

**MACHINE LEARNING AND DEEP LEARNING APPROACHES TO PREDICT
EARLY STAGE OF DIABETES**

BY

**SONNET BAIDYA
ID: 201-15-3272**

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Ms. Fabliha Haque
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Ms. Tania Khatun
Assistant Professor
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

22 JANUARY 2024

APPROVAL

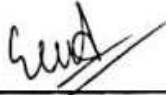
This Project/internship titled “Machine Learning and Deep Learning Approaches to Predict Early Stage of Diabetes”, submitted by Sonnet Baidya, ID No: 201-15-3272 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 22-01-2024.



BOARD OF EXAMINERS

Dr. S. M Aminul Haque (SMAS)
Professor & Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Chairman



Md. Sazzadur Ahmed (SZ)
Assistant professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Amatul Bushra Akhi (ABA)
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Md. Sazzadur Rahman (MSR)
Professor
Institute of Information Technology
Jahangirnagar University

External Examiner

DECLARATION

We hereby declare that this project has been done by us under the supervision of Ms. **Fabliha Haque, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Fabliha
24.01.24

Ms. Fabliha Haque
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Tania Khatun

Ms. Tania Khatun
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Sonnet Baidya

Sonnet Baidya
ID: 201-15-3272
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes me possible to complete the final year project/internship successfully.

I really grateful and wish our profound indebtedness to **Ms. Fabiliha Haque, Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning*” to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would like to express our heartiest gratitude to **Dr. Sheak Rashed Haider Noori**, Head, Department of CSE, for his kind help to finish my project and also to other faculty members and the staffs of CSE department of Daffodil International University.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

This study offers a thorough methodology that makes use of a variety of deep learning and machine learning algorithms to predict early stage of diabetes. Recurrent neural networks (RNN), Feedforward Neural Networks (FNN), Decision Trees (DT), Logistic Regression (LR), Random Forests (RF), K-Nearest Neighbors (KNN), Naive Bayes (NB), Support Vector Machines (SVM), and Long Short-Term Memory (LSTM) are all integrated in the suggested system. Nine health attributes for 100,000 entries are included in the dataset, which was obtained via Kaggle. Exploratory data analysis, quality checks, and encoding are all part of data pre-processing. For model evaluation, the dataset is divided into training and test sets, and a two-pronged feature selection technique is used. Notably, with 97% accuracy, the Decision Tree machine learning model shows greater accuracy in diabetes prediction. The study places a strong emphasis on moral issues with predictive modeling in healthcare. Prospective avenues for investigation encompass improving prediction models, augmenting openness, and tackling wider ethical considerations in the field of healthcare analytics.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
CHAPTER	
CHAPTER 1: INTRODUCTION	1-4
1.1 Introduction	1
1.2 Research Motivation	1
1.3 Rationale of the Study	2
1.4 Research Questions	2
1.5 Expected Output	2
1.6 Project Management and Finance	3
1.7 Report Layout	3
CHAPTER 2: BACKGROUND STUDY	5-14
2.1 Preliminaries	5
2.2 Diabetes Related Previous Works	5
2.3 Comparison Study of diabetes related previous works	10
2.4 Scope of the problem	11
2.5 Challenges	13

CHAPTER 3: RESEARCH METHODOLOGY	15-29
3.1 Proposed Model	15
3.2 Dataset Collection	15
3.3 Data Pre-processing	16
3.4 Dataset Description	16
3.5 Histogram Visualization	18
3.6 Box Plot Visualization	20
3.7 Training Data and Test Data	21
3.8 Feature Selection	21
3.9 Target Variable	23
3.10 Machine Learning and Deep Learning Algorithms Used	24
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	30-42
4.1 Experimental Setup	30
4.2 Experimental Results & Analysis	30
4.3 Machine Learning Models Outcome	31
4.4 Deep Learning Models Outcome	34
4.5 Discussion	40
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	43-44
5.1 Impact on Society	43
5.2 Impact on Environment	43

5.3 Ethical Aspects	43
5.4 Sustainability Plan	44
CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH	45-46
6.1 Summary of the Study	45
6.2 Conclusions	45
6.3 Recommendation	45
6.4 Implication for Further Study	46
APPENDIX	
REFERENCES	47-48
PLAGIARISM REPORT	49

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Block Diagram for Diabetes Prediction System	15
Figure 3.3: Data pre-processing Diagram	16
Figure 3.4: Descriptive Statistics of Dataset	18
Figure 3.5: Histogram Plot for Each of the Feature	19
Figure 3.6: Box Plot for Each Feature	20
Figure 3.8: Shows the Selected Features	22
Figure 3.9: Count plot of Diabetes Column	23
Figure 4.4.2: Confusion Matrix for RNN Model	35
Figure 4.4.3: Train- Test Model Loss (RNN)	36
Figure 4.4.5: Confusion Matrix for FNN Model	37
Figure 4.4.6: Train- Test Model Loss (FNN)	38
Figure 4.4.8: Confusion Matrix for LSTM Model	39
Figure 4.4.9: Train- Test Model Loss (LSTM)	40
Figure 4.5: Confusion Matrix for Decision Tree Model	42

LIST OF TABLES

FIGURES	PAGE NO
Table 2.3: Comparison Study of Diabetes Related Previous Works	11
Table 3.4: Sample of Dataset Column Headers	17
Table 4.3.1: Classification Report of KNN Model	32
Table 4.3.2: Classification Report of Naive Bayes (NB) Model	32
Table 4.3.3: Classification Report of Decision Tree (DT) Model	32
Table 4.3.4: Classification Report of Logistic Regression (LR) Model	33
Table 4.3.5: Classification Report of Random Forest (RF) Model	33
Table 4.3.6: Classification Report of SVM Model	34
Table 4.4.1: Classification Report of RNN Model	34
Table 4.4.4: Classification Report of FNN Model	36
Table 4.4.7: Classification Report of LSTM Model	38
Table 4.5: Comparison Table of ML and DL Classification Results	41

CHAPTER 1

Introduction

1.1 Introduction

Globally, the incidence of diabetes is rising, which highlights the urgent need for precise and early prediction models to enable prompt interventions and reduce related health risks. In order to meet this need, this research offers a thorough strategy that makes use of numerous machine learning and deep learning methods. The dataset used in the study, which was obtained from Kaggle and included nine essential health attributes, was pre-processed, visualized, and feature selected. Recurrent neural networks and long short-term memory are two examples of sophisticated deep learning models, while K-Nearest Neighbors, Decision Trees, and Support Vector Machines are notable methods. The study highlights the value of model transparency and explores ethical issues related to predictive healthcare analytics. Given the substantial societal and personal ramifications that diabetes prediction carries, this research seeks to advance the ongoing discussion on improving predictive models and promoting moral behavior in the field of healthcare data science.

1.2 Research Motivation

A promising avenue for developing non-invasive, early-stage diabetes prediction algorithms is made possible by the quick advances in machine learning (ML) and deep learning (DL). These methods can potentially identify complex patterns from readily available data, including demographics, health histories, and subtle lifestyle variations. The goal is to make accurate forecasts even in the early stages so that people who are at risk of diabetes can be identified more easily. By facilitating prompt preventative actions, this proactive strategy seeks to lower healthcare expenses and lessen related stress. These algorithms provide individualized risk evaluations by utilizing the data that is currently accessible, encouraging lifestyle changes for illness prevention. On the other hand, traditional diagnostic techniques like HbA1c and fasting blood glucose (FBG) sometimes depend on clinical presentations, which might result in missed chances for prompt

treatment and delayed diagnoses. Using cutting-edge ML and DL techniques improves prediction accuracy while enabling people to take early action and fostering wellbeing and personal development.

1.3 Rationale of the Study

The major objective of this study is to find out how effectively various ML and DL methods can predict diabetes in its early stages. assessing the effectiveness of several machine learning methods, such as SVM, Random Forest, KNN, Naive Bayes, Decision Trees, and Logistic Regression. evaluating the capabilities of DL algorithms such as LSTM, RNN, and FNN. determining the most important characteristics and looking at how various traits, including age, gender, blood sugar level, history of smoking, hypertension, heart disease, BMI, and HbA1c level, affect prediction accuracy. Creating several ML and DL models, then assessing how well they predict diabetes in its early stages.

1.4 Research Question

This research will address the following question:

- a) Which deep learning and machine learning algorithms predict early-stage diabetes with the highest accuracy?
- b) Which features play a major role in prediction models?
- c) Is it possible to interpret the developed models to comprehend the decision-making process?
- d) How can clinical workflows for early diabetes detection incorporate the proposed models?

1.5 Expected Output

Without assessing the algorithms, it is quite challenging to guarantee the desired result. However, machine learning algorithms such as KNN, Decision Trees, Logistic Regression, and others may not perform as well as deep learning algorithms. Through a variety of machine learning and deep learning techniques, the research attempts to anticipate diabetes

in its early stages. This study will assess various machine learning models and compare the accuracy, precision, recall, and F1-score of each model's performance and find the important factors that influence the risk of diabetes by examining feature importance. The interpretability of the top model will be assessed to comprehend how it makes decisions. Lastly, this study will discuss how to incorporate it into clinical processes for decision assistance and early detection, taking ethical issues like bias and data privacy into account. The goal of this project is to develop a potent early diabetes prediction tool that could save lives and enhance patient outcomes.

1.6 Project Management and Finance

Within the Department of Computer Science and Engineering at DIU, this research project is being carried out in order to fulfill criteria for the B.Sc. Eng. degree. This research project is not currently being supported by an outside sponsor; instead, **Ms. Fabliha Haque** is supervising this study. At this point, the study is entirely self-funded. To enable the advancement of our research, it is expected that our prestigious university and reputable authorities will provide financial support.

1.7 Report Layout

This thesis is organized as follows:

Chapter-1: Introduction: The introduction, research motivation, research purpose, and research question are the subjects covered in this chapter. This chapter is a reflection of the study.

Chapter-2: Background Study: This chapter will cover a variety of relevant themes, including the scope of work needed to address the issue and the difficulties in diagnosing the condition at an early stage.

Chapter-3: Research Methodology: This chapter will describe the data collection, preprocessing, and model development process, including details of the chosen.

Chapter-4: Experimental Results and Discussion: This chapter will present the results of the experiments, including the performance of each model on various evaluation metrics.

Chapter-5: Impact on Society, Environment and Sustainability: This chapter describes the detail impact of the research on society, environment and sustainability.

Chapter-6: Summary, Conclusion, Recommendation and Implication for Future Research: This chapter will provide an overview of the study, point out its shortcomings, and suggest future lines of inquiry for ML and DL-based early-stage diabetes prediction research.

CHAPTER 2

Background Study

2.1 Preliminaries

This work provides the basic knowledge needed to appreciate the complexity of diabetes prediction, including both traditional and cutting-edge approaches. Diabetes is a major worldwide health issue that highlights the critical need for early prediction techniques to allow for prompt therapies. The initial investigation entails clarifying basic principles of machine learning techniques applied in the healthcare industry. We introduce well-known algorithms with explanations of their functions, advantages, and disadvantages, such as K-Nearest Neighbors, Naive Bayes, Decision Tree, Logistic Regression, Random Forest, and Support Vector Machine (SVM). Furthermore, the presentation of deep learning techniques like long short-term memory (LSTM), feedforward neural networks (FNN), and recurrent neural networks (RNN) emphasizes their potential for identifying complex patterns in healthcare data. This study analyzes previous research on diabetes prediction as it wraps up its early phase with a comprehensive literature review. Through an analysis of current approaches and results, the project seeks to discover common patterns, pinpoint obstacles, and investigate prospects within the domain. This first investigation serves as the foundation for the study, directing the choice and explanation of the wide range of machines and deep learning algorithms utilized in the ensuing technique. This section's established understanding lays the groundwork for the following chapters' more thorough analysis of the research methodology used to forecast diabetes.

