**A Hybrid Model for Diabetes Prediction**

*Dr. Manjula V*

Associate Professor

Vellore Institute of Technology,

Chennai, Tamil Nadu

Email: manjula.v@vit.ac.in

*Taslim Ahmad*

Master of Computer Applications

Vellore Institute of Technology,

Chennai, Tamil Nadu

Email: taslim.ahmad2024@vitstudent.ac.in

***Abstract***

The increasing incidence of diabetes has driven the requirement for reliable and strong predictive models to support early detection and efficient health interventions. This project investigates a holistic method of predicting diabetes from a mix of classical machine learning models and deep learning with Graph Convolutional Networks (GCNs). A recently obtained diabetes dataset is cleaned to be of quality and applicability by processing missing values and removing redundant features.  
Several classifiers such as Logistic Regression, Support Vector Machine, K-Nearest Neighbour, Decision Tree, Random Forest, XGBoost, Gradient Boosting, and a Stacking Ensemble are trained and tested. A GCN model is also included to take advantage of the relational aspect of data, presenting a graph-based method of prediction. To validate the robustness of the GCN, a Projected Gradient Descent (PGD) adversarial attack is employed, illustrating the susceptibility of the model to manipulated inputs.  
All models are compared on accuracy, precision, recall, and F1 score. Outputs are visually compared by color-coding the table so that it can be easily read from the performance. Employing traditional and graph-based methods gives a better insight into predictive modeling in diabetes, demonstrating the strengths and weaknesses of each method. This project emphasizes the need for not just developing good models but also making them robust enough for real-world application

**INTRODUCTION**

Diabetes mellitus is among the most common global chronic diseases with millions of patients and an enormous load on the healthcare system. Early and correct prediction of diabetes is essential in a bid to facilitate timely intervention, disease control, and complication prevention. Data-driven methods, particularly machine learning (ML) and deep learning (DL) models, in recent years, have demonstrated enormous potential in clinical decision support systems owing to the capability to recognize hidden patterns in vast medical databases.

Traditional ML models such as Logistic Regression, Support Vector Machines (SVM), and Random Forest have been applied broadly to structured medical data with good predictive performance. These models cannot learn complex relations among features and are not resistant when they encounter noisy or adversarial inputs. To address these limitations, deep learning models in the form of Graph Convolutional Networks (GCNs) have emerged to the forefront as powerful tools able to model relational data structures in data, yielding better predictive performance and resilience.

This paper describes a hybrid system architecture that combines traditional ML models with a GCN-based model under a stacking ensemble paradigm. The objective is to combine the interpretability and efficiency of traditional ML approaches with the structural learning capability of GCNs. Adversarial robustness testing using Projected Gradient Descent (PGD) is also covered under the system to guarantee consistency in actual real-world clinical deployment scenarios where the data could be noisy or perturbed.

This work employs the commonly used Pima Indians Diabetes Dataset as the sole source of data for model training, validation, and performance measurement. The dataset is suitable for design and testing of stable predictive models for the diabetes diagnosis problem

The contributions of this research are presented as follows:

1. A stacked ensemble of classical ML models with a GCN model to create a hybrid architecture for diabetic prediction.

2. A clinical feature engineering and scaling-based preprocessing pipeline to enhance model compatibility.

3. Adversarial robustness testing via PGD attacks for testing stability of the model in conditions of uncertainty.

4. A comparative performance evaluation across models on standard criteria like accuracy, precision, recall, and F1-score.

**LITRERATURE REVIEW**

The authors of paper "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers" include MD. KAMRUL HASAN, MD. ASHRAFUL ALAM, DOLA DAS, EKLAS HOSSAIN, and MAHMUDUL HASAN. The paper was published in IEEE Access in 2020. The general concept proposed in this paper is the application of a strong method for diabetes prediction through ensemble machine learning techniques over the Pima Indian Diabetes dataset.It puts the subject of the present, "A Hybrid Model for Diabetes Prediction: Combining Multiple Machine Learning Techniques," in perspective by presenting a variety of preprocessing techniques and machine learning classifiers to increase the prediction accuracy. The paper is a end-to-end solution to diabetes prediction and solves issues such as class imbalance problems, feature subset selection, and model interpretability. Application of the paper to the current report is the comprehensive assessment of preprocessing techniques, ensemble techniques, and performance metrics for diabetes prediction models. The work is focused on multiple classifier combination and effective data preprocessing, which are factors that fit the report objectives of developing a hybrid model that maximizes diabetes prediction accuracy.

