International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Machine Learning and Deep Learning Models for Predicting the Onset of Diabetes: A Pilot Study

Reehana Sk[1]
[1]Research Scholar, School of Computer Science and Engineering,
VIT AP-University, Amaravati, Andhra Pradesh, India
email: skreehana.15@gmail.com

Dr. Siddique Ibrahim S[2*]
[*2]Assistant Professor, School of Computer Science and Engineering,
VIT AP-University, Amaravati, Andhra Pradesh, India
Correspondence author: * email: siddique.ibrahim@vitap.ac.in

Abstract— Diabetes currently one of the most significant worldwide concerns, and its prevalence is only expected to increase in the future years. In order to monitor glucose levels in the blood and set treatment protocols for diabetes, keeping a regular schedule for checking blood glucose levels is essential. The purpose of widespread adoption of digital health in recent years has been to enhance diabetic healthcare for patients, and as a result, a massive quantity of data has been collected that may be used in the ongoing management of this chronic condition. Deep learning, a relatively new kind of machine learning, is one method that has taken advantage of this trend, and its applications seem promising. In this research, we provide a thorough analysis of how deep learning has been used in the study of diabetes thus far. We conducted a comprehensive literature search and found that this method is most often used in the following settings: diabetes diagnosis, glucose control, and the identification of diabetes-related complications. We have described the most important details regarding the learning models used, the development process, the primary outcomes, and the baseline techniques for performance assessment from the 40 original research publications that we selected based on the search. In the reviewed literature, it becomes clear that several deep learning algorithms and frameworks have outperformed traditional machine learning methods to attain state-of-the-art performance on numerous problems involving diabetes. However, we point out several gaps in the existing literature, such as a dearth of readily available data and a lack of clarity in the interpretation of models. In the near future, these obstacles may be surmounted thanks to the fast advancements in deep learning methodologies which will allow for wider application of this technology in therapeutic settings.

Keywords- Diabetes, Glucose, Diabetic Complications, Deep Learning, CNN.

I. INTRODUCTION

Diabetes mellitus is a metabolic illness may be brought on by either a deficiency of insulin production or by insulin resistance and insufficient insulin synthesis. Both of these scenarios result in excessively high blood sugar levels [1]. Since it is a metabolic condition that worsens over time and impacts the patient in every aspect of their life, including their physical and emotional well-being, there is no treatment modality that can create remarkable improvements or halt the disease's progression [2]. In 2000, India had the highest number of diabetes (31.7 million); by 2011, that figure had risen to 62.4 million, and it is expected to rise to 69.9 million by 2025 [3]. India has rapid population and economic expansion are to blame for the country's high frequency. Indian's have a greater chance of acquiring diabetes than the general population does as a result of their abdominal obesity, low Body mass index (BMI), high percentage of body fat, and high insulin resistance [4]. Diabetes may lead to an assortment of signs and symptoms, including but not limited to blurred vision, weight loss, tiredness, raised appetite including dryness, disorientation, urine retention, poor healing, and continual infections.

Diabetes refers to a condition in which there is an excessive amount of sugar in the blood. The onset of difficulties associated with elevated sugar levels occurs whenever the human body is unable to produce a chemical or hormone known as insulin in sufficient quantities [5]. The literal meaning of this term is sweet urine, Healthy urine does not include sugar. The presence of sugar (or, more accurately, glucose) in the urine indicates an elevated blood glucose. concentration. As a result of the body's inability to metabolize glucose in an appropriate manner, glucose builds up in the bloodstream. Diabetes is a condition where the human body lacks the ability to process glucose properly as a result [6].

1.1 Diabetes and its Many Forms

The metabolic condition known as Diabetes mellitus (DM) may have a number of different root causes. It is distinguished by chronic hyperglycemia as well as changes in carbohydrate, lipid, and protein metabolism, all of which are brought on by a lack of insulin activity, an insulin deficit, or both. Diabetes is a condition that lasts for a long time [7]. If the condition is not successfully managed, diabetes may cause damage to the neurons and blood vessels in the lower legs, kidneys, heart, and eyes. If blood glucose levels stay high for a long period of time, there is a possibility that complications may arise. Mouth problems may take the form of gingivitis or tooth decay, for example. Diabetes may induce a disease called diabetic retinopathy, which can lead to visual loss and even blindness in

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

extreme situations. Diseases of the heart and blood vessels are collectively referred to as cardiovascular diseases or simply CVD. These conditions include heart attacks, strokes, and peripheral artery disease (insufficient blood flow to the feet and legs). Diabetes may cause a condition known as diabetic nephropathy, which is a condition in which the kidneys do not work effectively or at all. This condition is known as kidney disease [8] Diabetes mellitus type 1, diabetes type 2, and diabetes caused by pregnancy make up the three different types of diabetes.

1.1.1Type 1 diabetes (T1D)

In type 1 diabetes [9], the body is unable to generate an adequate amount of insulin. Without insulin, the cells of the body are unable to take up glucose from the circulation as a result, they are forced to depend on other sources of energy. Diabetes and its associated problems are caused when there is an excessive amount of glucose in the blood. Insulin-dependent diabetes mellitus is another name for this particular kind of diabetes (IDDM). Despite the fact that it strikes younger people, particularly adolescents and teens, it may strike anyone of any age. Injections of insulin (and, in some cases, oral medications), physical activity, strategic meal planning, and adjustments to one's lifestyle are required to achieve this delicate balance. Type 1 diabetes can cause a number of symptoms, such as a frequent need to urinate, unusually intense hunger and thirst, rapid weight loss, excessive exhaustion and fatigue, dizziness, and irritability.

1.1.2 Type 2 diabetes (T2D)

Insulin resistance may be the primary feature of type 2 diabetes [10], but secretory dysfunction may predominate in certain cases even in the absence of insulin resistance. Insulin is generated by the pancreas; however, this insulin may not be sufficient to maintain normal levels of glucose in the blood, or the cells themselves may be resistant to the insulin that is produced. The disease strikes people over the age of 40 at a higher rate than it does younger children and teens, although it is also growing more common in younger children. Drowsiness, dry, itchy skin, unusual weight gain or loss, vision problems, prickling, phantom pains, pain in the lower legs, easy exhaustion, sluggish recovery of cuts or scratches, and recurrent infections are some of the symptoms that are associated with type 1 diabetes. Type 1 diabetes is characterized by the following symptoms: (e.g., vaginal infections). In order to treat type 2 diabetes, it is required to make changes to one's diet, level of physical activity, and way of life, in addition to taking oral medications or injecting insulin as necessary [6].

