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

Diabetes Care: A Machine Learning Based Review Under Supervision and without Supervision

Dr. G B Hima Bindu[1], Dr.L.Thomas Robinson[2], Bingi Manorama Devi[3], Dr Kuppala Saritha[4], Dr. D.Ganesh[5*], Dr. P. Neelima[6]

Submitted: 05/02/2024 **Revised**: 13/03/2024 **Accepted**: 19/03/2024

Abstract: Diabetes is a persistent metabolic condition that affects millions of people globally. The effective management of diabetes care is crucial in order to prevent complications and improve patient outcomes. Recent years have seen a substantial increase in the use of machine learning techniques in the field of healthcare, especially the treatment of diabetes. This review seeks to offer a thorough examination of machine learning techniques used in diabetes treatment, both supervised and unsupervised. Algorithms for supervised machine learning have been widely used for a variety of diabetes care activities, including risk assessment, diagnosis, and medication recommendation. These algorithms utilize labelled data to train predictive models, allowing for accurate identification of high-risk individuals, early detection of diabetes, and personalized treatment plans. In particular, support vector machines, random forests, and synthetic neural networks have produced promising outcomes in these fields of contrast, unsupervised machine learning techniques have been used for pattern identification and exploratory analysis of big datasets without specified labels. The identification of patient subgroups based on shared traits using clustering techniques like k-means and hierarchical clustering has enabled personalised therapies and precision medicine approaches in the treatment of diabetes. Principal component analysis and t-distributed stochastic neighbour embedding are two examples of dimensionality reduction techniques that have been useful in visualising complex data and revealing hidden relationships. This review also discusses the challenges and limitations associated with the application of machine learning in diabetes care. Issues such as data quality, interpretability, and generalizability of models are addressed, highlighting the importance of addressing these concerns for successful implementation in clinical practice.
In conclusion, the integration of supervised and unsupervised machine learning techniques holds great potential in improving diabetes care. These methods provide valuable insights into risk assessment, diagnosis, treatment, and patient stratification. Nonetheless, further research and collaboration between data scientists, clinicians, and researchers are necessary to address the challenges and enhance the translation of machine learning algorithms into real-world clinical settings.

Keywords: *AdaBoost, XG Boost, Decision tree, Support vector classifier.*

1. Introduction

Diabetes is a complex and prevalent chronic disease that poses significant challenges to healthcare systems worldwide. There is significant interest in utilizing machine learning techniques to improve diabetes care due to the increased accessibility of electronic health records and massive datasets. In the area of diabetes care, in particular, supervised and unsupervised machine learning algorithms have become effective tools for analyzing and extracting significant insights from a variety of data sources. Various facets of diabetes treatment have widely used supervised machine learning algorithms, which learn from labelled data to produce predictions or classifications. These algorithms have the potential to assist in risk prediction, early diagnosis, and treatment recommendation by utilizing features derived from patient demographics, clinical measurements, and biomarkers. Supervised models can find patterns and associations that may not be immediately obvious to healthcare experts by training on past data, allowing for more precise and individualised interventions[1][2].

In contrast, unsupervised machine learning techniques provide valuable insights into large and unlabeled datasets. These methods are particularly useful in exploratory analysis and pattern recognition, allowing for the identification of underlying structures and subgroups within the diabetes population. Unsupervised learning can help uncover hidden patterns in patient data, enabling the discovery of novel phenotypes, risk factors, and potential therapeutic targets. By revealing the heterogeneity within the diabetic population, unsupervised methods support the development of targeted interventions and precision medicine approaches[3][5]. In this review, we set out to

[1] *Associate Professor Department of CSE School of Technology The Apollo University Chittoor, AP, India Email: himabindugbe@gmail.com*
[2] *Assistant professor Department of computer science (PG) Kristu jayanti college Autonomous, Bangalore, KA, India Email: son.mca@gmail.com*
[3] *Assistant professor, Department of CSE K.S.R.M College of Engineering (Autonomous), Kadapa,AP, India Email: bingimanorama@gmail.com*
[4] *Associate Professor Department of CSE School of CSE & IS Presidency University, Bangalore, KA, India Email: saritha.mphil@gmail.com*
[5*] *Associate Professor of CSE, Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College(Autonomous), Tirupati, Andhra Pradesh , India Email: dgani05@gmail.com*
[6] *Assistant Professor, Department of CSE, School of Engineering & Technology, SPMVV, Tirupati, AP, India. Email: neelima.pannem@gmail.com*