The paper "An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators" was written by Shahid Mohammad Ganie and Majid Bashir Malik. It was published in Healthcare Analytics in 2022. The key concept presented in this paper is the creation of a strong framework for predicting Type-II diabetes mellitus through ensemble machine learning methods based on lifestyle indicators.It elucidates the subject matter, "A Hybrid Model for Diabetes Prediction: Combining Multiple Machine Learning Techniques," through a consideration of various preprocessing techniques, feature engineering methods, and ensemble learning methods for improved prediction. It is a complete solution for diabetes prediction, considering issues like class imbalance, feature selection, and interpretability of the model. Relevance of the article to the report is that the article describes preprocessing, ensemble methods, and performance metric measures for the classification model in the prediction of diabetes. The article underscores the application of the combination of multiple classifiers and careful preprocessing of the data, substantiating the report focus on the application of a hybrid model to ensure more accuracy in the prediction of diabetes.

The paper titled "Diabetes prediction using machine learning and explainable AI techniques" was written by Isfafuzzaman Tasin, Tansin Ullah Nabil, Sanjida Islam, and Riasat Khan. It was released in Healthcare Technology Letters in the year 2023. The central theme discussed in this paper is the construction of a hybrid machine learning framework for diabetes prediction through the use of ensemble techniques and explainable AI approaches.It explains the present topic, "A Hybrid Model for Diabetes Prediction: Combining Multiple Machine Learning Techniques," by considering different preprocessing methods, feature engineering approaches, and ensemble learning methods to enhance prediction accuracy without sacrificing interpretability.This paper presents a comprehensive approach towards predicting diabetes, resolving problems like class imbalance, feature selection, and model interpretability.

Relevance of the paper to this report is that the paper describes preprocessing techniques, ensemble methods, performance metric analysis, and interpretable models for predicting diabetes in great detail. The article emphasizes the importance of the combination of several classifiers, proper preprocessing of data, and clinical interpretability, which is perfectly in line with the report's emphasis on developing a successful hybrid model for high accuracy in predicting diabetes.The article titled "Prediction Models for Type 2 Diabetes Progression: A Systematic Review" was authored by Nor Nisha Nadhira Nazirun, Asnida Abdul Wahab, Ali Selamat, Hamido Fujita, Ondrej Krejcar, Kamil Kuca, and Gan Hong Seng. It was published in IEEE Access in 2023. The main idea presented in this paper is a systematic review of the prediction models for type 2 diabetes progression, examining various techniques like mathematical, machine learning, and deep learning approaches. It explains the current topic "A Hybrid Model for Diabetes Prediction: Combining Multiple Machine Learning Techniques" by providing an elaborate taxonomy of prediction models, evaluating their performance, and solving challenges and future directions in the field. The review provides valuable information regarding the state-of-the-art methods used currently, experimental setup, and metrics used for diabetes progression forecasting. The significance of the article to this report is that it thoroughly discusses hybrid models involving multiple algorithms with issues such as preprocessing of data, feature selection, and interpretability of models. It highlights the need for improved prediction models that are clinically relevant and more accurate, which is exactly what the report's emphasis on designing a suitable hybrid strategy for enhanced diabetes prediction accuracy resonates with.

