**AI BASED DIABETES PREDICTION**

**SYSTEM**

**TEAM MEMBER**

**712221121001 – Aanesht Raj G**

**Phase 5 Document Submission**

**Project Title:** AI Based Diabetes Prediction System

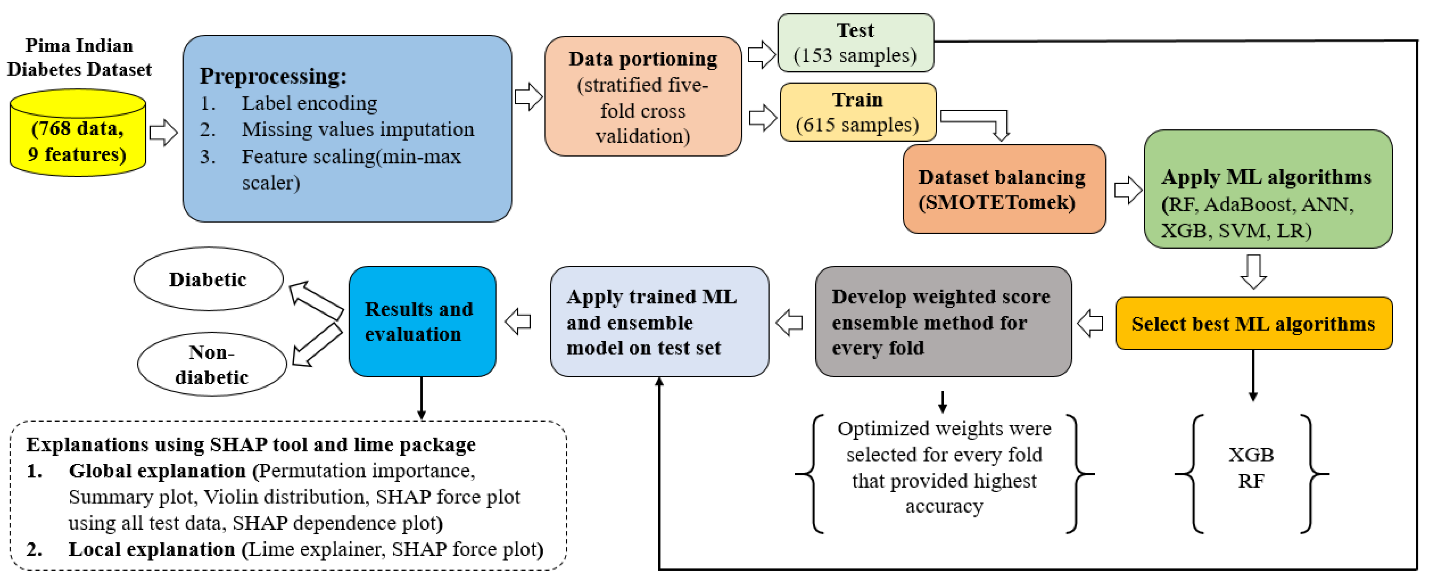
**Phase 5 :** Documentation & Submission

#### **Introduction :**

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar. Hyperglycaemia, or raised blood sugar, is a common effect of uncontrolled diabetes and, over time, leads to severe damage to many of the body’s systems, especially the nerves and blood vessels [1]. In 2014, 8.5% of adults aged 18 years and older had diabetes worldwide.

**Outline the problem Statement :**

The problem statement is to determine which algorithm is more accurate in predicting diabetes. The methodology involves implementing the SVM and decision tree algorithms on the dataset and evaluating their performance using metrics such as accuracy, precision, and recall.



**Design Thinking Process :**

Design thinking is an iterative, non-linear process which focuses on a collaboration between designers and users. It brings innovative solutions to life based on how real users think, feel and behave. This human-centered design process consists of five core stages Empathize, Define, Ideate, Prototype and Test.

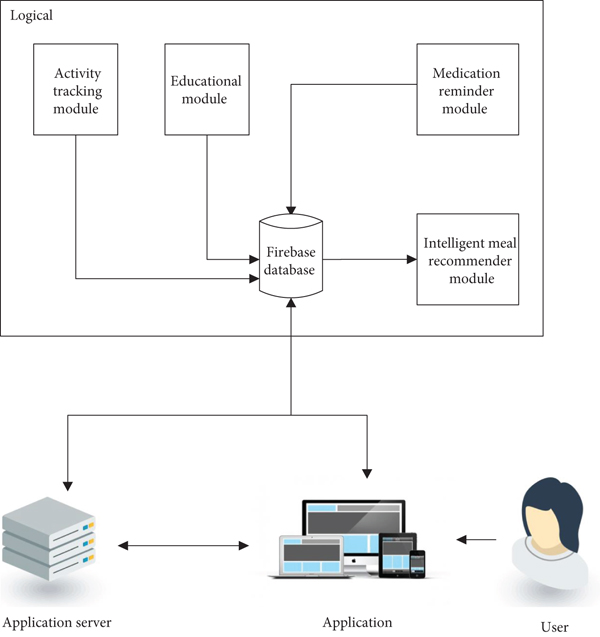
#### **Abstract :**

This paper describes the design and implementation of a software system to improve the management of diabetes using a machine learning approach and to demonstrate and evaluate its effectiveness in controlling diabetes. The proposed approach for this management system handles the various factors that affect the health of people with diabetes by combining multiple artificial intelligence algorithms.