2.2 Diabetes Related Previous Works

Their research proposed diabetes prediction using different types of machine learning and deep learning approaches like Decision Tree, ANN, Naive Bayes and SVM algorithms. This study uses the PIMA Indian Diabetes Dataset to test different machine learning techniques for precise diabetes prediction. It examines recall, accuracy, precision, and F1-score, but it ignores potentially useful methods like regression in favor of concentrating

only on classification algorithms. Although it presents insightful comparisons, there are drawbacks. Reproducible results are hampered by the study's absence of critical implementation details, such as feature selection or hyperparameter tweaking, and in-depth analysis of trained models. Despite these drawbacks, it makes a substantial contribution to the field of diabetes prediction research by pointing out areas in which further research is needed to increase the accuracy, interpretability, and generalizability of the model. Their best accuracy was 82%. But they can use alternative deep learning methods like FNN, RNN, LSTM. These methods will provide better accuracy [1].

This research proposed various machine learning methods, such as deep learning, support vector machines (SVMs), and artificial neural networks (ANNs), are examined for how well they analyse medical data and spot diabetes-related trends. The increasing prevalence of diabetes worldwide and the significance of early detection to avoid complications are highlighted in the paper. The benefits of applying machine learning are emphasized by the authors, including enhanced accuracy, customized risk assessment, and potential for early intervention. They do note certain drawbacks, though, such as poor data quality, a lack of generalizability to different populations, and the requirement for additional algorithmic validation and improvement [2].

The study examines six distinct machine learning algorithms (KNN, DT, SVM, RF, NB, and LR) for diabetes prediction. Because a dataset of patient medical records is used in the study, the results may have application in real-world settings. To help readers evaluate the algorithms' efficacy, the authors include each algorithm's accuracy, precision, recall, and F1-score. The study addresses the potential advantages for patients and medical professionals of early diabetes prediction using machine learning. The study's limited dataset (certain details may be absent from your summary) could restrict how broadly the results can be applied. The performance measures are presented in the paper, but the precise causes of some algorithms' superior performance are not discussed. Additional examination of the significance of features and model behaviour may yield insightful information. restricted comparison with previously published work Although the publication cites prior studies on machine learning-based diabetes prediction, a more

complete comparison with the body of existing literature would enhance the research's significance. Boosting and stacking are two methods that combine many algorithms to provide better results than single models. Examine the application of artificial neural networks and other deep learning techniques for potentially more accurate diabetes prediction. Using openly accessible medical datasets or working with healthcare organizations may improve the models' robustness and generalizability. Their research's best accuracy was 77% (KNN) [3].

This study's proposed work uses classification algorithms such as Decision Trees (DT), Random Forests, SVM, KNN, Gradient Boosting (GB), and Logistic Regression (LR) to classify individuals who have been diagnosed with diabetes. As demonstrated by the experiment, the KNN algorithm outperformed other classification algorithms in terms of output. An accuracy of 85% was attained, according to the data. But if they use alternative approaches like deep learning methods, they can get better output [4].

The study provides an insightful comparison of well-liked machine learning (ML) methods, such as logistic regression, decision trees, SVM, KNN, ANN, Naive Bayes, and ANN, for the prediction of diabetes. Investigating ontology-based machine learning approaches is novel and could advance our understanding of diabetes-related knowledge. The algorithms are evaluated by the authors using industry-standard criteria such as recall, accuracy, precision, and F-measure, which gives a clear foundation for comparison. The study shows that SVM and ontology classifiers had the best performance, indicating the potential of these methods for diabetes prediction. Because it only uses data from the Pima Indian Diabetes Database (PIDD), the study may not be able to adequately generalize to other populations. The use of ontology is mentioned in the paper, but neither its specifics nor its benefits to the ML models are explored. Model performance may be enhanced by investigating and maybe adding pertinent features outside of those included in PIDD. By using more extensive and varied datasets, using ensemble methods, and delving further into ontology, their accuracy would have grown. They achieved a maximum accuracy of 77.5% in Ontology [5].

Their research proposed Comparing the CLSTM model against current machine learning and deep learning techniques, it achieves greater accuracy on the PIMA Indian Diabetes Dataset. This implies that the model's ability to forecast diabetes is effective. The suggested model is compared in the research with a number of machine learning and deep learning techniques. This makes it possible for readers to comprehend the relative benefits and drawbacks of the CLSTM methodology. Because of the paper's excellent organization and structure, readers will find it simple to follow the methodology and findings. The PIMA Indian Diabetes Dataset, which is renowned for its shortcomings and possible biases, is the main source of data used in the research. The generalizability of the results would be strengthened by testing the model on additional datasets. Furthermore, Insufficient clarification regarding hyperparameter adjustment. By using a bigger and more varied dataset, investigating various hyperparameter tuning techniques, incorporating feature engineering, addressing class imbalance, and explaining and interpreting the model, they would have increased accuracy. The best accuracy of their LSTM research was 96.8% [6].

This paper provides a comprehensive review of the applications of deep learning in the field of diabetes. The research that employs deep learning algorithms for diabetes screening and early detection are analysed by the authors. The results are encouraging and outperform traditional approaches. Along with these constraints, they also point out data availability, interpretability of the model, and generalizability. There are a few viable ways to increase the precision of deep learning models for diabetes based on the limits found. Work together to exchange anonymised patient data with research centres and healthcare facilities, for example. Use explainable AI methods to comprehend how deep learning models make predictions. Develop models that are continuously adaptive by visualizing the significance of features and the decision-making process. gain knowledge and get better with time [7].

In this paper, a machine learning algorithm-based paradigm for early diabetes prediction is proposed. Three algorithms are compared by the authors: Decision Tree (DT), Support Vector Machine (SVM), and Naive Bayes (NB). They give a decent summary of various methods by contrasting three distinct machine learning algorithms. The results are only partially discussed, and the causes of the variations in algorithm performance are not

thoroughly investigated. Only three algorithms are examined by the writers. Other intriguing strategies, like Random Forest or Ensemble techniques, are not looked at. They are able to become more accurate. describing the dataset's limitations and how they may impact the conclusions' generalizability. By examining how feature selection and pre-processing affect algorithm performance, accuracy can be increased. Examine the causes of the variations in the algorithms' performance and by contrasting the suggested framework with other current methods for predicting diabetes, accuracy can also be increased. In Naive Bayes, their best accuracy was 74.28%. They can use deep learning methods also to improve their accuracy, but they do not use [8].

Several machine learning techniques were suggested in this study to forecast diabetes. The study emphasizes the potential advantages of machine learning for diabetes prediction. Only females over the age of 19 are the subject of this study. The results may be more broadly applicable if more men and a larger age range were included in the data set. In Decision Tree, the best accuracy was 79.3% [9].

This research proposed different types of machine learning algorithms like Naive Bayes, RF, LR, KNN, SVM, Dt and Hybrid model. The hybrid model enhances overall performance by utilizing the advantages of several machine learning techniques, including SVM, Decision Tree, and CNN. This research may not be representative of other populations because it uses data from the Pima Indians Diabetes Database. Because of this, the results might not be as broadly applicable. When two or more models are combined, there's a chance of overfitting, which occurs when the model works well with training data but not with undiscovered data. The use of is necessary to increase accuracy Employ a dataset that is bigger and more varied. Examine several ensemble techniques, include supplementary information, apply feature engineering in your Continually assess and revise the model. Updating the model is necessary to keep it current and accurate in light of new information and findings. Their research's best accuracy in the hybrid model was 90.62% [10].

This research presents a deep learning algorithm (DLA) based diabetes detection mechanism. This model, which went by the names Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and hybrid CNN-LSTM, employed three distinct DLA. Heart Rate Variability (HRV), which was obtained from the Electrocardiogram (ECG), served as the algorithm's input data. The assessment of the data using Support Vector Machine (SVM) methods for classification. Prior to applying SVM, the best results for CNN and LSTM-CNN were 0.03% and 0.06%, respectively. This categorization technique has a maximum accuracy of 95.7% when used to diagnose diabetes using ECG waves. KP Soman et al. The purpose of this work was to develop an architecture for automated diabetes detection employing heart rate signals (HRV), CNN, and LSTM networks [11].

The model that the authors suggest can determine whether a patient has diabetes. The prediction accuracy of strong machine learning algorithms, which employ metrics like recall, precision, and F1-measure, forms the basis of this model. Based on diagnostic methods, the authors forecast the onset of diabetes using the Pima Indian Diabetes (PIDD) dataset. Using the Logistic Regression (LR), Naive Bayes (NB), and K-nearest Neighbor (KNN) methods, the corresponding outcomes were 94%, 79%, and 69% [12].

2.3 Comparison Study of Diabetes Related Previous Works

Table 2.3 summarizes a thorough review of studies on diabetes prediction, including the machine learning and deep learning models used and their maximum levels of accuracy. Diverse models were used in twelve studies; the CLSTM model performed exceptionally well, obtaining an astounding 96.8% accuracy. The KNN model stands out because it has an accuracy of 85%, which is higher than that of its competitors. The table provides a quick reference by providing an overview of each study's expected performance. It offers insightful information on the diabetes prediction landscape and illustrates the efficacy of several deep learning and machine learning techniques. This field is dynamic, as seen by the variety of models chosen and the stated accuracy levels, and it has the potential to produce reliable and accurate predictive modelling for diabetes research.

Table 2.3: Comparison Study of Diabetes Related Previous Works

Study No.	Author name	Used ML/DL Models	Best Accuracy
[1]	Sonar P. et al.	Decision Tree, ANN, Naïve Bayes, SVM	82%
[2]	Jaiswal, Varun. et al.	Deep Learning, SVMs, ANNs	Not mentioned
[3]	Nasir Kamal. et al.	KNN, DT, SVM, RF, NB, LR	77% (KNN)
[4]	Kumar, Y. et al.	KNN, DT, SVM, RF, NB, LR	85% (KNN)
[5]	Hakim El Massari. et al.	SVM, Ontology Classifiers	77.5%
[6]	Chowdary, P. et al.	CLSTM and Others	96.8%
[7]	Taiyu Zhu. et al.	Deep Learning Algorithms	Not mentioned
[8]	Shafi, Salliah. et al.	DT, SVM, NB	74.28% (NB)
[9]	Llaha, Olta. et al.	Decision Tree	79.3%
[10]	Samet, Sarra. et al.	Hybrid Model (SVM, DT, CNN)	90.62%
[11]	Swapna, Goutham. et al.	CNN, LSTM, Hybrid CNN-LSTM	95.7%
[12]	Khaleel. et al.	LR, NB, KNN	94% (LR)

2.4 Scope of the Problem

The study's main goal is to create and evaluate models that predict the early stages of diabetes based on clinical and lifestyle data by utilizing both machine learning and deep learning approaches. The specific scopes of the problem are-

a) Data pre-processing

- Investigate any missing or unbalanced data and take appropriate action.
- Make sure that the gender and smoking history categorical variables are accurately encoded.
- Examine each feature's significance and, if necessary, take dimensionality reduction strategies into account.

b) Model development for machine learning

- To maximize performance for early-stage diabetes prediction, adjust the hyperparameters for each of the following algorithms: Random Forest, Naive Bayes, Decision Tree, Logistic Regression, KNN and SVM.
- Compare the performance of these algorithms in terms of accuracy, precision, recall, and other relevant metrics.

c) Model development for deep learning

- Create and train appropriate deep learning architectures for early-stage diabetes prediction, such as FNN, RNN and LSTM.
- Compare the performance of deep learning models with the best machine learning models.
- If applicable, think about including pertinent medical knowledge in the model's design.

d) Model Interpretation and Explainability

- Analyze the models to understand their decision-making process and identify key features influencing predictions.
- Evaluate the model's clinical relevance and healthcare practitioners' ability to understand them.

e) Generalization and Validation

- Evaluate the performance of the best models.
- Discuss the limitations of my study.
- Examine the effects of characteristics on the prediction of early-stage diabetes, such as HbA1c and blood glucose levels.
- Examine how my models might be used for customized risk assessment.
- Assess the viability and possible advantages of implementing your models in an actual clinical environment.

