classification success rate. The results achieved were marked as 68.04% for the Accuracy Score, 75.34% for the Sensitivity score, and 45.83% for the Specificity score.

Researchers (Ghanshyam Prasad Dubey et al., 2022) [4] in their study aimed at classifying Parkinson's disease utilizing ML algorithms. The analyser proposed that the unconventionality lies in the implementation of a dimensionality reduction technique before applying classification algorithms. This reduction method employs genetic algorithms for feature selection. The features chosen in their model served as input for the classification of Parkinson's. The data used for prediction involved the records of Parkinson's, and employing the ML algorithms with dimensionality reduction significantly improved performance. Especially, the gradient boosting algorithm achieved an impressive accuracy exceeding 97% during classification with the dimensionality reduction technique

hsuan-Ming Feng [29] investigated and concluded Dysphonia to be a common symptom in the early stages, affecting roughly 90% of patients. Dysphonia, often known as hoarseness, is a medical term for conditions characterized by an abnormal voice quality. This might result in breathy, raspy, strained, weak, or grave vocalizations, as well as pitch fluctuations and pauses. As a result, detecting chronic pronunciation or dysphonia in continuous speech can help with the prediction of Parkinson's. The study's dataset consisted of voice signals from 252 participants. In this study, language signal features served as inputs to machine learning algorithms, enhancing the accuracy of Parkinson's disease (PD) classification. The experiments showcased a diagnostic accuracy reaching 95% through these algorithms. Furthermore, a clinical experience-based feature extraction method was introduced for analyzing the language signals of the subjects.

This study by Researcher (Arunraj Gopalsamy at al.) [5] introduces a novel prognostic ensemble tree method, a machine learning-based classifier that incorporates a naïve Bayesian classifier, decision tree, and logit regression to predict Parkinson's using samples of audio dataset. To evaluate this model, an extensive experimental analysis was conducted across three publicly available Parkinson's datasets. The outcomes demonstrate a remarkable accuracy score of 98.01%, precision reaching 98.6%, and an impressive AUC score of 97.6% employing Ten-Fold cross-validation. Comparative analysis against diverse models highlights that the suggested model has improved performance across several assessment measures. Moreover, when compared to various existing models, this approach showcases minimal MSE, RMSE, and MAE, reaffirming its superiority in predictive capability.

Researchers (Tarigoppula V.S Sriram1 et al.) [26] utilised Orange v2.0b and weka v3.4.10 in their experimentation for the statistical analysis, classification, Evaluation and unsupervised learning methods claiming the diagnosis of the Parkinson disease through machine learning approach to be providing better understanding from the PD dataset in the present decade. The voice dataset for Parkinson's disease used for experimentation has been retrieved from the UCI Machine Learning repository from the Center for Machine Learning and Intelligent Systems. The dataset contains the following attributes name, MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz), along with MDVP:Jitter(%), MDVP: Jitter(Abs), MDVP:RAP, MDVP:PPQ, as well as Jitter:DDP, MDVP:Shimmer, MDVP:Shimmer(dB), Shim- mer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA, NHR, HNR, status, RPDE, DFA, spread1, spread2, D2, and, PPE. The parallel coordinates shown higher variation in Parkinson's disease dataset. SVM proved to provide good accuracy (88.9%) compared to Majority and k-NN algorithms. Classification algorithm like Random Forest also shown good accuracy (90.26) but the Naïve Bayes produced the lowest accuracy (69.23). A higher number of clusters in a healthy dataset in Fo and a smaller number in diseased data has been predicted by hierarchical clustering and SOM. The Parallel Coordinates as well as Sieve graph generated by the experiment has been highlighted in this paper.

B. Diabetes Prediction

Researchers Muhammad Zarar and Yulin Wang in their paper [30] aim to predict diabetes using machine learning by working on two datasets, one from PIMA India and the other from the Kaggle diabetes dataset. The study aimed to assess the efficacy of these classifiers with the Diabetes Dataset, which encompassed attributes with diverse value ranges. They have incorporated four different ML algorithms SVM, Decision Forest, Linear Regression, and Artificial Neural Network. Out of all these ANN resulted best scoring 98% accuracy score.

In their paper, researchers (Sumaia Rahman et al.) [21] worked upon the PIMA Indian dataset which includes medical predictor variables as well as lifestyle factors. Five ML algorithms— support vector machine, random forest, logistic regression, decision tree, and K-nearest neighbours were rigorously evaluated for predictive usefulness using criteria such as accuracy, sensitivity, and specificity. RF trowned the rest algorithms with a staggering 98% accuracy. This study highlighted the transformative potential of machine learning in illness management by providing a data- driven method for identifying at-risk patients and implementing preventative actions.

Asst. Prof. Moumita Dey et al. in the paper Binary Classification of Diabetes in Pima Indian Dataset: A Deep Learning Perspective employed the Keras with Theano as the backend, and established a binary classification algorithm which can effectively forecast the existence or absence of diabetes amongst the individuals. Their analysis uses Keras, a high-level neural network API, in combination with Theano to perform binary classification on the

Pima Indian diabetes dataset. Their work sheds light on the subject of medical data analysis, demonstrating the efficacy of deep learning approaches in developing diagnostic tools for proactive healthcare management. Diabetes, a common chronic illness worldwide, demands the creation of a system for early type 2 diabetic mellitus (T2DM) detection. In this paper, they present an in-depth review of Diabetic Retinopathy, covering its features, causes, various ML models, DL models, challenges, comparisons, and future directions for early DR detection.

Jawad Benabderrahmane et al. [3] intends to increase the results of diabetes mellitus prediction using a range of machine-learning techniques. The Pima Indians Diabetes dataset was used in this case. Several machine learning approaches were utilized, either alone or in conjunction with ensemble learning. The Support Vector Machine scored a prediction accuracy of 77.89%. Furthermore, a real evaluation of ensemble learning techniques employing the voting classifier and Bagging is carried out alongside new approaches and traditional classifiers. The ensemble approach is effective, with the voting classifier achieving a score of 90.11%. Notably, the Bagging Classifier emerges as the frontrunner, with an excellent accuracy rate of 96.69%.