1.1.3 Gestational diabetes

Gestational diabetes [11]is a condition that is identified in pregnant women who have not previously been affected by diabetes but who develop high blood glucose (sugar) levels throughout their pregnancy. It is a transitory ailment that affects two percent to four percent of all pregnant women and often goes away following the delivery of the baby. It is more common for women who have already had gestational diabetes to go on to acquire type 2 diabetes later in life. The exact cause of this kind of diabetes is not yet understood. Placental hormones aid in the development of the baby but also prevent the mother's insulin from functioning properly, leading to insulin resistance in the mother. The placenta is critical to the growth and development of the newborn. Gestational diabetes manifests itself when a pregnant woman's body is unable to produce and make full use of the insulin that is necessary for the pregnancy. The vast majority of pregnant women are completely oblivious to any and all warning signs and symptoms of gestational diabetes. Two symptoms include an increase in thirst as well as an increase in the frequency of urine. Numerous research examines the history of the development of glucose measuring devices. One of these studies also provides a review of the four generations of glucose monitoring [12], which are categorized according to the technology that was utilized evaluation of the manufactured medical devices began in the 1970s with the introduction of the first-generation glucose meters. These meters made use of reflectance technology and were cumbersome pieces of equipment that need a substantial volume of blood sample. The second generation of devices only required a little amount of blood, and because of the advancements in technology, they were able to be manufactured as more compact devices that were sold at more reasonable rates and allowed for individualized usage.

Patients with diabetes have a difficult time with finger pricking as the primary routine in these invasive methods since it may lead to scarring. This has motivated the development of devices that allow glucose measurement to be done in a manner that is both inexpensive and does not need invasive procedures. The third-generation devices began as minimally invasive devices that comprise an array of tiny needles on the skin and allowed constant glucose monitoring. Nowadays, these devices are known as Continuous glucose monitoring (CGM) systems.

Due to the fact that it is still in an early stage of development, the new generation of medical devices that have just emerged on the future may still be considered more of a potential alternative than an actual application of the devices that fall under this category of medical technology. In spite of this, designers shall refer to them as medical devices of the fourth generation since they integrate noninvasive procedures and provide an environment for remote real-time continuous monitoring. The procedures that are considered noninvasive do not involve invading the human body in any way, and they are based on a variety of methodologies, such as spectrometry or the measurement of other parameters that are connected with the glucose level [13].

In this study, we intend to discuss the existing techniques and current initiatives for noninvasive glucose monitoring [14], with an emphasis on the use of machine learning (ML)[15] and neural network (NN) approaches [16], which are employed in a lot of ongoing research to deal with estimate methods of the glucose level as shown in the Figure .1.

II. Related Work

2.1 Methods for Accessing Blood Glucose Levels

Measurements of glucose are almost often categorized according to the degree of intrusiveness required by the sensing equipment. Blood glucose monitoring may be simplified into two categories: invasive and non-invasive, with the former designating whether or not the human skin was harmed during the blood glucose test.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

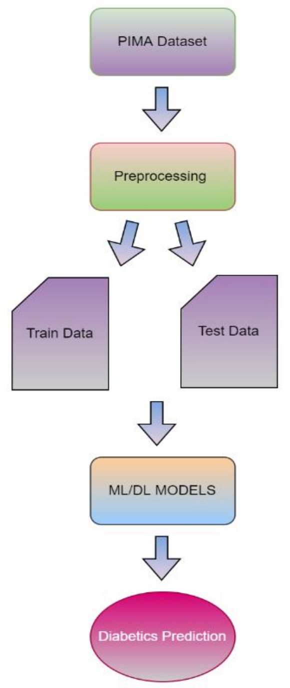

Fig.1: Flowchart of predicating Diabetics using ML/DL Models.

a) ***Invasive Blood Glucose Monitoring Current*** invasive blood glucose detection technology is robust, user-friendly, and applicable; as a result, both clinical and home glucometers use a blood sample followed by in vitro analysis to determine blood glucose levels. Accurate measurements of blood glucose concentration may be made using an automated biochemical analyzer from blood samples taken from people in hospitals first thing in the morning after they have fasted overnight. This Invasive approach [17] yields accurate findings and may be used as a significant foundation for diabetes diagnosis, However, due to the laborious procedure, lengthy detection time, and large amounts of venous blood extraction, it is not suitable for continuous diabetic monitoring Self-monitoring of blood glucose (SMBG) refers to the continuous measurement of the concentration of glucose in an individual's blood at a particular point of time [18]. This monitoring is typically carried out using an electronic glucose meter that is kept at home. Nonetheless, glucose oxidase biosensors are often used in home blood glucose monitors, which collect blood from a fingertip on a disposable paper strip and then utilize the current generated by the strip's chemical reaction to calculate blood glucose concentration. At the moment, the commercial glycemic meters produced by Roche, Sano, Omron, Johnson and Johnson, Bayer, Abbott, Echeng, Ecco, and a variety of other companies are the ones that dominate the market. The widespread use of commercial glucose meters in private settings can be attributed to the numerous benefits associated with these monitors, including their portability, low cost, user-friendliness, relatively accurate data, and the ability to perform monitoring multiple times per day. However, there is no denying the fact that it has a number of drawbacks. The patient's skin will be badly punctured since an accurate quantity of blood is collected from the fingertips prior to each test. It is becoming increasingly difficult for patients' fingertip wounds to heal in a timely manner as the number of times that blood is drawn from them increases. In addition to increasing the likelihood of infection from the outside environment, the patient will experience significant discomfort and anxiety in the hours leading up to each day's blood draw because of this. On the other hand, the test paper and the blood needle are both examples of consumables that should be thrown away after use. Families in developing nations and regions will have to spend a significant amount of money on them if they are used frequently. In addition to this, the strip does not have a particularly long shelf life, and if it is not properly stored, it can cause the blood glucose detection value to be inaccurate**.** In accordance with the recommendations provided by healthcare guidelines, SMBG should really be performed four times per day, but this number should increase to ten times per day in the event of illness or inadequate control [19]. However, according to some reports, approximately forty percent to fifty percent of diabetics don't really adhere to the recommendations. If an induvial do not inject insulin at the appropriate time, you risk of developing serious complications related to diabetes, including Diabetic ketoacidosis (DKA), cardiovascular diseases, blindness, stroke, and neurasthenia. In addition, receiving an excessive amount of insulin could result in a sudden and significant drop in blood glucose, which could have serious consequences including seizures, coma, and even death. It is abundantly clear that the existing SMBG method causes patients to experience discomfort and that it should be modified.