2.5 Challenges

There are various challenges while conducting this study, which are as follows:

- a) **Data quality:** As, Medical data can be messy, with missing values, inconsistencies, and errors. So, it is very important to handle the missing values, Clean and prepare the data that was very time-consuming and require domain expertise.
- b) **Data size and complexity:** The Dataset was very large and complex. So it was very difficult to analyze and extract meaningful insights.
- c) **Class imbalance:** Machine learning algorithms may have difficulties because there was a comparatively fewer number of diabetics than healthy individuals. So it was difficult to balance the class.
- d) **Heterogeneity:** Since each person's experience of diabetes is unique, developing a model that works for everyone is challenging.
- e) **Temporal dependency:** Models that take into consideration the dynamic nature of blood glucose levels and other risk variables are necessary since they fluctuate over time.
- f) **Choosing the appropriate algorithm:** Different algorithms have different strengths and weaknesses. So, it was very difficult for me to select the appropriate one for the specific task and data.
- g) **Overfitting:** Overfitting of the training set might cause a model to perform poorly on unobserved data. To avoid this, rigorous validation and regularization strategies are required.

- h) **Interpretability:** It might be tough to comprehend the underlying causes of complex models' predictions since they can be hard to interpret. This may impede their adoption and acceptance in medical environments.
- i) **Generalizability:** It is important to carefully assess the representativeness and generalizability of data since models trained on one population might not perform well on another.
- j) **Ethical Consideration:** When applying AI to prediction in healthcare, concerns like fairness, prejudice, and privacy must be taken into consideration.
- k) **Clinical Integration:** It is necessary to consider the needs and acceptance of healthcare practitioners when integrating prediction models into clinical processes and decision-making.
- l) **Cost effectiveness:** Prediction model development and implementation can be costly. Hence a cost-benefit analysis is required to guarantee their usefulness in healthcare systems.

CHAPTER 3

Research Methodology

3.1 Proposed Model

The main goal of the suggested system is to apply various machine learning and deep learning algorithm combinations, as depicted in the block diagram above. The foundation classification algorithms for accuracy authentication are LSTM, RNN, FNN, Decision Tree, Random Forest, KNN, Naive Bayes, Logistic Regression and SVM.

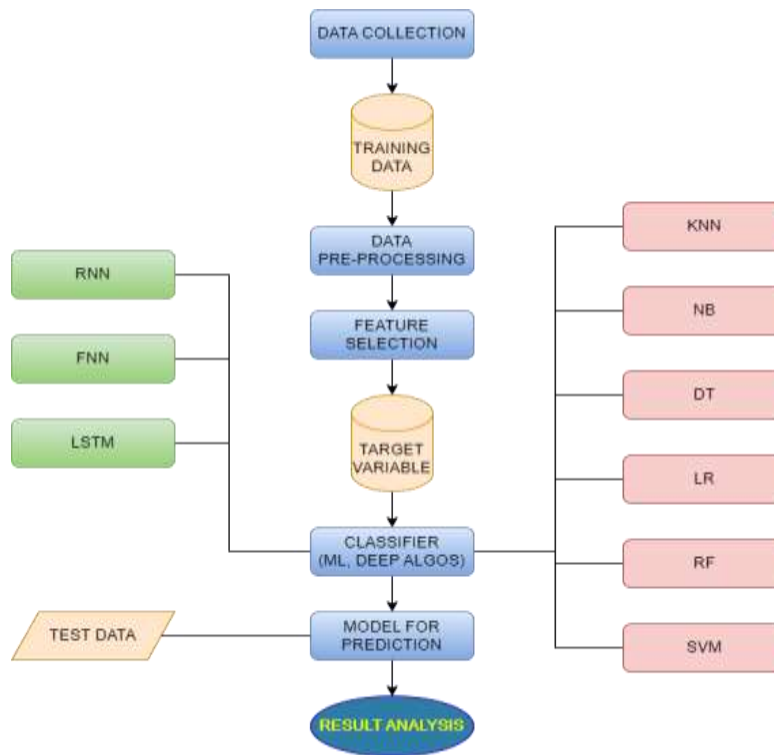


Figure 3.1: Block Diagram for Diabetes Prediction System

3.2 Dataset Collection

The initial action is to gather data. The dataset collected from Kaggle. Nine attributes (gender, age, blood glucose level, diabetes, heart disease, hypertension, smoking history, bmi, and HbA1c level) are included in the dataset with 100,000 raw values. In this instance,

certain factors are numerical, while others are categorical, such as gender and smoking history. Different types of qualities, such as male, female and others, are associated with gender. Additionally, smoking history includes a variety of features, including no info, never, former, current, not current, and ever.

3.3 Data pre-processing

The term “pre-processing” describes the changes we make to our data before giving it to the algorithm. Information The raw data is transformed into a set of interpretable data using preprocessing techniques. Stated differently, any time data is acquired in an unprocessed format from several sources, it becomes unusable for analysis.

As, the dataset which is collected contains two attributes with categorical values so first of all need to encode those into numerical values. Then need to check the duplicate values and dropped if there were any duplicate values. Then check the missing values and found there are no missing values in dataset.

Figure 3.3 shows the data pre-processing procedure-



Figure 3.3: Data pre-processing Diagram

3.4 Dataset Description

The dataset contains different types of features. There are 9 columns with 100000 rows. The attributes are gender, age, blood glucose level, diabetes, heart disease, hypertension, smoking history, bmi, and HbA1c level. There are also two columns gender and smoking history that contain categorical values. The other columns contain numerical values. Different types of qualities, such as male, female and others, are associated with gender. Additionally, smoking history includes a variety of features, including no info, never, former, current, not current, and ever.

Table 3.4 shows the sample of column headers below-

Table 3.4: Sample of Dataset Column Headers

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0
5	Female	20.0	0	0	never	27.32	6.6	85	0
6	Female	44.0	0	0	never	19.31	6.5	200	1
7	Female	79.0	0	0	No Info	23.86	5.7	85	0
8	Male	42.0	0	0	never	33.64	4.8	145	0
9	Female	32.0	0	0	never	27.32	5.0	100	0
10	Female	53.0	0	0	never	27.32	6.1	85	0
11	Female	54.0	0	0	former	54.70	6.0	100	0
12	Female	78.0	0	0	former	36.05	5.0	130	0
13	Female	67.0	0	0	never	25.69	5.8	200	0
14	Female	76.0	0	0	No Info	27.32	5.0	160	0
15	Male	78.0	0	0	No Info	27.32	6.6	126	0
16	Male	15.0	0	0	never	30.36	6.1	200	0
17	Female	42.0	0	0	never	24.48	5.7	158	0
18	Female	42.0	0	0	No Info	27.32	5.7	80	0
19	Male	37.0	0	0	ever	25.72	3.5	159	0

Feature index are-

- a) Gender: Indicates a person's biological sex, typically as "Male" or "Female or others".
- b) Age: Represents a person's age in years.
- c) Hypertension: Indicates whether a person has high blood pressure (1 = yes, 0 = no).
- d) Heart disease: Indicates whether a person has a history of heart disease (1 = yes, 0 = no).
- e) Smoking history: Indicates a person's smoking habits, as "never," "ever," "current," "former," "not current" or "no Info."
- f) Bmi: Body Mass Index, a measure of body fat based on height and weight.
- g) HbA1c level: Haemoglobin A1c, a blood test that measures average blood sugar levels over the past 2-3 months.

- h) Blood glucose level: The amount of glucose (sugar) present in the blood at a given time.
- i) Diabetes: Indicates whether a person has diabetes (1 = yes, 0 = no).

The count, mean, max, and standard deviation of the values in dataset have also been computed. The figure is given below-

	count	mean	std	min	25%	50%	75%	max
age	100000.0	41.885856	22.516840	0.08	24.00	43.00	60.00	80.00
hypertension	100000.0	0.074850	0.263150	0.00	0.00	0.00	0.00	1.00
heart_disease	100000.0	0.039420	0.194593	0.00	0.00	0.00	0.00	1.00
bmi	100000.0	27.320767	6.636783	10.01	23.63	27.32	29.58	95.69
HbA1c_level	100000.0	5.527507	1.070672	3.50	4.80	5.80	6.20	9.00
blood_glucose_level	100000.0	138.058060	40.708136	80.00	100.00	140.00	159.00	300.00
diabetes	100000.0	0.085000	0.278883	0.00	0.00	0.00	0.00	1.00
smoking_history_num	100000.0	0.274650	1.403060	-1.00	-1.00	0.00	1.00	4.00
gender_num	100000.0	0.585880	0.492937	0.00	0.00	1.00	1.00	2.00

Figure 3.4: Descriptive Statistics of Dataset

3.5 Histogram Visualization

The skewness of each class in the data distribution is another significant characteristic. Data visualization makes it easier to see the appearance of the data as well as the type of correlation that exists. We call this a histogram. An accurate graphical depiction of the distribution of numerical data is a histogram. It is a probability distribution estimate for a continuous variable. Understanding your data is greatly aided by using histograms. They make it simple to determine the locations of both huge and little amounts of data. To put it briefly, the histogram is made up of two axes: the x-axis displays the values on the x-axis, and the y-axis indicates how frequently those values appear in the data. The histogram for each of the features is given below in figure 3.5.

