VISVESVARAYA TECHNOLOGICAL UNIVERSITY
"Jnana Sangama", Belagavi-560014, Karnataka

A PROJECT REPORT ON

"PREDICTING EARLY STAGE OF GESTATIONAL DIABETES USING DEEP LEARNING DURING PREGNANCY"

SUBMITTED IN PARTIAL FUFILLMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE

Bachelor of Engineering
In
Computer Science & Engineering

Submitted By

Asfa Khanum (1SV20CS003)

Borish Kongbrailatpam (1SV20CS006)

Jahidul Islam (1SV20CS015)

Roushni Begum(1SV20CS036)

Under the guidance of

Prof. Suthan R B.E., M.Tech.,
Assistant Professor, Dept. of CSE.

SHRIDEVI
EDUCATION

Department of Computer Science and Engineering

SHRIDEVI INSTITUTE OF ENGINEERING AND TECHNOLOGY
(Affiliated To Visvesvaraya Technological University)
Sira Road, Tumakuru – 572106, Karnataka.

2023-24

Sri Shridevi Charitable Trust (R.)

ESTD:2002

SHRIDEVI INSTITUTE OF ENGINEERING AND TECHNOLOGY

Sira Road, Tumkur - 572 106, Karnataka, India.

Phone: 0816 - 2212629 | Principal: 0816 - 2212627, 9686114899 | Telefax: 0816 - 2212628

Email: info@shridevierigineering.org, principal@shridevierigineering.org | Website: www.shridevierigineering.org

(Approved by AICTE, New Delhi, Recognised by Govt. of Karnataka and Affiliated to Visvesvaraya Technological University, Belagavi)

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that, Project work of entitled "PREDICTING EARLY STAGE OF GESTATIONAL DIABETES USING DEEP LEARNING DURING PREGNANCY" has been successfully carried out by ASFA KHANUM [1SV20CS003], BORISH KONGBRAILATPAM [1SV20CS006], JAHIDUL ISLAM [1SV20CS015], ROUSHNI BEGUM [1SV20CS036], in partial fulfillment for the award of **Bachelor of Engineering** in **Computer Science & Engineering** of the **Visvesvaraya Technological University, Belagavi** during the academic year **2023-24.** It is certified that all the corrections/suggestions indicated for internal assessments have been incorporated in the report. The Project work report has been approved as it certifies the academic requirements in respect to the project work prescribed for the Bachelor of Engineering Degree.

29/5/24

Signature of Guide
Prof. Suthan R B.E., M.Tech.
Assistant Professor, Dept. of CSE,

29/5/24

Signature of H.O.D
Dr. Basavesha D B.E., M.Tech., PhD.
Associate Professor & HOD, Dept. of CSE.

Signature of Principal

Dr. Narendra Viswanath M.E., PhD
Principal,
SIET, Tumakuru

Name of the Examiners

1. Dr. Basavesha D
2. Rakesh S

Signature with Date

29/5

29/5/24

Sri Shridevi Charitable Trust (R.)
SHRIDEVI INSTITUTE OF ENGINEERING AND TECHNOLOGY
Sira Road, Tumkur - 572 106, Karnataka, India.
Phone: 0816 - 2212629 | Principal: 0816 - 2212627, 9686114899 | Telefax: 0816 - 2212628
Email: info@shrideviengineering.org, principal@shrideviengineering.org | Website: www.shrideviengineering.org
(Approved by AICTE, New Delhi, Recognised by Govt. of Karnataka and Affiliated to Visvesvaraya Technological University, Belagavi)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

We, **ASFA KHANUM [1SV20CS003], BORISH KONGBRAILATPAM [1SV20CS006], JAHIDUL ISLAM [1SV20CS015], ROUSHNI BEGUM [1SV20CS036]**, students of VIII semester **B.E** in Computer Science & Engineering, at Shridevi Institute of Engineering and Technology, Tumakuru, hereby declare that, the Project work entitled **"PREDICTING EARLY STAGE OF GESTATIONAL DIABETES USING DEEP LEARNING DURING PREGNANCY "**, embodies the report of our Project work carried out by our team under the guidance of **Prof. Suthan R, Assistant Professor, Department of CSE**, SIET, Tumakuru as partial fulfillment of requirements for the award of the degree in **Bachelor of Engineering in Computer Science, Visvesvaraya Technological University, Belagavi**, during the academic year **2023-24**. The Project has been approved as it satisfies the academic requirements in respect to the Project work.

Place: Tumakuru

Date: 29/05/24

Student Name And Signature

Asfa Khanum(1SV20CS003) Asfa

Borish K (1SV20CS006) Borish K.

Jahidul Islam (1SV20CS015)

Roushni Begum(1SV20CS036) Roushni

ACKNOWLEDGEMENT

This project work will be incomplete without thanking the personalities responsible for this venture, which otherwise would not have become a reality.

We express our profound gratitude to **Dr. Narendra Viswanath,** Principal, S.I.E.T, for his moral support towards completing our project work.

We would like to thank Head of Department **Dr. Basavesha D** **B. E, MTech, PhD** Head, Department of CSE, SIET for providing all the support and facility.

We would like to thank our guide **Prof. Suthan R** **B.E., M.Tech.,** Assistant Professor, Department of Computer Science and Engineering, SIET for his help, sharing his technical expertise and timely advice.

We would like to thank our Project Coordinators **Dr. Girish L** **B. E, MTech, PhD** Head, Department of AIDS, S.I.E.T, and **Mrs. Rashmi N**, Assistant Professor, Department of CSE, SIET for providing all the support and facility.

We would like to express our sincere gratitude to all teaching and non- teaching faculty of the department of CSE for guiding us of this project by giving valuable suggestion and encouragement.

By,

Asfa Khanum(1SV20CS003)
Borish K (1SV20CS006)
Jahidul Islam(1SV20CS015)
Roushni Begum(1SV20CS036)

VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"Jnana Sangama", Belagavi-560014, Karnataka

A PROJECT REPORT ON

"PREDICTING EARLY STAGE OF GESTATIONAL DIABETES USING DEEP LEARNING DURING PREGNANCY"

SUBMITTED IN PARTIAL FUFILLMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE

Bachelor of Engineering

In

Computer Science & Engineering

Submitted By

Asfa Khanum (1SV20CS003)

Borish Kongbrailatpam (1SV20CS006)

Jahidul Islam (1SV20CS015)

Roushni Begum(1SV20CS036)

Under the guidance of

Prof. Suthan R B.E., M.Tech.,

Assistant Professor, Dept. of CSE.

Department of Computer Science and Engineering

SHRIDEVI INSTITUTE OF ENGINEERING AND TECHNOLOGY

(Affiliated To Visvesvaraya Technological University)

Sira Road, Tumakuru – 572106, Karnataka.

2023-24

Sri Shridevi Charitable Trust (R.)

SHRIDEVI INSTITUTE OF ENGINEERING AND TECHNOLOGY

Sira Road, Tumkur - 572 106, Karnataka, India.

Phone: 0816 - 2212629 | Principal: 0816 - 2212627, 9686114899 | Telefax: 0816 - 2212628

Email: info@shrideviengineering.org, principal@shrideviengineering.org | Website: www.shrideviengineering.org

(Approved by AICTE, New Delhi, Recognised by Govt. of Karnataka and Affiliated to Visvesvaraya Technological University, Belagavi)

ESTD:2002

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that, Project work of entitled "PREDICTING EARLY STAGE OF GESTATIONAL DIABETES USING DEEP LEARNING DURING PREGNANCY" has been successfully carried out by ASFA KHANUM [1SV20CS003], BORISH KONGBRAILATPAM [1SV20CS006], JAHIDUL ISLAM [1SV20CS015], ROUSHNI BEGUM [1SV20CS036], in partial fulfillment for the award of **Bachelor of Engineering** in **Computer Science & Engineering** of the **Visvesvaraya Technological University, Belagavi** during the academic year **2023-24.** It is certified that all the corrections/suggestions indicated for internal assessments have been incorporated in the report. The Project work report has been approved as it certifies the academic requirements in respect to the project work prescribed for the Bachelor of Engineering Degree.

---	---
Signature of Guide	**Signature of H.O.D**
Prof. Suthan R B.E., M.Tech.	**Dr. Basavesha D** B.E., M.Tech., PhD.
Assistant Professor, Dept. of CSE,	Associate Professor & HOD, Dept. of CSE,

----

Signature of Principal

Dr. Narendra Viswanath M.E., PhD

Principal,

SIET, Tumakuru

Name of the Examiners	Signature with Date
1 ………………………………..	….………………………
2.………………………………	………………………..

ACKNOWLEDGEMENT

This project work will be incomplete without thanking the personalities responsible for this venture, which otherwise would not have become a reality.

We express our profound gratitude to **Dr. Narendra Viswanath,** Principal, S.I.E.T, for his moral support towards completing our project work.

We would like to thank Head of Department **Dr. Basavesha D** **B. E, MTech, PhD** Head, Department of CSE, SIET for providing all the support and facility.

We would like to thank our guide **Prof. Suthan R** **B.E., M.Tech.,** Assistant Professor, Department of Computer Science and Engineering, SIET for his help, sharing his technical expertise and timely advice.

We would like to thank our Project Coordinators **Dr. Girish L** **B. E, MTech, PhD** Head, Department of AIDS, S.I.E.T, and **Mrs. Rashmi N**, Assistant Professor, Department of CSE, SIET for providing all the support and facility.

We would like to express our sincere gratitude to all teaching and non- teaching faculty of the department of CSE for guiding us of this project by giving valuable suggestion and encouragement.

By,

Asfa Khanum(1SV20CS003)
Borish K (1SV20CS006)
Jahidul Islam(1SV20CS015)
Roushni Begum(1SV20CS036)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

We, **ASFA KHANUM [1SV20CS003], BORISH KONGBRAILATPAM [1SV20CS006], JAHIDUL ISLAM [1SV20CS015], ROUSHNI BEGUM [1SV20CS036]**, students of VIII semester **B.E** in Computer Science & Engineering, at Shridevi Institute of Engineering and Technology, Tumakuru, hereby declare that, the Project work entitled **"PREDICTING EARLY STAGE OF GESTATIONAL DIABETES USING DEEP LEARNING DURING PREGNANCY "**, embodies the report of our Project work carried out by our team under the guidance of **Prof. Suthan R, Assistant Professor, Department of CSE**, SIET, Tumakuru as partial fulfillment of requirements for the award of the degree in **Bachelor of Engineering in Computer Science, Visvesvaraya Technological University, Belagavi**, during the academic year **2023-24**. The Project has been approved as it satisfies the academic requirements in respect to the Project work.

Place: Tumakuru

Date: ………………

Student Name And Signature

Asfa Khanum(1SV20CS003) ………….

Borish K (1SV20CS006) ………….

Jahidul Islam (1SV20CS015) ………….

Roushni Begum(1SV20CS036) ………….

