Multiple Disease Prediction in Low-Power Edge Devices

M.Mary Adline Priya
Department of Medical Electronics
Saveetha Engineering College
Chennai, India
maryadlinepriyam@saveetha.ac.in

M.Nivethitha
Department of Medical Electronics
Saveetha Engineering College
Chennai, India
nivethitha77@gmail.com

S.Priyanka
Department of Medical Electronics
Saveetha Engineering College
Chennai, India
priyankalax.kkdi@gmail.com

S.Sivaraj Kumar
Department of Medical Electronics
Saveetha Engineering College
Chennai, India
sivarajs2510@gmail.com

***Abstract*—The growing prevalence of chronic and degenerative diseases necessitates efficient and accurate diagnostic tools for timely clinical decision-making. This study presents a Multiple Disease Prediction System that integrates machine learning–based models for heart disease, kidney disease, thyroid disorders, diabetes, and Parkinson's disease into a unified framework. Each disease is modelled using a dedicated classifier with optimized preprocessing and feature selection. The models are trained on publicly available datasets and evaluated using metrics such as accuracy, precision, recall, and F1-score, achieving performance between 90.5% and 99%. The trained models are converted into ONNX format and deployed on the ESP32 microcontroller for real-time inference using an Edge AI approach. User inputs are provided through a mobile-based IoT interface and transmitted via Wi-Fi, enabling on-device prediction with results displayed on an LCD module. The proposed system ensures low latency, improved data privacy, and reduced dependence on cloud infrastructure. Its modular design and embedded implementation make it suitable for portable, real-time healthcare applications in resource-constrained environments.**

Keywords—IoT, Wi-Fi, Graphical User Interface, Support Vector Machine (SVM), and Logistic Regression

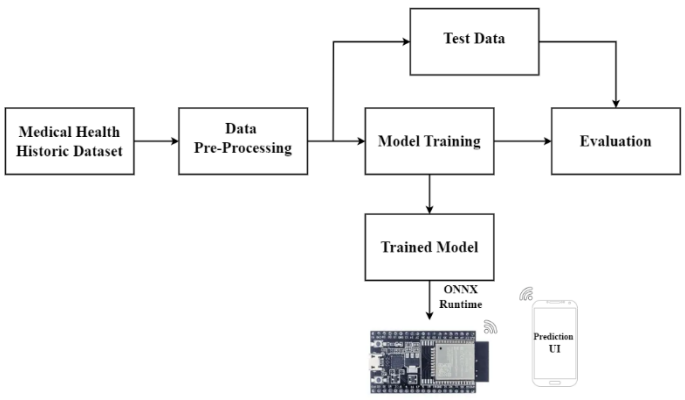
I. INTRODUCTION

This paper proposes a lightweight, edge-based system and machine learning for the prediction of different illnesses models deployed on low-power embedded devices. The system, built around an ESP32 microcontroller, processes clinical and biometric data locally to diagnose conditions such as Parkinson's, diabetes, and heart disease, and transmits results via a graphical user interface (GUI) with optional IoT connectivity. The machine learning framework employs a using logistic regression and support vector machines (SVM) to deliver real-time diagnostic alerts without reliance on cloud infrastructure. This novel approach aims to enhance healthcare accessibility and diagnostic efficiency while addressing challenges posed by limited medical resources in underserved regions. The urgent problem is the growing burden of non-communicable diseases in rural and low-resource settings, where traditional diagnostic tools are often unavailable or unaffordable [1]. Existing AI-driven diagnostic systems typically depend on cloud computation, introducing latency, privacy concerns, and connectivity barriers [2]. Prior research has demonstrated the feasibility of SVM for clinical classification [3], while studies on model compression have enabled neural network deployment on microcontrollers [4]. Additionally, federated learning approaches have been explored for privacy-preserving health analytics [5]. The methodology presented in this work involves training and optimizing a multi-disease prediction model using curated public health datasets, followed by deployment on an edge device to enable offline, real-time medical screening with minimal power consumption.

