Diabetes Detection and Forecasting using Machine Learning Approaches: Current State-of-the-art

Alwalid Alhashem [1], Aiman Abdulbaset [1], Faisal Almudarra [1], Hazzaa Alshareef [1], Mshari Alqasoumi [1], Atta-ur Rahman [1,*] and Maqsood Mahmud [2]

[1]Department of Computer Science (CS), College of Computer Science and Information Technology (CCSIT),
Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia
[2]Business Analytic Program, Department of Management and Marketing, College of Business Administration (CBA),
University of Bahrain, Sakhir 32038, Bahrain

Abstract
The emergence of COVID-19 virus has shaken almost every aspect of human life including but not limited to social, financial, and economic changes. One of the most significant impacts was obviously healthcare. Now though the pandemic has been over, its aftereffects are still there. Among them, a prominent one is people lifestyle. Work from home, enhanced screen time, limited mobility and walking habits, junk food, lack of sleep etc. are several factors that have still been affecting human health. Consequently, diseases like diabetes, high blood pressure, anxiety etc. have been emerging at a speed never witnessed before and it mainly includes the people at young age. The situation demands an early prediction, detection, and warning system to alert the people at risk. AI and Machine learning has been investigated tremendously for solving the problems in almost every aspect of human life, especially healthcare and results are promising. This study focuses on reviewing the machine learning based approaches conducted in detection and prediction of diabetes especially during and post pandemic era. That will help find a research gap and significance of the study especially for the researchers and scholars in the same field.

Keywords:
AI, Machine Learning, Deep Learning, Diabetes detection. Prediction

1. Introduction

In an era where data and technology hold unprecedented potential to transform lives, the intersection of healthcare and/or telemedicine and computational intelligence techniques such as machine learning offers a fertile ground for innovation [1,2,3]. This study resides at that nexus, aiming to provide an insight of the accuracy, timeliness, and effectiveness of diabetes detection and forecasting approaches investigated in the literature especially in the post COVID-19 period [4,5,6].

Such studies are designed to serve not just as a technological solution but as an embodiment of the collaborative strength of healthcare professionals and computer scientists. The importance of this interface is magnified in the face of the current global diabetes crisis, especially concerning Type 2 Diabetes, which is increasingly prevalent and debilitating. While the healthcare sector has seen advancements in technologies like Electronic Health Records (EHRs) and telemedicine [7,8], the focus is often on treatment rather than predictive analytics for preventive care [9,10].

Many people find out if they have diabetes when it's too late. Because of this, early diagnosis of Diabetes Type 2 can result in prompt intervention, which can stop or delay the onset of the disease's consequences. Cardiovascular conditions, neuropathies, and other complications can be fatal and drastically lower quality of life. We can give people the knowledge and resources they need to make lifestyle adjustments that may help them avoid developing Diabetes Type 2 by identifying when the condition will start to manifest [11,12]. In this regard, several approaches have been developed including some clinical decision support systems (CDSS). That are mainly used to provide assistive technology to the practitioners and medical professionals in a variety of ways [13,14,15]. Nonetheless, some studies are purely meant for theoretical investigations to develop machine learning and deep learning models to predict the disease [16,17,18,19,20].

This paper is organized as follows. The next section describes the background of the study.

Section 3 presents the extensive review of the related literature. In section 4, we discuss the results of the literature survey. Section 5 summarizes the results and gives some recommendations and provides some useful concluding remarks.

Manuscript received October 5, 2023
Manuscript revised October 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.10.24

2. Background

In this section we will explore two pillars of the current study: Diabetes Type 2 and Machine Learning. We will dive in an explanation of Diabetes Type 2 background, conditions, the complications, and symptoms. Additionally, we will introduce Gestational Diabetes, a variant that affects both the mother and child during pregnancy. Then we will go through some of the concepts of Machine Learning (ML) and the challenges of predicting the onset of Type 2 Diabetes. The results of the survey would be important in comparing these techniques to the new technique of mixed teaching strategies.

2.1 Diabetes:

Diabetes Type 2 is a chronic disease that affects how the body metabolizes glucose (sugar), often known as non-insulin dependent. In Type 2 diabetes, as opposed to Type 1 diabetes, when the body does not create insulin, the body either rejects the effects of insulin or does not produce enough insulin to keep blood glucose levels within normal range. This results in elevated blood sugar levels, which can damage organs, blood vessels, and nerves over time. The prevalence of Type 2 Diabetes has been steadily increasing on a global scale especially post pandemic period with diverse reasons. Both types are depicted in Figure 1.

Over 400 million individuals are thought to be affected by this illness worldwide, according to the World Health Organization (WHO), making it a serious public health issue. Urbanization, sedentary habits, bad diets, and an increase in life expectancy are a few causes of the surge. The development of Type 2 Diabetes has been linked to several risk factors. These include specific ethnic backgrounds, old age, family history of diabetes, inactivity, poor diet, and obesity. Additionally, diseases including polycystic ovarian syndrome (PCOS) and gestational diabetes can raise the risk.

