Current Trends in Technology and Science
www.ctts.in, ISSN: 2279-0535, Volume: 11, Issue: 4, July 2023

A Comparative Analysis of Machine Learning and Artificial Intelligence-based Models for Diabetes Prediction

Jitendra Sheetlani[1], Ajay Vyas[2], Harsh Gupta[3],
[1,2]Sri Satya Sai University of Technology and Medical Science, Sehore (MP.)
[3]IT Project Manager and Independent Researcher

Abstract

Diabetes has become a significant global public health concern as the prevalence of non-communicable diseases continues to rise. Heart disease claims the lives of approximately 18 million people annually, with diabetes and high blood pressure emerging as primary contributing factors. The social, physical, and economic consequences of diabetes are substantial. Elevated blood sugar levels characterize this chronic condition and occur due to the body's inability to produce or properly respond to insulin. Data mining enables analysts to efficiently analyze extensive datasets to identify patterns and trends associated with diabetes. In recent years, machine learning (ML) methods have been utilized for diabetes prediction. Data mining involves extracting essential information and leveraging it to enhance dynamic effectiveness. Various AI techniques, including Support Vector Machine (SVM), Random Forest, Decision Tree (DT), K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), and Naive Bayes (NB) classifiers, have been employed in diabetes prediction. This research focuses on analyzing different machine-learning models for diabetes prediction. The paper is structured into sections: the first section discusses the various machine learning models and their distinct activation functions employed in this study, while the second section presents a comparative analysis of these models.

Keywords: Diabetes prediction, Machine learning, Random Forest, Support Vector Machines (SVM), Logistic regression (LR), Gradient boosting (GB), k-nearest neighbor (k-NN)

1 Introduction

Diabetes stands as a pervasive health concern affecting populations worldwide, with a particularly heightened prevalence observed within industrialized nations. Its gravity and widespread impact necessitate substantial attention from medical professionals, patients, families, and society at large. The multifaceted ramifications of diabetes span across social, physical, and economic domains, rendering it a formidable chronic condition demanding comprehensive management strategies.

At its core, diabetes manifests as a chronic ailment characterized by persistently elevated levels of blood sugar or glucose. The genesis of diabetes stems from the body's inability to either produce insulin adequately or respond effectively to insulin's actions on cells—a phenomenon often referred to as "x syndrome" within the medical lexicon. Despite decades of research and medical advancements, the precise etiology of diabetes remains elusive, underscoring the complexity inherent in its pathology.

Traditionally, the paradigm of diabetes management has revolved around symptom alleviation rather than addressing the underlying pathological processes. However, contemporary healthcare imperatives have shifted towards a more holistic approach, emphasizing proactive disease management and prevention strategies.

Machine Learning (ML) techniques have emerged as invaluable tools in the domain of diabetes research and management. By harnessing the power of ML algorithms, researchers can glean insights from vast repositories of clinical data, discerning intricate patterns and trends that may elude traditional analytical approaches. The advent of ML has revolutionized data processing capabilities, facilitating expedited analysis and interpretation of complex datasets.

Of paramount importance within the realm of diabetes care is the early detection of the condition, a cornerstone in preemptive healthcare interventions. Recognizing the imperative of timely diagnosis, numerous studies have endeavored to leverage ML-based models and methodologies to forecast the onset of diabetes accurately. Notably, the PIMA Indian Diabetes Dataset (PIDD) has emerged as a cornerstone dataset in this regard, serving as a linchpin for the development and validation of predictive models.

Against this backdrop, our research endeavors to undertake a comprehensive analysis of diverse ML models tailored to the precise prediction of diabetes. By scrutinizing the efficacy and performance of various ML algorithms, our study seeks to furnish actionable insights that may inform clinical decision-making and healthcare policy formulation. Through rigorous experimentation and meticulous evaluation, we aim to elucidate the strengths and limitations of each model, thereby paving the way for enhanced diagnostic accuracy and patient care outcomes in the realm of

diabetes management.

In essence, our research endeavors to harness the transformative potential of ML in revolutionizing diabetes care paradigms. By harnessing the analytical prowess of ML algorithms, we aspire to fortify the armamentarium of healthcare practitioners with robust, data-driven tools capable of facilitating early detection, proactive intervention, and personalized management strategies for individuals afflicted by diabetes.