II. METHODOLOGY

A. Overview

The proposed Multiple Disease Prediction System integrates several machines learning based classification models into a single unified framework for the prediction of heart disease, kidney disease, thyroid disorders, diabetes, and Parkinson's disease. Each disease is modelled using a dedicated classification approach to capture its unique clinical characteristics. The complete methodology is structured into four major stages: collecting and preparing data analysis of specific model creation, and system integration through a unified graphical user interface (GUI). This structured pipeline ensures accurate disease prediction,



modular system design, and seamless interaction between

multiple disease models within a single platform. Figure 1 shows the block diagram for the proposed system.

Fig 1: Proposed System Block Diagram

B. Dataset description

The study's datasets were gathered from public accessible repositories on Kaggle**,** covering multiple disease domains including heart disease, kidney disease, thyroid disorders, diabetes, and Parkinson's disease**.** These datasets are widely used in healthcare machine learning research and contain clinically relevant features derived from medical examinations, laboratory tests, and patient health records. Each dataset corresponds to a specific disease and includes a set of input features along with a target label stating whether the illness is present or not. The datasets vary in terms of feature dimensionality, sample size, and data distribution, reflecting the heterogeneous nature of medical data. This diversity enables the development of disease-specific classification models tailored to the characteristics of each condition. The collected datasets include both numerical and categorical attributes such as physiological measurements, biochemical test results, and patient demographics. Prior to model training, all datasets undergo preprocessing steps including missing value handling, Feature extraction and normalization of categorical variables to guarantee compatibility with machine learning algorithms. Correlation matrix of each dataset is given in Figure 2.