b) ***Non-Invasive Blood Glucose Monitoring*** The term non-invasive blood glucose monitoring is used to describe the process of measuring blood sugar levels in humans without causing any tissue damage [20]. Many different non-invasive methods exist for measuring blood glucose, but they can be categorized into three broad categories: optical, microwave, and electrochemical. Near-infrared reflectance spectroscopy (NIRS), polarized optical rotation, Raman spectroscopy, fluorescence, Optical coherence tomography (OCT), and so on are all examples of optical techniques [21]. Glucose is not unique to human blood; it is present in high concentrations in a variety of other biofluids as well. The electrochemical approach makes use of the consistent association between biofluids and blood glucose value by first measuring the glucose content in body fluids and then indirectly deriving the blood glucose value after algorithm or data model calibration. ISF glucose levels are most similar to blood glucose levels in both normal and diabetic persons, giving a theoretical framework for the development of an ISF glucose biosensor [22]. Despite this, many studies have shown that ISF glucose lags behind blood glucose by a few minutes (between 4 and 10 minutes) when it comes to reflecting changes in blood glucose levels. The goal of rapid extraction of ISF can be attained through the use of Reverse iontophoresis (RI) technology, which is commonly used in transdermal bio fluid extraction. Specifically, the approach and the material are described in the following sections. Diabetics, who must inject insulin on a daily basis to maintain a healthy blood glucose balance, must monitor their blood glucose levels often and routinely. As a result, the novel painless and stress-free non-invasive blood glucose monitoring technology and precise closed-loop medicine administration system has the potential to immediately assist hundreds of millions of patients by eliminating the need for painful blood collection.

2.2 LITERATURE REVIEW

Technique	Dataset	Reference	Methodology	Result	Limitations
ML	PID	Yuvaraj et.al (2020)[23]	Hadoop cluster based distributed computing framework	Accuracy-88% Precision-87 Recall-77% f-measure-82%	Efficient for high volume of data
		Kumari et.al (2021)[24]	Ensemble soft voting classifier	Accuracy-79.04% Precision-73.48% Recall-71.45% f-measure-80.6%	Less accurate compared to basic ML algorithms
		Naz et.al (2020)[25]	Multilayer feed-forward perceptron based model	Accuracy-90.34% Precision-88.05% Recall-83.09% F measure-85.98% Specificity-91.43% Sensitivity-88.06%	No novel algorithm used
		Kalagotla et.al (2021)[26]	Stacking of MLP,SVM,LR	Accuracy-78.2% Precision-72.2% Recall-51% F-measure-59.4%	System complexity is high
		Nadesh et.al(2020)[27]	FI+DNN	Accuracy-96.77 Sensitivity-96 Specificity-98.5	High computational time
		Hasan et.al (2020)[28]	Ensemble ML classifiers	Accuracy-65% Sensitivity-78.9 Specificity-93.4	Ensemble is expensive in terms of time and space.
		Krishnamoorthi et.al (2022)[29]	IDMPF	Accuracy-90%	Only structured data is considered.
	EMR of Luzhou Municipal Health Commission data, China	Yang et.al(2021)[30]	XGBoost	Specificity-76.55% Recall-68.48% Accuracy-77.45% Precision-26.37% F1-37.66%	Feature selection are based upon the physical examinations.
	NHANES dataset	Maniruzzaman et.al (2020)[31]	Random Forest	Accuracy-92.54% AUC-0.91 Sensitivity-99.56% f-measure-96.30%	Usage of highly imbalanced dataset. For validation Indian liver dataset is used.
		Nadeem et.al (2021)[32]	Fusion of SVM and ANN	Accuracy-94.67% Specificity-97.32% Sensitivity-89.23% Precision-94.19%	High computational cost
	Hospital Frankfurt Germ	Haq et.al(2020)[33]	Filter method Decision Tree+a	Accuracy-98.2 Sensitivity-98 Specificity-97 Precision-99.8 Recall-98	Not a robust model.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

	any. Diabetes Data Set		da boost+random forest-ML	F1-98.6	
	Own dataset taken from Fitbit Versa 2 device	Adams et.al (2021)[34]	QDA,SVM	Accuracy-72.3,72.6 Precision-68.9,69.0 Recall-74.4,74.2	Potential errors and uncertainties from the wearable device from actions such as device drift
DL	PID	Ramezani el.al (2018)[35]	LANFIS	Accuracy-88.05%	Removing little patient data due to more missing qualities
		Kamalraj et.al (2021)[36]	IF-CNN+PD-SS	Accuracy-96.26% Precison-94.78% Recall-95.18% F-score-94.08	Magnitude of dataset and missing feature values
		Naveena et.al (2022)[37]	MF-CSA	Accuracy-96.4% Precision-98.4% F1-score-98.1% Sensitivity-97.8% Specificity-48.7%	The running time is not minimized.
		Zhou et.al (2020)[38]	DLPD	Correctly classified-763 Incorrectly classified-5	Past gradient accumulation is restricted..
	Augumentaed datset for 7 databases	Akpado et.al(2021)[39]	Mamdani fuzzy inference system	Accuracy-97% Specificity-95% Sensitivity-94% Precision-93%	Major limitation of SPSS (to fill missing data) cannot be used to analyse a very large dataset.
	MIMIC-III	Srinivasan et.al(2021)[40]	CNN based classifier	Accuracy-76.34% sensitivity of 76.66% specificity of 76.11%.	Dataset need to be appended to reach our model accuracy to FPG test level.
	AIDA	Asad et.al (2021)[41]	FF-NN, NAR-NN	RMSE-0.998 ml/dl, 0.606 ml/dl	Models are trained on 2 patients data only.
	Ohio T1DM	Martinsson et.al (2020)[42]	RNN	RMSE-31.403 SE-0.316	Distinguishing intra-patient variance from sensor faults is unlikely.
	Own dataset by observing 25 diabetic people	Rodríguez et.al (2019)[43]	SISAL	Symmetry among the features.	Values of insulin dosage and carbohydrate are recorded manually by monitoring people.