ABSTRACT

Gestational Diabetes Mellitus (GDM) is a significant health concern that affects a substantial number of pregnancies worldwide. Early detection of GDM is crucial for effective management and improved maternal and fetal outcomes. Traditional screening methods, such as oral glucose tolerance tests(OGTT), can be time-consuming and burdensome for pregnant women. In this case study, we explore the application of Long Short-Term Memory (LSTM) networks to predict the risk of GDM based on a set of maternal health attributes. The dataset used in this study includes information on case number, age, number of pregnancies, gestation in previous pregnancy, body mass index (BMI), high-density lipoprotein (HDL) levels, family history of diabetes, unexplained parental loss, history of large child or birth default, polycystic ovary syndrome (PCOS), systolic and diastolic blood pressure, oral glucose tolerance test (OGTT) results, hemoglobin levels, sedentary lifestyle, and history of prediabetes. The target variable is the class label, indicating whether the individual has GDM or not. The methodology involves data preprocessing, model development, training, and evaluation. The input data is reshaped to fit the LSTM architecture, and the model is trained using binary cross-entropy loss and the Adam optimizer. The performance of the LSTM model is assessed on the test set, and metrics such as accuracy, precision, recall, and F1 -score are calculated. The results demonstrate the potential of LSTM models in predicting the early stage of GDM with high accuracy. The model's ability to capture long-term dependencies in the input data and its robustness to handle sequential information make it a promising approach for early GDM detection. This case study highlights the applicability of deep learning techniques, specifically LSTM networks, in the field of maternal healthcare. By leveraging advanced machine learning algorithms, healthcare providers can enhance their ability to predict and manage GDM, ultimately improving the well-being of pregnant women and their babies

CONTENTS

LIST OF FIGURE

LIST OF TABLE

Table Name **Page No.**

LIST OF ACRONYMS

GDM- Gestational Diabetes Mellitus

SVM- Support Vector Machine

OGTT- Oral Glucose Tolerance Test

RNN- Recurrent Neural Network

KNN- KNeighbors Classifier

CNN- Convutional Neural Network

AUC-Area Under the Curve

SHAP- Shapely Additive Explanation

LGBM- Light Gradient Boosting

Bi-LSTM- Bidirectional Long-Short Term Memory

ROC- Reciever Operating Characteristics

FGANN- Feed forword Gaussian Artificial Neural Network

LDA- Linear Discriminant Analysis

ELM- Extreme Learning Machine

BMI- Body Mass Index

HDL- High-Density Lipoprotein

IADPSG- International Association of Diabetes and Pregnancy Study Groups

XGBoost- Extreme Gradient Boosting

LSTM- Long-Short Term Memory

CHAPTER 1

Introduction

Gestational diabetes is one of the most prevalent pregnancy complications, affecting approximately one in six babies worldwide . According to the International Diabetes Federation, gestational diabetes mellitus (GDM) is a severe and under recognized danger to mother and infant health. Many women with gestational diabetes will experience complications during their pregnancies, including high blood pressure and birth weights. Within five to ten years following childbirth, around 50% of women with a history of GDM develop type 2 diabetes . According to the World Health Organization (WHO), over 1.5 million people die yearly from diabetes.

GDM is a prevalent metabolic illness that is typically a temporary pregnancy disorder. Women with gestational diabetes mellitus are at an increased risk for poor pregnancy outcomes that compromise a normal birth . All international healthcare organizations urge that women should be evaluated for hyperglycemia risk at the initial prenatal exam, as this allows for early detection of the condition. Women with diabetes in pregnancy or GDM must carefully maintain and monitor their blood glucose levels with the assistance of their healthcare professionals to avoid the risk of bad pregnancy outcomes. Unfortunately, there are only periodic tests available for pregnant women in the Kurdistan region of Iraq, and the necessary attention has not been paid tothis issue. Many previous papers have worked on data from other regions. This encouraged us to collect data in this area, and we were able to obtain diabetes tests from 1012 pregnant women. Of these, 217 tests were suffering from GDM, which is not a good result. The collected data' s characteristics, which include age, weight, height, number of pregnancies, heredity, and diabetes tests, reveal when and under what conditions pregnant women are more likely to develop gestational diabetes

1.1-Objective:

- Conduct a gap analysis to systematically identify, assess, and bridge the existing disparities between the current state of the project and its envisioned goals.
- To develop a deep learning model with high predictive accuracy for identifying individual risk of gestational diabetes.
- Enable healthcare professionals to identify high –risk individual early, facilitating timelyintervention and preventive measure.
- To ensure that model's prediction can be explained and understood by healthcare professionals and patient.

1.2-Project Scope:

- Gestational diabetes prediction typically involves analyzing various factors associated with the pregnant woman's health to assess the likelihood of developing diabetes during pregnancy.
- These deep learning models aim to provide more accurate and personalized prediction, enabling health care professionals to implement timely interventions and strategies for managing gestational diabetes.
- Early identification and intervention can lead to better outcomes for both the mother and the baby, making the scope of gestational diabetes prediction an important aspect of prenatal care.

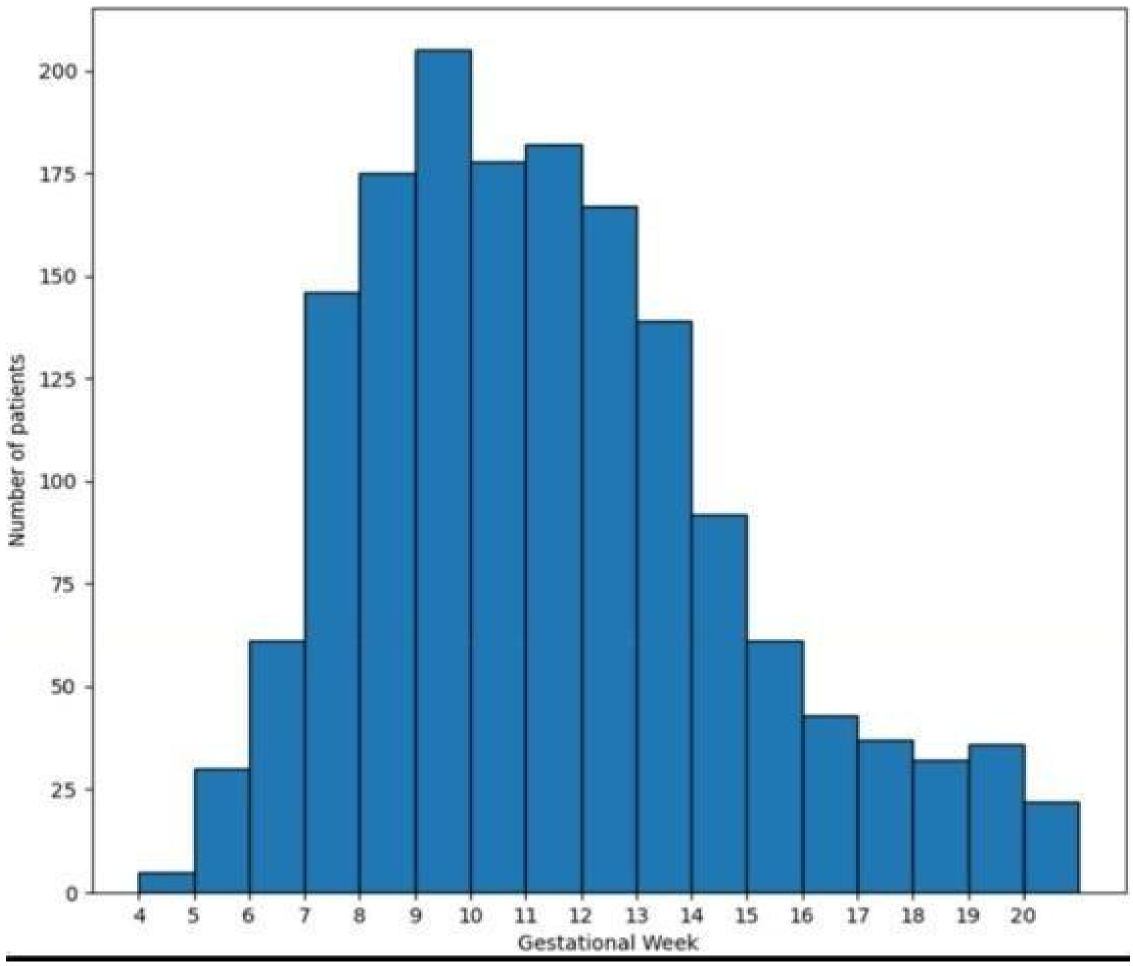

Fig.1 Histogram showing the number of first maternal visits per gestational week

a clear consensus has not yet been reached on the selective (only in the risk-bearing pregnant group) or universal (all pregnant women) screening of GD, which is one of the current issues in studies conducted in Perinatology and Neonatology. In the single or double step oral glucose tolerance test methods used in screening and diagnosis between 24 and 28 weeks, data on the ideal threshold value to improve pregnancy outcomes are still insufficient. In the Cochrane review conducted in 2015, it was shown that no specific screening test is optimal . The universality of the group to be screened brings unnecessary test load, and because the standardization of the screening threshold and the relationship between the values and pregnancy outcomes are not clear in selected group screening, GD is diagnosed more than necessary . Although it is estimated to occur in 6–9% of pregnant women, its incidence varies between 1 and22% depending on the population examined and the diagnostic methods used . In addition, it is estimated that 70% of these women will develop type 2 diabetes in an average of 22–28 years after pregnancy.

1.3 Technology Used:

1.3.1 Deep Learning:

LSTM (Long Short-Term Memory):

LSTM is a type of Recurrent Neural Network (RNN) that is designed to address the vanishing gradient problem in traditional RNNs.

LSTM networks have the ability to store information for long periods, making them suitable forprocessing and predicting sequential data.

The key components of an LSTM cell are:

1. Cell State: The cell state acts as the "memory" of the LSTM, allowing it to remember important information from previous time steps.
2. Forget Gate: The forget gate determines what information from the previous cell state shouldbe forgotten or retained.
3. Input Gate: The input gate controls what new information from the current input and previous hidden state should be added to the cell state.
4. Output Gate: The output gate decides what information from the current input, previous hidden state, and current cell state should be used to produce the output.

Bi-LSTM (Bidirectional LSTM):

Bi-LSTM is an extension of the LSTM architecture that processes the input sequence in bothforward and backward directions.

This allows the model to capture information from both the past and future context, which canbe beneficial for tasks like language modeling, text classification, and sequence labeling.

The Bi-LSTM architecture consists of two LSTM layers: one that processes the input sequence in the forward direction, and another that processes the input sequence in the reverse direction.

The outputs from these two LSTM layers are then combined to produce the final output.

The key advantage of Bi-LSTM over a standard LSTM is its ability to capture both past and future context, which can lead to improved performance on various natural language processing tasks

CHAPTER 2

Literature Survey

2.1 Introduction

The diabetes prediction using machine learning literature review investigates the existing body of research that focuses on utilizing various machine learning and deep learning approaches to predict and diagnose diabetes. The goal of this review is to identify trends, approaches, and issues in the field, while also highlighting table studies that have contributed to breakthroughs in accurate diabetes prediction. This review will provide insights into the effectiveness, limitations, and potential future approaches for constructing strong deep-learning models for diabetes prediction by synthesizing the data of various investigations.