Researchers (V. Jithendra et al.) [7] in their research worked upon the Kaggle Dataset. Seven algorithms were evaluated to determine which was best. The algorithms included Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbor, Support Vector Machine, Naïve Bayes, and Gradient Boosting. The assessment findings showed that Logistic Regression outperformed other algorithms for the provided data set, with an accuracy of 82%. After choosing the more accurate ML model, a user interface where users may add fresh data and retrieve results was constructed.

Andre Sun and Faneng Sun utilized six learning algorithms, LR, SVM (Support Vector Machine), k-nearest neighbors (k-NN), decision tree together with random forest as well as gradient boosting on the Pima Indians Diabetes Dataset. Each model's effectiveness in predicting diabetes in validation datasets was assessed using multiple evaluation metrics like accuracy, recall, and F1-score. Gradient boosting achieved an accuracy of 81.8%, beating all other classifi- cation algorithms on the majority of performance criteria. The gradient boosting model seemed to give an acceptable strategy for diabetes prediction with high accuracy based on the diagnostic measures acquired in this specific dataset. Yogendra Singh [24] and colleagues developed a new machine learning method that combines adaptive iterative imputation (AII) for missing value imputation, dynamic ensemble isolation forest (DE-IF) for outlier detection and removal, Iterated KMeans SMOTEENN (IKMSENN) for class imbalance, and an adaptive extra tree classifier (AETC) for classification. On the Pima Indian Diabetes Dataset (PIDD), their technique has a precision of 0.986, recall of 0.987, and ROC of 0.965, resulting in an accuracy of 98.58.

Clement Okolo et al. [16] through their paper present a pivotal effort to assist medical professionals in the detection and efficient diagnosis of Type 2 diabetes. For this, they worked upon several supervised machine-learning techniques to develop a machine model to predict diabetes with a low error rate. They attempted to create a prediction framework capable of accurately projecting diabetes onset by working upon the eight major indicators collected from the dataset, such as glucose levels, insulin resistance, and body mass index. The researchers tested the effectiveness of four popular machine learning algorithms: logistic regression, support vector machine (SVM), decision tree, and random forest. SVM performed best with 77.27% accuracy, 75.61% precision, 51.38% recall, and 82.47% roc-auc.

C. Breast Cancer Prediction

Through their research Maryam Poornajaf et al. [19] aims to shed light upon the utility of machine learning and deep learning algorithms for early detection of breast cancer among women. Using machine learning models performed on multidimensional datasets, their article aims to find the most efficient and accurate machine learning models for tumour classification as well as prediction. Many supervised machine learning algorithms were used to diagnose and forecast cancer tumours, including Logistic Regression Decision Tree, Random Forest, and KNN. The techniques were used on a dataset of 699 samples obtained from the UCI repository. The features of the dataset were examined, the feature significance score was calculated, and cross-validation was used to improve the algorithms' performance. As a result of the examination, the Logistic Regression method with an accuracy value of 99.14%, an AUC ROC value of 99.6%, and the Extra Tree algorithm with an accuracy value of 99.14% and an AUC ROC value of 99.1% outperformed the other algorithms. Therefore, these methodologies can be beneficial for diagnosing and predicting cancer tumours, as well as appropriately prescribing them.

Aruna Kumar Kavuru et al. [11] explored the role of machine learning models in predicting breast cancer in its early stages, when pain is not a frequent symptom. Their research focused on combining intelligent learning models with nature-inspired optimization techniques to increase prediction accuracy. They combined four machine learning models (KNN, NB, SVM, and ANN) with three adaptive optimization methodologies (APSO, AGA, and AVFO), yielding twelve prediction models. These models were tested using breast cancer datasets (WBC, WDBC, and WPBC) from the UCI data repository. The models were assessed using performance metrics such as accuracy, F-measure, and G-mean. The study discovered that AVFO-SVM, AVFO-ANN, and APSO-ANN outperformed other models on the WBC, WPBC, and WDBC datasets, respectively. Overall, the study found that ANN and SVM machine learning algorithms paired with AVFO for feature selection performed well across all three datasets.

In their work Thilaka et al. stated Breast Cancer is the most prevalent and significant causes for malignancies in women. The paper claims that the greatest method for managing breast cancer symptoms is early identification. In this research, the Random Forest and SVC algorithms were used and their performance was compared. The dataset was obtained from the UCI repository. Furthermore, the classifiers' performance was analyzed and compared in terms of accuracy, precision, and f1-Score. SVC model outscored the other by achieving 93% accuracy.

According to the World Health Organization's (WHO) 2020 report, 2.3 million new instances of breast cancer were reported, and 685,000 women died as a result of the disease. This paper by (Kapil Dev Mahato et al.)[12] highlighted the use of 13 supervised machine learning (SML) techniques, namely: Decision Tree (DT), Logistic REgression (LR), RF, Naive Bayes (NB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Gradient Boosting (GB), Extreme Gradient Boosting (XGB), Adaptive Boosting (AB), Categorical Boosting (CB), Light Gradient Boosting Machine (LGBM), Multi-Layer Perceptron (MLP), and Extra Trees (ET) to predict the outcomes of the Wisconsin Breast Cancer Origina MLP outperformed all thirteen algorithms with a score of 98.76%. This accuracy rate is 0.56% higher than the MLP classifier's previously reported accuracy of 98.2% on the same dataset.

Researchers (Sundarambal Balaraman et al.) [2] in their study worked upon determining the efficiency of machine learning techniques in the detection research of breast cancer. They applied machine learning models on the UCI machine Wisconsin breast cancer dataset. The dataset was analyzed, and the revamped dataset was constructed by eliminating redundant features and appending new features essential for the prediction. Logistic regression, K closest neighbors (KNN), support vector machine (SVM), decision trees, random forest, and XGBoost are examples of machine learning algorithms. Standard for calculating the accuracy rate. In the trial, these classifications were found to function for breast cancer with greater than 97% accuracy. Logistic regression, XGBoost, and Adaboost had the highest accuracy at 99.28 percent.

