EARLY PREDICTION OF DIABETES MELLITUS ON PIMA DATASET USING ML AND DL TECHNIQUES

Ovass Shafi Zargar[1], Avinash Bhagat[2], Tawseef Ahmed Teli[3*], Sophiya Sheikh[4]
[1,2,4]*School of Computer Applications, LPU, India*
[1]*owaisfour03@gmail.com*
[2]*avinash.baghat@lpu.co.in*
[4]*sophiyasheikh@gmail.com*
[3*]*Department of Higher Education, J&K, India*
[3]**mtawseef805@gmail.com*

Abstract— Ranked fourth among the top fatal diseases with a significantly higher mortality rate is Diabetes Mellitus (DM), which is triggered by inadequate insulin production by the pancreas or human body insulin resistance resulting in high insulin demand that is not met by the pancreas. Diabetes Mellitus can lead to various other diseases, such as kidney ailments, cardiac problems, blindness, and brain damage. Advances in technology, particularly in AI, have greatly increased the use of various data mining techniques in healthcare, providing a boon for patients. Data mining techniques also find applications in extracting useful patterns and features from diabetes datasets to assist in the process of classifying and diagnosing DM in its early stages. This work provides a detailed review of various ML and DL techniques and their contribution to predicting diabetes mellitus in its early stages. An in-depth analysis of several machine-learning techniques is provided. A comparative analysis of the most promising ML-based techniques has been provided.

Index Terms— Diabetes, prediction, pre-processing, SVM, DT, RF, KNN, GNB, GB.

INTRODUCTION

DM is a metabolic ailment that is prompted by high concentrations of glucose, and if left untreated, it can lead to various diseases related to the heart, eyes, kidneys, liver, and brain. Every passing day, the rate of diabetes patients is growing at a distressing rate, and based on the prediction made by the International Diabetes Federation, by 2035, the number will get to 592 million. Type I (Juvenile diabetes) is caused by damage to insulin-releasing cells by our immune system, ultimately inhibiting insulin production in the body. This type of diabetes accounts for only 10% of diabetes patients globally. Type II (insulin-independent diabetes) is caused due to insulin resistance in the body, leading to increased insulin demand in the body. This type of diabetes accounts for almost 90% of diabetes patients worldwide. Another variant is Gestational diabetes, which mainly occurs in pregnancy and may pose a threat to both mothers and babies.

The global high mortality rate caused by DM has caused havoc worldwide. However, with the increasing use of ML and DL algorithms in making predictions in eCommerce and better business decisions, there is a ray of hope for using these techniques in medical science to assist in the timely prediction of various diseases. Today, with a vast volume of medical data available, there is a possibility to apply machine learning algorithms to these datasets and find useful patterns and hidden information that can later be used to predict diseases much earlier before their onset.

In machine learning, classification involves building a model that identifies and categorizes a dataset into distinct classes, while clustering is a process that examines data objects without utilizing class labels, grouping samples into new classes by maximizing the similarity between them. Association Rule Learning (ARL) is another approach that mines frequent patterns from data.

The remainder of this paper is organized as section 2 discusses the importance of this study. Section 3 gives a review of the current literature followed by a complete analysis of ML and DL techniques and the results achieved in various studies. Section 5 discusses the experiment and results followed by a discussion given in section 6. Lastly, the final remarks are given in section 7.

SIGNIFICANCE OF THE STUDY

In 2018, around 11% of the United States population was affected by diabetes, with 1/5 of those cases being undiagnosed. Unfortunately, many individuals are unaware of their susceptibility to diabetes until the disease has progressed significantly. Therefore, early detection of diabetes is essential to avoid severe problems. While diabetes cannot be cured completely, early detection can aid in reversing some of its effects and help patients achieve remission by maintaining normal blood sugar levels without long-term medication. Machine learning (ML) and deep learning (DL) can play a major role in early detection. The medical industry generates vast amounts of data from hospitals, nursing homes, clinical health centres, and polyclinics, making it challenging to process manually. Employing various DL algorithms can extract hidden relations and information from the datasets and forecast the onset of diabetes before the disease progresses. This proactive approach enables necessary measures to be taken to prevent multiple health-related problems in patients and help them lead healthy and fulfilling life. However, the raw dataset may contain multiple anomalies, such as missing values, redundant information, null values for some attributes, and erroneous values, making it challenging to apply DL algorithms to it. Therefore, the dataset must be processed and converted into a usable form that aids informed decision-making.

REVIEW OF RELATED STUDIES

In a research study [1], a Firefly Optimized Neural Network was proposed and compared with multiple ML algorithms. The algorithm had better accuracy than ANN, achieving an accuracy of 95.07%. In a separate research study [2], commonly used machine learning algorithms were compared, and the authors concluded that SVM outperformed other techniques. Another study [3] conducted a review of ML algorithms including DT, NB, BN, KNN, KStar, LR, ANN, and SVM on the PIMA dataset. The authors concluded that KNN and Logistic Regression provide better accuracy. In a study [4], an ensemble framework was proposed to predict diabetes mellitus, and its performance was compared to commonly used machine learning algorithms such as LR, ANN, DT, NB, DNN, BayesNet, AdaBoost, Decision Bagging, and RF. It was successfully shown that the framework outperformed other ML algorithms.

ML techniques have also found applications in various fields such as cryptography and networks [5], navigation [6], [7], and predictive analysis and healthcare. Deep learning techniques [8], which have applications in secure healthcare [9], [10], [11], have gained significant attention in recent times. In addition to disease prediction [12], ML and DL approaches are also suitable for drug design.