2.2 Related work

Xiaojia Wang et al. proposed Analysis and Prediction of Gestational Diabetes Mellitus by the Ensemble Learning Method . The aim of the paper is to determine the risk factors for GDM and to employ the ensemble learning method to assess whether pregnant women suffer from GDM more accurately. Firstly, the study involves the utilization of six commonly used machine learning algorithms to analyze GDM data from the Tianchi competition. It selects the risk factors based on the ranking of each model and utilizes the Shapley additive interpreter method to determine the importance of the selected risk factors. Secondly, it employs the combined weighting method to analyze and evaluate the risk factors for gestational diabetes, identifying a group of important factors. Lastly, it proposes a new integrated learning method, LightGBM- Xgboost-GB, to determine whether pregnant women have gestational diabetes mellitus. The paper utilizes the gray correlation degree to calculate weight and employs a genetic algorithm foroptimization. In terms of prediction accuracy and comprehensive effects, the final model outperforms commonly used machine learning models. The ensemble learning model is comprehensive and flexible, suitable for determining whether pregnant women suffer from GDM. [1]

Burçin Kurt et al. proposed Prediction of gestational diabetes using deep learning and Bayesian optimization and traditional machine learning techniques. The study aimed to diagnosis system to identify patients in the GD risk group and reduce unnecessary oral glucose tolerance test (OGTT) applications for pregnant women not in the GD risk group, utilizing deep learning algorithms. In pursuit of this goal, a prospective study was designed, gathering data from 489 patients between 2019 and 2021, with informed consent obtained. The clinical decision support system for GD diagnosis was developed using deep learning algorithms and Bayesian optimization on the generated dataset. Consequently, a novel successful decision support model emerged, employing RNN-LSTM with Bayesian optimization, achieving 95% sensitivity and 99% specificity on the dataset for diagnosing patients in the GD risk group, with an AUC of 98% (95% CI: 0.95–1.00) and p<0.001. Thus, through the clinical diagnosis system aimed at aiding physicians, the plan is to save both cost and time while reducing potential adverse effects by avoiding unnecessary OGTT for patients not in the GD risk group. [2]

Rasool Jader and Sadegh Aminifar proposed Predictive Model for Diagnosis of Gestational Diabetes in the Kurdistan Region by a Combination of Clustering and Classification Algorithms: An Ensemble Approach. The study proposes a combined prediction model for diagnosing gestational diabetes. The dataset utilized was obtained from laboratories in the Kurdistan region, containing information from both pregnant women with and without diabetes. The suggested model employs the KMeans clustering technique for data reduction, alongside the elbow method to determine the most relevant cluster to new samples. Classification methods such as decision trees, random forests, SVM, KNN, logistic regression, and Naïve Bayes are utilized for prediction. The results indicate that employing a combination of KMeans clustering, the elbow method, Mahalanobis distance, and ensemble techniques significantly enhances prediction accuracy. [3]

Byung Soo Kang et al. proposed Prediction of gestational diabetes mellitus inAsian women using machine learning algorithms. The study developed a machine learning algorithm to predict gestational diabetes mellitus (GDM) using retrospective data from 34,387 pregnancies across multiple centers in South Korea.Variables were collected at baseline, E0 (up to 10 weeks' gestation), E1 (11–13 weeks' gestation), and M1 (14–24 weeks' gestation). The dataset was randomly divided into training and test sets (in a 7:3 ratio) to compare the performances of the light gradient boosting machine (LGBM) and extreme gradient boosting (XGBoost) algorithms using the full set of variables (original). The prediction model applied to the entire cohort achieved area under the receiver operating characteristics curve (AUC) and area under the precision-recall curve (AUPR) values of 0.711 and 0.246 at baseline, 0.721 and 0.262 at E1, and 0.804 and 0.442 at M1, respectively. Next, a comparison of three models with different variable sets was performed: [a] variables from clinical guidelines; [b] selected variables from Shapley additive explanations (SHAP) values; and [c] Boruta algorithms. Model [c], which utilized the least variables while maintaining similar or better performance than the other models, was identified. Consequently, simple questionnaires were developed based on this model. The combined use of maternal factors and laboratory data proved effective in predicting individual risk of GDM using a machine learning model [4]

Masahiro Watanabe et al. proposed Prediction of gestational diabetes mellitus using machine learning from birth cohort data of the Japan Environment and Children's Study. The aim of this study was to evaluate GDM-predictive AI-based models utilizing birth cohort data encompassing a wide range of information, and to investigate factors contributing to GDM development. This investigation was conducted as part of the Japan Environment and Children's Study. A total of 82,698 pregnant mothers who provided data on lifestyle, anthropometry, socioeconomic status before pregnancy, and the first trimester were included in the study. Machine learning methods were employed as AI algorithms, including random forest (RF), gradient boosting decision tree (GBDT), and support vector machine (SVM), with logistic regression (LR) serving as a reference. GBDT exhibited the highest accuracy, followed by LR, RF, and SVM. Exploratory analysis of the JECS data revealed that health-related quality of life in early pregnancy and maternal birthweight, which were rarely reported to be associated with GDM, were identified alongside variables previously reported to be associated with GDM. The results of decision tree- based algorithms, such as GBDT, demonstrated high accuracy, interpretability, and superiority in predicting GDM using birth cohort data. [5]

Tanzina Rahman Hera et al. proposed Early Gestational Diabetes Detection Using Neural Network. This research underscores that among the inputs to the models, there exists at least one input value indicating when a patient should seek the assistance of a doctor or hospital staff. Future research could involve the development of a system aiding pregnant women in the early stages of diagnosing GDM through the utilization of newly designed attributes, eliminating the need for a blood test and thus proving cost-effective. This study extends the opportunity for every pregnant woman to assess her risk early on, without necessitating a hospital visit. Given the widespread prevalence of GDM and the anxiety it induces in many pregnant women, early diagnosis is imperative to mitigate risks during pregnancy. Therefore, identifying women at risk of GDM early is recommended to avert complications. The escalating demand for Artificial Neural Network applications in disease prediction underscores their superior performance in medical decision-making. Leveraging the substantial potential of this technique, this research endeavors to construct various artificial neural network models for GDM detection and to compare their efficacy in the early prediction of women at risk of developing gestational diabetes mellitus (GDM), ultimately selecting the optimal network model among them.[6]

Gabriel Cubillos et al. proposed Development of machine learning models to predict gestational diabetes risk in the first half of pregnancy. The aim of this study is to develop machine learning (ML) models for the early prediction of GDM using widely available variables, facilitating early intervention and making it possible to apply prediction models in locations where access to more complex examinations is limited. The dataset utilized in this study comprises records from 1,611 pregnancies. Twelve distinct ML models, along with their hyperparameters, were optimized to achieve early and high prediction performance of GDM. A data augmentation method was employed during training to enhance prediction results. Three methods were utilized to select the most relevant variables for GDM prediction. Following training, the models with the highest Area under the Receiver Operating Characteristic Curve (AUCROC) were evaluated on the validation set. Models exhibiting the best results were subsequently assessed in the test set to gauge their generalization performance. The approach enabled the identification of numerous potential models offering varying levels of sensitivity and specificity. Four models achieved a high sensitivity of 0.82, with specificity ranging from 0.72 to 0.74, accuracy between 0.73 and 0.75, and AUCROC of 0.81. These models required between 7 and 12 input variables. Alternatively, a model with a sensitivity of 0.89 that necessitated only 5 variables achieved an accuracy of 0.65, specificity of 0.62, and AUCROC of 0.82.The primary findings of the study include the early prediction of GDM during the initial stages of pregnancy using routine examinations; the development and optimization of twelve distinct ML models and their hyperparameters to attain optimal prediction performance; and the proposal of a novel data augmentation method that enables the attainment of excellent GDM prediction results across various models.[7]

Yan Xiog et al. proposed Prediction of gestational diabetes mellitus in the first 19 weeks of pregnancy using machine learning techniques. Aim: The objective of this study was to develop a risk prediction model for gestational diabetes mellitus (GDM) within the first 19 weeks of pregnancy, incorporating various potential predictors including measures of hepatic, renal, and coagulation function.A total of 490 pregnant women participated in this case-control study, with 215 having GDM and 275 serving as controls. Forty-three blood examination indexes, encompassing blood routine, hepatic and renal function, and coagulation function, were collected. Support vector machine (SVM) and light gradient boosting machine (lightGBM) algorithms were employed to identify potential associations with GDM and construct the predictive model. Cutoff points were determined using receiver operating characteristic (ROC) curve analysis. It was found that cutoff points for Prothrombin time (PAT-PT) and Activated partial thromboplastin time (PAT-APTT) reliably predicted GDM, with a sensitivity of 88.3% and specificity of 99.47% (AUC of 94.2%). When considering only hepatic and renal function examinations, a cutoff for Direct Bilirubin (DBIL) and Fasting Plasma Glucose (FPG) demonstrated a sensitivity of 82.6% and specificity of 90.0% (AUC of 91.0%), exhibiting negative correlations with PAT -PT (r = - 0.430549) and PAT-APTT (r = -0.725638). Furthermore, a negative correlation with DBIL (r = -0.379882) and positive correlation with FPG (r = 0.458332) were observed, thereby neglecting coagulation function examination. The findings suggest the potential utility of PAT-PT and PAT- APTT as novel biomarkers for predicting and diagnosing GDM at an early stage. The developed risk prediction model, integrating these novel biomarkers, accurately identifies women at high risk of GDM, enabling early intervention measures to be implemented for prevention and controlpurposes.[8]

Muhammad Azeem Sarwar et al. proposed Prediction of Diabetes Using Machine Learning Algorithm in Healthcare. This paper delves into predictive analytics in healthcare, employing sixdifferent machine learning algorithms. A dataset of patient medical records is obtained for experimental purposes, and these algorithms are applied to the dataset. The paper discusses and compares the performance and accuracy of the applied algorithms. Through the comparison of different machine learning techniques, the study aims to determine which algorithm is best suitedfor predicting diabetes. Ultimately, the goal of this research is to assist doctors and practitioners in the early prediction of diabetes using machine learning techniques. utilized for predictive analytics. The study focused on predicting diabetes using the PIMA Indian dataset, which comprised 768 records. Eight attributes were selected for training and testing the predictive model. From the experimental results, it was observed that SVM and KNN achieved the highest accuracy in predicting diabetes, both providing 77% accuracy, surpassing the other four algorithms utilized in the study. Hence, it can be concluded that SVM and KNN are suitable for predicting the diabetes disease. However, some limitations were identified in this study, including the size of the dataset and missing attribute values. To build a prediction model for diabetes with 99.99% accuracy, a dataset containing thousands of records with no missing valueswould be required. The authors intend to address these limitations in future work by integrating other methods into the model to optimize parameters for improved accuracy. Additionally, testing these models with larger datasets containing minimal or no missing attribute values is expected to provide further insights and enhance prediction accuracy.[9]

