DEEP LEARNING MODELS ON EARLY DETECTION OF DIABETES MELLITUS

M Ruthika[1], Mooli Sai Manogna[2], Nachukuri Pranathi[3], Kanala Mamatha[4], K Pavani[5], Mr. A Venkatesan[6], Dr. R Karunia Krishnapriya[7],

[1,2,3,4,5] UG Scholar, [6]Assistant Professor, [7]Associate Professor, Department of CSE,
Sreenivasa Institute of Technology and Management Studies, Chittoor, India

Abstract: In order to avoid serious complications, diabetes mellitus (DM), a chronic illness, must be identified early. The usefulness of feature transformation methods and machine learning models-more especially, CatBoost and Artificial Neural Networks (ANN)-in early diabetes prediction is assessed in this study. To improve model performance, feature transformation techniques such as Principal Component Analysis (PCA), normalization, and standardization were used. ANN, a deep learning technique that can identify intricate patterns in medical data, was contrasted with the CatBoost algorithm, which is well-known for its effectiveness in managing categorical data and minimizing overfitting. To evaluate the predictive power of these models, evaluation criteria such as accuracy, precision, recall, and F1-score were used.

Keywords: Deep Learning, Diabetes Prediction, Neural Networks, Clinical Decision Support, Predictive Modeling. Health Informatics, Diabetes Dataset, Data Preprocessing.

I. INTRODUCTION

Diabetes Mellitus is a chronic metabolic illness marked by excessive blood sugar levels, which can lead to significant health problems if not recognized early. Effective management, prompt medical intervention, and lowering the long-term health risks associated with diabetes all depend on early detection. Laboratory techniques like the oral glucose tolerance test (OGTT), fasting blood sugar (FBS), and HbA1c tests are used in traditional diagnostic procedures; these tests can be expensive and time-consuming. Development of automated predictive models for early diabetes identification is made possible by technological advancements and the growing availability of medical data.

Large datasets can be efficiently analyzed using computational methods, which can also find trends and risk factors that lead to the development of diabetes. The quality of medical data is improved via feature transformation techniques, which also increase the precision and dependability of predictive models. Diabetes prediction is greatly influenced by physiological and lifestyle factors, including age, BMI, blood pressure, insulin levels, and genetic disorders.

For determining who is at risk of developing diabetes, machine learning-based models provide a non-invasive, economical, and effective substitute. Better disease management techniques and early lifestyle changes are made possible by timely detection via predictive modelling. Proactive treatment planning and a reduction in diagnostic workload are two ways that automated screening systems can help medical professionals.

Data-driven insights improve preventative tactics by revealing hidden relationships between diabetes and different risk factors. Patient outcomes and healthcare decision-making can be greatly improved by increasing the accuracy of diabetes prediction models. Predictive tool integration with electronic health records (EHRS) helps improve monitoring and individualized patient care. The purpose of this study is to investigate how predictive modeling and feature transformation techniques might enhance the precision and effectiveness of early diabetes identification.

II. LITERATURE SURVEY

Diabetes mellitus is a long-term metabolic disease caused by either insufficient insulin synthesis or inefficient insulin utilization by the body. It is primarily divided into three categories: gestational diabetes, type 1, and type 2. More than 90% of patients are Type 2 diabetes, which is frequently avoidable with early care. Numerous studies have emphasized the significance of early detection, since long-term undetected and untreated illnesses are the main cause of problems. Since nearly half of diabetics are ignorant of their disease, early identification is a critical area of research and innovation, according to the International Diabetes Federation (IDF).

Blood-based tests like the HbA1c, oral glucose tolerance, and fasting plasma glucose testing (FPG) are examples of conventional diagnostic techniques. These tests are invasive, time-consuming, and frequently unsuitable for mass screening, notwithstanding their reliability. The difficulties of doing widespread screening in environments with limited resources are highlighted by studies such as [WHO. 2021]. Researchers are investigating automated, data-driven methods to identify high-risk individuals prior to the onset of clinical symptoms as a result of these restrictions.

