JADAVPUR UNIVERSITY

KOLKATA - 700032

---------------PROJECT REPORT TITLE---------------

Prediction of Diabetes Mellitus at Early Stage using Deep Learning Methods

2021-23

Presented by:

Shreeparna Debnath

Roll No. – 002130201005

JADAVPUR UNIVERSITY

School Of Bio-Science and Engineering

This is to certify that the project report

Prediction of Diabetes Mellitus at Early Stage using Deep Learning Methods

Has been successfully completed by

SHREEPARNA DEBNATH

ROLL NO: – 002130201005

In partial fulfillment for the award of the degree in

Master Of Engineering

In

BIO-MEDICAL ENGINEERING

2021-23

Under the Supervision of

Prof. Dr. Anasua Sarkar

(Professor, Department of Computer Science and Engineering,

Jadavpur University)

Certificate of Recommendation

This is to certify that the Thesis entitled ***"*Prediction of Diabetes Mellitus at Early Stage using Deep Learning Methods*"***submitted by **Shreeparna Debnath** under the supervision of ***Prof. Dr Anasua Sarkar*** (Professor, Dept. of CSE, JU), has been prepared according to the regulations of **M.E Degree** in **School Of Bio-Science and Engineering Department**, awarded by Jadavpur University and he/she has fulfilled the requirements for submission of thesis report and that neither his/her thesis report has been submitted for any degree/diploma or any other academic award anywhere before.

…………………………………………………….………

Prof. Dr. Anasua Sarkar

(Professor, Dept of CSE, JU)
Project Supervisor

………………………………

Prof. Piyali Basak
(HOD, Dept of BME, JU)

JADAVPUR UNIVERSITY

*Certificate of Approval**

The foregoing thesis report is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned don't necessarily endorse or approve any statement made opinion expressed or conclusion drawn therein but approve the project report only for the purpose for which it is submitted.

Signature of the Examiners:

1……………………………………………….

2……………………………………………….

**Only in the case the thesis report is approved*

Acknowledgement

The successful completion of this project marked the beginning of an ever-going learning experience of converting ideas and concepts into real life, practical system. This report was quite satisfying with respect to learning experience at each and every step. At the same time it has given me confidence to work in professional setup which will lead me to gain the bright prospect in future. With the deep sense of gratitude, I express my sincere thanks to Prof. Anasua Sarkar and Prof. Piyali Basak for her active support and continuous guidance for this report.

Signature:

Shreeparna Debnath

ABSTRACT

Diabetes Mellitus is one of the world's leading cause of heart attacks, kidney failure, lower limb ablation, stroke and blindness. The number of affected people increased from 108 million to 422 million from 1980 to 2014 and will go as high as 629 million by 2045.Automated testing, early prediction and diagnosis of diabetes mellitus is therefore essential for improvement of patient's survival rate. In this paper, we have explored various deep learning methods like single LSTM, Auto-Encoder, optimized CNN + LSTM and LSTM Skip connection model over the same two datasets PIMA Indian Diabetes dataset and dataset collected from the victims in Sylhet Diabetes Hospital, Bangladesh. We have achieved highest accuracy of 93.26% in LSTM with Skip connection algorithm and highest Precision of 98.53% in Optimized LSTM + CNN algorithm.

CONTENTS

5. RESULTS

6. CONCLUSION

REFERENCES

CHAPTER 1: INTRODUCTION

1.1 Motivation

Diabetes Mellitus is one of the world's leading cause of heart attacks, kidney failure, lower limb ablation, stroke and blindness, according to World Health Organization (WHO) and it was the major cause of 1.5 million deaths in 2019 [1]. The number of affected people increased from 108 million to 422 million from 1980 to 2014 and will go as high as 629 million by 2045 [2].It is predicted that approximately 592 individuals will die from diabetes by the year 2035.Hence, our proposed method focuses on the early detection and diagnosis of diabetes mellitus with improved accuracy and further analysis.

1.2 Objectives

The objectives of this thesis are:

I. Study of different types of diabetes and its symptoms.
II. Gain insight into the different deep learning methods.
III. In specific, prediction of diabetes at early stage and focus on improving the accuracy by our proposed methodology.

1.3 Outline

This thesis is divided into six chapters.
In this chapter, we presented the motivation behind doing this project thesis and the broad framework of the thesis.
Chapter 2 gives an overview of the symptoms, risk factors and different types of diabetes and deep learning methods and finally a study of previous literature.
Chapter 3 presents the databases that were used in the method proposed in this thesis.
Chapter 4 presents the models used in our proposed methodology.
Chapter 5 presents the results obtained from the methods used for diabetes prediction.
Finally, we conclude the thesis presenting its main conclusions in **Chapter 6**. The future scope of our work is also discussed.

CHAPTER 2: BACKGROUND

2.1 Diabetes Mellitus

Diabetes Mellitus is a medical condition that results from either insufficient insulin production by the pancreas or ineffective utilization of insulin by the body, leading to elevated levels of glucose (blood sugar)[3]. People of all ages are affected by it. It is one of the world’s leading cause of heart attacks, kidney failure, lower limb ablation, stroke and blindness, according to World Health Organization (WHO).

2.1.1 Symptoms Of Diabetes

Symptoms of Diabetes includes the following

- Thirst
- Frequent Urination
- Unintended weight loss
- Fatigue
- Irritability or mood changes
- Blurred vision
- Slow – healing wounds and frequent infections

2.1.2 Causes of Diabetes

The occurrence of Diabetes Mellitus, irrespective of its type, is due to an excessive amount of glucose present in the blood [3].

Diabetes is caused by various factors, such as autoimmune disease, damage in pancreas, insulin resistance, hormonal imbalances and genetic mutations. Insulin resistance, which occurs when cells in the liver, muscles, and fat do not respond adequately to insulin, is primarily responsible for type 2 diabetes, and factors such as obesity, hormonal imbalances, genetics, and diet contribute to varying degrees of insulin resistance. In contrast, type 1 diabetes and LADA result from autoimmune attacks on insulin-producing cells in the pancreas, whereas gestational diabetes occurs during pregnancy due to the hormones released by the placenta causing insulin resistance. Additionally, Type 2 diabetes can result from other hormone-related conditions such as acromegaly and Cushing's syndrome, and physical damage to the pancreas due to a condition, surgery, or injury can cause Type 3c diabetes. Furthermore, specific genetic mutations can lead to neonatal diabetes and MODY.

2.2 **Types Of Diabetes**

2.2.1 Type 1 Diabetes

Insulin-dependent diabetes is another name for Type 1 diabetes, which was previously referred to as juvenile-onset diabetes because it typically develops during childhood. Type 1 diabetes is an autoimmune disorder that occurs when the immune system produces antibodies that attack the pancreas, causing damage to the organ and impeding its ability to produce insulin. Genetic factors can contribute to the development of this form of diabetes, or it may occur due to issues with insulin-producing cells within the pancreas. Individuals with Type 1 diabetes are at an increased risk of developing various health issues, including diabetic retinopathy, diabetic neuropathy, and diabetic nephropathy, which stem from damage to the small blood vessels in the eyes, nerves, and kidneys, respectively. Furthermore, those with Type 1 diabetes have a greater likelihood of developing heart disease and experiencing strokes [4].

The medicaments for Type 1 diabetes involve administering insulin injections into the subcutaneous fatty tissue beneath the skin.

2.2.2 Type 2 Diabetes

Type 2 diabetes was previously referred to as adult-onset or non-insulin-dependent diabetes, but over the last two decades, it has become increasingly prevalent among children and adolescents, primarily due to rising rates of obesity and overweight. Approximately 90% of individuals with diabetes are diagnosed with Type 2 diabetes. Type 2 diabetes typically has less severe symptoms compared to Type 1 diabetes, although it can still result in significant health complications, particularly in the small blood vessels of the kidneys, eyes, and nerves. In addition, Type 2 diabetes increases the likelihood of experiencing heart disorder and strokes [4].

The medicaments for Type 2 diabetes involve maintaining a healthy weight through proper nutrition and exercise, while some individuals may also require medication.

2.2.3 Gestational Diabetes

During pregnancy, women typically experience insulin resistance to some extent. When this condition progresses to diabetes, it is referred to as gestational diabetes, which is commonly detected during the middle or late stages of pregnancy by medical professionals. As a woman's blood sugar levels are transmitted through the placenta to the growing baby, it is essential to manage gestational diabetes to safeguard the baby's proper growth and development.

