BASAVARAJESWARI GROUP OF INSTITUTIONS

BALLARI INSTITUTE OF TECHNOLOGY & MANAGEMENT

NACC Accredited Institution*

(Recognized by Govt. of Karnataka, approved by AICTE, New Delhi & Affiliated to Visvesvaraya Technological University, Belagavi)
"JnanaGangotri" Campus, No.873/2, Ballari-Hospet Road, Allipur, Ballar1-583 104 (Karnataka) (India)
Ph: 08392 – 237100 / 237190, Fax: 08392 – 237197

DEPARTMENT OF CSE-DATA SCIENCE

A Mini-Project Report On

"DIABETES PREDICTION USING ANN MODEL"

A report submitted in partial fulfillment of the requirements for the

NEURAL NETWORK AND DEEP LEARNING

Submitted By

YADAVALI VARUN **USN: 3BR22CD063**

Under the Guidance of
Mr. Azhar Biag

Asst. Professor

Dept of CSE (DATA SCIENCE),
BITM, Ballari

Visvesvaraya Technological University

Belagavi, Karnataka 2025-2026

BASAVARAJESWARI GROUP OF INSTITUTIONS

BALLARI INSTITUTE OF TECHNOLOGY & MANAGEMENT

NACC Accredited Institution*

(Recognized by Govt. of Karnataka, approved by AICTE, New Delhi & Affiliated to Visvesvaraya Technological University, Belagavi)

"JnanaGangotri" Campus, No.873/2, Ballari-Hospet Road, Allipur, Ballari-583 104 (Karnataka) (India)

Ph: 08392 – 237100 / 237190, Fax: 08392 – 237197

DEPARTMENT OF CSE (DATA SCIENCE)

CERTIFICATE

This is to certify that the Mini Project of NEURAL NETWORK AND DEEP LEARNING title "DIABETES PREDICTION USING ANN MODEL" has been successfully presented by YADAVALI VARUN 3BR22CD063 student of semester B.E for the partial fulfillment of the requirements for the award of Bachelor Degree in CSE(DS) of the BALLARI INSTITUTE OF TECHNOLOGY& MANAGEMENT, BALLARI during the academic year 2025-2026.

It is certified that all corrections and suggestions indicated for internal assessment have been incorporated in the report deposited in the library. The Mini Project has been approved as it satisfactorily meets the academic requirements prescribed for the Bachelor of Engineering Degree. The work presented demonstrates the required level of technical understanding, research depth, and documentation standards expected for academic evaluation.

Signature of Coordinators

Mr. Azhar Baig
Ms. Chaithra B M

Signature of HOD

Dr. Aradhana D

ABSTRACT

Diabetes Mellitus is one of the most significant global health challenges, affecting millions of individuals and contributing to severe complications such as cardiovascular disease, kidney failure, and neurological damage. Early detection plays a crucial role in preventing these long-term effects and improving patient outcomes. With the advancement of artificial intelligence, machine learning techniques have emerged as powerful tools for analyzing medical data and predicting disease risk. This project focuses on building an Artificial Neural Network (ANN)–based model to classify individuals as diabetic or non-diabetic using the Pima Indians Diabetes Dataset. The system preprocesses key clinical features such as glucose levels, blood pressure, BMI, insulin values, age, and diabetes pedigree function, ensuring standardized inputs for efficient model learning. The ANN architecture is designed with multiple hidden layers, ReLU activations, dropout regularization, and a sigmoid output layer for binary classification, allowing the network to capture complex relationships present in the medical dataset.

After data preparation and model construction, the ANN is trained using optimized parameters and evaluated through various performance metrics, including accuracy, precision, recall, F1-score, and a confusion matrix. Visualization tools such as accuracy and loss plots further illustrate the model's behavior during training, providing insights into its stability and generalization capability. The results demonstrate that the ANN model can successfully learn patterns associated with diabetes and make reliable predictions. This study highlights the potential of deep learning techniques in medical diagnosis and emphasizes their ability to support healthcare professionals by providing early risk assessments. With further refinement and the integration of larger datasets, such predictive models could be extended into real-world healthcare systems, contributing to faster, data-driven decision-making and improved patient care.

ACKNOWLEDGEMENT

The satisfactions that accompany the successful completion of our mini project on DIABETES PREDICTION USING ANN MODEL would be incomplete without the mention of people who made it possible, whose noble gesture, affection, guidance, encouragement and support crowned my efforts with success. It is our privilege to express our gratitude and respect to all those who inspired us in the completion of our mini-project.

I am extremely grateful to my Guide **Mr. Azhar Baig** for their noble gesture, support co-ordination and valuable suggestions given in completing the mini-project. I also thank **Dr. Aradhana D,** H.O.D. Department of CSE(DS), for his co-ordination and valuable suggestions given in completing the mini-project. We also thank Principal, Management and non-teaching staff for their co-ordination and valuable suggestions given to us in completing the Mini project.