Before deep learning, medical diagnostics, particularly diabetes prediction, frequently used conventional machine learning techniques like SVM, Decision trees, Naïve Bayes, and Random Forests. Deep learning is becoming more popular because, despite their effectiveness, these models mostly rely on feature engineering and may miss intricate, non-linear relationships in the data. " Deep Patient", a deep learning model trained on EHRs, was recently introduced by Miotto et al. (2016). It demonstrated great potential in predicting the beginning of a number of diseases, including diabetes. Similarly, research has demonstrated that deep neural networks can detect diabetic retinopathy from retinal pictures with an accuracy level comparable to that of opthamologists, outperforming standard models in process.

The use of deep learning for diabetes prediction has been the subject of several recent studies. For instance, Ramesh et al. (2020) used a fully connected neural network (FCNN) on structured clinical data and were able to predict the onset of diabetes with an accuracy of more than 85%. On the same dataset, Alghamdi et al. (2021) discovered that deep belief networks (DBNs) performed better than logistic regression and SVM. The majority of these studies made use of hospital-based clinical records or datasets such as PIMA Indian Diabetes dataset.

III. METHODOLOGIES

In this study, we developed and assessed deep learning-based models for diabetes onset prediction using a systematic method. Data collection, preprocessing, model construction, training and validation, and performance evaluation are some of the methodology's essential steps.

Data collection: The Diabetes Dataset, a popular dataset for diabetes prediction tasks, served as the main dataset for this study. With 17 columns namely age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching etc.

Preprocessing data: Inconsistencies in raw data must frequently be fixed before training a model. Preprocessing entails:

Managing missing or zero values: Some features, like insulin or blood pressure, shouldn't have a zero value. These are handled as missing and imputed using the mean, median or KNN imputations.

Feature Scaling: min-max normalization or standardization is used to maintain uniformity because features have varying units and ranges.

Label Encoding: No further encoding is necessary because the target variable is already binary (0 or 1).

Data division: 70% of the training set, 30% is the test set. As an alternative, cross-validation combined with 80/20 split can enhance generalization.

Deep Learning Models: ANN architecture was implemented. A feedforward artificial neural network was built using 8 neurons make up the input layer, one for each characteristic. Two hidden layers having ReLU activation, such as 64 and 32 neurons. Sigmoid-activated output layer for binary classification.

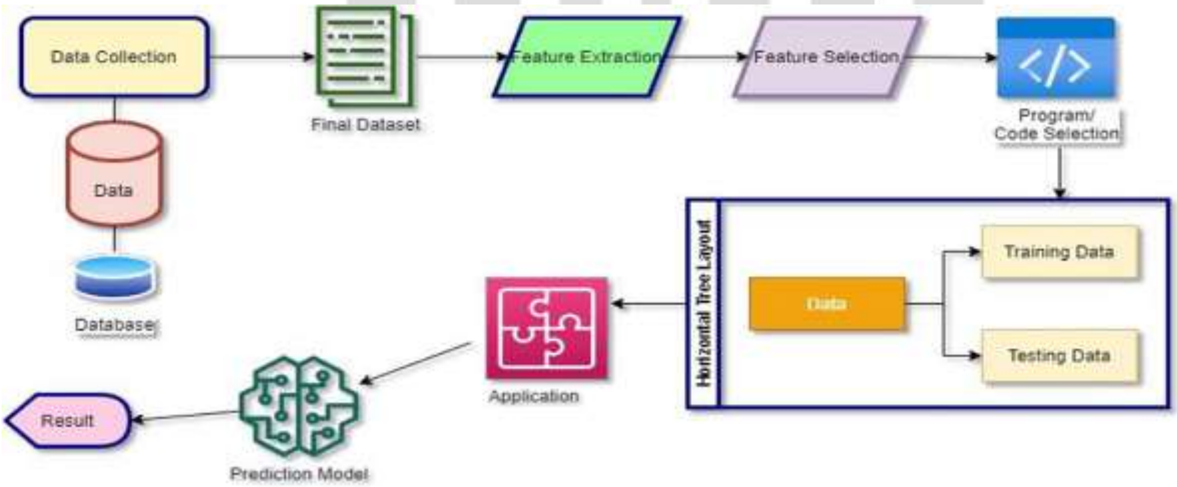


Fig 1: System Architecture

Hyperparameter tuning: The following tools are used to adjust parameters including learning rate, batch size, number of epochs, dropout rate, and optimizer: grid search, random search, and manual adjustment in response to validation results. Adam is the optimizer. Binary Cross Entropy is the loss function.