MACHINE AND DEEP LEARNING

With technological advancements, the lifestyle of modern individuals has become increasingly comfortable, leading to a reduction in physical activity and a rise in various health issues, including diabetes mellitus (DM), which has become a significant problem in the last two decades. The diagnosis and efficient treatment are challenging due to its complex mechanisms and related symptoms. Artificial intelligence (AI) has been applied in healthcare, including the use of ML and DL to process large datasets generated by medical industries such as hospitals, nursing homes, and clinical laboratories. ML and DL algorithms can extract hidden patterns and information from these datasets, which are too large to be

processed manually. By applying ML algorithms to diabetes datasets, researchers can predict the onset of diabetes and potentially improve the health outcomes of the population.
Table I and Table II summarize some of the ML and DL-based research and comparative analysis conducted in this field respectively. The number of the most relatable paper on the PIMA dataset using ML and DL techniques is given in Figure 1 and Figure 2 respectively.

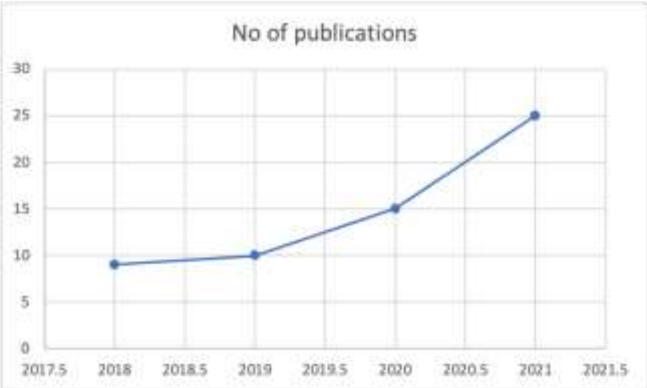

Figure 1. ML-based publications (Year wise)

There has been an upward trend in publishing articles on implementing ML techniques for DM prediction while DL-based techniques see a dip. The reason for this could generally be attributed to the smaller dataset size of PIMA.

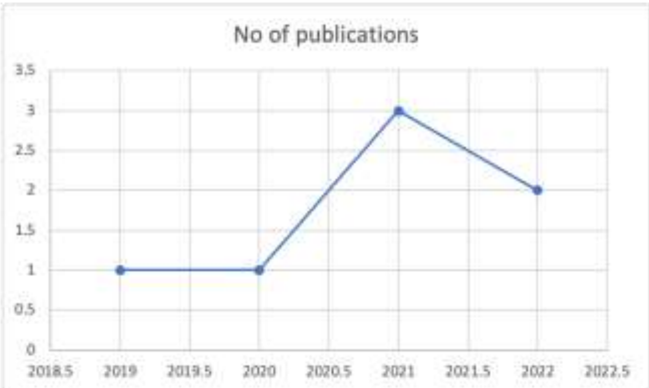

Figure 1. DL-based publications (Year wise)

EXPERIMENT AND RESULTS

Most of the above-discussed techniques have used the PIMA dataset for their studies. A total of 9 attributes, that are significant to the process of prediction are used**.** With a total of 768 samples, 500 records are values with outcome attribute 0 (non-diabetic) and 268 records have a value of 1(diabetic). Based on the results obtained in various studies done by researchers and discussed in the section, some of the most significant models have been implemented on the PIMA. The results are compared on various evaluation parameters to provide detailed insights into the optimality of these methods for diabetes prediction. The following ML techniques were implemented in Keras and TensorFlow:

- KNN
- SVC
- LR
- DT

- GNB
- RF
- GB

The following evaluation criteria have been used to ascertain the performance of different ML methods:

The confusion matrix (CM) is used to analyze the performance and accuracy of any supervised learning algorithm. The confusion matrix will be used for the same. The performance of the algorithm/technique can be calculated as:

$$Acc = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

$$Sen = \frac{(TP)}{(TP+FN)} \quad (2)$$

$$Spec = \frac{(TN)}{(TN+FP)} \quad (3)$$

$$F1\ Score = \frac{(2TP)}{(2TP+ FP+FN)} \quad (4)$$

The CM is given in Figure 3:

TABLE I: ML ALGORITHMS

Ref.	Year	Method	Dataset	Best Performance Method	Result of Best Performance Method
[2]	2021	Various ML Techniques	PIMA	SVM	Accuracy: 98%
[3]	2021	Various ML Techniques	PIMA	KNN, Logistic Regression	Accuracy: 80%
[4]	2021	LR, ANN, DT, NB DNN BayesNet, AdaBoost, Decision Bagging, RF, Proposed Ensemble Model	PIMA	Proposed Ensemble Model	Accuracy: 79.22%
[12]	2019	Decision Trees SVM Naive Bayes	PIMA	Naive Bayes	Precision: 0.757 Recall: 0.761 Measure: 0.758 Accuracy: 74.28% ROC: 0.817
[13]	2020	Neural Network, FONN	PIMA	FONN	Accuracy: 95.07% Precision: 88%

					Recall: 88%
[14]	2021	J48, CART and Naive Bayes	PIMA	J48 and CART	Accuracy: 99%
[15]	2021	SVM Logistic Red. Logistic Step Elastic Net LGBM: BstLinTree LDA XGB: Tree LGBM: Boost Tree XGB: Linear C5.0 Rand F. Red. LGBM: RF CART Naive Bayes Red. K/TF DenseNN	PIMA	LGBM: Boost Tree	Accuracy: 93.44%
[16]	2021	Various ML Techniques	PIMA	SVM	Accuracy 97.87%
[17]	2021	LR, LDA, NB, K-NN, CART, SVM	PIMA	Naive Bayes	Accuracy 95%
[18]	2021	Naive Bayes Neural Network AdaBoost kNN Random SVM	Early-stage diabetes risk prediction dataset	Random Forest	Accuracy 99.3%
[19]	2018	Back propagation J48 NB, SVM	PIMA	Back Propogation	Accuracy 83.11%
[20]	2021	RF, LR, DT, SVM, NB, KNN, EM	PIMA	Ensemble Method	Accuracy 87.09%
[21]	2021	DT	PIMA	Decision Tree	Accuracy: 71.35%
[22]	2021	DT, KNN, SVM, RF, NB, LR	PIMA	Random Forest Linear Regression	Accuracy: 90%
[23]	2021	RF KNN MLP Ada boost D tree Classifier NB	PIMA	Stacked ensemble combined with genetic algorithms	Accuracy: 98%