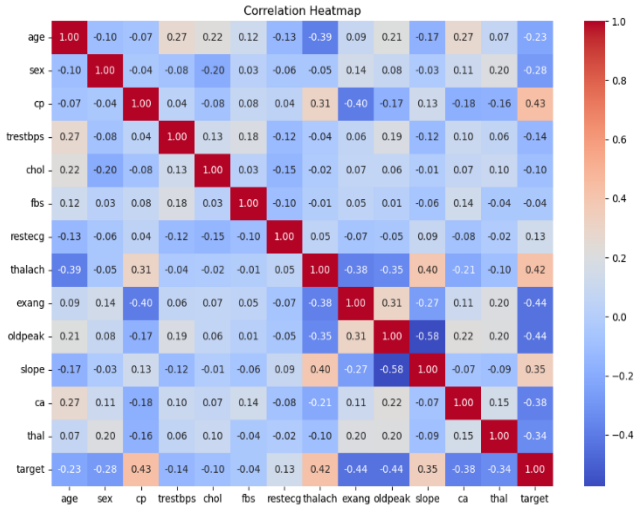


Fig.2. a. Correlation Matrix of Heart Disease Dataset

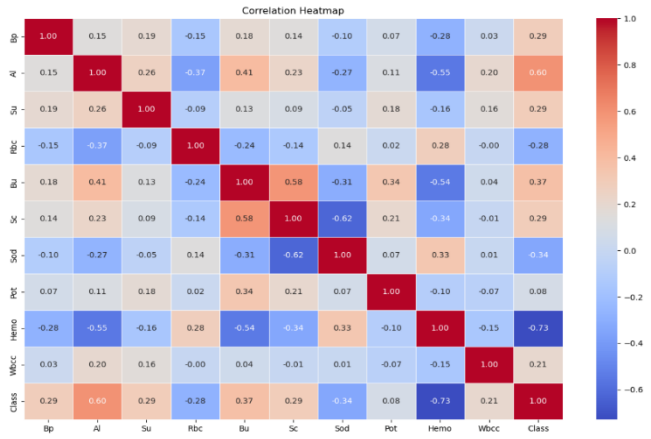


Fig.2. b. Correlation Matrix of Kidney Disease Dataset

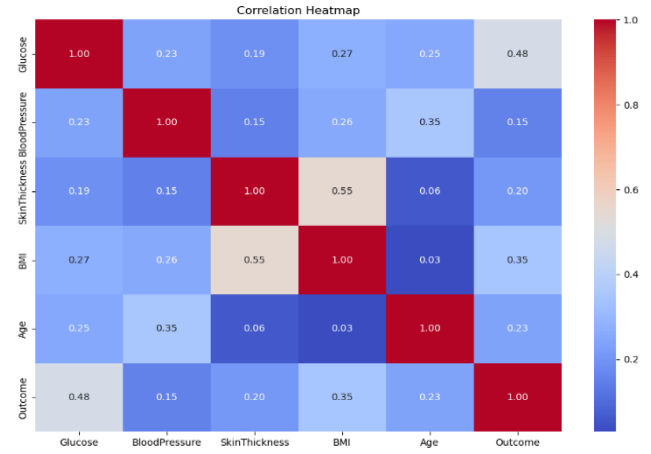


Fig.2. c. Correlation Matrix of Diabetes Dataset

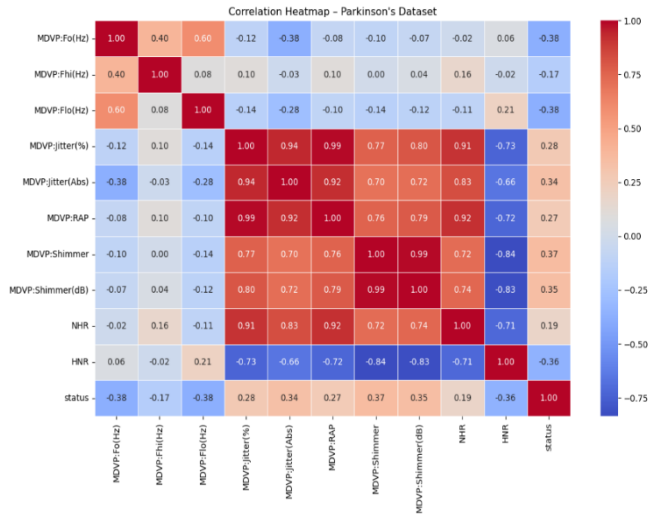


Fig.2. d. Correlation Matrix of Parkinson's Disease Dataset

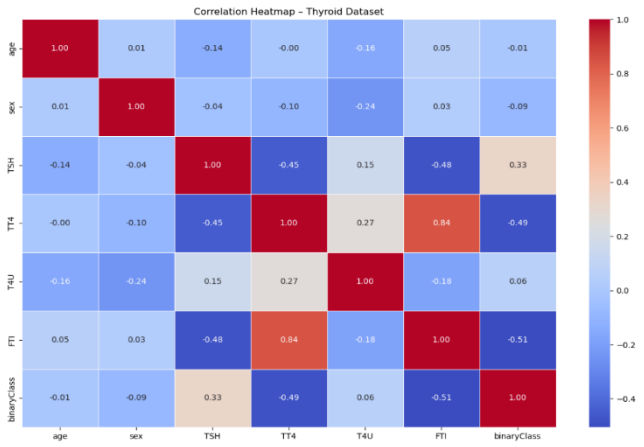


Fig.2. e. Correlation Matrix of Thyroid Disease Dataset

C. Data Pre-Processing

Before model development, each disease-specific dataset undergoes thorough a preprocessing stage to the enhanced the learning and data quality efficiency. Missing and inconsistent values are handled using statistical imputation, followed by feature normalization to scale numerical attributes into a common range using min–max normalization, expressed as

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

Where x stands for the initial feature value, and $x_{\min}$ and $x_{\max}$ denote the feature's lowest and highest values, respectively. Subsequently, feature extraction and selection are carried out using correlation matrix analysis, in which Pearson's correlation coefficient is computed as

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (2)$$

Where x_i and y_i represent individual feature samples and corresponding class labels, and $\bar{x}$ and $\bar{y}$ denote their mean values. Features exhibiting strong correlation with the disease outcome are retained, while redundant highly correlated attributes are minimized to reduce multicollinearity, resulting in optimized feature sets that improve classification accuracy and generalization across multiple disease prediction models.

D. Classification Model

Supervised machine learning techniques are employed to this system to develop reliable classification models for multiple disease prediction. Among various classifiers, the use of logistic regression and support vector machines (SVM) are selected because of the way they work in handling medical datasets, robustness to high-dimensional feature spaces, and strong generalization capability. These models are trained independently for each disease to capture disease-specific characteristics and patterns present in clinical data. Flow chat for training the models is given in figure 2

a) Support Vector Machine (SVM): Often utilized for classification applications, SVM is a potent supervised learning technique in medical decision-support systems due to its strong generalization ability and effectiveness in high-dimensional feature spaces. The primary objective of SVM is to construct an ideal choice boundary that optimizes the margin with different classes, thereby improving classification robustness.