#### The System’s Requirements and Design Analysis :

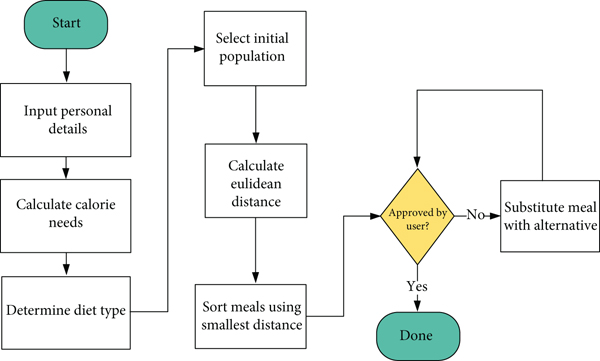
In the design and development of the architecture for the diabetes management system, the clinical requirements and design analysis of the system were based on discussions with collaborators from the Department of Nutrition and Food Science of the University of Ghana and Kwame Nkrumah University of Science and Technology (KNUST). From these discussions, the diet type of patients was determined to be an essential approach suitable for the diabetes management system. The following functionalities were mentioned: (1) Scheduling and reminding diabetic patients to take their medication and blood glucose readings, (2) recommending healthy meals for diabetics to keep their blood glucose levels in check, (3) encouraging and tracking the activity of diabetic patients, (4) providing a visual interface to help them make meaning of their readings and establishing a sufficient connection between the doctor and the diabetic patient using e-mail.

Providing the diabetic patient with a data visualization tool to display the data in tables, charts, and an educational program for newly diagnosed and ongoing diabetes treatment is valuable for the treatment and management of diabetes.



##### The Diabetes Intelligent Meal Recommender Module :

This module schedules a diet for diabetes management by generating whole meals for breakfast, lunch, and supper to meet the nutritional requirements of the diabetic patient. The recommender system uses the knowledge-based approach, where meals are recommended using the user’s profile and a knowledge base. The steps followed for the meal recommendation system are outlined in the flowchart given below in Figure [3](https://www.hindawi.com/journals/ijta/2020/8870141/fig3/).

[[](https://www.hindawi.com/journals/ijta/2020/8870141/fig3/)](https://www.hindawi.com/journals/ijta/2020/8870141/fig3/" \t "_blank)

###### **Data Collection :**

###### The used data in this research consists of two data types, the patient data obtained from an interface provided to the user to input personal details like age, sex, weight, height, and level of activity. The food nutrition data was obtained from the Department of Nutrition and Food Science, University of Ghana, and from the MyFitnessPal database [29]. The diet type of the patient is determined from the obtained data, and calorie needs calculated using the Harris-Benedict’s equation .

### Data Preprocessing

The research process included data preprocessing, model description, hyperparameter tuning, and model evaluation. Data preprocessing entails various methods for improving data quality, such as outlier detection, missing-value imputation, and data normalization (Alasadi and Bhaya [Citation2017](https://www.tandfonline.com/doi/full/10.1080/08839514.2022.2145644)). ML models can improve performance through data preprocessing.

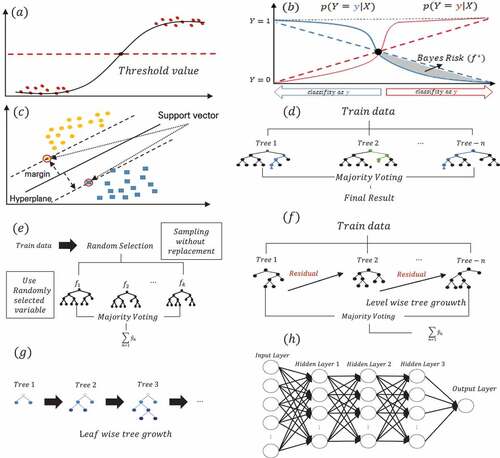
### Data Split

The data were split at an 8:2 ratio to test the training and performance of the predictive model. In addition, GridSearchCV, which simultaneously performs the parameter optimization algorithm grid search and *k*-fold cross-validation, was conducted. In this study, *k* was set to 5, and stratification was set to match the label distribution of each fold to accurately reflect the response variable distribution of the entire dataset.

**Machine Learning Algorithms**

Based on the previous diabetes status, we used logistic regression, naive Bayes, support vector machines, random forests, extremely randomized tree, XG Boost, Light GBM, and multi-layer perceptron to predict the occurrence of prediabetes and diabetes at certain time points, as shown in Figure 2. The multicollinearity problem was addressed before this study by removing variables that were higher than 10.

Figure 2. Machine learning model architecture: (a) Logistic regression; (b) naïve bayes; (c) support vector machine; (d) random forest; (e) extremely randomized tree (f) extreme gradient boosting; (g) light gradient boosting machine; (h) multilayer perceptron.