In recent years, numerous academics have developed several machine-learning methods for effectively detecting breast cancer. In this study, researchers (Nitish Biswas et al.) [14] used the Wisconsin Breast Cancer Dataset (WBCD) as a training set from the UCI machine learning repository to investigate the performance of various machine learning algorithms. Several machines learning classifiers, including SVM, RF, K-NN, DT, NB, LR, AdaBoost, Gradient Boosting, MLP, NCC, and VC, have been used to differentiate between benign and malignant breast cancer tumours. These classifiers use a variety of mathematical and statistical methodologies to find patterns and correlations in the data, allowing for precise tumour categorization. Several assessment criteria, such as error rate, accuracy, precision, F1-score, and recall, were used to analyze the model's performance. These measures shed light on several facets of the model's predictive capacity, including its ability to properly categorize occurrences of malignant and benign tumours, as well as its ability to reduce false positives and false negatives. The accuracy of each approach was evaluated to determine the best match. The investigation shows that the Voting classifier has the best accuracy (98.77%) and the lowest error rate. This shows that the Voting classifier is very good at differentiating between benign and malignant breast cancer tumours, making it a potential technique for breast cancer prediction tasks.

D. Heart Disease Prediction

Researchers Shilpa Sharma et al. [23] underlined the significance of early identification and prevention in lowering mortality from heart disease. They advocated utilizing PCA to reduce the number of tests necessary to predict cardiac disease. The study assessed the performance of several machine learning algorithms, such as SVM, NB, NN, DT, RF, LR, and XGBoost, using two datasets: the Framingham and UCI datasets. Before employing these classifiers, the dataset was dimensionally reduced using PCA. On the UCI dataset, their model obtained 98% accuracy, 98% precision, 98% recall, and a 98% f1 score using Random Forest and XGBoost. On the Framingham dataset, the model achieved 85% accuracy, 85% recall, 80% f1 score, and 79% precision with XGBoost, exceeding other recommended models in predicting coronary heart disease.

Researchers Agung Muliawan et al. [15] investigated cardiac disease prediction with ensemble classifiers and parameter tuning. They used PSO for feature selection and PCA for feature extraction to minimize the dimensions of the UCI heart disease dataset's 13 variables. To obtain the best accuracy, parameter optimization was applied to several machine-learning classifiers such as SVM (Radial Basis Function), Deep Learning, and Ensemble Classifier (Bagging and Boosting). The adjusted deep learning parameters produced a maximum accuracy of 84.47%, while the SVM RBF parameters obtained 83.56%. Bagging SVM with the ensemble classifier resulted in the best accuracy of 83.51%, a 0.5% improvement over SVM without bagging.

In their study, researchers (CMM. Mansoor et al. [13]) worked on the development of ML model for cardio- vascular disease (CVD) forecast based on correlated problems. They employed multiple DL approaches to compare the findings and analyze the UCI Machine Learning Heart Disease dataset. They focused on three types of deep learning models: convolutional neural network (CNN), artificial neural network (ANN), and long short-term memory (LSTM). When compared to other DL approaches, the results revealed that LSTM achieves higher prediction accuracy in less time. This LSTM approach resulted in 91% accuracy. The models were fitted to the test dataset and trained on the training dataset to evaluate which performed the best.

Nandakumar Pandiyan et al. [18] worked upon Euclidean Distance for data pre-processing to clean the unwanted data and employed metaheuristics bio-inspired algorithms such as elephant herding optimization (EHO) for feature selection. Thereafter the data was passed to deep learning models such as convolutional neural network (CNN) and Inception-ResNet-v2 model for the prediction of heart disease from the benchmark dataset such as the UCI Cleveland heart dataset. Finally, the proposed hybrid model used a convolutional neural network with an Inception-ResNet-v2 in the third layer of the architecture to identify heart disease with a promising result of 98.77% accuracy for the Cleveland dataset.

Ahmed Alkurdi [1] aimed to automate heart disease prediction with the Heart Disease UCI dataset. They used a systematic strategy to train machine-learning models for heart disease diagnosis. Data preparation procedures included MEAN missing value imputation, normalization, SMOTE, and correlation analysis. The preprocessed data was then analyzed using four commonly used classification algorithms: Decision Tree, SVM, Random Forest, and k-nearest Neighbors. These models carefully evaluated the dataset and surpassed current models, with up to 100% accuracy.

Through their work researchers (Zahraa Chaffat Oleiwi et al.) [17] suggested an AI model that can help diagnose this disease early. The suggested system focuses on three objectives:

First: Adaptive feature selection with SVM, LR, and RF. The ideal number of features was identified using Mutual Information (MI) and Recursive Feature Elimination (RFE).

Second: Creation of arrhythmia detection models using the MIT-BIH arrhythmia collection of ECG signals. Hybrid models that included machine and deep learning classifiers beat traditional ones. Notably, the OWSK model, which uses a cascade technique, obtained great accuracy.

Third: A binary classification model based on DenseNet121 was developed to detect cardiomegaly using chest X-ray pictures from the CheXpert dataset. The adaptive feature selection technique combining SVM-MI with RF-MI resulted in improved classification accuracy with less features. The OWSK model performed well under several assessment techniques.

The OWSK model used a cascade technique that includes the One-Sided Selection (OSS) method for down-sampling, the Wavelet Transform for feature extraction, SVM, and k-nearest neighbor's algorithms. Furthermore, a binary classification model based on DenseNet121 was created to detect cardiomegaly using the CheXpert dataset of chest X-ray images. The recommended adaptive feature selection technique, which combined SVM-MI with RF-MI, was extremely effective, yielding greater classification accuracy with less features. Under the inter-patient and intra-patient schemes, the OWSK model earned weighted accuracy, recall, precision, and F1 scores of 90%, 90%, 93%, 91%, and 98%, respectively.

III. METHODOLOGY

This study's activities were conducted out on a laptop with an i5 CPU, utilizing Python in Google Colab with Spyder settings. Our work made use of a range of ML algorithms, including but not limited to Decision Tree, Logistic Regression, Neural Networks, Support Vector Classifiers, as well as Random Forest. These algorithms were chosen based on their adaptability of handling various types of data and their track record in predictive modelling jobs. In addition, we used preprocessing techniques like feature scaling and dimensionality reduction to improvise the estimation ability of our models.

A ten-fold cross-validation technique was utilized to provide robustness in accuracy evaluation. This approach entails partitioning the dataset into 10 equal parts, training the algorithm on nine of them, and testing it on the residual part, then repeating the procedure ten times using various partitions. The average performance over all iterations is a more trustworthy indicator of the model's accuracy and generalizability.