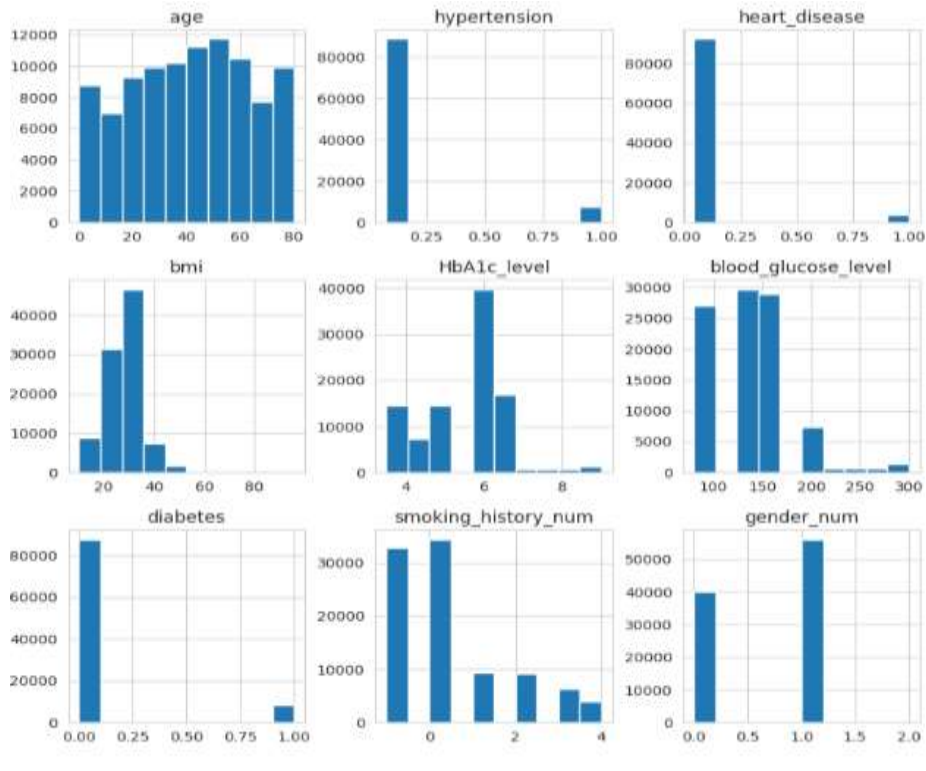


Figure 3.5: Histogram Plot for Each of the Feature

A right-skewed distribution is seen in Figure 3.5, highlighting a greater concentration of people in younger age groups. There are about equal numbers of people with and without hypertension, indicating that this specific health indicator has a symmetrical distribution. With regard to heart disease, however, a noteworthy left-skewed pattern appears that suggests a higher proportion of people without the illness. In contrast, the distribution of diabetes has a tendency that is skewed to the right, indicating a higher percentage of people who do not have the condition. There is a left-skewed distribution of smokers, while the non-smokers are more common. There is a virtually equal number of males and females in the symmetrical distribution of genders. The distribution and prevalence of different health issues across the population shown are clarified by the full view of numerous skewness patterns provided by Figure 3.5, which spans multiple health indicators. The graphic highlights the different levels of skewness in health-related variables, offering insightful information for additional research and interpretation.

3.6 Box Plot Visualization

Box Plots, often called box-and-whisker charts, are a useful visual aid for summarizing and comprehending a dataset's distribution. They are very helpful in machine learning for comparing distributions, locating outliers, and evaluating the quality of data. We can also make a box plot since the input variables are numerical. In a box plot, the outlier points are explicitly separated, and the median, 25th, 50th, and 75th percentiles, as well as the min/max that is not an outlier, are often displayed. The box plots for each feature are given below in figure 3.6.

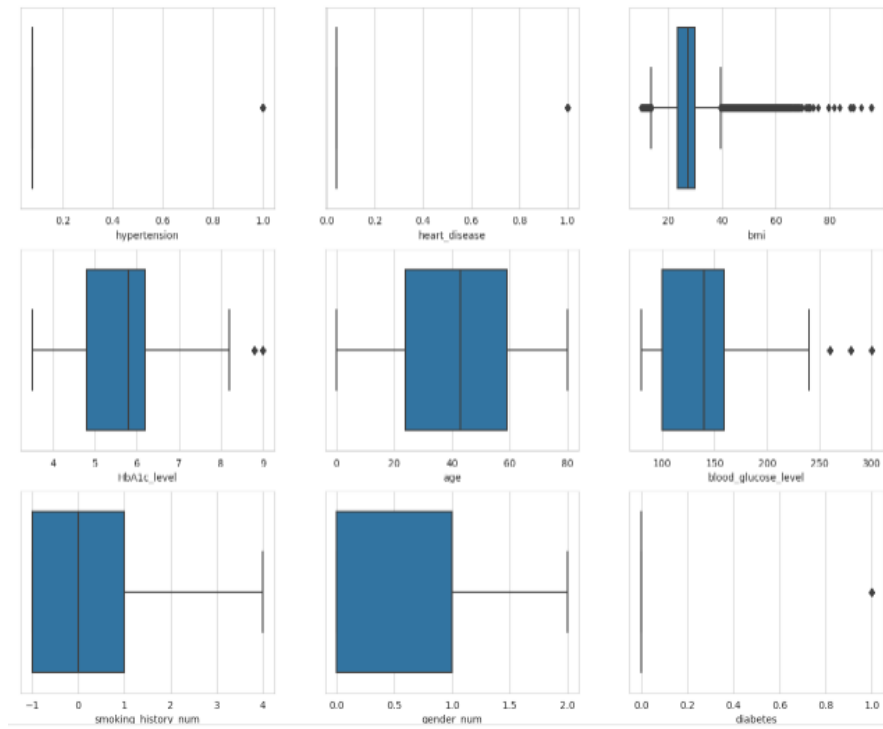


Figure 3.6: Box Plot for each Feature

For every feature, the box plot is displayed in figure 3.6. Data for distinct variables are presented in each boxplot. With a few outliers on the high end for bmi, the data is right skewed. With a few outliers on the high and low ends for hypertension, the data is essentially symmetrical. Regarding cardiac conditions, there are a few low-end outliers in the left-skewed data. Although there are a few outliers on both the high and low ends, the

data is generally balanced. With a few outliers on the top end of the age distribution, the data is right skewed. In the glucose level of the brood Although there are a few outliers on both the high and low ends, the data is generally balanced. The distribution of gender_Num is left-skewed, with a few low-end outliers. smoking_history_num There are a few outliers in the column on both the high and low ends, but overall it is symmetrical. With a few outliers on the low end of the diabetes spectrum, the data is left-skewed. In order to deal with these outliers, need to used quantiles to alter a dataset in order to uniformize its features, lessen the effect of outliers, and make them comparable. After that, a new Data Frame with modified features and illustrative column names is created.

3.7 Training Data and Test Data

To guarantee an objective assessment of the model and avoid overfitting, the dataset was carefully split into training and testing sets. 80% of the data was used for training and 20% for testing, which left a representative percentage for performance evaluation and provided enough data for the model to learn from. The model's capacity to generalize to new data and faithfully represent its predicting powers in the actual world is protected by this deliberate division.

3.8 Feature Selection

In this study used two-pronged feature selection strategy to find the most important informational nuggets inside the data. Initially, correlation analysis was employed to identify traits that demonstrated a strong association with the target variable. By eliminating characteristics that were unnecessary or redundant, it was able to simplify the model and lower its noise level. The use of model-based feature significance scores allowed for more refinement. It was able to rank the most informative features and exclude the least useful ones by examining each feature's contribution to the model's predictions. In addition to improving the model's performance, this careful selection procedure also made it easier to grasp the main ideas underlying its predictions. There are different types of features in the dataset (gender, age, blood glucose level, diabetes, heart disease,

hypertension, smoking history, bmi, and HbA1c level). For better accuracy selected all the features between those features. But the diabetes column which did not select because this is the target variable.

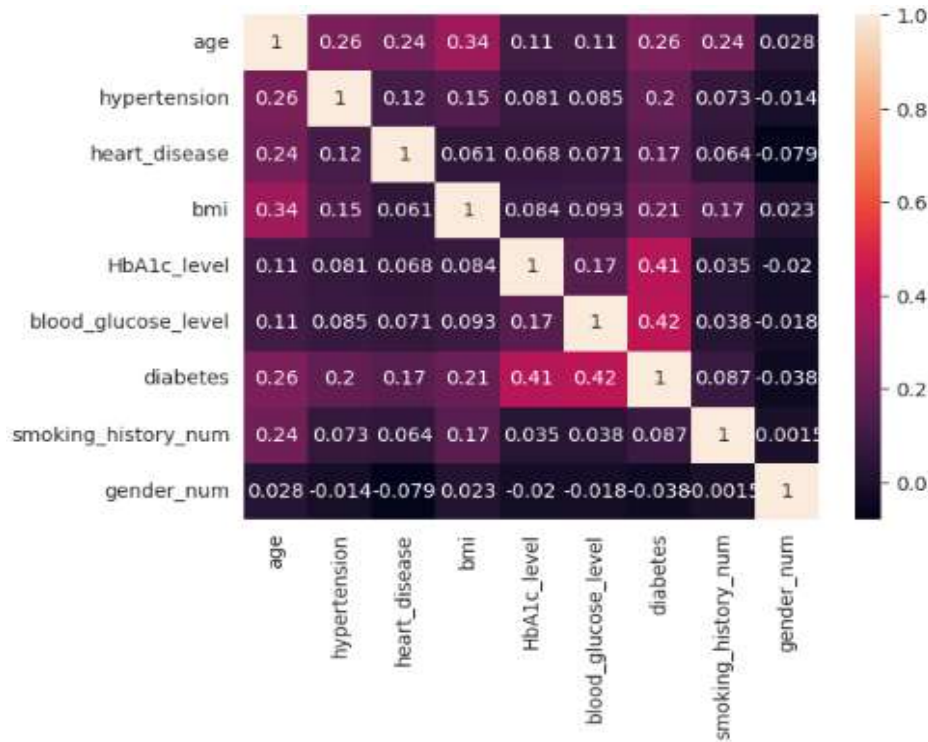


Figure 3.8: Shows the Selected Features

The figure 3.8 displays the distribution of actual and predicted classes for a machine learning classification model as a heatmap of a confusion matrix. The projected classes are represented by the columns of the matrix, and the actual classes are represented by the rows. The percentage of samples that are actually in the row class but were projected to belong to a different class is indicated by the colour intensity in each matrix cell. The matrix's diagonal cells have the deepest colours, indicating that the model has accurately predicted the sample's class in these cells. The model's forecast appears to be inaccurate, as indicated by the lighter cells off the diagonal. As an illustration, the light blue cell in the row labelled "hypertension" and the column labelled "heart disease" shows that certain samples, which were hypertensive in reality, were predicted to have heart disease. Overall, the confusion matrix demonstrates that although the machine learning model is not perfect in its predictions, it is still performing well in accurately identifying the data.

3.9 Target Variable

As the target variable in this medical prediction effort, diabetes is the main focus. This key outcome variable has enormous clinical implications and directs the learning process of our model to precisely identify people who are susceptible to this chronic illness. Through a thorough analysis of an extensive collection of features, the model aims to establish a strong correlation between these predictors and the existence or lack of diabetes, providing medical professionals with important information for prompt intervention and better patient outcomes.

