Journal of Robotics and Control (JRC)
Volume 6, Issue 1, 2025
ISSN: 2715-5072, DOI: 10.18196/jrc.v6i1.22258

Advancements in Artificial Intelligence Techniques for Diabetes Prediction: A Comprehensive Literature Review

Emad Majeed Hameed [1*], Hardik Joshi [2], Qusay Kanaan Kadhim [3]
[1, 2] Gujarat University, Ahmadabad, Gujarat, India
[1] Middle Technical University, Baghdad, Iraq
[3] Department of Computer Science, University of Diyala, Baqubah 32001, Diyala, Iraq
Email: [1] emadhameed@gujaratuniversity.ac.in, [2] hardikjoshi@gujaratuniversity.ac.in, [3] dr.qusay.kanaan@uodiyala.edu.iq
*Corresponding Author

***Abstract*—Diabetes mellitus (DM) is a chronic condition requiring lifelong management due to inadequate insulin secretion or inefficacy of insulin. Its global prevalence has led to extensive research focusing on diagnosis, prevention, and treatment. The developments in artificial intelligence (AI) have improved diabetes management and prediction. This paper provides a comprehensive review of the contributions of machine learning (ML) algorithms in predicting and classifying diabetes. The review examines research on artificial intelligence techniques used to predict diabetes over the past six years, intending to identify the latest innovations and trends in this field. This time frame reflects recent methodological advances and new applications that exemplify the current state of artificial intelligence in diabetes prediction. It covers dataset selection, preprocessing, AI algorithms application, and evaluation methodologies. The results of this review show that the most predominant methods used in diabetes prediction are Random Forest, Logistic Regression, Decision Trees, Support Vector Machine, Naïve Bayes, and K-Nearest Neighbors, each with distinct advantages and limitations. The review also shows through its examination that the highest accuracy provided by the hybrid approach was 99.4%, the ensemble approach (ada boost) was 98.8%, deep learning (DNN) was 98.04%, and traditional machine learning (decision tree_ ID3) was 99%. Most studies conducted for diabetes prediction trained the models on specific datasets, which makes their generalizability to diverse populations and healthcare settings limited. The future directions must address ensuring the robustness and generalizability of predictive models through comprehensive external validation across various populations, settings, and geographic areas.**

***Keywords*—Diabetes Prediction; AI Techniques; Machine Learning; Ensemble Learning; Deep Learning.**

I. INTRODUCTION

One of the most common chronic diseases in people is diabetes mellitus, which arises from either an insufficient insulin production or insufficient sensitivity of cells to the action of insulin. Diabetes mellitus is expressed as a group of metabolic disorders characterized and defined by hyperglycemia. Energy is a fundamental requirement for human bodily functions, derived primarily from the conversion of food into glucose. The role of insulin, secreted by the pancreas, is pivotal in facilitating the transportation of glucose into cells for energy production [1]. Diabetes manifests as chronic hyperglycemia, stemming from inadequate insulin production by the pancreas or the ineffective utilization of produced insulin [2]. Elevated blood glucose levels, stemming from inefficient glucose utilization, are linked to various health complications, including diabetes-related conditions such as cardiovascular disease, kidney disease, ocular issues, nerve complications, and vascular impairment [3], [4]. High blood sugar presents noticeable symptoms including frequent urination, excessive thirst (polydipsia), weight loss, blurred vision, and growth impairment.

Diabetes is generally classified into two primary types: Type 1 diabetes arises from pancreatic insufficiency in producing insulin and is often associated with autoimmune processes and genetic predispositions. Alternatively, Type 2 diabetes results from a combination of insulin sensitivity issues and a failure in the body's response to compensatory insulin production. The glucose content in ingested food directly influences blood sugar levels, while biological factors like digestion, circulation, insulin levels, and cell responsiveness contribute to the dynamic process of movement of glucose from the bloodstream into cells. These physiological factors vary among individuals and in different physiological states. Emotional states or changes in activity affecting circulation rates further impact blood sugar dynamics, complicating the estimation of blood glucose values.

The timely identification and diagnosis of the ailment rely significantly on the expertise and clinical acumen of the medical practitioner. The health sector generates a great deal of data on health services, but this data is not used effectively in undetected cases. Human decisions pose significant risks in the early detection of diseases due to their reliance on healthcare specialists' subjective observations and judgments, which may lack consistent accuracy [5]. Therefore, various advanced mechanisms and software-based programs are considered necessary for automatic diagnosis and early detection of diseases with better accuracy.

Advancements in computational methodologies, particularly Machine Learning (ML) and Artificial Intelligence (AI), have revolutionized early-stage diabetes identification and diagnosis, surpassing the abilities of humans [5]. ML, a branch of AI, learns from examples rather

than relying solely on programmed instructions, enabling the analysis and interpretation of extensive patient data stored in computer systems. The health sector has progressively utilized these ML techniques for disease diagnosis and prediction, including applications in diabetes research. Despite the growing research interest in diabetes prediction, there is still a need for a comprehensive evaluation of the various artificial intelligence techniques used in this field. To guide this evaluation, we ask the following research questions:

1. What are the latest developments in artificial intelligence approaches for diabetes prediction?
2. Is there a promising machine learning model suitable for all diabetes databases?

Current AI techniques for diagnosing diabetes have some limitations in their application and are not tested on different datasets or different populations, making them limited for use in the prediction process. This paper aims to address and summarize the literature concerned with machine learning techniques applied to predict diabetes and its associated challenges. This work will therefore be useful for better prediction of the disease and improved understanding of the diabetes pattern, making it useful for treatment and reducing the risk of other complications of diabetes. The contribution of this study is to comprehensively assess the diverse applications of machine learning methods in diabetes prediction and classification through an extensive literature review.

II. BACKGROUND

Diabetes Mellitus (DM) encompasses a cluster of conditions affecting the body's energy metabolism processes following food consumption. Upon food ingestion, the body converts it into glucose, a vital energy source transported via the bloodstream. Insulin, a pancreatic hormone, aids in the movement of glucose from the bloodstream into cells, essential for generating energy [6]. Disruption in the body's glucose management mechanisms characterizes diabetes, leading to elevated blood sugar levels (hyperglycemia), precipitating severe health complications like diabetic retinopathy, nephropathy, and neuropathy [7]. The interplay between insulin and glucagon, hormones respectively secreted by Beta and Alpha cells in the islets of Langerhans, regulates plasma glucose levels. Normal bodily functions maintain glucose levels within a defined range. In individuals without diabetes, high blood glucose prompts insulin release, enabling glucose absorption by target cells. Conversely, in low-glucose scenarios, glucagon triggers the conversion of glycogen into glucose. However, in diabetic individuals, this coordinated system malfunctions, resulting in persistent high blood glucose levels (hyperglycemia), alongside concerns about hypoglycemia [8].

Diabetes exists in two primary forms based on pathophysiology: Type I and Type II. Type I diabetes involves the destruction of Beta cells, causing rapid and near-total insulin deficiency [9], [10]. Contrastingly, Type II diabetes, a chronic metabolic disorder, progresses over time. Individuals with Type II diabetes maintain some capacity for insulin production, but their bodies exhibit reduced efficiency in utilizing insulin. Over time, beta cell function diminishes, contributing to a gradual decline in insulin production [11]. Unlike Type I diabetes, where severe insulin deficiency occurs, Type II diabetes patients can often be managed effectively for prolonged periods through lifestyle modifications or oral medications.

A. Fundamentals of Intelligent Techniques

Machine Learning (ML) and Artificial Intelligence (AI) represent the scientific inquiry into computational systems' ability to acquire knowledge from experiences. Scholars often perceive "ML" as a subset within the broader field of "AI," positing that the capacity to learn embodies a fundamental trait of human cognition. Machine learning endeavors to fashion computer systems capable of learning from prior observations and subsequently adapting their responses. The overarching aim of artificial intelligence involves crafting intelligent agents or aides that harness diverse machine-learning methodologies [12].

Machine learning, a subset entrenched within artificial intelligence, explores data patterns through varied techniques and methodologies. Initially, it assimilates knowledge by scrutinizing labeled data associated with these patterns, thus facilitating the construction of systems equipped to make inferences based on past experiences. This potential is enabled by a spectrum of algorithms employing diverse mathematical and statistical approaches. Through analysis of instances within the dataset, these algorithms acquire the capacity to generalize solutions by comprehending the underlying patterns [13].

Artificial intelligence encompasses the scientific and engineering disciplines directed at fabricating intelligent machines capable of achieving human-like objectives. Machine learning stands as a form of artificial learning that seeks to emulate human learning processes. Within the realm of neural networks—particularly within deep learning—areas characterized by networks boasting more than three layers, denoting multiple hidden layers, constitute the field of deep learning (DL). DL represents a contemporary subfield of machine learning reliant on computationally intensive methodologies and extensive datasets to discern intricate relationships within data. In recent years, the multifaceted applications of machine learning and artificial intelligence have garnered global attention.

Numerous classifications exist for machine learning techniques. A prevalent categorization delineates them into three primary types: supervised, unsupervised, and reinforcement learning.

- Supervised learning entails the creation of a function from labeled training data. In this paradigm, the training dataset comprises input-output pairs, where the input constitutes a feature vector, and the output represents the desired function outcome. The output is an estimation or classification label. Supervised learning branches into two categories based on objectives: Classification, where targets exhibit similarities or correlations, and Regression, designed for discerning relationships between quantitative variables [14], [15].

- Unsupervised learning involves algorithms devoid of labeled data, tasked solely with discovering inherent structures within datasets. These algorithms operate on unlabeled, unclassified, or ungrouped data. Fundamental unsupervised techniques include dimensionality reduction methods like PCA and t-SNE, used for data preprocessing and visualization, respectively. A more advanced subset involves clustering algorithms that uncover latent patterns within data, such as K-means clustering, Gaussian mixture models, and latent Markov models [16].

- Reinforcement learning, inspired by behaviorism, focuses on actions necessary to maximize rewards within an environment. This versatile approach finds applications in diverse fields like game theory, control theory, and statistics. Unlike supervised learning, reinforcement learning lacks explicit input/output matches and relies on internal correction for non-optimal actions [17].

Within the medical domain, machine learning algorithms have gained prominence for disease prediction and diagnosis. Researchers extensively employ these approaches, particularly in predicting and categorizing diabetes, aiming for the most accurate and dependable prognostications. The subsequent section provides a literature review highlighting intelligent techniques utilized in the classification and prediction of DM.

III. MATERIALS AND METHODS

A literature search of various academic databases was performed to find pertinent research articles related to enhancing and supporting diabetes prediction using machine learning and deep learning techniques.

A. Search Strategy

The search queries considered the title, abstract, and keywords sections. The search criteria included four keywords: 'Diabetes', 'Prediction', 'Machine Learning', and 'Deep Learning', combined using the AND, OR operators. Various databases such as PubMed/MEDLINE, IEEE Xplore, SpringerLink, Elsevier ScienceDirect, IOP Science, and BMC (BioMed Central) were queried for articles published from 2018 onwards. PubMed was chosen because it contains a large biomedical literature, and this allows us to include clinical and experimental studies. As for the rest of the other databases, such as Scopus, IEEE, and Springer, they were chosen to index and provide extensive scientific literature in various disciplines. This time frame reflects recent methodological advances and new applications that exemplify the current state of artificial intelligence in diabetes prediction. However, the effectiveness of using 'DiabetesType' as a search keyword was limited, as many research articles did not specify the type of diabetes in their titles or abstracts. The considered studies vary as 85 Journal Articles, 13 Conference Papers, 4 Books, and 1 Dissertation. Table I illustrates exemplars of search strings intended for querying electronic repositories of scientific publications.

B. Eligibility Criteria

The criteria used to determine which articles were eligible for selection are outlined below:

1. Only papers composed in the English language.
2. Due to rapid technological advancements in diabetes prediction, only articles published within the past six years were considered.
3. Only papers that focus on diabetes prediction and classification.
4. Any paper addressing the prediction of diabetes, including type 1, type 2, or gestational diabetes.
5. Only articles focused on the topic utilizing machine learning and deep learning methodologies will be considered.