This section explains the methodology employed in our proposed study. We want to design a machine learning-powered application that can identify diseases including breast cancer, diabetes, Parkinson's, and heart disease. Figure 1 depicts a flowchart of the workflow, highlighting the many phases involved in application development and deployment. Our proposed study employs the following machine learning methodologies:

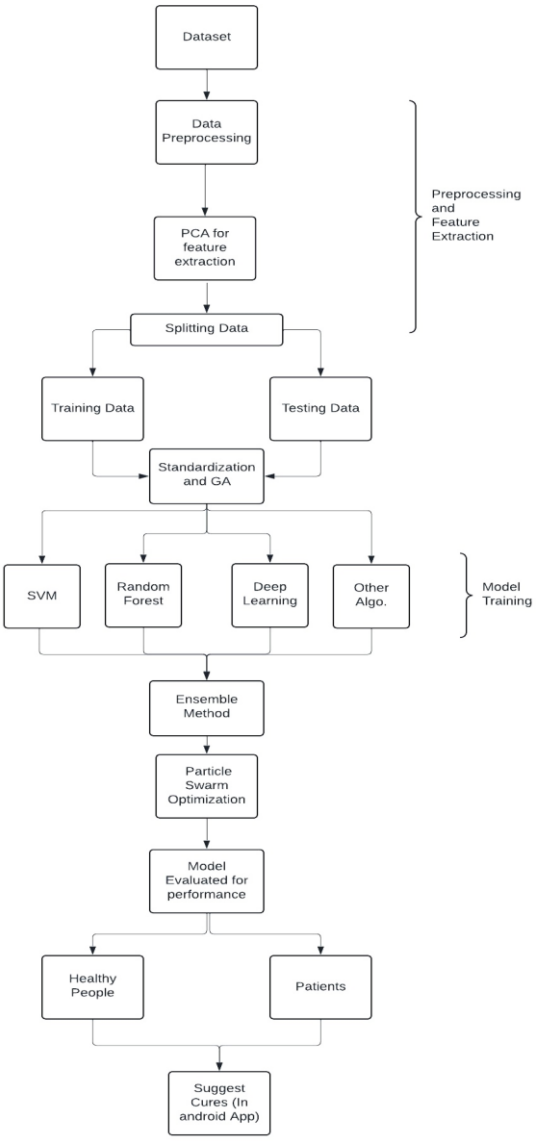

Figure 1. Flowchart of the process followed in general for the experimentation

A. Datasets Used

1) Heart Disease Dataset: We utilized UCI's "Heart Disease Dataset"[6] to forecast the occurrence of heart illnesses. This dataset neholds 13 medical predictor characteristics and 1 result feature. The qualities are as follows: chol, cp, trestbps, age, fbs, sex, respecg, exang, slope, thal, ca, oldpeak, and thalach. The collection consists of 303 instances and 75 characteristics.
2) Diabetes Dataset: To predict the incidence of diabetic diseases, we used Kaggle's "Pima Indians Diabetes Dataset" [25]. This dataset contains eight medically significant prediction factors and one outcome feature: blood pressure, pregnant status, glucose levels, skin thickness, BMI, insulin, age, and diabetes history.
3) Breast Cancer Dataset: We utilized Sklearn's Dataset for Breast Cancer[27], which is accessible for download from here. It has 31 medical features, such as the radii, texture, overall perimeter, total area, smoothness as well as compactness, the concavity, and concave points.
4) Parkinson's Dataset: We utilized UCI's "Oxford Parkinson's Disease Detection Dataset"[9] to augur the occurrence of Parkinson's illnesses. This dataset consists of the biological vocal measurements of thirty-one people, twenty-three of whom have the disease. Individual columns in table indicate a distinct voice calculation, additionally, each row relates to one of the 195 audio recordings from the people ("name" column). The preliminary intent of the data is to distinguish between healthy and Parkinson's disease patients using the "status" column, which results 0 for healthy and 1 for PD.

B. Feature Extraction - PCA

1) Principal Component Analysis:

It is a statistical approach that employs an orthogonal transformation to turn correlated variables into uncorrelated variables. PCA is the most often used method for exploratory data analysis and machine learning prediction models. Furthermore, Principal Component Analysis (PCA) is an unsupervised learning approach that analyzes the relationships between variables. It is also known as a generic factor analysis, which uses regression to find the optimum line of fit.
The primary purpose of Principal Component Analysis (PCA) is to reduce a dataset's dimensionality while keeping the most essential patterns or correlations between variables, regardless of prior knowledge of the target variables. The primary components are linear combinations of the original variables in the dataset, arranged in decreasing order of importance. The overall variance recorded by all primary components equals the entire variance of the original dataset. The first principal component captures the largest variation in the data, whereas the second principal component collects the maximum variance orthogonal to the first principal component, and so on.
PCA was applied in our work to improve the performance of the model by curating better features to work upon. It operates on the assumption that when data in a higher dimensional space is mapped to data in a lower dimensional space, the variance of the data in the lower dimensional space is maximized.
Following is the algorithm for the same.

Algorithm 1: Principal Component Analysis (PCA)
1: Input: Data matrix X of size $a \times b$, where a is the samples' number and b is the feature quantity.
2: Output: Principal components W.
3: 1. Subtract the mean from each feature: $X \leftarrow X - \text{mean}(X)$.
4: 2. Compute the covariance matrix: $C \leftarrow \frac{1}{n-1} X^T X$
5: 3. Perform eigen decomposition of C: $C = V \Lambda V^T$, where V contains the eigenvectors and Λ is a diagonal matrix of eigenvalues.
6: 4. Select the first k eigenvectors corresponding to the largest k eigenvalues to form the matrix W.
7: 5. Project the data onto the new subspace: $Y = XW$.
8: return W.