Given a training dataset

$$\{(x_i, y_i)\}_{i=1}^{N}, x_i \in \mathbb{R}^d, y_i \in \{-1, +1\} \quad (3)$$

SVM aims to determine a separating hyperplane defined by a weight vector $w$and a bias term $b$such that

$$w \cdot x + b = 0 \quad (4)$$

To achieve maximum margin separation while allowing limited misclassification, SVM formulates the learning process as a convex optimization problem expressed as

$$\min_{w,b,\xi} \ \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \xi_i \quad (5)$$

Subject to the constraints

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \quad (6)$$

Here, ξ_i denotes slack variables that permit certain samples to violate the margin constraints, enabling the model to handle noisy and overlapping data. The parameter for regularization Ccontrols the compromise between increasing the margin and decreasing classification mistake. To address non-linear separability in medical datasets, kernel functions are employed to map the input feature space into a space with more dimensions where linear separation turns into feasible. This makes SVM particularly effective for complex disease prediction tasks involving multiple physiological parameters.

b) Logistic Regression: One of the popular supervised learning approach for binary classification is logistic regression problems, particularly in medical diagnosis due to its simplicity, interpretability, and probabilistic output. Unlike linear regression, Logistic Regression models the probability of class membership using a non-linear activation function, making it suitable for disease prediction tasks where outcomes are categorical.

Given an input feature vector

$$x = [x_1, x_2, \ldots, x_d] \in \mathbb{R}^d \quad (7)$$

Logistic regression calculates the likelihood that positive class by using the logistic function (sigmoid) defined as

$$P(y = 1 \mid x) = \sigma(z) = \frac{1}{1 + e^{-z}}, \text{where} z = w \cdot x + b \quad (8)$$

Here, the weight vector is represented by w, and the bias term is denoted by b. The predicted probability is mapped to a class label by using a threshold for decisions, typically set to the 0.5. These parameters are learned by minimizing the logistic loss (cross-entropy loss) function given by

$$J(w,b) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \quad (9)$$

Where $\hat{y}_i$ is predicted probability for the i^{th} sample.

Logistic Regression provides interpretable coefficients that indicate the influence of each feature on disease prediction, which is particularly valuable in healthcare applications. Its computational efficiency and robustness make it well-suited for integration into multi-disease prediction systems, where reliable and transparent decision-making is essential.

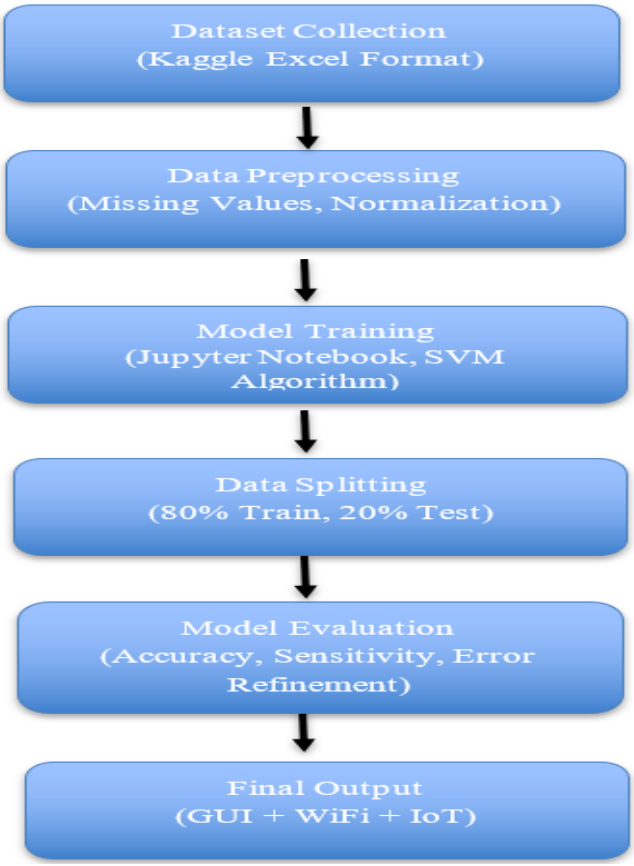


Fig.3. Flow Chart for Training the Model

E. Edge AI Implementation on ESP32

The proposed ML models are deployed on the ESP32 microcontroller to enable real-time, on-device disease prediction using Edge AI principles. The trained ML models, including Logistic Regression and Support Vector Machine (SVM), are optimized using ONNX runtime and embedded into the microcontroller memory to perform inference locally without relying on cloud infrastructure.