TABLE I. SEARCH STRATEGIES FOR THE SELECTED DATABASES

Database	Search Query
PubMed/MEDLINE	(("Diabetes" [Title/Abstract/Keywords]) AND ("Prediction" [Title/Abstract/Keywords]) AND ("Machine Learning" [Title/Abstract/Keywords]) AND ("Deep Learning" [Title/Abstract/Keywords]))
IEEE Xplore	(abstract:"Diabetes" AND abstract:"Prediction" AND abstract:"Machine Learning" AND abstract:"Deep Learning")
SpringerLink	((title:"Diabetes" OR abstract:"Diabetes" OR keyword:"Diabetes") AND (title:"Prediction" OR abstract:"Prediction" OR keyword:"Prediction") AND (title:"Machine Learning" OR abstract:"Machine Learning" OR keyword:"Machine Learning") AND (title:"Deep Learning" OR abstract:"Deep Learning" OR keyword:"Deep Learning"))
Elsevier ScienceDirect	(TITLE-ABS-KEY("Diabetes") AND TITLE-ABS-KEY("Prediction") AND TITLE-ABS-KEY("Machine Learning") AND TITLE-ABS-KEY("Deep Learning"))
IOP Science	(diabetes AND prediction AND "machine learning" AND "deep learning")
BMC (BioMed Central)	(("Diabetes" [Title/Abstract/Keywords]) AND ("Prediction" [Title/Abstract/Keywords]) AND ("Machine Learning" [Title/Abstract/Keywords]) AND ("Deep Learning" [Title/Abstract/Keywords]))

C. Study Selection

We employed the Rayyan web-based tool for review management to remove duplicate records and establish a unique database of references. In selecting articles from the main database, we followed a three-step process recommended [62] in the "Inclusion and Exclusion Criteria" section:

1. Assessing the title;
2. Reviewing the abstract and keywords;
3. Analyzing the full text.

The objective was to filter out irrelevant searches in phases (1) and (2), followed by the assessment of the remaining documents based on the specified eligibility criteria in phase (3). Ultimately, during the eligibility stage, we assembled the selected studies into our final database.

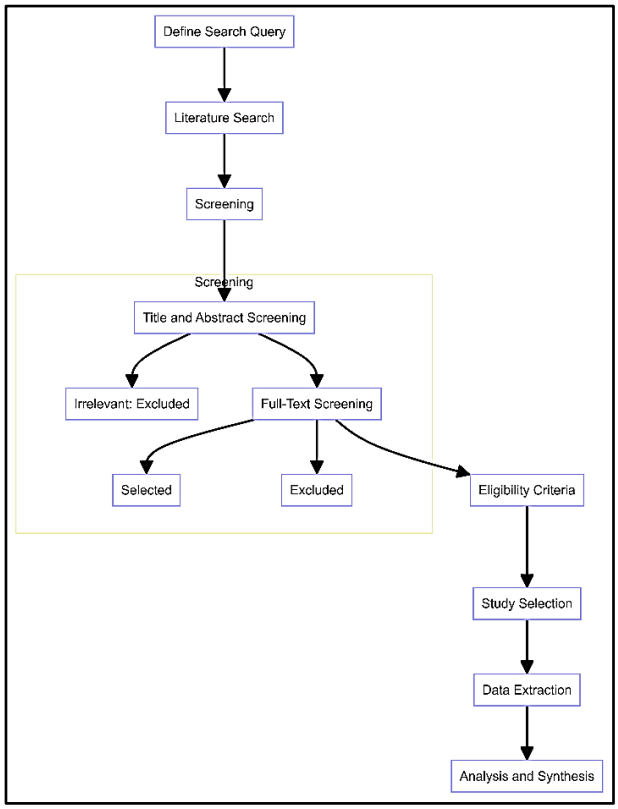

Fig. 1. Shows the study selection process

D. Quality Assessment

To evaluate the quality and reliability of the selected studies, we used the following criteria as a basis for quality assessment according to the principles adopted in the review of artificial intelligence research in healthcare.

1. The study must use and describe AI methods to classify or predict diabetes, apply appropriate validation methods, give a report on a range of performance metrics, and compare the performance of AI models.
2. The aim and hypothesis of the study must be clearly defined.
3. The study must contain appropriate answers and solutions to address the research question.
4. The study must contain a clear description of the source of the data and that data must be relevant and reliable.

IV. MACHINE LEARNING BASED DIABETES PREDICTION METHODS

Several studies have delved into the realm of machine learning for diabetes prediction, each employing various techniques and datasets. In a study conducted in 2020 by [18], Naive Bayes, Sequential Minimal Optimization (SMO), RepTree, and Simple Logistic Regression were applied to the Pima Indians diabetes dataset. Using the SMOTE method for dataset balancing, the evaluation metrics focused on accuracy, precision, and recall, with Logistic Regression achieving an accuracy of 75.70%. In contrast, a 2019 study [19] explored Fuzzy SVM techniques on the same dataset, achieving an accuracy of 89.02%, while [20] in 2021 employed Backward Elimination and Support Vector Machine (SVM) techniques, attaining an accuracy of 85.71%.

Another notable study in 2021 [21] investigated Decision trees (DT), KNN, Logistic Regression (LR), Naive Bayes, and Random Forest on a dataset comprising 70,000 records from USA hospitals, highlighting Random Forest as the best performer for classification. Similarly, [22] in 2019 employed ANN, Random forest, and K-means clustering on the PID dataset, achieving accuracies of 75.7%, 74.7%, and 73.6%, respectively, after preprocessing techniques such as data cleaning, reduction, and normalization.

Further enhancing predictive capabilities, [23] in 2020 utilized Sequential Minimal Optimization (SMO) and the Farthest first clustering algorithm, attaining an impressive accuracy of 99.4% after outlier detection and removal using Inter Quartile Range (IQR). In a more recent study in 2021 [27], Artificial Neural Networks (ANN) were applied to the PID dataset, yielding an accuracy of 85%.

However, not all studies provided performance metrics. For instance, [29] in 2021 explored KNN, SVM, DT, Gradient Boosting (GB), and RF algorithms without specifying accuracy results. Similarly, [45] in 2020 employed Glmnet, LightGBM, LM, and RF on electronic health records from Slovenia without discussing specific metrics.

In a 2021 study by [24], various ML models including Auto regression, kalman filter, Recurrent Neural Networks, and LSTM were applied to a private dataset consisting of Continuous Glucose Monitoring (CGM) records. Evaluation metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were utilized, with LSTM demonstrating the lowest MAE and RMSE values. Meanwhile, a 2018 study [25] employed data mining and statistical approaches on a private dataset, highlighting hypoglycemia and Insulin as key factors for type 1 diabetes prediction.

Similarly, a 2020 study [26] explored SVM, KNN, Naïve Bayes, ANN, and ensemble methods on a private dataset, achieving notable accuracies across different models. In contrast, a 2019 study [28] utilized SVM, RF, DT, Extra Tree Classifier, and Ada Boost on a combined private and Pima Indian Diabetes (PID) dataset, reporting high accuracy for Ada Boost.

Another 2021 study [30] introduced a Mediative Fuzzy Logic based inference system on the PID dataset, emphasizing its suitability for medical applications due to its handling of contradictory elements. These studies exemplify diverse approaches to diabetes prediction, spanning from traditional statistical methods to sophisticated ML algorithms.

Furthermore, research has been conducted on various datasets, including the PIDD, private datasets, and datasets from medical institutes. Techniques such as imputation of missing values, normalization, and feature selection have been commonly applied for data preprocessing. Evaluation metrics encompass a wide range, from traditional accuracy and precision to more specialized metrics like AUC-ROC and sensitivity.

In 2018, a study by [31] utilized Decision Trees (DT), Support Vector Machines (SVM), and Naive Bayes on the Pima Indians Diabetes Database (PIDD), concluding that Naive Bayes achieved the highest accuracy of 76.30%. Similarly, in 2020, [32] experimented with SVM, Decision Tree, Deep Learning (DL), Naive Bayes, Logistic Regression (LR), and K-Nearest Neighbors (KNN) on the same dataset, highlighting Random Forest as the best performer with an accuracy rate of 74.4%. In contrast, a 2021 study by [33] employed Logistic Regression (LG), Random Forest (RF), SVM, XGBoost, and ensemble techniques on a dataset from Hanaro Medical/Korea, revealing SVM and RF as the top performers with 73% accuracy.

Other studies explored different datasets and techniques. For instance, [34] focused on Multivariable Logistic Regression (LG) using combined datasets from Exeter cohorts, achieving an impressive ROC AUC of 0.94 by employing the SMOTE method for imbalanced data. Moreover, [35] employed an improved K-means algorithm and logistic regression on various datasets, reporting high precision (0.954), recall (0.954), Mathews correlation coefficient (0.899), and ROC (0.979). In addition, studies like [36] and [37] experimented with deep learning models such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Artificial Neural Networks (ANN) on private datasets, achieving accuracies ranging from 82.61% to 95.1%.

Furthermore, recent studies have explored more advanced techniques like Joint Bagging Boosting with Stacking Algorithm [44], LSTM with Bayesian Optimization [53], and twice-growth deep neural network (2GDNN) [55], achieving notable accuracies ranging from 80.11% to 99.57%. Moreover, ensemble techniques like Voting Classifier [48] and combination methods such as Empirical Mode Decomposition (EDM) and LSTM [40] have been investigated, revealing promising results in terms of accuracy and predictive performance. These studies collectively demonstrate the diverse range of machine learning techniques applied to diabetes prediction, showcasing advancements in accuracy, feature selection, and model interpretability.

In 2019, [38] employed a Deep Belief Neural Network on the Pima Indians Dataset, achieving high recall, precision, and F1 measure values, indicating its effectiveness. Similarly, in 2019, [41] utilized a Deep Neural Network on the same dataset, obtaining remarkable accuracies of 98.04% for five-fold and 97.27% for ten-fold cross-validation.

In 2018, [42] utilized LSTM and GRU on a dataset comprising records of over 14,000 patients, achieving an accuracy of over 97%. [43] in the same year employed a Neural Network on datasets from Luzhou hospital physical examinations and the PIMA Indian Diabetes dataset, achieving accuracies of 74.14% and 74.75% respectively.

A comprehensive comparison of various machine learning algorithms was conducted by [46] in 2019 on a diabetes dataset, with logistic regression exhibiting the highest accuracy of 96%. [47], in 2020, explored KNN, MDR, and SVM on the PIDD dataset, with SVM achieving the highest accuracy of 89%.

More recent studies include [57], who in 2021, utilized DT, KNN, RF, NB, AB, LR, SVM, and NN on PIDD after preprocessing steps, with NN providing the highest accuracy of 88.6%. Additionally, [54], in 2023, employed CNN on the PIDD dataset, achieving an accuracy of 96.13%.

Further research has explored ensemble techniques like Gradient Boosting Machine (GBM) along with Logistic Regression, as seen in [58] (2019), achieving AUC\ROC of 84.7% for GBM and 84.0% for LR. Moreover, [60], through statistical techniques and feature selection, achieved remarkable accuracies using Decision Trees (ID3), attaining a test accuracy of 99% and an average accuracy of 99.8% across k-folds, with a validation accuracy of 99.9%.

In 2019, [39] employed Long Short-Term Memory (LSTM) neural networks, specifically Recurrent Neural Networks (RNNs), on the Direct Net Inpatient Accuracy Study dataset, comprising around 110 instances. They reported Root Mean Square Error (RMSE) values ranging from 4.67 to 29.12, with 4.67 being the minimum and 29.12 the maximum obtained.

Similarly, in the same year, [50] experimented with Support Vector Machines (SVM) and Convolutional Neural Networks (CNN) on the PIDD dataset. They achieved an accuracy of 65.38% with SVM and 76.81% with CNN. Moving to 2021, [51] employed CNN and Multilayer Perceptron (MLP) on the PIDD dataset, reaching accuracy rates of 92.31% with CNN and 79.22% with MLP.

In 2020, [52] focused on Deep Neural Networks (DNN) using the NHANES dataset, evaluating the Area Under the Curve (AUC) which amounted to 80.11%. The landscape expanded further in 2023, as [56] explored Naïve Bayes, Logistic Regression, KStar, and Random Forest algorithms on a dataset sourced from Kaggle. They found Random Forest to outperform others, achieving an accuracy of 96.53%.

Lastly, [59] in 2022, experimented with Logistic Regression (LR), K-Nearest Neighbors (KNN), SVM, and Random Forest on the PIDD dataset, employing techniques such as filling missing values, outlier removal, and standardizing data. Their LR model demonstrated the highest performance, achieving an accuracy of 86%, evaluated alongside other metrics like Receiver Operating Characteristic (ROC).

The ensuing table, Table II, encapsulates the research endeavors detailed herein, specifically about the prediction of diabetes.

It is noted from Table II that diabetes prediction models are trained on specific datasets, which makes their generalizability to other populations limited. The future directions must address ensuring the generalizability of predictive models through comprehensive external validation across various populations, settings, and geographic areas.