C. Algorithmic Models Used

1) Logistic Regression

It is a statistical technique used to compute binary classification issues. It utilizes a logistic function (LF) to fit data and estimate the probability of an event occurring. The LF, often known as the sigmoid function, turns any real-number to a value between 0 and 1 that may be used to compute probability. It is especially useful when the dependent variable is categorical, as it can handle both linear and nonlinear interactions. In classification issues, logistic regression calculates the likelihood that input data belongs to different classes. After setting a threshold (often 0.5), the data points are assigned to the class with the highest likelihood.

i. Binary Classification: In binary classification, logistic regression is used to predict outcomes that have only two potential values, such as Yes/No, True/False, or 0/1.
ii. Multiclass Classification: Logistic regression may also be used to solve multiclass classification issues with strategies such as one-vs-rest or multinomial logistic regression.

The following equation represents the logistic regression model:

$$P(Y = 1/\gamma) = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1\gamma_1 + \alpha_2\gamma_2 + \cdots + \alpha_n\gamma_n)}} \quad (1)$$

here:
- $P(Y = 1|\gamma)$ is the probability of the outcome being 1 (diabetes) given the input features X.
- $\alpha_0, \alpha_1, \alpha_2, ..., \alpha_n$ = the coefficients of the model.
- $\gamma_1, \gamma_2, ..., \gamma_n$ = the input features.
- e = base of the natural logarithm.

The LR model estimates the log-odds of the probability of the event:

$$\log\left(\frac{P(Y=1/\gamma)}{1-P(Y=1/\gamma)}\right) = \alpha_0 + \alpha_1\gamma_1 + \alpha_2\gamma_2 + \cdots + \alpha_n\gamma_n \quad (2)$$

The output is the estimation that the fed data belongs to the positive class (diabetes in this case).

- Hyperparameter Tuning:

Hyperparameter adjustment is critical for improving the performance of the LR models. Some popular hyperparameters to tweak are:

i. Regularization: Overfitting is a potential problem with logistic regression. Regularization approaches such as L_1 (Lasso) and L_2 (Ridge) regularization help reduce overfitting by punishing large coefficients.
ii. Solver: Varying solutions offer varying benefits depending on the size of the dataset and the computing resources available. Common solvers are 'liblinear', 'newton-cg', 'lbfgs','sag', and'saga'.
iii. C-Inverse of Regularization Strength: The parameter 'C' governs the deal between fitting the training data accurately and keeping the model simple. Lower 'C' values enhance regularization strength, perhaps preventing overfitting.
iv. Max Iterations: The most number of iterations that the solver should do. Increasing the number of iterations may increase convergence but may result in longer training times.

Using approaches like as Grid Search or Random Search, we may determine the ideal combination of hyperparameters that optimizes the accuracy and generalization of previously unknown data.

Example:

Let's assume we have two features, γ_1 (blood sugar level) and γ_2 (body mass index), to estimate whether a patient has diabetes or not. The logistic regression model for this example is:

$$P\,(Y = 1/\gamma) = \frac{1}{1+ e^{-(\alpha_0+\alpha_1\gamma_1+\alpha_2\gamma_2)}} \quad (3)$$

In this example, the coefficients α_0, α_1, and α_2 would be estimated from the data for training to fit the algorithm/model.

2) Random Forest

Random Forest is an ensemble learning technique that creates predictions using a collection of decision trees. It incorporates predictions from many trees to improve overall accuracy while decreasing overfitting. Random Forest is a useful approach for categorization jobs. It features creating a forest of DT, with each DT trained on a random subsection of the data as well as the attributes. The resultant forecast gets generated via integration of the projections of all trees.

i. Decision Łrees: The Random Forest's individual trees are decision trees that create predictions based on a sequence of feature judgments.
ii. Ensemble Learning: Random Forest's strength is in its ensemble learning technique. Combining many trees' predictions minimizes variation and increases the model's overall accuracy.

- Hyperparameter Tuning

Tuning hyperparameters of a Random Forest model is critical for reaching peak performance. Some significant hyperparameters to consider are:

i. Number of Łrees (n_estimators): The total no. of trees in the forest. Increasing the no. of trees improves model performance along with unfortunately raising the computational cost.
ii. Maximum Depth (max_depth): The utmost depth of any tree. Deeper trees may learn more complicated patterns in data, but they are more prone to overfitting.
iii. Minimum Samples Split (min_samples_split): The smallest no. of samples necessary to divide a node. Increasing this value can help avoid overfitting.
iv. Minimum Samples Leaf (min_samples_leaf): The minimal no. of samples needed to be at a leaf node. Increasing this option may also avoid overfitting.
v. Maximum Features (max_features): The maximum amount of characteristics to evaluate while determining the optimum split. This option allows you to control the model's unpredictability.

By tweaking these hyperparameters using approaches such as Grid Search or Random Search, we may determine the ideal configuration that maximizes the Random Forest's accuracy and generalization to new data sets.

Example:

RF is an ensemble learning approach which creates several decision trees and then combines their predictions to increase accuracy. The formula for forecasting the outcome with Random Forest is:

Let the variable $X = (X_1, X_2, ..., X_n)$ = the input features, and $T_1, T_2, ..., T_m$ be the individual decision trees in the forest. The prediction $\hat{y}$ for a new input X is calculated as:

$$\hat{y} = \frac{1}{m} \sum_{i=1}^{m} T_i(X) \quad (4)$$

Here, Ti(X) is the prediction of the i^{th} decision tree for input X. Each DT Ti is built using a subsection of the training data and an arbitrary batch of features at each split, ensuring variations amongst the trees.

3) Support Vector Machine

It is an effective supervised learning technique for the task of classification. It works by identifying the appropriate hyperplane for separating data points into multiple classifications. SVM seeks to optimize the margin between the border points called support vectors (the data points that are closest to the decision border) of various classes. SVM is successful for both linearly and nonlinearly separable datasets. It can identify data points by transforming them into an increased-dimension feature space making linear separation possible. The decision boundary is established by the SVs, which are the key locations for finding the best hyperplane.

i. Linear SVM: Linear SVM assumes that the data is linearly separable. The method chooses the appropriate hyperplane to divide the classes while maximizing the margin.
ii. Nonlinear SVM with Kernel Łrick: SVM may employ the kernel approach to translate non-linear datasets into higher-dimensional spaces where linear separation is achievable. Common kernels include the following:

- Polynomial Kernel: The formula for polynomial kernel $K(x,y) = (x^T y+c)^d$, here d refers to the polynomial degree & c, a constant.
- Radial Basis Function (RBF) Kernel: The formula for radial basis function $K(x, y) = \exp(-\gamma\|x - y\|^2)$, where γ is a parameter controlling the kernel's width.
- Sigmoid Kernel: The formula for sigmoid kernel $K(x,y) = \tanh(\alpha x^T y+c)$, which can handle nonlinearity and similarity-based features.