Input data is provided through a mobile-based IoT interface and transmitted through Wi-Fi to the ESP32. The input parameters, the device processes the data and executes the prediction algorithm in real time. The classification results are then displayed on an LCD module connected to the microcontroller. The implementation is designed to operate within the resource constraints of the ESP32, with efficient memory utilization and low computational overhead. This makes the proposed solution particularly effective for deployment in remote and resource-limited environments, where access to centralized medical infrastructure is limited.

F. Evaluation Metrices

To evaluate the effectiveness of the proposed multiple disease prediction system, standard classification Assessment metrics are used. These measurements offer a thorough comprehension of the model effectiveness by measuring predictive accuracy, class-wise performance, and error distribution. The evaluation is true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) according to the confusion matrix.

Accuracy symbolizes the overall correctness of the classification model and is defined as the proportion of accurate samples to the total number of samples:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

Although Accuracy offers a broad performance indicator, it may be confusing when there is a difference in the class.

The accuracy measures the proportion of correctly predicted samples that are positive among all the positive samples and shows how reliable positive forecasts are:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (11)$$

Low false-positive rates are indicative of high precision, which is essential for medical diagnosis to avoid incorrect disease predictions.

Recall, sometimes referred to as sensitivity, quantifies the percentage of true positive samples that the models are accurately identifies:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (12)$$

A high recall value ensures that most diseased cases are detected, which is essential in healthcare applications where missing a diagnosis can have serious consequences.

The F1-score is the precision and recall harmonic mean, providing an assessment of classification performance:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}} \quad (13)$$

When working with imbalanced datasets, this statistic is quite helpful.

The confusion matrix offers the analysis of classification outcomes by the compared labels with the correct labels. This will provide insight into the type of mistakes the model made and serves as the basis for computing all evaluation metrics.

III. Discussion of The Results

This section displays the experimental findings derived from the suggested multiple disease prediction system and offers a detailed discussions of the performance by the

created classification models. The system's effectiveness is evaluated and standard performance indicators across datasets corresponding to heart disease, kidney disease, thyroid disorders, diabetes, and Parkinson's disease. Confusion matrix of all the models is given in Table 1 and Figure 4 describes a classification report off the suggested framework; A GUI of the described system is given in the Figure 5. Hardware module is given in figure 6 and Mobile GUI is given in figure 7.

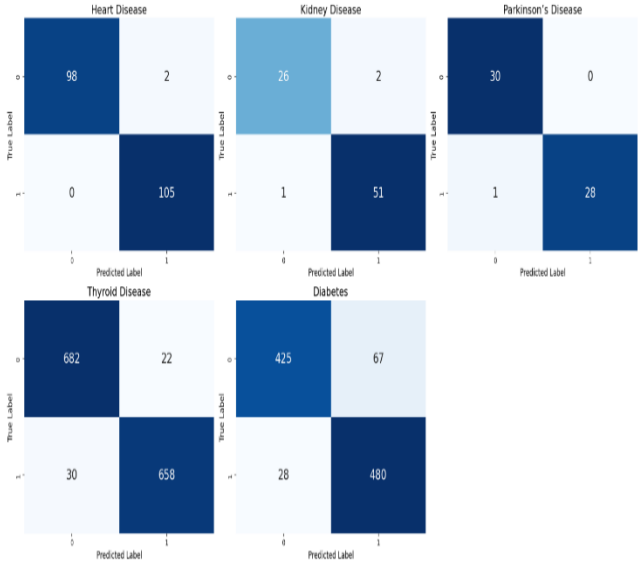


Fig.4. Confusion Matrices of Classification Models

TABLE.1. CLASSIFICATION REPORT OF THE PROPOSED PREDICTION MODELS

S.no	*Disease*	*Precision*	*Recall*	*F1 Score*	*Support*	*Accuracy*
1.	Diabetes	0.91	0.91	0.90	1000	0.90
2.	Thyroid	0.96	0.96	0.96	1392	0.96
3.	Parkinson	0.98	0.98	0.98	59	0.98
4.	Kidney	0.96	0.96	0.96	80	0.96
5.	Heart	0.99	0.99	0.99	205	0.99