Ramezani et al. [35] proposed LANFIS is adaptive network works on fuzzy logic system with logistic regression. Previously existing systems(ANFIS) gave continuous outputs, this problem has been resolved by LANFIS by using logistic regression and produces the binary output. The main focus of the work is on handling the missing data and reducing the high dimensionality of clinical data present in the training dataset using multiple imputations and orthogonal transformations, this produces more accurate results in the prediction of diabetes. While handing the missing values in the Pima Indian diabetic dataset 15 patients data has been eliminated because more attribute values are missed.

Akpado et al. [39] suggested a fuzzy logic approach for spotting type 2 diabetes. This framework contains four stages: fuzzification, rule assessment, aggregation, and difuzzification. An expanded data set culled from seven sources and combined with information from a single database maintained by the Federal medical institution in Nigeria. The missing data in the expanded dataset is imputed using a multiple imputation approach. When compared to the Gizem Koca system, this fuzzy inference system performed at a higher level, with a 95% specificity, 94% sensitivity, and 93% accuracy.

Srinivasan et al. [40] proposed a CNN-based classifier to track diabetes. Before moving further with medical examinations, this method is helpful to the physicians. The MIMIC-III data is sent to the deep learning model as an input. This model solves the issues that arose when the machine learning models were being developed. The model has a sensitivity of 76.66% while also having a specificity of 76.11%.

Early diabetes predictions on PIMA Indian diabetes datasets were proposed by Kamalraj et al. [36] using a CNN prediction model and a Pet Dog-Smell Sensing (PD-SS) method. This model uses an Interpretable Filter before delivering the pre-processed PIMA dataset to the CNN model. The seven neurons in this CNN stand for the seven attributes in the data set. The CNN's output is sent into the auto encoder, where it is reduced in dimension. Productive capabilities are ultimately selected using the PD-SS learning paradigm.

The exponential growth of healthcare systems throughout the world has resulted in a deluge of data. Nowadays, the accuracy of the model is of equal importance to the issue of dealing with massive datasets for computation. Hadoop, a distributed computing platform built on top of clusters, facilitates the processing and storage of massive information in the cloud. When it comes to predicting cases of diabetes, Yuvaraj et al. [23] suggest an unique method that uses machine learning algorithms deployed on Hadoop-based clusters.

The ensemble soft voting classifier presented by Kumari et al. [24] combines the strengths of logistic regression, naive bays, and random forest in order to classify and predict cases of diabetes mellitus. In addition to the traditional computational methods, soft voting makes use of a voting classifier. Each of the underlying models is equipped with its own unique categorization probability. Using a voting aggregator and a soft voting approach, each model calculates its own forecast, and then the majority voting is calculated, yielding the final prediction. To differentiate diabetics from healthy individuals, Yang et al. [30] proposed a model based on eXtreme Gradient Boosting (XGBoost). The author combined information from three distinct areas of a physical examination into a single computer model: patient demographics, vital signs, and laboratory results. These combined characteristics may be all that's needed to determine whether someone is at risk for developing diabetes. After the aforementioned feature selection strategies, XGBoost was utilized as a basic classifier to provide the best features, and a diabetes risk scorecard was constructed using logistic regression.

The authors Naz et al. [25] proposed a method for predicting diabetes by employing a variety of machine learning algorithms with the PIMA dataset. Prior to application of our data mining algorithms: decision tree, Naive Bayes, artificial neural network, and deep learning data cleaning for the PIMA dataset is performed. Considering all the machine learning methods, deep learning shows the best outcome.

A comparative analysis of traditional ML-based classifiers and those using features extracted using Logistic Regression was proposed by Maniruzzaman et al. [31] using four different classifiers, including naive Bayes (NB), decision tree (DT), Ada Boost (AB), and random forest (RF), to identify potential diabetics. These methods have been implemented and replicated over 20 trials, using three different partition protocols K2, K5, and K10. The data for this research comes from the National Health and Nutrition Examination Survey, with 657 people with diabetes and 5904 without, the dataset has a total of 6561 participants. High-risk indicators may be identified by LR feature extraction. Overall, an ML-based system has a 90.62% accuracy.92.54% ACC is achieved by combining LR-based feature selection with an RF-based classifier.

E-healthcare for diabetes diagnosis using Machine Learning data mining was suggested by Haq et al. [33] The Decision Tree (ID3) technique is utilized to choose relevant characteristics in this approach. In addition, DT ensemble learning, the feature selection methods Ada Boost and Random Forest are used. We further validated the prediction model using a variety of cross-validation methods, including the hold out, K-fold, and LOSO. According to the results of the statistical analysis, the suggested technique outperforms the previously presented methods in terms of accuracy.

Asad et al. [41] worked using data from continuous glucose monitoring (CGM) to forecast blood glucose levels. Both an optimum feed forward neural network (FF-NN) and an optimal nonlinear auto regressive neural network (NAR-NN) are constructed for the CGM system. Two patient’s cases are taken from AIDA for study, and these systems validated on UCI datasets- Abalone and Servo. Minimal inputs in the prediction horizon (PH) for the model set to 15-30 minutes and root mean square (RMSE) generated for the PH have been found to improve outcomes on validation datasets. When compared to the FF-NN model, NAR-NN fares quite well.

The CGM system developed by Martinsson et al. [42] may foretell blood glucose levels up to an hour in advance. The Ohio T1DM diabetes dataset is utilized to create a Gaussian-Distributed RNN model using patients' historical glucose levels as inputs. There is no need for costly data pre-processing with this methodology. In addition to root-mean-squared error, the glucose specific metric of Surveillance error grid (SEG) is used to evaluate effectiveness of the model.