present a thorough examination of the use of supervised and unsupervised machine learning approaches in the treatment of diabetes. We will explore the latest advancements, methodologies, and challenges associated with the use of these algorithms. We want to emphasize the possible advantages and limitations of machine learning in enhancing diabetes care, as well as the future directions and prospects for research and clinical implementation, by critically analyzing the present literature. The prevalence and complexity of diabetes present significant challenges to healthcare systems worldwide. With the growing availability of electronic health records and vast datasets, researchers and healthcare professionals are increasingly turning to machine learning techniques to enhance diabetes care[4][6].

By conducting exploratory analysis and pattern recognition, these methods unveil underlying structures and subgroups within the diabetic population. Unsupervised learning aids in the discovery of novel phenotypes, risk factors, and potential therapeutic targets by revealing hidden patterns within patient data. The identification of the heterogeneity within the diabetic population enables the development of tailored interventions and precision medicine approaches[8][10].

2. Literature Survey

2.1. Comparative Analysis of Ml Algorithms

Throughout history, humanity has been captivated by the pursuit of knowledge about the intricate details of the universe. From celestial bodies like stars, planets, asteroids to the discovery of exoplanets, each advancement in astronomy has expanded our understanding of the cosmos. NASA's Kepler Mission represents a significant milestone in this ongoing quest. Through telescopic surveys of the Milky Way galaxy, it aims to identify thousands of earth-sized and smaller planets located within or near the habitable zone of their respective stars. This data helps estimate the number of stars in our galaxy that might host such orbiting planets.

Exoplanets, or planets outside of our solar system, provide important insights on how planets are formed. However, in the past, extracting information about exoplanets from mission data was a labor-intensive task, reliant on traditional algorithms. This procedure has been shortened and made more effective with the introduction of several machine learning techniques. However, not all algorithms produce equally promising results when applied to different types of data. Conducting a comparative study of these algorithms is essential to identify their strengths and weaknesses in analyzing specific data forms[11]..

2.2. Unsupervised Clustering for Diabetes Care

Diabetes is a complex and prevalent chronic disease that poses significant challenges to healthcare systems worldwide. Unsupervised clustering techniques have emerged as powerful tools for extracting meaningful patterns and subgroups within large and diverse diabetes datasets. This research paper presents a comprehensive review of unsupervised clustering techniques applied to diabetes care, aiming to provide insights into their applications, methodologies, and challenges.

Unsupervised clustering algorithms enable the exploration of hidden structures and patterns in unlabeled diabetes data, without the need for predefined class labels. These algorithms group similar patients or data points together based on shared characteristics, allowing for the identification of distinct subgroups within the diabetic population. Such subgroups can reveal important insights into disease progression, treatment response, and potential risk factors [12][13].

2.3. Deep Learning for Diabetic Retinopathy Detection

One of the main factors contributing to vision loss and blindness in people with diabetes is diabetic retinopathy (DR). For effective intervention and management of DR, early recognition and precise diagnosis are essential. The science of ophthalmology is undergoing a revolution as a result of the significant promise deep learning techniques have in automating the identification and classification of DR from retinal pictures. The goal of this research study is to summarize the most recent developments, methodology, difficulties, and potential future directions in the field of deep learning for the identification of diabetic retinopathy[14][16].

2.4. Decision Support System for Diabetes Management

Effective management of diabetes requires continuous monitoring, personalized treatment plans, and informed decision-making. Decision support systems (DSS) have emerged as valuable tools in diabetes care, providing clinicians with timely and evidence-based recommendations for optimal patient management. This research paper presents a comprehensive review of decision support systems for diabetes management, aiming to summarize the latest advancements, methodologies, challenges, and future directions in this field [17].

To help healthcare professionals make well-informed decisions, decision support systems use a variety of computational techniques, such as machine learning, expert systems, and data-driven algorithms. The paper reviews different components of DSS, including data acquisition and integration, knowledge representation, inference engines, and user interfaces. In order to present a comprehensive picture of the patient's health status, it investigates the integration of many data sources, including wearable technology, electronic health records, and patient-reported data [18][23][24].

