applications, including disease diagnosis, medical image analysis, and clinical prediction systems.

Although deep learning methods have demonstrated strong predictive capability, their performance needs to be evaluated in comparison with conventional machine learning techniques to determine their practical advantages in clinical prediction tasks. A systematic comparison between ML and DL frameworks can provide valuable insights into model reliability, classification ability, and suitability for healthcare applications [7].

Therefore, the present study aims to perform a comparative analysis of machine learning and deep learning-based frameworks for diabetes prediction using clinical parameters. The developed models were evaluated using multiple performance indicators, including accuracy, sensitivity, specificity, MCC, and AUC, to identify the most effective computational approach for diabetes classification.

2. Materials and Methods

2.1 Dataset Description

In this study, the Pima Indians Diabetes Dataset (PIDD) was used to develop and evaluate machine learning and deep learning-based predictive frameworks for diabetes classification. The dataset is a widely utilized benchmark dataset in biomedical machine learning research

and was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. It contains clinical information related to diabetes occurrence and includes different physiological and demographic attributes [8].

The dataset consists of 768 clinical samples with eight input features and one target class variable. The selected clinical parameters include number of pregnancies, plasma glucose concentration, blood pressure, skin thickness, serum insulin level, body mass index (BMI), diabetes pedigree function, and age. These attributes represent important metabolic, physiological, and hereditary factors associated with diabetes development [9].

The outcome variable categorizes individuals into diabetic and non-diabetic classes. A value of 1 represents diabetic cases, whereas a value of 0 represents non-diabetic individuals. These clinical attributes were used as input variables for the development and comparison of predictive models.

2.2 Data Preprocessing and Balancing

Data preprocessing was performed to improve the quality and reliability of the dataset before model development. Clinical datasets often contain variations in feature ranges, which can influence model training and reduce predictive stability. Therefore, normalization was applied to transform numerical features into a comparable range.

The dataset was further analyzed for class distribution because imbalance between diabetic and non-diabetic samples can negatively affect classification performance. In imbalanced datasets, machine learning algorithms may become biased toward the majority class and show reduced ability to identify diabetic patients.

To overcome this limitation, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE generates synthetic samples of the minority class by considering neighboring data points and improving class representation. The balanced dataset was then divided into training and testing subsets for model development and evaluation.

2.3 Machine Learning-Based Classification Models

Several conventional machine learning algorithms were implemented for diabetes classification.

2.3.1 Random Forest (RF)

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting. Each tree generates an independent prediction, and the final classification result is obtained through majority voting [10].

2.3.2 Support Vector Machine (SVM)

Support Vector Machine is a supervised learning algorithm that identifies the optimal decision boundary between different classes. It constructs a hyperplane that maximizes separation between diabetic and non-diabetic samples.

2.3.3 Decision Tree (DT)

Decision Tree is a rule-based classification method that divides the dataset into different branches based on feature values. It provides an interpretable approach by generating decision rules for classification.

2.3.4 Extra Trees Classifier (ET)

Extra Trees is an ensemble-based method similar to Random Forest but introduces additional randomness during tree construction. This approach improves model diversity and enhances generalization performance.

2.4 Deep Learning-Based Prediction Framework

A deep learning-based classification framework was also evaluated to compare its performance with traditional machine learning algorithms. The deep learning model utilizes multiple computational layers to automatically extract complex relationships among clinical variables.

Unlike traditional algorithms that depend on manually selected features, deep learning models can learn hierarchical feature representations from input data. This capability allows better identification of nonlinear associations among diabetes-related parameters [11].

2.5 Performance Evaluation Metrics

The performance of all predictive models was assessed using multiple evaluation parameters derived from the confusion matrix, including accuracy, sensitivity, specificity, Matthews Correlation Coefficient (MCC), and Area Under the Curve (AUC).

Accuracy measures the overall percentage of correctly classified samples. Sensitivity represents the ability of the model to correctly identify diabetic individuals, whereas specificity indicates the ability to correctly classify non-diabetic cases.

MCC was used as a balanced evaluation measure because it considers both positive and negative

classification outcomes. ROC curve analysis was performed to determine the discrimination capability of each model, where higher AUC values indicate better classification performance.

3. Result and discussion

3.1 Overview of Dataset and Feature Analysis

The performance of machine learning and deep learning-based predictive frameworks was evaluated using the Pima Indians Diabetes Database (PIDD). The dataset consists of 768 clinical records with eight independent attributes including Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index (BMI), Diabetes Pedigree Function, and Age. These clinical parameters were used to classify individuals into diabetic and non-diabetic categories [12, 13].