The medicaments of gestational diabetes comprises several measures, including adhering to a well-planned dietary regime that provides adequate nutrients while limiting fat and calories. Engaging in regular physical activity and monitoring weight gain are also essential components of the treatment plan [4]. In case of elevated blood sugar levels, insulin therapy may be necessary to regulate glucose levels effectively.

2.2.4 Prediabetes

Prediabetes refers to a medical condition in which an individual's blood sugar levels exceed the recommended range but are insufficient for a diabetes diagnosis by a medical practitioner [4]. Over one-third of the US population has prediabetes, but a majority of them remain unaware of it. Having prediabetes elevates the likelihood of developing type 2 diabetes and heart disease.

Engaging in physical activity and shedding some extra pounds, even if it is just 5% to 7% of one's body weight, can effectively reduce the associated risks.

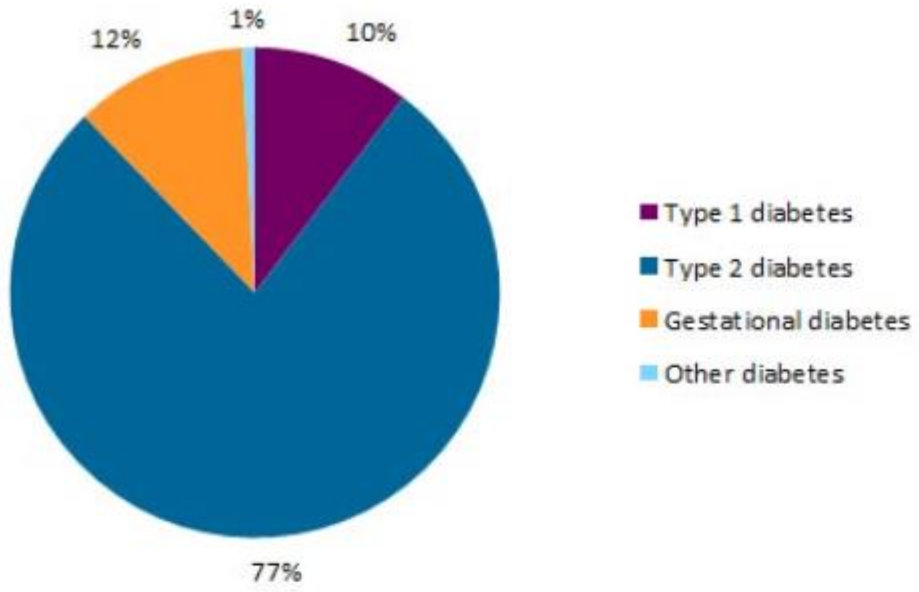

Fig – 1: Percentage of types of diabetes

Source: https://diabetesmellitusucd.wordpress.com/what-exactly-is-diabetes/

2.3 **Literature Survey**

Diabetes, being a non-communicable ailment, can give rise to long-term complexities and grave health concerns. The World Health Organization's report [5] sheds light on the impact of diabetes and its associated complexities, which can adversely affect individuals, their families, and their finances. The study also revealed that an uncontrolled stage of diabetes leads to approximately 1.2 million deaths, while roughly 2.2 million deaths are caused by other diseases related to diabetes, such as cardiovascular ailments, which arise due to its risk factors.

This paper [6] explores several classifiers and proposes a decision support system that employs the AdaBoost algorithm with Decision Stump as the ground classifier for classification. Additionally, Naive Bayes, Support Vector Machine and Decision Tree have also been implemented as ground classifiers for the AdaBoost algorithm to validate its accuracy. The precision attained for the AdaBoost algorithm with Decision Stump as the ground classifier is 80.72%, which is significantly higher than that of Support Vector Machine, Naive Bayes, and Decision Tree. In [7] research was conducted to detect diabetes using various machine learning (ML) algorithms including MLP, J48, Sequential Minimal Optimization (SMO) and Reduced Error Pruning Tree (REP-tree)methods. Their study concluded that the SMO algorithm resulted in the highest accuracy of 76.80%. In [8] the author employed Artificial Neural Network for diabetes prediction. Training of the data was done utilizing the MATLAB and performed regression analysis using distinct algorithms, namely BFGS Quasi-Newton, and Levenberg-Marquardt and Bayesian Regulation in which they achieved 88.8% accuracy. Ensemble of ML models (DT + XGB + LGB + RF) was done by the authors in [9] in which they achieved highest accuracy of 0.735 and area under ROC curve of 0.832. In [10] Convolutional Long Short-term Memory (Conv-LSTM) model was developed and the performance was compared with Convolutional Neural Network (CNN), Traditional LSTM (T-LSTM), and CNN-LSTM. Highest accuracy of 91.38% was attained in Conv-LSTM model. An Optimization Based CNN-Bi-LSTM deep learning model for diabetes prediction was developed by [11] which achieved accuracy of 98% exceeded other deep learning models.

Likewise, even a slight modification in data could significantly impact the entire structure of the decision tree model [12]. Additionally, Support Vector Machine (SVM) may encounter minor challenges with noisy data [13].

Diabetes is detected by scrutinizing Heart Rate Variability (HRV) signals derived from ECG signals in [14]. To automatically detect the abnormality, they employed Convolutional Neural Network (CNN) and a combination of CNN and Long Short-Term Memory (LSTM), referred to as CNN-LSTM deep learning networks. They obtained the highest accuracy of 90.9% in CNN-LSTM on the test data.

CHAPTER 3: MATERIALS

3.1 Diabetes Datasets

Publicly available datasets such as PIMA Indian Diabetes Dataset and dataset collected from the victims in Sylhet Diabetes Hospital, Bangladesh were used to experiment and evaluate the results of various algorithms.

- The National Institute of Diabetes and Digestive and Kidney Diseases provided the PIMA Indian Diabetes dataset [15] with the aim of predicting, through specific diagnostic measurements included in it, whether a patient has diabetes or not. This dataset consists of 768 rows (data) and a total of 9 columns (features).Mostly all the data here are of females of at least 21 years old. Outcome is either 1 (‘has diabetes’) or 0 (‘does not have diabetes’).

- The second dataset used for this experiment is dataset collected from the victims in Sylhet Diabetes Hospital, Bangladesh [16] which consisted of 520 rows (data) and 17 column (features) including the Outcome column.

3.2 Software Used

Python 3.7 has been used for this project along with OpenCV 4.5.1 and Scikit Image 0.17.2 package for different image operations. Matplotlib 3.4 package has been used for visualization of various images and plotting graphs and results. Jupyter Notebook and Pycharm have been used as an Integrated Development Environment (IDE). This entire project was performed on a Laptop with specifications: Windows 10 64bit platform Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz 2.59 GHz, 32GB RAM, NVIDIA Cuda GeForce GTX 1660 Ti 6GB GPU.