TABLE II. MACHINE LEARNING TECHNIQUES FOR DIABETES PREDICTION

Authors	Year	ML Techniques	Dataset	Preprocessing Techniques	Evaluation Metric	Findings
[18]	2020	Sequential Minimal Optimization (SMO), Simple Logistic Regression, Naive Bayes, RepTree	Pima Indians diabetes data set	SMOTE technique for achieving dataset balance	Accuracy, Precision and recall	The accuracy for Logistic Regression, REPTree, SMO, and Naive Bayes classifiers are 75.70%, 75.10%, 74.00%, and 73.60%, respectively
[19]	2019	Fuzzy SVM techniques	Pima Indian Diabetes (PID) dataset	- Identify the features that exhibit a significant amount of missing data - Using F score for feature selection	Accuracy, Precision and recall	The accuracy is 89.02%
[20]	2021	The Backward Elimination and Support Vector Machine SVM	Pima Indian Diabetes (PID) dataset	feature selection using Backward Elimination	Accuracy	The accuracy is 85.71%
[21]	2021	KNN, Decision Trees, Random Forest, Naive Bayes, Logistic Regression	Diabetes 130-US hospitals for years 1999-2008	Features selection, outliers removing	Precision Specificity Sensitivity AUC	Naïve Bays (0.758, 0.708, 0.676, 0.791), KNN (0.798, 0.772, 0.761, 0.839), logistic regression (0.793, 0.772, 0.764, 0.850), decision tree (0.840, 0.786, 0.767, 0.832), random forest (0.890, 0.814, 0.793, 0.912)
[22]	2019	ANN, Random forest, K-means clustering	The Pima Indian Diabetes (PID)	Data cleaning (medianvalue), Data reduction (PCA), smoothing data (binning method). Association Rules (Apriori), Min-Max normalization technique	A ccuracy and AUROC curve	ANN(75.7%, 0.81), Random forest (RF) (74.7, 0.806), K means clustering (73.6)
[23]	2020	Farthest first clustering algorithm, Sequential minimal optimization (SMO)	Pima Indians diabetes data set	They employed the Interquartile Range (IQR) method to identify and eliminate outliers within the dataset	Accuracy, F measure, ROC area, Kappa statistics	The findings demonstrated an accuracy rate of 99.4%, affirming the efficacy of the hybrid model in aiding physicians to enhance decision-making in diagnosing diabetic patients
[24]	2021	Auto regression, kalman filter, Recurrent Neural Networks, LSTM (Long Short Term Memory	The CGM series and Bolus Data file contain glucose level recordings of an individual in 5-minute intervals spanning a period of 6 months. The file comprises approximately 55,000 records in total, some of which may contain empty values	The null values were excluded	Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)	SARIMA (3.87, 6.90), KALMAN (3.21, 4.36), RNN (0.32, 0.56), LSTM (11.0, 20.15)
[25]	2018	Data mining approach (Probability of sub factors, χ2Test , Info gain), Statistical approach(P-value and Confidence Interval)	Private data set with the total number of records 306, 152 affected and 154 unaffected	They used WEKA tools to clean data	X^2test, info gain	They found that factors hypoglycemia (103.342, 0.464) and Insulin (154, 1) are the key factors of type 1 diabetes. It is straightforward to determine whether

						the individual has diabetes or not based on the resultant decision tree
[26]	2020	SVM, KNN, Na¨ıve Bayes, ANN, the ensemble approach	Private dataset of 400 instances, 10 factors associated with it	N\A	Accuracy using ten fold cross validation technique	SVM (94%), KNN (91.23%), Na¨ıve Bayes (95%), ANN(96%), the ensemble approach (98.60%)
[27]	2021	ANN	Pima Indian Diabetes (PID) dataset		Accuracy	accuracy 85%
[28]	2019	SVM, RF, DT, Extra Tree Classifier, Ada Boost algorithm	Private dataset+Pima Indian Diabetes (PID)	Imputing missing values and scale the dataset to normalize all values	Accuracy, confusion matrix and f1-score	The accuracy of Logistic Regression 96%. While the Ada Boost 98.8%
[29]	2021	SVM, Decision Trees, Random Forest, Gradient Boosting (GB)	Private dataset+Pima Indian Diabetes (PID)	Filtering out the noisy data and replacing missing values	The authors didn't provide any details regarding the performace metric	The authors didn't provide any details regarding the accuracy of the algorithms employed in their study
[30]	2021	Mediative Fuzzy Logic based inference system	Pima Indian Diabetes (PID) dataset	Setting the triangular intuitionistic fuzzy number	The authors didn't provide any details regarding the performace metric	MFL is essential in medical area since it has contradicting element, essential feature to take into account when employing this method
[31]	2018	Naive Bayes, SVM, and DT	Pima Indians Diabetes Database (PIDD)		Precision, Accuracy, F-Measure, and Recall	In comparison to other methods, The research revealed that Naive Bayes achieved the highest level of accuracy at 76.30%
[32]	2020	SVM, decision Tree DT, Naive Bayes, Logistic Regression LR and KNN	Pima Indians Dataset Database (PIDD)		Accuracy	random forest has come with highest accuracy rate of 74.4%
[33]	2021	XG Boost, SVM, LG, RF	The data was gathered over a span of six years from the private medical institution Hanaro Medical in Korea, sourced from electronic records	hey chose the primary features through the utilization of techniques such as chi-squared tests, ANOVA tests, and recursive feature elimination	Accuracy, recall, precision, and F1 score	The findings indicated that SVM and RF achieved the highest level of accuracy, standing at 73%, while the lowest accuracy was 71% given by LR
[34]	2019	Multivariable Logistic Regression (LG)	The dataset consists of participants identified from 4 Exeter cohorts(DARE, PRIBA, MRC Pro/RetroMaster, MRCcrossover) and combined in one dataset	The SMOTE method was used to work with imbalanced data	ROC AUC	ROC AUC=0.94
[35]	2018	Improved K-means algorithm, logistic regression algorithm	Pima Indians Dataset, Dataset presented by Dr. Schorling, Dataset	Transforming the pregnancies attribute from numerical to nominal, replacing the missing value by the mean of attribute, normalization	Precision, recall, Mathews correlation coefficien, ROC	Precision (0.954), recall (0.954), MCC(0.899), ROC(0.979)
[36]	2018	CNN and LSTM	Private dataset called Electrocardiograms		Accuracy	CNN=90.9 LSTM=95.1
[37]	2020	ANN, SVM, Decision tree, Linear Regression, Logistic Regression	Private dataset involved a number of laboratory records	Filling the missing values with the medians, Removing features that have too many missing	Precision, recall, F1 score, accuracy AUC	Using the neural network classifier, it is able to get up accuracy of 82.61

				values		
[38]	2019	Deep belief neural network	Pima Indians Dataset		F1 measure, Recall, Precision	This network achieved impressive recall, precision, and F1 measure scores, indicating its effectiveness as a robust model
[39]	2019	Long shortterm memory (LSTM) neural networks – RNNs	The author employed the Direct Net Inpatient Accuracy Study dataset, comprising around 110 examples	Remove unnecessary repetition and outliers in successive measurements	RMSE	The RMSE values range from a minimum of 4.67 to a maximum of 29.12
[40]	2019	Empirical mode decomposition (EDM) and LSTM	The data was gathered from a hospital in Shanghai, comprising a total of 174 cases	The researchers employed interpolation techniques to address the missing data and eliminated any outliers from the dataset	The Mean Absolute Error (MAE), the Root Mean Square Error (RMSE)	The MAE assesses the error in predictions, while the RMSE indicates the disparity between the observed and true values. These evaluations are conducted across time intervals of 30, 60, 90, and 120 minutes
[41]	2019	DNN	Pima Indians Dataset(PIDD)		accuracy	The accuracy achieved by this model with a five-fold approach was 98.04%, while with a tenfold approach, it reached 97.27%
[42]	2018	LSTM and GRU	The dataset utilized comprised data from 2010 to 2015, encompassing records of more than 14,609 patients	Filling the missing values	accuracy	They gained an accuracy of over 97%
[43]	2018	Neural network	The author acquired data from Luzhou through hospital physical examinations. A separate test group was extracted, comprising 164,431 samples, each featuring 14 attribures. Additionally, another dataset utilized was the PIMA Indian diabetes dataset	Deleting the abnormal and missing values	accuracy	The neural network achieved an accuracy of 74.14% using the Luzhou dataset and 74.75% using the PIMA Indian dataset
[44]	2021	Joint bagging boosting with stacking algorithm	The authors used dataset of Demographic, medical and family history. The size of this dataset is 37,730	Data cleaning, Resampling	AUC	AUC—0.885
[45]	2020	Glmnet, LightGBM, LM, RF	EHR, Slovenia		N/A	N/A
[46]	2019	J48 decision tree, AdaBoostM1, Sequential Minimal Optimization, Bayes Net, Naïve Bayes	the electronic health records database in Shengjing Hospital of China Medical University (4205 records, 9 features)	İmputing the missing values, normalization	accuracy	J48 decision tree provided highest accuracy with 0.9503
[47]	2020	KNN, MDR, SVM	PIDD		accuracy	88%(KNN), 83% (MDR), 89(SVM)
[48]	2021	LDA, RF, Voting Classifier, SVM	PIDD		accuracy	79%(LDA), 82%(RF),

						80%(VotingClassifier), 79%(SVM)
[49]	2021	LR, RF, MLP	Self-Prepared	handle the missing values, Encoding of Categorical Data, Normalization, Feature Selection	accuracy	97.115(LR), 98.076(RF), 77.9221(MLP)
[50]	2019	SVM, CNN	PIDD		accuracy	65.38%(SVM), 76.81%(CNN)
[51]	2021	CNN, MLP	PIDD		accuracy	92.31%(CNN), 79.22%(MLP)
[52]	2020	DNN	KNHANES \ 2013–2016	Analysis of basic characteristics, select non-invasive variables, and correlation analysis	AUC	80.11%
[53]	2023	RNN-LSTM with Bayesian optimization	Self-Prepared from 489 patients between the years 2019 and 2021	missing data imputation	Sensitivity, specificity, and AUC	95% (sensitivity), 99% (specificity), 98% (AUC)
[54]	2023	CNN	PIDD		accuracy	96.13%
[55]	2022	twice-growth deep neural network (2GDNN)	PIDD and the diabetes dataset from the Laboratory of the Medical City Hospital (LMCH)	The framework incorporates the utilization of Spearman correlation for feature selection and polynomial regression for handling missing values	precision, sensitivity, F1-score, train-accuracy, and test-accuracy scores	PIDD [97.34%, 97.24%, 97.26%, 99.01%, 97.25] and LMCH [97.28%, 97.33%, 97.27%, 99.57%, 97.33]
[56]	2023	Naïve Bayes, Logistic Regression, KStar, and Random Forest	The dataset is collected from Kaggle		accuracy, f1-measure, precision, and recall	Random Forest performs better with accuracy of 96.53%
[57]	2021	NN, DT, RF, AB, LR, NB, SVM, KNN	PIDD	Filling missing values, Outlier removal Data normalization, Feature selection	accuracy, f1-measure, precision, and recall	The NN provides highest accuracy of 88.6%
[58]	2019	Logistic Regression and Gradient Boosting Machine (GBM) techniques	records of 13,309 Canadian patients with ages 18 - 90 years	Removing the missing values	AUC\ROC	84.7%9(GBM), 84.0%(LR)
[59]	2022	LR, KNN, SVM, RF	PIDD	Filling the missing values, removal of outliers and standardizing the data	Accuracy, ROC	The proposed LR has highest performance with accuracy 86%
[60]	2020	DT(ID3)	Hospital Frankfurt Germany Diabetes Data Set	different statistical techniques and feature selection	accuracy, specificity, sensitivity, precision, recall, F1-score, MCCandROC-AUC	DT (ID3) attained a test accuracy of 99%, an average accuracy of 99.8% across k-folds, and achieved a 99.9% accuracy through LOSO validation
[61]	2024	Logistic Regression, Decision Tree, SVM, Ada Boost, Random Forest, Gradient Boosting, and KNN	PIMA dataset and the Iraqi Society in Medical City Hospital and the Diabetes-Al Kindy	filling the missing values, nomalization, and eliminating noisy data	Accuracy	Logistic regression achieves the highest accuracy performance of 79% when applied to the Pima dataset. On the other hand, when the Iraqi society dataset is utilized, gradient boosting demonstrates superior performance with an accuracy rate of 97.7%

V. DISCUSSION

A. Dataset

Diabetes research benefits greatly from diverse datasets that encompass a wide array of demographics, clinical parameters, and monitoring techniques. Among the notable datasets available, the Pima Indians Diabetes Dataset (PIDD) stands out with its 768 samples and 8 features, including crucial indicators like age, body mass index (BMI), and plasma glucose concentration, providing invaluable insights into diabetes prevalence among Pima Indians. Additionally, the CGM dataset comprising 55,000 samples provides continuous glucose monitoring data, offering a detailed perspective on glucose fluctuations over 6 months in 5-minute intervals.