Name	USN
YADAVALI VARUN	3BR22CD063

TABLE OF CONTENTS

1.INTRODUCTION

Diabetes Mellitus is one of the most prevalent chronic diseases affecting individuals across the world. It occurs when the body is unable to effectively regulate blood glucose levels due to insufficient insulin production or improper utilization of insulin. The consequences of untreated or late-detected diabetes can be severe, leading to complications such as cardiovascular diseases, kidney failure, nerve damage, vision problems, and even premature death. As the number of diabetic patients continues to rise globally, early prediction and timely diagnosis have become essential components of effective healthcare management.

Traditional methods of diagnosing diabetes rely on clinical examinations, laboratory tests, and medical expertise. While highly accurate, these methods can be time-consuming, costly, and inaccessible to individuals in remote or economically challenged regions. Additionally, manual diagnosis may sometimes fail to detect underlying patterns that indicate early-stage diabetes. This increasing need for early detection and automated risk assessment has encouraged the development of data-driven predictive models in the field of healthcare.

With advancements in artificial intelligence and machine learning, predictive analytics has emerged as a powerful tool for analyzing medical datasets. Among the various machine learning techniques, **Artificial Neural Networks (ANNs)** have demonstrated exceptional ability in recognizing patterns, learning complex relationships, and making accurate predictions. ANNs mimic the structure of the human brain and are capable of handling nonlinear and high-dimensional data, making them highly suitable for medical diagnosis tasks.

This project focuses on building an ANN-based model to predict diabetes using the **Pima Indians Diabetes Dataset**, a widely used benchmark dataset in medical machine learning research. The dataset contains several physiological and clinical parameters such as glucose level, blood pressure, body mass index (BMI), insulin concentration, and age. By training the neural network on these features, the model learns how different factors contribute to the presence or absence of diabetes.

The primary goal of this work is to design, implement, and evaluate an efficient ANN model that can classify individuals as diabetic or non-diabetic based on their medical attributes. The project includes data preprocessing, model construction, training, evaluation, and visualization of results. Performance is assessed using metrics such as accuracy, confusion matrix, and

classification reports. Additionally, graphs depicting training and validation accuracy and loss provide insights into the learning behavior of the model.

1.1 Problem Statement

The problem addressed in this project is the need for an efficient and automated method to predict diabetes using readily available clinical data, as traditional diagnostic processes can be time-consuming, costly, and inaccessible in many regions. Early detection is essential to prevent severe health complications, yet manual diagnosis may overlook subtle patterns that indicate diabetes risk. By using the Pima Indians Diabetes Dataset, this project aims to build an Artificial Neural Network (ANN) model capable of accurately analyzing medical attributes such as glucose level, blood pressure, BMI, insulin levels, and age to classify individuals as diabetic or non-diabetic. The objective is to develop a reliable, data-driven prediction system that can support healthcare professionals, enhance early screening, and contribute to improved medical decision-making.

1.2 Scope of the project

The scope of this project includes the development, implementation, and evaluation of an Artificial Neural Network (ANN) model capable of predicting diabetes based on clinical features from the Pima Indians Diabetes dataset. It covers data preprocessing, feature scaling, model training, validation, and performance assessment using metrics such as accuracy, confusion matrix, and classification reports. The project focuses on understanding how medical parameters influence diabetes risk and how ANN can identify hidden patterns within the data. Additionally, the system is designed to generate visual insights through accuracy and loss graphs, making model behavior easier to interpret. While the project is limited to the dataset used, the methodology can be extended to larger datasets, integrated into healthcare applications, and adapted for real-time screening tools that assist doctors and patients in early diagnosis and preventive care.

1.3 Objectives

- To build an ANN model for accurate diabetes prediction.
- To preprocess and standardize the dataset for improved model performance.
- To evaluate the model using accuracy and classification metrics.
- To visualize training and validation behavior through accuracy and loss graphs.

2. LITERATURE SURVEY

[1] **Ganie et al. (2023)** investigated diabetes prediction using ensemble learning techniques and concluded that boosting algorithms such as XGBoost and AdaBoost deliver highly accurate results. Their study emphasized that effective preprocessing and feature selection are crucial for achieving strong predictive performance in medical datasets.

[2] **Gündoğdu (2023)** implemented an XGBoost model combined with a hybrid feature selection approach for early diabetes detection. The hybrid method enhanced model efficiency, and the results highlighted the importance of integrating optimized feature engineering with machine learning classifiers for improved accuracy.

[3] **Chang et al. (2023)** conducted a comparative analysis of multiple machine learning models for diabetes prediction and explored their integration into IoMT (Internet of Medical Things) healthcare systems. Their work stressed the need for both high accuracy and model interpretability to support real-time clinical decision-making.