CHAPTER 4: METHODOLOGY

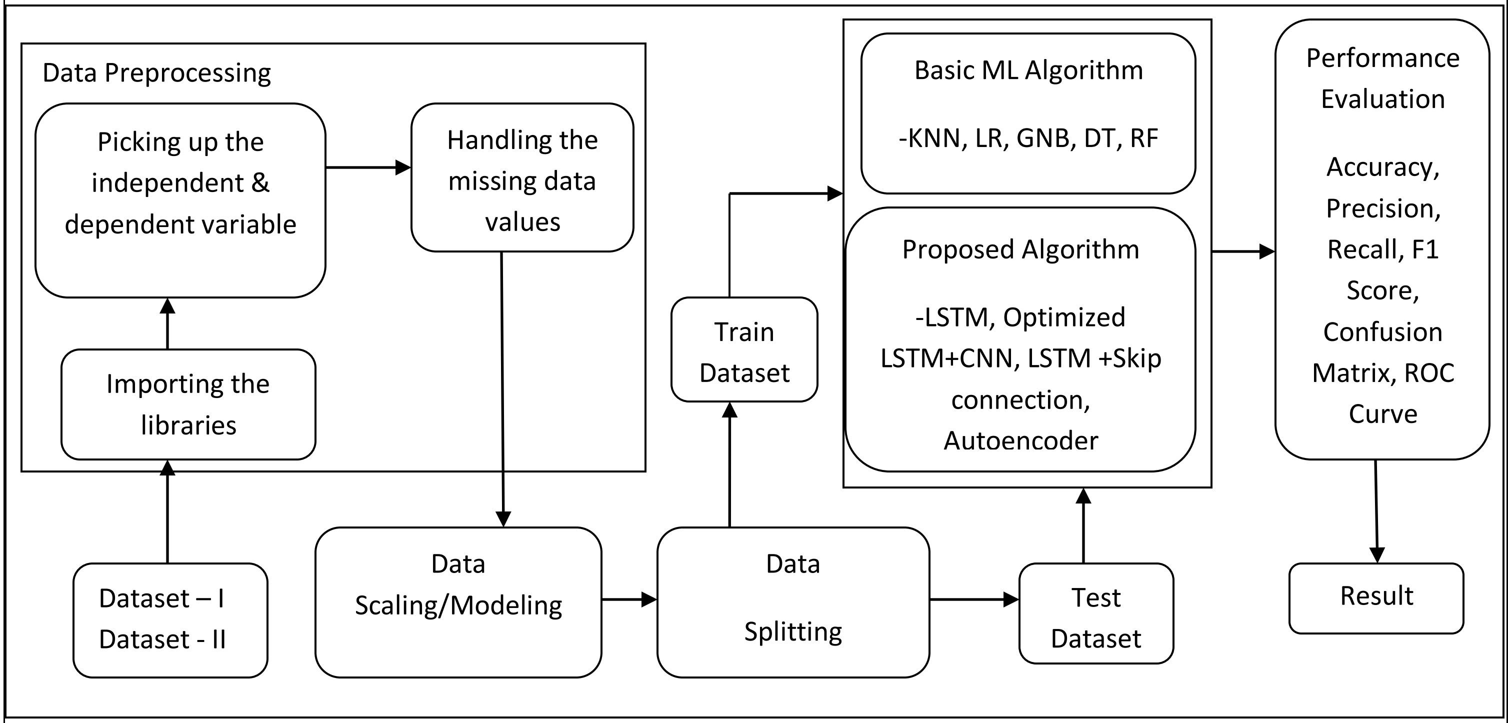

Fig -2: Overall Methodology of the proposed work.

4.1 **Implementation Details**

The initial step of the methodology is data pre-processing & data modeling. As the data obtained from various sources is often collected in raw format that is unsuitable for analysis that is why data needs to be pre-processed before feeding it into various algorithms for obtaining better results [17].It includes the following steps in the below sequence:

Data Pre-processing

4.1.1 Importing the Libraries and dataset

To carry out data preprocessing with Python, we must import pre-existing Python libraries that have specific functions. The libraries that we have used for data preprocessing are numpy, pandas, and matplotlib. And then finally we import the dataset using the read_csv() function.

4.1.2 Picking up the independent & dependent variables

4.1.3 Managing the missing data values

The subsequent phase of data preprocessing involves managing any missing data present in the dataset. Failure to address missing values can pose significant challenges to the machine learning model's effectiveness. Therefore, it is crucial to handle missing values in the dataset. There are two primary methods of managing missing data: deleting the entire row or column that contains null values or calculating the mean of the column or row containing missing data and substituting it in place of the missing value [36].

Data Modeling/Data Scaling

4.1.4 Data Transformation Using Quantile Transformer

Quantile transforms refers to a method used to convert numerical input or output variables in order to achieve a Gaussian or uniform probability distribution [21].
It aims to map the data to a specified probability distribution, such as Gaussian or uniform, by estimating the quantiles of the data and applying a transformation accordingly. This transformation can be useful for various purposes, such as normalizing the data or reducing the impact of outliers.
Data transformation can also be done by various other scaling functions like StandardScaler function that standardizes the data and ensures that the data has a mean of zero and standard deviation of one.Similarly, there are other scaling functions like MinMaxScaler(), RobustScaler(), MaxAbsScaler() and PowerTransformer().Data Transformation can be improved by Box-Cox transformation as in [20]. These scaling functions improve the performance and stability of Machine Learning Models

Data Splitting

4.1.5 Data Splitting

Data splitting refers to the process of separating a dataset into separate subsets for training, validation, and testing [33]. The dataset is typically split into two or three portions, depending on the specific use case and available data:
Training Set: This subset is used to train the machine learning model. The model learns patterns and relationships in the data from this set.

Validation Set: Also known as the development set, this subset is used to fine-tune the model and optimize its hyperparameters. It helps in assessing the model's performance and making adjustments to improve its generalization ability.

Testing Set: This subset is used to evaluate the final performance of the trained model. It provides an unbiased estimate of the model's performance on unseen data.

Data splitting is a critical step in machine learning as it allows for proper evaluation, validation, and optimization of models. It is also important for avoiding Overfitting and Hyperparameter Tuning. It helps ensure the model's ability to generalize well and perform reliably on unseen data, making it a fundamental practice in the field of machine learning.

Model Building

4.1.6 Basic ML Algorithms

i. K-Nearest Neighbor(K-NN)
K-NN is a straightforward machine learning algorithm that falls under the category of supervised learning techniques. It relies on the assumption of similarity between new data points and the existing labeled data to classify the new data into the most similar category. The K-NN algorithm stores all the available data points in its memory and determines the category of a new data point by comparing its similarity with the existing data [38]. Essentially, it finds the K nearest neighbors to the new data point and assigns the category that is most prevalent among those neighbors.

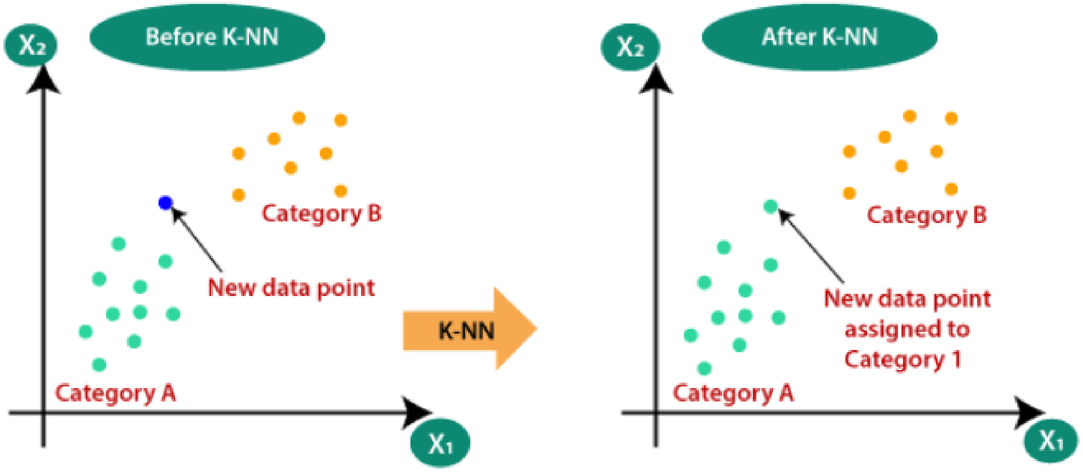

Fig 3- : K-NN Algorithm (Source -https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning)

ii. Logistic Regression (LR) –

Logistic regression is an extensively employed machine learning algorithm, categorized under supervised learning techniques. Its purpose is to predict a categorical dependent variable using a given set of independent variables[31]. Logistic regression is specifically designed for predicting the outcome of a categorical dependent variable. Therefore, the predicted outcome falls into discrete categories, such as Yes or No, 0 or 1, true or false, among others. Instead of providing an exact value of 0 or 1, logistic regression produces probabilistic values that range between 0 and 1[22].