Incorporating diverse perspectives, a dataset sourced from Dhaka presents 306 samples with 22 features, encompassing factors such as HbA1c, pancreatic disease history, and symptoms like frequent urination and increased thirst. Similarly, datasets collected based on various criteria, such as age above 35 years or specific health parameters, offer nuanced insights into diabetes risk factors and symptoms.

Furthermore, regional datasets like the King Abdullah International Research Centre Diabetes (KAIMRCD) dataset and datasets from Luzhou, China, and Iraq, offer context-specific insights into diabetes prevalence, risk factors, and associated health indicators.

Notably, datasets like the CPCSSN dataset and the Frankfurt hospital dataset provide clinical parameters such as BMI, blood sugar levels, and lipid profiles, essential for assessing diabetes management and associated cardiovascular risks.

Additionally, comprehensive datasets like the Diabetes 130 US hospital dataset from the UCI Machine Learning Repository offer extensive clinical and demographic information spanning over a decade, facilitating longitudinal studies and predictive modeling.

Overall, the wealth of diabetes datasets available, ranging from population-based surveys to clinical records and monitoring data, empowers researchers to explore multifaceted aspects of diabetes epidemiology, management, and outcomes, contributing significantly to advancements in diabetes research and patient care. These data collections consist of different quantities of cases (patients), yet they possess common fundamental features, Table II provides an overview of the datasets considered in this study. Based on the Table III, the common features across different datasets in the field of diabetes and related health indicators are:

1. Age,
2. Gender
3. Body mass index (BMI)
4. Blood glucose level (Glucose)
5. Family history
6. Hypertension
7. Medical history (including cardiovascular disease, stroke, diabetes)
8. Smoking status
9. Blood pressure (Systolic and Diastolic)
10. Insulin
11. Cholesterol levels (including HDL, LDL, total cholesterol)
12. Triglycerides
13. Weight
14. Height
15. Physical activity

These features appear across multiple datasets and are relevant for studying diabetes, its risk factors, and associated health indicators.

TABLE III. DATASETS USED IN DIABETES PREDICTION AND CLASSIFICATION

Dataset	Reference	Number of Samples	Number of Features	Features	Link
Pima Indians Diabetes Dataset (PIDD)	[18], [19], [20], [22], [23], [38], [41], [47], [48], [50], [51], [54], [56], [57], [59]	768	8	Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Body mass index (weight in kg/(height in m)^2), Number of times pregnant, Triceps skin fold thickness (mm), Age (years), Diabetes pedigree function, 2-Hour serum insulin (mu U/ml), Diastolic blood pressure (mmHg), Class variable (0 or 1)	Pima Indians Diabetes Database | Kaggle
CGM dataset	[24]	55000	1	Monitoring blood sugar levels every five minutes for a duration of six months.	Private
A dataset acquired from Dhaka, originating from a specific questionnaire.	[25]	306	22	Frequent Urination, Family History affected in Type 1 Diabetes (Father and Mother), Increased thirst, Impaired glucose metabolism, Education of Mother, Standardized birth weight, Standardized growth-rate in infancy, Extreme Hunger, Adequate Nutrition, Age, Autoantibodies, Sex, HbA1c, Hypoglycemia, Unintended weight loss, Pancreatic disease affected in child, Fatigue and Weakness, Area of Residence, Family History affected in Type 2 Diabetes (Father and Mother)	Private
A dataset selected based on a random and varied assortment of individuals aged over 35 years.	[26]	400	10	Urination, Fatigue, Gender, Drinking, Weight, Height, Age, Family history, Smoking, Thirst	Private
Dataset Collection	[28]	800	10	Glucose Level, Age, Blood Pressure, Job Type, Skin Thickness(mm), Insulin, Number of Pregnancies, BMI, Work/Machine-work,	Private
Dataset gathered at Hanaro Medical foundation in Seoul, Korea as electronic records for 6 years	[33]	535,169	12	HbA1c, BMI, physical activity, age, family history, sex, uric acid, gamma-GTP, Triglycerides, smoking, drinking, FPG	Private

Electrocardiograms (ECG)	[36]	142 000	1	HRV, derived from ECG signals, refers to the fluctuations in instantaneous heart rate.	Private
A dataset collected from Almazov specialized medical center's medical information system data set , Russia.	[37]	238,590	31	Low density lipoproteins (LDL), Cholesterol, Troponin, Glucose, Alanine transaminase (ALT), Mean corpuscular volume (MCV),Hemoglobin (HGB), Procalcitonin (PCT),Red blood cell distribution width (RDW), Neutrophils (NEUT),Platelet distribution width (PDW), Hematocrit (HCT),Aspartate aminotransferase (AST),Mean cell hemoglobin (MCH),Platelets (PLT),Bilirubin,Gender,Red blood cell count (RBC),Blood in urine (BLD),White blood count (WBC),Monocytes,pH,Leukocytes (LEU),Age,Nephropathy,High-density lipoprotein (HDL),Creatinine,Mean platelet volume (MPW),Retinopathy, Triglycerides	Private
Dataset from the King Abdullah International Medical Research Center Diabetes (KAIMRCD).	[40]	174	1	The Continuous Glucose Monitoring (CGM) system has the capability to capture the patient's blood glucose levels at intervals of 5 minutes for 3 days.	Private
KAIMRCD	[42]	14,609 cases	30	Gender, Age and 28 Vital Signs and Lab Readings	The KAIMRCD dataset is accessible upon formal request to KAIMRC
The dataset concerning hospital physical examinations conducted in Luzhou, China.	[43]	164431	14	Height, Low Density Lipoprotein (LDL), Left Diastolic Pressure (LDP), Right Systolic Pressure (RSP), Pulse Rate, Waistline, Right Diastolic Pressure (RDP), Weight, Breathe, Age, Left, Systolic Pressure (LSP), High, Density Lipoprotein (HDL), Physique Index, Fasting Glucose	Private
The dataset from the Henan Rural Cohort Study	[44]	37,730	20	Demographic(Gender, Education Level, Per Capita Monthly Income, Age, Marital Status), Medical And Family History(Family History Of Dyslipidemia, Family History Of CHD, Family History Of T2DM, Family History Of Hypertension), Anthropometric Indicators (Diastolic Blood Pressure, Heart Rate, Waist To Hip Ratio, Systolic Blood Pressure), Lifestyle And Dietary Indicators (Drinking Tea Frequently, High Salt Diet, Adequate Vegetable And Fruit Intake, Smoking, Physical Activity, High Fat Diet, Drinking)	The dataset is accessible upon the formal request
Electronic health record (EHR) information gathered during routine preventive healthcare check-ups among a healthy demographic at ten primary healthcare facilities in Slovenia.	[45]	3,723	59	N\A	The datasets can be obtained from the corresponding author upon a reasonable request and subject to approval from the data providers.
The database containing electronic health records at Shengjing Hospital affiliated with China Medical University.	[46]	4205	9	Family History of Diabetes, Work Stress, Age, Hypertension, Gender, Body Mass Index (BMI), History of Cardiovascular Disease or Stroke, Physical Activity, Salty Food Preference	The datasets can be obtained from the corresponding author upon a reasonable request and subject to approval from the data providers.
Self-Prepared	[49]	N\A	15	Delayed Healing, Visual Blurring, Muscle Stiffness, Polyphagia, Sudden Weight Loss, Genital Thrush, Weakness, Age, Alopecia, Irritability, Itching, Polyuria, Partial Paresis, Polydipsia, Obesity	Self-Prepared
The Korean National Health and Nutrition Examination Survey	[52]	15900	7	Gender, Family history of diabetes, Smoking status, Hypertension, Body mass index Waist circumference, Age	If researchers wish to make use of non-public KNHANES data or data

(KNHANES)2013–2016					associated with other governmental bodies, they must carry out their analysis within the confines of a research data center.
The data was acquired from Karadeniz Technical University Medical Faculty Farabi Hospital (located in Trabzon city) during the period spanning January 2019 to March 2021.	[53]	489	73	N\A	Private
Diabetes dataset gathered from Iraqi society in the Laboratory of Medical City Hospital (LMCH)	[55], [61], [63], [64], [65], [66], [67], [68], [69], [70]	1000	10	Body Mass Index (BMI), Urea, Triglycerides, Low-Density Lipoprotein (LDL), Cholesterol, Creatinine Ratio, Hemoglobin A1c (HBA1C), Age, High-Density Lipoprotein (HDL), Very-Low-Density Lipoprotein (VLDL)	https://data.mendeley.com/datasets/wj9rwkp9c2/1
CPCSSN dataset	[58], [71], [72], [73]	13,309	8	FBS (Fasting Blood Sugar), LDL (Low Density Lipoprotein), Age, BMI, HDL (High Density Lipoprotein), sBP (Systolic Blood Pressure), TG (Triglycerides), Sex	www.cpcssn.ca
Data on diabetes patients collected from Frankfurt Hospital in Germany.	[60], [74], [75], [76], [77], [78], [79], [80]	2000	9	Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome	https://www.kaggle.com/datasets/johndasilva/diabetes
Diabetes 130 US hospital 1999-2008	[21], [81], [82], [83], [84], [85], [86], [87], [88]	100000	55	diag_3,admission_source_id,glipizide,discharge_disposition_id,gender,encounter_id,glimepiride-pioglitazone,metformin-rosiglitazone,num_lab_procedures,diag_2,pioglitazone,examide,admission_type_id,A1Cresult,glimepiride,number_diagnoses,rosiglitazone,nateglinide,metformin,acetohexamide,tolbutamide,repaglinide,metformin-pioglitazone, change,race,troglitazone,weight,medical_specialty,age,insulin,glyburide-metformin,glipizide-metformin,time_in_hospital,diabetesMed,patient_nbr,number_emergency,payer_code,number_procedures,number_inpatient,glyburide,number_outpatient,citoglipton,number_inpatient,diag_1,acarbose,max_glu_serum,metformin,glipizide-metformin, discharge_disposition_id,number_emergency,num_medications,number_outpatient	Diabetes 130-US hospitals for years 1999-2008 - UCI Machine Learning Repository
The historical data set DirecNet Inpatient Accuracy Study provided by Diabetes Research in Children Network (DirecNet)	[39], [89], [90], [91], [92], [93], [94]	110 T1DM patients from CGM device	5	Patient ID, Number of measurements, Min BGL, Max BGL, Mean BGL	https://public.jaeb.org/direcnet/stdy/

B. Data Preprocessing

The effectiveness of Machine Learning (ML) in addressing a specific problem is impacted by various factors. Foremost among these is the development and quality of the dataset. Real-world data is frequently messy, insufficient, and untrustworthy. Analyzing data during the training phase becomes more demanding when it contains excessive redundant content, noise, or inaccuracies. Any manipulation of raw data to render it more manageable and beneficial for subsequent processing is termed data preprocessing. Data preprocessing encompasses several categories including [95]:

1) Missing Value Imputation: One of the fundamental challenges in data preprocessing is handling missing values. Missing data can significantly impact the performance of machine learning models. Techniques such as mean imputation, median imputation, and predictive imputation are commonly used to address missing values.

Mean imputation is the process of filling the missing values with the mean of a given feature. This method is easy to apply and suitable for data with normal distribution. Removing the relationships between features and changing the distribution of data when contains outliers are disadvantages of this approach. It may decrease the performance of the model when the value of the imputed mean does not represent the missing data.

The process of filling the missing data with the median of a given feature is called median imputation. It is robust to outliers and gives better results for data of skewed distributions. As in mean imputation, it removes the relationships between features and may decrease the

performance of the model when the value of the imputed mean is not representative in multimodal data.

The predictive imputation depends on other available features for handling the missing data. This technique considers the relationships between features leading to a better imputation. It requires an accurate model and more steps of preprocessing. This method can improve the performance of the model by giving a more accurate imputation.

2) Data Normalization and Standardization: Data normalization and standardization involve methods employed to adjust numerical attributes to a common scale. Normalization rescales the data to fit within the range of 0 to 1, whereas standardization alters the data to have an average of 0 and a standard deviation of 1. These techniques help in improving the convergence of optimization algorithms and make features comparable.Encoding Categorical Variables: Many machine learning algorithms require numerical input, making it necessary to encode categorical variables. Techniques such as one-hot encoding, label encoding, and target encoding are used to transform categorical data into a numerical format that can be easily understood by machine learning algorithms.

3) Feature Selection and Dimensionality Reduction: In datasets with a large number of features, feature selection and dimensionality reduction techniques are employed to reduce the complexity of the data. Techniques like Principal Component Analysis (PCA), feature importance ranking, and recursive feature elimination (RFE) help in selecting the most relevant features and reducing redundant information.