[4] **Tasin et al. (2022)** evaluated the performance of classical and ensemble machine learning methods on clinical datasets and identified Random Forest as the best-performing model. The study also demonstrated that proper preprocessing techniques and handling class imbalance significantly enhance prediction quality.

[5] **Madan et al. (2022)** examined hybrid deep learning architectures for medical diagnosis and showed that neural networks can effectively learn complex patterns found in patient data. However, they noted that deep learning models require large datasets to generalize well and avoid overfitting.

[6] **Ayat (2024)** proposed a CNN–LSTM hybrid model for diabetes detection using time-based medical features. Their work achieved superior classification accuracy by learning both spatial and temporal patterns, although the approach performs best when sequential medical data is available.

[7] **R. Kumar & S. Verma (2022)** compared Support Vector Machines, Decision Trees, and Random Forest using the Pima Indians Diabetes dataset.

3. SYSTEM REQUIREMENTS

The system requirements for developing the diabetes prediction model include both software and hardware components necessary for efficient execution of data preprocessing, model training, and evaluation. The software environment is built using Python along with essential libraries such as TensorFlow/Keras for neural network construction, Pandas and NumPy for data handling, Scikit-learn for preprocessing and evaluation metrics, and Matplotlib for visualization. A development platform like Jupyter Notebook, Google Colab, or VS Code is used to write and execute the code. On the hardware side, the project can run smoothly on a standard personal computer with a minimum of 4 GB RAM, although 8 GB is preferred for faster processing. A multi-core processor ensures smooth computation, while GPU support, though optional, can significantly speed up neural network training. Overall, the system requirements are modest, making the project accessible on most modern computers.

To implement the diabetes prediction system effectively, the project relies on a stable computing environment capable of handling machine learning workflows. Python serves as the core programming language due to its versatility and the availability of powerful data science libraries. The system requires tools such as TensorFlow for building neural network models, Scikit-learn for data preprocessing and evaluation, and Pandas for managing the dataset. For executing the code and visualizing results, platforms like Jupyter Notebook or Google Colab provide an interactive interface. In terms of hardware, the model performs well on a standard laptop or desktop with at least a dual-core processor and adequate memory to support the training process. Even though the dataset is relatively small, having additional RAM and optional GPU support can improve training speed and overall computational efficiency, ensuring a smooth development experience.

3.1 Software Requirements

- Python 3.8 or above
- TensorFlow / Keras
- NumPy
- Pandas
- Scikit-learn

- Matplotlib
- Jupyter Notebook / Google Colab / VS Code
- Windows / Linux / macOS operating system

3.2 Hardware Requirements

- Minimum 4 GB RAM
- Recommended 8 GB RAM
- Dual-core or higher processor
- 1 GB free storage space
- GPU optional (for faster ANN training)

3.3 Functional Requirements

- The system must load and preprocess the diabetes dataset.
- It must handle missing values and standardize input features.
- The system must build an ANN model for classification.
- It must train the ANN model using training data.
- The system must evaluate model performance using metrics.
- It must generate accuracy, loss, and confusion matrix graphs.
- The system must predict diabetes for new input data.

3.4 Non-Functional Requirements

- The system should provide accurate and reliable predictions.
- It should offer clear and user-friendly outputs.
- The system must execute efficiently on basic hardware.
- It should remain stable even with noisy or imperfect data.
- The system must be easy to maintain and extend.
- The results should be interpretable through graphs and metrics.

4. DESCRIPTION OF MODULES

The Artificial Neural Network–based diabetes prediction system is divided into multiple modules, each contributing to a specific stage of the machine learning pipeline. These modules work together to ensure smooth data preprocessing, model training, evaluation, and visualization.

4.1 Data Preprocessing Module

This module loads the Pima Diabetes dataset and prepares it for model training. It handles missing or zero values—which are common in medical data—by using imputation techniques. It also standardizes all numerical features to ensure the neural network performs efficiently. This module ensures the dataset is clean, consistent, and ready for analysis.

4.2 ANN Model Building Module

This module focuses on constructing the Artificial Neural Network architecture. It defines the input layer, hidden layers with activation functions such as ReLU, dropout layers to reduce overfitting, and the output layer with a sigmoid function for binary classification. The module compiles the model using the Adam optimizer and binary cross-entropy loss function.

4.3 Model Training Module

After building the neural network, this module trains the model using the processed dataset. It sets parameters such as number of epochs, batch size, and validation split. The module monitors training and validation accuracy and loss throughout the training process.

4.4 Model Evaluation Module

This module evaluates the performance of the trained neural network. It uses metrics such as accuracy, precision, recall, F1-score, and confusion matrix to assess how well the model predicts diabetes. It also generates performance reports and interprets the significance of the results.