In logistic regression, rather than fitting a regression line, we establish a curve following an "S" shape known as the logistic function. This logistic function predicts maximum values of 0 or 1.The curve derived from the logistic function represents the likelihood of a certain outcome

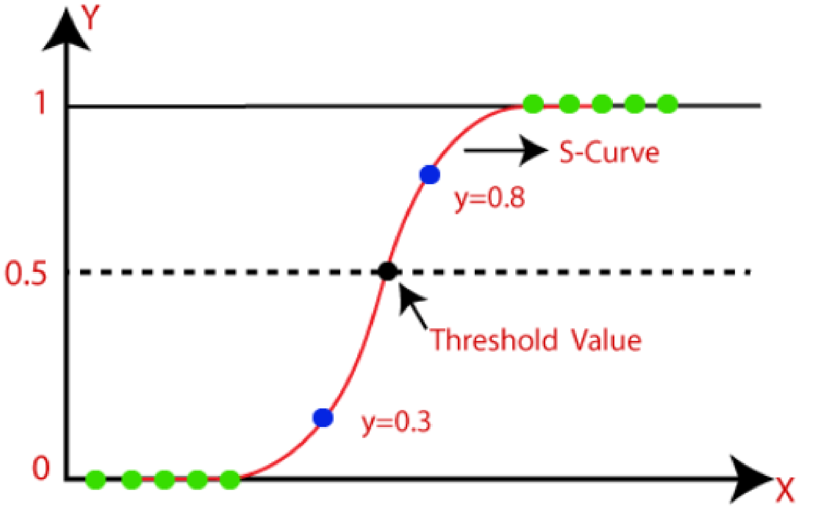

Fig -4: LR function (Source - https://www.javatpoint.com/logistic-regression-in-machine-learning)

iii. Decision Tree –

The Decision Tree is a supervised learning technique that is versatile and applicable to both classification and regression problems, although it is predominantly used for classification tasks.[32] It is characterized by a tree-like structure, where internal nodes represent the features of a dataset, branches represent decision rules, and each leaf node represents an outcome. Within the Decision Tree, two types of nodes exist: Decision Nodes and Leaf Nodes. Decision nodes are responsible for making decisions and possess multiple branches that correspond to different conditions or rules. On the other hand, leaf nodes serve as the final output of these decisions and do not contain any additional branches. They represent the predicted class or value based on the applied decision rules.

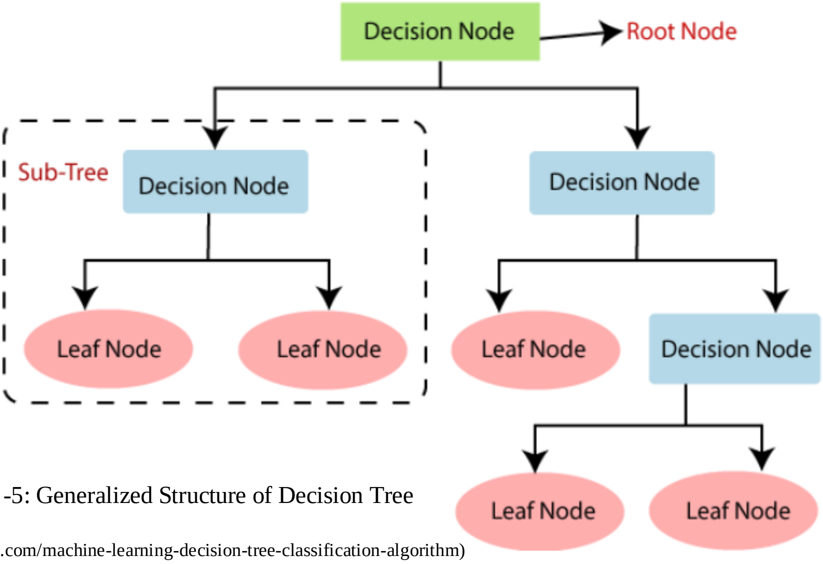

Fig -5: Generalized Structure of Decision Tree

(Source- https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm)

iv. Gaussian Naive Bayes –

The Naive Bayes algorithm is a popular technique for supervised classification tasks, utilizing the Bayes theorem .[25] It finds wide application in text categorization tasks that involve a considerable amount of training data. The Naive Bayes Classifier stands out as a simple and effective algorithm for classification, enabling the rapid creation of machine learning models that deliver accurate predictions. As a probabilistic classifier, it uses the probability of an event occurring to make predictions.

v. Random Forest (RF) –

It is a supervised learning method that is applied to both Classification and Regression tasks in machine learning. It is founded on the concept of ensemble learning, which involves combining multiple classifiers to solve complex problems and enhance model performance. As its name implies, it is a classifier comprising numerous decision trees built on different subsets of the provided dataset. [23] By averaging the predictions from each tree, it enhances the accuracy of predictions. Instead of relying on a single decision tree, the random forest considers the majority votes from the predictions made by each tree to determine the final output.

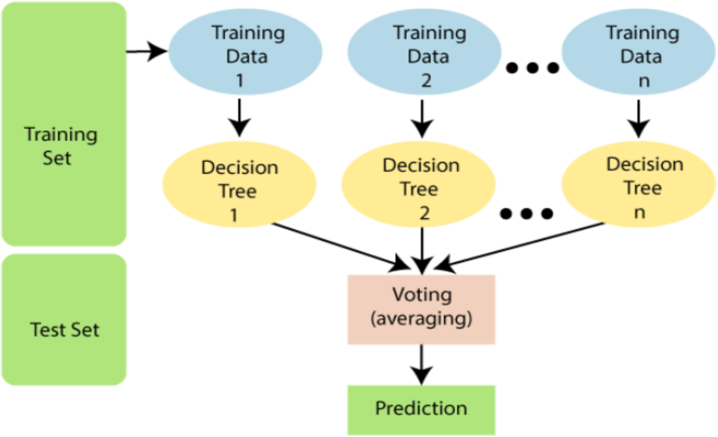

Fig - 6: Working Principle of RF

(Source - https://www.javatpoint.com/machine-learning-random-forest-algorithm)

4.1.7 Proposed Model

I. Long Short Term Memory Networks (LSTM) –

The LSTM (Long Short-Term Memory) model is a type of recurrent neural network (RNN) that is designed to address the vanishing gradient problem and capture long-term dependencies in sequential data. It is widely used in tasks such as natural language processing, speech recognition, and time series analysis [28]. The working principle of an LSTM model involves the use of specialized memory cells, known as LSTM cells, which are capable of retaining information over long periods of time. An LSTM cell's essential elements are:

Cell State (Ct): The cell state acts as a memory that stores information from previous time steps, allowing the LSTM to capture long-term dependencies. It can selectively retain or discard information through gates.
Input Gate (i): The input gate determines which information from the current time step is relevant and should be stored in the cell state. It combines input features with the previous hidden state to decide what to update.
Forget Gate (f): The forget gate determines which information should be forgotten from the cell state. It selectively removes irrelevant information from the previous cell state by using a sigmoid activation function.
Output Gate (o): The output gate controls which parts of the cell state are revealed as the output. It combines the input features and previous hidden state to decide what to output.
Hidden State (h): The hidden state is the output of the LSTM cell at each time step. It is a filtered version of the cell state and carries relevant information to the next time step [35].
The LSTM model processes sequential data by iteratively applying the operations of the LSTM cells to each time step. It takes the input at each time step, updates the cell state, and produces an output. The cell state is updated based on the input gate, forget gate, and output gate, which are all learned during the training process.
The LSTM model's ability to retain long-term dependencies and mitigate the vanishing gradient problem makes it well-suited for tasks that involve sequential data with complex temporal relationships [40].

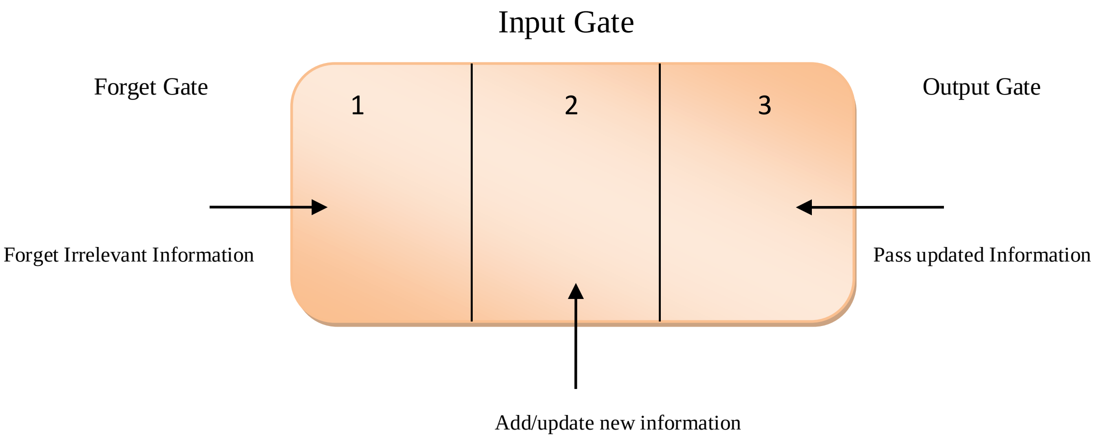

Fig -7 : Working Principle of LSTM

II. Optimized LSTM +CNN –

The optimized LSTM + CNN algorithm combines the strengths of both Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) architectures to improve the performance of various tasks, such as image classification, natural language processing, and time series analysis. The working principle of this algorithm involves integrating LSTM and CNN components in a complementary manner [24].