[24]2020	GBC SVM Extra Tree Suggest Method (ST-GA) CNN CUSTOMIZED CNN Radial Basis Neural Network Function Genetic Algorithm	PIMA	CUSTOMIZED CNN	Accuracy: 80%
[25]2021	DUNN Index Davies Bouldin Index Silhouette Index DT LR SVM KNN NB GB	PIMA	Logistic Regression	Accuracy: 80%
[26]2021	LR, KNN, SVM, NB, DT, RF, Soft Voting Classifier, AdaBoost, Bagging, GradientBoost, XGBoost, CatBoost	PIMA	Soft Voting Classifier	Accuracy:79.08% Precision: 73.13% F1 Score:71.56% Recall:70%
[27]2021	ANN Random Forest Clustering	PIMA	ANN	Accuracy: 75.8%
[28] 2019	SVM RF CNN	PIMA	RF	Accuracy: 83.67%
[29] 2019	SVM RF, NB, DT, KNN	PIMA	SVM	Accuracy: 77.73%
[30]2019	J48 NB RF, LR,	PIMA	Logistic Regression	Accuracy: 77% Precision: 0.77 Recall:0.77 F-Score:0.76 AUC: 0.83
[31] 2021	ADA BOOST with RF ADA BOOST with Extra Tree	Image Dataset taken from local Clinic	Ada Boost with RF	Accuracy: 96.71% Precision: 97.55 Sensitivity: 97.95 F1-Score: 97.75

[32]2020	KNN CNN SVM LR Extreme Learning	PIMA	Extreme Learning	Accuracy: 90.54%
[33]	2018 NB SVM DT	PIMA	Naive Bayes	Accuracy: 76.3% F-Measure: 0.76 Precision: 0.759 Recall:0.763
[34]2020	LR KNN SVM NB DT RF	PIMA	Random Forest	Accuracy: 75.0% Sensitivity: 0.250 Specificity:0.789 Precision:0.661
[35]2021	RF SVM AdaBoost Gradient Boosting	PIMA	RF	Accuracy 99.35% SEN 99.01 % SPE 100% FPR 0% FNR 0.99% NPV 98.15%
[36]2021	LR KNN SVM NB DT RF	PIMA	KNN	Precision: 0.747 Recall: 0.751 F-Measure: 0.749 Accuracy: 75.10%
[37]2020	KNN LR DT RF, SVM MLP classifier Proposed CNN	PIMA	Proposed CNN	Accuracy: 93.2%
[38]2018	Linear Kernel SVM Radial Basis Kernel SVM KNN ANN MDR	PIMA	Linear Kernel SVM	Accuracy: 89% Precision: 0.87 Recall: 0.88 F1-Score 0.87 AUC: 0.90

[39]2021	RF (cross-validation) NB (cross-validation) KNN (cross-validation) J48(cross-validation) RF (split method) NB (split method) KNN (split method) J48 (split metho	PIMA	Random forest (cross-validation)	F-measure: 0.983 MCC: 0.9654 AOC RUC: 0.999 PR AUC: 0.999 Accuracy: 98.3055%
[40]2021	NB DT SVM	PIMA	decision tree	Accuracy: 85%
[41]2018	SVM Bayes Net DecisionStumb AdaBoostM1 Proposed method (PM)	PIMA	Proposed Method (PM)	Accuracy: 90.36%
[42]2021	LR RF SVM ANN ANN	Pregnant cohort study in eastern China	Random Forest	Accuracy: 86.91% Sensitivity: 63.30 Specificity: 97.53 AUC: 0.80
[43]2019	SVM K-NN DT NB LR	PIMA	Logistic regression	Accuracy:77.61% Recall:0.8902 Precision:0.7979
[44]2021	DLCNN, CTCPN, LVQOAC, MODLNN	PIMA	DLCNN	Accuracy: 98.42%
[45]2020	KNN NB RF SVM DT LR	Wisconsin dataset (University of Wisconsin Hospit	Decision Tree and Logistic Regression	Accuracy: 97%

		als, USA)		
[46]2021	LR SVM DT RF	HbA1c-labeled and FPG-labelled datasets	SVM	Accuracy: 82.10% Precision: 82.30 Recall: 82.10 F1 Score: 82.05
[47]2018	SVM	MESSIDOR	SVM	Accuracy: 90.04%
[48]2018	RT SVM LR MLP	Chronic Kidney Disease Dataset from Apollo Hospital	Logistic Regression and Multilayer Perceptron	Accuracy:98.1% F1 score:98.4
[49]2018	RF LR MLP neural network	PIMA	MLP neural network	Accuracy: 77.08%
[50]2020	Ensemble of ADA Boot XG Boost	PIMA	Ensemble of ADA Boot XG Boost	Accuracy: 95.0%
[51]2021	RF DT NB LR ADA Boost	PIMA	Random Forest	Accuracy: 94.0%
[52]2019	SVM NB KNN C4.5 DT	Diagnostic dataset from medic	C4.5 Decision Tree	Accuracy: 74.0%

		al Center		
[53]2020	KNN SVM RF	PIMA	Random Forest	Accuracy: 74.47% Precision: 80.48 Recall: 79.83 F1-Score: 80.16
[54]2020	K-Means Algorithm LR SVM KNN RF DT NB	PIMA	SVM	Accuracy: 93%
[55]2018	NB SVM RF Simple CART	PIMA	SVM	Accuracy: 79.13%
[56]2018	SVM KNN LR DT RF NB	PIMA	SVM and KNN	Accuracy: 77%
[57]2019	RF	UCI Learning Repository	Random Forest	Accuracy: 90%
[58]2019	RF	Clinical Dataset	Random Forest	Accuracy: 95.1%
[59]2020	Glmnet RF XGBoost LightGBM)	Clinical Dataset	Glmnet	Accuracy: 95%
[60]2019	Various ML Techniques	Diabetes	RF	Accuracy: 99%