2.1.1 Symptoms:

The beginning of Diabetes Type 2 is often gradual, and its symptoms can be subtle at first and may develop with various speed depending on the person. Common symptoms include:

- High thirst
- Frequent urination
- Weight loss
- Fatigue
- Blurred vision
- Darkened skin areas, known as acanthosis nigricans.

2.1.2 Complications:

Diabetes Type 2 can have a variety of problems that can damage practically all the body's major organs like lungs, kidneys etc. if it is not treated properly. These issues can include:

- Cardiovascular disease
- Neuropathy
- Nephropathy
- Retinopathy
- Foot damage
- Skin and mouth conditions
- Hearing impairment
- Alzheimer’s disease

There is also another type of diabetes called Gestational diabetes which is a type that can develop during pregnancy in women, who most of the time does not have diabetes. And could cause problems to the baby if blood sugar levels are untreated or unchecked. Moreover, there could be complications in the fetus/baby caused by gestational diabetes, including:

- Excess growth
- Low blood sugar
- Type 2 diabetes later in life
- Death

Not to mention the complications in the mother, which include:

- Preeclampsia
- Gestational diabetes

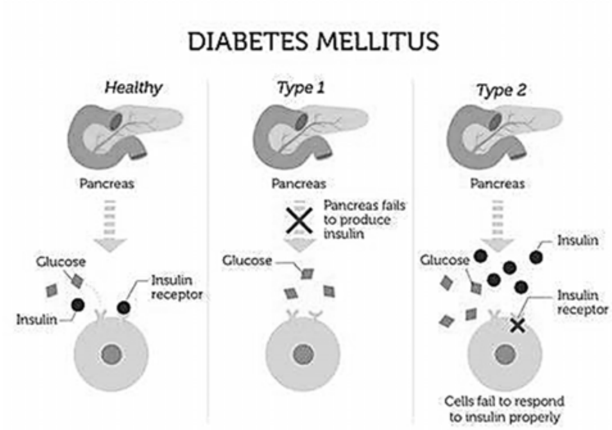

Figure 1: Diabetes Type I & II

2.2 Machine Learning:

A branch of artificial intelligence known as machine learning (ML) concentrates on creating algorithms that let computers learn from data and make decisions based on that data. These systems are trained utilizing a lot of data, which gives them the capacity to learn how to accomplish the task, as opposed to being explicitly programmed to do so. Machine learning (ML) has recently made considerable strides in the healthcare industry [21-22].

It can be difficult to forecast the start of diabetes type 2 using conventional statistical approaches because the disease is complex. A more complex approach is provided by machine learning, which enables the incorporation of a wide range of data, including genetic markers, lifestyle characteristics, and even data from wearable devices. Early research in this field has demonstrated that ML models can perform better than conventional risk assessment methods in foretelling the onset of Type 2 Diabetes [23].

3. State-of-the-art

In their study Rehman et al. (2020) [23] proposed the modelling, simulation and optimization approach to diabetes type II prediction using deep extreme learning machine (DELM) approach. In the study, they stated that, “In the human body, glucose is the main element that boosts cells. However, insulin is a key that enters the cells to control blood sugar. People with diabetes type I do not have the ability to produce insulin. Whereas people with diabetes type II lack the ability to react to insulin and frequently do not make enough insulin. For adequate analysis of such a fatal disease, techniques with a minimum error rate must be utilized. By using the DELM approach, a high level of reliability with a minimum error rate is achieved. The approach shows significant improvement in results compared to previous studies. In this regard, a refined dataset of 1400 samples have been utilized after due preprocessing. It is observed that during the investigation the proposed approach has the highest accuracy rate of 92.8% with 70% of training (9500 samples) and 30% of test and validation (4500 examples) [23].

In research by Tigga and Garg [24], authors have investigated prediction of Type 2 Diabetes using Machine Learning classification methods. The study focuses on using machine learning algorithms for early diagnosis and treatment of Type 2 diabetes in India. The study collected data from 952 participants through an online and offline questionnaire that included 18 questions related to health, lifestyle, and family background. Various machine learning algorithms such as Random Forest Classifier were employed, and the paper claims these algorithms to be highly accurate for this healthcare application. In the literature review, the authors discuss the extensive application of machine learning, data mining, neural networks, and genetic algorithms in diabetes prediction, citing several studies that have used different datasets and methods, each with their own levels of accuracy. The Random Forest Classifier was found to be the most accurate for both the collected dataset and the Pima Indian Diabetes database [24].

Authors in [25], in their research proposed prediction of Diabetes in healthful inhabitants through Machine Learning. The research focuses on employing machine learning techniques to predict the future risk of type-2 diabetes. The study revisits data from the San Antonio Heart Study and uses Support Vector Machines (SVM) along with ten features known to be strong predictors of future diabetes. The dataset used is unbalanced, with a higher number of healthy individuals compared to those with diabetes. To address this, the authors use 10-fold cross-validation for training and a hold-out set for validation. The results show a validation accuracy of 84.1% with a recall rate of 81.1%, averaging over 100 iterations. The best method used in this study is the Support Vector Machines (SVM) model. The accuracy of the SVM model is 84.1% [25].