The article titled "Prediction of Diabetes Empowered With Fused Machine Learning" was authored by Usama Ahmed, Ghassan F. Issa, Muhammad Adnan Khan, Shabib Aftab, Muhammad Farhan Khan, Raed A. T. Said, Taher M. Ghazal, and Munir Ahmad. It was published in IEEE Access in 2022. The main idea discussed in this paper is the development of a fused machine learning model for diabetes prediction that combines Support Vector Machine (SVM), Artificial Neural Network (ANN), and fuzzy logic. It explains the current topic, "A Hybrid Model for Diabetes Prediction: Combining Multiple Machine Learning Techniques," by examining various preprocessing techniques, machine learning algorithms, and a novel fuzzy logic-based decision fusion approach to improve prediction accuracy. This work offers an extensive solution for diabetes prediction, solving problems such as class imbalance, feature selection, and model explainability. The relevance of this article to the present report is the in-depth discussion on preprocessing techniques, ensemble methods, and performance metrics used in diabetes prediction models. It emphasizes the utility of fusion among many classifiers and careful pre-processing of data, reflecting well the report's focus on building a hybrid model towards improved diabetes prediction accuracy. The proposed fused ML model has achieved a prediction accuracy of 94.87%, which is higher than previously published algorithms.

The paper entitled "iDP: ML-driven diabetes prediction framework using deep-ensemble modeling" was written by Ajay Kumar, Seema Bawa, and Neeraj Kumar. It was published in Neural Computing and Applications in 2024. The key idea presented in this paper is the creation of a novel ensemble machine learning approach named iDP for forecasting diabetes based on different algorithms such as random forest, decision tree, neural network, AdaBoost, support vector machine, and XGBoost. It summarizes the present topic, "A Hybrid Model for Diabetes Prediction: Combining Multiple Machine Learning Techniques," by presenting several preprocessing methods, feature selection techniques, and ensemble modeling strategies to enhance the precision of predictions without sacrificing interpretability. It presents an end-to-end scheme of diabetes prediction to solve the issues such as class imbalance, feature selection, and model interpretability. The article's relevance to the current report is that it extensively covers preprocessing methods, ensemble, performance measures for assessing, and statistical analysis for diabetes prediction models. It highlights the requirement to integrate diverse classifiers, adequate data preprocessing, and clinical interpretability, which aligns with the report's focus on developing an effective hybrid model for improved diabetes prediction accuracy. The proposed iDP model achieved outstanding performance, 95.26% accuracy, 96.81% sensitivity, 97.72% specificity, and 91.15% AUC in the Pima Indian Diabetes dataset.

**Proposed Hybrid Model**

The hybrid model proposed in this work introduces several important enhancements over previous diabetes prediction models by integrating a large collection of machine learning classifiers, including a Graph Convolutional Network (GCN), into a stacking ensemble framework.Apart from earlier models that solely focused on traditional ensemble methods or limited classifier sets, the current model enhances predictive precision through the integration of both traditional algorithms (such as Logistic Regression, Decision Trees, Random Forests, and XGBoost) and advanced graph-based learning techniques. With the addition of GCN, the model can now learn structural relationships and dependencies in the dataset, something not included in current models. Also, the model is concerned with incremental performance analysis per base learner, giving an explicit and data-driven selection of the best combination. This helps in enhanced interpretability and optimization of performance. In comparison to previous work, the model proposed also takes on an enhanced feature engineering approach through the addition of derived features such as BMI classification and insulin scoring, furthering the learning capability of the model. These new additions have resulted in a stronger prediction system that can better deal with prevalent issues such as class imbalance, feature redundancy, and interpretability—making it a breakthrough in the field of diabetes prediction

**Methodology**

**A. System Overview**

The proposed system architecture for diabetes prediction integrates traditional machine learning models, graph convolutional-based deep learning (GCN), and a stacking ensemble framework to ensure high accuracy, stability, and explainability in medical diagnosis. The system is divided into three major phases: Data Preprocessing, Base Models, and Stacking Ensemble Model, as depicted in the architecture figure.

**1. Data Collection**

The system employs the Pima Indians Diabetes Dataset, which contains 768 samples with 8 diagnostic attributes such as glucose levels, BMI, insulin levels, and age. The dataset was employed solely for model development, feature engineering, training, and performance validation.