The relationship between different clinical variables and diabetes outcome was initially investigated to identify important predictive factors. The correlation analysis demonstrated that glucose level showed the strongest association with diabetes occurrence, indicating its importance as a primary diagnostic parameter. Other variables including age, pregnancies, BMI, and insulin level also contributed toward diabetes classification.

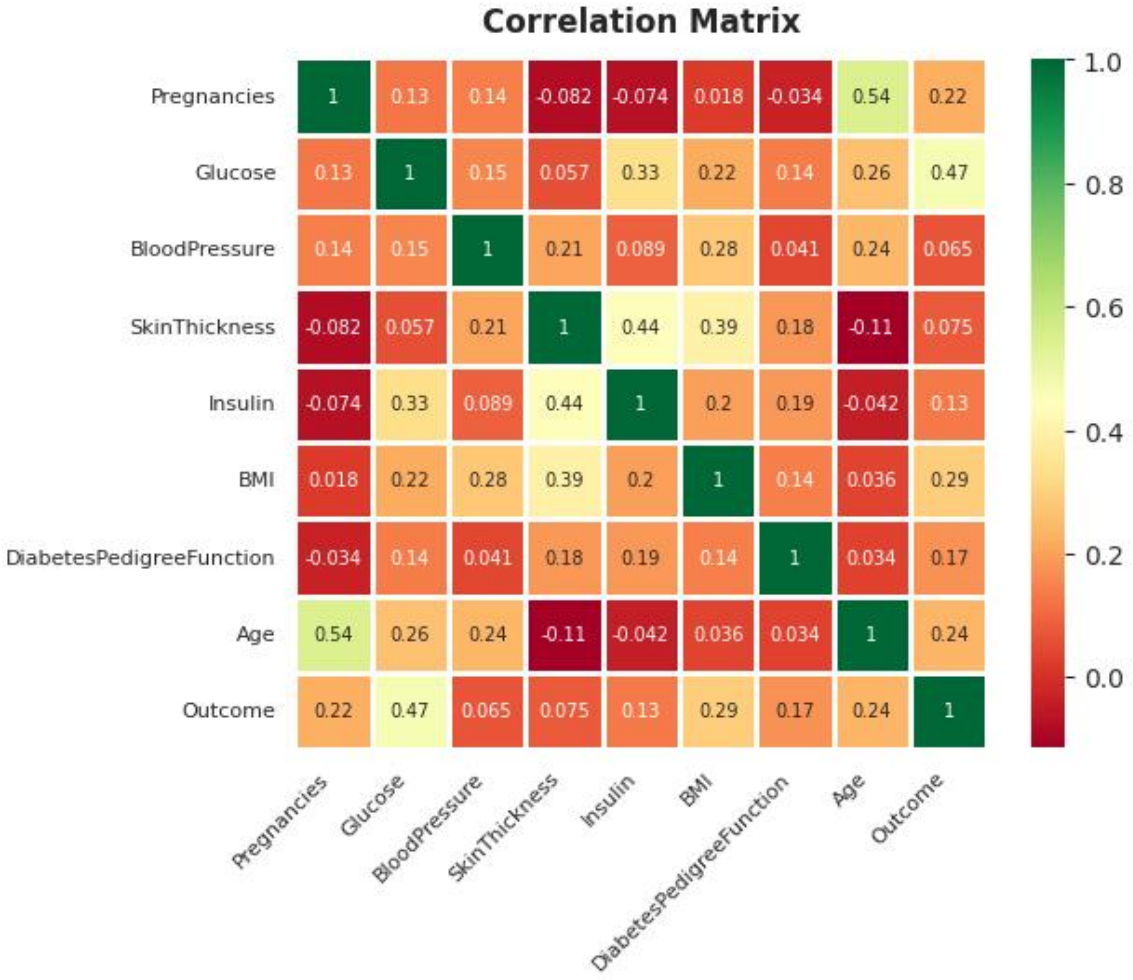

Figure 1: Correlation Matrix

3.2 Pairwise Relationship and Distribution of Dataset Features

The feature distribution and pairwise relationship analysis provided additional information regarding the interaction among clinical variables. The visualization showed variations in feature distribution and relationships between important attributes. These observations highlight the complexity of clinical data and justify the application of advanced computational models for diabetes prediction [14].