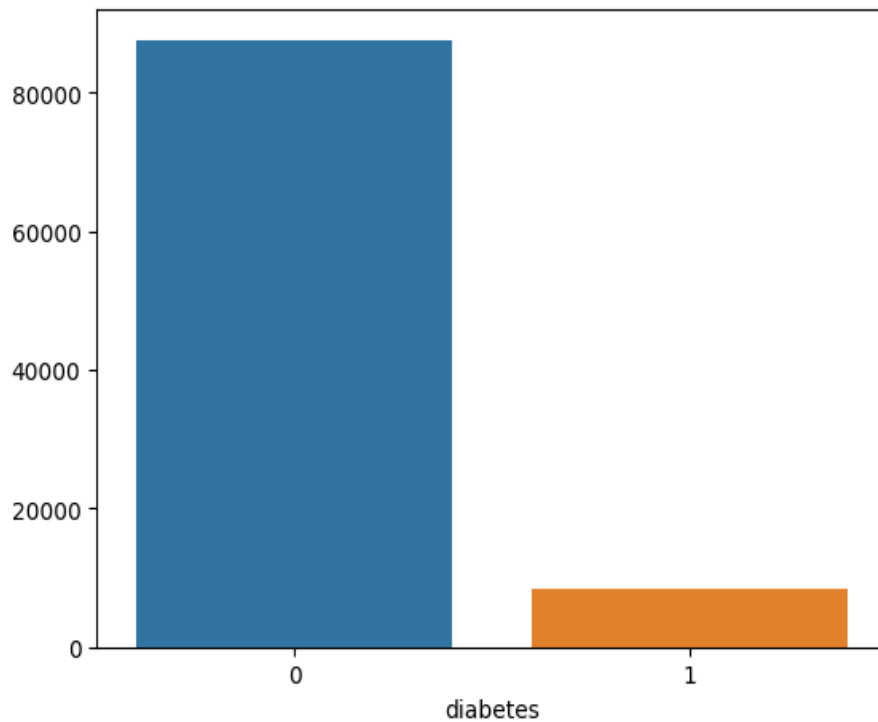


Figure 3.9: Count plot of Diabetes Column

A count plot is used as a bar chart in Figure 3.9 to represent the frequency of data for both those with diabetes and those without it. The number of people without diabetes is shown by the blue line, and the number of people with diabetes is shown by the orange line. This figure provides a clear visual assessment of the difference in distribution between the two groups, providing a brief summary of the incidence of diabetes in the dataset and emphasizing the difference in the numbers of individuals with and without diabetes.

3.10 Machine Learning and Deep Learning Algorithms Used

As the main goal is to predict the early stage of diabetes, The model went through a process of trial and error to come up with a short list of algorithms that yield better results. In this research, Applied various machine learning classification methods. The data visualization charts provide a sense of the kinds of algorithms that will work well for the classification task. The machine learning system employs test data to assess the predictive quality of the trained model and uses training data to train models to recognize patterns. A machine learning system compares predictions on the evaluation data set with actual values (often referred to as ground truth) using a range of criteria to assess predicted performance.

In this study, the model will assess nine distinct categories of machine learning and deep learning algorithms which are given below-

1. K-Nearest Neighbor (KNN)
2. Naive Bayes (NB)
3. Decision Tree (DT)
4. Logistic Regression (LR)
5. Random Forest (RF)
6. Support Vector Machine (SVM)
7. Recurrent Neural Network (RNN)
8. Feedforward Neural Network (FNN)
9. Long Short- Term Memory (LSTM)

K-Nearest Neighbor (KNN)

K-Nearest Neighbor, or KNN, is a straightforward but effective machine learning algorithm for regression and classification problems. It predicts the class or value of a new data point by locating the K most similar data points (neighbor) to it. This is done by looking at most of the neighbor. KNN, also known as K-Nearest Neighbor, functions similarly to a detective by recognizing new diabetes cases based on how similar they are

to previous patients. It selects the K most comparable patients (witnesses) from the known instances and uses the majority vote among these neighbor to forecast the result for the new patient. Envision KNN collects health indicators such as age, BMI, and blood sugar, then applying these to identify the most similar records in the medical database. KNN estimates the new patient's diabetes status as best it can by aggregating data from these comparable cases. KNN appears to be a useful detective in this early-stage diabetes diagnosis challenge, even though it requires some tweaks to function at its best.

Naive Bayes (NB)

A family of probabilistic classifiers called Naive Bayes is employed in machine learning for classification applications. Think of a detective using a more straightforward method to solve a riddle, such as the diagnosis of diabetes. Rather than thoroughly examining every hint collectively, Naive Bayes concentrates on each clue (elements such as age, blood sugar, etc.) separately, presuming that they do not impact one another, much way witness statements do when examine separately. With a twist, Naive Bayes solves the diabetic problem in this model much like a detective would. Rather than concentrating on identifying similar suspects (as in KNN), it examines indicators (features) such as age, blood sugar, and BMI separately, presuming they don't affect one another in the same way that a witness's testimony wouldn't depend on another's. Naive Bayes determines the likelihood of having diabetes based on each of these discrete cues, then use that information to predict the outcome. This Naive Bayes model's objective is to correctly identify patients as having diabetes or not by evaluating each patient's health data separately and computing the total likelihood of having diabetes from these individual probabilities. Consider computing the likelihood of rain depending on wind speed, humidity, and cloud cover individually, then combining them to produce a final forecast.

Decision Tree (DT)

A strong machine learning technique called a decision tree is utilized for regression (which predicts continuous values) as well as classification (which predicts categories). Think

about it like a tree with leaves and branches, where each branch is a choice or query depending on the characteristics of a data item. A leaf that indicates the expected class or value for the data point can be found by following these decision routes. In this model, Decision Tree diagnoses diabetes by posing a series of yes/no questions regarding a patient's health in the manner of a medical adviser. With each inquiry at a branch and the final diagnosis—diabetic or not—at the leaves, it constructs a structure resembling a tree. Upon arrival, a new patient is guided through the tree by the code, which asks questions based on their data (age, blood sugar, etc.) until it reaches a leaf, which indicates whether the patient has diabetes. The model was fine-tuned to select the most illuminating queries and critical decision points for precise diagnosis. This model mimics a doctor's diagnosis process by asking a series of pertinent questions regarding a patient's health in order to appropriately identify the patient as diabetes or not. By creating a decision tree that effectively divides patients according to their characteristics and provides a clear yes/no response regarding their diabetes status, it seeks to accomplish this.

Logistic Regression (LR)

A strong and popular statistical method for machine learning classification tasks is logistic regression. Think of it as a mathematical model that takes a set of independent variables (like age, blood sugar, etc.) and uses them to estimate the chance that an event (like getting diabetes) will occur. It determines the likelihood of falling into a specific category rather than making categorization predictions like "Yes" or "No". Based on the patient data that is currently available, Logistic Regression predicts the probability of diabetes in this model, much like a statistician would. It employs statistics and probabilities to predict the "odds" of a patient acquiring diabetes given their health information, as opposed to creating a decision tree or identifying nearest neighbor. Consider computing the probability of rolling a particular number on a die by considering its weight and form. Here, blood sugar, age, and other factors are used to determine a person's likelihood of having diabetes using logistic regression. This Logistic Regression model's objective is to precisely forecast a patient's likelihood of acquiring diabetes by evaluating their medical records and allocating a score ranging from 0 to 1. A number nearer 1 denotes a higher possibility of diabetes,

whereas a number nearer 0 denotes a lesser probability. Physicians can choose the best course of action for additional diagnostic testing or treatment by knowing these probabilities.

Random Forest (RF)

In machine learning, Random Forest is a strong and adaptable ensemble learning technique that can be applied to both classification and regression problems. Consider a group of varied investigators, each with a special method for cracking cases. The detectives in this case are individual decision trees, and the case involves determining whether someone has diabetes. Random Forest functions as a group of investigators looking into diabetes, each with an own strategy. In this model. Consider a scenario in which each investigator analyses patient data, such as age, blood sugar, and BMI, using a different decision tree. The twist is that each tree makes use of a random selection of attributes and decision points. Because of this diversity, there is collective wisdom that surpasses that of any one detective. The team casts vote upon the arrival of a new patient, with most predicting "diabetic" or "not diabetic." This Random Forest model uses the combined knowledge of several decision trees to accurately categorize patients as either diabetic or not. The model avoids overfitting and guarantees a more reliable and generalizable prediction for unknown data by incorporating randomness into the creation of each tree.

Support Vector Machine (SVM)

In the machine learning toolkit, support vector machines (SVMs) are particularly effective for classification tasks. If distinct classes are represented by a scattering of data points in a high-dimensional space. Finding the ideal separation line, also known as a hyperplane, that maximizes the distance between these classes is the goal of an SVM. Imagine it as creating a broad moat between two armies. The model is more reliable and accurate thanks to this wide margin, even when dealing with previously untested data points. The hyperplane is a straight line if you must build an SVM with a linear kernel. This works well if the data is naturally divided into linear segments; but, in more complicated scenarios, SVMs can map

the data into higher dimensions where separation is easier by using "kernel tricks". The goal is to use my training data to train the SVM ("building the moat") so that it can be used to predict new data point classes with accuracy ("crossing the enemy lines"). The accuracy, confusion matrix, and classification report may all be used to assess how well our SVM performed in achieving the goal of a definite and secure class separation.

Recurrent Neural Network (RNN)

Recurrent neural networks, or RNNs, are a unique class of artificial neural networks intended to process sequential input, including speech, text, and time series data. RNNs could "remember" prior inputs and utilize that knowledge to process the current input and provide predictions, in contrast to typical neural networks that handle each input independently. Using patient data, the model attempts to forecast diabetes in its early stages. It makes use of RNNs' capacity to handle sequential data, possibly identifying significant patterns in the data that more straightforward models would overlook. The model is composed of a single neuron layer for binary classification after an RNN layer with 32 units. The model learns from labelled data during training to find patterns that differentiate people with diabetes from those without the disease. Accuracy and confusion matrix measures are used to evaluate its performance. By potentially identifying pertinent temporal correlations within the data, our RNN technique presents a possible option for early diabetes prediction. Still, to ascertain which strategy works best for this particular assignment, it is imperative that you evaluate its performance in comparison to other approaches.

Feedforward Neural Network (FNN)

A basic kind of artificial neural network known as a feedforward neural network (FNN) is distinguished by its uncomplicated information flow and layered, basic structure. In a FNN, data flows across connected layers without loops or feedback, in contrast to more intricate structures such as Recurrent Neural Networks (RNNs). Unlike the RNN, the FNN (Feedforward Neural Network) model uses a different strategy to forecast diabetes in its

early stages. Rather than concentrating on consecutive patterns, it examines every patient trait separately to find non-linear correlations between them and the risk of developing diabetes. The model uses Dropout layers to avoid overfitting and numerous Dense layers with activation functions to gradually extract these associations. Ultimately, it uses a single neuron with sigmoid activation to produce a binary prediction—diabetic or non-diabetic.

Long Short- Term Memory (LSTM)

In contrast to the FNN, which concentrates on features' static associations, the LSTM delves deeply into the temporal dynamics of patient data. With memory cells and gates, this potent recurrent neural network is excellent at recalling and applying previous knowledge found in the data sequence. Imagine it carefully monitoring the ebb and flow of features throughout time, perhaps revealing minute patterns associated with the onset of diabetes. When compared to models that handle each feature separately, the LSTM achieves improved accuracy because of its emphasis on temporal dependencies. The system's capacity to unearth latent temporal insights in the data may result in more accurate and ultimately more insightful forecasts for early-stage diabetes, despite its complexity and processing requirements being higher than those of the FNN. When selecting one of these models, you should carefully weigh the advantages and disadvantages of each in light of your particular set of data and objectives.

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Setup

The suggested diabetes prediction model is trained and tested in this experiment utilizing a nine-attribute dataset that was obtained from Kaggle. Duplicate values handling and encoding are examples of data preprocessing. The research makes use of a wide range of deep learning and machine learning methods, assessing results using confusion matrices and accuracy metrics.