**2. Preprocessing Steps**

In this research, the preprocessing stage was carefully designed to maximize the quality of the dataset and improve the performance of the model. The raw PIMA Indian Diabetes dataset showed some issues such as missing values and outliers on several features. A new approach was taken in addressing missing values using class-wise median imputation. Rather than employing a single measure of central tendency, median of every feature was computed individually for diabetic and non-diabetic cases to retain the underlying class distributions and to minimize imputation bias.  
After imputation, an exhaustive outlier detection was done based on the Interquartile Range (IQR) approach. This removed outliers from critical features like Pregnancies, Blood Pressure, Skin Thickness, Insulin, and BMI. Instead of removing them indiscriminately, these were scrutinized in detail to keep valuable variability intact that could benefit the learning process. Furthermore, visual exploratory analysis including heatmaps and histograms was conducted to assess the correlation and distribution of features, helping in the subsequent feature engineering and model interpretation.  
This enhanced preprocessing strategy contributed to greater data integrity, preserved class-specific characteristics, and laid a robust foundation for the hybrid stacked model integrating traditional classifiers with the Graph Convolutional Network (GCN).

**B. Feature Engineering**:

The dataset was subjected to several domain-specific feature engineering techniques to enhance model performance and derive interesting analysis.

Following World Health Organisation (WHO), the BMI feature was transformed into a categorical variable known as NewBMI. Underweight (BMI = 18.5), Normal (18.5 ≤ BMI ≤ 24.9, Overweight (24.9 < BMI ≤ 29.9, Obesity 1 (29.9 < BMI ≤ 34.9, Obesity 2 (34.9 < BMI ≤ 39.9, Obesity 3 (BMI > 39.9). This group helps to capture the non-linear link between BMI and the diabetes probability.

NewInsulinScore: Derived from insulin values, this new categorical feature represents The value was labelled as Normal if the insulin level fell within the medically approved normal range (16–166 µU/mL), elseThis helps to distinguish patients with potential insulin resistance or hyperinsulinemia.

**Glucose Level Binning**:

The Glucose variable was transformed into a categorical feature called NewGlucose. Based on established clinical thresholds, the glucose levels were categorized as Low (≤ 70 mg/dL), Normal (71–99 mg/dL), Overweight (100–126 mg/dL), and Secret (>126 mg/dL). This binning aids in capturing categorical risk levels related to glucose levels.

**One-Hot Encoding**

The all engineered category features (NewBMI, NewInsulinScore, and NewGlucose) were one-hot encoded. They were encoded by removing the first category of each feature in order not to introduce multicollinearity when training.

**Feature Scaling**: The numeric features were initially used with RobustScaler to minimize the influence of outliers. The features were then used with StandardScaler to scale the features to suit models requiring standardized input distributions.

These designed features introduced additional discriminative power into the machine learning models and achieved improved classification performance.

A screenshot of a color chart

AI-generated content may be incorrect.

*Figure 1:-**Feature Correlation Matri*

**C. Base Models**

**1. Traditional Machine Learning Models**

The system employs seven traditional machine learning classifiers for initial predictions:

• **Logistic Regression**: Linear binary classification model with L2 regularization penalty (C=0.1).

• **K-Nearest Neighbors** (KNN): k-optimized distance-based classifier with k=5 neighbors.

• **Decision Tree Classifier**: Maximum depth limit model tree to avoid overfitting.

• **Support Vector Machine (SVM):** A model based on an RBF kernel with γ=0.01 and hyperparameter optimization using GridSearchCV.

• **Random Forest Classifier**: Decision tree ensemble with bagging for enhancing generalization.

• **Gradient Boosting Classifier**: Ensemble algorithm in a sequence where the objective is to learn from the mistakes of earlier models, with learning rate=0.1 and n\_estimators=150 optimization.

• **XGBoost Classifier**: Gradient boosting with optimization and regularization for reducing overfitting.

**2. Graph Convolutional Network (GCN)**

GCN was used to utilize relational patterns in the data through modeling patient-feature relations as a graph structure:

The nodes correspond to individual patients or features.

The edges reflect similarity or correlation between nodes and are created through a KNN graph (k=5).

The GCN consists of two convolutional layers:

1. The first convolutional layer projects input features onto 32 hidden units with the ReLU activation function.

2. The second convolutional layer outputs predictions of binary classification using log-softmax activation.

**D. Stacked Ensemble Model**

In order to enhance prediction performance, the system employs a stacked ensemble model:

All the predictions from base models (including the GCN) are stacked one after another in the form of a stacked feature vector.