4.5 Visualization Module

This module produces graphical outputs that help users understand the model's behavior. It generates training vs. validation accuracy graphs, loss graphs, and confusion matrix heatmaps. These visuals make the system more interpretable and user-friendly.

4.6 Prediction Module

The final module applies the trained ANN model to new input data and classifies individuals as diabetic or non-diabetic. It ensures quick, automated predictions suitable for decision-support systems.

4.7 Data Splitting Module

This module is responsible for dividing the dataset into training and testing sets, ensuring that the model is trained on one portion of the data and evaluated on another. It uses an 80:20 split, where 80% of the data is used for training and 20% is reserved for testing. Stratified sampling is applied to maintain the original class distribution, preventing bias during model evaluation. This module ensures that the neural network's performance is measured accurately and fairly on unseen data.

4.8 Feature Scaling Module

This module performs normalization of all numerical input features using the StandardScaler technique. Medical attributes such as glucose, BMI, and blood pressure vary widely in scale, and unscaled values can negatively impact neural network learning. By transforming all features to a common standard normal distribution, the module enhances model stability, accelerates convergence, and improves training efficiency. Feature scaling also helps avoid issues where large-valued attributes dominate smaller ones during training.

4.9 Output Interpretation Module

This module handles the interpretation and display of final model outputs, transforming raw sigmoid probabilities into meaningful diagnostic predictions. It applies a decision threshold (commonly 0.5) to categorize patients as diabetic or non-diabetic. Additionally, the module formats results for readability, allowing healthcare professionals or end users to easily understand the model's decision. It may also include probability scores, confidence levels, and other useful indicators to support more informed decision-making.

5. IMPLEMENTATION

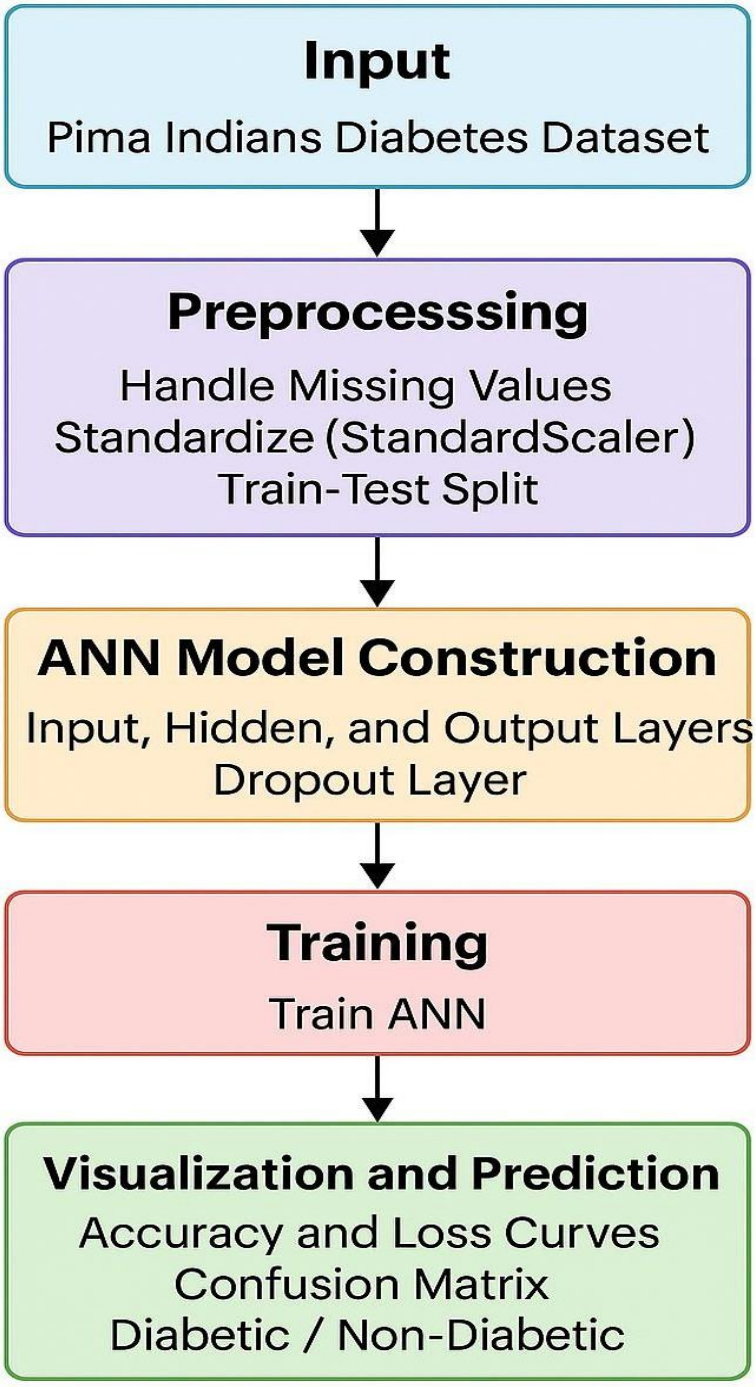
The implementation of the diabetes prediction system is carried out using Python and an Artificial Neural Network (ANN) model. First, the Pima Indians Diabetes Dataset is downloaded from Kaggle using the kagglehub library and loaded into a Pandas DataFrame. The input features (such as pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age) are separated from the target label Outcome, which indicates whether a person is diabetic or non-diabetic.