Adams et al. [34] compared the accuracy of QDA and SVM in classifying glucose levels using data collected from a person with type 1 diabetes's wearable sensors. The numbers are read from the Fitbit Versa 2 and then utilized to make a determination. Individuals with type 1 diabetes had 60 days' worth of data collected throughout the summer. Daily cumulative values of each characteristic supplied by a Fitbit device used for categorization include calories(kcal), steps(steps), miles(miles), minutes(sedentary), and activity calories (activity calories).

A Sequential Input Selection Algorithm (SISAL), proposed by Rodrguez et al. [43] is used to determine the degree of similarity between various characteristics and how they impact glucose amounts in the blood. Twenty-five persons with diabetes participate in this study by having their insulin levels, blood sugar levels, carbohydrates consumed, exercise, heart rate, sleep, and routine monitored. Using SISAL's Z-score test evaluation and related lag, an optimal value is determined for each patient. This value takes into consideration the significance conveyed by the p-value. The more important a characteristic is, the smaller its p-value. Each patient's corresponding time stamp is considered, yielding just a probability value (p-value). Ultimately, when compared to exercise, heart rate, sleep, and routine, factors like insulin, food, and glycaemia have a more substantial impact.

To improve classification on the PIMA Indian diabetes dataset, Kalagotla et al. [26] introduced a new stacking strategy. The author first implemented the correlation technique for feature selection and then classification is done by applying Ada Boost on those correlated features. At last multi-layer perceptron, support vector machine and logistic regression are stacked. Using the PIMA Indian diabetes dataset, the Stacking ensemble classifier outperformed the other ML techniques that were claimed to have done well.

Naveena et al. [37] proposed a new approach for identifying diabetics and predicting their blood sugar levels. The Moth-Flame Optimization (MFO) and Crowd Search Algorithm

(CSA) work together called MFCSA to choose features optimally using a deep learning models such as CNN. To achieve lowest correlation between the features and prevent duplicate information, the hybrid MF-CSA is utilized to optimize the number of hidden neurons in two convolutional layers of CNN. The PIMA public-use dataset was used for this analysis.

Unsupervised learning using a Deep Neural Network (DNN) classifier was suggested by Nadesh et al. [27] for making precise predictions on the Pima Indian Diabetes dataset, while a Feature Importance model packed with Extra Trees and Random Forest(RF) was utilized to narrow down candidate features. For important characteristics, Randomized Trees is a great option to consider. It uses bagged DT like Extra Trees and RF. A DNN classifier is a multi-layered, neural network model. Because of this, the suggested model has a very high time complexity.

Zhou et al. [38] suggested a novel DLPD (Deep Learning Predicting Diabetes) model for diabetes type prediction. This model not only predicts who will get diabetes in the future, but it also specifies the kind of diabetes. For the sake of long-term data storage, a normalization layer is added to the model. The model's dropout layer prevents overfitting. The training loss function is the binary cross entropy. The PIMA dataset was used for this study because it achieved a training accuracy of 99.4 percent.

For diabetes prediction, Nadeem et al. [32] presented a fusion of the support vector machine (SVM) and artificial neural network (ANN) methods of machine learning. Dataset is based on the National Health and Nutrition Examination Survey (NHANES). The dataset undergoes minimal preparation, including the addition of missing values and normalization. The fusion rules are implemented using a posterior probability method. As compared to simple machine learning classifiers, the system performed better after using fused architecture.

With the goal of improving classification accuracy, Hasan et al. [28] primarily concentrated on preprocessing the PIMA dataset and then using ensemble machine learning classifiers. The pre-processing phase involves a number of steps, including as removing outliers, normalizing the data, and performing a cross-validation test, filling missing values. For a more central tendency, utilizes mean values to fill in missing data rather than median. Soft weighted voting is used to choose the optimal ensemble of machine learning classifiers. When combined, Ada Boost and XGBoost provide the highest accuracy of any known combination on PIMA dataset.

In the research on the development of diabetes during pregnancy, Krishnamurthy et al. [29] suggested an original intelligent diabetes mellitus prediction framework (IDMPF). Models from the LR, RF, SVM, and KNN families of machine learning were applied to the PIMA dataset. The dataset undergoes preprocessing in order to deal with inconsistent data before the IDMP framework is implemented. Grid search and random search are used to fine-tune hyper parameters afterwards. At long last, retrieved hyper parameters are used in ML models for categorization.

Aliberti et al. [44] proposed two models for continuous glucose monitoring systems(CGMS) for the purpose of anticipating glucose levels. The performance of these models is tested using RT-CGM, a dataset. Pre-processing the dataset by removing sequences with missing data or more than thirty consecutive samples improved prediction accuracy. The first NAR model is a modification of the classic linear autoregressive model, which it extends beyond by avoiding the inherent limitations of the linearity assumption and fixing the stability issues of earlier formulations. The second type, Long Short-Term Memory (LSTM), is excellent for processing time series data. It solves the vanishing gradients problem and other issues that have plagued other models.

2.3 Datasets

PID: The Pima Indian Dataset (PID) was utilized for the analysis in this paper. Material is compiled from the machine learning repository at UCI. The National Institute of Diabetes and Digestive and Kidney Diseases was the original source of this dataset. The PID dataset has a single output class with a binary value indicating whether or not a person has diabetes, and it contains eight characteristics. In addition, there are a total of 768 cases, 500 of which do not have diabetes and the remaining 268 who have. As PIMA is widely used to compare the results of different approaches, it was selected as the benchmark dataset for this investigation.

2.4 PERFORMANCE EVALUATION METRICS

The effectiveness of these investigations has been evaluated based on the following four outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

The following metrics will then be used to assess the state of art models.

Accuracy: It measures the overall percentage of accurate predictions made by the model and may be assessed as a proportion between the number of right predictions and the overall number of test cases made by any model in the following manner.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Recall: Recall is the proportion of sample data that correctly classifies “positive class' out of the overall samples tested for that class. It is also referred as the true positive rate (TPR).