3. Existing System

The current system faces challenges in implementing machine learning algorithms due to limited knowledge about data visualization and the complexity involved in mathematical calculations for model building. This often leads to time-consuming processes and increased complexity. To address these issues, we can leverage machine learning packages provided by the scikit-learn library[19][20].

By utilizing scikit-learn, we can simplify the implementation of machine learning algorithms **Fig 1**. The library offers a wide range of pre-built functions and tools that facilitate data visualization, making it easier to understand and interpret the data. This helps in gaining valuable insights and making informed decisions during the model building process [21][22].

3.1. Disadvantages of Existing System:

1. Potential for misclassification and misdiagnosis due to limited training data and lack of diversity in the dataset
2. High computational complexity and processing time required for the machine learning algorithms, which can limit the scalability.

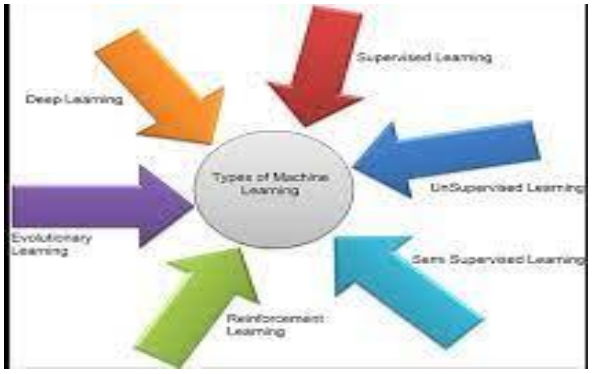

Fig 1: Examination of Machine Learning Methods for Disease Diagnosis

4. Proposed Work

By utilizing these machine learning techniques, we sought to assess their accuracy and performance in distinguishing whether a patient has diabetes or not. Our results revealed that AdaBoost yielded the highest accuracy among the tested classifiers.

To evaluate the efficacy of the different models, we conducted extensive classifier tests. Through these tests, we analyzed the performance of Decision Tree, AdaBoost, and XG Boost in accurately identifying diabetes in patients. Among the models examined, AdaBoost consistently demonstrated superior accuracy, providing the most reliable and precise outcomes.

4.1. Advantages of Current System:

1. Requires less time
2. Good Accuracy
3. Easy to Handle

5. Methodology and Algorithm

5.1. Data-Set:

The dataset I mentioned in the previous response is commonly referred to as the "Pima Indians Diabetes Database." It comes from the National Institute of Diabetes and Digestive and Kidney Diseases and is a well-known dataset. The dataset's main goal is to determine a patient's likelihood of having diabetes using a variety of diagnostic metrics that are supplied inside the dataset.

The dataset has specific constraints regarding patient selection. The dataset only contains patients who are female, at least 21 years old, and of Pima Indian ancestry. These constraints were imposed to focus the analysis on a particular demographic group for research purposes.

Researchers and practitioners often use this dataset to develop machine learning models for diabetes prediction and explore various data analysis techniques. Due to its accessibility and applicability to the job of diabetes diagnosis, it has been frequently utilized as a benchmark dataset in the fields of machine learning and diabetes research.

SourceLink: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

5.2. Parameters:

Input: here user/patient need to provide the parameters like (sugar, height, weight) to predict diabetes.

Output: After the input received form the user the model will predict the output from the data it will predict the diabetes **Fig 2** .

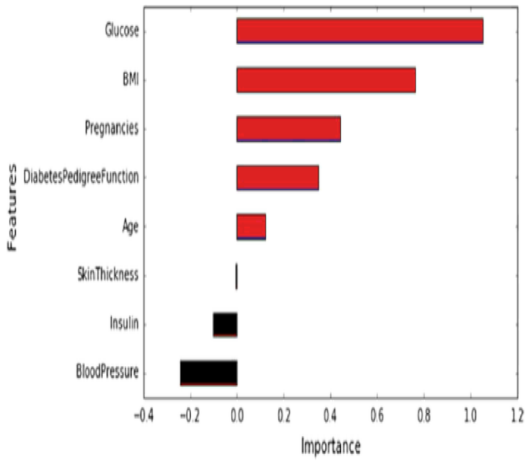

Fig 2: Output for prediting diabetes.