		Hospital of Sylhet, Bangladesh.		
[61]2019	RF XGBoost	PIMA	XGBoost	Accuracy: 74.10%
[62]2020	Linear Discriminant Analysis (LDA)	PIMA	LDA	Precision:0.701 Recall: 0.817 Specificity: 0.720 F-Score: 0.755 Accuracy: 76.86%
[63]2020	ANN NB DT SVM	Data collected from android application and PIMA Dataset	SVM	Accuracy:81.6% Sensitivity:87.32 Specificity:73.46
[64]2021	RBF	PIMA	RBF	TP Rate: 0.459 FP Rate: 0.819 Precision:0.792 Recall:0.860 F-Measure: 0.825 MCC:0.459 Recall: 0.792 ROC Area: 0.890
[65]2020	KNN	PIMA	K-Nearest Neighbor	TPR: 77.36 TNR: 89.11 FPR: 10.89 FNR: 22.64 F1 score:78.10% Accuracy:85.06% Recall:77.36 % Precision:78.85%

[66]2020	DT AdaBoost RF	PIMA	Random Forest	Specificity:89.11% Sensitivity: 99.56 Positive predictive value: 93.25 Negative predictive value: 89.98 F-measure: 96.30
[67]2020	SVM XG Boost	PIMA	XG Boost	Accuracy: 77.0%

*Refer to Appendix I for acronyms

TABLE II: DL ALGORITHMS

Ref.Year	Method	Dataset	Best Performance Method	Result of Best Performance Method
[68]2022	Deep Neural Network	PIMA	Deep Neural Network with missing values handling	Accuracy: 80.0 (MAX)
[69]2020	Deep Learning Decision Tree Artificial Neural Network Naïve Bayes	PIMA	Deep Learning	Accuracy: 98.07
[70]2021	Deep Learning SVM	PIMA	Deep Neural Network	Accuracy: 77.474
[71]2021	Deep Learning Perceptron SVM	PIMA	Deep Learning Perceptron	Accuracy: 65.10
[72]2019	Logistic Regression Improved GA Modified K-Means + SVM SVM with efficient coding Deep Neural Network	PIMA	Deep Neural Network	Accuracy: 98.35
[73]2021	Deep Learning TLSTM CLSTM	PIMA	Deep learning	Accuracy: 93.7% Accuracy: 95.6%
[74]2022	DNN + 10-fold cross-validation	PIMA	Deep Neural Network	Sensitivity: 87% Specificity: 91% Accuracy: 89%

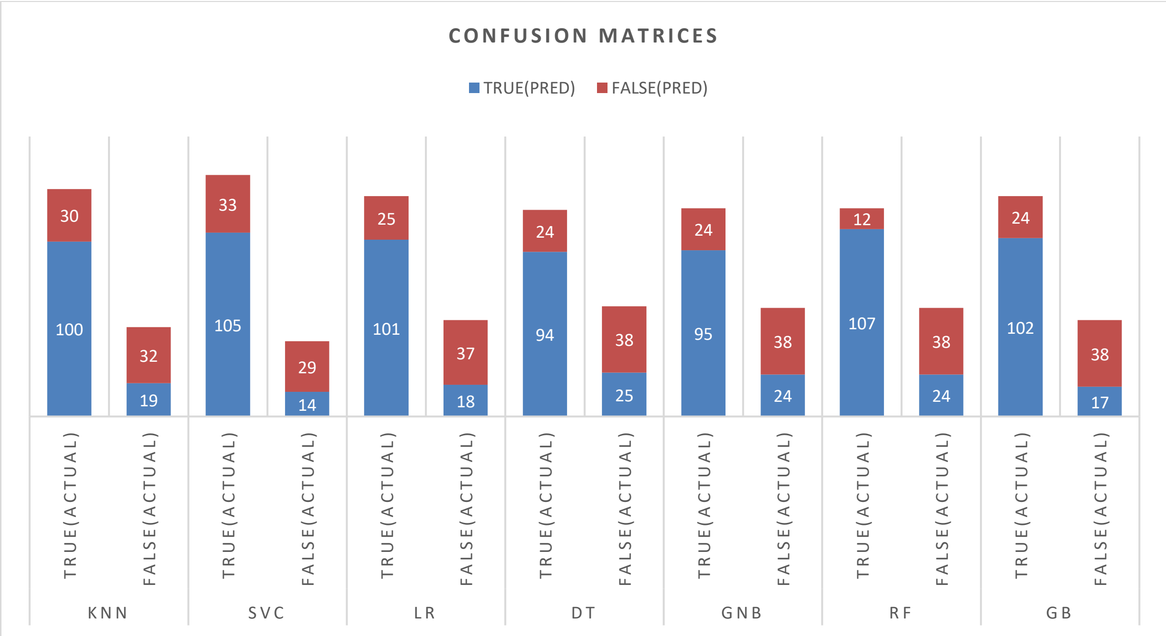

Figure 3. Confusion Matrix of various ML techniques.