2 LITERATURE REVIEW

Researchers have widely used the Pima Indian Diabetes Dataset (PIDD) to develop machine learning-based models for predicting diabetes, which describe the characteristics of diabetes patients. This section reviews the literature that explores various machine-learning approaches for building an autonomous diabetes prediction system. Michie et al. provide an updated review of different classification approaches, comparing their performance on challenging datasets and discussing their applicability to real-world industrial problems. Constructing a classification procedure from known data with true classes is referred to as pattern recognition, discrimination, or supervised learning [1]. Jeatrakul et al. conducted a research paper comparing the performance of different neural network techniques for binary classification problems. They compared five types of neural networks, including Back Propagation Neural Networks (BPNN), Radial Basis Function Neural Networks (RBFNN) [2], General Regression Neural Networks (GRNN) [3], Probabilistic Neural Networks (PNN) [4], and Complementary Neural Networks (CMTNN). The comparison was based on three benchmark datasets from the UCI machine learning repository [5].

Estebanez et al. utilised genetic programming-based data projections in their research. They presented a method based on genetic programming (GP) to automatically evolve projections, making data classification easier in the projected space. The approach involved evolving independent sub-trees, allowing for the construction of relevant attributes and potential dimensionality reduction or increase depending on the feasibility of classification in higher dimensional spaces [6].

Anjli Negi (2021) emphasises that diabetes is a metabolic disease characterised by impaired glucose utilisation, leading to consistently high blood glucose levels. Complications of diabetes include diabetic ketoacidosis, nonketotic hyperosmolar coma, heart disease, stroke, chronic renal failure, retinal damage, and foot ulcers. The global prevalence of diabetes is rapidly increasing, posing a significant public health concern. Early detection of diabetes plays a crucial role in reducing the risk of major complications and facilitating effective treatment [7].

A. M. Abdulazeez (2021) highlights the wide application of machine learning (ML) in computational work for algorithm development and performance improvement. Learning from unbalanced datasets has been a significant challenge in machine learning, appearing in various applications such as computer security, Swarm Intelligence, remote sensing, and biomedicine [8].

Jyotismita Chaki et al. [9] highlight the transformative potential of machine learning and AI in enabling early detection and diagnosis of DM, thereby averting the dire consequences associated with advanced-stage diabetes. The review comprehensively analyzes the various facets of DM detection, diagnosis, and self-management, spanning six key dimensions: datasets utilized in DM research, preprocessing methodologies, feature extraction techniques, machine learning-based identification and classification of DM, AI-driven intelligent DM assistants, and performance evaluation metrics.

Drawing upon a meticulously curated selection of 107 primary publications sourced from reputable repositories such as Scopus and PubMed, this literature survey offers a comprehensive overview of the state-of-the-art techniques employed in DM detection and self-management. By synthesizing insights from diverse research endeavors, the review seeks to furnish valuable insights to the scientific community engaged in the domain of automatic DM detection and self-management.

The review not only presents a detailed exposition of existing methodologies but also extrapolates upon the conclusions drawn from prior studies, shedding light on the significance and implications of their findings. Furthermore, the literature survey identifies three pressing research issues warranting attention in the realm of DM detection, diagnosis, and self-management, thus delineating future research directions for the scientific community.

These seminal works collectively underscore the transformative potential of machine learning and artificial intelligence in revolutionizing healthcare analytics, particularly in the domain of diabetes prediction and management. By leveraging innovative methodologies and cutting-edge techniques, researchers are poised to unlock new insights into the pathophysiology of diabetes, thereby facilitating more accurate diagnoses and personalized treatment strategies. As the field of predictive healthcare analytics continues to evolve, these foundational studies serve as guiding beacons, illuminating the path towards a future where data-driven insights empower clinicians and patients alike in the fight against diabetes and other chronic diseases.

3 COMPARATIVE ANALYSIS OF DIFFERENT MODELS

Researchers have extensively investigated the Pima Indian Diabetes dataset using various machine learning algorithms to predict diabetes. Table 1 displays the perfor-

Table 1: Pima Indian Diabetes Dataset Michie, Spiegelhalter, and Taylor Classification Results

S. No.	Algorithm	CC (%)	ER (%)
1	Discrim	77.5	22.5
2	Quadisc	73.8	26.2
3	Logdisc	77.7	22.3
4	SMART	76.8	23.2
5	ALLOC80	69.9	30.1
6	K-NN	67.6	32.4
7	CASTLE	74.2	25.8
8	CART	74.5	25.5
9	IndCART	72.9	27.1
10	NewID	71.1	28.9
11	AC2	72.4	27.6
12	Baytree	72.9	27.1
13	NaiveBay	73.8	26.2
14	CN2	71.1	28.9
15	C4.5	73	27
16	Itrule	75.5	24.5
17	Cal5	75	25
18	Kohonen	72.7	27.3
19	DIPOL92	77.6	22.4
20	Backprob	75.2	24.8
21	RBF	75.7	24.3
22	LVQ	72.8	27.2
23	Average	73.8	26.2
CC= Correct Classification, ER Error rate			

mance of different algorithms in terms of correct classification and miss classification (error rate) on the dataset. According to the table, some algorithms achieved the highest correct classification percentages. The "Logdisc [10]" algorithm achieved a correct classification of 77.7% with a miss classification (error rate) of 22.3%. Similarly, the "DIPOL92 [11]" algorithm achieved a correct classification of 77.6% with a miss classification of 22.4%, while the "Discrim [12]" algorithm achieved a correct classification of 77.5% with a miss classification of 22.5%.