The kernel trick helps SVM learn complex decision boundaries, making it apt for an expansive range of classification tasks.

$$\hat{y} = RF(X) = \frac{1}{\eta} \sum_{i=1}^{\eta} f_i(X) \quad (5)$$

Here:

- $\hat{y}$ = predicted outcome (diabetes or not).
- η = no. of trees in the forest.
- $f_i(X)$ = prediction of the i^{th} decision tree.

Each DT in the RF, in this example, is trained using an arbitrarily chosen subset of data and features. The end result is obtained by averaging the predictions of all trees.

4) Deep Learning using Artificial Neural Networks (ANN)

Motivated from the neural structure of human brain, the artificial neural networks are nothing but basic deep learning models . They are composed of layers of linked nodes, each of which performs a basic calculation. ANNs are commonly employed for classification jobs due to their ability to understand complicated data patterns. A basic artificial neural network (ANN) contains an output layer, single or multiple hidden layers, and an input layer. Each layer consists of nodes (neurons) that process input data.

i. Feedforward Neural Networks: In a feedforward NN, data moves from the first layer i.e. input layer to the second layer i.e. hidden layers, and then to the final output layer. Every node in a layer is linked to all nodes in the next layer, and each of the connections has a weight.

ii. Activation Functions: Activation functions add nonlinearity to the model, enabling it to learn compli- cated mappings between inputs and outputs. Common activation functions are:

- ReLU (Rectified Linear Unit): f (x) = max(0,x), which is effective in handling sparse data and prevents the vanishing gradient problem.
- Sigmoid: $f(X) = \frac{1}{1+e^{-x}}$, which maps values to a range between 0 and 1, suitable for binary classification.
- Softmax: Used in the final output layer for multi-class sorting, it normalizes the output to probabilities.

D. Optimization - PSO

1) Particle Swarm Optimization:

It's a powerful meta-heuristic method for optimization influenced by natural swarm activity, like fish and bird schools. It's an imitation of a simple social system. The initial goal of the PSO algorithm was to graphically represent the elegant but unexpected dance of a bird flock.

In nature, the bird's viewable surroundings are constrained in some way. However, having multiple birds in a swarm features an awareness of the wider surface of a fitness function.

PSO is a computer approach for iteratively improving a proposed solution. The population here is based on a stochastic algorithm. Similar to the Genetic Algorithm it has a random population, and fitness evaluation and based on fitness it updates the population but unlike the former, it lacks crossover, mutation or other genetic operator. PSO was applied in our work to improve the work quality of the model by curating better features to compute upon.

Following is the pseudo code/algorithm of PSO:

Algorithm 2: Particle Swarm Optimization

Input: Population size N, maximum no. of iterations T, cognitive parameter c_1, inertia weight ω, social parameter c_2
Output: Global best solution gbest
1 Initiate particles with random positions and velocities;
2 Initialize pbest for each particle as its current position;
3 Initialize gbest as the best particle in the population;
4 for t = 1 to T do
5 for each particle do
6 Update velocity: $v_i(t+1) = \omega\, v_i(t) + c_1 r_1(pbest_i - x_i) + c_2 r_2(gbest - x_i)$;
7 Update position: $x_i(t+1) = x_i(t) + v_i(t+1)$;
8 Update pbest if necessary;
9 Update gbest if necessary;
10 end
11 end
12 return gbest;

2) Genetic algorithm:

It was used to identify the most relevant characteristics in the dataset. The GA tries to increase classifier performance while lowering data dimensionality by optimizing a subset of features. This stage includes:

- Creating an initial population of possible feature subsets.
- Evaluating each individual's fitness (feature subset) using a classifier (e.g., SVM) and its accuracy.
- Using genetic operators (selection, crossover, and mutation) to develop a population over several generations.
- Choosing the optimal feature subset with the highest fitness score (accuracy).

Algorithm 3: Genetic Algorithm

1: Start with initial population: population ← initialize_population(population_size)
2: Assess fitness: fitness_scores ← evaluate_fitness(population)
3: while not termination_condition do
4: selected_parents ← select_parents(population)
5: offspring ← crossover(selected_parents)
6: offspring ← mutate(offspring)
7: fitness_scores ← evaluate_fitness(offspring)
8: population ← replace_population(population, offspring)
9: end while
10: best_solution ← get_best_solution(population)
11: return best_solution

IV. RESULTS

Datasets were processed using a variety of techniques, including Support Vector Classifier (Radial Basis Function), Deep Learning, Logistic Regression, Ensemble Classifiers, Random Forest, and Decision Tree. Their performances were assessed using metrics such as accuracy, F1-score, precision, and Area under the Receiver Operating Characteristic Curve (ROC).

The outputs were examined using bar graphs for each condition to offer a visual representation of the accuracy. The figures in Figs. 2–6 show a clear juxtaposition of the accuracy achieved by various machine learning algorithms. The datasets for various diseases were classified into training and testing sets, as is customary in classification tasks, to test the model's generalization capabilities.

Figures 2-6 exhibit accuracy score charts designed to offer a full knowledge of each algorithm's performance. These graphs show the accuracy ratings obtained by the Logistic Regression, Random Forest, Neural Network, Decision

Tree, and SVM models for predicting various diseases. This comprehensive study offers an educated evaluation of each algorithm's efficacy in sickness prediction tasks.

Table 1 compares all datasets and algorithms along with highlighting the best accuracy technique. In the table: LR-Logistic Regression, DL-Deep Learning model, DT-Decision Tree, RF-Random Forest

Figures 7–10 demonstrate the output screens of both the existing web portal and the to-be-developed Android application.