The CNN component plays a crucial role in extracting significant features from the input data, whether it's images or sequential data. It comprises convolutional layers, pooling layers, and activation functions. The convolutional layers employ filters to analyze the input data, capturing local patterns and distinctive features.[29] Meanwhile, the pooling layers downsample the resulting feature maps, reducing the spatial dimensions while preserving essential information. Additionally, activation functions introduce non-linear transformations to the CNN, empowering the network to acquire intricate representations and learn complex relationships within the data [37]. In Optimized LSTM + CNN, the CNN extracts relevant spatial features from the input data, capturing patterns and local dependencies.

The LSTM processes the extracted features sequentially, capturing long-term dependencies and temporal patterns. The CNN component typically serves as a feature extractor, and the LSTM component acts as a sequential processor. The outputs of the CNN component are fed into the LSTM component for further analysis and prediction. The LSTM can effectively learn from the spatial features extracted by the CNN, making predictions or classifications based on the sequential context.

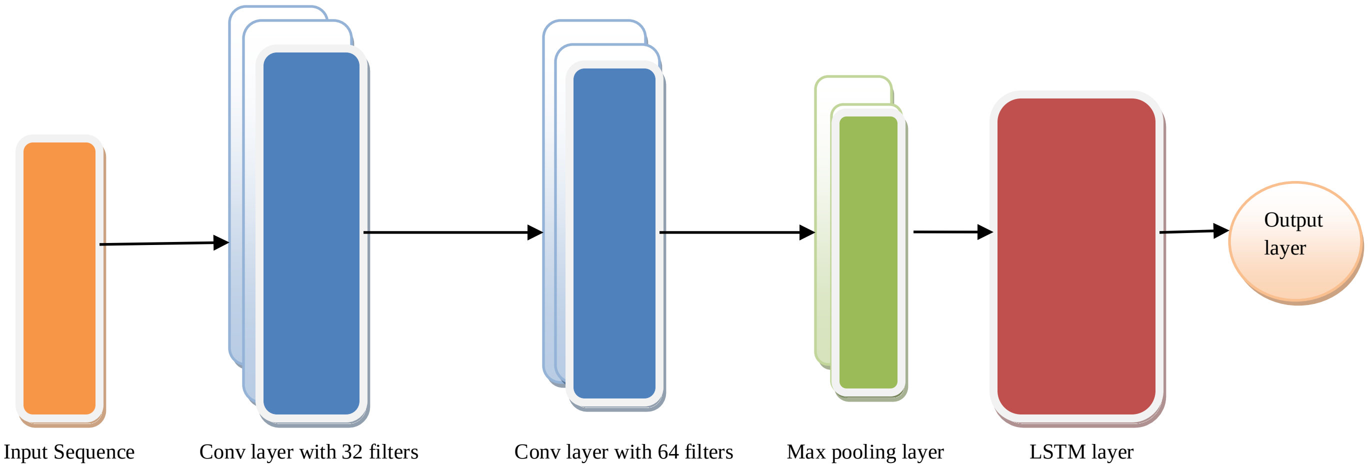

Fig- 8: LSTM-CNN model

III. LSTM with Skip Connection –

LSTM with skip connection refers to a variant of the Long Short-Term Memory (LSTM) architecture that incorporates skip connections, also known as residual connections. In LSTM with skip connection, skip connections are added to the standard LSTM architecture. Skip connections allow direct connections from one LSTM unit to another, bypassing intermediate layers. These connections [39] enable the gradient to flow more easily during training and help mitigate the vanishing gradient problem. The skip connections can be implemented by adding the output of a previous LSTM unit to the input of a subsequent LSTM unit [27].

Advantages of Skip Connections:

a. Skip connections facilitate the flow of information from earlier layers to later layers, preserving valuable information and reducing the risk of information loss.
b. They enable the model to access both low-level and high-level features at different stages of processing [34].
c. Skip connections promote better gradient propagation, allowing for more effective learning and training of deep LSTM networks.
d. They can help improve the overall performance and convergence speed of the LSTM model.

By incorporating skip connections, LSTM with skip connection enhances the ability of the network to capture and utilize information from various depths within the architecture. This can result in improved model performance, especially when dealing with complex tasks and long-term dependencies in sequential data.

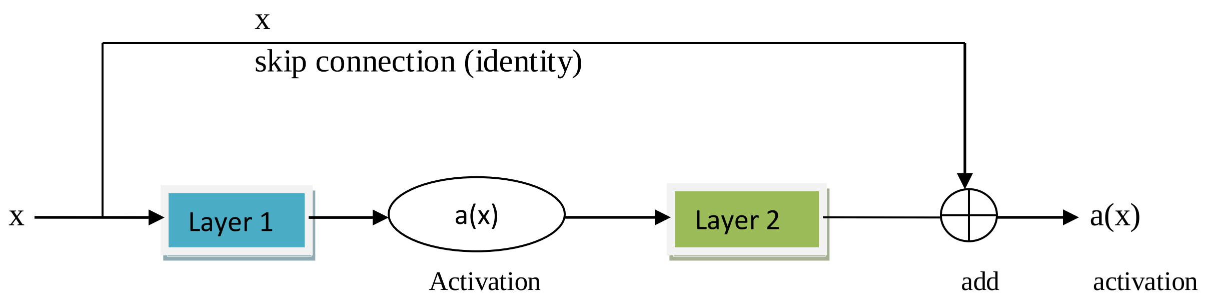

Fig - 9: Skip Connection

IV. Autoencoder –

An autoencoder is a type of unsupervised learning model in machine learning that aims to learn efficient representations of input data by training a neural network to reconstruct the input itself. It is primarily used for dimensionality reduction, data compression, and feature extraction [26].

The working principle of an autoencoder can be described as follows:

a. Encoder :

The encoder component of an autoencoder takes the input data and maps it to a lower-dimensional representation, also known as the latent space or encoding. It typically consists of one or more fully connected layers, where each layer applies a linear transformation followed by a non-linear activation function. The encoder progressively reduces the dimensionality of the input data, capturing the most important features.

b. Decoder :

The decoder component reconstructs the input data from the encoded representation. It mirrors the architecture of the encoder but in reverse order, with each layer expanding the dimensionality of the encoded representation. [30] The decoder's objective is to generate output that closely resembles the original input.

c. Reconstruction Loss :

During training, an autoencoder aims to minimize the difference between the input and the reconstructed output. The most common loss function used for this purpose is the mean squared error (MSE) between the input and the output. By optimizing the reconstruction loss, the autoencoder learns to capture the most salient features of the input data in the encoded representation.

d. Bottleneck Layer :
The bottleneck layer refers to the layer(s) in the encoder-decoder architecture with the lowest dimensionality. This layer acts as a compressed representation of the input data and serves as a bottleneck through which the data must pass.

e. Latent Space Representation:
The latent space, also known as the encoding or compressed representation, is the low-dimensional space where the input data is mapped by the encoder. This representation captures the most important features of the input data in a compressed form. The latent space can be used for tasks such as dimensionality reduction, anomaly detection, or generating new data samples.