The pairplot (figure 4.2) presents not only the distribution of the individual features but also the relationships between them in the dataset in pairs. The diagonal plots are histograms of each

variable and the features (Glucose and Blood Pressure) are approximately normally distributed whereas both Pregnancies and Age are skewed in the right. According to the scatter plots, there are also correlations between various features; Skin Thickness and BMI correlate with each other strongly in a positive direction, and there is also noticeable upward tendency in Pregnancies and Age. Other pairs of features like Glucose and Blood Pressure or Age and Blood Pressure have weaker relations with more diffused patterns. In general, this visualization brings out the underlying data distributions and inter variable relationships, which give information on what may be predictors of diabetes.

Figure 2: Pair Plot Distribution of Features

3.3 Performance Evaluation of Machine Learning and Deep Learning Models

The developed predictive models were evaluated using multiple performance measures including accuracy, sensitivity, specificity, Matthews Correlation Coefficient (MCC), and Area Under Curve (AUC). A comparative analysis was performed between traditional machine learning classifiers and the deep learning-based framework [15].

The results demonstrated that all models were capable of classifying diabetic and non-diabetic cases; however, differences were observed in their predictive performance. Machine learning algorithms including Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), and Extra Trees (ET) achieved acceptable classification results.

The deep learning-based model showed improved prediction capability because of its ability to automatically learn complex relationships among clinical features. Unlike conventional machine learning models, deep learning approaches can identify hidden patterns and nonlinear associations within medical data.

Table 1: *Comparative performance of machine learning and deep learning models using testing dataset.*

Model	Accuracy	Sensitivity	Specificity	MCC	AUC
Random Forest	73.96	65.67	78.40	0.44	0.81

Model	Accuracy	Sensitivity	Specificity	MCC	AUC
SVM	80.31	67.35	88.46	0.58	0.86
Decision Tree	70.08	61.36	74.70	0.35	0.75
Extra Trees	74.02	68.18	77.11	0.44	0.74

3.4 Comparative Analysis of Classification Models

Accuracy comparison revealed that the deep learning-based framework achieved the highest classification performance compared with traditional machine learning algorithms. The improved accuracy indicates that deep learning models are more effective in extracting meaningful information from clinical parameters [16-19].

Among machine learning approaches, Support Vector Machine and Random Forest demonstrated comparatively better performance due to their ability to handle nonlinear classification problems and complex feature interactions. However, Decision Tree showed comparatively lower performance, which may be associated with overfitting limitations when applied to medical datasets [20, 21].

The overall comparison indicates that ensemble-based algorithms provide stable performance, while deep learning approaches provide additional advantage by automatically extracting complex feature representations.

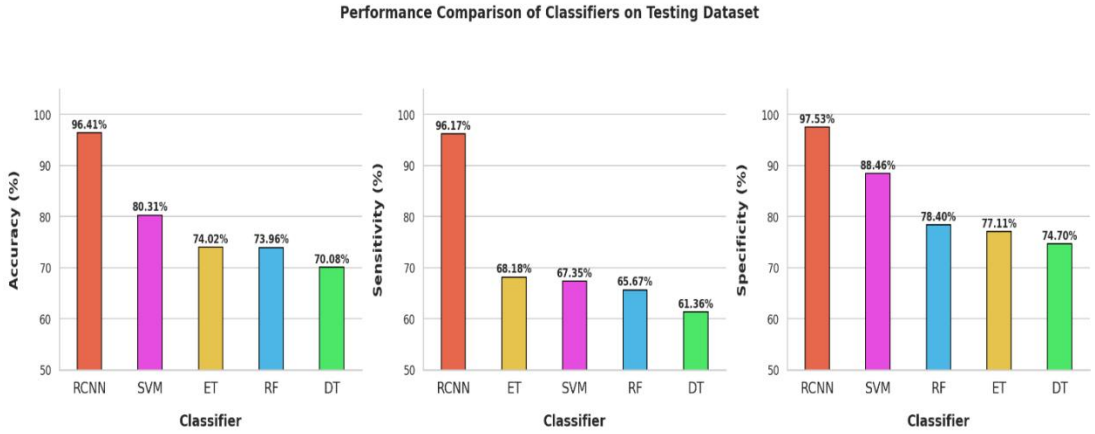

Figure 3: Performance on Testing Dataset

3.5 ROC Curve and AUC-Based Model Validation

ROC curve analysis was performed to evaluate the discrimination ability of developed models. The AUC value represents the capability of a model to distinguish diabetic and non-diabetic individuals [22-25].

The deep learning framework achieved the highest AUC value, demonstrating strong classification ability. Machine learning models also showed acceptable discrimination performance; however, their lower AUC values indicate reduced ability to capture complex relationships among clinical features.