PCA proves particularly efficient when working with big databases with lots of features, especially when dimensionality reduction is desired for computational efficiency or graphical representations. When the interpretability of the transformed features is important, it is less appropriate. RFE is useful in situations where the selected features' interpretability is important to figure out each risk factor's role. Combining the RFE and PCA may be informative in some circumstances. The most important factors or features from the reduced dataset can then be chosen using RFE after PCA has first reduced the dimensionality.

4) Outlier Detection and Removal: Outliers can significantly affect the performance and accuracy of machine learning models. Various statistical techniques such as z-score, interquartile range (IQR), and isolation forests are used to detect and remove outliers from the dataset, ensuring the robustness of the model.

Outliers are identified using the z-score method by counting the number of standard deviations of a data point from the mean. Data are considered outliers if the z-score is greater than a predefined threshold. This method is effective if data sets have a normal distribution and may not be suitable for skewed data or when relationships between variables are non-linear.

In the IQR method, the data set is divided into quartiles. The data is sorted in ascending order and divided into 4 equal parts. Q1, Q2 and Q3 which are called the first, second and third quartiles are the values that separate the four equal parts. IQR is difference between the 75th percentile (Q3) and the 25th percentile (Q1). Data points falling outside 1.5 times the IQR above Q3 or below Q1 are considered outliers. The IQR method is effective for data that have non-normal distributions and is efficient for detecting outliers in data sets with different distributions.

5) Data Balancing: In classification problems, imbalanced datasets, characterized by a substantial disproportion between classes, can result in models that exhibit bias. Methods such as oversampling, undersampling, and synthetic data generation (SMOTE) are used to balance the distribution of classes, improving the model's ability to generalize across different classes.

The oversampling approach seeks to achieve balance in class distribution by replicating minority class instances. With this method, the model may result in overfitting since the data will be repeated. This method enhances the performance of the model if the risk of overfitting is controlled.

In contrast to oversampling, undersampling is performed by reducing the number of majority class instances until the balance of class distribution is achieved. undersampling can lose useful information from the majority class. undersampling can be effective with datasets involving high majority redundancy.

SMOTS interpolates between existing instances to create synthetic examples for the minority class. Actual data may not be accurately represented by synthetic samples, particularly if the feature space is complicated.

Table IV presents the preprocessing techniques used by the studies considered in this paper. Based on this table, certain studies didn't undertake any data preprocessing, whereas others engaged in preprocessing across all six mentioned categories.

Addressing missing values constituted a crucial step, about 32 studies following methodologies involved various strategies such as imputation and deletion to mitigate the impact of missing data on subsequent analyses. Normalization and standardization procedures were also implemented, drawing upon methodologies used in 16 studies. These techniques ensure that variables are brought to a comparable scale, mitigating issues related to the disparate magnitudes of different features within the dataset. Categorical variables were encoded utilizing methodologies detailed in reference [49], [72], [85], facilitating their integration into subsequent analyses. This process involved transforming categorical data into numerical representations, enabling the inclusion of such variables in mathematical models effectively. Feature selection and dimensionality reduction techniques were applied to streamline the dataset and enhance model performance. 23 studies guided the implementation of these methodologies, which involved identifying and retaining the most informative features while reducing redundancy and computational complexity.

Outlier detection and removal were performed to mitigate the influence of erroneous data points on subsequent analyses. 16 studies provided methodologies for identifying

and appropriately handling outliers, ensuring the robustness and reliability of the analytical process. Additionally, data balancing techniques were employed to address potential biases stemming from class imbalance within the dataset. Strategies outlined in 8 works were utilized to ensure equitable representation of different classes, thereby enhancing the generalizability of the model. Finally, data smoothing techniques were applied to mitigate noise and fluctuations within the dataset, as guided by methodologies described in references [22], [35], [85].

TABLE IV. THE PREPROCESSING TECHNIQUES

Preprocessing Categories	Frequency	References
Missing Value Handling	32	[19], [22], [24], [25], [28], [29], [35], [37], [40], [42], [43], [44], [46], [49], [53], [55], [57], [58], [59], [61], [65], [69], [72], [75], [76], [78], [79], [82], [84], [85], [86], [95]
Data Normalization and Standardization	16	[22], [28], [35], [46], [49], [57], [59], [61], [65], [72], [75], [84], [85], [88], [91], [95]
Encoding Categorical Variables	3	[49], [72], [85]
Feature Selection and Dimensionality Reduction	23	[19], [20], [22], [30], [33], [49], [52], [55], [57], [60], [65], [66], [67], [68], [69], [70], [72], [73], [76], [82], [87], [88], [95]
Outlier Detection and Removal	16	[23], [29], [39], [40], [57], [59], [61], [64], [70], [72], [82], [85], [86], [89], [90], [95]
Data Balancing	8	[18], [34], [44], [74], [80], [82], [85], [88]
Smoothing Data	3	[22], [35], [85]

C. Machine Learning Techniques of Diabetes Prediction

The frequency of employing various intelligent methods for predicting diabetes was established based on studies conducted over the past six years, as indicated in Table IV. Random Forest, Logistic Regression, Decision tree, Support Vector Machine (SVM), Naïve Bayes, and KNN emerged as the predominant methods for diabetes prediction, as depicted in Fig. 2. Table V presents a summary of the pros and cons associated with each of these algorithms.

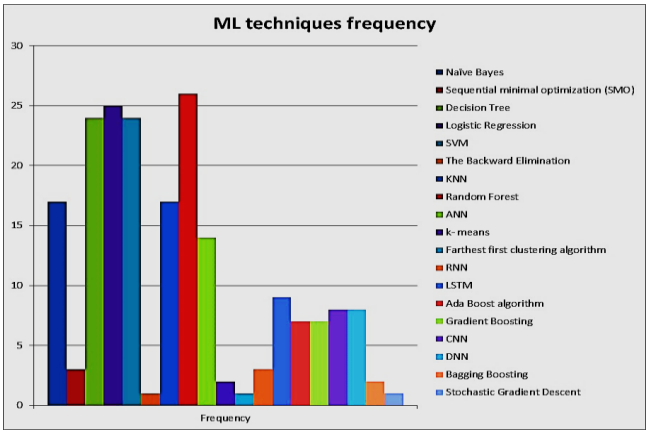

Fig. 2. Frequencies of machine learning techniques

TABLE V. ML TECHNIQUES FREQUENCY

ML Technique	Frequency	References
Naïve Bayes	17	[18], [21], [26], [31], [32], [46], [56], [57], [64], [66], [70], [72], [74], [75], [77], [85], [86]
Sequential minimal optimization (SMO)	3	[18], [23], [46]
Decision Tree	24	[18], [21], [28], [29], [31], [32], [37], [46], [57], [60], [61], [64], [70], [72], [73], [74], [76], [77], [78], [79], [80], [85], [86], [87]
Logistic Regression	25	[18], [21], [32], [33], [34], [35], [37], [49], [56], [57], [58], [59], [61], [66], [70], [71], [72], [73], [74], [75], [77], [80], [83], [86], [87]
SVM	24	[19], [20], [26], [28], [29], [31], [32], [33], [37], [47], [48], [50], [57], [59], [61], [64], [68], [72], [74], [75], [77], [79], [80], [83]
The Backward Elimination	1	[20]
KNN	17	[21], [26], [29], [32], [47], [57], [59], [61], [64], [68], [69], [74], [75], [77], [78], [80], [83]
Random Forest	26	[21], [22], [28], [29], [33], [45], [49], [56], [57], [59], [61], [64], [66], [67], [68], [69], [70], [74], [75], [76], [77], [79], [80], [83], [85], [87]
ANN	14	[22], [26], [27], [37], [43], [49], [51], [57], [75], [77], [78], [86], [89], [91]
k- means	2	[22], [35]
Farthest first clustering algorithm	1	[23]
RNN	3	[24], [39], [53]
LSTM	9	[24], [36], [39], [40], [42], [53], [72], [90], [94]
Ada Boost algorithm	7	[28], [46], [57], [61], [64], [75], [77]
Gradient Boosting	7	[29], [33], [58], [61], [70], [75], [87]
CNN	8	[36], [50], [51], [54], [72], [84], [88], [94]
DNN	8	[38], [41], [52], [55], [64], [68], [82], [85]
Bagging Boosting	2	[44], [77]
Stochastic Gradient Descent	1	[75]

1) Random Forest

Breiman (2001) introduced the popular ensemble classification technique known as the random forest method, which is widely utilized in various application domains within the realms of machine learning and data science [96]. The RF algorithm is a type of supervised machine learning technique applicable for both regression and classification tasks, depending on the nature of the problem at hand. This method integrates numerous decision trees through random aggregation, subsequently amalgamating their predictions to facilitate decision-making processes. This technique operates by assembling predictions generated by individual decision trees and subsequently determining the final decision through a majority voting mechanism. The amalgamation of decision trees via random aggregation presents a promising avenue for enhancing decision-making accuracy in various domains. It is of significance to acknowledge that the Random Forest

(RF) algorithm exhibits optimal performance under conditions characterized by a substantially larger number of variables relative to the number of observations within the environment. The primary concern regarding models employing decision trees is that when the dataset contains a limited number of samples, overfitting becomes a prominent issue. The random forest algorithm is a meta-learner that examines different subsets of the dataset with a range of decision tree classifiers, and then combines their outcomes to enhance predictive accuracy and reduce the risk of overfitting. The size of the subsample typically matches that of the source data [97].

2) Logistic Regression

Logistic regression analysis constitutes one of the methodologies employed for categorizing observations within a dataset. In the realm of statistics, logistic regression serves as a methodology aimed at classifying binary variables into two distinct classes [98]. In logistic regression, a statistical technique widely employed in scientific studies, the association between a collection of predictor variables and a categorical outcome variable is depicted through a curve. This curve illustrates the probability of an event taking place, providing insights into the likelihood of occurrence based on the specified independent variables. In scientific literature, it is commonly acknowledged that while independent variables within a study may exhibit a spectrum of values ranging from continuous to categorical, the dependent variable is inherently categorical [99]. Logistic regression is well-suited for modeling scenarios characterized by binary outcomes, where the focus lies on distinctions between two discrete states, such as success versus failure, affirmative versus negative responses, or healthy versus unhealthy conditions. The sigmoid function is commonly applied in logistic regression methodologies to derive binary outcome probabilities from one or multiple predictors, thereby facilitating the selection of optimal parameter values. Equation (1) and Fig. 3 illustrates the sigmoid function denoted as σ, along with its input variable z [100].

$$\sigma(z) = \frac{1}{1+e^{-z}}, \qquad z \epsilon R \; and \; \sigma(z) \epsilon \; (0,1) \tag{1}$$

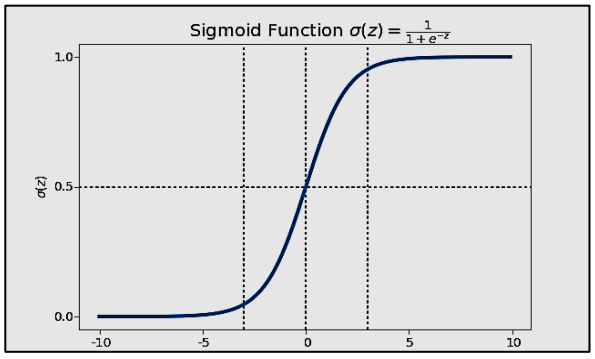

Fig. 3. Sigmoid function

3) Decision Trees

A supervised learning approach commonly employed for tackling classification tasks is the decision tree method. Decision trees iteratively partition the provided dataset into two or more subsets based on attributes, aiming to predict the class value of the target variable [101]. By segregating the dataset, decision trees construct a model to forecast unknown class labels. This technique accommodates both binary and continuous variables. Decision trees select the optimal root node by maximizing entropy, thereby favoring the most consistent hypothesis within the training data. Input to the decision tree consists of attributes and instance values, while the output yields the decision model. Challenges in decision model construction include attribute selection, determination of splits, defining stopping criteria, pruning, adequacy of training samples in terms of quality and quantity, and the sequence of splits, among others.

The decision model is structured as a tree, consisting of nodes including decision nodes (split nodes with conditions) and leaf nodes. Fig. 4 illustrates the representation of this decision tree. Selecting the appropriate attribute as the root node to initiate the split poses a significant challenge among the various attributes within the dataset. The decision node may bifurcate into 2 or more branches. The process begins with the selection of the root node, wherein the model identifies the optimal attribute or predictor node from the available set. Various methodologies exist to determine the best attribute for the root node, typically relying on measures of impurity within child nodes such as Entropy, Gini index, and classification error. These performance measures are computed for all attributes, and subsequent comparison aids in selecting the most favorable split [102].