[](https://www.tandfonline.com/doi/full/10.1080/08839514.2022.2145644)

**Evaluvation Metrics :**

1. **Accuracy**:
   * Accuracy measures the proportion of correct predictions over the total predictions.
   * It is suitable when the class distribution is roughly balanced.
   * Formula: (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives).
2. **Precision**:
   * Precision, also known as positive predictive value, measures the proportion of true positive predictions out of all positive predictions.
   * It helps in assessing how well the model identifies true positives without many false positives.
   * Formula: True Positives / (True Positives + False Positives).
3. **Recall** (Sensitivity or True Positive Rate):
   * Recall measures the proportion of true positives that were correctly identified by the model out of all actual positives.
   * It helps in assessing how well the model avoids missing positive cases.
   * Formula: True Positives / (True Positives + False Negatives).
4. **F1-Score**:
   * The F1-Score is the harmonic mean of precision and recall. It provides a balance between precision and recall.
   * It is useful when you want to consider both false positives and false negatives in your evaluation.
   * Formula: 2 \* (Precision \* Recall) / (Precision + Recall).
5. **Specificity** (True Negative Rate):
   * Specificity measures the proportion of true negatives that were correctly identified by the model out of all actual negatives.
   * It is particularly relevant when false positives are costly.
   * Formula: True Negatives / (True Negatives + False Positives).
6. **Area Under the Receiver Operating Characteristic (ROC-AUC)**:
   * ROC-AUC measures the ability of the model to distinguish between positive and negative cases.
   * It plots the True Positive Rate (Sensitivity) against the False Positive Rate as the discrimination threshold varies.
   * A higher AUC indicates a better model performance.
7. **Area Under the Precision-Recall Curve (PR-AUC)**:
   * PR-AUC is particularly useful when dealing with imbalanced datasets.
   * It measures the area under the precision-recall curve.
   * A higher PR-AUC indicates better performance in terms of precision and recall.
8. **Matthews Correlation Coefficient (MCC)**:
   * MCC takes into account both true and false positives and negatives.
   * It ranges from -1 (completely wrong predictions) to +1 (perfect predictions) and 0 (random predictions).
   * It is suitable for imbalanced datasets.
9. **F-beta Score**:
   * The F-beta Score is a variant of the F1-Score that allows you to adjust the balance between precision and recall using a parameter beta.
   * When beta > 1, it places more weight on recall, and when beta < 1, it places more weight on precision.

**Innovative Techniques :**

1. **Deep Learning and Neural Networks**:
   * Deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are being applied to diabetes prediction.
   * These models can handle complex, high-dimensional data, including medical images and time series data like glucose level readings.
2. **Generative Adversarial Networks (GANs)**:
   * GANs can be used to generate synthetic medical data for training AI models when real data is limited or privacy concerns exist.
   * GANs can create realistic synthetic data, which can be used to enhance the training of predictive models.
3. **Explainable AI (XAI)**:
   * XAI techniques aim to provide interpretable and transparent models for healthcare applications.
   * In the context of diabetes prediction, it's important to understand the model's decision-making process, especially for clinical use.
4. **Personalized Medicine**:
   * AI models are increasingly being tailored to individual patient profiles, taking into account a patient's genetic, lifestyle, and medical history.
   * Personalized medicine can improve prediction accuracy and treatment recommendations.
5. **Feature Engineering**:
   * Innovative feature engineering methods, such as dimensionality reduction and feature selection techniques, can help in extracting relevant information from high-dimensional healthcare data.
6. **Longitudinal Data Analysis**:
   * Utilizing longitudinal patient data to track changes over time is a growing trend.
   * It helps in predicting the progression of diabetes and tailoring treatment plans for individual patients.
7. **IoT and Wearable Devices**:
   * The use of Internet of Things (IoT) and wearable devices allows continuous monitoring of vital signs and glucose levels.
   * Real-time data from these devices can be integrated into AI models for more accurate predictions.
8. **Explanatory Models for Diabetes Risk Factors**:
   * AI systems are being used to discover new risk factors and relationships between various health and lifestyle factors and diabetes.
   * These insights can help in early diagnosis and prevention strategies.
9. **Ensemble Models and Model Stacking**:
   * Combining multiple machine learning models into ensemble models and stacking them can improve prediction accuracy.
   * This technique leverages the strengths of different algorithms.
10. **Extrapolation and Forecasting**:
    * AI models are being used to forecast future diabetes trends, helping healthcare providers and policymakers plan for the future.
11. **Telemedicine and Remote Monitoring**:
    * With the growth of telemedicine, AI models are being integrated into remote monitoring systems to provide continuous care and early detection of diabetes-related issues.
12. **Natural Language Processing (NLP)**:
    * NLP techniques are used to analyze clinical notes, electronic health records, and patient narratives, extracting valuable information for prediction.
13. **Blockchain for Data Security**:
    * Blockchain technology can enhance the security and privacy of patient data, which is crucial in healthcare applications.
14. **Quantum Machine Learning**:
    * As quantum computing technology advances, it may have applications in optimizing complex healthcare models and accelerating data analysis tasks.