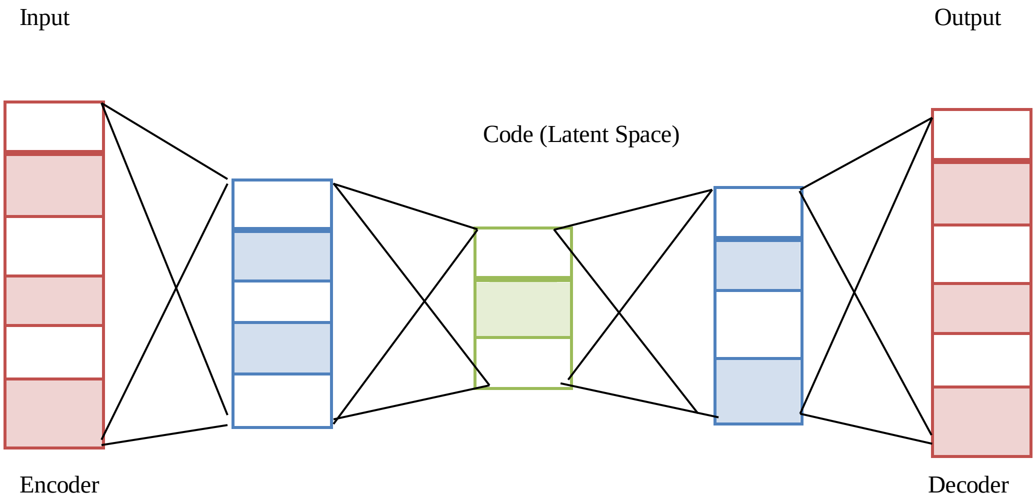

Fig :10 - Working Principle of Autoencoder

4.3 Performance Evaluation

Four Parameters have been used for determining the performance/efficiency of the algorithm. These four parameters Accuracy, Precision, Recall and F1 Score.

Accuracy is defined as the proportion of correct predictions out of the total number of predictions made, and it can be calculated using the following formula.

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

Precision, in the context of identifying patients with diabetes, refers to the proportion of correctly identified diabetic patients (positive instances) out of all patients with diabetes. It is calculated by dividing the number of true positives (TP) by the sum of TP and false positives (FP).

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP + FP}}$$

Recall is also known as sensitivity or true positive rate. The recall score is determined by dividing the number of true positives (TP) by the sum of true positives and false negatives (FN). In simpler terms, recall represents the proportion of actual positive instances accurately identified by the model.

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

The F1-Score is a metric that combines precision and recall by calculating their weighted average. By doing so, the F1-Score takes into account both false positives and false negatives, providing a balanced evaluation of the model's performance.

$$\textbf{F1 Score} = \frac{\textbf{2 x Precision x Recall}}{\textbf{Precision + Recall}}$$

Here TP means true positive, TN means true negative, FP is false positive, and FN is false negative.

CHAPTER 5 : RESULTS

After implementing different Machine Learning Algorithms on the datasets, we obtained the following result.

1. Basic ML Algorithms(Model)

   We have first implemented the basic ML Algorithms on the PIMA dataset. Following are the accuracy Table and the ROC curve in fig-

Algorithms	Accuracy
KNN	73.2%
Logistic Regression	83%
Decision Tree	73.2%
Gaussian NB	80.3%
Random Forest	81%

Table -1 : Accuracy Table of the basic ML algorithms applied.

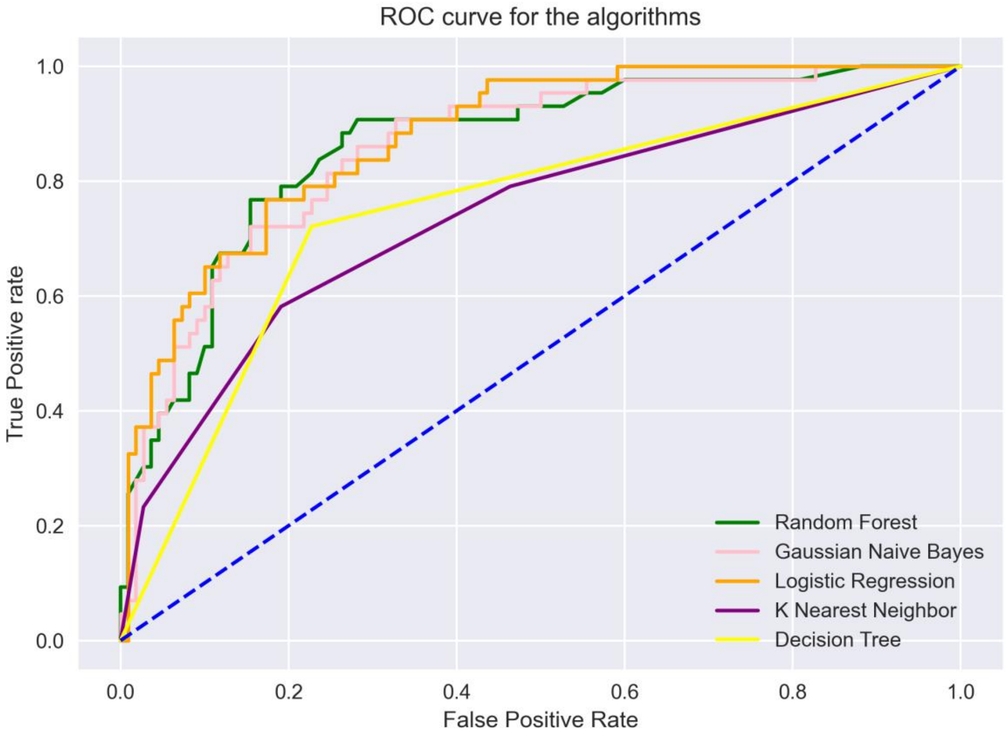

Fig -11 : ROC Curve of the basic ML algorithms

It can be observed from the Accuracy Table and the ROC curve that Logistic Regression (LR) gives the highest accuracy.

Confusion Matrix for the algorithm with highest accuracy i.e. Logistic Regression is given below:

	Diabetic	**Non – Diabetic**
Diabetic	101	9
Non-Diabetic	17	26

Table -2 : Confusion Matrix for Logistic Regression

2. Proposed Algorithms (Models)

We have explored various deep learning methods and have proposed 4 such models - single LSTM, optimized CNN + LSTM ,LSTM with Skip connection and Autoencoder model over the same two datasets PIMA Indian Diabetes dataset and dataset collected from the victims in Sylhet Diabetes Hospital, Bangladesh.

The Performance metrics score table for both the datasets of our proposed methodology is given below:

a) Dataset I – PIMA Dataset

Models	**Accuracy**	**Precision**	**Recall**	**F1 Score**
LSTM	79.95 %	69.57%	62.88%	66.23%
Optimized LSTM + CNN	78.57%	69.64%	70.91%	70.27%
LSTM + skip connection	80.52 %	69.77%	63.83%	66.66%
Autoencoder	75.62%	67.84%	69.09%	68.46%

Table-3 : Performance metrics of our proposed methodology on 1st Dataset

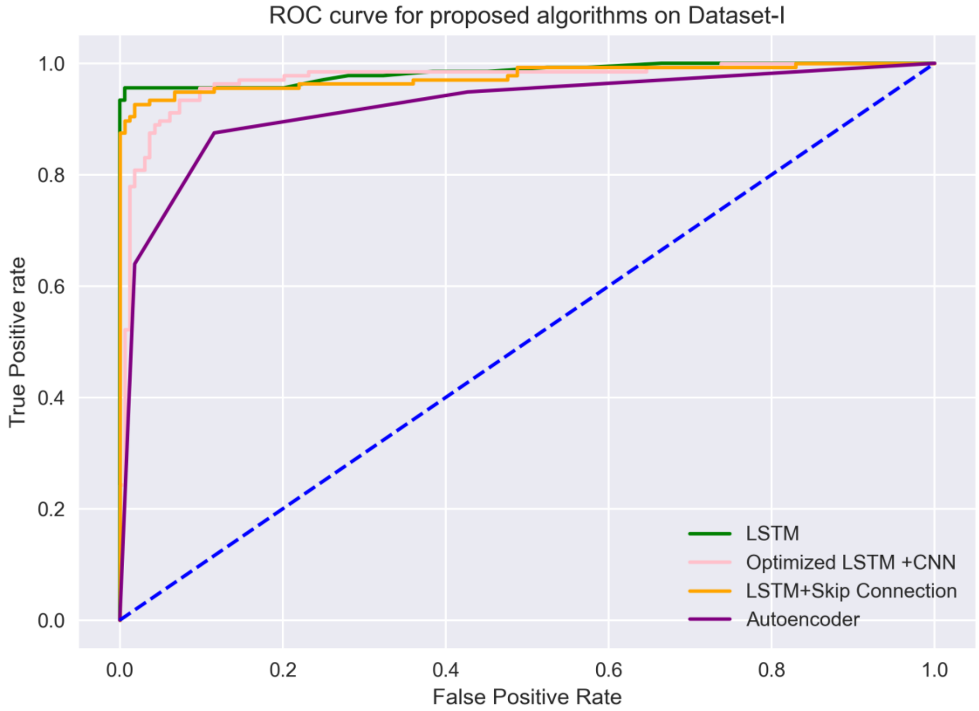

Fig - 12: ROC curve for proposed methodology on 1st Dataset

It can be observed that the Accuracy and Precision score is highest for LSTM with Skip connection algorithm whereas Recall and F1 score is highest for Optimized LSTM + CNN algorithm on PIMA Dataset. Below is the confusion matrix for the algorithm with the highest accuracy score:

	Positive	**Negative**
Positive	94	13
Negative	17	30

Table -4 : Confusion Matrix for LSTM with Skip connection model on Dataset-I

b) Dataset II - Dataset collected from the victims in Sylhet Diabetes Hospital, Bangladesh.