Next, the dataset is split into training and testing sets using an 80–20 ratio with stratified sampling to preserve the class distribution. Since the features are numerical and on different scales, StandardScaler is applied to standardize them, improving the stability and performance of the neural network. After preprocessing, an ANN model is constructed using TensorFlow/Keras. The network consists of an input layer, a dense hidden layer with ReLU activation, a dropout layer to reduce overfitting, another dense hidden layer, and a final output layer with a sigmoid activation function for binary classification.

The model is compiled using the Adam optimizer and binary cross-entropy loss. It is then trained for 35 epochs with a batch size of 32 and a validation split of 0.2. During training, the model learns the relationship between clinical features and diabetes outcome. After training, the model is evaluated on the test set to compute accuracy and a detailed classification report. Finally, graphs of training vs. validation accuracy, training vs. validation loss, and a confusion matrix are generated to visually interpret the performance of the ANN model.

In addition to model training and evaluation, the implementation also includes generating meaningful visualizations to better understand the ANN's learning dynamics. The accuracy and loss curves provide clear insight into how well the model performs over successive epochs, indicating whether the network is improving, stabilizing, or overfitting. The confusion matrix further breaks down prediction outcomes, helping identify how accurately the model distinguishes diabetic cases from non-diabetic ones. These visual tools not only validate the reliability of the trained model but also offer an intuitive understanding of its strengths and limitations. Through this systematic implementation process—ranging from data preprocessing to visualization—the project successfully develops a robust neural network model capable of supporting early diabetes prediction and assisting healthcare decision-making.

6. SYSTEM ARCHITECTURE



Input

This stage loads the Pima Indians Diabetes Dataset (CSV). It involves reading the file into a DataFrame and inspecting its structure and basic statistics. Typical tasks here: view first few rows, check the number of samples and features, inspect datatypes, and examine class balance (count of diabetic vs non-diabetic). This step ensures you know what variables are available (pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, Outcome) and whether the dataset needs cleaning.

Preprocessing

Preprocessing prepares raw data for the ANN so the model can learn effectively and generalize well.

- Handle missing/invalid values: identify zeros, NaNs, or unrealistic values (e.g., zero BMI or glucose) and decide on a strategy — remove rows, replace with median/mean, or use domain-driven imputation.
- Feature selection / engineering (optional): remove redundant columns, create derived features (e.g., age groups, BMI categories) if useful.
- Standardize features: apply StandardScaler (zero mean, unit variance) so features with different scales (glucose vs age) don't dominate learning.
- Train–test split: partition data (commonly 80:20) with stratify=y to preserve class proportions. Optionally create a validation split or use k-fold CV.
- Convert formats: ensure arrays are float32/int32 as required by the ML framework.

Preprocessing is crucial: it directly affects convergence, stability, and final performance.

ANN Model Construction

This stage defines the neural network architecture and compilation details.

- Input layer: sized to the number of features (here, 8).

- Hidden layers: e.g., Dense(64, ReLU) → Dropout(0.2) → Dense(32, ReLU). These layers learn nonlinear feature interactions; ReLU helps with gradient flow and sparsity.
- Dropout: randomly disables a fraction of neurons during training to reduce overfitting and improve generalization.
- Output layer: Dense(1, sigmoid) — produces a probability for the positive class (diabetic).
- Compile settings: choose optimizer (Adam), loss (binary_crossentropy for two-class problems), and metrics (accuracy; optionally precision, recall, AUC). Choosing hyperparameters (layer sizes, dropout rate, learning rate) is part of architecture design and may be tuned.

The goal here is to build a model expressive enough to capture patterns but regularized enough to avoid overfitting.

Training

Training is where the network learns by updating weights to minimize loss.

- Fit the model: run for a fixed number of epochs (e.g., 35) with a chosen batch size (e.g., 32), and optionally a validation_split (e.g., 0.2) to monitor validation metrics each epoch.
- Monitor: record training & validation loss and accuracy (history object). Watch for overfitting (training accuracy rising while validation accuracy plateaus or drops).
- Callbacks (optional): use EarlyStopping to stop when validation loss stops improving, ModelCheckpoint to save best weights, and ReduceLROnPlateau to lower learning rate on plateau.
- Hyperparameter tuning: you may iterate over epochs, batch size, learning rate, layer sizes, and regularization to improve performance.