The analysis of ML-based methods reveals that some of the techniques provide better results than others. Based on this thorough comparative analysis, it was observed that few techniques including KNN, SVC, LR, DT, GNB, RF and GB provide results that are consistent among all studies. Subsequently, these techniques provide a promising roadmap to approaching this predictive problem. With this vision, the techniques were implemented for comparative analysis, as shown in Figure. 3. The CM provides a clear summary of good classifiers for DM diagnosis.

Among the methods, RF gives better results as enumerated in Table II.

The performance matrix for various ML techniques is given in Table III:

TABLE III: PERFORMANCE OF ML ALGORITHMS

Algorithm	Precision	Recall	Specificity	F1-Score	Accuracy
KNN	0.8403	0.7692	0.6275	0.8032	0.7293
SVC	0.8824	0.7609	0.6744	0.8171	0.7403
LR	0.8487	0.8016	0.6727	0.8245	0.7624
DT	0.7899	0.7966	0.6032	0.7932	0.7293
GNB	0.7983	0.7983	0.6129	0.7983	0.7348
RF	0.8992	0.8168	0.7600	0.8560	0.8011
GB	0.8571	0.8095	0.6909	0.8327	0.7735

The rates of various parameters of these ML techniques are given in Table IV:

TABLE IV: RATES OF ML ALGORITHMS

Algorithm	TPR	FNR	TNR	FPR
KNN	0.7692	0.2308	0.6275	0.3725
SVC	0.7609	0.2391	0.6744	0.3256
LR	0.8016	0.1984	0.6727	0.3273
DT	0.7966	0.2034	0.6032	0.3968
GNB	0.7983	0.2017	0.6129	0.3871
RF	0.8168	0.1832	0.7600	0.2400
GB	0.8095	0.1905	0.6909	0.3091

The RF algorithm performs fairly well in terms of the rates as well. In future, the same technique could be used in tandem with DN-based hybrid techniques to achieve higher accuracies.

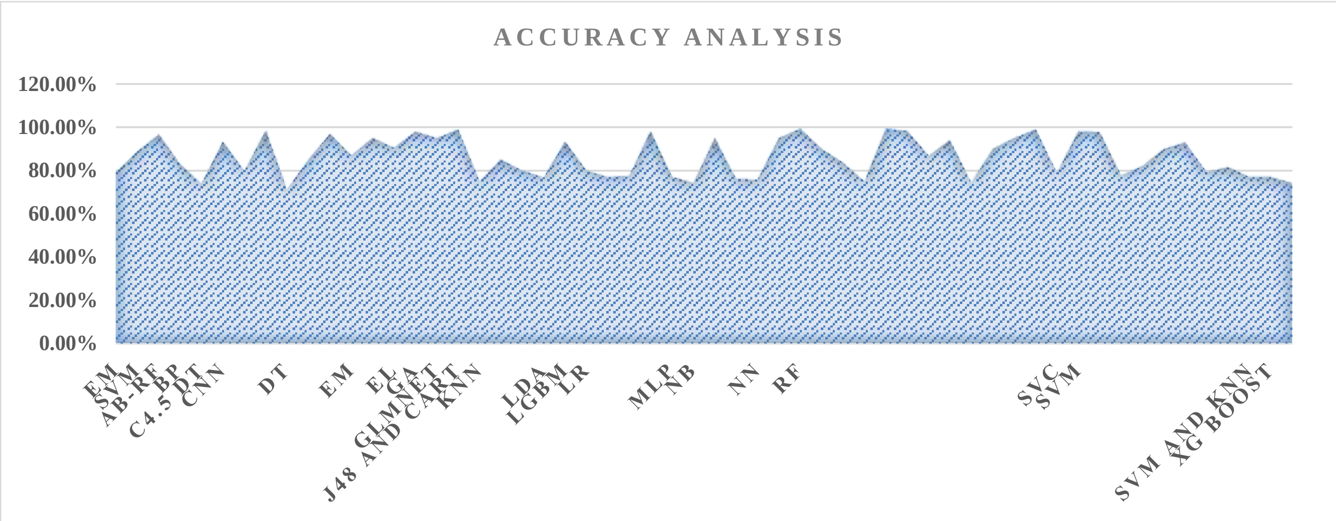

Figure 4. Accuracy Analysis.

Figure 4 summarizes the changes in accuracies achieved by various ML techniques applied by researchers over many years. The accuracy falls between 65% to touching 100%. Some of the better-performing algorithms include RF, SVM and DT etc. Many of these basic ML algorithms have been modified to achieve better results. It is evident from the figure that many versions of the same algorithm perform variedly with different rectifications and hybrid models perform better frequently.

Discussion

Many research studies do not provide a single classification model for predicting both Type I and Type II diabetes [43]. There has been the use of a single dataset with few records which doesn't provide reliable results [41]. Many researchers use the PIMA dataset without any pre-processing technique like normalization. As a result, the results suffer from outliers, overfitting, underfitting and other anomalies [53], [35] There have been studies that used limited machine learning algorithms to diagnose diabetes and didn't handle the missing values. [12]. In other studies, some authors have inhibited the application of feature extraction fully. The feature extraction process could be enhanced by the application of an automatic process of deep feature extraction [28].

Some studies do not consider the importance of all the attributes of the dataset. Attributes like body size, height and BMI and their contribution to the diagnosis of diabetes mellitus have not been considered, which affects the performance of the classifier [39]. Many authors investigated only matricellular proteins as biomarkers however there are multiple biomarkers like microRNAs, angiographic vasospasm etc. Some models suffer from the anomaly of oversampling [58]. It has also been brought to light that the medication affects the attributes of the patients, many researchers in their research did not collect any data regarding the medication of patients which limits the performance of the classifier [48].

The results suggest that the Random Forests provide better accuracy followed by Gradient Boost and Logistic Reasoning. For early detection of diabetes mellitus, these algorithms may be considered while making efforts to keep in the queue the above-discussed insights for better performance and results.