Models	Accuracy	Precision	Recall	F1 Score
LSTM	89.16 %	87.42%	88.83%	88.12%
Optimized LSTM + CNN	91.19%	98.53%	94.36%	96.40%
LSTM + skip connection	93.26 %	97.05%	92.95%	94.96%
Autoencoder	87.50%	87.84%	89.09%	88.46%

Table- 5: Performance metrics of our proposed methodology on 2nd Dataset

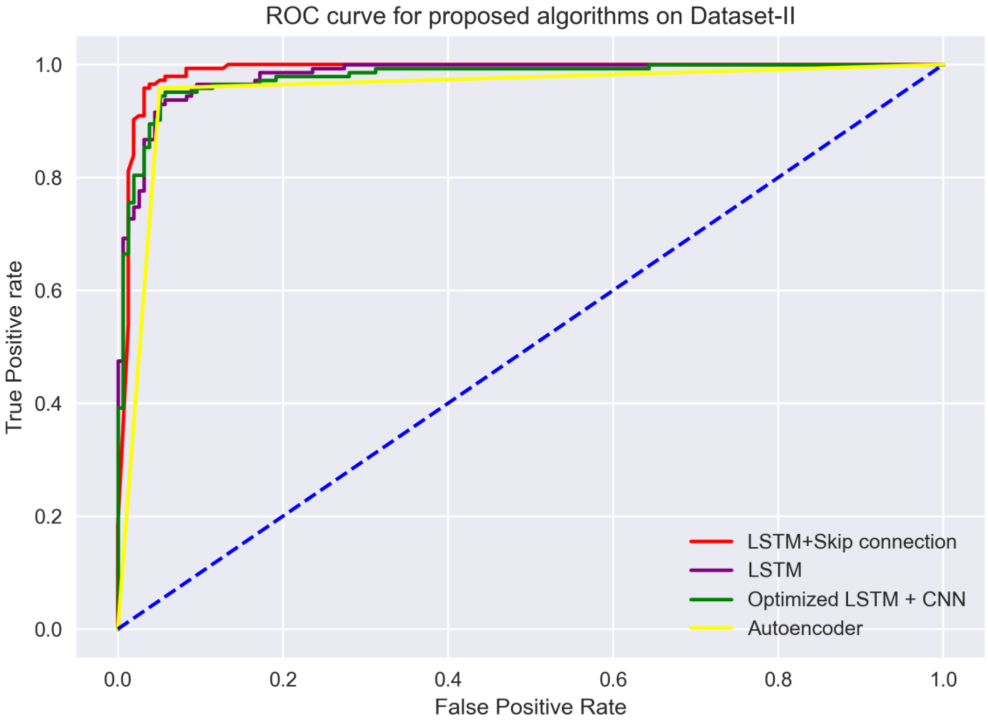

Fig -13 : ROC curve for proposed methodology on 2nd Dataset

It can be observed that the Accuracy score is highest for LSTM with Skip connection algorithm whereas Precision, Recall and F1 score is highest for Optimized LSTM + CNN algorithm on the 2nd Dataset. Below is the confusion matrix for the algorithm with the highest accuracy score:

	Positive	**Negative**
Positive	31	2
Negative	5	66

Table -6 : Confusion Matrix for LSTM with Skip connection model on Dataset-II

CHAPTER 6: CONCLUSION

In this thesis, we have proposed four models for classification of diabetes dataset i.e., LSTM, Optimized LSTM +CNN, LSTM with Skip connection and Autoencoder Model. The observed result shows that from both the datasets, the highest accuracy score i.e. 80.52%(dataset-I) and 93.26%(dataset-II) is obtained for LSTM with Skip connection algorithm and the highest Recall and F1 Score is obtained for Optimized LSTM + CNN algorithm. Precision score is highest for Autoencoder model in Dataset-I (67.84%) and Optimized LSTM + CNN model in Dataset-II (98.53%).The highest accuracy obtained on dataset-II has surpassed the accuracy obtained using Convolution based LSTM model as in [10].Also our proposed methodology has exceeded the accuracy 85.09% obtained in [18] by artificial neural network. In [19] artificial backpropagation scaled conjugate gradient neural network (ABP-SCGNN) algorithm has been used to predict diabetes at an early stage obtaining an accuracy of 93% which is less than the accuracy obtained by our proposed methodology i.e. LSTM with Skip connection on dataset-II. Implementation of LSTM with Skip connection algorithm in diabetes dataset is being done for the first time and has not been used earlier for diabetes prediction. This algorithm has proved to be quite efficient compared to other algorithms in early prediction of diabetes. Adding skip connection in LSTM algorithm has helped to improve the accuracy and the efficiency of the model.

It would be more advantageous to delve into deeper networks that imitate the combination of LSTM+CNN with skip connections in future research. This approach has the potential to enhance the learning capacity and precision of the model. Additionally, the model parameters can be optimized using techniques such as Grid Search or similar methods.

REFERENCES

[1] https://www.who.int/health-topics/diabetes#tab=tab_1

[2] Naz H, Ahuja S. Deep learning approach for diabetes prediction using PIMA Indian dataset. J Diabetes Metab Disord. 2020 Apr 14;19(1):391-403. doi: 10.1007/s40200-020-00520-5. PMID: 32550190; PMCID: PMC7270283.

[3] https://my.clevelandclinic.org/health/diseases/7104-diabetes

[4] https://www.webmd.com/diabetes/guide/types-of-diabetes-mellitus

[5] Global report on diabetes by World Health Organisation. 2016, ISBN 978 92 4 156525 7.

[6] VeenaVijayan V, Anjali C. Prediction and diagnosis of diabetes mellitus—a machine learning approach. Recent Adv. 2015. https://doi.org/10.1109/raics.2015.7488400.

[7] D. Verma, N. Mishra, "Analysis and predicion of breast cancer and diabetes disease datasets using data mining classification techniques", International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, 533-538, 2017.

[8] Muhammad Akmal Sapon, Khadijah Ismail and Suehazlyn Zainudin - "Prediction of Diabetes by using Artificial Neural Network". 2011 International Conference on Circuits, System and Simulation IPCSIT vol.7 (2011) (2011) IACSIT Press, Singapore.

[9] Dutta A, Hasan MK, Ahmad M, Awal MA, Islam MA, Masud M, Meshref H. Early Prediction of Diabetes Using an Ensemble of Machine Learning Models. *International Journal of Environmental Research and Public Health*. 2022; 19(19):12378. https://doi.org/10.3390/ijerph191912378

[10] Motiur Rahman, Dilshad Islam, Rokeya Jahan Mukti, Indrajit Saha,A deep learning approach based on convolutional LSTM for detecting diabetes,Computational Biology and Chemistry,Volume 88,2020,107329,ISSN 1476-9271, https://doi.org/10.1016/j.compbiolchem.2020.107329

[11] Madan P, Singh V, Chaudhari V, Albagory Y, Dumka A, Singh R, Gehlot A, Rashid M, Alshamrani SS, AlGhamdi AS. An Optimization-Based Diabetes Prediction Model Using CNN and Bi-Directional LSTM in Real-Time Environment. *Applied Sciences*. 2022; 12(8):3989. https://doi.org/10.3390/app12083989

[12] Kannadasan, K.; Edla, D.R.; Kuppili, V. Type 2 diabetes data classification using stacked autoencoders in deep neural networks. *Clin. Epidemiol. Glob. Health* **2019**, *7*, 530–535.