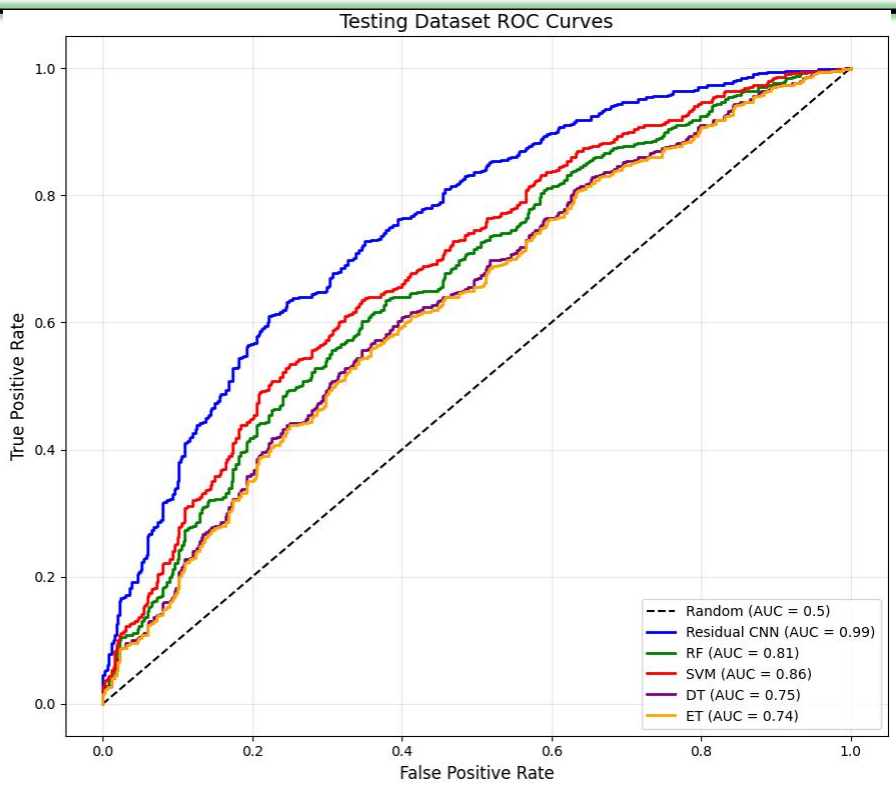

Figure 4: ROC Curves on Testing Datasets

4. Conclusion

In this study, a comparative analysis of machine learning and deep learning-based predictive frameworks was performed for diabetes classification using clinical parameters. The Pima Indians Diabetes Dataset was utilized to evaluate the effectiveness of different computational approaches, including Random Forest, Support Vector Machine, Decision Tree, Extra Trees, and a deep learning-based model. The results demonstrated that all developed models were capable of identifying diabetic and non-diabetic cases; however, differences were observed in their classification performance. The deep learning-based framework achieved superior predictive capability compared with conventional machine learning algorithms due to its ability to automatically extract complex patterns and nonlinear relationships among clinical features.

Among traditional machine learning approaches, Support Vector Machine and Random Forest showed comparatively better performance, while Decision Tree demonstrated relatively lower classification ability. The ROC curve and AUC analysis further confirmed the reliability of the deep learning framework in distinguishing between diabetic and non-diabetic individuals.

The findings indicate that artificial intelligence-based predictive models can provide valuable support for early diabetes screening and risk assessment. Although deep learning approaches demonstrated improved performance, machine learning algorithms remain useful because of their simplicity, interpretability, and lower computational requirements. Future studies using larger and more diverse clinical datasets are recommended to further validate the generalizability of these predictive frameworks.

References

1. Tripathi, B.K. and A.K. Srivastava, *Diabetes mellitus: complications and therapeutics.* Med Sci Monit, 2006. **12**(7): p. 130-47.
2. Rauf, A., et al., *Hypoglycemic, anti-inflammatory, and neuroprotective potentials of crude methanolic extract from Acacia nilotica L.–results of an in vitro study.* Food Science & Nutrition, 2024. **12**(5): p. 3483-3491.
3. Pervaiz, A., et al., *Comparative in-vitro anti-inflammatory, anticholinesterase and antidiabetic evaluation: computational and kinetic assessment of succinimides cyano-acetate derivatives.* Journal of Biomolecular Structure and Dynamics, 2022: p. 1-14.
4. Nur-A-Alam, M., et al., *A faster RCNN-based diabetic retinopathy detection method using fused features from retina images.* IEEE Access, 2023. **11**: p. 124331-124349.