Deep Neural based algorithms show promising results but fail to achieve higher accuracy without bias due to the smaller dataset size. A valuable insight into the problem is to use k-fold cross-validation techniques in tandem with a DNN to achieve promising results.

Conclusion

Several researchers had extensively studied diabetes mellitus, a life-threatening disease, due to its widespread grip on the world population. Many trials had been conducted to improve ML techniques for better accuracy. This study focused on analyzing and comparing various techniques to uncover their limitations and drawbacks. Many parameters, such as missing values, inadequate datasets, inefficient feature extraction, reduced biomarkers, and medication effects on significant parameters, were often disregarded while using ML and DL classifiers for diabetes mellitus diagnosis. The study established that Random Forests, Gradient Boost, and Logistic Reasoning classifiers performed better and should be considered for future research, incorporating all significant parameters that may limit classifier performance. The DL-based techniques require a larger dataset, a DNN shall be preferred only when cross-validation is integrated.

Appendix

Appendix I is added to the glossary.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Each author contributed equally to the paper.

References

[1] V. Vaidya and L. K. Vishwamitra, “Data Mining Based Prediction of Diabetes Using Firefly Optimized Neural Network”, 1997. doi:10.1109/CSNT51715.2021.9509590.

[2] T. Sharma and M. Shah, “A comprehensive review of machine learning techniques on diabetes detection,”, *Vis. Comput. Ind. Biomed. Art*, vol. 4, no. 1, 30, 2021. doi:10.1186/s42492-021-00097-7.

[3] L. Ismail et al., “Type 2 diabetes with artificial intelligence machine learning: Methods and evaluation,” *Arch. Comp. Methods Eng.*, vol. 89, 01234567, 2021. doi:10.1007/s11831-021-09582-x.

[4] Education, M. Prakash et al., "An ensemble technique for early prediction of type 2 diabetes mellitus – A normalization approach," vol. 12, no. 9, pp. 2136-2143, 2021.
[5] T. A. Teli and F. Masoodi. "Blockchain in Healthcare: Challenges and Opportunities.", *2nd International Conference on IoT Based Control. Networks and Intelligent Systems (ICICNIS 2021)* Available at: https://ssrn.com/abstract=3882744, 2021.
[6] T. A. Teli and M. A. Wani, "A fuzzy-based local minima avoidance path planning in autonomous robots," *Int. J. Inf. Technol.*, vol. 13, no. 1, pp. 33-40, 2021. doi:10.1007/s41870-020-00547-0.
[7] T. A. Teli et al. "Security Concerns and Privacy Preservation in Blockchain-based IoT Systems: Opportunities and Challenges". ,*International Conference on IoT based Control. Networks and Intelligent Systems (ICICNIS 2020)* Available at: https://ssrn.com/abstract=3769572, 2020.
[8] S. J. Sidiq et al., "Big data and deep learning in healthcare" in *Appl. Artif. Intell. Big Data Internet Things Sustain. Dev.*, 2022. doi:10.1201/9781003245469.
[9] T. A. Teli et al., "MANET routing protocols, attacks and mitigation techniques: A review" in *Int. J. Mech. Eng.*, vol. 7, no. 2, 2022.
[10] T. A. Teli et al., "HIBE: hierarchical identity-based encryption in *Functional Encryption." EAI/Springer Innovations in Communication and Computing*, K. A. B. Ahmad, K. Ahmad and U. N. Dulhare, Eds. Cham: Springer, 2021, 187-203. doi:10.1007/978-3-030-60890-3_11.
[11] T. A. Teli et al., "Ensuring secure data sharing in IoT domains using blockchain in Cyber Security and Digital Forensics", *M. M. Ghonge, S. Pramanik, R. Mangrulkar and D.-N. Le, Eds.*, 2022. doi:10.1002/9781119795667.ch9.
[12] O. Shafi et al., K. (2022), "Effect of preprocessing techniques in predicting diabetes mellitus with focus on artificial neural network" in *Adv. Appl. Math. Sci.*, vol. 21, no. 8.
[13] H. Yang et al., "Risk Prediction of Diabetes: Big data mining with fusion of multifarious physical examination indicators," *Inf. Fusion*, vol. 75, pp. 140-149, 2021. doi:10.1016/j.inffus.2021.02.015.
[14] M. Nabeel et al., "Review on effective disease prediction through data mining techniques," *October*, 2021. doi:10.15676/ijeei.2021.13.3.13.
[15] A. Adler, "Using Machine Learning Techniques to Identify Key Risk Factors for Diabetes and Undiagnosed Diabetes", 2021.
[16] F. A. Khan et al., "Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review", *IEEE Access*, vol. 9, pp. 43711-43735, 2021. doi:10.1109/ACCESS.2021.3059343.
[17] S. Sankar and S. Sathyalakshmi (N.D.), "Prediction of Endocrine Disorders Using Machine Learning Classification Algorithms: A Comprehensive", Jul. 2021, pp. 151-163. doi:10.17605/OSF.IO/FYS93.
[18] A. C. Study, 2021. *applied sciences* "Data Mining Techniques for Early Diagnosis of Diabetes" : 1–12.
[19] F. G. Woldemichael and S. Menaria, "Prediction of diabetes using data mining techniques" *2nd International Conference on Trends in Electronics and Informatics (ICOEI), Icoei*, vol. 2018, 2018, pp. 414-418. doi:10.1109/ICOEI.2018.8553959.
[20] J. February et al., "Classifier algorithms and ensemble models for diabetes mellitus prediction: A review,", *IJATCSE*, Mar., no. 1, 430-439, 2021. doi:10.30534/ijatcse/2021/641012021.