[13] Kotsiantis, S.B. Decision trees: A recent overview. *Artif. Intell. Rev.* **2013**, *39*, 261–283.

[14] Swapna G, Soman Kp, Vinayakumar R,Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals,Procedia Computer Science,Volume 132,2018,Pages 1253-1262,ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.05.041

[15] https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[16] https://www.kaggle.com/datasets/andrewmvd/early-diabetes-classification

[17] https://www.javatpoint.com/data-preprocessing-machine-learning

[18] Nitesh Pradhan, Geeta Rani, Vijaypal Singh Dhaka, Ramesh Chandra Poonia,14 - Diabetes prediction using artificial neural network,Editor(s): Basant Agarwal, Valentina Emilia Balas, Lakhmi C. Jain, Ramesh Chandra Poonia, Manisha,Deep Learning Techniques for Biomedical and Health Informatics,Academic Press,2020,Pages 327-339,ISBN 9780128190616, https://doi.org/10.1016/B978-0-12-819061-6.00014-8

[19] Muhammad Mazhar Bukhari, Bader Fahad Alkhamees, Saddam Hussain, Abdu Gumaei, Adel Assiri, Syed Sajid Ullah, "An Improved Artificial Neural Network Model for Effective Diabetes Prediction", *Complexity*, vol. 2021, Article ID 5525271, 10 pages, 2021. https://doi.org/10.1155/2021/5525271

[20] Osborne, Jason (2019) "Improving your data transformations: Applying the Box-Cox transformation," *Practical Assessment, Research, and Evaluation*: Vol. 15, Article 12. DOI: https://doi.org/10.7275/qbpc-gk17

[21] https://machinelearningmastery.com/quantile-transforms-for-machine-learning/

[22] Satish Kumar Kalagotla, Suryakanth V. Gangashetty, Kanuri Giridhar,A novel stacking technique for prediction of diabetes,Computers in Biology and Medicine,Volume 135,2021,104554,ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2021.104554

[23] https://www.javatpoint.com/machine-learning-random-forest-algorithm

[24] P. Bharath Kumar Chowdary and R. Udaya Kumar, “An Effective Approach for Detecting Diabetes using Deep Learning Techniques based on Convolutional LSTM Networks” International Journal of Advanced Computer Science and Applications(IJACSA), 12(4), 2021. https://dx.doi.org/10.14569/IJACSA.2021.0120466

[25] arun Jaiswal, Anjli Negi, Tarun Pal,A review on current advances in machine learning based diabetes prediction,Primary Care Diabetes,Volume 15, Issue 3,2021,Pages 435-443,ISSN 1751-9918, https://doi.org/10.1016/j.pcd.2021.02.005

[26] https://neptune.ai/blog/representation-learning-with-autoencoder

[27] Miyoshi, R., Nagata, N. & Hashimoto, M. Enhanced convolutional LSTM with spatial and temporal skip connections and temporal gates for facial expression recognition from video. *Neural Comput & Applic* **33**, 7381–7392 (2021). https://doi.org/10.1007/s00521-020-05557-4

[28] Masatoshi Nagata, Kohichi Takai, Keiji Yasuda, Panikos Heracleous, and Akio Yoneyama. 2018. Prediction Models for Risk of Type-2 Diabetes Using Health Claims. In *Proceedings of the BioNLP 2018 workshop*, pages 172–176, Melbourne, Australia. Association for Computational Linguistics.

[29] Yahyaoui, Amani & Rasheed, Jawad & Jamil, Akhtar & Yesiltepe, Mirsat. (2019). A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques. 10.1109/UBMYK48245.2019.8965556.

[30] Wang W, Huang Y, Wang Y, Wang L. Generalized autoencoder: A neural network framework for dimensionality reduction. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops 2014 (pp. 490 -497).

[31] Priyanka Rajendra, Shahram Latifi,Prediction of diabetes using logistic regression and ensemble techniques,Computer Methods and Programs in Biomedicine Update,Volume 1,2021,100032,ISSN 2666-9900, https://doi.org/10.1016/j.cmpbup.2021.100032

[32] A. M. Posonia, S. Vigneshwari and D. J. Rani, "Machine Learning based Diabetes Prediction using Decision Tree J48," *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Thoothukudi, India, 2020, pp. 498-502, doi: 10.1109/ICISS49785.2020.9316001.

[33] https://machinelearningmastery.com/training-validation-test-split-and-cross-validation-done-right/

[34] Chang, Shiyu & Zhang, Yang & Han, Wei & Yu, Mo & Guo, Xiaoxiao & Tan, Wei & Cui, Xiaodong & Witbrock, Michael & Hasegawa-Johnson, Mark & Huang, Thomas. (2017). Dilated Recurrent Neural Networks.

[35] https://www.geeksforgeeks.org/long-short-term-memory-networks-explanation/

[36] https://www.geeksforgeeks.org/detect-and-remove-the-outliers-using-python/

[37] P. Nagabushanam, N. C. Jayan, C. Antony Joel and S. Radha, "CNN Architecture for Diabetes Classification," *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, Coimbatore, India, 2021, pp. 166-170, doi: 10.1109/ICSPC51351.2021.9451724.

[38] P. Tumuluru, L. R. Burra, K. K. Sushanth, S. N. Vali, C. H. M. H. SaiBaba and P. Yellamma, "DPMLT: Diabetes Prediction Using Machine Learning Techniques," *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, Tuticorin, India, 2022, pp. 1127-1133, doi: 10.1109/ICEARS53579.2022.9751944.

[39] Lindemann, Benjamin & Müller, Timo & Vietz, Hannes & Jazdi, Nasser & Weyrich, Michael. (2020). A Survey on Long Short-Term Memory Networks for Time Series Prediction. 10.13140/RG.2.2.36761.65129/1.

[40] https://towardsdatascience.com/lstm-networks-a-detailed-explanation-8fae6aefc7f9

● 8% Overall Similarity

Top sources found in the following databases:

- 2% Internet database
- Crossref database
- 7% Submitted Works database
- 1% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	WorldQuant University on 2023-06-06 Submitted works	1%
2	Cardinal Newman College on 2023-02-23 Submitted works	<1%
3	Aston University on 2022-01-07 Submitted works	<1%
4	Bradley University on 2023-05-11 Submitted works	<1%
5	Liverpool John Moores University on 2023-04-21 Submitted works	<1%
6	UOW Malaysia KDU University College Sdn. Bhd on 2023-03-15 Submitted works	<1%
7	nith on 2023-04-26 Submitted works	<1%
8	dspace.vutbr.cz Internet	<1%

9	University of Bristol on 2023-04-20 Submitted works	<1%
10	University of Portsmouth on 2023-05-12 Submitted works	<1%
11	coursehero.com Internet	<1%
12	diva-portal.org Internet	<1%
13	sec.gov Internet	<1%
14	University of Sunderland on 2013-03-22 Submitted works	<1%
15	VIT University on 2023-05-25 Submitted works	<1%
16	semanticscholar.org Internet	<1%
17	"Computer Vision – ECCV 2016 Workshops", Springer Science and Busi... Crossref	<1%
18	Liverpool John Moores University on 2022-05-20 Submitted works	<1%
19	Southern New Hampshire University - Continuing Education on 2023-0... Submitted works	<1%
20	University College London on 2023-05-12 Submitted works	<1%

21	Callaghan Campus on 2005-05-23 Submitted works	<1%
22	E. Sabitha, M. Durgadevi. "Improving the Diabetes Diagnosis Prediction... Crossref	<1%
23	University of Venda on 2022-03-15 Submitted works	<1%

● Excluded from Similarity Report

- Bibliographic material
- Cited material
- Quoted material
- Manually excluded text blocks

EXCLUDED TEXT BLOCKS

Finally, we conclude the thesis presenting its main conclusions in Chapter 6

Andrés Yesid Díaz Pinto. "Machine Learning for Glaucoma Assessment using Fundus Images", Universitat P...

OutlineThis thesis is divided into six chapters.In this chapter, we presented the

Andrés Yesid Díaz Pinto. "Machine Learning for Glaucoma Assessment using Fundus Images", Universitat P...

proposed in this thesis.Chapter 4 presents the

Andrés Yesid Díaz Pinto. "Machine Learning for Glaucoma Assessment using Fundus Images", Universitat P...