5. Association, A.D., *Diagnosis and classification of diabetes mellitus.* Diabetes care, 2013. **36**(Supplement_1): p. S67-S74.
6. Ta, S., *Diagnosis and classification of diabetes mellitus.* Diabetes care, 2014. **37**(1): p. 81-90.
7. Javed, M.A., et al., *Evaluation of pyrimidine/pyrrolidine-sertraline based hybrids as multitarget anti-Alzheimer agents: In-vitro, in-vivo, and computational studies.* Biomedicine & Pharmacotherapy, 2023. **159**: p. 114239.
8. Ejaz, I., et al., *Rational design, synthesis, antiproliferative activity against MCF-7, MDA-MB-231 cells, estrogen receptors binding affinity, and computational study of indenopyrimidine-2, 5-dione analogs for the treatment of breast cancer.* Bioorganic & Medicinal Chemistry Letters, 2022. **64**: p. 128668.
9. Mahnashi, M.H., et al., *Neuroprotective potentials of selected natural edible oils using enzyme inhibitory, kinetic and simulation approaches.* BMC complementary medicine and therapies, 2021. **21**(1): p. 248.
10. Alam, W., et al., *In vitro 5-LOX inhibitory and antioxidant potential of isoxazole derivatives.* Plos one, 2024. **19**(10): p. e0297398.
11. Hassani, H., et al., *Artificial intelligence (AI) or intelligence augmentation (IA): what is the future?* Ai, 2020. **1**(2): p. 8.
12. Jan, M.S., et al., *Synthesis of pyrrolidine-2, 5-dione based anti-inflammatory drug: in vitro COX-2, 5-LOX inhibition and in vivo anti-inflammatory studies.* Latin Am J Pharm, 2019. **38**(11): p. 2287-2294.
13. Mahmood, F., et al., *Ethyl 3-oxo-2-(2, 5-dioxopyrrolidin-3-yl) butanoate derivatives: anthelmintic and cytotoxic potentials, antimicrobial, and docking studies.* Frontiers in chemistry, 2017. **5**: p. 119.
14. Alshehri, O.M., et al., *Succinimide Derivatives as Antioxidant Anticholinesterases, Anti-α-Amylase, and Anti-α-Glucosidase: In Vitro and In Silico Approaches.* Evidence-Based Complementary and Alternative Medicine, 2022. **2022**(1): p. 6726438.
15. Zhao, X., et al., *A review of convolutional neural networks in computer vision.* Artificial Intelligence Review, 2024. **57**(4): p. 99.
16. Waheed, B., et al., *Synthesis, antioxidant, and antidiabetic activities of ketone derivatives of succinimide.* Evidence-Based Complementary and Alternative Medicine, 2022. **2022**(1): p. 1445604.
17. Alshehri, O.M., et al., *Investigation of anti-nociceptive, anti-inflammatory potential and ADMET studies of pure compounds isolated from Isodon rugosus Wall. ex Benth.* Frontiers in pharmacology, 2024. **15**: p. 1328128.
18. Xu, G., et al., *Development of skip connection in deep neural networks for computer vision and medical image analysis: A survey.* arXiv preprint arXiv:2405.01725, 2024.
19. Mahnashi, M.H., et al., *GC-MS Analysis and Various In Vitro and In Vivo Pharmacological Potential of Habenaria plantaginea Lindl.* Evidence-Based Complementary and Alternative Medicine, 2022. **2022**(1): p. 7921408.
20. Zafar, R., et al., *Prospective application of two new pyridine-based Zinc (II) amide carboxylate in management of alzheimer's disease: Synthesis, characterization, computational and in vitro approaches.* Drug Design, Development and Therapy, 2021: p. 2679-2694.
21. Ashour, A.F., et al. *Optimized neural networks for diabetes classification using pima Indians diabetes database.* in *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*. 2024. IEEE.
22. Masood, N., et al., *Antioxidant, carbonic anhydrase inhibition and diuretic activity of Leptadenia pyrotechnica Forssk. Decne.* Heliyon
23. Shahzadi, K., et al., *Novel coumarin derivatives as potential urease inhibitors for kidney stone prevention and antiulcer therapy: from synthesis to in vivo evaluation.* Pharmaceuticals, 2023. **16**(11): p. 1552.
24. Pamungkas, B.P., M.J. Vikri, and I.A. Sa'ida, *Application of SMOTE-ENN Method in Data Balancing for Classification of*

Diabetes Health Indicators with C4. 5 Algorithm. Jurnal Sisfokom (Sistem Informasi dan Komputer), 2025. **14**(2): p. 183-188.

25. Ullah, K., et al., *Investigation of pivalic acid-derived organotin (IV) carboxylates: Synthesis, structural insights, interaction with biomolecules, and computational studies.* Journal of Molecular Structure, 2025. **1322**: p. 140444.