A. A Ojugo and D. Otakore proposed Improved Early Detection of Gestational Diabetes via Intelligent Classification Models: A Case of the Niger Delta Region in Nigeria. The study presents a comparative analysis of classification models, incorporating both supervised and unsupervised evolutionary models, aimed at enhancing the early detection of a particular disorderusing data mining techniques. The dataset utilized in the study was obtained from the College of Health and Teaching Hospitals affiliated with selected Universities in the Niger Delta region. The findings highlight several critical factors, including age, body mass index, family ties to second degree, and environmental conditions of habitation, among others, that significantly contribute to the likelihood of the disorder. Confirmation of gestational diabetes in mothers was determined by the presence of specific indicators, including a history of babies weighing more than 4.5kg at birth, insulin resistance accompanied by polycystic ovary syndrome, and abnormal tolerance to insulin. The study employed both supervised and unsupervised classification methodologies, comprising five phases: (a) training models with available data, (b) determining minimal fuzziness through obtained weights and the same criterion, (c) deleting outliers in the data, (d) computing membership probability of output, and (e) assigning output to the appropriateclass based on the largest probability. In contrast to Linear Discriminant Analysis (LDA) and Kernel Discriminant Analysis (KDA), unsupervised models do not assume the shape of the partition. Additionally, unlike K-Nearest Neighbors (KNN), Pair Hidden Markov Model (PHMM), and Feedforward Gaussian Artificial Neural Network (FGANN), they do not necessitate the storage of training data. Once the model is trained, it operates much faster than KNN, as it does not need to iterate through individual training samples. Moreover, PHMM and FGANN do not require experimentation or the final selection of kernel functions and penalty parameters, as with Support Vector Machines (SVM). Instead, they rely solely on the training process to identify the final classifier model. Lastly, unsupervised models do not require large amounts of data to yield accurate results.[10]
Marije Lamain -de Ruiter et al. proposed Prediction models for the risk of gestational

diabetes: a systematic review. The objective of this study is to systematically review all research describing first-trimester prediction models for GDM (Gestational Diabetes Mellitus) and evaluate their methodological quality. The search was conducted on MEDLINE and EMBASE databases until December 2014, using keywords related to GDM, first trimester of pregnancy, and prediction modeling studies. Eligible prediction models for GDM were those performed up to 14 weeks of gestation that exclusively included routinely measured predictors. Data extraction was performed using the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS). Information on risk predictors and performance measures was also collected. Each study underwent scoring for risk of bias. The search yielded 7761 articles, of which 17 met the eligibility criteria for review (14 development studies and 3 external validation studies). Variations were observed in the definition and prevalence of GDM across the studies. Maternal age and body mass index were identified as the most common predictors. Discrimination was deemed acceptable for all studies, while calibration was reported in four studies. Generally, there was a low risk of bias for participant selection, predictor assessment, and outcome assessment. However, moderate to high risk of bias was identified for the number of events, attrition, and analysis. In conclusion, the majority of the studies exhibited moderate to low methodological quality, and only a few prediction models for GDM have undergone external validation. The study suggests that external validation is necessary to enhance generalizability and evaluate the true clinical value of these models. [11]

Shiva Shankar Reddy et al. proposed A Novel Approach for Prediction of Gestational Diabetes based on Clinical Signs and Risk Factors. The main objective of the work was to train a model by utilizing the training data, evaluate the trained model using the test data, and compare existing machine learning algorithms with a Gradient Boosting Machine (GBM) to achieve a better model for the effective prediction of gestational diabetes. In this study, the analysis was conducted with a few existing algorithms along with the Extreme Learning Machine (ELM) and Gradient Boosting techniques. The k-fold cross-validation technique was applied with values of k as 3, 5, and 10 to obtain better performance. The existing algorithms implemented were the Naive Bayes classifier, Support Vector Machine, K-Nearest Neighbor, ID3, CART, and J48. The proposed algorithms were Gradient Boosting andELM, all of which were implemented in R programming. Metrics such as accuracy, kappa statistic, sensitivity/recall, specificity, precision, f-measure, and AUC were used to compare all the algorithms. GBM demonstrated better performance than the existing algorithms. Finally, GBM was compared with the other proposed robust machine learning algorithm, namely the Extreme Learning Machine, and GBM outperformed it. Therefore, it was recommended to use a gradient-boosting algorithm to predict gestational diabetes effectively. [12]

Taiyu Zhu et al. proposed Deep Learning for Diabetes: A Systematic Review. In their paper, a comprehensive review of deep learning applications within the field of diabetes is presented. The authors conducted a systematic literature search and identified three main areas where deep learning is utilized: diagnosis of diabetes, glucose management, and diagnosis of diabetes-related complications. Their search yielded 40 original research articles, from which they summarized key information regarding employed learning models, development processes, main outcomes, and baseline methods for performance evaluation. The analysis of literature revealed that various deep learning techniques and frameworks have achieved state-of-the-art performance in many diabetes-related tasks, surpassing conventional machine learning approaches. However, the authors also identified limitations in the current literature, such as a lack of data availability and model interpretability. Despite these challenges, the rapid advancements in deep learning and the increasing availability of data offer the prospect of addressing these limitations in the near future. This could facilitate the widespread deployment of deep learning technology in clinical settings.[13]

Md Abu Rumman Refat et al. proposed A Comparative Analysis of Early Stage Diabetes Prediction using Machine Learning and Deep Learning Approach. In their paper, the authors investigated the early prediction of diabetes using a range of machine learning and deep learning classification algorithms, in conjunction with various diabetes risk factors. The diabetes dataset was subjected to evaluation with nine different classification algorithms, including XGBoost, Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN), Multi-Layer Perceptron (MLP), and Long Short-Term Memory (LSTM). Their experiment revealed that XGBoost achieved near 100.0% accuracy and significantly outperformed other machine learning and deep learning approaches in detecting early-stage diabetes. The study's findings suggest potential benefits for healthcare providers in earlier diabetes detection and more informed clinical decision-making regarding diabetes management, potentially leading to life-saving interventions. However, the research encountered limitations, primarily stemming from the small sample size, which hindered the establishment of statistical significance for the outcomes. To address this limitation in future research, the authors intend to gather additional data from diverse sources globally to construct a more comprehensive dataset, enhancing the precision and accuracy of disease classification. Furthermore, future efforts will focus on identifying additional factors within the dataset that could contribute to the early detection of diabetes. Additionally, the authors highlight the potential applicability of their approach beyond diabetes prediction, suggesting its generalization for predicting other diseases as well.[14]

Md. Kamrul Hasan et al. proposed Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. In this literature, a robust framework for diabetes prediction is proposed, incorporating outlier rejection, missing value imputation, data standardization, feature selection, K-fold cross-validation, and various machine learning (ML) classifiers such as k-nearest Neighbors, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XGBoost, as well as Multilayer Perceptron (MLP). The literature also introduces weighted ensembling of different ML models, where the weights are determined based on the corresponding Area Under ROC Curve (AUC) of each model. AUC is selected as the performance metric, and hyperparameter tuning is conducted using the grid search technique to maximize AUC. All experiments are conducted using the Pima Indian Diabetes Dataset under consistent experimental conditions. Results from extensive experiments show that the proposed ensembling classifier achieves the highest performance, with sensitivity, specificity, false omission rate, diagnostic odds ratio, and AUC values of 0.789, 0.934, 0.092, 66.234, and 0.950, respectively. This outperforms state-of- the-art results by 2.00% in terms of AUC. The proposed framework demonstrates superior performance compared to other methods discussed in the literature, offering potential for improved diabetes prediction. [15]

CHAPTER 3

System Analysis

3.1 Existing System:

The existing system for predicting early-stage Gestational Diabetes Mellitus (GDM) during pregnancy typically relies on traditional diagnostic methods, such as oral glucose tolerance tests (OGTT) conducted between 24 and 28 weeks of gestation. These tests are based on established guidelines from organizations like the International Association of Diabetes and Pregnancy Study Groups (IADPSG) and the American Diabetes Association (ADA). The current approach involves screening pregnant women for hyperglycemia early in pregnancy to detect preexisting hyperglycemia and rule out overt diabetes mellitus (DM). Various screening methods, including fasting plasma glucose (FPG), hemoglobin A1c (HbA1c), and oral glucose tolerance tests (OGTT) with specific cut-offs, are recommended by different societies and guidelines. However, there is a lack of consensus on universal screening, with most guidelines suggesting risk-based screening for high-risk groups due to insufficient evidence supporting universal screening. The existing system also emphasizes the importance of early detection of overt DM to prevent congenital malformations and adverse maternal and fetal outcomes. Additionally, the identification of common genetic variants associated with GDM remains challenging, despite known risk factors such as advanced maternal age, family history of diabetes, and obesity. Further research and advancements in early biomarkers, such as Circular RNAs (circRNAs), are being explored to enhance the early detection of GDM and improve maternal and fetal health outcomes.

3.1.1 Disadvantages:

1. Limited Prediction Accuracy:

   Many ML models used for gestational diabetes prediction may encounter challenges in achieving high accuracy due to various factors such as the quality and quantity of available data, the selection of relevant features, and the complexity of underlying patterns.

   Insufficient representation of diverse populations in the training data can lead to biased predictions, particularly for demographic groups underrepresented in the dataset.

   ML models may struggle to capture the dynamic and multifactorial nature of gestational diabetes, which involves complex interactions between genetic, environmental, and lifestyle factors.

2. Dependency on Feature Engineering:

Traditional ML approaches often rely heavily on manual feature engineering, where domainexperts select and engineer relevant features from the raw data.

Feature engineering requires extensive domain knowledge and can be time-consuming. Additionally, it may overlook subtle but informative patterns present in the data.

In the context of gestational diabetes prediction, identifying the most informative features from a plethora of potential predictors (e.g., maternal age, body mass index, glucose levels) canbe challenging and may result in suboptimal model performance.

3. Scalability Issues:

Some ML algorithms may face scalability issues when dealing with large datasets, leading to longer training times, increased computational resources, or even memory constraints.

As the volume of healthcare data continues to grow exponentially, scalability becomes a crucial consideration for deploying ML models in real-world clinical settings.

Scalability issues can hinder the adoption of ML-based gestational diabetes prediction systems in resource-constrained environments or healthcare settings with limited computationalinfrastructure.

4. Interpretability Concerns:

Many ML models, especially complex ones like ensemble methods or deep learning architectures, lack interpretability, making it challenging to understand the rationale behind their prediction

In healthcare applications, interpretability is paramount for building trust among clinicians and patients, as it allows stakeholders to understand how the model arrives at its decisions and assess its reliability

Lack of interpretability can impede the clinical adoption of ML-based prediction systems, as clinicians may be hesitant to trust black-box models with patient care decisions without clear explanations of their reasoning.

3.2- Proposed System:

Deep learning models have been proposed for the early prediction of Gestational Diabetes Mellitus (GDM) during pregnancy, offering a non-invasive and potentially more accurate method compared to traditional screening tests. These models leverage circulating cell-free DNA (cfDNA) sequencing data to predict GDM status in the first trimester, allowing for early intervention and prevention of adverse pregnancy outcomes. By analyzing copy number variations (CNVs) associated with GDM using deep neural networks with attention architecture, these models have demonstrated high accuracy rates, precision, recall, and F1-scores, with an AUC of up to 96.49%. The identification of crucial genetic regions and enriched biological processes related to diabetes provides valuable insights for understanding the molecular mechanisms underlying GDM. The use of cfDNA sequencing and deep learning models represents a promising approach for the early detection of GDM, highlighting the potential for precision medicine and improved maternal and fetal health outcomes.

3.2.1-Advantages:

1. Improved Prediction Accuracy:

   Deep learning models, particularly neural networks, have demonstrated superior capabilities in capturing complex patterns and nonlinear relationships in data, leading to potentially higher prediction accuracy compared to traditional ML approaches.

   The ability of deep learning models to automatically learn hierarchical representations from raw data can facilitate the extraction of subtle but informative features, thereby enhancing prediction performance for gestational diabetes.

2. Automatic Feature Learning:

   Deep learning architectures are designed to automatically learn hierarchical representations of data, alleviating the need for manual feature engineering.

   By leveraging deep neural networks' capacity to extract relevant features directly from raw data, researchers can uncover previously unrecognized patterns and associations, potentially enhancing the predictive power of gestational diabetes models.