5.3. Algorithms used:-

5.3.1. Using Support Vector Machines (SVM):

In both supervised and unsupervised machine learning, Support Vector Machines (SVM) have become potent tools for analyzing and addressing a variety of difficulties in diabetes treatment. In the realm of supervised learning, SVMs have been extensively employed in different aspects of diabetes management. These algorithms utilize labeled data, comprising features derived from patient demographics, clinical measurements, and biomarkers, to make predictions or classifications related to diabetes risk, diagnosis, and treatment outcomes.

5.3.2. Decision Tree:

Decision Trees have proven to be effective tools in both supervised and unsupervised machine learning for reviewing and addressing various aspects of diabetes care. In the context of supervised learning, Decision Trees are widely used for classification and regression tasks related to diabetes management. These algorithms learn from labeled data, consisting of patient characteristics, clinical measurements, and medical histories, to make predictions and inform decision-making processes.

5.3.3. AdaBoost Classifier:

For reviewing and treating many facets of diabetes treatment, the AdaBoost (Adaptive Boosting) Classifier is a potent machine learning method that has been widely employed in both supervised and unsupervised contexts. In supervised learning, the AdaBoost Classifier is commonly employed for classification tasks related to diabetes management. This algorithm combines multiple weak classifiers to create a strong ensemble classifier that can accurately classify patients into different categories based on their features and attributes.

6. Architecture

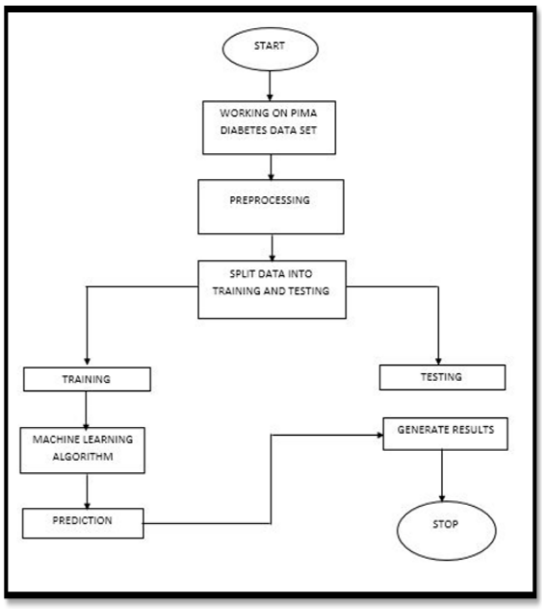

Fig 3: Block Diagram for System Architecture

Here's an explanation of how these components work together which mentioned in **Fig3.**

6.1. Data Preprocessing:

Data Cleaning: We employ techniques to handle missing values, outliers, and inconsistent data in the diabetes dataset.

Feature Engineering: We extract relevant features from the dataset, including patient demographics, clinical measurements, laboratory results, and medical history.

6.2. Supervised Learning Component:

Algorithm Selection: Support Vector Machines (SVM), Random Forest, and Logistic Regression are some examples of suitable supervised learning algorithms that we assess and choose.

Model Training: We train the selected algorithms on labeled data, utilizing features derived from the diabetes dataset.

Performance Evaluation: Using evaluation criteria including accuracy, precision, recall, and F1-score, we evaluate the models' performance.

6.3. Unsupervised Learning Component:

Clustering Analysis: We use unsupervised learning methods to find subgroups or trends in the diabetic population, such as K-means clustering or hierarchical clustering.

Dimensionality Reduction: We lower the dataset's dimensionality and determine the most crucial factors influencing diabetes care using methods like Principal Component Analysis (PCA) Table 2.

6.4. Model Validation and Interpretation:

Cross-Validation: To verify the effectiveness of the supervised learning models and assure their generalizability, we cross-validate them.

Model Interpretation: We interpret the results and provide insights into the important features and relationships learned by the models.

6.5. Decision Support and Recommendations:

Risk Prediction: The system can predict the risk of diabetes or its complications for individual patients based on their characteristics.

Treatment Recommendation: The system can provide personalized treatment recommendations based on the patient's profile and historical data.