The research paper in [26] emphasized on building risk prediction models for Diabetes Type 2 using Machine Learning techniques. The study used data from the 2014 Behavioral Risk Factor Surveillance System, which included 138,146 participants. Several machine learning models were built, including Support Vector Machine (SVM), Decision Tree, Logistic Regression, Random Forest, Neural Network, and Gaussian Naive Bayes. The models were evaluated based on accuracy, sensitivity, specificity, and Area Under the Curve (AUC). The Neural Network model had the highest accuracy at 82.4%, specificity at 90.2%, and AUC at 0.7949. However, the Decision Tree model had the highest sensitivity at 51.6%. Study also identified two new potential risk factors: sleeping time and checkup frequency [26].

The research study by [27] conducted a comparative analysis of Machine-Learning Approaches Type 2 Diabetes predictions by cross-validation methods. The study focuses on the application of machine learning (ML) techniques for predicting Type 2 diabetes in Pima Indian females. The study used a Pima Indian diabetes dataset (PIDD) with 768 female patients and compared four different ML classifiers: Naïve Bayes (NB), J48, Logistic Regression (LR), and Random Forest (RF). These models were evaluated using different cross-validation techniques (K = 5, 10, 15, and 20) and performance metrics like accuracy, precision, F-score, recall, and AUC were calculated for each model. The study found that Logistic Regression (LR) had the highest accuracy of 0.77 across all 'k' values. The study aims to identify an optimized ML model for predicting diabetes and concluded that LR, RF, and NB were the best models for this purpose, with AUC values of 0.83, 0.82, and 0.81, respectively [27].

Another similar approach on diabetes detection using machine learning methods was conducted by [28]. It focuses on the application of machine learning techniques for the early diagnosis of Type-2 diabetes. The paper uses the Pima Indian Diabetes Dataset for its experiments and employs two machine learning algorithms: Logistic Regression and Random Forest Classifier. The study aims to compare the performance of these algorithms in predicting Type-2 diabetes. The paper concludes that Random Forest provides the highest accuracy for the PIMA dataset. Specifically, the accuracy was improved to 95% for Random Forest and 97% for Logistic Regression [28].

Han et al. (2018) [29] tackled the pressing issue of accurately predicting Type 2 Diabetes Mellitus (T2DM). A condition that has significant public health implications. While existing models and algorithms have been employed for this purpose, they often fall short in terms of prediction accuracy and adaptability to multiple datasets. Traditional methods have used single algorithms like K-means or logistic regression and have been limited by the datasets they were trained on. Utilizing the Waikato Environment for Knowledge Analysis (WEKA) toolkit and the Pima Indians Diabetes Dataset, the authors introduce a novel two-level algorithmic model that combines an improved K-means algorithm for initial data clustering with logistic regression for final classification. This multi-faceted approach achieves a remarkable accuracy rate of 95.42%. What makes their achievement great is the high accuracy they managed. Also, designing their model to adapt to more datasets is a plus for their record. However, the model may fall short due to the complexity, due to the multi-level algorithmic structure, could be a barrier to implementation. Also, the model's primary testing on the Pima Indians Diabetes Dataset raises questions about its applicability to other populations [29].

In 2018, a study by Deepti Sisodia and Dilip Singh Sisodia address the critical challenge of predicting diabetes, particularly in pregnant women a demographic that is especially vulnerable to this chronic condition [30]. The authors leverage the Waikato Environment for Knowledge Analysis (WEKA) toolkit and the Pima Indians Diabetes Database (PIDD) to evaluate the performance of three machine learning algorithms: Naive Bayes, Support Vector Machine (SVM), and Decision Tree. While existing studies have often focused on single algorithms, this research provides a comparative analysis, thereby offering a more comprehensive understanding of algorithmic effectiveness in diabetes prediction. Among the algorithms tested, Naive Bayes stands out with the highest accuracy of 76.30%, setting a notable benchmark in the field. The study's strength lies in its multi-algorithmic approach and its focus on a high-risk group, offering potentially valuable insights for healthcare providers. However, the research is not without limitations. The moderate accuracy levels indicate room for improvement, and the use of the PIDD dataset raises questions about the model's generalizability to broader populations [30].

A study by Swapna et al. (2018) [31] addresses the critical issue of early diabetes detection, focusing on a non-invasive methodology that leverages Heart Rate Variability (HRV) signals derived from Electrocardiograms (ECG). While traditional methods have often relied on single algorithms and limited datasets, this study employs a multi-algorithmic approach using deep learning architectures like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM), in conjunction with Support Vector Machine (SVM) for final classification. This comprehensive methodology achieves an impressive accuracy rate of 95.7% with the CNN 5-LSTM with SVM architecture. The study's strength lies in its high accuracy and non-invasive approach, offering a potentially transformative tool for healthcare providers. However, the research is not without limitations; the complexity of the deep learning architectures used could pose challenges for real-world implementation, and the relatively small dataset raises questions about the model's generalizability to broader populations [31]. This makes the research less suitable for real time environments.