[21] T. Dudkina et al., “Classification and prediction of diabetes disease using decision tree method,” vol. 2836, pp. 0-1, 2021.
[22] I. Journal (N.D.), *IRJET*- “Various Data Mining Techniques Analysis to Predict Diabetes Mellitus.”
[23] J. Abdollahi and B. Nouri-Moghaddam (N.D.), “Hybrid stacked ensemble combined with genetic algorithms for Prediction of Diabetes,”, *Iran J. Comput. Sci.*, vol. 5, no. 3, 205-220, 2022. doi:10.1007/s42044-022-00100-1.
[24] Al-Hadhrami, T, & Mohammed, F, *Advances on Smart and Soft Computing*. 2020
[25] V. Prudhvi (N.D.), “Prediction of Diabetes Mellitus Using RBF Neural Model and Genetic”, vol. 3, 1524-1531 Algorithm. 32.
[26] S. Kumari et al., “An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier.” *Int. J. Cogn. Comput. Eng.*, vol. 2, pp. 40-46, 2021. doi:10.1016/j.ijcce.2021.01.001.
[27] T. Mahboob Alam et al., “A model for early prediction of diabetes,” *Inform. Med. Unlocked*, vol. 16, no. Jul., p. 100204, 2019. doi:10.1016/j.imu.2019.100204.
[28] A. Yahyaoui et al., “A decision support system for diabetes prediction using machine learning and deep learning techniques,” vol. 2, pp. 1-4, 2019. doi:10.1109/UBMYK48245.2019.8965556.
[29] N. Sneha and T. Gangil, “Analysis of diabetes mellitus for early prediction using optimal features selection,” *J. Big Data*, vol. 6, no. 1, 2019. doi:10.1186/s40537-019-0175-6.
[30] G. Battineni et al., “Comparative machine-learning approach: A follow-up Study on Type 2 diabetes predictions by cross-validation methods,”, *Machines*, vol. 7, no. 4, pp. 1-11, 2019. doi:10.3390/machines7040074.
[31] A. Khandakar et al., “A machine learning model for early detection of diabetic foot using thermogram images,” *Comput. Biol. Med.*, vol. 137, no. Sept., p. 104838, 2021. doi:10.1016/j.compbiomed.2021.104838.
[32] J. Chaki et al., “Machine learning and artificial intelligence-based Diabetes Mellitus detection and self-management: A systematic review,”. *Journal of King Saud University - Computer and Information Sciences*, xxxx, 2020. doi:10.1016/j.jksuci.2020.06.013.
[33] [49] D. Sisodia and D. S. Sisodia, “Prediction of diabetes using classification algorithms,” *Procedia Comput. Sci.*, vol. 132(Iccids), pp. 1578-1585, 2018. doi:10.1016/j.procs.2018.05.122.
[34] N. P. Tigga and S. Garg, “prediction of Type 2 diabetes using machine learning prediction of Type 2 diabetes using machine learning classification methods classification methods,” *Procedia Comput. Sci.*, vol. 167, pp. 706-716, 2020. doi:10.1016/j.procs.2020.03.336.
[35] M. K. Hasan et al., “Diabetes prediction using ensembling of different machine learning classifiers,”, *IEEE Access*, vol. 8, 76516-76531, 2020. doi:10.1109/ACCESS.2020.2989857.
[36] J. J. Khanam and S. Y. Foo, “A comparison of machine learning algorithms for diabetes prediction,” *ICT Express*, vol. 7, no. 4, pp. 432-439, 2021. doi:10.1016/j.icte.2021.02.004.
[37] M. Rout et al., *Nature Inspired Computing for Data Science*.
[38] H. Kaur and V. Kumari, “Predictive modelling and analytics for diabetes using a machine learning approach,” vol. 18, no. 1, pp. 92-102, 2020. doi:10.1016/j.aci.2018.12.004.

[39] O. O. Oladimeji et al., "Classification models for likelihood prediction of diabetes at early stage using feature selection,", *ACI*, 2021. doi:10.1108/ACI-01-2021-0022.
[40] I. M. Ibrahim and A. M. Abdulazeez, "The role of machine learning algorithms for diagnosing diseases,", *JASTT*, vol. 2, no. 1, pp. 10-19, 2021. doi:10.38094/jastt20179.
[41] M. Alehegn, "Analysis and prediction of diabetes mellitus using machine learning," vol. 9, pp. 871-878, 2018 Algorithm. 118.
[42] J. Wang et al., "Machine Learning Approaches for Early Prediction of Gestational Diabetes Mellitus Based on Prospective Cohort Study", pp. 1-14.
[43] A. Saxena et al., "Data mining techniques based diabetes prediction,", *IJAINN*, vol. 1, no. 2, pp. 29-35, 2021. doi:10.35940/ijainn.B1012.041221.
[44] R. Murugadoss, "Early prediction of diabetes using deep learning convolution neural network and Harris Hawks," *Optimization*, vol. 1, pp. 88-100, 2021.
[45] F. M. J. M. Shamrat et al., "An Analysis on Breast Disease Prediction Using Machine Learning Approaches an Analysis on Breast Disease Prediction Using Machine Learning Approaches."
[46] H. F. Ahmad et al., "Investigating Health-Related Features and Their Impact on the Prediction of Diabetes Using Machine Learning.". *applied sciences*, 2021
[47] M. Chetoui et al., "Diabetic Retinopathy Detection Using Machine Learning and Texture Features", 2018. doi:10.1109/CCECE.2018.8447809.
[48] A. J. Aljaaf, Al-jumeily, D. Haglan, H.M., Alloghani, M., Baker, T, & Hussain, A.J. "Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics" *IEEE Congress on Evolutionary Computation (CEC)*, vol. 2018, 2018., pp. 1-9.
[49] S. Y. Rubaiat et al., "Important feature selection & accuracy comparisons of different machine learning models for early diabetes detection," vol. 1, pp. 1-6. doi:10.1109/CIET.2018.8660831.
[50] V. Krishnapraseeda et al., "Predictive Analytics on Diabetes Data Using Machine Learning Techniques", 2021, pp. 1670-1673. doi:10.1109/ICACCS51430.2021.9441972.