4.2 Experimental Results and Analysis

Precision:

In machine learning and deep learning, precision is a statistic used to assess a classification model's performance, especially in binary classification issues. The precision of the model's positive predictions is a measure of their accuracy. The ratio of true positive predictions to the total of true positive and false positive predictions is how it is defined. The number of TP on the number of TP "+" number of FP is the definition of precision. False positives are instances in which a model that is truly negative is mistakenly classified as positive. To put it mathematically, the following formula is used to determine precision:

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

Recall:

Recall is a statistic used in machine learning and deep learning to assess a classification model's performance, especially in binary classification tasks. It is sometimes referred to as sensitivity or true positive rate. The number of genuine positive predictions divided by the total of true positives and false negatives is known as recall. Stated differently, it

assesses a model's capacity to accurately detect every pertinent instance present in a dataset. The number of true TP divided by the TP "+" FN is known as the recall. The recall formula can be defined by:

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

F1- Score:

A popular measure in deep learning and machine learning for assessing a classification model's effectiveness is the F1 score. When working with imbalanced datasets—that is, datasets with unequal numbers of samples in each class—it is especially helpful. Precision and recall's harmonic mean is the F1 score. Recall and precision are two crucial variables that highlight various facets of categorization performance. F1 Score is required when there is an unequal class distribution (more real negatives) and you want to find a compromise between Precision and Recall. The recall formula can be defined by:

$$\text{F1} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy:

Another popular metric in machine learning and deep learning that is used to assess a classification model's overall performance is accuracy. It shows the proportion of accurately predicted cases to all instances in the dataset. The accuracy formula is:

$$\text{Accuracy} = (\text{TN} + \text{TP}) / (\text{TN} + \text{FP} + \text{FN} + \text{TP})$$

4.3 Machine Learning Models Outcome

The accuracy, recall, F1-Score, and support values of the KNN model's classification performance are compiled in Table 4.3.1. This thorough analysis provides insights into the accuracy of the model by emphasizing its precision (the ability to classify examples properly), recall (the ability to catch relevant instances), and F1-Score (the ability to balance the two), all supported by appropriate support values.

Table 4.3.1: Classification Report of K- Nearest Neighbor (KNN) Model

	Precision	Recall	F1- Score	Support
0.0	0.96	0.99	0.97	17562
1.0	0.82	0.58	0.68	1668
Accuracy			0.95	19230
Macro Avg	0.89	0.78	0.83	19230
Weighted	0.95	0.95	0.95	19230

Table 3.4.2 provides a detailed evaluation of the Naive Bayes model's classification performance by presenting the model's precision, recall, F1-Score, and support values.

Table 4.3.2: Classification Report of Naive Bayes (NB) Model

	Precision	Recall	F1- Score	Support
0.0	0.93	0.99	0.96	17562
1.0	0.63	0.21	0.31	1668
Accuracy			0.92	19230
Macro Avg	0.78	0.60	0.63	19230
Weighted	0.90	0.92	0.90	19230

Table 4.3.3 provides a thorough assessment of the Decision Tree model's classification performance by displaying the model's precision, recall, F1-Score, and support values.

Table 4.3.3: Classification Report of Decision Tree (DT) Model

	Precision	Recall	F1- Score	Support
0.0	0.97	1.00	0.98	17562
0.1	1.00	0.68	0.81	1668
Accuracy			0.97	19230
Macro Avg	0.99	0.84	0.90	19230
Weighted	0.97	0.97	0.97	19230

Table 4.3.4 provides a thorough overview of the Logistic Regression model's precision, recall, F1-Score, and support values. It provides information on the model's capacity to capture true positives, balance between precision and recall overall, and number of instances in each class, among other things.

Table 4.3.4: Classification Report of Logistic Regression (LR) Model

	Precision	Recall	F1- Score	Support
0.0	0.95	0.98	0.97	17562
0.1	0.71	0.47	0.57	1668
Accuracy			0.94	19230
Macro Avg	0.83	0.73	0.77	19230
Weighted	0.93	0.94	0.93	19230

A detailed analysis of the Random Forest model's precision, recall, F1-Score, and support metrics is provided in Table 4.3.5. This table sheds light on the model's overall classification capabilities by offering useful information on its recall ability, accuracy in detecting positive cases, harmonic mean of precision and recall, and distribution of instances across different classes.

Table 4.3.5: Classification Report of Random Forest (RF) Model

	Precision	Recall	F1- Score	Support
0.0	0.97	1.00	0.98	17562
1.0	0.95	0.69	0.80	1668
Accuracy			0.97	19230
Macro Avg	0.96	0.84	0.89	19230
Weighted	0.97	0.97	0.97	19230

Table 4.3.6 presents the SVM model's precision, recall, F1-Score, and support values, providing a thorough assessment of its classification performance. It explores the memory capacity of the model, how well it predicts positive examples, how well accuracy and recall are balanced, and how instances are distributed across various classes.

Table 4.3.6: Classification Report of SVM Model

	Precision	Recall	F1- Score	Support
0.0	0.95	0.98	0.97	17562
0.1	0.72	0.45	0.55	1668
Accuracy			0.94	19230
Macro Avg	0.84	0.71	0.76	19230
Weighted	0.93	0.94	0.93	19230

4.4 Deep Learning Models Outcome

A thorough assessment of the RNN model's classification performance is provided by the precision, recall, F1-Score, and support numbers shown in table 4.4.1. It offers information about the recall capacity, the harmonic mean of precision and recall, the distribution of examples among various classes, and the model's accuracy in forecasting positive cases.

Table 4.4.1: Classification Report of RNN Model

	Precision	Recall	F1- Score	Support
0.0	0.97	1.00	0.98	17562
0.1	1.00	0.67	0.80	1668
Accuracy			0.97	19230
Macro Avg	0.98	0.84	0.89	19230
Weighted	0.97	0.97	0.97	19230

Figure 4.4.2 shows an RNN model's confusion matrix. The model's classification performance is represented on the blue heatmap. The most common outcomes—accurate negative forecasts and wrong positive predictions, respectively—are shown by the darkest hues at the top left and bottom right places. The model's overall accuracy and propensity to commit all kinds of errors can be determined by examining the value distribution. Details on the task of the model, the classes involved, and the properties of the dataset, however, are essential for a more nuanced understanding. We can explore the findings from the

matrix in more detail and pinpoint areas that can be used to improve with this extra context. in the upper left cell, is True Negatives (TN): 17558. The cell on top right had 547 False Negatives (FN). The model predicts 4 false positives. Positioned in the lower left cell, True Positives (TP): 1121.

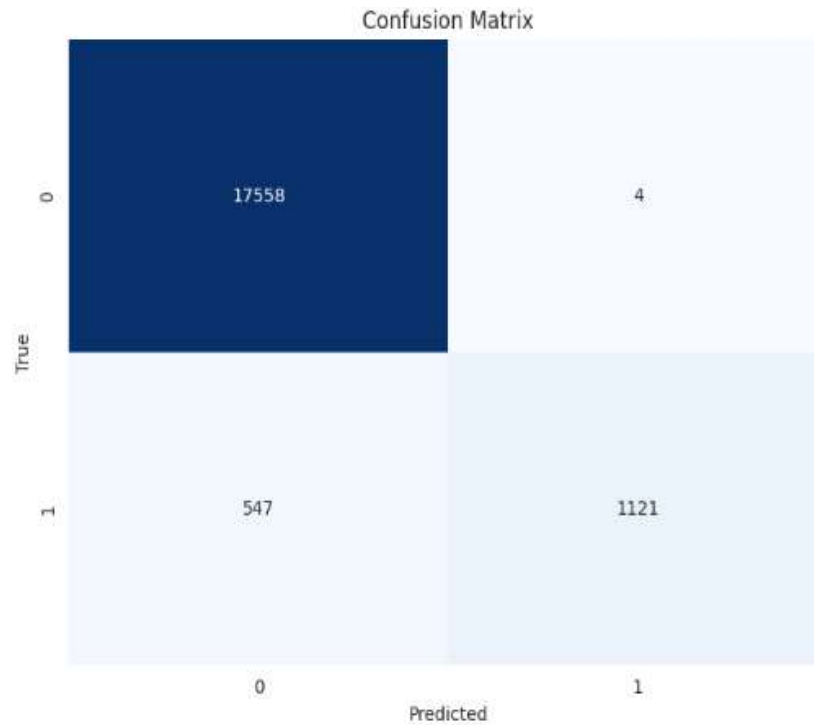


Figure 4.4.2: Confusion Matrix for RNN Model

An RNN model's learning curve is shown in Figure 4.4.3, which shows an early increase in training and validation accuracy, indicating quick understanding. Only five epochs later, though, a plateau becomes apparent, suggesting an excess of fit. Notably, the validation accuracy of the model only reaches 96%, lagging behind the training accuracy of 97% by epoch 25, indicating that the model struggles with generalization to new data. This discrepancy highlights how difficult it is for the model to successfully extrapolate learned patterns to new cases, highlighting its shortcomings in obtaining strong generalization over a variety of datasets and highlighting the necessity of resolving overfitting issues in order to improve overall model performance.



Figure 4.4.3: Train- Test Model Loss (RNN)

For the Feedforward Neural Network (FNN) model, Table 4.4.4 summarizes a thorough examination of the precision, recall, F1-Score, and support metrics. With regard to the model's classification efficacy, this presentation provides a thorough knowledge by revealing important details like its recall capacity, precision in detecting affirmative cases, distribution of instances across various classes, and harmonic mean of precision and recall.

Table 4.4.4: Classification Report of FNN Model

	Precision	Recall	F1- Score	Support
0.0	0.97	1.00	0.98	17562
0.1	0.99	0.68	0.80	1668
Accuracy			0.97	19230
Macro Avg	0.98	0.84	0.89	19230
Weighted	0.97	0.97	0.97	19230

A confusion matrix summarizing a classification model's performance is displayed in Figure 4.4.5. In this instance, the model is attempting to categorize a person as confused or

not. The target variable's actual values are represented by the rows of the matrix, while its anticipated values are represented by the columns. The number of accurate predictions the model produced is displayed in the matrix's diagonal cells. The off-diagonal cells display how many of the model's predictions were off.

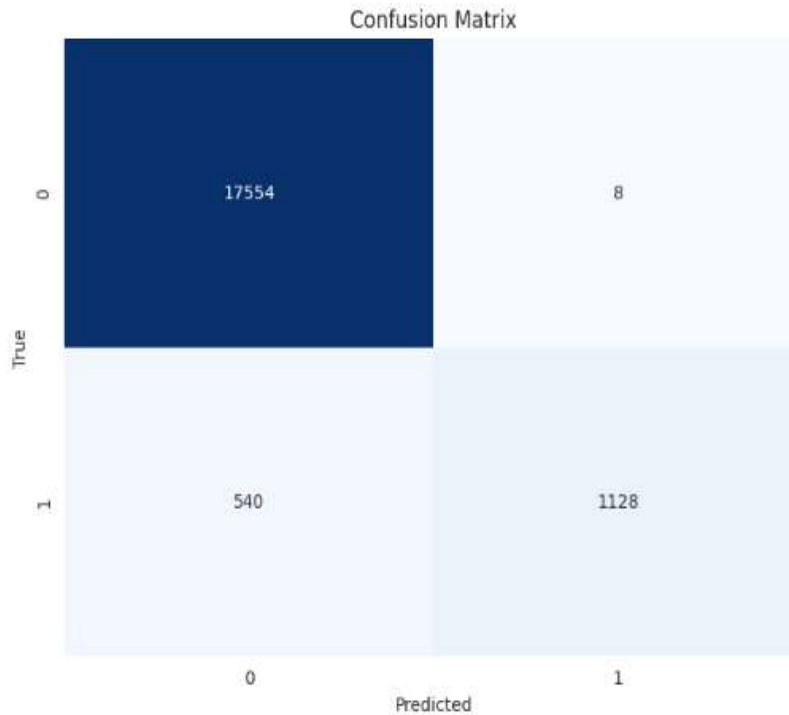


Figure 4.4.5: Confusion Matrix for FNN Model

A line graph representing a feedforward neural network (FNN) model's training and validation accuracy may be found in Figure 4.4.6. The model's accuracy on the training set of data is known as the training accuracy, and its accuracy on the validation data, which is a different set of data, is known as the validation accuracy. The reason behind the training accuracy being higher than the validation accuracy is usually the overfitting of the model to the training set. When a model exhibits overfitting, it is unable to generalize to new data because it has learned the training set too thoroughly. The graphic illustrates that the validation accuracy begins at approximately 0.93 and rises to approximately 0.96, while the training accuracy begins at roughly 0.92 and rises to approximately 0.97.