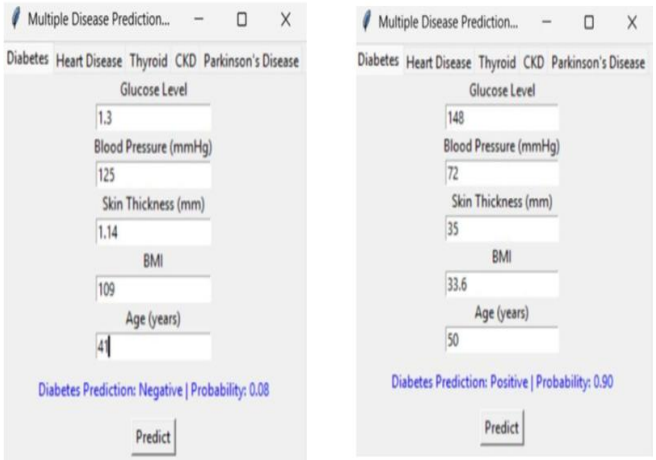


Fig.5. GUI of the Proposed System

Fig. 6 Hardware Module

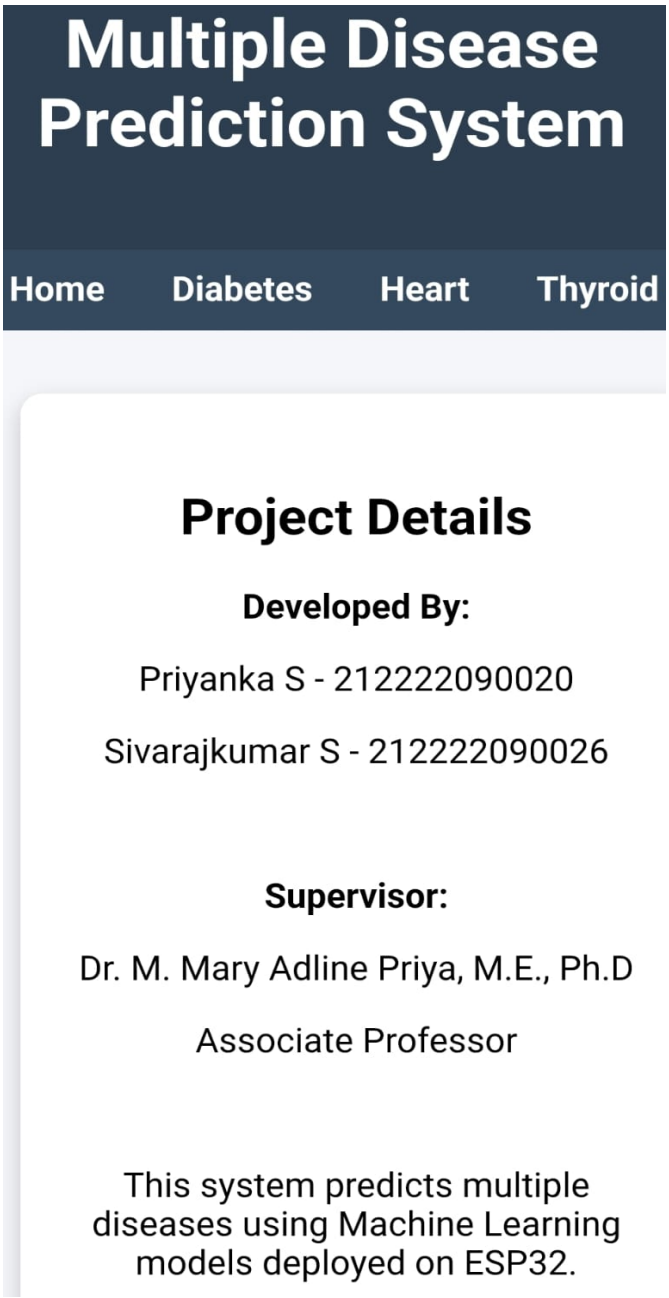


Figure 7. Mobile GUI

IV. Conclusion

The multiple disease prediction system was developed by combining machine learning models for heart, kidney, thyroid, diabetes, and Parkinson's diseases into a single platform. The models achieved high accuracy across all datasets, showing their effectiveness for early diagnosis. To make the system practical, the trained models were converted into ONNX format and deployed on the ESP32 microcontroller, enabling real-time predictions using an Edge AI approach. The system takes input through a mobile interface and displays results instantly on an LCD, reducing dependency on cloud services and improving data privacy. The proposed system shows that low-cost, portable devices can support real-time healthcare solutions. In the future, the system can be expanded with more diseases and improved using larger datasets and advanced models.

References

[1] A. S. Zamani, A. H. A. Hashim, A. S. A. Shatat, M. M. Akhtar, M. Rizwanullah, and S. S. I. Mohamed, "Implementation of machine learning techniques with big data and IoT to create effective prediction models for health informatics," *Biomedical Signal Processing and Control,* vol. 94, Art. no. 106247, 2024, ISSN: 1746-8094.

[2] A. Singh, A. Yadav, S. Shah, and R. Nagpure, "Multiple disease prediction system," *International Research Journal of Engineering and Technology (IRJET),* vol. 9, no. 3, pp. 1697, Mar. 2022. [Online], e-ISSN: 2395-0056, p-ISSN: 2395-0072.

[3] M. S. Arif, A. U. Rehman, and D. Asif, "Explainable machine learning model for chronic kidney disease prediction," Algorithms, vol. 17, no. 10, Art. no. 443, Oct. 2024.