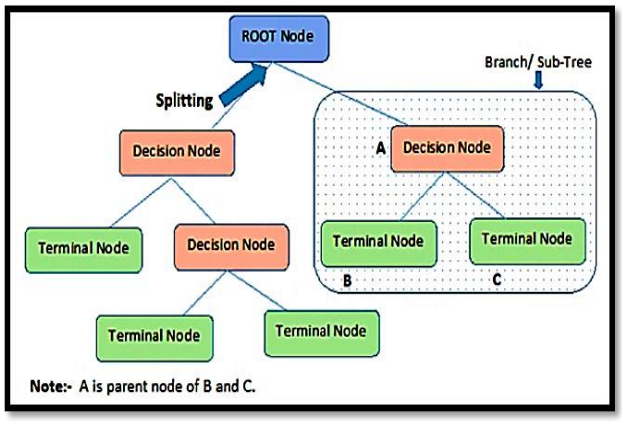

Fig. 4. Decision tree model

4) Support Vectors Machine SVM

SVM, a classification algorithm rooted in the principles of statistical learning theory, was originally devised by Vapnik in the 1990s. Recognized for its efficacy, SVM stands as one of the foremost methods within the realm of data mining algorithms [103]. Notably, SVM exhibits proficiency in segregating data into two or more classes through linear separation mechanisms across different dimensions: linear in two-dimensional space, plan in three-dimensional space, and hyperplane in multidimensional space. Linear separability arises when data clusters can be effectively partitioned by a line. An innovative concept proposed within SVM is that the demarcation between two classes is not merely a line but a margin, with its width optimized by certain data vectors to attain maximum separation. This scenario is illustrated in Fig. 5.

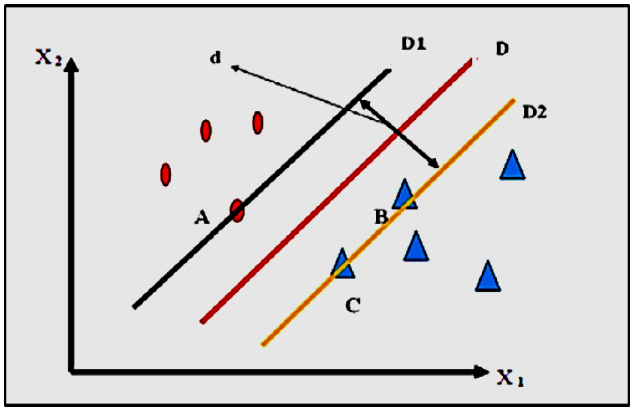

Fig. 5. The case where two classes can be linearly separated

Here, the lines D, D1, D2 are determined by the equations (2), (3), and (4) respectively, where w is the weight vector and b is the constant (bias) value.

$$w.x + b = 0 \quad (2)$$

$$w.x + b = 1 \quad (3)$$

$$w.xi + b = -1 \quad (4)$$

If the distance (margin) between the lines D1 and D2 is denoted by d, with analytical geometry information, it is easy to see that d can be calculated as in equation (5).

$$d = \frac{2}{\|w\|} \quad (5)$$

The value of $\|w\|$ is calculated as in equation (6).

$$\|w\| = \sqrt{{w_1}^2 + {w_2}^2} \quad (6)$$

To get the maximum value of d, naturally the minimum value of $\|w\|$ or $\|w\|^2$ must be found.

$$Min.\frac{1}{2}\|w\|^2 \quad (7)$$

Constraints:

$$w.x + b \geq 1\,, if\ y_i = 1\ (Class\ 1 - Red) \quad (8)$$

$$w.x + b \leq -1\,, if\ y_i = -1\ (Class\ 2 - blue) \quad (9)$$

5) Naïve Bayes

The theoretical framework of the Naive Bayes algorithm draws upon principles pioneered by Thomas Bayes during the 18th century. These principles enable the assessment of event probabilities and their revision in response to new data [85]-[86]. The algorithm relies on the fundamental notion of conditional probability as elucidated by Bayes' Theorem. This concept is exemplified through a scenario, accompanied by the formula and a commonly cited medical testing illustration. Consider a hypothetical cancer screening test administered to a cohort of 1000 individuals:

Let A represent the probability of an individual having cancer, and B denote the probability of receiving a positive outcome from a cancer screening examination. Within this context, P(A|B) signifies the probability that an individual who tests positive for cancer actually has the disease. This concept epitomizes the essence of Bayesian theory, highlighting the potential for false positives in cancer screening tests. Moreover, it underscores the existence of individuals with cancer who may not register as positive on the screening test. Consequently, Bayes' conditional probability formula elucidates the computation of $P(A|B)$ utilizing the probabilities $P(B)$, $P(A)$, and $P(A \cap B)$. In essence, the Bayesian formula enables the assessment of the reliability of disease screening tests based on historical data [104], [105]. The Bayesian formula can be expressed as follows:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)} \quad (10)$$

6) K Nearest Neighbors KNN

The K-Nearest Neighbors (KNN) algorithm is a widely used non-parametric method in machine learning and pattern recognition for classification and regression tasks. It operates on the principle of instance-based learning, where the algorithm doesn't explicitly learn a model but instead memorizes the training instances to make predictions for unseen data points based on their proximity to the training examples in the feature space. In KNN, the 'K' denotes the number of nearest neighbors considered for classification or regression [106]. To predict the label or value for a new data point, the algorithm identifies the 'K' nearest neighbors in the training set based on a chosen distance metric, typically Euclidean distance. The most common approach is to assign the majority class label (for classification) or compute the average (for regression) among these neighbors and assign it to the new data point.

One of the notable characteristics of the KNN algorithm is its simplicity and intuitive nature. It doesn't require explicit training or complex optimization procedures, making it easy to implement and understand. However, its performance can be sensitive to the choice of the distance metric and the value of 'K', which need to be carefully tuned based on the specific dataset and problem at hand.

Despite its simplicity, KNN can be quite effective, especially in low-dimensional feature spaces or when the decision boundaries are complex and nonlinear. Moreover, it can handle multi-class classification and regression tasks without any modifications, making it versatile across various applications.

However, KNN's computational complexity grows linearly with the size of the training set, making it less efficient for large datasets. Additionally, it can struggle with high-dimensional feature spaces due to the curse of dimensionality, where the notion of distance becomes less meaningful as the number of dimensions' increases.

In conclusion, while the KNN algorithm offers simplicity and versatility, its performance can vary based on the dataset characteristics and parameter choices. It remains a valuable tool in the machine learning toolbox, particularly for small to moderate-sized datasets and tasks where interpretability and ease of implementation are prioritized [61], [102]. Table VI summarizes the pros and cons of the most used machine learning algorithms in diabetes prediction.

TABLE VI. THE PROS AND CONS OF ML ALGORITHMS

Algorithm	Pros	Cons	Performance
Random Forest	- Handles large datasets effectively - Reduces overfitting - Can handle missing values - Provides feature importance	- Complexity and computational overhead - Less interpretable compared to decision trees - Biased towards categorical variables with more levels	RF's high usage reflects its robustness and efficiency in dealing with diverse datasets and reducing overfitting risks. The ensemble method's ability to manage large feature sets makes it a preferred choice
Logistic Regression	- Simple and easy to implement - Interpretable results - Works well with small datasets - Outputs probabilities	- Assumes linearity between dependent and independent variables - Not suitable for non-linear relationships - Sensitive to outliers	Effective for binary/multiclass classification problems and for linearly separable data
Decision Tree	- Easy to interpret and visualize - Handles both numerical and categorical data - Requires little data preprocessing - Captures non-linear relationships	- Prone to overfitting - Instability: small changes in data can lead to different trees - Biased with imbalanced datasets	High accuracy but has scalability issues with large data sets
Support Vector Machine	- Effective in high-dimensional spaces - Versatile: Different kernel functions for customization - Robust against overfitting - Effective in cases where number of features > number of samples	- Not suitable for large datasets - Complex to fine-tune parameters - Computationally intensive	SVM's strength in high-dimensional data and robustness against overfitting justify its frequent use. However, its complexity and resource demands can be limiting factors
Naïve Bayes	- Simple and easy to implement - Works well with high-dimensional data - Efficient in training and prediction - Performs well in multi-class prediction - Robust to irrelevant features and noise	- Assumes independence of features, which may not be true in real-world data - Sensitivity to irrelevant features - Requires a relatively large amount of training data for accurate estimates of probabilities - Prone to the zero probability problem	- Promising accuracy for certain types of data - Excellent scalability
KNN	- No training phase, simple to implement - No assumptions about the underlying data distribution - Effective with non-linear data and boundary - Performs well with small training datasets - Handles multi-class cases naturally	- Computationally expensive during prediction, especially with large datasets - Sensitive to the choice of distance metric and k value - High memory requirements, as it stores all training data points - Not suitable for high-dimensional data - Classifying an unseen instance can be slow if the dataset is large	KNN is valued for its simplicity and effectiveness in certain conditions, particularly where interpretability and ease of implementation are crucial. Its performance is highly dependent on parameter tuning and dataset characteristics
Artificial Neural Networks (ANN)	Captures complex patterns, effective for large datasets, adaptable through various architectures	Requires substantial computational resources, prone to overfitting, challenging to interpret	ANN's ability to model complex patterns makes it suitable for diverse applications. Its resource intensity and interpretability issues need careful consideration
Long Short-Term Memory (LSTM)	Excellent for sequential data, mitigates vanishing gradient problem, captures long-term dependencies	Computationally demanding, complex architecture	LSTM's usage highlights its suitability for time-series and sequence-based tasks, leveraging its ability to remember long-term dependencies

When developing a predictive model for diabetes, preprocessing and machine learning algorithms face some difficulties, which can affect the effectiveness of the models.

When a model learns very efficiently on training data, noise and small changes may be present in the model, which negatively affects the model's performance when applied to new data. These obstacles are important in diabetes prediction models, as diabetes prediction for new data must be accurate and efficient.

One approach to reduce the aforementioned overfitting is to evaluate the performance of model on different subsets of data using cross-validation techniques, such as k-fold cross-validation, to ensure the generalizability of the model. The implementing k-fold cross-validation involves splitting the dataset into k subsets. After that, the model is trained on k-1 subsets, and validation is performed on the remaining subset. During each iteration, the validation subset is rotated.

Another way to address overfitting is to enhance the diversity of the training data using augmentation methods

like bootstrapping or generating synthetic data. When applying data augmentation, multiple samples from the original dataset are created using bootstrapping, and synthetic data points are generated to enrich the training set.

Working with data containing class imbalance is another obstacle when developing a diabetes prediction model. This occurs when the data set has a skewed distribution of classes, biasing the models toward the majority class. In the field of diabetes prediction, this bias can lead to reduced performance in predicting the early stages of diabetes in patients with rare conditions. The strategies to handle Class Imbalance are Oversampling and Undersampling. The oversampling approach seeks to achieve balance in class distribution by replicating minority class instances. Undersampling is performed by reducing the number of majority class instances until the balance of class distribution is achieved.

Interpreting complex models such as deep neural networks or ensemble methods are also challenges that must be addressed, and are also crucial in healthcare settings. Interpretation techniques such as SHAP (Shapley Additive explanations) or LIME (Local Interpretable Model-agnostic Explanations) can be used to clarify predictions made by complex models, providing an understanding of the effects of features and model performance.

It is noted in current and previous studies that diabetes prediction models are trained on specific datasets, which makes their generalizability to diverse populations and healthcare settings limited. The future directions must address ensuring the robustness and generalizability of predictive models through comprehensive external validation across various populations, settings, and geographic areas.

D. Implications and Recommendations

Algorithm Selection: The choice of algorithm should align with the dataset characteristics and the problem domain. RF and SVM are robust choices for many applications, but their limitations must be managed. For instance, SVM might not be ideal for extremely large datasets due to computational constraints.

Parameter Tuning: Effective parameter tuning is critical for optimizing algorithm performance. This is particularly true for algorithms like SVM and KNN, where parameter choices significantly impact results.

Computational Resources: Consider the computational cost and resource availability, especially for ANN and LSTM, which can be resource-intensive.

Interpretable Models: When interpretability is key, simpler models like Decision Trees or Logistic Regression might be preferable despite their limitations.

VI. CONCLUSION

In this literature review, we explored various machine learning techniques for diabetes prediction across diverse datasets. The methodologies reviewed ranged from traditional algorithms like Naive Bayes and Logistic Regression to complex models such as RNNs and LSTMs. Our findings highlight that Decision Trees, Logistic Regression, SVM, Random Forests, and ANN are the most utilized methods in the field.

Decision Trees offer excellent interpretability, making them suitable for clinical applications where understanding the decision-making process is crucial.

Logistic Regression, SVM, and Random Forests are noted for their high accuracy, making them reliable choices for diabetes prediction tasks.