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: The F1-score takes into consideration both accuracy and your recall, and it may be stated as follows:

$$F1 - Score = 2 * (\frac{precision * recall}{precision + recall})$$

III. Limitations and Challenges

Despite the fact that AI, machine learning, and deep learning have advanced the state of the art in several aspects of diabetes, [45] the applications of these technologies in healthcare practices need to be reliable, reliable, and convincing in order to prevent safety hazards and provide effective therapeutic aids. In this regard, there are still a multitude of limitations and obstacles that need to be overcome before computer-aided systems may be further implemented in real clinical settings. In real-world circumstances, human error and sensor artefacts are likely to contaminate the data acquired from persons with diabetes. In certain cases, it might be time consuming and costly to get authentic data. Sharing data sets across different research teams might be challenging at times as a result of rules governing data privacy. Because of these constraints, the quantity of data used in many research is sometimes decreased, and the results might be considered inconclusive. A further difficulty that comes as a result of the intricacy of glucose fluctuations is determining how to analyses the data that is available in order to describe persons who have diabetes. Moreover, the models of deep learning are not transparent. It is necessary for physicians to understand, from their point of view, why the models create the output that they do for a given input situation, especially when it comes to specific applications that involve crucial decision-making. The intricate structures that are included in DNN layers allow for successful learning of patterns from non-linear data, but at the expense of the model's capacity to be interpreted. Consequently, while looking at deep learning for diabetes, it is essential to weigh the trade-off between effectiveness and accuracy. In conclusion, it is anticipated that future advancements in both algorithmic and hardware practices will result in an increase in the effectiveness of the training of deep learning models [46].

IV. OPPORTUNITIES AND FUTURE WORK

In the outset, multi-modal solutions including wearables and smartphone assistance are progressively collecting digital data and vital signs. The vast majority of these records are stored in streamlined, easy-to-access databases on the cloud. Data quantities and variety of data sources are predicted to rapidly rise in many healthcare application areas, especially in diabetes treatment, with the widespread use of the Internet of Things and the advent of 5 G networks. More data means more opportunities to identify and eliminate low-quality data samples from training sets, and better wearables (like CGM) mean fewer measurement mistakes. To deal with this influx of data, deep learning performs well. Section IV describes certain datasets that are accessible to the public, and additional datasets will be made available to the communities following appropriate post-processing and de-anonymization. Tools like TensorFlow Lite [47], [48] make it simple to adapt the frameworks discussed in Part II for use on mobile devices, allowing for rapid deployment of deep learning in a mobile scenario.

Rather of relying exclusively on models that are driven by data, including the expert knowledge into the learning process may be beneficial in assisting with a better understanding of the underlying processes of a health condition such as diabetes. In particular, there are two different approaches that are viable options. The first approach is to integrate the physiological characteristics as an input feature of the systems, and the second is to employ technical expertise as a guidance throughout the training process. Both approaches are discussed more below. In order to draught safety constraints and compute the confidence of the model outputs, expert knowledge is also needed. Rather of relying exclusively on models that are driven by data, including the expert knowledge into the learning process may be beneficial in assisting with a better understanding of the underlying processes of a health condition such as diabetes. In particular, there are two different approaches that are viable options. The first approach is to integrate the physiological characteristics as an input feature of the systems, and the second is to employ technical expertise as a guidance throughout the training process. Both approaches are discussed more below. In order to draught safety constraints and compute the confidence of the model outputs, expert knowledge is also needed.

Studies from several of the publications acknowledged the need for further validation in real-world settings. A group at Google made some progress in this direction. Human-centered research using deep learning to treat diabetic eye problems was undertaken across 11 clinics. According to the findings, various societal and environmental considerations must be met before such automated systems may be widely used.

V. RESULTS

Figure.2,3,4 displays the results of many different methods for predicting diabetes. The objective of diabetes prediction is shown in Figure by comparing the results of various algorithms published by different authors on a variety of characteristics. These parameters may include the data set utilized, the features extracted, the classifier used, and the results produced. From existing literature, it is observed that ML and DL are widely used algorithms for detection of diseases. Both algorithms offer the better accuracy as compare to other algorithms. Artificial Neural network is also very useful for prediction. It also shows the maximum output but it takes more time as compared to other algorithms. They also show enhanced accuracy when it responded correctly to the attributes of data set.

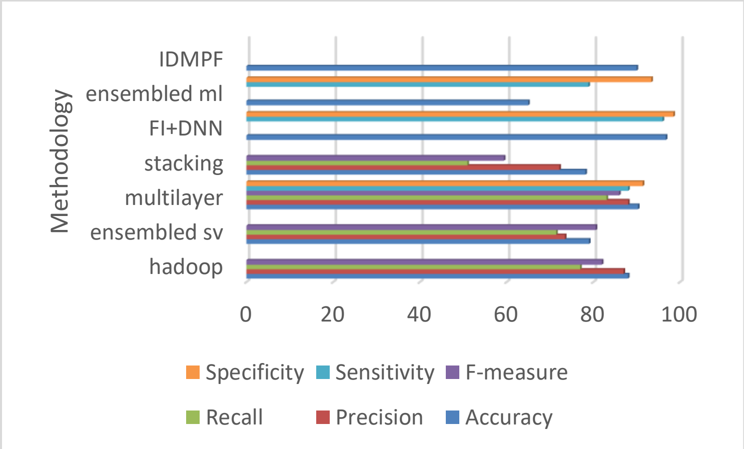

Figure.2: Comparison of Performance measures of various pervious ML techniques on PIMA dataset.

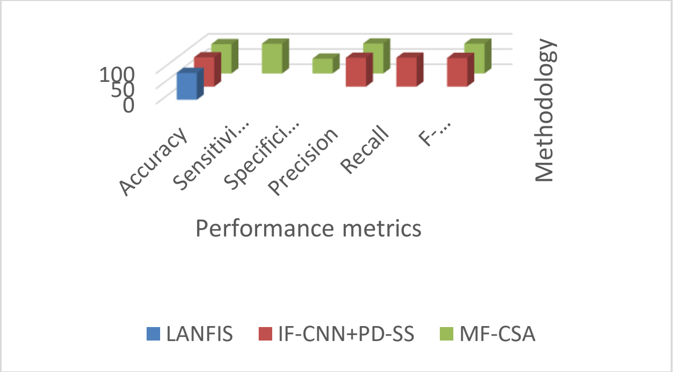

Figure.3: Comparison analysis of Performance metrics of various pervious DL techniques on PIMA dataset.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

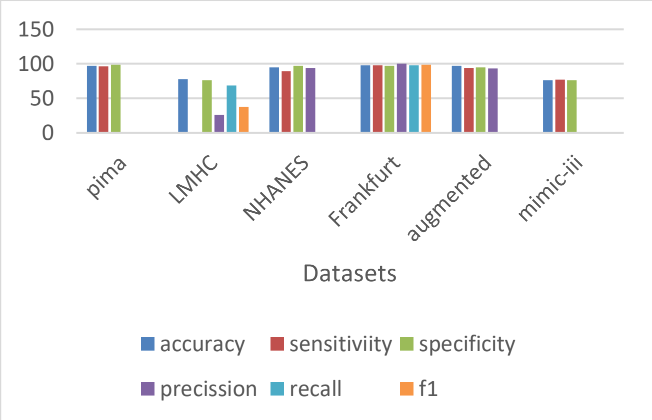

Figure.4: Comparison of Performance measures of various current techniques on existing datasets.