[51] R. Ferdousi et al., "Early-stage risk prediction of non-communicable disease using machine learning in health CPS," *IEEE Access*, vol. 9, pp. 96823-96837, 2021. doi:10.1109/ACCESS.2021.3094063.
[52] M. F. Faruque et al., "Performance analysis of machine learning techniques to predict diabetes mellitus" *International Conference on Electrical, Computer and Communication Engineering (ECCE)*, vol. 2019, 2019, pp. 1-4. doi:10.1109/ECACE.2019.8679365.
[53] J. R. Raut, "Performance evaluation of various supervised machine learning algorithms for diabetes," vol. 7, no. 8, pp. 4921-4925, 2020.
[54] P. M. S. Sai et al., 2020, "Survey on Type 2 Diabetes Prediction Using Machine learning," Iccmc, pp. 770-775. doi:10.1109/ICCMC48092.2020.ICCMC-000143.
[55] A. Mir and S. N. Dhage, "Diabetes disease prediction using machine learning on big data of healthcare" *Fourth International Conference on Computing Communication Control. and Automation (ICCUBEA)*, vol. 2018, 2018, pp. 1-6. doi:10.1109/ICCUBEA.2018.8697439.
[56] M. A. Sarwar et al., "Prediction of diabetes using machine learning algorithms in healthcare" *24th International Conference on Automation and Computing (ICAC)*, vol. 2018, 2018. doi:10.23919/IConAC.2018.8748992.

[57] K. Vijiyakumar et al., “Random forest algorithm for the prediction of diabetes” *IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, vol. 2019, 2019, pp. 1-5. doi:10.1109/ICSCAN.2019.8878802.
[58] S. Tanioka et al., “Machine learning analysis of matricellular proteins and clinical variables for early prediction of delayed cerebral ischemia after aneurysmal subarachnoid hemorrhage”. Ml, pp. 9-12, 2019.
[59] L. Kopitar et al., “Early detection of type 2 diabetes mellitus using machine learning-based prediction models,” *Sci. Rep.*, vol. 10, no. 1, pp. 11981, 2020. doi:10.1038/s41598-020-68771-z.
[60] J. Omana and M. Moorthi, “Prediction of diabetes mellitus using measure of insulin resistance: A combined classifier approach,” vol. 12, no. 11, pp. 4793-4801, 2021.
[61] S. Borah, *Soft Computing Techniques and Applications (Issue Ic3)*, 2020.
[62] Q. Xu et al., “A systematic literature review of predicting diabetic retinopathy, nephropathy and neuropathy in patients with type 1 diabetes using machine learning”. Ml, *J. Med. Artif. Intell.*, vol. 3, 6-6. doi:10.21037/jmai.2019.10.04.
[63] O. AlShorman et al., “A review of wearable sensors-based monitoring with daily physical activity to manage type 2 diabetes,”, *IJECE*, vol. 11, no. 1. doi:10.11591/ijece.v11i1.pp646-653.
[64] P. Ghosh et al., & Detecting, “A Comparative Study of Different Machine Learning Tools in Detecting Diabetes” *Procedia Comput. Sci.*, vol. 192, pp. 467-477, 2021. doi:10.1016/j.procs.2021.08.048.
[65] A. Choudhury and D. Gupta (N.D.), “A Survey on Medical Diagnosis of Diabetes Using Machine Learning”. Singapore: *Springer*. doi:10.1007/978-981-13-1280-9.
[66] S. Shafi et al., “Early Prediction of Diabetes Disease & Classification of Algorithms Using Machine Learning Approach.”, Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021), Available at SSRN: https://ssrn.com/abstract=3852590 or http://dx.doi.org/10.2139/ssrn.3852590 May 25, 2021.
[67] H. Thakkar et al., “Clinical ehealth Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis. Clinical ehealth,”, *Clinical eHealth*, vol. 4, pp. 12-23, 2021. doi:10.1016/j.ceh.2020.11.001.
[68] T Shafi Zargar, O., Baghat, A. & Teli, T. A. *A DNN Model for Diabetes Mellitus Prediction on PIMA Dataset*.https://infocomp.dcc.ufla.br/index.php/infocomp/article/view/2476 (2022).
[69] H. Naz and S. Ahuja, “Deep learning approach for diabetes prediction using PIMA Indian dataset Deep learning approach for diabetes prediction using PIMA Indian dataset,” no. April, 2020, doi: 10.1007/s40200-020-00520-5.
[70] Thaiyalnayaki, K. Classification of diabetes using deep learning and svm techniques. *Int J Curr Res Rev* **13**, 146–149 (2021).
[71] K. Thaiyalnayaki, “Classification of Diabetes Using Deep Learning and SVM Techniques,” vol. 13, no. 01, 2021.
[72] S. I. Ayon and M. Islam, “Diabetes Prediction : A Deep Learning Approach,” no. March, pp. 21–27, 2019, doi: 10.5815/ijieeb.2019.02.03.
[73] Bharath, P., Chowdary, K. & Udaya Kumar, R. An Effective Approach for Detecting Diabetes using Deep Learning Techniques based on Convolutional LSTM Networks. IJACSA)

International Journal of Advanced Computer Science and Applications vol. 12 www.ijacsa.thesai.org.

[74] Hounguè, P. & Bigirimana, A. G. Leveraging Pima Dataset to Diabetes Prediction: Case Study of Deep Neural Network. Journal of Computer and Communications 10, 15–28 (2022).