On the other hand, certain algorithms exhibited higher miss classification percentages or error rates. For instance, the "K-NN [13]" algorithm achieved a correct classification of 67.6% with a miss classification of 32.4%. Similarly, the "ALLOC80 [14]" algorithm achieved a correct classification of 69.9% with a miss classification of 30.1%. The average correct classification percentage across all algorithms was 73.8%, with an average miss classification of 26.2%. When selecting the most suitable algorithm for the task, it's crucial to consider factors such as specific prediction requirements, dataset characteristics, and the trade-off between correct and miss classification rates.

Extensive research has been conducted to improve the order precision of predictions on the Pima Indian Diabetes dataset using artificial neural networks. In this regard, Jeatrakul and Wong examined the performance of various neural network architectures, including Back Propagation Neural Network (BPNN), General Regression Neural Network (GRNN), Radial Basis Function Neural Network (RBFNN), Probabilistic Neural Network (PNN), and Complementary Neural Network (CNN), also known as Complementary Multi-task Neural Network (CMTNN).

Table 2 presents the performance results of these architectures. The table provides the precision or accuracy achieved by each neural network architecture across multiple tests on the Pima Indian Diabetes dataset. Here is a summary of the results: Test 1: BPNN achieved a precision of 77.27%, GRNN - 74.68%, RBFNN - 79.22%, PNN - 74.68%, and CMTNN - 77.92%. Test 2: BPNN achieved a precision of 76.62%, GRNN - 79.87%, RBFNN - 79.22%, PNN - 79.87%, and CMTNN - 76.62%. Test 3: BPNN achieved a precision of 70.13%, GRNN - 70.13%, RBFNN - 74.03%, PNN - 70.13%, and CMTNN - 72.08%. Test 4: BPNN achieved a precision of 85.71%, GRNN - 81.82%, RBFNN - 79.22%, PNN - 81.82%, and CMTNN - 83.77%. Test 5: BPNN achieved a precision of 75.97%, GRNN - 75.97%, RBFNN - 77.27%, PNN - 75.97%, and CMTNN - 75.32%.

These findings offer valuable insights into the performance of different neural network architectures on the Pima Indian Diabetes dataset. The average precision ranges from 75.26% to 76.56% across all architectures. Estebanez, Alter, and Valls employed genetic programming-based data projections to analyse clustering tasks. They utilised the Pima Indian diabetes dataset and reduced the data dimensionality from 8 to 3. They employed Support Vector Machine (SVM), Simple Logistics, and Multilayer Perceptron algorithms for the classification task, utilising the Pima Indian Diabetes data. The results of their analysis are presented in Table 3. The Multilayer Perceptron of the Fake Neural Network achieved a characterisation performance of 76.69 percent. Single Logistics, on the other hand, obtained the highest rating of 77.86 percent. In a study by Lena Kallin Westin, she investigated various preprocessing approaches to handle missing data in the Pima Indian Diabetes dataset. She developed a preprocessing perceptron for decision support using the diabetes dataset. The trained decision support network achieved an overall classification performance of 79 percent.

In another research by lander, different classification approaches, including Naive Bayes, decision trees, and two types of ensemble methods, were employed on the Pima Indian Diabetes dataset. Table 4 presents the various classification approaches by lander and their corresponding classification performance. Misra and Dehuri made a Functional Link Artificial Neural Network for Classification Task in Data Mining. They stood apart its depiction execution from other AI calculations in their review Functional Link Artificial Neural Network for Classification Task in Data

Table 2: Results of Jeatrakul and Wong's classification of the Pima Indian Diabetes Dataset

TN	BPNN	GRNN	RBFNN	PNN	CMTNN
1	77.27	74.68	79.22	74.68	77.92
2	76.62	79.87	79.22	79.87	76.62
3	70.13	70.13	74.03	70.13	72.08
4	85.71	81.82	79.22	81.82	83.77
5	75.97	75.97	77.27	75.97	75.32
6	70.78	70.13	72.08	70.13	72.08
7	75.32	72.73	76.62	72.73	75.97
8	79.22	78.57	77.27	78.57	79.22
9	74.68	74.68	76.62	74.68	75.32
10	75.97	74.03	74.03	74.03	76.62
AVG	76.17	75.26	76.56	75.26	76.49
TN=Test Number and AVG=Average					

Table 3: Pima Indian Diabetes Dataset Classification by Estebanez, Alter, and Valls