Advanced techniques like LSTM, CNN, and DNN show promise, particularly for sequential data and image-based diagnostics, due to their ability to capture complex patterns.

Ensemble methods and clustering algorithms further enhance prediction performance, with studies indicating that models using Random Forest and Ada Boost can achieve accuracies up to 98.60%. Models incorporating CNNs and LSTMs also show high accuracy, often exceeding 95%.

We examined various datasets, such as PIDD and CGM, highlighting the importance of feature selection in improving model performance. The use of diverse data sources, including clinical parameters and lifestyle habits from platforms like Kaggle and UCI, contributes to robust model training. However, challenges like data privacy and heterogeneity must be addressed. Collaborative efforts for data sharing and standardization are essential, with transparency and reproducibility being crucial for advancing research.

Effective preprocessing techniques, such as feature selection, missing value imputation, data normalization, encoding, dimensionality reduction, outlier detection, and data balancing, significantly enhance model efficiency and robustness. Smoothing data to reduce noise also contributes to better accuracy.

Evaluation metrics vary, with accuracy, precision, recall, F1 score, and AUC-ROC being commonly used to provide comprehensive insights into model performance in clinical settings. These metrics are crucial for assessing the effectiveness and reliability of predictive models.

Through examining the literature in this review, it was found that there is no ideal model for diabetes prediction that is suitable for all types of databases, and this conclusion is the answer to the question of this study, “Is there a promising machine learning model suitable for all diabetes databases?”. Most studies conducted for diabetes prediction trained the models on specific datasets, which makes their generalizability to diverse populations and healthcare settings limited. The future directions must address ensuring the robustness and generalizability of predictive models through comprehensive external validation across various populations, settings, and geographic areas.

VII. SUGGESTIONS FOR FUTURE RESEARCH

Refinement of Models: Future research should focus on refining existing models to improve their accuracy and robustness, particularly in diverse and real-world datasets.

Exploration of Novel Methodologies: There is a need to explore novel machine learning methodologies that can offer improved performance and interpretability.

Validation Through Clinical Trials: Models should be validated through extensive clinical trials to ensure their applicability in real-world healthcare settings. This step is crucial for developing personalized healthcare interventions.

Data Sharing and Standardization: Encouraging collaborative efforts for data sharing and establishing standards for data formats will help in building comprehensive datasets that enhance model training and evaluation.

Addressing Data Privacy: Research should focus on developing methods to ensure data privacy and security, which is vital for gaining access to more diverse and comprehensive datasets.

Improving Transparency and Reproducibility: Ensuring that studies are transparent and reproducible will build trust in machine learning models and facilitate their adoption in clinical practice.

REFERENCES

[1] R. D. H. Devi, A. Bai, and N. Nagarajan, "A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms," *Obesity Medicine*, vol. 17, p. 100152, Mar. 2020.

[2] H. A. A. Mohammed, I. Nazeeh, W. C. Alisawi, Q. K. Kadhim, and S. T. Ahmed, "Anomaly Detection in Human Disease: A Hybrid Approach Using GWO-SVM for Gene Selection," *Rev. d'Intelligence Artif.*, vol. 37, no. 4, pp. 913–919, 2023, doi: 10.18280/ria.370411.

[3] A. I. Veresiu, C. I. Bondor, B. Florea, E. J. Vinik, A. I. Vinik, and N. A. Gâvan, "Detection of undisclosed neuropathy and assessment of its impact on quality of life: a survey in 25,000 Romanian patients with diabetes," *Journal of Diabetes and Its Complications*, vol. 29, no. 5, pp. 644–649, 2015.

[4] V. A. Kumari and R. Chitra, "Classification Of Diabetes Disease Using Support Vector Machine," *International Journal of Engineering Research and Applications (IJERA)*, vol. 3, no. 2, pp. 1797-1801, 2013.

[5] Q. K. Kadhim, A. Altameemi, and S. Jasim, "Artificial Intelligence Techniques for Colon Cancer Detection: A Review," *J. Yarmouk*, vol. 21, no. 2, pp. 11–18, 2023.

[6] I. Qureshi, J. Ma, and Q. Abbas, "Recent development on detection methods for the diagnosis of diabetic retinopathy," *Symmetry*, vol. 11, no. 6, p. 749, 2019.

[7] American Diabetes Association, "Diagnosis and classification of diabetes mellitus," *Diabetes Care*, vol. 37, no. Supplement 1, pp. S81-S90, 2014.

[8] G. Gustin and B. Macq. *Diabetes management through artificial intelligence and gamification: blood glucose prediction models and mHealth design considerations*. MSc Dissertation, Catholic University of Louvain, pp. 10, 2016.

[9] American Diabetes Association, "Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes," *Diabetes Care*, vol. 42, no. Supplement 1, p. S13, 2019.

[10] R. A. Oram *et al*., "The majority of patients with long-duration type 1 diabetes are insulin microsecretors and have functioning beta cells," *Diabetologia*, vol. 57, no. 1, pp. 187-191, 2014.

[11] S. E. Inzucchi *et al*., "Management of hyperglycaemia in type 2 diabetes: a patient-centered approach. Position statement of the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD)," *Diabetologia*, vol. 55, no. 6, pp. 1577-1596, 2012.

[12] 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*, vol. 34, no. 6, pp. 3204-3225, 2020.

[13] Z. K. Maseer, Q. K. Kadhim, B. Al-Bander, R. Yusof, and A. Saif, "Meta-analysis and systematic review for anomaly network intrusion detection systems: Detection methods, dataset, validation methodology, and challenges," *IET Networks*, pp. 1–38, 2024, doi: 10.1049/ntw2.12128.

[14] S. H. Dhahi, E. H. Dhahi, S. T. Ahmed, and Q. K. Kadhim, "Predicting Parkinson's disease using filter feature selection method," in *3RD International Conference On Engineering And Science*, vol. 3104, pp. 2–10, 2024, doi: 10.1063/5.0191620.

[15] O. F. Alwan, Q. K. Kadhim, R. B. Issa, and S. T. Ahmed, "Early Detection and Segmentation of Ovarian Tumor Using Convolutional Neural Network with Ultrasound Imaging," *Rev. d'Intelligence Artif.*, vol. 37, no. 6, pp. 1503–1509, 2023, doi: 10.18280/ria.370614.

[16] D. Krotov and J. J. Hopfield, "Unsupervised learning by competing hidden units," *Proceedings of the National Academy of Sciences*, vol. 116, no. 16, pp. 7723-7731, 2019.

[17] K. Shameer *et al*., "Machine learning in cardiovascular medicine: are we there yet?," *Heart*, vol. 104, no. 14, pp. 1156-1164, 2018.

[18] N. Razali *et al*., "Analyzing Diabetic Data using Classification," *Journal of Physics: Conference Series*, vol. 1529, p. 22105, 2020.

[19] R. B. Lukmanto, A. Nugroho, and H. Akbar, "Early detection of diabetes mellitus using feature selection and fuzzy support vector machine," *Procedia Computer Science*, vol. 157, pp. 46-54, 2019.

[20] F. Maulidina *et al*., "Feature optimization using Backward Elimination and Support Vector Machines (SVM) algorithm for diabetes classification," *Journal of Physics: Conference Series*, vol. 1821, p. 012006, 2021.

[21] Ö. B. Bilge, Y. Metin, and S. E. Selin, "Classification of Diabetes Mellitus with Machine Learning Techniques," *Journal of Natural and Applied Sciences*, vol. 25, no. 1, pp. 112-120, 2021.

[22] T. M. Alam *et al*., "A model for early prediction of diabetes," *Informatics in Medicine Unlocked*, vol. 16, p. 100204, 2019.

[23] R. D. H. Devi, A. Bai, and N. Nagarajan, "A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms," *Obesity Medicine*, vol. 17, p. 100152, 2020.

[24] J. A. Jose, T. Waggoner, and S. Manikandan, "Continuous Glucose Monitoring Prediction," *Arxiv Preprint Arxiv:2101.02557*, 2021.

[25] S. Asaduzzaman *et al*., "Dataset on significant risk factors for Type 1 Diabetes: A Bangladeshi perspective," *Data in Brief*, vol. 21, pp. 700-708, 2018.

[26] V. Maan, J. Vijaywargiya, and M. Srivastava, "Diabetes Prognostication–An Aptness of Machine Learning," in *2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3)*, pp. 1-5, 2020.

[27] N. Nerkar, V. Inamdar, L. Kajrolkar, and R. Barve, "Diabetes Prediction using Neural Network," *International Research Journal of Engineering and Technology (IRJET)*, vol. 8, no. 2, pp. 330-333, Feb. 2021.

[28] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Procedia Computer Science*, vol. 165, pp. 292-299, 2019.

[29] R. Roy, A. Prasad, and S. M. Andrews, "Diabetes Prediction Using Machine Learning," *International Journal of Research Publication and Reviews*, vol. 2, no. 4, pp. 134-136, 2021.

[30] M. K. Sharma, N. Dhiman, and V. N. Mishra, "Mediative fuzzy logic of sugeno-tsk model for the diagnosis of diabetes," in *Journal of Physics: Conference Series*, vol. 1724, no. 1, p. 012028, 2021.

[31] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," in *International Conference on Computational Intelligence and Data Science (ICCIDS 2018)*, vol. 132, pp. 1578–1585, 2018.

[32] K. G. Naveen, V. Rajesh, A. A. Reddy, K. Sumedh, and T. S. Reddy, "Prediction of Diabetes Using Machine Learning Classification Algorithms," *International Journal of Scientific and Technology Research*, vol. 9, no. 1, pp. 1805–1808, Jan. 2020.

[33] H. M. Deberneh and I. Kim, "Prediction of Type 2 Diabetes Based on Machine Learning Algorithm," *International Journal of Environmental Research and Public Health*, vol. 18, no. 6, p. 3317, 2021.

[34] A. Lynam. *Developing clinical prediction models for diabetes classification and progression*. Ph.D. dissertation, University of Exeter, 2020.

[35] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Informatics in Medicine Unlocked*, vol. 10, pp. 100-107, 2018.

[36] G. Swapna, K. P. Soman, and R. Vinayakumar, "Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals," *Procedia Comput. Sci.*, vol. 132, pp. 1253–1262, 2018.

[37] O. Metsker, K. Magoev, A. Yakovlev, S. Yanishevskiy, G. Kopanitsa, S. Kovalchuk, and V. V. Krzhizhanovskaya, "Identification of risk factors for patients with diabetes: diabetic polyneuropathy case study," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, pp. 1-15, 2020.

[38] P. Prabhu and S. Selvabharathi, "Deep Belief Neural Network Model for Prediction of Diabetes Mellitus," *2019 3rd International Conference on Imaging, Signal Processing and Communication (ICISPC)*, pp. 138-142, 2019, doi: 10.1109/ICISPC.2019.8935838.

[39] T. E. Idriss, A. Idri, I. Abnane, and Z. Bakkoury, "Predicting Blood Glucose using an LSTM Neural Network," *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pp. 35-41, 2019, doi: 10.15439/2019F159.

[40] W. Song, W. Cai, J. Li, F. Jiang, and S. He, "Predicting Blood Glucose Levels with EMD and LSTM Based CGM Data," *2019 6th International Conference on Systems and Informatics (ICSAI)*, pp. 1443-1448, 2019, doi: 10.1109/ICSAI48974.2019.9010318.

[41] S. I. Ayon and M. Islam, "Diabetes prediction: a deep learning approach," *Int. J. Inf. Eng. Electron. Bus*., vol. 11, no. 2, pp. 21-27, 2019.

[42] Z. Alhassan, A. S. McGough, R. Alshammari, T. Daghstani, D. Budgen, and N. A. Moubayed, "Type-2 diabetes mellitus diagnosis from time series clinical data using deep learning models," in *Artificial Neural Networks and Machine Learning - ICANN 2018, 27th International Conference on Artificial Neural Networks*, pp. 468-478, 2018.

[43] Q. Zou, K. Y. Qu, Y. M. Luo, D. H. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Front. Genet*., vol. 9, p. 515, 2018.

[44] L. Zhang *et al*., "Nonlaboratory-based risk assessment model for type 2 diabetes mellitus screening in Chinese rural population: a joint Bagging-Boosting Model," *IEEE J. Biomed. Health Inform*., vol. 25, no. 10, pp. 4005–4016, 2021.

[45] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," *Sci. Rep*., vol. 10, no. 1, pp. 1–12, 2020.

[46] D. Pei *et al*., "Accurate and rapid screening model for potential diabetes mellitus," *BMC Med. Inform. Decis. Mak*., vol. 19, p. 41, 2019.

[47] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," *Appl. Comput. Inform*., vol. 1330, 2020.