A study conducted in 2021 by Joshi and Dhakal [32] tackles the critical issue of predicting type 2 diabetes, a condition with significant public health implications. They employ a combined approach of logistic regression and machine learning, specifically a classification tree, to analyze the Pima Indian dataset. Their model identifies five main predictors of type 2 diabetes: glucose, pregnancy, body mass index (BMI), diabetes pedigree function, and age. The 6th model of Logistic Regression achieved a prediction accuracy of 78.26%% and a cross-validation error rate of 21.74%. The model's strength lies in its reasonable prediction accuracy and its potential to complement existing preventive measures. The integration of Logistic Regression and Classification Tree methodologies allows for a more comprehensive analysis, leading to enhanced prediction accuracy. This combined approach leverages the strengths of both methodologies to provide a robust predictive model. However, its primary testing on a specific dataset could limit its applicability to broader populations. Authors aimed to address the issue in their subsequent work [32].

In another 2020 study by Leon Kopitar et al. [33], the GLMNet model was employed to predict T2DM. The model achieved impressive results in terms of the area under the curve (AUC), with an AUC of 0.859. One of the primary advantages of the GLMNet model is its superior performance in datasets with highly correlated and sparse predictor variables. This makes it particularly suitable for complex datasets where predictors might have interdependencies in certain ways [33].

In another 2021 study by Nikos Fazakis et al. [34], the Weighted Voting LRRFs model was utilized for T2DM prediction. The model achieved an AUC of 0.884, indicating a high level of accuracy. A significant advantage of the Weighted VotingLRRFs model is its ability to optimize multiple classification quality measures, such as AUC and Sensitivity, simultaneously. This ensures a well-rounded prediction model that considers various aspects of the data [34].

Research focusing on Type 2 Diabetes with Artificial Intelligence Machine Learning with the Methods and Evaluation has catalyzed a proliferation of studies employing machine learning algorithms for early detection and prognosis, given the imperative nature of timely intervention [35]. Rooted in varied datasets, numerous studies have delved into machine learning paradigms, ranging from tree-based to meta heuristic-based methods, aiming to optimize predictive accuracy based on diverse risk factors. However, the disparate datasets and evaluation metrics across these studies have obfuscated a clear comparative understanding of their effectiveness. The juxtaposition of these algorithms in extant literature underscores the imperative for a holistic, standardized assessment to discern the most efficacious models, with the prevailing emphasis being on balancing accuracy with real-world applicability, especially in the face of imbalanced healthcare datasets. The study compared 35 algorithms and highlighted the best performers, emphasizing the importance of accuracy and F-measure. Key predictors for diabetes were pinpointed. We utilized Weka 3.8 for algorithm evaluation across three datasets: PIMA Indian, UCI, and MIMIC III. The PIMA dataset predicts diabetes presence based on diagnostic metrics, the UCI dataset encompasses patient outcomes from 130 US hospitals (1999-2008), while MIMIC III details 40,000 patients from Berth Israel Deaconess Medical Center's critical care units between 2001 and 2012. Performance was assessed with and without feature selection [35].

The research paper [36] titled: "Prediction of Type 2 Diabetes Based on Machine Learning Algorithm "the increasing prevalence of diabetes, a chronic metabolic disorder, has elevated the need for early detection methods, with type 2 diabetes (T2D) being of particular concern due to its potential long-term organ damage. Electronic medical records have been recognized for their revolutionary potential in diagnostics, especially when integrated with machine learning (ML) and artificial intelligence (AI) techniques. Prior research has tapped into ML models for prediabetes screening in certain populations, while other studies have extended these techniques to predict treatment success in various chronic ailments. Furthermore, the American Diabetes Association (ADA) provides specific guidelines to identify T2D based on fasting plasma glucose levels, emphasizing the role of lifestyle factors in the disease's development. The feature selection process in ML, crucial for enhancing computational efficiency and model interpretability, has also been emphasized. Using health records from 2013-2018, a model was developed to anticipate Type 2 Diabetes in the upcoming year, with accuracies between 71% and 73%. This study utilized a six-year electronic medical record from Hanaro Medical Foundation in Seoul, South Korea, spanning 2013-2018. It comprises 535,169 records from 253,395 patients, with each record containing 1,444 attributes. For the experiment, the dataset includes patients who visited the institute between one to six times during the span of study period [36].

The research paper titled: "Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis [37]". There is a burgeoning interest in leveraging machine learning (ML) to predict Type 2 Diabetes Mellitus (T2DM), especially within community settings. Historically, non-ML methods dominated diabetes prediction, but no comprehensive review covers the newer ML-focused approaches. This gap is crucial as community-based predictions can identify early cases, even asymptomatic ones, facilitating timely interventions and potentially mitigating severe complications and hospitalizations. Moreover, traditional models have faced challenges in methodology, reporting, and validation which ML models aim to address. The systematic review delves into these ML models, aiming to discern their efficacy, strengths, and areas needing refinement (Kushan De Silva a, 2020) [37].