The goal of our proposed system is to improve risk assessment, diagnosis, and therapy recommendation by incorporating supervised and unsupervised machine learning approaches. Informed judgements made by

healthcare professionals with the help of the system could ultimately improve diabetes patient care and results

7. Experimentation and Results

Among the most dangerous disorders is diabetes mellitus. Diabetes is mostly caused by a combination of factors, including but not limited to: advanced age, obesity, inactivity, genetics, lifestyle, bad nutrition, hypertension, etc. Table 1 shows that the most popular supervised learning classification algorithms are decision trees or variants thereof, including XG Boost, AdaBoost, and RF. From machine learning to deep learning, the trend is moving. In a number of areas, ANN has been performing admirably, making it a favorite ML method in recent times [5] [12]. When data comes to dimensionality reduction, unsupervised learning methods like principal component analysis and linear discriminant analysis are typically leveraged. The diabetic data set has unnecessary features that are affecting the accuracy of the classifier. Therefore, a mix of supervised and unsupervised learning can improve diabetes detection and prediction. In their study [23] applied logistic regression, PCA, and K-Mean. LR was used for classification, k-means for clustering, and principal component analysis (PCA) for dimensionality reduction.

Table 1: major findings of diabetes prediction using supervised & unsupervised learning with accuracy

S.N	FINDINGS	ACCURACY
1.	PCA,K-Means and LR algorithm. PCA boosted the KMeans.	80%
2.	Author compared SVM and k-means & SVM on PID data set f	K-Mean & SVM with 99.64 % accuracy

Table 2: major findings of diabetes prediction using supervised & unsupervised Learning with Best Algorithm

S. N	FINDINGS	BEST ALGORITHM
1.	Apriori method has been used to establish a strong relationship of diabetes BMI and blood glucose level. ANN, RF and K-means clustering methods were applied for the prediction of diabetes	ANN with accuracy of 75.7%
2.	K-Mean was applied for outlier detection and then SVM was applied for the classification	K-Mean & SVM

8. Conclusion

Finally, it can be said that the use of supervised and unsupervised machine learning techniques in the context of diabetes treatment has the potential to completely transform the industry. By leveraging large datasets and sophisticated algorithms, these approaches have provided valuable insights into disease prediction, diagnosis, and personalized treatment.

Support vector machines, decision trees, and neural networks are examples of supervised machine learning algorithms that have shown successful in forecasting diabetes outcomes and locating high-risk patients. These models can analyze various patient characteristics and medical variables to generate accurate predictions, allowing healthcare professionals to intervene early and implement targeted interventions to improve patient outcomes.

Unsupervised machine learning methods, including clustering and anomaly detection, have enabled the identification of distinct patient subgroups and patterns within diabetes datasets. These approaches have shed light on previously unrecognized phenotypes, contributing to a better understanding of the disease and the development of tailored treatment strategies.

References

[1] J. Chaki, S. T. Ganesh, S. K. Cidham and S. A.Theertan, "Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review," King Saud University Journal of Computer and Information Sciences, 2020.

[2] T. M. Alam, M. A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T. I. Baig and Z. Abbas, “A model for early prediction of diabetes,” Informatics in Medicine Unlocked, 16, 100204, 2019.

[3] Sisodia's article "Prediction of diabetes using classification algorithms." Computer science procedia, vol. 132, pp. 1578-1585, 2018.

[4] M. Alehegn, R. Joshi and P. Mulay, “Analysis and prediction of diabetes mellitus using machine learning algorithm,” International Journal of Pure and Applied Mathematics, vol. 118, pp. 871-878, 2018.

[5] N. Sneha and T. Gangil, “Analysis of diabetes mellitus for early prediction using optimal features selection,” Journal of Big data, vol. 6, pp. 13,2019.

[6] Davanam, G., Pavan Kumar, T., & Sunil Kumar, M. (2021). Novel Defense Framework for Cross-layer Attacks in Cognitive Radio Networks. In International Conference on Intelligent and Smart

Computing in Data Analytics (pp. 23-33). Springer, Singapore.

[7] Ganesh, Davanam, Thummala Pavan Kumar, and Malchi Sunil Kumar. "Optimised Levenshtein centroid cross-layer defence for multi-hop cognitive radio networks." IET Communications 15.2 (2021): 245-256.

[8] Natarajan, V. Anantha, et al. "Segmentation of nuclei in histopathology images using fully convolutional deep neural architecture." 2020 International Conference on computing and information technology (ICCIT-1441). IEEE, 2020.