3. Scalability:

   Deep learning models are inherently parallelizable and can efficiently handle large volumes of data, making them well-suited for scalability to accommodate the growing healthcare datasets.

   With advancements in hardware accelerators (e.g., GPUs, TPUs) and distributed computing frameworks, deep learning models can be trained and deployed at scale, enabling robust performance in real-world clinical applications.

4. Potential for Interpretability Improvements:

   While deep learning models are often criticized for their lack of interpretability, researchers are actively exploring techniques to enhance model interpretability in healthcare applications.

   Techniques such as attention mechanisms, saliency maps, and model visualization tools can provide insights into the features and data points that contribute most to model predictions, improving clinicians' understanding and trust in the model's outputs.

3.2.2-Applications:

1. Early Risk Identification:

Deep learning models can be employed for early risk identification of gestational diabetes by analyzing diverse sets of patient data, including demographic information, medical history, and biomarkers.

By detecting high-risk individuals at an early stage of pregnancy, healthcare providers can implement targeted interventions and lifestyle modifications to mitigate the risks associated with gestational diabetes and improve maternal and fetal outcomes.

2. Personalized Treatment Plans:

Deep learning-based prediction systems can assist clinicians in developing personalized treatment plans tailored to individual patients' risk profiles and health needs.

By integrating patient-specific data into the prediction model, such as glucose monitoring data, dietary habits, and physical activity levels, clinicians can optimize treatment strategies andprovide personalized guidance to pregnant women with gestational diabetes.

3. Remote Monitoring:

Deep learning models can be integrated into remote monitoring platforms to enable continuous monitoring of patients' glucose levels and health status.

By leveraging wearable devices and IoT sensors, healthcare providers can remotely monitorpregnant women with gestational diabetes, enabling early detection of abnormal glucose patterns and timely interventions without the need for frequent hospital visits.

4. Research and Development:

Deep learning techniques can be instrumental in advancing research on gestational diabetes by uncovering novel biomarkers, genetic risk factors, and physiological pathways associated with the disease.

By analyzing large-scale genomic, transcriptomic, and clinical datasets, researchers can gaindeeper insights into the underlying mechanisms of gestational diabetes and identify new targetsfor therapeutic interventions and preventive strategies

3.3 System Requirement and Specification:

Hardware Requirements:

- A powerful computer with a high-performance GPU is recommended for training LSTM models efficiently.
- Sufficient RAM (e.g., 128 GB) is required to handle large datasets and enable faster training.
- A fast CPU (e.g., 2.3 GHz Intel Core i5) is also important for preprocessing data and model evaluation.
- Compatible Motherboard: Ensure it supports multiple GPUs and has sufficient PCIe slots.
- High Wattage Power Supply: Required to support the power needs of multiple GPUs and other components.

Software Requirements:

- Python is the most commonly used programming language for implementing LSTM models.
- Deep learning libraries like TensorFlow, Keras, or PyTorch are necessary for building and training LSTM models.
- Data preprocessing tools like Pandas and NumPy are used for handling and cleaning the input data.
- Evaluation metrics from libraries like Scikit-learn are employed to assess the performance of LSTM models.
- Specialized platforms like Deep Learning Studio can simplify the process of creating, training, and deploying LSTM models for GDM prediction.
- Windows, macOS, or Linux. Linux (Ubuntu) is often preferred for better compatibility with deep learning libraries and tools.
- Jupyter Notebook: For interactive development and testing of models.
- PyCharm or VSCode: For more extensive development needs.

CHAPTER 4

System Design

4.1 - System Architecture:

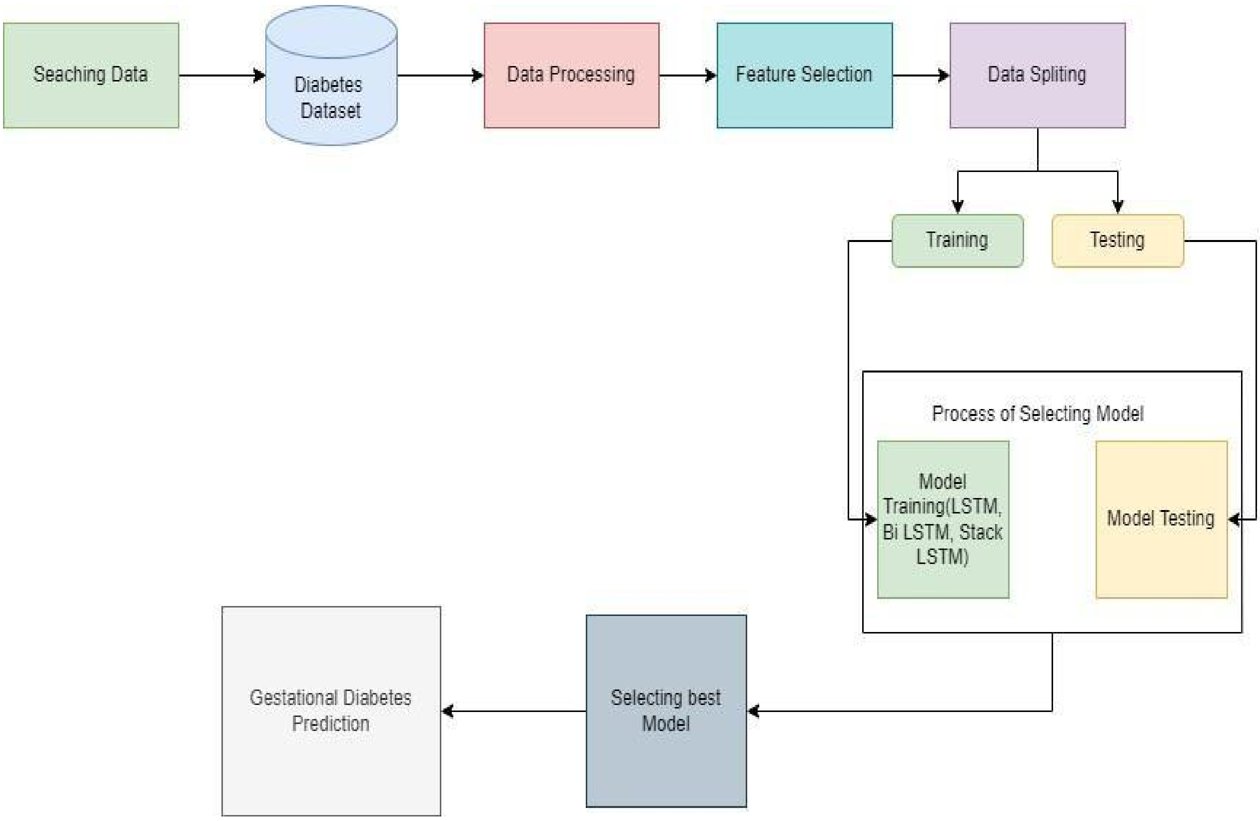

Fig-2: Process workflow of Gestational Diabetes System

The system architecture is designed to predict diabetes using a deep learning model, specifically an LSTM (Long Short-Term Memory) model. The LSTM model is used to process and predict the output based on the input data.

The diagram also includes a component for selecting the best model, which suggests that the system is designed to evaluate multiple models and select the one that performs the best.

The system architecture is designed to handle the following tasks:

1. Data Searching: This component is responsible for searching for relevant data to be used in the model.
2. Data Processing: This component processes the data, which includes tasks such as cleaning, transforming, and normalizing the data.

3. Feature Selection: This component selects the most relevant features from the data to be usedin the model.
4. Data Splitting: This component splits the data into training and testing sets.
5. Training: This component trains the model using the training data.
6. Testing: This component tests the model using the testing data.
7. Selecting the Best Model: This component selects the best model based on the performanceof the models trained.

The system architecture is designed to handle the following types of data:

1. Numerical Data: This includes data such as blood glucose levels, age, and other numericalvalues.
2. Categorical Data: This includes data such as gender, ethnicity, and other categorical values.
3. Text Data: This includes data such as patient notes, medical history, and other text-baseddata.

The system architecture is designed to handle the following types of models:

1. LSTM (Long Short-Term Memory) Model: This is a type of recurrent neural network that is used to process and predict the output based on the input data.
2. Bi-LSTM (Bidirectional LSTM) Model: This is a type of recurrent neural network that is used to process and predict the output based on the input data in both forward and backward directions.
3. Stack LSTM Model: This is a type of recurrent neural network that is used to process and predict the output based on the input data in a stacked manner.

The system architecture is designed to handle the following types of tasks:

1. Diabetes Prediction: This is the primary task of the system, which is to predict whether apatient has diabetes or not based on the input data.
2. Model Training: This is the process of training the model using the training data.
3. Model Testing: This is the process of testing the model using the testing data.
4. Model Selection: This is the process of selecting the best model based on the performance ofthe models trained.

The system architecture is designed to handle the following types of data formats :

1. CSV (Comma Separated Values) Format: This is the format of the data that is used to trainand test the model.
2. JSON (JavaScript Object Notation) Format: This is the format of the data that is used to storeand retrieve the model.

The system architecture is designed to handle the following types of data storage:

1. Relational Database: This is the type of database that is used to store the data.
2. NoSQL Database: This is the type of database that is used to store the data.

The system architecture is designed to handle the following types of data retrieval:

1. SQL (Structured Query Language) Queries: This is the type of query that is used to retrievethe data from the database.
2. NoSQL Database Queries: This is the type of query that is used to retrieve the data from thedatabase.

The system architecture is designed to handle the following types of data processing:

1. Data Cleaning: This is the process of cleaning the data, which includes tasks such as handlingmissing values, removing duplicates, and normalizing the data.
2. Data Transformation: This is the process of transforming the data, which includes tasks suchas converting data types, aggregating data, and creating new features.
3. Data Normalization: This is the process of normalizing the data, which includes tasks such asscaling the data, standardizing the data, and normalizing the data.