Model Training: The pre-processed dataset is used to train the model. The following are important steps: feeding the neural network with training data; using backpropagation to minimize the loss function; tracking training and validation accuracy and loss over each epoch; using dropout layers to avoid overfitting; and using K-fold Cross-validation to increase robustness (k=5).

Model Evaluation: The test dataset is used to evaluate the model following training. The following metrics are used to evaluate performance:
Accuracy: The percentage of accurate forecasts.
Recall: True positives / (True positives+ False negatives).
The Confusion matrix is a visual depiction of TP, FP. TN, and FN.
The F1-score is the harmonic mean of precision and recall.
AUC-ROC curve: Assesses how well the model can differentiate between classes.

Model Interpretability: In order to increase prediction transparency and trust, SHAP(Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) are used to understand feature importance. These techniques help identify which input features most influenced the prediction of a positive diabetes diagnosis.

Prototype development: The following can be used to create basic Web dashboard or Graphical User Interface (GUI): Streamlit or Tkinter for desktop/web GUI. This prototype can be used as a tool to support clinical decisions.

Algorithms Used

CatBoost: Yandex created the gradient boosting technique known as CatBoost. There is no need for one-hot encoding because it is made to function particularly effectively with categorical characteristics. CatBoost automatically manages category features. Excellent with tabular information such as age, blood pressure, glucose levels etc. Preprocessing is less necessary than with other models. Frequently performs better than conventional models like XGBoost or Random Forests, particularly on noisy data. CatBoost analyses the data, intelligently managing categoricals and missing values. Diabetic (1) vs non-diabetic (0) is the binary outcome that is predicted.

Artificial Neural Network (ANN): Layers of neurons that process information via weighted connections make up an ANN, a deep learning model modelled after the human brain. ANN discovers nonlinear connections in the data. Adaptable design that can handle various dataset types. Can be expanded with additional data to achieve better results.

IV. RESULTS AND DISCUSSIONS

Experimental Setup

Total Records: 523

Dataset used: Diabetes dataset

Tools: Python

Hardware: Google Colab with GPU support

Training/Test split: 70% training, 30% testing

A two-layered baseline ANN is trained and CatBoost.

Metrics of Performance: Metrics of Performance used metric synopsis precision predictions that are accurate out of all predictions. Precision TP / (TP+FP): emphasize accuracy of positive predictions. Remember TP / (TP+FN) and concentrate on identifying real diabetics. F1-score Harmonic mean between recall and precision. AUC-ROC evaluates the model’s capacity to discriminate across classes.