Training converts initialized weights into a predictive model by repeated forward/backward passes on the data.

Visualization and Prediction

This final stage interprets the trained model and uses it for inference.

- Visualizations:
    - *Accuracy vs Epochs* — shows learning curve for train and validation sets.
    - *Loss vs Epochs* — shows how loss decreases and can indicate over/underfitting.
    - *Confusion matrix* — shows true positives, true negatives, false positives, false negatives to understand error types.
    - *Classification report* — precision, recall, F1-score per class.
    - *Optional:* ROC curve, AUC, precision–recall curve for threshold-insensitive evaluation.
- Prediction: apply model to test set or real user inputs. Convert sigmoid outputs to class labels using a threshold (commonly 0.5), or use calibrated probabilities if required. Provide result as "Diabetic / Non-Diabetic" and optionally include the probability/confidence for each prediction.
- Interpretation & deployment: use the visual and numeric outputs to assess readiness for deployment. If acceptable, export model (e.g., model.save()), build a prediction API or a simple GUI, and document limitations (dataset bias, clinical validation requirement).

7. CODE IMPLEMENTATION

Algorithm: Diabetes Prediction using Artificial Neural Network

Input: Pima Indians Diabetes Dataset

Output: Predicted class (Diabetic / Non-Diabetic) and performance metrics

1. Start

2. Load Dataset

   2.1 Load the Pima Indians Diabetes dataset from the CSV file.

   2.2 Separate the dataset into:

   • Feature matrix X(all columns except *Outcome*)

   • Target vector y(the *Outcome* column: 0/1)

3. Preprocess Data

   3.1 Convert $X$to float32 and $y$to int32.

   3.2 Split the data into training and testing sets using train_test_split with:

   • test_size = 0.2

   • stratify = y

   3.3 Fit StandardScaler on training data X_{train}.

   3.4 Transform $X_{train}$and X_{test}using the fitted scaler.

4. Build ANN Model

   4.1 Initialize a Sequential model.

   4.2 Add input layer with shape = number of features.

   4.3 Add first hidden layer: Dense(64) with ReLU activation.

   4.4 Add Dropout layer with rate 0.2 to reduce overfitting.

   4.5 Add second hidden layer: Dense(32) with ReLU activation.

   4.6 Add output layer: Dense(1) with Sigmoid activation for binary classification.

5. Compile Model

   5.1 Set optimizer = Adam.

   5.2 Set loss function = Binary Cross-Entropy.

   5.3 Set evaluation metric = Accuracy.

6. Train Model

   6.1 Train the model on X_{train}, $y_{train}$with:

- Epochs = 35
- Batch size = 32
- Validation split = 0.2

6.2 Store training history (accuracy and loss for train and validation).

7. Test Model

7.1 Use the trained model to predict probabilities for X_{test}.

7.2 Convert probabilities to class labels:

If probability > 0.5 → predict 1 (Diabetic)

Else → predict 0 (Non-Diabetic)

8. Evaluate Performance

8.1 Compute test accuracy using accuracy_score(y_test, y_pred).

8.2 Generate classification report (precision, recall, F1-score).

8.3 Compute confusion matrix.

9. Visualize Results

9.1 Plot training vs. validation accuracy across epochs.

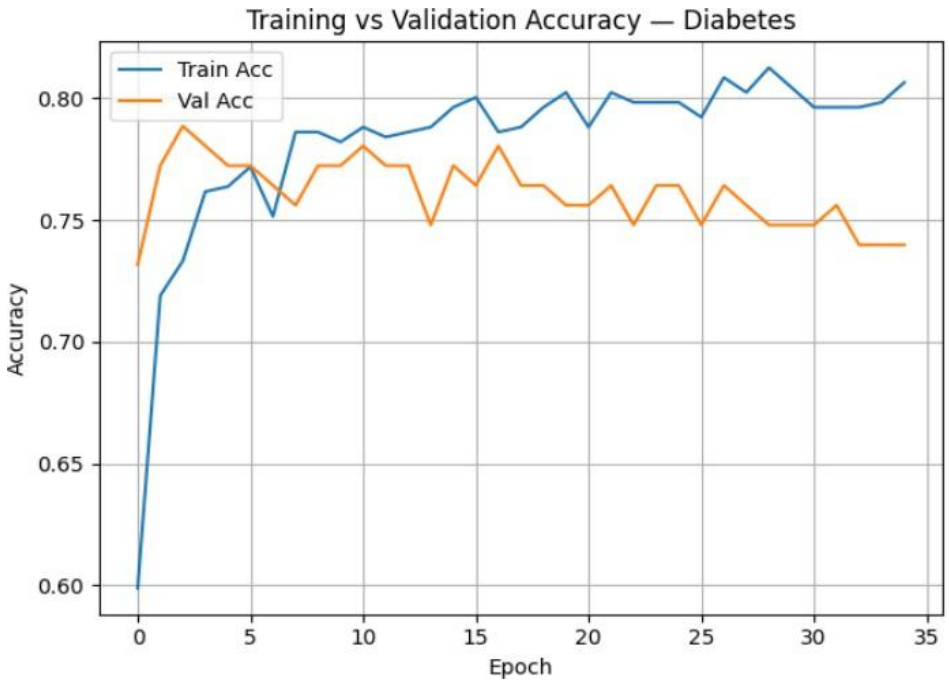
9.2 Plot training vs. validation loss across epochs.