TABLE I:
COMPARISON OF PREDICTION ACCURACY FOR VARIOUS DISEASES

Disease	**LR**	**DL**	**DT**	**RF**	**Best Accuracy**
Parkinson's	83	71	84	**99**	99
Heart Disease	84	86	**97**	96	97
Diabetes	**95**	81	90	94	95
Breast Cancer	-	**96**	-	-	96

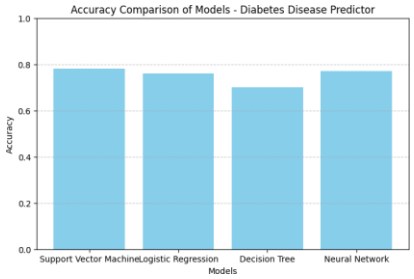

Figure 2. Model accuracy comparison chart for Diabetes Predictor

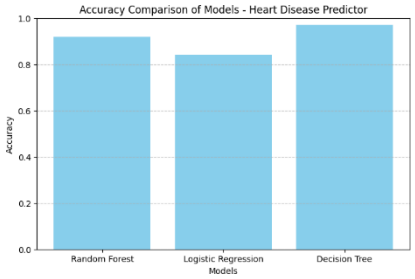

Figure 3. Model accuracy comparison chart for Heart Disease Predictor

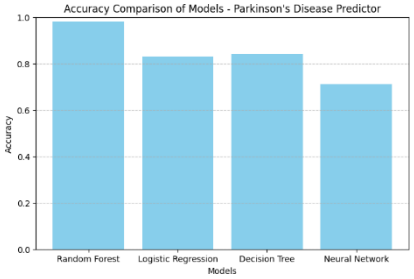

Figure 4. Model accuracy comparison chart for Parkinson's Disease Predictor

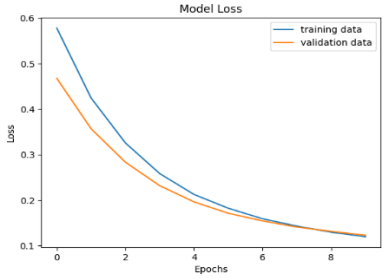

Figure 5. Model loss chart for Breast Cancer Disease Predictor

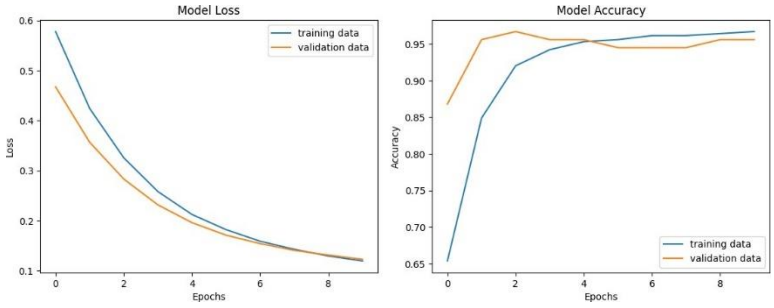

Figure 6. Model accuracy and loss chart for Breast Cancer Disease Predictor

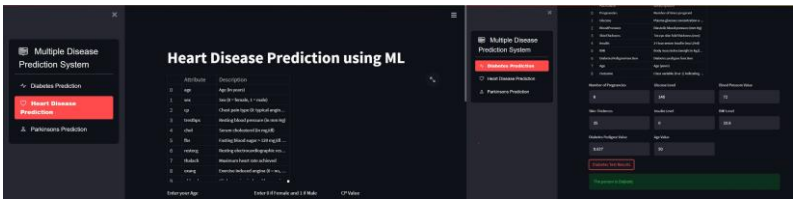

Figure 7. Output web Screen of the multidisease predictor

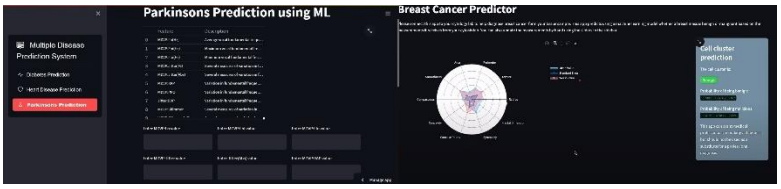

Figure 8. Output web Screen of the multidisease predictor

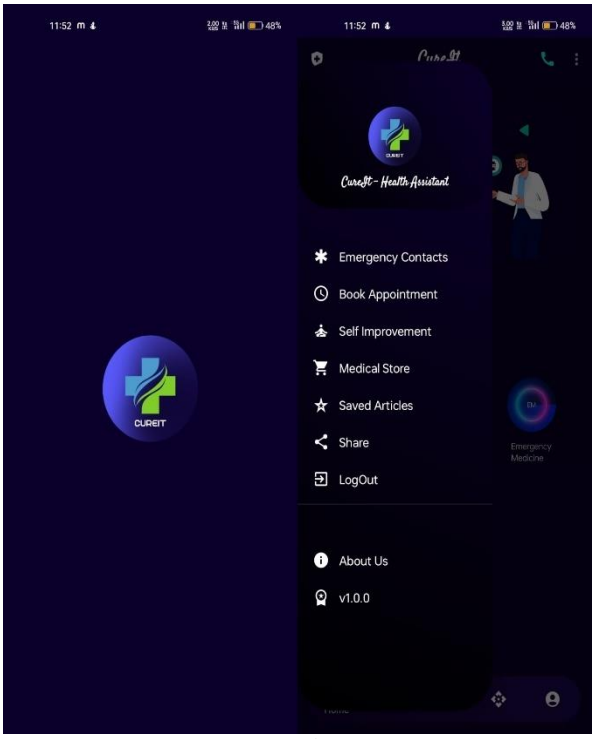

Figure 9. Output web Screen of the multidisease predictor

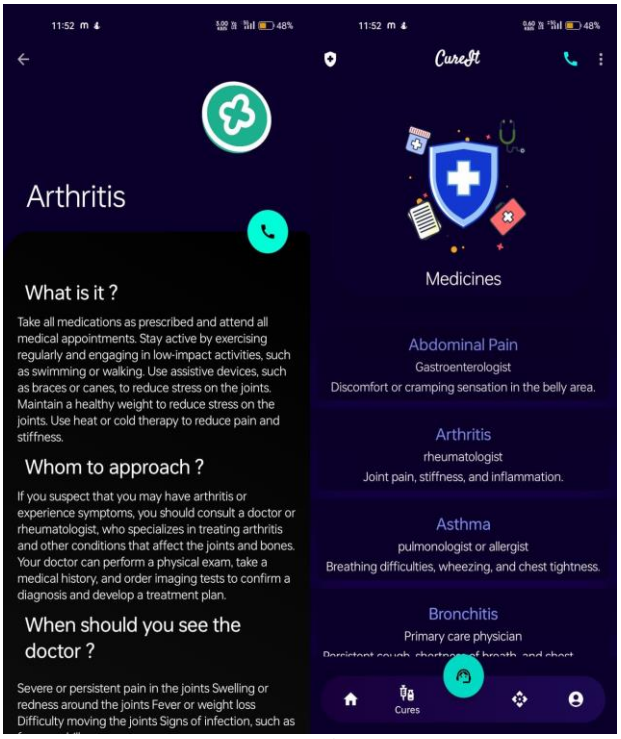

Figure 10. Output web Screen of the multidisease predictor

V. CONCLUSION

Our work, as mentioned in the paper, unifies multiple AI models for predicting diabetes mellitus, Breast malignancy, heart disease, as well as Parkinson's into one single platform. These trained models are then programmed to be deployed as Streamlit website using the lightweight Flask framework. Several classification methods are employed to train the models, in which the random forest gave the best result for Parkinson's prediction reaching up to 98%, Deep Learning provided an accuracy of 96% in predicting breast cancer, SVM worked better to track diabetes and also logistic regression scored above 90% accuracy, and Decision Tree scored 97% accuracy in predicting heart diseases.

VI. FUTURE WORK

We aim to broaden this effort by including numerous other predictable diseases as well as its cures all compiled up into one Android Application - CureIT which would also feature natural home-based remedies for certain diseases that can be either cured or whose nature can be controlled without allopathic medications. Also, the app would have better accuracy results as well as image-based analysis of disease.

ACKNOWLEDGMENT

I want to express my heartfelt gratitude to everyone who helped me complete my project, which aimed to create a website for multi-disease prediction using machine learning. First and foremost, I would like to convey my deepest

thanks to my supervisor, Mrs. Rupal R. Chaudhari, for her essential assistance, encouragement, and support during this project. Their skills and counsel have helped shape the course of our venture. I am also grateful to the academic members of L.D. College of Engineering and Sankalchand Patel College of Engineering for their helpful feedback and support in refining the strategy and results of this work. I thank the creators of the UCI Machine Learning Repository for giving access of the datasets utilized in this work. Their efforts to create and manage these datasets have been critical in furthering machine learning research and innovation. Furthermore, I want to thank my colleagues and friends for their encouragement and talks, which have helped me gain better knowledge and perspective on the topic. Finally, I am grateful to my family for their ongoing support, patience, and understanding during this journey. This project would not have been attainable without the joint efforts and help of the persons listed above. Many thanks.