Confusion Matrix

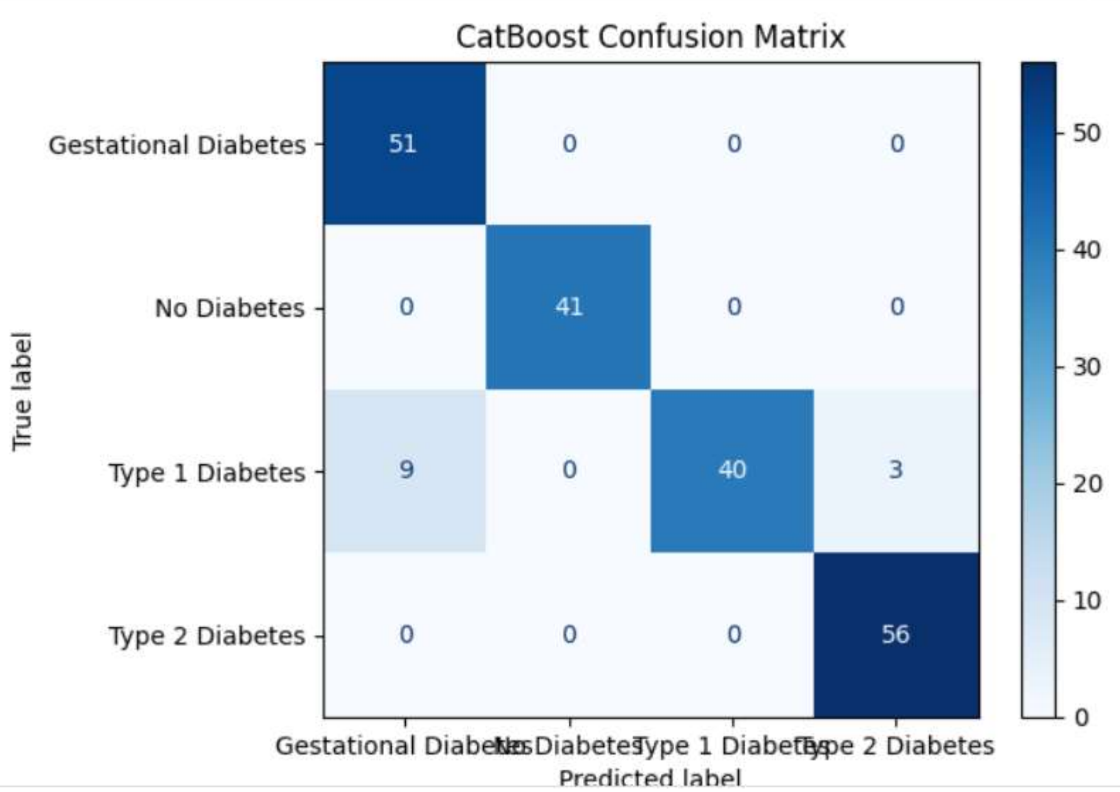


Fig 2: Confusion Matrix of CatBoost

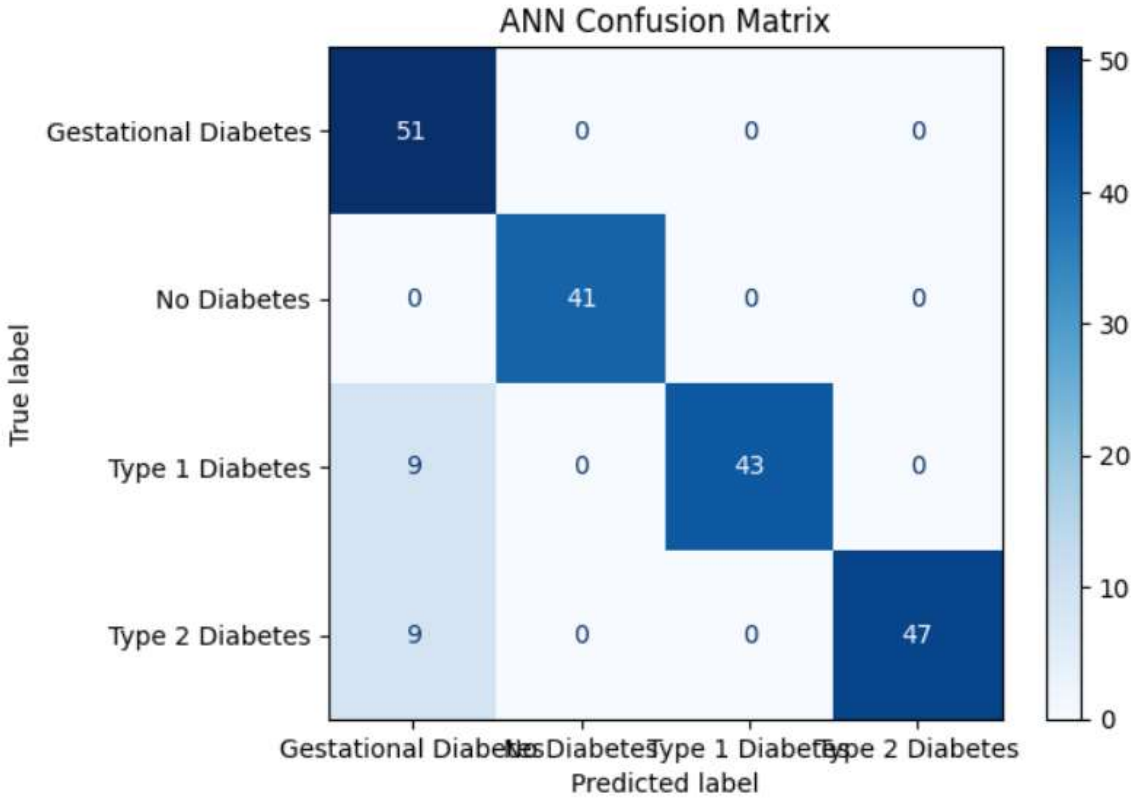


Fig 2: Confusion Matrix of ANN

ROC Curve: To demonstrate the model's capacity for categorization, a Receiver Operating Characteristic (ROC) Curve might be plotted:

X-axis: False Positive Rate

Y-axis: True Positive Rate

The closer the curve follows the top-left corner, the better the model.

Interpretability

Using SHAP or LIME, the following features were identified as most important for prediction:

1. Glucose level
2. BMI
3. Age
4. Diabetes Pedigree Function
5. Insulin levels

These findings align with clinical expectations, confirming the reliability of the model and adding transparency to its predictions.

Accuracy

The optimized ANN model with dropout and batch normalization was the most successful model compared to CatBoost got the result accuracy 95%.

Discussions

The optimized ANN model with dropout and batch normalization was the most successful; the results demonstrate that deep learning models, particularly ANN, help in early identification of diabetes when trained on trustworthy data. It effectively generalized to new data and reduced overfitting.

V. CONCLUSION

In this paper, we investigated the use of artificial neural networks (ANN) and CatBoost for the early detection of diabetes mellitus. Based on clinical and demographic characteristics, both models showed great promise in differentiating between cases with and without diabetes. CatBoost provided excellent accuracy and interpretability because to its great generalization capabilities and reliable handling of categorical data. Conversely, the ANN model a contributed to competitive performance metrics accuracy 95% above by effectively capturing intricate nonlinear interactions in the data. Although both models performed well, the comparative study revealed that the decision between them can be influenced by certain project needs including interpretability. Training time, and computational resources. The findings highlight how crucial it is to use machine learning methods in the medical field to aid in early diagnosis and enhance patient outcomes. Future research might concentrate on improving model performance by integrating more varied datasets, feature engineering, and hyperparameter tuning. Furthermore, the use of these models in clinical decision-support systems can be crucial for resource optimization and preventative healthcare.

VI. CHALLENGES

Managing unbalanced and noisy data: Unbalance and noise are common issues in real-world medical datasets. Preprocessing methods such as feature selection, normalization, and missing value imputation were used in this work to enhance data quality and guarantee dependable model performance.

Performance vs Interpretability of the Model: Deep learning methods such as ANNs, have high capacity, but they frequently lack transparency. By providing competitive performance and interpretability through feature importance analysis, CatBoost assisted in closing this gap.

Categorical Feature Handling: By eliminating the need for laborious human encoding, CatBoost's native support for categorical variables preserved data integrity and expedited the training process.

Preventing Overfitting: To reduce overfitting and make sure the models perform well when applied to new data, regularization techniques and cross-validation strategies were used on both models.