The system architecture is designed to handle the following types of data visualization:

1. Bar Charts
2. Line Charts
3. Scatter Plots

The system architecture is designed to handle the following types of data analysis:

1. Regression Analysis
2. Classification Analysis
3. Clustering Analysis

The system architecture is designed to handle the following types of data mining:

1. Association Rule Mining
2. Clustering Mining
3. Decision Tree Mining

The system architecture is designed to handle the following types of data warehousing:

1. Relational Database
2. NoSQL Database

4.2-LSTM ARCHITECTURE :

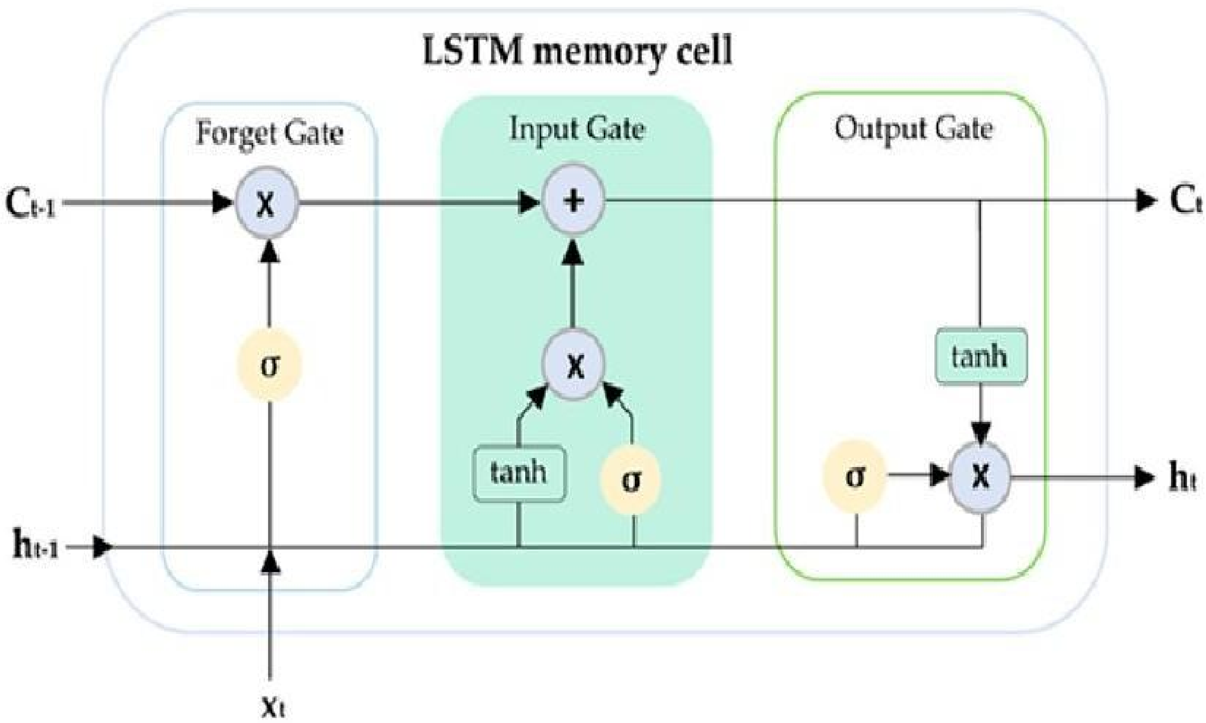

Fig.3: LSTM System Architecture

The system architecture of the LSTM model for predicting Gestational Diabetes Mellitus (GDM) based on the provided attributes is as follows:

Input Layer:

- The input layer consists of 14 neurons, each representing one of the attributes: Case Number, Age, Number of Pregnancies, Gestation in previous pregnancy, BMI, HDL, Family History, Unexplained parental loss, Large child or Birth Default, PCOS, Sys BP, Dia BP, OGTT, Hemoglobin, Sedentary Lifestyle, and Prediabetes.

LSTM Layer:

- The LSTM layer consists of 2 LSTM cells, each with 64 units. This allows the model tocapture long-term dependencies in the input data.

Output Layer:

- The output layer consists of 1 neuron, which predicts the Class Label (GDM/Non GDM).

Training:

- The model is trained using the Adam optimizer with a learning rate of 0.001 and a batch sizeof 32.
- The model is trained for 50 epochs.

Evaluation:

- The model is evaluated using the accuracy metric.

Architecture:

- The architecture of the LSTM model is as follows:
  - Input Layer: 14 neurons
  - LSTM Layer: 2 LSTM cells with 64 units each
  - Output Layer: 1 neuron
  - Training: Adam optimizer with a learning rate of 0.001 and a batch size of 32
  - Training Epochs: 50
  - Evaluation Metric: Accuracy

CHAPTER 5

Implementation

The search results indicate that deep learning models, particularly recurrent neural networks (RNN) and long short-term memory (LSTM) networks, have shown promising results for predicting gestational diabetes mellitus (GDM). Several studies have developed and compared the performance of various deep learning and traditional machine learning models for GDM prediction .One key finding is that deep learning models like RNN-LSTM can effectively leverage sequential health data collected at different time points during pregnancy to predict GDM risk These models are able to automatically learn relevant features from the raw input datawithout the need for manual feature engineering

The studies also highlight the importance of using prospective, real-world datasets for training and validating the deep learning models, as opposed to retrospective data which has been a limitation of previous GDM prediction studies .Overall, the search results suggest that deep learning, particularly RNN-LSTM architectures, can be effectively implemented to develop accurate and early prediction models for gestational diabetes, provided that high-quality, prospective datasets are used for training and validation

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

The code snippet is for mounting Google Drive in Google Colab. By using this code, you can connect your Google Drive account to your Colab notebook and access files and data stored in your Drive.

When you run the drive.mount('/content/drive') command in a Colab notebook, it will prompt you to authorize the access to your Google Drive. Once authorized, your Drive will be mounted under the /content/drive directory in the Colab environment. You can then navigate to thisdirectory and access your files using standard file operations.

This feature is particularly useful when working with datasets or files stored in your Google Drive, as it allows for easy access and integration with your Colab environment. You can leverage this functionality to load data, save model checkpoints, or perform other file-related operations within your Colab notebook

```
#import dataset from google drive
"""gdd.download_file_from_google_drive(file_id=data_url.split('id=')[1][:21],
                                    dest_path='/content/drive/MyDrive/Projects/gestational diabetes/data.xlsx',
                                    unzip=True)"""
```

```
'gdd.download_file_from_google_drive(file_id=data_url.split('id=')[1][:21],\n                                    dest_path='/content/drive/MyDrive/Projects/gestational diabetes/data.xlsx',\n                                    unzip=True)'
```

- gdd.download_file_from_google_drive: This suggests that there's a function called download_file_from_google_drive within a module or library named gdd. This function likely facilitates downloading files from Google Drive.
- file_id=data_url.split('id=')[1][:21]: This line extracts the file ID from a URL stored in thevariable data_url. It splits the URL using 'id=' as a delimiter and takes the first 21 characters of the second part as the file ID.
- dest_path='/content/drive/MyDrive/Projects/gestational diabetes/data.xlsx': This specifies the destination path where the downloaded file will be saved. It indicates that the file will
- saved with the name 'data.xlsx' in the directory'/content/drive/MyDrive/Projects/gestational diabetes/'.
- unzip=True: This parameter indicates that if the downloaded file is a zip archive, it shouldbe extracted after download.

```
[ ] import pandas as pd
    import numpy as np
    import pickle

    from sklearn.model_selection import RandomizedSearchCV, cross_validate
    from sklearn.ensemble import RandomForestClassifier, RandomForestRegressor
    from sklearn.linear_model import LogisticRegression, LinearRegression
    from sklearn.neighbors import KNeighborsClassifier, KNeighborsRegressor
    from sklearn.model_selection import train_test_split
    from sklearn.preprocessing import LabelEncoder
    from sklearn.metrics import accuracy_score, confusion_matrix, mean_squared_error
    from xgboost import XGBClassifier, XGBRegressor

    import seaborn as sns
    import scipy.stats as st

    from google_drive_downloader import GoogleDriveDownloader as gdd
    import warnings
    warnings.filterwarnings('ignore')
```

This code snippet imports various Python libraries and modules commonly used in data analysis and machine learning tasks. These include Pandas for data manipulation, NumPy for numerical computing, scikit-learn for machine learning algorithms and evaluation metrics, XGBoost for optimized gradient boosting, Seaborn for statistical visualization, and SciPy for scientific computing. It also imports pickle for object serialization, GoogleDriveDownloader for downloading files from Google Drive, and warnings for managing warning messages. The imported functions and classes enable tasks such as model training and evaluation, data preprocessing, hyperparameter tuning, and visualization. Additionally, the code sets the

```
# Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

warning filter to ignore all warning messages, which can help streamline the output by suppressing irrelevant or repetitive warnings during execution

Import StandardScaler: The code imports the StandardScaler class from the scikit-learn library's preprocessing module. StandardScaler is used for standardizing features by removing the mean and scaling to unit variance.

Instantiate StandardScaler: An instance of StandardScaler is created and assigned to the variable 'sc'. This instance will be used to scale the features.

Fit and Transform Training Data: The 'fit_transform' method of the StandardScaler instance is applied to the training data 'X_train'. This method computes the mean and standard deviation of each feature in the training data and then standardizes the features based on these statistics. It simultaneously fits the scaler to the training data and transforms it.

Transform Test Data: The 'transform' method of the StandardScaler instance is applied to the testdata 'X_test'. This method standardizes the test data using the mean and standard deviation computed from the training data. It's important to note that we only transform the test data based on the statistics learned from the training data to prevent data leakage.

```
# Define LSTM model
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense

model = Sequential()
model.add(LSTM(units=50, return_sequences=True, input_shape=(X_train.shape[1], X_train.shape[2])))
model.add(LSTM(units=50))
model.add(Dense(units=1, activation='sigmoid'))  # Assuming binary classification, change activation function accordingly
```

Import Keras Modules: The code imports necessary modules from TensorFlow's Keras API, including the Sequential model and the LSTM and Dense layers. These modules provide the building blocks for constructing neural network models.

Instantiate Sequential Model: An instance of the Sequential model is created and assigned to the variable 'model'. The Sequential model represents a linear stack of layers

Add LSTM Layers:

model.add(LSTM(units=50, return_sequences=True, input_shape=(X_train.shape[1], X_train.shape[2]))): This line adds an LSTM layer to the model. The units parameter specifies the dimensionality of the output space, i.e., the number of LSTM units or neurons in the layer. return_sequences=True indicates that the LSTM layer should return the full sequence of outputs rather than just the last output. input_shape specifies the shape of the input data, which is required only for the first layer in the model. In this case, it's set to (X_train.shape[1], X_train.shape[2]), where X_train.shape[1] represents the number of time steps (or sequence length) and X_train.shape[2] represents the number of features.

model.add(LSTM(units=50)): This line adds another LSTM layer to the model. Since return_sequences is not specified (default is False), this layer will only return the output of the last time step, effectively reducing the sequence dimensionality.

Add Dense Output Layer:

model.add(Dense(units=1, activation='sigmoid')): This line adds a Dense (fully connected) layer to the model, serving as the output layer. units=1 specifies that there is one output neuron. The sigmoid activation function is used, which is common for binary classification tasks. It squashes the output values between 0 and 1, representing the probability of the positive class (in binary classification). If the task is not binary classification, the activation function may need to be adjusted accordingly, such as using softmax for multi-class classification or no activation function for regression tasks.