S. No.	Algorithms	Accuracy
1	SVM	77.21
2	Simple Logistics	77.86
3	Multilayer Perceptron	76.69

Table 4: Pima Indian Diabetes Dataset By lander Classification Performance

S. No.	Methods	Accuracy
1	Belief Network (Laplace)	72.50%
2	Belief Network	72.30%
3	Decision Tree	72.00%
4	Naïve Bayes	71.50%

Table 5: Performance of Misra and Dehuri Classification on Pima Indian Diabetes Dataset

S. No.	Models	Accuracy
1	NN	65.1
2	KNN	69.7
3	FSS	73.6
4	BSS	67.7
5	MFS1	68.5
6	MFS2	72.5
7	CART	74.5
8	C4.5	74.7
9	FID3.1	75.9
10	MLP	75.2
11	FLANN	78.13

Mining. Their FLANN gathering execution was 78.13 percent, while their MLP demand execution was 75.2 percent. On the Pima Indian Diabetes dataset, Table 5 shows the strategy execution of different AI calculations. KNN from case-based thinking can be utilised to recover comparative authentic models and eliminate exceptions, bringing about enhanced brain network arrangement execution.

4 Conclusion

We conducted a comparative analysis of machine learning and AI-based models for predicting diabetes with high accuracy. We employed several machine learning algorithms, such as Decision Tree (DT), K-Nearest Neighbor (KNN), and Logistic Regression (LR), on the PIMA Indian Diabetes Dataset (PIDD) to predict diabetes and evaluated their performance based on different parameters. The results obtained from these models were promising, showing good performance across various metrics. The early detection of diabetes is important in addressing the health challenges associated with the disease. By leveraging machine learning and AI techniques, we can effectively predict diabetes and enable timely intervention and treatment, thus mitigating the potential complications associated with the condition. Our study highlights the potential of machine learning and AI-based approaches in accurately predicting diabetes. Further research and advancements in this field can contribute to developing more robust and efficient models for the early detection and management of diabetes, ultimately improving the overall healthcare outcomes for individuals affected by the disease.

References

[1] D. Michie, D. J. Spiegelhalter, C. C. Taylor, and J. Campbell, *Machine learning, neural and statistical classification*. Ellis Horwood, 1995.

[2] S. I. Pasha and H. P. Singh, "A novel model proposal using association rule based data mining techniques for indian stock market analysis," *Annals of the Romanian Society for Cell Biology*, pp. 9394–9399, 2021.

[3] A. R. Md, H. P. Singh, and K. N. Reddy, "Data mining approaches to identify spontaneous homeopathic syndrome treatment," *Annals of the Romanian Society for Cell Biology*, pp. 3275–3286, 2021.

[4] V. Naiyer, J. Sheetlani, and H. P. Singh, "Software quality prediction using machine learning application," in *Smart Intelligent Computing and Applications: Proceedings of the Third International Conference on Smart Computing and Informatics, Volume 2*. Springer, 2020, pp. 319–327.

[5] P. Jeatrakul and K. W. Wong, "Comparing the performance of different neural networks for binary classification problems," in *2009 Eighth International Symposium*

on Natural Language Processing. IEEE, 2009, pp. 111–115.

[6] C. Estébanez Tascón, R. Aler, and J. M. Valls, "Genetic programming based data projections for classification tasks," 2005.

[7] B. Misra and S. Dehuri, "Functional link artificial neural network for classification task in data mining," 2007.

[8] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and structural biotechnology journal*, vol. 15, pp. 104–116, 2017.

[9] J. Chaki, S. T. Ganesh, S. Cidham, and S. A. Theertan, "Machine learning and artificial intelligence based diabetes mellitus detection and self-management: A systematic review," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 3204–3225, 2022.

[10] V. Jaiswal, A. Negi, and T. Pal, "A review on current advances in machine learning based diabetes prediction," *Primary Care Diabetes*, vol. 15, no. 3, pp. 435–443, 2021.

[11] N. M. Abdulkareem and A. M. Abdulazeez, "Science and business," *International Journal*, vol. 5, no. 2, pp. 128–142, 2021.

[12] T. M. Alam, M. A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T. I. Baig, A. Hussain, M. A. Malik, M. M. Raza, S. Ibrar *et al.*, "A model for early prediction of diabetes," *Informatics in Medicine Unlocked*, vol. 16, p. 100204, 2019.

[13] T. Sharma and M. Shah, "A comprehensive review of machine learning techniques on diabetes detection," *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, no. 1, p. 30, 2021.

[14] B. Priyanka and H. P. Singh, "A review on big data analysis of tobacco consuming trends in india," *Annals of the Romanian Society for Cell Biology*, pp. 3516–3527, 2021.