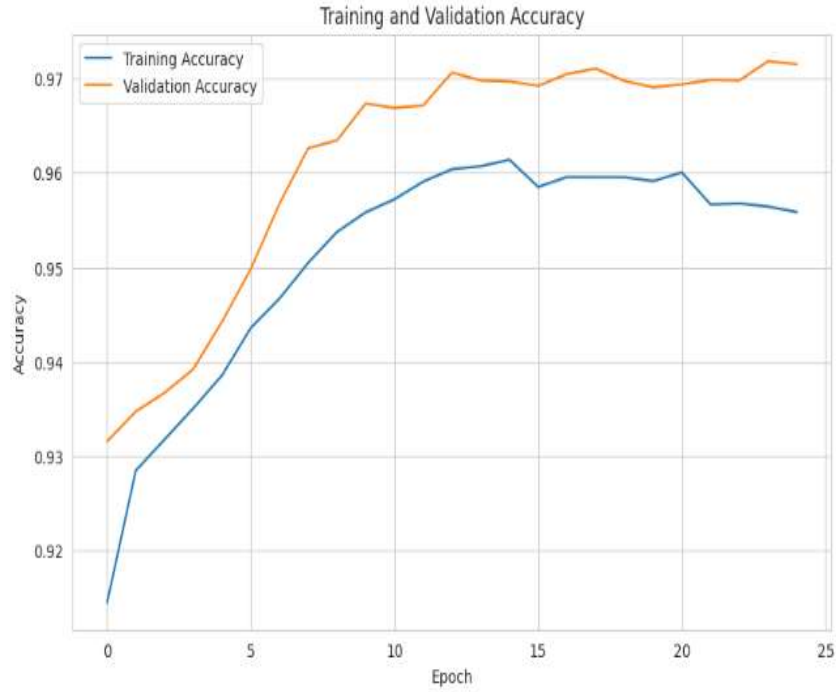


Figure 4.4.6: Train- Test Model Loss (FNN)

Table 4.4.7 presents a detailed analysis of the Long Short-Term Memory (LSTM) model's precision, recall, F1-Score, and support metrics. This comprehensive presentation shows the model's memory ability, balance between recall and precision, distribution of cases across different classes, and accuracy in recognizing positive instances, offering a comprehensive view of its categorization capabilities.

Table 4.4.7: Classification Report of LSTM Model

	Precision	Recall	F1- Score	Support
0.0	0.97	1.00	0.98	17562
0.1	1.00	0.68	0.81	1668
Accuracy			0.97	19230
Macro Avg	0.98	0.84	0.90	19230
Weighted	0.97	0.97	0.97	19230

Figure 4.4.8 shows how well an LSTM model performed in dividing data into two groups, denoted by 0 and 1. It displays the number of times the model identified data points correctly (hits) and the number of times it misclassified data points (misses). 17560 True negatives, 1127 True positives, 541 False negatives, and 2 False positives are predicted by the model.



Figure 4.4.8: Confusion Matrix for LSTM Model

Figure 4.4.9 shows the training and validation accuracy curve for an LSTM model. The accuracy of a model on training data is called training accuracy; on a different set of data that was not used for training, it is called validation accuracy. Validation accuracy offers a more accurate estimate of the model's performance on new data since it is resistant to overfitting of the training set. The graph shows that the training accuracy is higher than the validation accuracy, which is an adversative indication. It is likely that the model is learning new abilities from the training set. The accuracy difference between it and RNN and FNN is not very great, yet it still outperforms them.

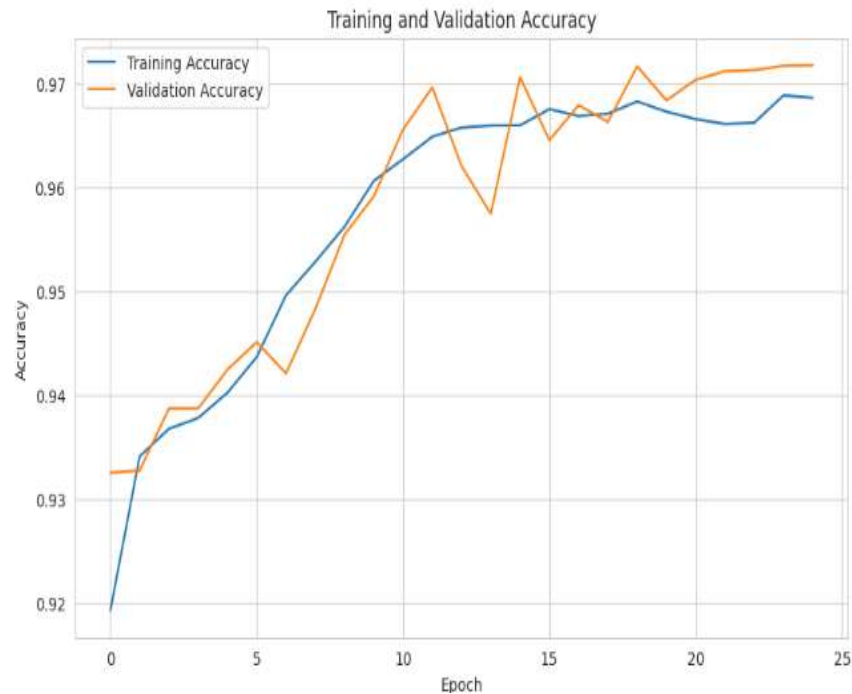


Figure 4.4.9: Train- Test Model Loss (LSTM)

4.5 Discussion

Commendable accuracy of 95% indicates the K-Nearest Neighbors (KNN) model's dependability in accurately classifying cases. However, a closer look at the metrics for precision, recall, and F1-Score shows a trade-off, particularly in recall for positive cases (1.0), suggesting potential difficulties in accurately identifying people with diabetes. In spite of this, the model performs quite well overall. 92% of the time, the Naive Bayes (NB) model shows impressive prediction ability. Nonetheless, it has recall deficits for positive examples, which reflects difficulties correctly diagnosing diabetics. Although the NB model is good at predicting cases that are not diabetic, it may overlook some cases of diabetes, according to the precision-recall trade-off. The models that provide the highest accuracy of 97% are the Decision Tree (DT) and Random Forest (RF). Their high precision, recall, and F1-Score values for both classes demonstrate how well these models capture complex relationships within the dataset. In comparison to Random Forest, the Decision Tree specifically exhibits better precision and F1-Score, highlighting its efficacy in diabetes prediction. At a remarkable 94% accuracy rate, the Support Vector Machine

(SVM) and Logistic Regression (LR) models achieve a good balance between recall and precision. While recall for positive examples might be improved, these models do a good job of predicting both diabetic and non-diabetic patients. This suggests potential areas for further model refining. With an accuracy rate of 97%, the Recurrent Neural Network (RNN), Feedforward Neural Network (FNN), and Long Short-Term Memory (LSTM) models are the most accurate deep learning models. Nonetheless, the noteworthy trade-off between precision and recall in positive cases indicates that possible overfitting needs to be carefully taken into account. It is possible to investigate regularization strategies to improve generalization without sacrificing accuracy. In conclusion, the assessment emphasizes the trade-offs between recall and precision while highlighting the complex performance of each model. While deep learning models show great accuracy but need more examination for possible overfitting, the decision tree stands out for its higher precision and F1-Score. The efficacy of diabetes prediction models in practical applications will be enhanced by additional model optimization, investigation of regularization techniques, and cautious assessment of interpretability.

Table 4.5: Comparison Table of ML and DL Classification Results

Models	Accuracy	Precision (Avg)	Recall (Avg)	F-1 score (Avg)
KNN	95%	0.89	0.78	0.83
NB	92%	0.78	0.60	0.63
DT	97%	0.99	0.84	0.90
LR	94%	0.83	0.73	0.77
RF	97%	0.96	0.84	0.89
SVM	94%	0.84	0.71	0.76
RNN	97%	0.98	0.84	0.89
FNN	97%	0.98	0.84	0.89
LSTM	97%	0.98	0.84	0.90

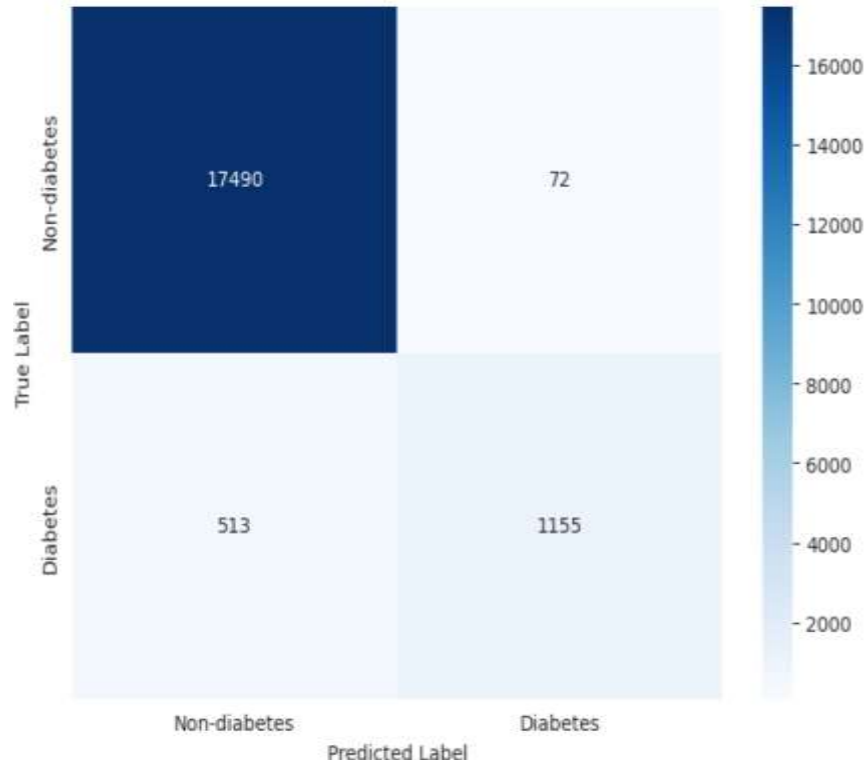


Figure 4.5: Confusion Matrix for Decision Tree Model

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

For society as a whole, the application of diabetes prediction models is important. Potentially lessening the burden of complications associated to diabetes, the models, particularly with their high accuracy rates, can aid in early detection and intervention. Better patient outcomes result from these models' ability to empower timely and well-informed decision-making among individuals and healthcare providers. The influence on healthcare resources, policy, and public health policies is a result of the societal impact that goes beyond direct health. To make sure that different socioeconomic groups may benefit from these models, it is imperative to address concerns of inclusion and accessibility.