[9] Sreedhar, B., BE, M. S., & Kumar, M. S. (2020, October). A comparative study of melanoma skin cancer detection in traditional and current image processing techniques. In 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 654-658). IEEE.

[10] Ganesh, D., Kumar, T. P., & Kumar, M. S. (2021). Optimised Levenshtein centroid cross-layer defence for multi-hop cognitive radio networks. IET Communications, 15(2), 245-256.

[11] Balaji, K., P. Sai Kiran, and M. Sunil Kumar. "Resource aware virtual machine placement in IaaS cloud using bio-inspired firefly algorithm." Journal of Green Engineering 10 (2020): 9315-9327.

[12] Peneti, S., Sunil Kumar, M., Kallam, S., Patan, R., Bhaskar, V., & Ramachandran, M. (2021). BDN-GWMNN: internet of things (IoT) enabled secure smart city applications. Wireless Personal Communications, 119(3), 2469-2485.

[13] Balaji, K., P. Sai Kiran, and M. Sunil Kumar. "Power aware virtual machine placement in IaaS cloud using discrete firefly algorithm." Applied Nanoscience (2022): 1-9.

[14] Davanam, G., Kumar, T. P., & Kumar, M. S. (2021). Efficient energy management for reducing cross layer attacks in cognitive radio networks. Journal of Green Engineering, 11, 1412-1426.

[15] Kumar, M. Sunil, and K. Jyothi Prakash. "Internet of things: IETF protocols, algorithms and applications." Int. J. Innov. Technol. Explor. Eng 8.11 (2019): 2853-2857.

[16] AnanthaNatarajan, V., Kumar, M. S., & Tamizhazhagan, V. (2020). Forecasting of Wind Power using LSTM Recurrent Neural Network. Journal of Green Engineering, 10.

[17] Rupesh, B., & Kumar, M. S. (2015). Predicting the Hard Keyword Queries over Relational Databases. International Journal of Applied Engineering Research, 10(10), 26629-26640.

[18] Prasad, T. G., Turukmane, A. V., Kumar, M. S., Madhavi, N. B., Sushama, C., & Neelima, P. (2022). CNN BASED PATHWAY CONTROL TO PREVENT COVID SPREAD USING FACE MASK AND BODY TEMPERATURE DETECTION. Journal of Pharmaceutical Negative Results, 1374-1381.

[19] Sangamithra, B., Manjunath Swamy, B.E., Sunil Kumar, M. (2022). Personalized Ranking Mechanism Using Yandex Dataset on Machine Learning Approaches. In: Kumar, A., Ghinea, G., Merugu, S., Hashimoto, T. (eds) Proceedings of the International Conference on Cognitive and Intelligent Computing. Cognitive Science and Technology. Springer, Singapore. https://doi.org/10.1007/978-981-19-2350-0_61

[20] Burada, S., Swamy, B.E.M., Kumar, M.S. (2022). Computer-Aided Diagnosis Mechanism for Melanoma Skin Cancer Detection Using Radial Basis Function Network. In: Kumar, A., Ghinea, G., Merugu, S., Hashimoto, T. (eds) Proceedings of the International Conference on Cognitive and Intelligent Computing. Cognitive Science and Technology. Springer, Singapore. https://doi.org/10.1007/978-981-19-2350-0_60

[21] Burada, Sreedhar, Manjunathswamy Byranahalli Eraiah, and M. Sunil Kumar. "Optimal hybrid classifier with fine-tuned hyper parameter and improved fuzzy C means segmentation: skin cancer detection." International Journal of Ad Hoc and Ubiquitous Computing 45.1 (2024): 52-64.

[22] Godala, Sravanthi, and M. Sunil Kumar. "A weight optimized deep learning model for cluster based intrusion detection system." Optical and Quantum Electronics 55.14 (2023): 1224.

[23] Sangamithra, B., BE Manjunath Swamy, and M. Sunil Kumar. "Evaluating the effectiveness of RNN and its variants for personalized web search." Optical and Quantum Electronics 55.13 (2023): 1202.

[24] N. P. Tigga and S. Garg, “Prediction of Type 2 Diabetes using Machine Learning Classification Methods,” Procedia Computer Science, vol. 167, pp. 706-716, 2020.