VI. CONCLUSION

In this work, we provide a detailed analysis of the recent advancements in the application of deep learning techniques to the study of diabetes. We conducted a comprehensive literature review, selected a set of relevant publications, and synthesized their major findings into three categories: diabetes diagnosis, glucose control, and consequences of diabetes. Experiments conducted in these domains using a variety of DNN architectures and learning methodologies have shown results that are superior to those achieved using traditional machine learning methods. Nevertheless, the literature has pointed out a few obstacles, such as data availability, feature processing, and model interpretability. The future holds enormous promise for addressing these issues by using cutting-edge deep learning technology to large, multi-modal datasets related to diabetes care. In the near future, we anticipate that deep learning technologies will become more commonplace in clinical settings, significantly enhancing the care provided to persons with diabetes.

REFERENCES

[1] Cole, J.B. and Florez, J.C., 2020. Genetics of diabetes mellitus and diabetes complications. Nature reviews nephrology, 16(7), pp.377-390.

[2] Mauricio, D., Alonso, N. and Gratacòs, M., 2020. Chronic diabetes complications: the need to move beyond classical concepts. Trends in Endocrinology & Metabolism, 31(4), pp.287-295.

[3] Gregory, G.A., Robinson, T.I., Linklater, S.E., Wang, F., Colagiuri, S., de Beaufort, C., Donaghue, K.C., Magliano, D.J., Maniam, J., Orchard, T.J. and Rai, P., 2022. Global incidence, prevalence, and mortality of type 1 diabetes in 2021 with projection to 2040: a modelling study. *The Lancet Diabetes & Endocrinology*, *10*(10), pp.741-760.

[4] Malone, J.I. and Hansen, B.C., 2019. Does obesity cause type 2 diabetes mellitus (T2DM)? Or is it the opposite?. Pediatric diabetes, 20(1), pp.5-9.

[5] Diabetics:https://www.who.int/news-room/fact-sheets/detail/diabetes,(accessed on 14 th march).

[6] Diabetes symptoms: When diabetes symptoms are a concern, https://www.mayoclinic.org/diseases-conditions/diabetes/in-depth/diabetes-symptoms/art-20044248,(accessed on 14 th march).

[7] Diabetes long-term effects, https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/diabetes-long-term-effects,(accessed on 14 th march).

[8] Diabetic Kidney Disease, https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/diabetic-kidney-disease,(accessed on 14th march).

[9] Norris, J.M., Johnson, R.K. and Stene, L.C., 2020. Type 1 diabetes—early life origins and changing epidemiology. The lancet Diabetes & endocrinology, 8(3), pp.226-238.

[10] Pearson, E.R., 2019. Type 2 diabetes: a multifaceted disease. Diabetologia, 62(7), pp.1107-1112.

[11] Choudhury, A.A. and Rajeswari, V.D., 2021. Gestational diabetes mellitus-A metabolic and reproductive disorder. Biomedicine & Pharmacotherapy, 143, p.112183.

[12] Freckmann, G., Pleus, S., Grady, M., Setford, S. and Levy, B., 2019. Measures of accuracy for continuous glucose monitoring and blood glucose monitoring devices. Journal of diabetes science and technology, 13(3), pp.575-583.

[13] Jain, P., Joshi, A.M. and Mohanty, S.P., 2019. iGLU: An intelligent device for accurate noninvasive blood glucose-level monitoring in smart healthcare. IEEE Consumer Electronics Magazine, 9(1), pp.35-42.

[14] Bolla, A.S. and Priefer, R., 2020. Blood glucose monitoring-an overview of current and future non-invasive devices. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(5), pp.739-751.

[15] Hasan, M.K., Alam, M.A., Das, D., Hossain, E. and Hasan, M., 2020. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, pp.76516-76531.

[16] Deng, Y., Lu, L., Aponte, L., Angelidi, A.M., Novak, V., Karniadakis, G.E. and Mantzoros, C.S., 2021. Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients. NPJ Digital Medicine, 4(1), p.109.

[17] Sempionatto, J.R., Moon, J.M. and Wang, J., 2021. Touch-based fingertip blood-free reliable glucose monitoring: Personalized data processing for predicting blood glucose concentrations. ACS sensors, 6(5), pp.1875-1883

[18] Ardilouze, A., Bouchard, P., Hivert, M.F., Simard, C., Allard, C., Garant, M.P., Ménard, J., Ouellet, A., Houde, G., Pesant, M.H. and Baillargeon, J.P., 2019. Self-monitoring of blood glucose: a complementary method beyond the oral glucose tolerance test to identify hyperglycemia during pregnancy. Canadian journal of diabetes, 43(8), pp.627-635.

[19] Alsunaidi, B., Althobaiti, M., Tamal, M., Albaker, W. and Al-Naib, I., 2021. A review of non-invasive optical systems for continuous blood glucose monitoring. Sensors, 21(20), p.6820.

[20] Jin, Y., Yin, Y., Li, C., Liu, H. and Shi, J., 2022. Non-invasive monitoring of human health by photoacoustic spectroscopy. Sensors, 22(3), p.1155.

[21] Delbeck, S., Vahlsing, T., Leonhardt, S., Steiner, G. and Heise, H.M., 2019. Non-invasive monitoring of blood glucose using optical methods for skin spectroscopy—Opportunities and recent advances. Analytical and bioanalytical chemistry, 411, pp.63-77.