5.2 Impact on Environment

Although diabetes prediction models have an indirect effect on the environment, their use can support the sustainability of healthcare systems as a whole. These models may lessen the need for substantial and resource-intensive medical treatments by supporting early illness prevention and management, hence reducing the environmental impact of healthcare practices. It is important to take into account the environmental effects of the computing resources and data centers that make up the technical infrastructure that underpins these models. Potential environmental risks can be reduced by putting energy-efficient procedures into place and looking into eco-friendly computer options.

5.3 Ethical Aspects

Diabetes prediction models raise a number of ethical issues, such as consent, bias, interpretability, privacy of data, and interpretability. Ensuring the confidentiality and privacy of personal health information is critical, necessitating strong data anonymization and security protocols. Ethical deployment requires addressing model biases to avoid differences in predictions between demographic groups. Moreover, building confidence

between healthcare providers and users depends on transparent model interpretability. To address ethical problems, it is essential to implement informed consent methods, educate users about the capabilities and limitations of the models, and develop explicit guidelines for responsible model use.

5.4 Sustainability Plan

For diabetes prediction models to continue to be successfully used and responsibly deployed, a thorough sustainability plan is essential. This strategy should include continual model monitoring and assessment to gauge its effectiveness in practical situations. Prioritizing regular upgrades and enhancements based on growing medical knowledge and technology breakthroughs is important. To adapt the concept to various healthcare contexts, cooperation amongst healthcare stakeholders—such as practitioners, legislators, and community representatives—is crucial. The model's long-term viability will also be enhanced by a dedication to moral principles and continual public involvement programs, which will build confidence and guarantee the model's conformity with society norms. The deployment of the approach will be even more sustainable overall if eco-friendly technologies are adopted and regular environmental impact evaluations are conducted.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the Study

This study used a variety of machine learning and deep learning techniques to create and assess diabetes prediction models. Data preprocessing methods, such as handling duplicate values and encoding, were used on a nine-attribute dataset obtained from Kaggle. K-Nearest Neighbors (KNN), Naïve Bayes (NB), decision tree (DT), random forest (RF), support vector machine (SVM), logistic regression (LR), recurrent neural network (RNN), feedforward neural network (FNN), and long short-term memory (LSTM) models were used in the study. The models were evaluated according to their overall accuracy, recall, F1-Score, and precision.

6.2 Conclusion

With differing degrees of precision, the experimental findings show how well the models predict diabetes. Among these, KNN, Naive Bayes, SVM, and Logistic Regression models performed exceptionally well, with accuracy rates exceeding 90%. Decision Tree and Random Forest provide 97% accuracy also. RNN, FNN, and LSTM, three deep learning models, outperformed with amazing 97% accuracy rates. But the possibility of overfitting in FNN and RNN models points to the necessity of regularization techniques. Different models performed differently when it came to correcting class disparities, which highlights the need for more development. The interpretability of traditional models, such as Decision Trees, is essential for real-world applications. In general, the models exhibit potential in aiding in the early identification and treatment of diabetes.

6.3 Recommendation

The results suggest that more research be done on ensemble approaches, which integrate the advantages of several models to improve overall prediction performance. Additionally,

to reduce the risk of overfitting, efforts should be focused on improving deep learning models. The practical value of the models will be improved by addressing class imbalances and enhancing interpretability, particularly in deep learning models. Model accuracy can be further improved by working with healthcare practitioners to gain domain-specific insights and by adding more pertinent features to the dataset. Continuous observation and model updates ought to be essential to guarantee the models' flexibility in responding to changing healthcare situations.

6.4 Implication for Further Study


The study provides opportunities for more research in a number of areas. Examining how various feature sets affect model performance may shed light on the significance of particular features and direct the gathering of data. Ensuring model generality requires investigating the transferability of models across various demographic groupings and healthcare systems. A thorough grasp of the relative advantages and disadvantages of sophisticated machine learning/deep learning models and conventional statistical models can be obtained through comparative studies between them. Assessing the models' socioeconomic impact and evaluating their actual application in clinical settings will provide useful information for politicians and healthcare professionals. In order to address ethical problems and promote broader adoption of predictive models in healthcare, it is imperative that ongoing research be conducted in the areas of model interpretability, fairness, and openness.

REFERENCES

- [1] Sonar P, JayaMalini K, “Diabetes prediction using different machine learning approaches,” *In2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 367-371, 27 March 2019.
- [2] Jaiswal, Varun, Anjali Negi, and Tarun Pal, “A review on current advances in machine learning based diabetes prediction,” *Primary Care Diabetes*, vol. 15, pp. 435-443, 26 February 2021. Available at: <https://doi.org/10.1016/j.pcd.2021.02.005>
- [3] Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, Munam Ali Shah, “Prediction of diabetes using machine learning algorithms in healthcare,” *2018 24th international conference on automation and computing (ICAC)*, pp. 1-6, September 2018. Available at: <https://doi.org/10.23919/ICAC.2018.8748992>
- [4] Kumar, Y. Jeevan Nagendra, N. Kameswari Shalini, P. K. Abhilash, K. Sandeep, D. Indira, “Prediction of diabetes using machine learning,” *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, pp. 2547-2551, May 2019.
- [5] Hakim El Massari, Zineb Sabouri, Sajida Mhammedi, Noreddine Gherabi, “Diabetes prediction using machine learning algorithms and ontology,” *Journal of ICT Standardization*, vol. 10, pp. 319-337, May 2022. Available at: <https://doi.org/10.13052/jicts2245-800X.10212>
- [6] Chowdary, P. Bharath Kumar, 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*, vol. 12, pp. 519-525, 2021.
- [7] Taiyu Zhu, Kezhi Li, Pau Herrero, Pantelis Georgiou, “Deep learning for diabetes: a systematic review,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 2744-2757, 2020.
- [8] Shafi, Salliah, Gufran Ahmad Ansari, “Early prediction of diabetes disease & classification of algorithms using machine learning approach,” *Proceedings of the International Conference on Smart Data Intelligence*, vol. 12, pp. 519-525, 2021.
- [9] Llah, Olta, Amarildo Rista, “Prediction and Detection of Diabetes using Machine Learning,” *RTA-CSIT*. 2021.
- [10] Samet, Sarra, Mohamed Ridda Laouar, Issam Bendib, “Diabetes mellitus early stage risk prediction using machine learning algorithms,” *2021 International Conference on Networking and Advanced Systems (ICNAS)*, pp. 1-6, December 2021. Available at: <https://doi.org/10.1109/ICNAS53565.2021.9628955>

[11] Swapna, Goutham, Soman Kp, and Ravi Vinayakumar. "Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals." *Procedia computer science* 132 (2018): 1253-1262. Available at: <https://doi.org/10.21203/rs.3.rs-3145599/v1>

[12] Khaleel, Fayroza Alaa, and Abbas M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms," *Materials Today: Proceedings*, vol. 80, pp. 3200-3203, 2023. Available at: <https://doi.org/10.1016/j.matpr.2021.07.196>

<p>Plagiarism Checked by Ms. Fabliha Haque, Lecturer Department of CSE Daffodil International University</p>	
--	---

ORIGINALITY REPORT		
<div style="font-size: 2em; color: #c00000; font-weight: bold;">24%</div> <p style="font-size: 0.8em; color: #c00000;">SIMILARITY INDEX</p>	<div style="font-size: 2em; color: #c00000; font-weight: bold;">20%</div> <p style="font-size: 0.8em; color: #c00000;">INTERNET SOURCES</p>	<div style="font-size: 2em; color: #c00000; font-weight: bold;">10%</div> <p style="font-size: 0.8em; color: #c00000;">PUBLICATIONS</p>
<div style="font-size: 2em; color: #c00000; font-weight: bold;">13%</div> <p style="font-size: 0.8em; color: #c00000;">STUDENT PAPERS</p>		
PRIMARY SOURCES		
<div style="background-color: #c00000; color: white; padding: 5px; width: 20px; margin: 0 auto;">1</div>	<p style="color: #c00000; font-weight: bold;">dspace.daffodilvarsity.edu.bd:8080</p> <p style="font-size: 0.8em; color: #c00000;">Internet Source</p>	<div style="font-size: 1.5em; color: #c00000; font-weight: bold;">5%</div>
<div style="background-color: #800080; color: white; padding: 5px; width: 20px; margin: 0 auto;">2</div>	<p style="color: #800080; font-weight: bold;">Submitted to Daffodil International University</p> <p style="font-size: 0.8em; color: #800080;">Student Paper</p>	<div style="font-size: 1.5em; color: #800080; font-weight: bold;">3%</div>
<div style="background-color: #4b0082; color: white; padding: 5px; width: 20px; margin: 0 auto;">3</div>	<p style="color: #4b0082; font-weight: bold;">Submitted to Huntington Beach Union High School District</p> <p style="font-size: 0.8em; color: #4b0082;">Student Paper</p>	<div style="font-size: 1.5em; color: #4b0082; font-weight: bold;">1%</div>
<div style="background-color: #008080; color: white; padding: 5px; width: 20px; margin: 0 auto;">4</div>	<p style="color: #008080; font-weight: bold;">dokumen.pub</p> <p style="font-size: 0.8em; color: #008080;">Internet Source</p>	<div style="font-size: 1.5em; color: #008080; font-weight: bold;">1%</div>
<div style="background-color: #008000; color: white; padding: 5px; width: 20px; margin: 0 auto;">5</div>	<p style="color: #008000; font-weight: bold;">github.com</p> <p style="font-size: 0.8em; color: #008000;">Internet Source</p>	<div style="font-size: 1.5em; color: #008000; font-weight: bold;">1%</div>
<div style="background-color: #806400; color: white; padding: 5px; width: 20px; margin: 0 auto;">6</div>	<p style="color: #806400; font-weight: bold;">ir.lib.uwo.ca</p> <p style="font-size: 0.8em; color: #806400;">Internet Source</p>	<div style="font-size: 1.5em; color: #806400; font-weight: bold;">1%</div>
<div style="background-color: #654321; color: white; padding: 5px; width: 20px; margin: 0 auto;">7</div>	<p style="color: #654321; font-weight: bold;">assets.researchsquare.com</p> <p style="font-size: 0.8em; color: #654321;">Internet Source</p>	<div style="font-size: 1.5em; color: #654321; font-weight: bold;">1%</div>
<div style="background-color: #000080; color: white; padding: 5px; width: 20px; margin: 0 auto;">8</div>	<p style="color: #000080; font-weight: bold;">doctorpenguin.com</p> <p style="font-size: 0.8em; color: #000080;">Internet Source</p>	<div style="font-size: 1.5em; color: #000080; font-weight: bold;"><1%</div>
<div style="background-color: #800080; color: white; padding: 5px; width: 20px; margin: 0 auto;">9</div>	<p style="color: #800080; font-weight: bold;">Submitted to Damonte Ranch High School</p> <p style="font-size: 0.8em; color: #800080;">Student Paper</p>	<div style="font-size: 1.5em; color: #800080; font-weight: bold;"><1%</div>