The integration of machine learning (ML) and deep learning (DL) in healthcare has shown promise in the early and accurate detection of Diabetes Mellitus (DM) [38]. This study delves into various algorithms and techniques, highlighting their effectiveness in DM detection. While these advancements underscore the transformative potential of ML and DL in medical diagnostics, challenges such as data scarcity, model construction complexities, and real-world deployment persist. However, with further research, more comprehensive datasets, and the exploration of hybrid models, the union of AI and healthcare offers a promising avenue for enhanced disease detection in the future (Toshita Sharma, 2021) [38].

In a study using the Pima Indians Diabetes database consisting of 768 samples, researchers compared the diabetes prediction capabilities of three techniques: Support Vector Machine (SVM), Random Forest (RF), and Convolutional Neural Network (CNN). Each method was tested on an equivalent set of training and testing samples, drawn from a dataset characterized by eight attributes related to diabetes risk. The results indicated that the Random Forest method, with an accuracy of 83.67%, was the most effective in predicting diabetes, outperforming the Support Vector Machine and Deep Learning techniques, which achieved accuracies of 65.38% and 76.81% respectively (Amani Yahyaoui, 2019) [39].

4. Analysis of State-of-the-art

This section is dedicated to the analysis of comprehensively reviewed literature. Several studies have been reviewed in this regard mainly in the post pandemic era e.g., year 2020 onward. Table 1 presents the summary of the literature review including the methodology and pros and cons of reviewed study.

Table 1: Summary of Literature Review

Study	Author/s	Methodology	Accuracy	Advantages	Limitation
1	Neha Prerna Tigga and Shruti Garg	Random Forest Classifier	The exact accuracy 0.941 for Dataset and 0.750 for Pima dataset	Extensive application in healthcare; high accuracy for both collected dataset and Pima Indian Diabetes database.	Limited to Indian population.
2	Lejla Alic et al.	Support Vector Machines (SVM)	84.1%	High validation accuracy of 84.1% and a recall rate of 81.1%; robust to unbalanced datasets.	Limited to San Antonio Heart Study data; may not generalize well.
3	Zidian Xie et al.	Multiple models including Neural Network	Neural Network: 82.4%	High specificity (90.2%) and AUC (0.7949); identified new potential risk factors like sleeping time and frequency of checkups.	Neural Network model had low sensitivity (37.8%); Decision Tree model had high sensitivity but lower accuracy.
4	Gopi Battineni et al.	Multiple models including Logistic Regression (LR)	LR: 0.77	High accuracy across all 'k' values in cross-validation; AUC of 0.83 for LR.	Limited to Pima Indian females; may not generalize well.
5	Valasapalli Mounika et al.	Random Forest and Logistic Regression	Random Forest: 95%, Logistic Regression: 97%	Extremely high accuracy; robust algorithms.	Limited to Pima Indian Diabetes Dataset; may not generalize well.
6	Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, and Xiaoyi Wang	Two-level algorithmic model consists of a combination of improved K-means algorithm and the Logistic regression algorithm	95.42%	High accuracy. Adaptability: the model able to adapt to more than one dataset.	Complexity: The model involves multiple steps and algorithms. Dataset Limitations: The model was primarily tested on the Pima Indians Diabetes Dataset.
7	Deepti Sisodia and Dilip Singh Sisodia	Naïve Bayes (NB), Decision Tree (DT), and Support Vector Machine (SVM)	NB: 76.30% DT: 73.82% SVM: 65.10%	Versatility: more than one machine learning algorithm has been used and worked. Focus on pregnant women.	Limited Dataset: The study uses the Pima Indians Diabetes Database (PIDD) dataset, which might not be representative of the broader population. Moderate Accuracy: Although Naïve Bayes achieved the highest accuracy, it is still below 80%, indicating room for improvement.

8	Swapna G., Vinayakumar R., and Soman K.P.	Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Support Vector Machine (SVM)	CNN 5 with SVM: 93.9% CNN 5-LSTM with SVM: 95.7%	High Accuracy: The CNN 5-LSTM with SVM architecture achieved a very high accuracy of 95.7% Non-Invasive Method: The study uses HRV signals for diabetes detection, which is a non-invasive approach.	Limited Dataset: The study was conducted on a relatively small dataset, which may affect its generalizability. Complexity: The deep learning architectures used are complex, which might be a barrier for real-world implementation.
9	Ram D. Joshi and Chandra K. Dhakal	logistic regression and classification tree	Logistic regression: 78.26% Classification Tree: 73.96%	It helps in identifying high-risk individuals	The model is only tested on Pima Indians Dataset.
10	Leon Kopitar et al.	Glmnet	In terms of the area under the curve (AUC), Glmnet achieved the best results with an AUC of 0.859	Glmnet performs better in datasets with highly correlated and sparse predictor variables	It can result in higher instability of the selected variables.
11	Ram D. Joshi, Chandra K. Dhakal	Logistic Regression, Model Selection Criteria, Validation and Cross-Validation, Classification Tree	Prediction accuracy of 78.26% and a cross-validation error rate of 21.74%	The combined approaches of Logistic regression and Classification Tree allows for more comprehensive analysis and better prediction accuracy.	Only a few predictors were considered due to data limitation.
12	Nikos Fazakis et al,	Weighted-VotingLRRFs	It resulted in AUC = 0.884	It optimizes more than one classification quality measures like AUC and Sensitivity simultaneously	Excluded family history of diabetes and women with gestational diabetes from the features set for training the ML models.
13	Leila Ismail	Bagging-LR	Our experimental results show that algorithm is the most accurate for a balanced dataset with and without feature selection	Performs well on balanced datasets.	requires multiple models to be stored, which can be space-consuming
14	Henock M. Deberneh and Intaek Kim	RF (Random Forest)	The best accuracy achieved for predicting the occurrence of diabetes was 73% on the test dataset	Utilization of diverse machine learning models and ensemble classifiers.	The majority of the classification errors were from the prediabetes class.
15	Kushan De Silva	ML-based	prediction models for T2DM in community settings achieved "good discrimination ability"	Good discrimination ability of ML-based prediction models for T2DM in community settings.	Need for improvement in methodology and reporting.