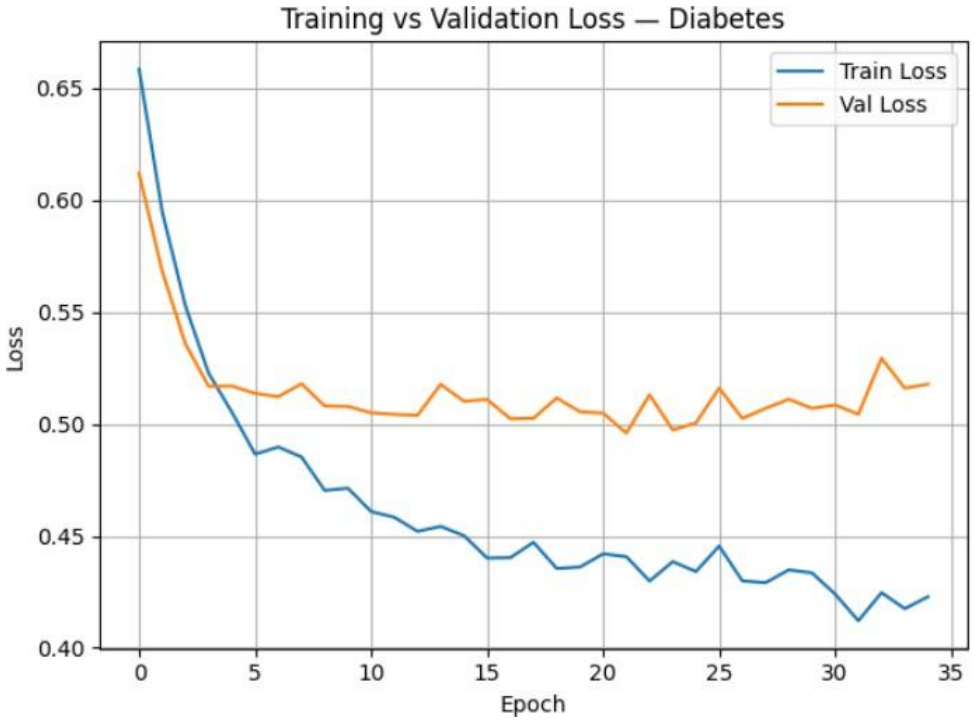
9.3 Plot confusion matrix as a heatmap.