[4] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, S. Saha, R. Hossen, S. Ahmed, M. A. T. Rony, and M. F. Akter, "ML-CKDP: Machine learning-based chronic kidney disease prediction with smart web applications," *Journal of Pathology Informatics,* vol. 15, Art. no. 100371,2024.

[5] H. Zhu, S. Qiao, D. Zhao, K. Wang, B. Wang, Y. Niu, S. Shang, Z. Dong, W. Zhang, Y. Zheng, and X. Chen, "Machine learning model for cardiovascular disease prediction in patients with chronic kidney disease," *Frontiers in Endocrinology,* vol. 15, Art. no. 1390729, 2024.

[6] A. S. Tang, K. P. Rankin, G. Cerono, S. Miramontes, H. Mills, J. Roger, B. Zeng, C. Nelson, K. Soman, S. Woldemariam, Y. Li, A. Lee, R. Bove, M. Glymour, N. Aghaeepour, T. T. Oskotsky, Z. Miller, I. E. Allen, S. J. Sanders, S. Baranzini, and M. Sirota, "Leveraging electronic health records and knowledge networks for Alzheimer's disease prediction and sex-specific biological insights," *Nature Communications,* vol. 15, Art. No. 1390, Feb. 2024.

[7] Halder RK, Uddin MN, Uddin MA, Aryal S, Saha S, Hossen R, Ahmed S, Rony MAT, Akter MF. "ML-CKDP: Machine learning-based chronic kidney disease prediction with a smart web application." *JPathol Inform*. 2024 Feb 22;15: 100371.

[8] R. H. Khan, J. Miah, M. Tayaba, and M. M. Rahman, "Comparative study of machine learning algorithms for detecting breast cancer," in Proc. 2023 *IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC),* 2023.

[9] D. Goyal, B. Pratap, S. Gupta, S. Raj, R. R. Agrawal, and I. Kishor, Eds., "Recent Advances in Sciences, Engineering," Information Technology & Management. *Springer,*2023.

[10] Zamani, A. S., Hashim, A. H. A., Shatat, A. S. A., Akhtar, M. M., Rizwanullah, M., and Mohamed, S. S. I., "Implementation of machine learning techniques with big data and IoT to create effective prediction models for health informatics," *Biomedical Signal Processing and Control,* vol. 94, Art. no. 106247, 2024.