16	Toshita Sharma, Manan Shah	- Literature Review - Case Studies Analysis - Comparative Analysis - Challenges Identification - Future Recommendations		- Comprehensive Review - Detailed Case Studies - Challenges Highlighted - Future Recommendations	- Over-reliance on PIMA Dataset
17	Amani Yahyaoui, Akhtar Jamil, Jawad Rasheed, Mirsat Yesiltepe	- Support Vector Machine (SVM) - Random Forest (RF) - Convolutional Neural Network (CNN)	- RF: 83.67% - DL: 76.81% - SVM: 65.38%	- Comprehensive Comparison - Use of Renowned Dataset - Robust Evaluation Metrics - Cross-validation - Detailed Attribute Information - Previous Work Review	- Limited to One Dataset - Discrepancy in Reported Accuracies - Depth of Analysis

5. Conclusion

In the rapidly evolving landscape of medical technology, understanding the existing state of play is crucial for innovation. As part of this study, a comprehensive review of existing studies in diabetes detection using machine learning has been conducted, including an examination of their technical documentation, user reviews, and related scholarly articles. This multi-faceted evaluation led to the identification of critical gaps in current solutions, revealing both missing features and areas for improvement in usability and user experience. While some gaps may have more immediate implications for healthcare outcomes, others serve as opportunities for differentiation and value addition. In addition, the fluid nature of medical technology demands a forward-looking approach. Based on the research, in future, authors are intended to propose a system that aims not only to bridge these identified gaps but is designed with the flexibility to adapt to future technological advancements and clinical needs. Moreover, in the future, hybrid intelligent technologies and transfer learning approaches can be investigated to address the highlighted challenges [40-50].

References

[1] A. Rahman, M.I.B. Ahmed, “Virtual Clinic: A CDSS Assisted Telemedicine Framework,” Telemedicine Technologies, 1, 250, Chapter 15: pp. 227-238, 2019.

[2] Atta-ur-Rahman, M. H. Salam and S. Jamil, "Virtual clinic: A telemedicine proposal for remote areas of Pakistan," 2013 Third World Congress on Information and Communication Technologies (WICT’ 13), Hanoi, Vietnam, 2013, pp. 46-50.

[3] Atta-ur-Rahman et al. (2019). A Comprehensive Study of Mobile Computing in Telemedicine. In Advanced Informatics for Computing Research. ICAICR 2018. Communications in Computer and Information Science, vol 956. Springer, Singapore.

[4] Ahmed, M.S.; Rahman, A.; AlGhamdi, F.; AlDakheel, S.; Hakami, H.; AlJumah, A.; AlIbrahim, Z.; Youldash, M.; Alam Khan, M.A.; Basheer Ahmed, M.I. Joint Diagnosis of Pneumonia, COVID-19, and Tuberculosis from Chest X-ray Images: A Deep Learning Approach. Diagnostics 2023, 13, 2562. https://doi.org/10.3390/diagnostics13152562.

[5] Alqarni, A.; Rahman, A. Arabic Tweets-Based Sentiment Analysis to Investigate the Impact of COVID-19 in KSA: A Deep Learning Approach. Big Data Cogn. Comput. 2023, 7, 16. https://doi.org/10.3390/bdcc7010016.

[6] A Rahman, K Sultan, I Naseer, R Majeed, D Musleh et al., “Supervised machine learning-based prediction of COVID-19,” Computers, Materials and Continua 69 (1), 21-34, 2021.

[7] R Zagrouba, MA Khan, A Rahman, MA Saleem et al., “Modelling and Simulation of COVID-19 Outbreak Prediction Using Supervised Machine Learning,” Computers, Materials & Continua 66 (3), 2397-2407, 2021.

[8] RA Naqvi, MF Mushtaq, NA Mian, MA Khan et al., “Coronavirus: A "Mild" Virus Turned Deadly Infection,” Computers, Materials & Continua 67 (2), 2631-2646, 2021.

[9] N Aldowesh, A Alfaleh, M Alhejazi, H Baghdadi, A Rahman, “Electronic Data Interchange Framework for Financial Management System,” IJCSNS, 22(6), pp. 275-287, 2022.