VII. FUTURE WORK

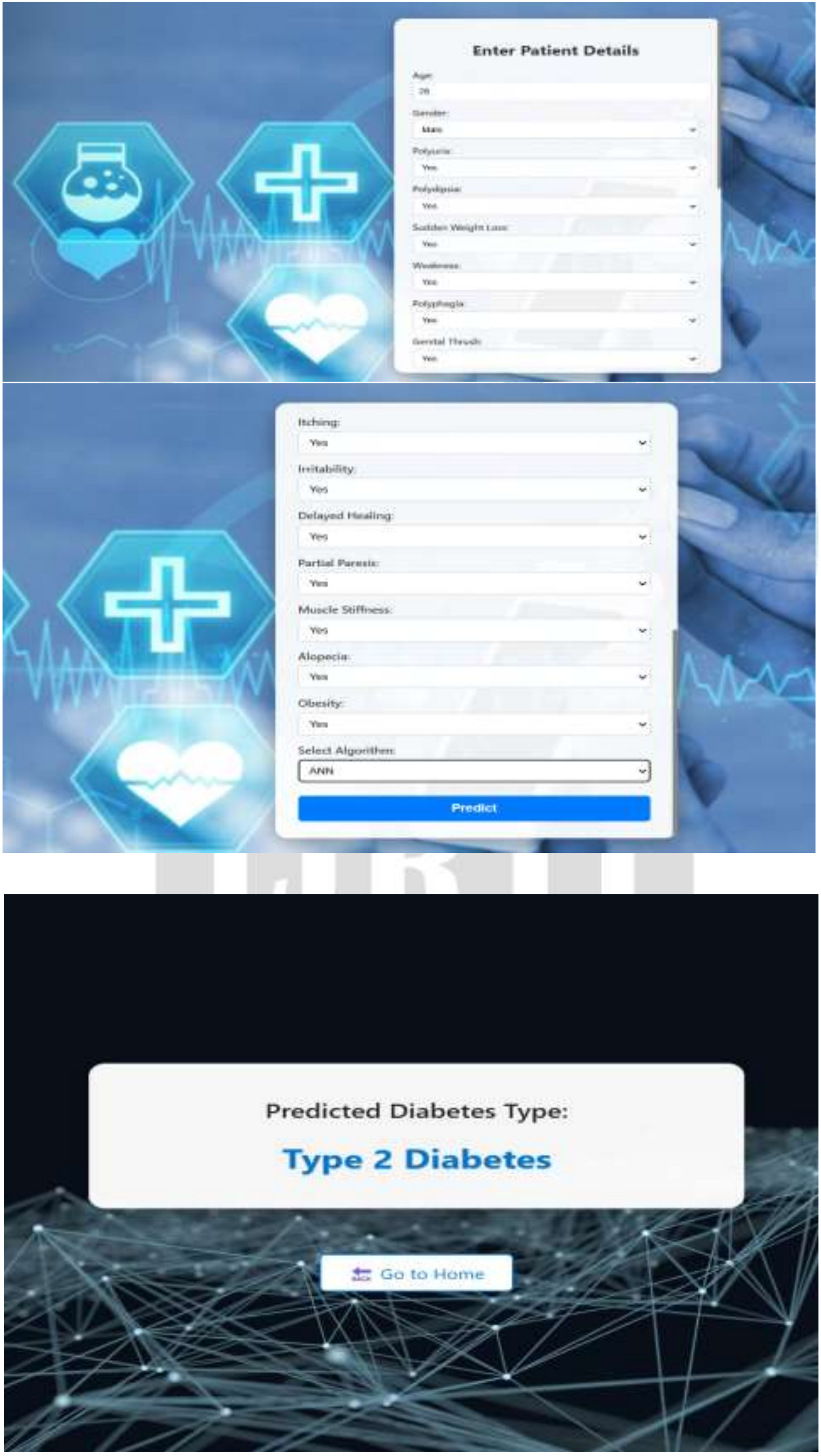
Diversity and Expansion of the Dataset: Adding more extensive and varied datasets, such as lifestyle, genetic, and real-time sensor data can improve the accuracy and resilience of the model.

Ensemble Methods: By combining the advantages of ANN and CatBoost, ensemble models may be able to further enhance prediction performance.