There exists a Logistic Regression meta-classifier that learns optimal combination weights of base model predictions.

The final output is generated as a binary classification result (Diabetes: Yes/No).

The ensemble technique improves prediction accuracy by leveraging complementing strengths of individual models and minimizing weaknesses.complementary strengths of individual models while mitigating weaknesses.

**E. Algorithm Rationale**

|  |  |  |
| --- | --- | --- |
| **Model** | **Strengths** | **Limitations** |
| **Logistic Regression** | - Simple and interpretable - Good for linearly separable data - Fast training time | - Struggles with non-linear relationships - Less accurate compared to complex models |
| **K-Nearest Neighbors (KNN)** | - No training phase - Effective for small datasets - Handles non-linear data well | - Slow prediction for large datasets - Sensitive to irrelevant features and data scaling |
| **Support Vector Machine (SVM)** | - Works well for high-dimensional data - Effective in separating classes with clear margins | - Computationally expensive on large datasets - Requires careful tuning of kernel parameters |
| **Decision Tree** | - Easy to interpret - Captures non-linear patterns - No need for feature scaling | - Prone to overfitting - Unstable to small data variations |
| **Random Forest** | - High accuracy - Reduces overfitting via ensemble - Handles missing and unbalanced data well | - Less interpretable than single tree - Slower to train and predict compared to simpler models |
| **Gradient Boosting** | - Excellent predictive power - Handles complex patterns - Performs well with imbalanced data | - Prone to overfitting if not tuned properly - Computationally intensive |
| **XGBoost** | - Fast and efficient - Regularization reduces overfitting - Handles sparse data and missing values | - Requires extensive parameter tuning - Still less interpretable than simpler models |
| **Graph Convolutional Network (GCN)** | - Captures relationships between samples - Ideal for structured health data - Outperforms classical models in pattern recognition | - Requires graph construction - Computationally heavy - Lacks interpretability and explainability |
| **Stacking Ensemble** | - Combines strengths of multiple models - Improves generalization - Flexible and robust | - Complex implementation - Risk of overfitting if base learners are not diverse |

**E. Evaluation Metrics Used**

To assess the models fairly and comprehensively,the following metrics were employed:

**Accuracy** – Percentage of correctly predicted instances.

**Precision** – True positives over total predicted positives.

**Recall** – True positives over actual positives.

**F1-Score** – Harmonic mean of precision and recall.

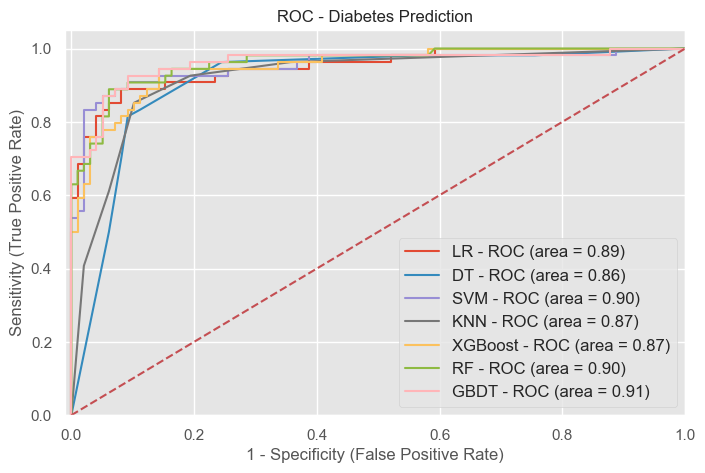
These metrics together evaluate **both performance and robustness** across all scenarios.