[10] M. Mukhtar, F. Yunus, J. Li, Atta-ur-Rahman, T. Mahmood and Y. A. A. Ali, "Future Prospects and Challenges of On-Demand Mobility Management Solutions," in IEEE Access,

[11] Ahmed, M.I.B.; Saraireh, L.; Rahman, A.; Al-Qarawi, S.; Mhran, A.; Al-Jalaoud, J.; Al-Mudaifer, D.; Al-Haidar, F.; AlKhulaifi, D.; Youldash, M.; et al. Personal Protective Equipment Detection: A Deep-Learning-Based Sustainable Approach. Sustainability 2023, 15, 13990.

[12] RA Qamar, M Sarfraz, A Rahman, SA Ghauri, “Multi-Criterion Multi-UAV Task Allocation under Dynamic Conditions,” Journal of King Saud University-Computer and Information Sciences 35 (9), 101734, 2023.

[13] Ahmed, M.I.B.; Alotaibi, R.B.; Al-Qahtani, R.A.; Al-Qahtani, R.S.; Al-Hetela, S.S.; Al-Matar, K.A.; Al-Saqer, N.K.; Rahman, A.; Saraireh, L.; Youldash, M.; et al. Deep Learning Approach to Recyclable Products Classification: Towards Sustainable Waste Management. Sustainability 2023, 15, 11138. https://doi.org/10.3390/su151411138.
[14] Ibrahim, N.M.; Gabr, D.G.; Rahman, A.; Musleh, D.; AlKhulaifi, D.; AlKharraa, M. Transfer Learning Approach to Seed Taxonomy: A Wild Plant Case Study. Big Data Cogn. Comput. 2023, 7, 128. https://doi.org/10.3390/bdcc7030128.
[15] A Albassam, F Almutairi, N Majoun, R Althukair, et al., "Integration of Blockchain and Cloud Computing in Telemedicine and Healthcare," IJCSNS, 23 (6), 17-26, 2023.
[16] Sajid, N.A.; Rahman, A.; Ahmad, M.; Musleh, D.; Basheer Ahmed, M.I.; Alassaf, R.; Chabani, S.; Ahmed, M.S.; Salam, A.A.; AlKhulaifi, D. Single vs. Multi-Label: The Issues, Challenges and Insights of Contemporary Classification Schemes. Appl. Sci. 2023, 13, 6804.
[17] T. A. Khan et al., "Secure IoMT for Disease Prediction Empowered with Transfer Learning in Healthcare 5.0, the Concept and Case Study," in IEEE Access, vol. 11, pp. 39418-39430, 2023, doi: 10.1109/ACCESS.2023.3266156.
[18] Alghamdi, A.S.; Rahman, A. Data Mining Approach to Predict Success of Secondary School Students: A Saudi Arabian Case Study. Educ. Sci. 2023, 13, 293.
[19] NA Sajid, M Ahmad, A Rahman, G Zaman, MS Ahmed et al., "A Novel Metadata Based Multi-Label Document Classification Technique," Computer Systems Science and Engineering 46 (2), 2195-2214, 2023.
[20] Rahman, A. GRBF-NN based ambient aware realtime adaptive communication in DVB-S2. J Ambient Intell Human Comput 14, 5929–5939 (2023).
[21] MA Qureshi, M Asif, S Anwar, U Shaukat, MA Khan, A Mosavi, "Aspect Level Songs Rating Based Upon Reviews in English," Computers, Materials & Continua 74 (2), 2023.
[22] S Abbas, SA Raza, MA Khan, A Rahman, K Sultan, A Mosavi, "Automated File Labeling for Heterogeneous Files Organization Using Machine Learning," Computers, Materials & Continua 74 (2), 3263-3278, 2023.
[23] A Rehman, A. Athar, MA Khan, S Abbas, A Fatima, A Rahman, A Saeed, "Modelling, simulation, and optimization of diabetes type II prediction using deep extreme learning machine," Journal of Ambient Intelligence and Smart Environments 12 (2), 125-138, 2020.
[24] Neha Prerna Tigga, S. G. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. Retrieved from https://doi.org/10.1016/j.procs.2020.03.336.
[25] Hasan Abbas, L. A.-G. (2019). Predicting Diabetes in Healthy Population through Machine Learning. Retrieved from https://ieeexplore.ieee.org/abstract/document/8787404.
[26] Zidian Xie, O. N. (2019). Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6795062/pdf/PCD-16-E130.pdf
[27] Gopi Battineni, G. G. (2019). Comparative Machine-Learning Approach: A Follow-Up Study on Type 2 Diabetes Predictions by Cross-Validation Methods. Retrieved from
[28] C. Charitha, A. Devi Chaitrasree, P. C. Varma and C. Lakshmi, "Type-II Diabetes Prediction Using Machine Learning Algorithms," 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp. 1-5.
[29] Han Wu, S. Y. (2018). Type 2 diabetes mellitus prediction model based on data mining. Retrieved from
[30] D. Sisodia, D. S. Sisodia (2018). Prediction of Diabetes using Classification Algorithms. Procedia Computer Science, 132, pp. 1578-1585, 2018.
[31] Swapna G., Vinayakumar R., Soman K.P. (2018). Diabetes detection using deep learning algorithms. ICT Express, 4 (4), 243-246, 2018. https://doi.org/10.1016/j.icte.2018.10.005.
[32] Joshi, R.D.; Dhakal, C.K. Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. Int. J. Environ. Res. Public Health 2021, 18, 7346.
[33] Kopitar, L., Kocbek, P., Cilar, L. et al. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Sci Rep 10, 11981 (2020).
[34] N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis and K. Moustakas, "Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction," in IEEE Access, vol. 9, pp. 103737-103757, 2021.
[35] Ismail, L., Materwala, H., Tayefi, M. et al. Correction to: Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation. Arch Computat Methods Eng 28, 5039 (2021
[36] Deberneh, H.M.; Kim, I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. Int. J. Environ. Res. Public Health 2021, 18, 3317.
[37] Silva K, Lee WK, Forbes A, Demmer RT, Barton C, Enticott J. Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis. Int J Med Inform. 2020 Nov; 143:104268.
[38] Sharma T, Shah M. A comprehensive review of machine learning techniques on diabetes detection. Vis Comput Ind Biomed Art. 2021 Dec 3;4(1):30. doi: 10.1186/s42492-021-00097-7. PMID: 34862560; PMCID: PMC8642577.
[39] A. Yahyaoui, A. Jamil, J. Rasheed and M. Yesiltepe, "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques," 2019 1st International Informatics and Software Engineering Conference (UBMYK), Ankara, Turkey, 2019, pp. 1-4.
[40] Atta-ur-Rahman, Dash, S., Luhach, A.K. et al. A Neuro-fuzzy approach for user behaviour classification and prediction. J Cloud Comp 8, 17 (2019). https://doi.org/10.1186/s13677-019-0144-9.
[41] M Mahmud, A Rahman, M Lee, JY Choi, "Evolutionary-based image encryption using RNA codons truth table," Optics & Laser Technology 121, 105818, 2020.
[42] A. Rahman, "Memetic computing based numerical solution to Troesch problem," Journal of Intelligent and Fuzzy Systems 36 (6), 1-10, 2019.
[43] A. Rahman, "Optimum information embedding in digital watermarking," Journal of Intelligent & Fuzzy Systems 37 (1), 553-564, 2019.
[44] Atta-ur-Rahman et al. (2019). A Comprehensive Study of Mobile Computing in Telemedicine. In: Luhach, A., Singh, D., Hsiung, PA., Hawari, K., Lingras, P., Singh, P. (eds) Advanced Informatics for Computing Research. ICAICR 2018. Communications in Computer and Information Science, vol 956. Springer, Singapore. https://doi.org/10.1007/978-981-13-3143-5_34.