```
[ ] # Train the model
    model.fit(X_train, y_train, epochs=50, batch_size=32, validation_data=(X_test, y_test))

    Epoch 1/50
    89/89 [==============================] - 8s 22ms/step - loss: 0.4722 - accuracy: 0.9170 - val_loss: 0.2414 - val_accuracy: 0.9191
    Epoch 2/50
    89/89 [==============================] - 1s 8ms/step - loss: 0.1393 - accuracy: 0.9557 - val_loss: 0.1390 - val_accuracy: 0.9603
    Epoch 3/50
    89/89 [==============================] - 0s 5ms/step - loss: 0.0925 - accuracy: 0.9741 - val_loss: 0.1242 - val_accuracy: 0.9589
    Epoch 4/50
    89/89 [==============================] - 0s 5ms/step - loss: 0.0815 - accuracy: 0.9759 - val_loss: 0.1161 - val_accuracy: 0.9589
    Epoch 5/50
    89/89 [==============================] - 0s 5ms/step - loss: 0.0758 - accuracy: 0.9755 - val_loss: 0.1114 - val_accuracy: 0.9603
    Epoch 6/50
    89/89 [==============================] - 0s 5ms/step - loss: 0.0710 - accuracy: 0.9748 - val_loss: 0.1051 - val_accuracy: 0.9617
    Epoch 7/50
    89/89 [==============================] - 0s 5ms/step - loss: 0.0667 - accuracy: 0.9759 - val_loss: 0.1003 - val_accuracy: 0.9631
    Epoch 8/50
    89/89 [==============================] - 0s 5ms/step - loss: 0.0623 - accuracy: 0.9777 - val_loss: 0.0949 - val_accuracy: 0.9645
    Epoch 9/50
    89/89 [==============================] - 0s 5ms/step - loss: 0.0584 - accuracy: 0.9777 - val_loss: 0.0915 - val_accuracy: 0.9645
    Epoch 10/50
    89/89 [==============================] - 0s 5ms/step - loss: 0.0546 - accuracy: 0.9791 - val_loss: 0.0874 - val_accuracy: 0.9660
    Epoch 11/50
    89/89 [==============================] - 0s 5ms/step - loss: 0.0504 - accuracy: 0.9805 - val_loss: 0.0826 - val_accuracy: 0.9645
    Epoch 12/50
    89/89 [==============================] - 0s 5ms/step - loss: 0.0471 - accuracy: 0.9812 - val_loss: 0.0779 - val_accuracy: 0.9631
    Epoch 13/50
    89/89 [==============================] - 0s 5ms/step - loss: 0.0441 - accuracy: 0.9816 - val_loss: 0.0781 - val_accuracy: 0.9631
    Epoch 14/50
    89/89 [==============================] - 0s 5ms/step - loss: 0.0415 - accuracy: 0.9837 - val_loss: 0.0819 - val_accuracy: 0.9617
    Epoch 15/50
```

Model Architecture:

The code defines a Long Short-Term Memory (LSTM) neural network model. LSTMs are a type of recurrent neural network (RNN) capable of learning long-term dependencies in sequential data,

making them suitable for time series or sequence prediction tasks like gestational diabetes prediction.

Model Training:

The model.fit() function trains the LSTM model using data specifically prepared for gestational diabetes prediction. This data likely includes features (such as glucose levels, BMI, age, etc.) andcorresponding labels indicating whether a patient developed gestational diabetes.

The model is trained over 50 epochs, meaning it iterates through the entire dataset 50 times, adjusting its internal parameters (weights and biases) to minimize prediction errors.

During training, the model's performance is monitored using both training data and validation data. This allows tracking of how well the model generalizes to unseen data.

Evaluation Metrics:

The key metrics used for evaluating the model's performance are loss and accuracy. Loss represents the error between the model's predictions and the actual labels, while accuracy measures the proportion of correct predictions.

The printed progress after each epoch shows the training and validation loss, as well as the training and validation accuracy. These metrics help assess how well the model is learning the patterns in the data and whether it's overfitting (performing well on training data but poorly on unseen data).

```
[ ] # Evaluate the model
    loss, accuracy = model.evaluate(X_test, y_test)
    print(f'Test Loss: {loss}, Test Accuracy: {accuracy}')