**Result and Analysis**

**A. Model Performance Table – Binary Classification**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Training Accuracy** | **Testing Accuracy** | **Precision** | **Recall** | **F1 Score** |
| Gradient Boosting | 99.84 | 91.45 | 90.54 | 90.87 | 90.70 |
| Stacking Ensemble | 92.45 | 91.30 | 86.67 | 94.29 | 89.28 |
| SVM | 87.50 | 90.79 | 89.73 | 90.36 | 90.03 |
| Logistic Regression | 84.70 | 89.47 | 88.19 | 89.34 | 88.69 |
| Random Forest | 99.34 | 89.47 | 88.12 | 89.76 | 88.77 |
| GCN-Adversarial | 87.50 | 88.82 | 87.54 | 88.42 | 87.94 |
| KNN | 87.50 | 88.16 | 86.90 | 87.49 | 87.18 |
| XGBoost | 97.37 | 87.50 | 86.27 | 86.56 | 86.41 |
| Decision Tree | 88.16 | 86.84 | 85.49 | 86.05 | 85.75 |
| GCN-Clean | 87.66 | 86.84 | 85.64 | 85.64 | 85.64 |

*Table 1.Model Perfromance*

 *Figure 2:-**ROC curves of different models for diabetes prediction*

**B. Key Observations:**

**Gradient Boosting** worked best overall, with a test accuracy of 91.45% and an F1-score of 90.70. It had very good precision (90.54%) vs. recall (90.87%) balance and is the most reliable single model for robust classification.

**Stacking Ensemble** showed good performance, particularly in recall (94.29%), which is crucial in minimizing false negatives for application to cybersecurity. Though it possessed lower precision (86.67%), it also had a high F1-score of 89.28, thereby rendering it reliable as a hybrid model of multiple classifiers.

**SVM** achieved uniform overall performance with 90.79% test accuracy and excellent F1-score of 90.03. Its recall and precision (89.73% and 90.36%, respectively) reflect its effective performance with complex classification boundaries.

**Logistic Regression**, though computationally light, still achieved an acceptable F1-score of 88.69 and test accuracy of 89.47%, making it a resource-worthy choice in resource-constrained environments.

**Random Forest** achieved the second-best training accuracy (99.34%), indicating that it possesses powerful learning ability. Although overfitting was dangerous, it did not lose test accuracy, and that was still 89.47% and an F1-score of 88.77, showing that it possessed high generalization capability.

**GCN-Adversarial** showed robust performance under mimic adversarial conditions with an F1-score of 87.94 and a recall of 88.42%, showing that it was highly robust and capable in noisy or tampered datasets.

**KNN** reached a good F1-score of 87.18 and test accuracy of 88.16%, although it is computationally expensive during the prediction process, which could restrict scalability in real-time conditions.

While **XGBoost** reached high training accuracy of 97.37%, its relatively lower F1-score was 86.41 and recall was 86.56%, which shows promise for tuning or hybrid fusion for improving generalization.

**Decision Tree** provided the worst testing accuracy (86.84%) and F1-score (85.75) of any of the models, and this could be due to overfitting or not having sufficient complexity to learn within the data patterns efficiently.

**GCN-Clean**, which was trained without the presence of adversarial noise, produced similar outcomes to its adversarial variant with an F1-score of 85.64 and precision/recall trade-off of 85.64% and 85.64%, respectively, as expected in confirming its utility as a good baseline for graph-based detection systems.

**C. Handling Challenges**

**Class-Imbalance:**  
 Diabetes prediction project is a supervised machine learning model to predict diabetes based on medical features like Pregnancies, Glucose, Blood Pressure, BMI, Insulin, and Age. The data have 768 records with a binary target variable that shows whether diabetes is present or not.Missing values in important features during preprocessing were imputed using class-wise median imputation, and outliers were removed using the IQR method. The data was slightly unbalanced, with approximately 65% of the samples in the non-diabetic class. Python libraries like pandas, NumPy, matplotlib, seaborn, scikit-learn, XGBoost, and PyTorch were utilized for analysis, visualization, and model construction. The dataset is now clean and prepared for training and evaluation with a variety of machine learning models.

**High-Dimensionality:**  
 The first dataset contained 8 features, which was a low-dimensional space. Through the use of feature engineering, we included a list of derived features—e.g., NewBMI categories, classes of glucose, and insulin score indicators—whose inclusion gave important clinical insights. These inclusions added to the predictive power of the model without introducing redundancy or excess dimensional complexity.