[45] S Dash, S BISWA, D BANERJEE, A Rahman, “EDGE AND FOG COMPUTING IN HEALTHCARE – A REVIEW,” Scalable Computing: Practice and Experience 20 (2), 191-205, 2019.
[46] Azam, M., Atta-ur-Rahman, Sultan, K., Dash, S., Khan, S.N., Khan, M.A.A. (2019). Automated Testcase Generation and Prioritization Using GA and FRBS. In: Luhach, A., Singh, D., Hsiung, PA., Hawari, K., Lingras, P., Singh, P. (eds) Advanced Informatics for Computing Research. ICAICR 2018. Communications in Computer and Information Science, vol 955. Springer, Singapore. https://doi.org/10.1007/978-981-13-3140-4_52.
[47] Alhiyafi, J., Atta-ur-Rahman, Alhaidari, F.A., Khan, M.A. (2019). Automatic Text Categorization Using Fuzzy Semantic Network. In: Benavente-Peces, C., Slama, S., Zafar, B. (eds) Proceedings of the 1st International Conference on Smart Innovation, Ergonomics and Applied Human Factors (SEAHF). SEAHF 2019. Smart Innovation, Systems and Technologies, vol 150. Springer, Cham. https://doi.org/10.1007/978-3-030-22964-1_3.
[48] Alhaidari, F.A., Atta-ur-Rahman, Alghamdi, A., Dash, S. (2019). Motion Detection in Digital Video Recording Format with Static Background. In: Benavente-Peces, C., Slama, S., Zafar, B. (eds) Proceedings of the 1st International Conference on Smart Innovation, Ergonomics and Applied Human Factors (SEAHF). SEAHF 2019. Smart Innovation, Systems and Technologies, vol 150. Springer, Cham. https://doi.org/10.1007/978-3-030-22964-1_2.
[49] M Zaheer, IM Qureshi, K Sultan, A Rahman, MZ Muzaffar, R Alnanih, “High Capacity Image Steganography Based on Prime Series Representation and Payload Redundancy Removal,” Journal of Information Assurance and Security 14 (2), 40-47, 2019.
[50] M Ahmad, U Farooq, A Rahman, A Alqatari, S Dash, A Luhach, “Investigating TYPE constraint for frequent pattern mining,” Journal of Discrete Mathematical Sciences and Cryptography 22 (4), 605-626, 2019.