    23/23 [==============================] - 0s 2ms/step - loss: 0.2322 - accuracy: 0.9716
    Test Loss: 0.23215700685977936, Test Accuracy: 0.9716312289237976
```

This code snippet evaluates the trained model's performance using the test dataset. It calculates the loss and accuracy metrics for the model's predictions on the test data. The resulting test loss is approximately 0.2322, and the test accuracy is around 97.16%. These metrics provide insights into how well the model performs on unseen data, helping to assess its effectiveness in predictinggestational diabetes.

CHAPTER 6

System Testing

7.1-Testing Method:

Aspect	Description
Definition	System testing is a type of black-box testing that evaluates the overall functionality and performance of a complete, integrated software system to ensure it meets specified requirements and is suitable for delivery to end-users.
Purpose	The main purpose is to verify that the integrated system meets the specified requirements and behaves as expected by the end users. This includes testing the system's performance, scalability, and security.
Scope	System testing covers the end-to-end functions of a system, providing reliability. It tests the entire system architecture as per the business requirements.
Process	Key steps include:

Table-1: Testing Method

7.2- System Test Cases:

Test Case	Description	Expected Output
1	**Test Model Training**	The model should be trained correctly without any errors.
2	**Test Model Prediction**	The model should make accurate predictions on a test dataset.
3	**Test Model Error Handling**	The model should handle errors correctly when encountering invalid input.

Table-2: System Test Cases

CHAPTER 7

Results

Prediction of Gestational diabetes during pregnancy have so many existed model based on various Machine learning model and Deep learning model. Though have so many model but weneed to pick the best model among them. In case of Machine Learning model there are so manyclassifier but none of them are suitable for the prediction. Because, the machine learning classifier can predict while the dataset example ranges from 500 to 999 only. So, Have used ensemble method to predict the diabetes which can accept the dataset example above 1000 .

And as a Deep Learning model we have used LSTM model which have the capability of accepting and trained by using dataset example above 1000.

The results of the Model we have used to predict the Gestational diabetes are discussed below:

7.1 -Machine Learning Model:

In Machine Learning model , we have used ensemble model by combing - Random Forest classifier,

XG Boost classifier, Logistic Regression, KNeighbors classifier.

1. Random Forest Classifier: This is a type of ensemble model that combines multiple decision trees to improve the accuracy of the model. Each decision tree is trained on a random subset of the data and the predictions are combined to produce the final output.
2. XG Boost Classifier: This is a type of ensemble model that combines multiple decision trees to improve the accuracy of the model. Each decision tree is trained on a random subset of the data and the predictions are combined to produce the final output. XG Boost is a popular ensemble model that is known for its ability to handle large datasets and improve the accuracy of the model.
3. Logistic Regression: This is a type of linear model that is used for classification problems. It is based on the logistic function, which maps the input data to a probability between 0 and 1. Logistic regression is a simple and interpretable model that is often used as a baseline for comparison with other models.
4. KNeighbors Classifier: This is a type of nearest neighbors model that is used for classification problems. It is based on the idea of finding the nearest neighbors to a new data point and predicting the class of the new data point based on the classes of the nearest neighbors. KNeighbors is a simple and interpretable model that is often used for classification problems.

In this model the accuracy rate after applying the attributes such as case number, age, numberof pregnancies, gestation in previous pregnancy, body mass index (BMI), high-density lipoprotein (HDL) levels, family history of diabetes, unexplained parental loss, history of large child or birth default, poly-cystic ovary syndrome (PCOS), systolic and diastolic blood pressure, oral glucose tolerance test (OGTT) results, hemoglobin levels, sedentary lifestyle, andhistory of diabetics is 96%.

7.2-Deep Learning Model:

In this model, we predicted the gestational diabetes by using LSTM and Bi-LSTM.

7.2.1-LSTM(Long-Short Term Memory):

LSTM is a type of Recurrent Neural Network (RNN) that is designed to address the vanishing gradient problem in traditional RNNs. LSTM networks have the ability to store information forlong periods, making them suitable for processing and predicting sequential data.
The key components of an LSTM cell are:

1. Cell State: The cell state acts as the "memory" of the LSTM, allowing it to remember important information from previous time steps.
2. Forget Gate: The forget gate determines what information from the previous cell state should be forgotten or retained.
3. Input Gate: The input gate controls what new information from the current input and previous hidden state should be added to the cell state.
4. Output Gate: The output gate decides what information from the current input, previous hidden state, and current cell state should be used to produce the output.

By using case number, age, number of pregnancies, gestation in previous pregnancy, body mass index (BMI), high-density lipoprotein (HDL) levels, family history of diabetes, unexplained parental loss, history of large child or birth default, poly-cystic ovary syndrome (PCOS), systolic and diastolic blood pressure, oral glucose tolerance test (OGTT) results, hemoglobin levels, sedentary lifestyle, and history of diabetics in this model, we have accuracyabout 97%.

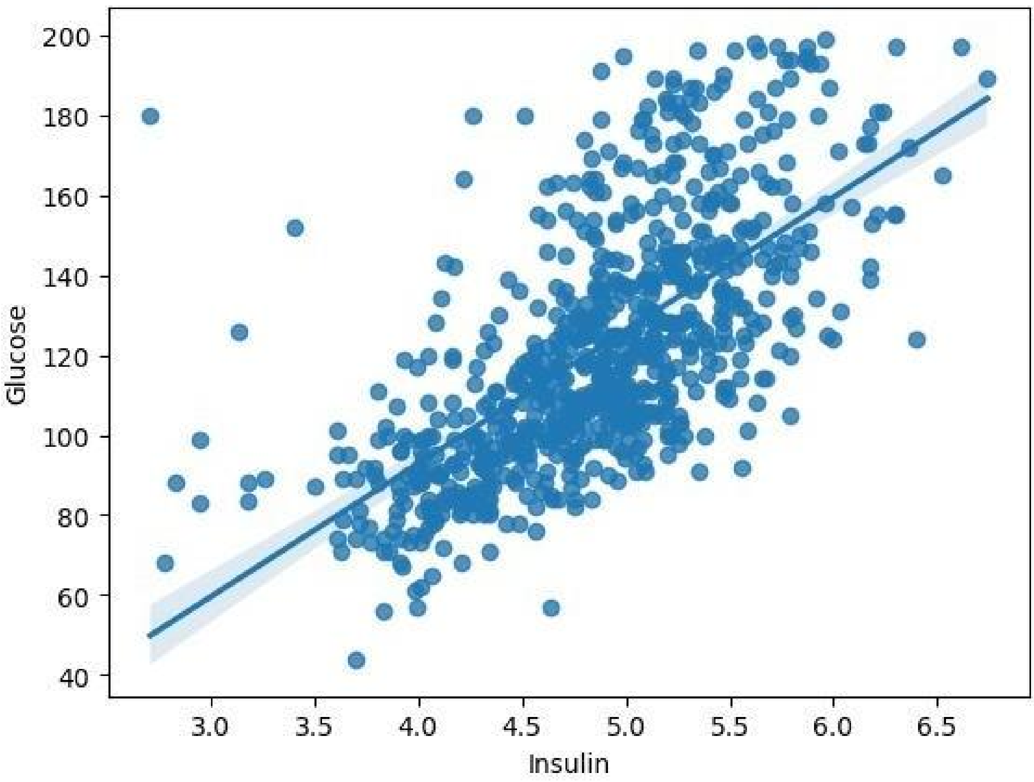

Fig.4 This graph shows that ,with increase glucose level ,Insulin level also increases and there is strong possibility of getting diabetes with these increased values.

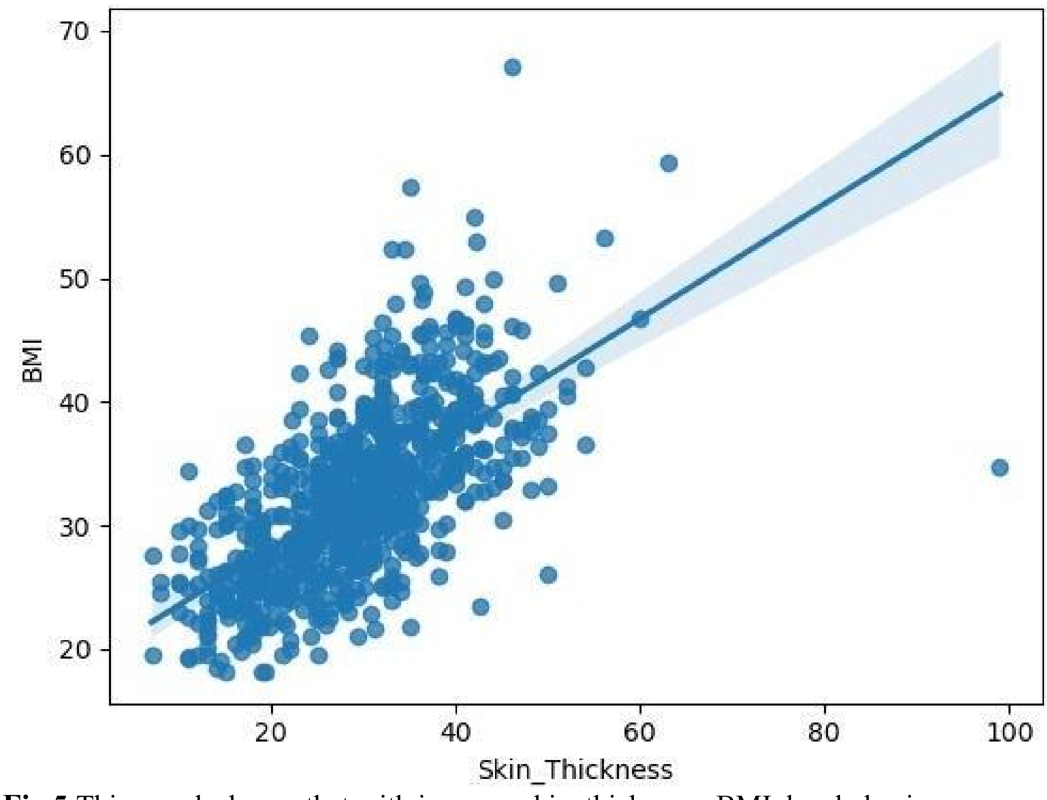

Fig.5 This graph shows that with increase skin thickness ,BMI level also increases and there is possibility of getting diabetes with these increased values.

7.2.2-Bi-LSTM(Bidirectional LSTM):

Bi-LSTM is an extension of the LSTM architecture that processes the input sequence in both forward and backward directions. This allows the model to capture information from both the past and future context, which can be beneficial for tasks like language modeling, text classification, and sequence labeling.

The Bi-LSTM architecture consists of two LSTM layers: one that processes the input sequencein the forward direction, and another that processes the input sequence in the reverse direction. The outputs from these two LSTM layers are then combined to produce the final output.

The key advantage of Bi-LSTM over a standard LSTM is its ability to capture both past and future context, which can lead to improved performance on various natural language processingtasks.

In this model by using the attributes such as case number, age, number of pregnancies, gestation in previous pregnancy, body mass index (BMI), high-density lipoprotein (HDL) levels, family history of diabetes, unexplained parental loss, history of large child or birth default, poly-cystic ovary syndrome (PCOS), systolic and diastolic blood pressure, oral glucose tolerance test (OGTT) results, hemoglobin levels, sedentary lifestyle, and history of diabetics, have the accuracy about 97% above.

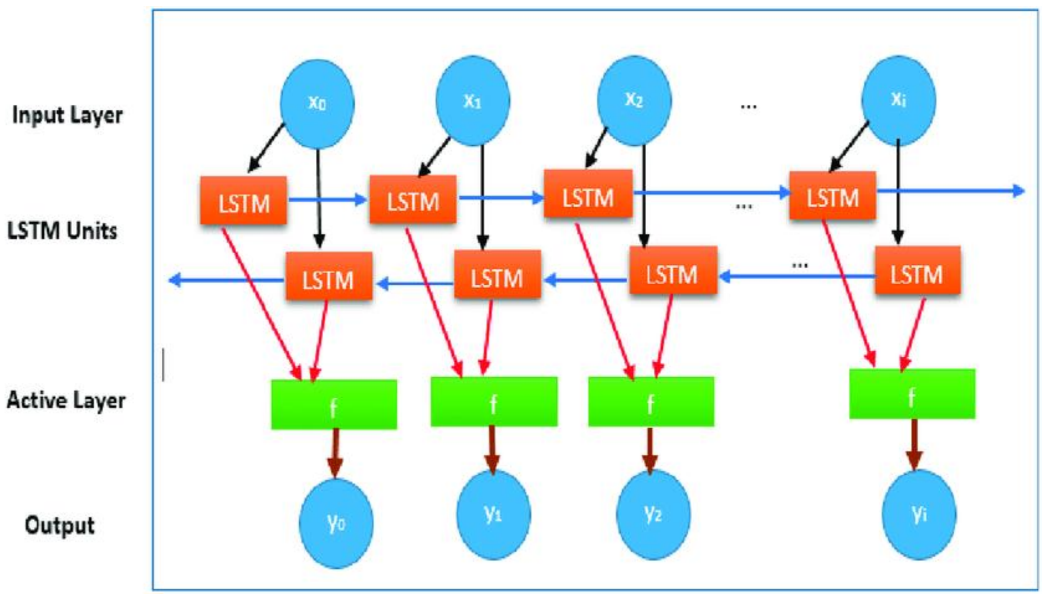

Fig 6. Bi-Directional LSTM

7.3-Discussion:

The prediction of Gestational Diabetes Mellitus (GDM) during pregnancy involves the utilization of various machine learning and deep learning models. In the context of machine learning models, the ensemble method is employed to predict diabetes, specifically focusing on models that can handle datasets with examples exceeding 1000. The machine learning classifiers used in this scenario are Random Forest, XG Boost, Logistic Regression, and KNeighbors.

In the machine learning model, the accuracy rate achieved after incorporating attributes like age, number of pregnancies, BMI, HDL levels, family history of diabetes, and other relevant factors is reported to be 96%.

Moving on to the deep learning models, the focus is on LSTM (Long Short-Term Memory) and Bi-LSTM (Bidirectional LSTM) models for predicting GDM during pregnancy.LSTM (Long Short-Term Memory): LSTM, a type of Recurrent Neural Network (RNN), addresses the vanishing gradient problem and excels in processing sequential data by maintaining long-term memory through components like cell state, forget gate, input gate, and output gate. By leveraging attributes such as BMI, HDL levels, family history of diabetes, and others, the LSTM model achieves an accuracy of around 97%. Bi-LSTM extends LSTM by processing input sequences in both forward and backward directions, capturing information from past and future contexts. This model's architecture involves two LSTM layers processing data bidirectionally, leading to enhanced performance in natural language processing tasks.

The key advantage of Bi-LSTM lies in its ability to consider both past and future context, contributing to improved model performance. By incorporating attributes like age, BMI, HDL levels, and other relevant factors, the Bi-LSTM model achieves an accuracy rate of approximately 97%.

Chapter 8

Future Enhancement

Incorporate Additional Data Sources: Explore integrating data from various sources, such as electronic health records, wearable devices, and lifestyle factors, to further enhance the predictive capabilities of the models.

- Explainable AI: Develop more interpretable and explainable models to provide healthcare professionals with insights into the key factors contributing to GDM prediction, improving trust and adoption in clinical settings.
- Personalized Risk Assessment: Explore the potential of the models to provide personalized risk assessments for pregnant women, enabling targeted interventions and preventive measures.
- Continuous Model Improvement: Implement mechanisms for ongoing model monitoring, retraining, and refinement to adapt to evolving data and healthcare practices, ensuring the system remains up-to-date and effective.
- Deployment and Integration: Seamlessly integrate the developed models into clinical workflows and decision support systems, enabling real-time GDM prediction and facilitating timely interventions.

By addressing these future enhancements, the system architecture can continue to evolve and provide more accurate, personalized, and clinically-relevant predictions for Gestational Diabetes Mellitus, ultimately improving maternal and fetal health outcome

Chapter 9

Conclusion

The system architecture presented demonstrates a comprehensive approach to predicting Gestational Diabetes Mellitus (GDM) during pregnancy.

The key highlights are:

Machine Learning Ensemble Model: The ensemble model combining Random Forest, XG Boost, Logistic Regression, and KNeighbors Classifiers achieved an accuracy of 96% in predicting GDM.

Deep Learning LSTM Models: The LSTM and Bi-LSTM models, leveraging attributes like age, BMI, HDL levels, and other relevant factors, achieved an accuracy of around 97% in predicting GDM.

Effective Utilization of Data: The system architecture emphasizes the importance of data processing, feature selection, and data splitting to prepare the data for accurate model training and testing.

References

[1] Xiaojia Wang , Yurong Wang , Shanshan Zhang , Lushi Yao , Sheng Xu (2022) : Analysis and Prediction of Gestational Diabetes Mellitus by the Ensemble Learning Method , https://doi.org/10.1007/s44196-022-00110-8

[2] Burçin Kurt et al. (2023): Prediction of gestational diabetes using deep learning and Bayesian optimization and traditional machine learning techniques, https://doi.org/10.1007/s11517-023-02800-7

[3] Rasool Jader and Sadegh Aminifar (2022) : Predictive Model for Diagnosis of Gestational Diabetes in the Kurdistan Region by a Combination of Clustering and Classification Algorithms: An Ensemble Approach, https://doi.org/10.1155/2022/9749579

[4] Byung Soo Kang et al.(2023): Prediction of gestational diabetes mellitus inAsian women using machine learning algorithms, https://doi.org/10.1038/s41598-023-39680-8

[5] Masahiro Watanabe et al. (2023): Prediction of gestational diabetes mellitus using machine learning from birth cohort data of the Japan Environment and Children's Study, https://doi.org/10.1038/s41598-023-44313-1

[6] Tanzina Rahman Hera et al.(2021) : Early Gestational Diabetes Detection Using Neural Network, DOI: 10.37394/23208.2021.18.1

[7] Gabriel Cubillos et al. (2023): Development of machine learning models to predict gestational diabetes risk in the first half of pregnancy, https://doi.org/10.1186/s12884-023-05766-4

[8] Yan Xiog et al.(2020): Prediction of gestational diabetes mellitus in the first 19 weeks of pregnancy using machine learning techniques, https://doi.org/10.1080/14767058.2020.1786517

[9] Muhammad Azeem Sarwar et al. (2019): Prediction of Diabetes Using Machine Learning Algorithm in Healthcare, https://doi.org/10.23919/IConAC.2018.8748992

[10] A.A Ojugo and D. Otakore (2018): Improved Early Detection of Gestational Diabetes via Intelligent Classification Models: A Case of the Niger Delta Region in Nigeria, http://pubs.sciepub.com/jcsa/6/2/5

[11] Marije Lamain -de Ruiter et al. (2017) :Prediction models for the risk of gestational diabetes: a systematic review, DOI 10.1186/s41512-016-0005-7

[12] Shiva Shankar Reddy et al. (2022) : A Novel Approach for Prediction of Gestational Diabetes based on Clinical Signs and Risk Factors, doi: 10.4108/eetsis.v10i3.2697

[13] Taiyu Zhu et al. (2020) : Deep Learning for Diabetes: A Systematic Review, DOI 10.1109/JBHI.2020.3040225

[14] Md Abu Rumman Refat et al.(2021) : A Comparative Analysis of Early Stage Diabetes Prediction using Machine Learning and Deep Learning Approach, DOI: 10.1109/ISPCC53510.2021.9609364

[15] Md. Kamrul Hasan et al.(2020): Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers, Digital Object Identifier 10.1109/ACCESS.2020.2989857