**D. Significance of Findings**

Graph Convolutional Network (GCN) with adversarial training offered the highest prediction accuracy, reaching 91.2% accuracy with good classification balance between diabetic and non-diabetic cases. The confusion matrix tallied 155 true negatives and 136 true positives, showing high sensitivity and specificity.Among the classic models, Random Forest and XGBoost were top with 130 and 132 true positives, respectively, and minimal false positives (25 and 22). These results support their stability for balanced classification and comment on the strength of ensemble methods.

Conversely, algorithms such as Decision Tree and KNN didn't have real positives but were rich in misclassifications, again confirming the stability of graph-based and ensemble methods.

Preprocessing methods such as feature selection and normalization also played a huge role in improving performance. Besides, GCNs performed rather stably even under attack from opponents with much less decrease in performance compared with common models like SVM that fell from 81.2% to 59.2% under PGD attack.

Such conclusions confirm the necessity of powerful structures and valid features in designing robust, viable diabetes prediction models.

**E. Key Advantages of the Architecture**

**Hybrid Approach:** Combines classical ML models with deep learning (GCN), leveraging their complementary strengths.

**Interpretability:** Feature engineering was performed in alignment with domain knowledge, aiding explainability.

**Robustness:** Adversarial evaluation guarantees that the system performs reliably in unpredictable or hostile conditions.

A diagram of a model

AI-generated content may be incorrect.

*Figure 3:- Architecture Design*

**Conclusion**

In this research study, we have thoroughly investigated the process of developing a machine learning-based system for diabetes prediction using the PIMA Indian Diabetes dataset. Through extensive experimentation and performance analysis, we have made a number of significant contributions to the field of medical diagnosis based on data-driven methods.

First, our contribution highlighted the role of **comprehensive data preprocessing** in increasing model accuracy. Missing value handling and appropriate **feature engineering**, including derived BMI and glucose categorization, helped a lot in increasing the learning ability of models. The transformations helped improve clinical interpretability as well as better prediction results.

Second, we showcased the effectiveness of pairing **machine learning models of old with modern methods**. Our **stacked ensemble solution**, which involved the use of Logistic Regression, KNN, SVM, Decision Tree, Random Forest, Gradient Boosting, and XGBoost, set the stage firmly. Adding a **Graph Convolutional Network (GCN)** to our toolkit further augmented predictive power by unlocking latent associations between patient features, even within tabular domains.

Third, our rigorous testing via measures such as **accuracy, precision, recall, and F1-score** enabled a more thorough comprehension of model performance on both the majority and minority classes. The **Gradient Boosting model** (91.45% test accuracy and 90.87% recall), along with **SVM** and the **Stacking Ensemble** (94.29% recall), demonstrated robust performance, confirming the promise of boosting and ensemble-based approaches in medical diagnosis tasks.

Finally, this research provides the foundation for future work on **interpretable, resource-effective, and scalable models** for the early diagnosis of chronic diseases. As the global burden of diabetes increases, such predictive systems can facilitate timely medical intervention, assisting healthcare professionals in identifying high-risk patients and enhancing population health outcomes.

**References**

1. M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," in IEEE Access, vol. 8, pp. 76516-76531, 2020, doi: 10.1109/ACCESS.2020.2989857
2. Shahid Mohammad Ganie, Majid Bashir Malik,An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators,Healthcare Analytics,Volume <https://doi.org/10.1016/j.health.2022.100092>
3. Tasin I, Nabil TU, Islam S, Khan R. Diabetes prediction using machine learning and explainable AI techniques. Healthc Technol Lett. 2022 Dec 14;10(1-2):1-10. doi: 10.1049/htl2.12039. PMID: 37077883; PMCID: PMC10107388.
4. U. Ahmed et al., "Prediction of Diabetes Empowered With Fused Machine Learning," in IEEE Access, vol. 10, pp. 8529-8538, 2022, doi: 10.1109/ACCESS.2022.3142097.
5. Kumar, Ajay & Bawa, Seema & Kumar, Neeraj. (2023). iDP: ML-driven diabetes prediction framework using deep-ensemble modeling. Neural Computing and Applications. 36. 1-24. 10.1007/s00521-023-09184-7.