[22] Zhu, B., Li, X., Zhou, L. and Su, B., 2022. An overview of wearable and implantable electrochemical glucose sensors. Electroanalysis, 34(2), pp.237-245.

[23] Yuvaraj, N. and SriPreethaa, K.R., 2019. Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. *Cluster Computing*, *22*(Suppl 1), pp.1-9.
[24] Kumari, S., Kumar, D. and Mittal, M., 2021. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. International Journal of Cognitive Computing in Engineering, 2, pp.40-46.
[25] Naz, H. and Ahuja, S., 2020. Deep learning approach for diabetes prediction using PIMA Indian dataset. Journal of Diabetes & Metabolic Disorders, 19, pp.391-403.
[26] Kalagotla, S.K., Gangashetty, S.V. and Giridhar, K., 2021. A novel stacking technique for prediction of diabetes. *Computers in Biology and Medicine*, *135*, p.104554.
[27] Nadesh, R.K. and Arivuselvan, K., 2020. Type 2: diabetes mellitus prediction using deep neural networks classifier. *International Journal of Cognitive Computing in Engineering*, *1*, pp.55-61.
[28] Hasan, M.K., Alam, M.A., Das, D., Hossain, E. and Hasan, M., 2020. Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, *8*, pp.76516-76531.
[29] Krishnamoorthi, R., Joshi, S., Almarzouki, H.Z., Shukla, P.K., Rizwan, A., Kalpana, C. and Tiwari, B., 2022. A novel diabetes healthcare disease prediction framework using machine learning techniques. *Journal of Healthcare Engineering*, *2022*.
[30] Yang, H., Luo, Y., Ren, X., Wu, M., He, X., Peng, B., Deng, K., Yan, D., Tang, H. and Lin, H., 2021. Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators. Information Fusion, 75, pp.140-149.
[31] Maniruzzaman, M., Rahman, M.J., Ahammed, B. and Abedin, M.M., 2020. Classification and prediction of diabetes disease using machine learning paradigm. Health information science and systems, 8, pp.1-14.
[32] Nadeem, M.W., Goh, H.G., Ponnusamy, V., Andonovic, I., Khan, M.A. and Hussain, M., 2021, October. A fusion-based machine learning approach for the prediction of the onset of diabetes. In *Healthcare* (Vol. 9, No. 10, p. 1393). MDPI.
[33] Haq, A.U., Li, J.P., Khan, J., Memon, M.H., Nazir, S., Ahmad, S., Khan, G.A. and Ali, A., 2020. Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data. Sensors, 20(9), p.2649.
[34] Adams, D. and Nsugbe, E., 2021. Predictive Glucose Monitoring for People with Diabetes Using Wearable Sensors. *Engineering Proceedings*, *10*(1), p.20.
[35] Ramezani, R., Maadi, M. and Khatami, S.M., 2018. A novel hybrid intelligent system with missing value imputation for diabetes diagnosis. *Alexandria engineering journal*, *57*(3), pp.1883-1891.
[36] Kamalraj, R., Neelakandan, S., Kumar, M.R., Rao, V.C.S., Anand, R. and Singh, H., 2021. Interpretable filter based convolutional neural network (IF-CNN) for glucose prediction and classification using PD-SS algorithm. *Measurement*, *183*, p.109804.
[37] Naveena, S. and Bharathi, A., 2022. A new design of diabetes detection and glucose level prediction using moth flame-based crow search deep learning. *Biomedical Signal Processing and Control*, *77*, p.103748.
[38] Zhou, H., Myrzashova, R. and Zheng, R., 2020. Diabetes prediction model based on an enhanced deep neural network. *EURASIP Journal on Wireless Communications and Networking*, *2020*, pp.1-13.
[39] Akpado, K.A., Njonu, U.J., Obioma, P.C. and Isizoh, A.N., An Improved Method for Predicting Diabetes Mellitus Using Adaptive Neuro-Fuzzy Inference System.
[40] Srinivasan, V.B. and Foroozan, F., 2021, August. Deep Learning based non-invasive diabetes predictor using Photoplethysmography signals. In *2021 29th European Signal Processing Conference (EUSIPCO)* (pp. 1256-1260). IEEE.
[41] Asad, M., Qamar, U. and Abbas, M., 2021. Blood Glucose Level Prediction of Diabetic Type 1 Patients Using Nonlinear Autoregressive Neural Networks. *Journal of Healthcare Engineering*, *2021*, pp.1-7.
[42] Martinsson, J., Schliep, A., Eliasson, B. and Mogren, O., 2020. Blood glucose prediction with variance estimation using recurrent neural networks. *Journal of Healthcare Informatics Research*, *4*, pp.1-18.
[43] Rodríguez-Rodríguez, I., Rodríguez, J.V., González-Vidal, A. and Zamora, M.Á., 2019. Feature selection for blood glucose level prediction in type 1 diabetes mellitus by using the sequential input selection algorithm (SISAL). *Symmetry*, *11*(9), p.1164.
[44] Aliberti, A., Pupillo, I., Terna, S., Macii, E., Di Cataldo, S., Patti, E. and Acquaviva, A., 2019. A multi-patient data-driven approach to blood glucose prediction. *IEEE Access*, *7*, pp.69311-69325.
[45] Zednik, C. and Boelsen, H., 2022. Scientific exploration and explainable artificial intelligence. Minds and Machines, 32(1), pp.219-239.
[46] Sahoo, A.K., Pradhan, C. and Das, H., 2020. Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making. Nature inspired computing for data science, pp.201-212.
[47] David, R., Duke, J., Jain, A., Janapa Reddi, V., Jeffries, N., Li, J., Kreeger, N., Nappier, I., Natraj, M., Wang, T. and Warden, P., 2021. Tensorflow lite micro: Embedded machine learning for tinyml systems. Proceedings of Machine Learning and Systems, 3, pp.800-811.
[48] Louis, M.S., Azad, Z., Delshadtehrani, L., Gupta, S., Warden, P., Reddi, V.J. and Joshi, A., 2019, June. Towards deep learning using tensorflow lite on risc-v. In Third Workshop on Computer Architecture Research with RISC-V (CARRV) (Vol. 1, p. 6).