[48] N. Abdulhadi and A. Al-Mousa, "Diabetes Detection Using Machine Learning Classification Methods," *2021 International Conference on Information Technology (ICIT)*, pp. 350-354, 2021, doi: 10.1109/ICIT52682.2021.9491788.

[49] D. V. V. Rani, D. Vasavi, and K. Kumar, "Significance of multilayer perceptron model for early detection of diabetes over ML methods," *J. Univ. Shanghai Sci. Technol*., vol. 23, no. 8, pp. 148–160, 2021.

[50] 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)*, pp. 1-4, 2019, doi: 10.1109/UBMYK48245.2019.8965556.

[51] M. T. García-Ordás, C. Benavides, J. A. Benítez-Andrades, H. Alaiz-Moretón, and I. García-Rodríguez, "Diabetes detection using deep learning techniques with oversampling and feature augmentation," *Comput. Methods Programs Biomed*., vol. 202, p. 105968, 2021.

[52] K. S. Ryu, S. W. Lee, E. Batbaatar, J. W. Lee, K. S. Choi, and H. S. Cha, "A deep learning model for estimation of patients with undiagnosed diabetes," *Appl. Sci.*, vol. 10, no. 1, p. 421, 2020.

[53] B. Kurt *et al*., "Prediction of gestational diabetes using deep learning and Bayesian optimization and traditional machine learning techniques," *Med. Biol. Eng. Comput.*, vol. 61, no. 7, pp. 1649-1660, Jul. 2023.

[54] K. K. Patro *et al*., "An effective correlation-based data modeling framework for automatic diabetes prediction using machine and deep learning techniques," *BMC Bioinformatics*, vol. 24, no. 1, p. 372, Oct. 2023.

[55] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Comput. Methods Programs Biomed.*, vol. 220, p. 106773, Jun. 2022.

[56] A. Saini, K. Guleria, and S. Sharma, "Predictive Machine Learning Techniques for Diabetes Detection: An Analytical Comparison," in *2023 2nd Edition of IEEE Delhi Section Flagship Conference (DELCON)*, pp. 1-5, 2023.

[57] 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.

[58] H. Lai *et al*., "Predictive models for diabetes mellitus using machine learning techniques," *BMC Endocrine Disorders*, vol. 19, no. 1, p. 101, 2019.

[59] R. Krishnamoorthi *et al*., "A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques," *J. Healthc. Eng*., vol. 2022, p. 1684017, Jan. 2022.

[60] A. U. Haq *et al*., "Intelligent Machine Learning Approach for Effective Recognition of Diabetes in E-Healthcare Using Clinical Data," *Sensors (Basel)*, vol. 20, no. 9, p. 2649, May 2020.

[61] E. M. Hameed and H. Joshi, "Performance comparison of machine learning techniques in prediction of diabetes risk," in *AIP Conference Proceedings*, vol. 3051, no. 1, 2024.

[62] H. A. A. Mohammed, A. A. Kasim Jizany, I. M. Mahmood, and Q. K. Kadhim, "Predicting Alzheimer's Disease Using a Modified Grey Wolf Optimizer and Support Vector Machine," *Ing. des Syst. d'Information*, vol. 29, no. 2, pp. 669–676, 2024, doi: 10.18280/isi.290228..

[63] R. Kowsar and A. Mansouri, "Multi-level analysis reveals the association between diabetes, body mass index, and HbA1c in an Iraqi population," *Scientific Reports*, vol. 12, no. 1, p. 21135, 2022.

[64] H. A. Ismael, N. H. Al-A'araji, and B. K. Shukur, "Enhanced the prediction approach of diabetes using an autoencoder with regularization and deep neural network," *Periodicals of Engineering and Natural Sciences*, vol. 10, no. 6, pp. 156-167, 2023.

[65] K. Abnoosian, R. Farnoosh, and M. H. Behzadi, "Prediction of diabetes disease using an ensemble of machine learning multi-classifier models," *BMC bioinformatics*, vol. 24, no. 1, p. 337, 2023.

[66] R. Alhalaseh, D. A. G. AL-Mashhadany, and M. Abbadi, "The Effect of Feature Selection on Diabetes Prediction Using Machine Learning," in *2023 IEEE Symposium on Computers and Communications (ISCC)*, pp. 1-7, 2023.

[67] A. A. Alhussan *et al*., "Classification of Diabetes Using Feature Selection and Hybrid Al-Biruni Earth Radius and Dipper Throated Optimization," *Diagnostics*, vol. 13, no. 12, p. 2038, 2023.

[68] P. Nuankaew, S. Chaising, and P. Temdee, "Average weighted objective distance-based method for type 2 diabetes prediction," *IEEE Access*, vol. 9, pp. 137015-137028, 2021.

[69] X. Li *et al*., "Optimized Computational Diabetes Prediction with Feature Selection Algorithms," in *Proceedings of the 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, pp. 36-43, 2023.

[70] M. R. Rajput and S. S. Khedgikar, "Diabetes prediction and analysis using medical attributes: A Machine learning approach," *Journal of Xi'an University of Architecture & Technology*, vol. 14, no. 1, pp. 98-103, 2022.

[71] G. Alix, H. Huang, A. Guergachi, K. Keshavjee, and X. Gao, "An Online Risk Tool for Predicting Type 2 Diabetes Mellitus," *Diabetology*, vol. 2, no. 3, pp. 123-129, 2021.

[72] I. Naveed, M. F. Kaleem, K. Keshavjee, and A. Guergachi, "Artificial intelligence with temporal features outperforms machine learning in predicting diabetes," *PLOS Digital Health*, vol. 2, no. 10, p. e0000354, 2023.

[73] B. C. Lethebe. *Using machine learning methods to improve chronic disease case definitions in primary care electronic medical records*. Unpublished master's thesis, University of Calgary, Calgary, Alberta, Canada, 2018.

[74] K. Kangra and J. Singh, "Comparative analysis of predictive machine learning algorithms for diabetes mellitus," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1728-1737, 2023.

[75] M. M. Nishat, F. Faisal, M. A. Mahbub, M. H. Mahbub, S. Islam, and M. A. Hoque, "Performance assessment of different machine learning algorithms in predicting diabetes mellitus," *Biosc. Biotech. Res. Comm*., vol. 14, no. 1, pp. 74-82, 2021.

[76] K. Azbeg, M. Boudhane, O. Ouchetto, and S. J. Andaloussi, "Diabetes emergency cases identification based on a statistical predictive model," *Journal of Big Data*, vol. 9, no. 1, pp. 1-25, 2022.

[77] S. Malik, S. Harous, and H. El-Sayed, "Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women," in *International Symposium on Modelling and Implementation of Complex Systems*, pp. 95-106, Sep. 2020.

[78] O. Daanouni, B. Cherradi, and A. Tmiri, "Type 2 diabetes mellitus prediction model based on machine learning approach," in *Innovations in Smart Cities Applications Edition 3: The Proceedings of the 4th International Conference on Smart City Applications 4*, pp. 454-469, 2020.

[79] A. Yaganteeswarudu, "Multi disease prediction model by using machine learning and Flask API," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pp. 1242-1246, 2020.

[80] K. Sidana, "Prediction of Diabetes using Machine Learning Algorithms," in *2023 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON)*, pp. 1-6, 2023.

[81] A. Mao and M. O. Shafiq, "On the analysis of a public dataset for diabetes," in *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*, pp. 88-93, 2018.

[82] T. Goudjerkan and M. Jayabalan, "Predicting 30-day hospital readmission for diabetes patients using multilayer perceptron," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 2, 2019.

[83] S. M. Kuriakose, P. B. Pati, and T. Singh, "Prediction of Diabetes Using Machine Learning: Analysis of 70,000 Clinical Database Patient Record," in *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1-5, 2022.

[84] R. Shakil, B. Akter, F. Faisal, T. R. Chowdhury, T. Roy, and A. Khater, "A promising prediction of diabetes using a deep learning approach," in *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 923-927, 2022.

[85] C. Neto *et al*., "Different scenarios for the prediction of hospital readmission of diabetic patients," *Journal of Medical Systems*, vol. 45, pp. 1-9, 2021.

[86] H. N. Pham *et al*., "Predicting hospital readmission patterns of diabetic patients using ensemble model and cluster analysis," in *2019 International Conference on System Science and Engineering (ICSSE)*, pp. 273-278, 2019.

[87] A. Doğru, S. Buyrukoğlu, and M. Arı, "A hybrid super ensemble learning model for the early-stage prediction of diabetes risk," *Medical & Biological Engineering & Computing*, vol. 61, no. 3, pp. 785-797, 2023.

[88] A. Hammoudeh, G. Al-Naymat, I. Ghannam, and N. Obied, "Predicting hospital readmission among diabetics using deep learning," *Procedia Computer Science*, vol. 141, pp. 484-489, 2018.

[89] G. Alfian *et al*., "Blood glucose prediction model for type 1 diabetes based on artificial neural network with time-domain features," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 4, pp. 1586-1599, 2020.

[90] T. El Idriss, A. Idri, I. Abnane, and Z. Bakkoury, "Predicting blood glucose using an LSTM neural network," in *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pp. 35-41, 2019.

[91] G. Annuzzi *et al*., "Impact of Nutritional Factors in Blood Glucose Prediction in Type 1 Diabetes Through Machine Learning," *IEEE Access*, vol. 11, pp. 17104-17115, 2023.

[92] T. Zhu, W. Wang, and M. Yu, "A novel blood glucose time series prediction framework based on a novel signal decomposition method," *Chaos, Solitons & Fractals*, vol. 164, p. 112673, 2022.

[93] X. Chen, J. Tuo, and Y. Wang, "A prediction method for blood glucose based on grey wolf optimization evolving kernel extreme learning machine," in *2019 Chinese Control Conference (CCC)*, pp. 3000-3005, 2019.

[94] B. J. Khadhim, Q. K. Kadhim, W. K. Shams, S. T. Ahmed, and W. A. Wahab Alsiadi, "Diagnose COVID-19 by using hybrid CNN-RNN for chest X-ray," *Indones. J. Electr. Eng. Comput. Sci*., vol. 29, no. 2, pp. 852–860, 2023, doi: 10.11591/ijeecs.v29.i2.pp852-860.

[95] S. A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, "Data preprocessing in predictive data mining," *The Knowledge Engineering Review*, vol. 34, p. e1, 2019.

[96] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, p. 160, 2021.

[97] E. M. Hameed, I. S. Hussein, H. G. Altameemi, and Q. K. Kadhim, "Liver Disease Detection and Prediction Using SVM Techniques," in *2022 3rd Information Technology To Enhance e-learning and Other Application (IT-ELA)*, pp. 61-66, 2022.

[98] S. Kost, O. Rheinbach, and H. Schaeben, "Using logistic regression model selection towards interpretable machine learning in mineral prospectivity modeling," *Geochemistry*, vol. 81, no. 4, p. 125826, Nov. 2021, doi: 10.1016/j.chemer.2021.125826.

[99] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *Journal of Big Data*, vol. 6, no. 1, pp. 1-19, 2019.

[100] Q. K. Kadhim, O. F. Alwan, and I. Y. Khudhair, "Deep Learning Methods to Prevent Various Cyberattacks in Cloud Environment," *Rev. d'Intelligence Artif*., vol. 38, no. 3, pp. 893–900, Jun. 2024, doi: 10.18280/ria.380316.

[101] B. Lantz. *Machine Learning with R: Expert Techniques for Predictive Modeling*. Packt publishing ltd, 2019.

[102] P. Thareja and R. S. Chhillar, "Comparative Analysis of Data Mining Algorithms for Cancer Gene Expression Data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 10, pp. 322–328, 2021, doi: 10.14569/IJACSA.2021.0121035.

[103] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2019.

[104] D. Q. K. Kadhim, R. M. Abd ul kader, A. ismaeel Altameemi, and R. jassim Mohammed, "Identification of Alzheimer's Disease Hub Genes Based on Improved HITS Algorithm," *J. Kufa Math. Comput.*, vol. 11, no. 1, pp. 25–31, Mar. 2024, doi: 10.31642/jokmc/2018/110105.

[105] T. H. Hadi, J. Kadum, Q. K. Kadhim, and S. T. Ahmed, "An Enhanced Cloud Storage Auditing Approach Using Boneh-Lynn- Shacham ' s Signature and Automatic Blocker Protocol," *Ingénierie des Systèmes d'Information*, vol. 29, no. 1, pp. 261–268, 2024, doi: 10.18280/isi.290126.

[106] Q. K. Kadhim, S. H. Dhahi, E. G. Abdulkadhim, and W. A. W. Alsiadi, "COVID-19Disease Diagnosis using Artificial Intelligence based on Gene Expression : A Review," *Sumer J. Pure Sci.*, vol. 2, no. 2, pp. 88–102, 2023.