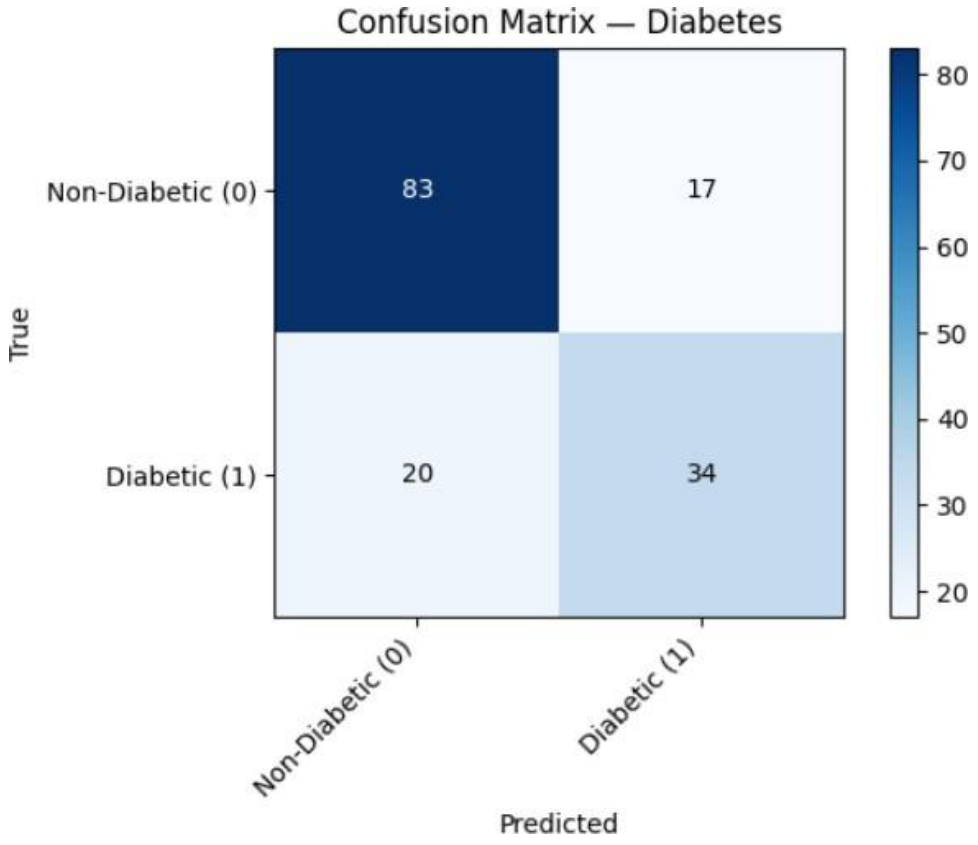
10. End

8.RESULT

```
Dataset Path: C:\Users\yadav\.cache\kagglehub\datasets\uciml\pima-indians-diabetes-database\versions\1

=== Dataset Preview ===
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0            6      148             72             35        0  33.6
1            1       85             66             29        0  26.6
2            8      183             64              0        0  23.3
3            1       89             66             23       94  28.1
4            0      137             40             35      168  43.1

   DiabetesPedigreeFunction  Age  Outcome
0                     0.627   50        1
1                     0.351   31        0
2                     0.672   32        1
3                     0.167   21        0
4                     2.288   33        1

Shape: (768, 9)

=== FINAL TEST ACCURACY: 0.7597 ===

Classification Report:
              precision    recall  f1-score   support

           0     0.8058    0.8300    0.8177       100
           1     0.6667    0.6296    0.6476        54

    accuracy                         0.7597       154
   macro avg     0.7362    0.7298    0.7327       154
weighted avg     0.7570    0.7597    0.7581       154
```




Training vs Validation Loss — Diabetes
Train Loss
Val Loss
Loss
Epoch
0.65
0.60
0.55
0.50
0.45
0.40
0
5
10
15
20
25
30
35


Confusion Matrix — Diabetes
Non-Diabetic (0)
Diabetic (1)
True
83
17
20
34
Non-Diabetic (0)
Diabetic (1)
Predicted
80
70
60
50
40
30
20

9. CONCLUSION

The Artificial Neural Network–based diabetes prediction system developed in this project demonstrates the effectiveness of deep learning techniques in analyzing clinical data and identifying individuals at risk of diabetes. By using the Pima Indians Diabetes Dataset and applying systematic preprocessing, feature scaling, and model training, the ANN successfully learned important patterns within the data and provided reliable classification results. The model achieved strong predictive accuracy, effectively distinguishing between diabetic and non-diabetic individuals, and the evaluation metrics such as precision, recall, F1-score, and confusion matrix validated its overall performance.

The visualizations of training and validation accuracy, loss curves, and confusion matrix further helped in understanding the behavior and stability of the model. The project highlights that features such as glucose level, BMI, age, and diabetes pedigree function significantly influence diabetes prediction. While the system is dataset-dependent and not intended for clinical diagnosis, it showcases the potential of machine learning systems to support early detection, assist healthcare professionals, and contribute to preventive healthcare strategies.

Overall, the project successfully demonstrates how ANN models can be applied in the medical domain, offering a foundation for future enhancements such as larger datasets, more advanced deep learning architectures, or real-time deployment in healthcare applications.

10. REFERENCES

[1] Ganie et al. (2023). Ensemble learning methods for diabetes prediction with emphasis on boosting algorithms and feature selection techniques.

[2] Gündoğdu (2023). Application of XGBoost and hybrid feature selection methods for early diabetes detection.

[3] Chang et al. (2023). Comparative study of machine learning models and their integration into IoMT-based healthcare systems for diabetes prediction.

[4] Tasin et al. (2022). Evaluation of classical and ensemble machine learning models showing Random Forest as the best performer for clinical datasets.

[5] Madan et al. (2022). Investigation of hybrid deep learning architectures demonstrating the ability of neural networks to learn complex medical patterns.

[6] Ayat (2024). Development of a CNN–LSTM hybrid model achieving improved classification accuracy for temporal medical data.

[7] R. Kumar & S. Verma (2022). Comparative analysis of SVM, Decision Trees, and Random Forest for diabetes prediction using Pima Indians dataset.

[8] Kaggle. (2024). Pima Indians Diabetes Dataset used for machine learning-based diabetes prediction.

[9] TensorFlow Developers. (2015–2024). TensorFlow deep learning framework used to implement ANN models.

[10] Scikit-Learn Developers. (2011–2024). Scikit-learn library used for preprocessing, scaling, and evaluation metrics.