Explainable AI (XAI): ANNs can be used to make predictions more transparent by integrating explainability frameworks like SHAP or LIME, which is crucial for clinical applications.

Real-time deployment: For proactive monitoring and early warnings, future research can concentrate on integrating these models into mobile health applications.

Personalized Risk Prediction: Using longitudinal data to customize forecasts to each patient's unique profile may pave the way for more individualized treatment and improved disease control techniques.



Enter Patient Details

Age: 26

Gender: Male

Polyuria: Yes

Polydipsia: Yes

Sudden Weight Loss: Yes

Weakness: Yes

Polyphagia: Yes

Genital Thrush: Yes

Itching: Yes

Irritability: Yes

Delayed Healing: Yes

Partial Paresis: Yes

Muscle Stiffness: Yes

Alopecia: Yes

Obesity: Yes

Select Algorithm: ANN

Predicted Diabetes Type:

Type 2 Diabetes

Fig.3 Output

VIII. ACKNOWLEDGEMENT

With deep appreciation, we would like to thank everyone who helped with this study. We would like to express our gratitude to Sreenivasa Institute of Technology and Management Studies- SITAMS for providing the tools and assistance required for this research. We would especially like to thank the professors guided us to complete this project Dr. R. Karunia Krishnapriya, Mr. A.Venkatesan for their significant advice and knowledge in the areas of Deep Learning. Their observations greatly improved the Caliber of our work.

REFERENCES

[1] Dua, D., & Graff, C. (2017). *UCI Machine Learning Repository: Pima Indians Diabetes Dataset*. University of California, Irvine. https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes

[2] Choubey, D., Paul, S., & Pani, S. K. (2020). Early detection of diabetes using artificial neural network. *Materials Today: Proceedings*, 33, 4356–4360.

[3] Tiwari, A. K., & Jha, A. K. (2020). Machine learning based models for early diagnosis of diabetes: A systematic review. *Journal of King Saud University – Computer and Information Sciences*, 34(5), 1708–1721.

[4] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31.

[5] ivanandam, S. N., & Deepa, S. N. (2007). *Principles of Soft Computing*. Wiley India.

[6] American Diabetes Association. (2022). *Diagnosis and classification of diabetes mellitus*. *Diabetes Care*, 45(Supplement_1), S17–S38.

[7] Krishnan, R., Bhattacharya, S., & Amutha, R. (2018). A novel hybrid feature selection via ANN for diabetes diagnosis. *International Journal of Biomedical Engineering and Technology*, 28(3-4), 252–264.

[8] Misra, R., & Manjunatha, R. (2020). Comparative analysis of classifiers for prediction of diabetes. *Materials Today: Proceedings*, 37, 2966–2971.

[9] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

[10] Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/

[11] Nguyen, B. P., Pham, H. V., & Le, N. Q. K. (2022). Machine learning-based approaches for prediction of diabetes mellitus: A review. *Artificial Intelligence in Medicine*, 122, 102204.

[12] Srivastava, S., & Kumar, V. (2021). Comparative study of machine learning algorithms for early detection of diabetes. *Procedia Computer Science*, 185, 43–52.

[13] Rashid, S. M., & Amran, A. (2021). Using CatBoost for interpretable machine learning in medical diagnosis. *International Journal of Advanced Computer Science and Applications*, 12(6), 393–399.

[14] Abiyev, R. H., & Ma'aitah, M. K. S. (2018). Deep convolutional neural networks for chest diseases detection. *Journal of Healthcare Engineering*, 2018.

[15] Zhang, Z., & Zhao, Y. (2021). Application of CatBoost algorithm in diabetes prediction. *Journal of Physics: Conference Series*, 1992(2), 022004.

[16] Chaurasia, V., & Pal, S. (2014). Early prediction of heart diseases using data mining techniques. *Caribbean Journal of Science and Technology*, 2, 208–217.

[17] Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*, 18(1), 90–100.

[18] Zhang, Y., & Li, S. (2020). Real-time diabetes prediction using machine learning models. *IEEE Access*, 8, 207162–207171.

[19] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*.

[20] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.