REFERENCES

[1] Ahmed Alkurdi. Enhancing heart disease diagnosis using machine learning classifiers. Fusion Practice and Appli- cations, 13:8–18, 09 2023.
[2] Sundarambal Balaraman, Ramesh Ramamoorthy, and Raja Krishnamoorthi. Breast cancer detection with re- vamped dataset using machine learning techniques. Journal of Medical Imaging and Health Informatics, 11:2996– 3009, 12 2021.
[3] Jawad Benabderrahmane, Mohammed Kasri, Inssaf Guabassi, Anas El-Ansari, and Abderrahim beni hssane.
[4] Improving Machine Learning Performance for Diabetes Prediction, pages 361–371. 02 2024.
[5] Ghanshyam Dubey, vishal chourasia, and Akhil Phadnis. Detection of parkinson disease through hybridizing machine learning algorithms and genetic algorithm. 11 2022.
[6] Arunraj Gopalsamy and B. Radha. Machine Learning-Based Ensemble Classifier Using Naïve Bayesian Tree with Logit Regression for the Prediction of Parkinson's Disease, pages 451–469. 03 2022.
[7] Steinbrunn William Pfisterer Matthias Janosi, Andras and Robert Detrano. Heart Disease. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C52P4X.
[8] V. Jithendra, M Madhusudhan, B Jagadeesh, and S Kusuma. Diabetes prediction using machine learning tech- niques. Journal of Artificial Intelligence and Capsule Networks, 5:190–206, 06 2023.
[9] R. Kumar and R. Pal. India achieves who recommended doctor population ratio: A call for paradigm shift in public health discourse! Journal of Family Medicine and Primary Care, 7(5):841–844, Sep-Oct 2018.
[10] Max Little. Parkinsons. UCI Machine Learning Repository, 2008. DOI: https://doi.org/10.24432/C59C74.
[11] LiveMint. Heart attacks account for 28 News Article, 2024. Accessed on April 12, 2024.
[12] Tamilselvi Madeswaran, Aruna Kavuru, Padma Theagarajan, Nasser Hadrami, Ohm Rambabu, and Maya Foori. Breast cancer prediction using hybrid machine learning and nature-inspired adaptive optimization algorithms. Journal of Chemical Health Risks, 13, 10 2023.
[13] Kapil Mahato, Chitra Saini, Chandrashekhar Azad, and Uday Kumar. Breast cancer prediction using different machine learning algorithms: A comparative study. 07 2023.
[14] CMM Mansoor, Sarat Chettri, and HMM Naleer. Efficient prediction model for cardiovascular disease using deep learning techniques. Migration Letters, 20:449–459, 12 2023.
[15] Khandaker Mohammad Mohi Uddin, Nitish Biswas, Sarreha Rikta, and Samrat Dey. Machine learning-based di- agnosis of breast cancer utilizing feature optimization technique. Computer Methods and Programs in Biomedicine Update, 3:100098, 02 2023.
[16] Agung Muliawan, Achmad Rizal, and Sugondo Hadiyoso. Heart disease prediction based on physiological pa- rameters using ensemble classifier and parameter optimization. Journal of Applied Engineering and Technological Science (JAETS), 5:258–267, 12 2023.
[17] Clement Okolo. Diabetes Prediction Using Machine Learning Algorithm. PhD thesis, 12 2022.
[18] Zahraa Oleiwi, Ebtesam Alshemmary, and Salam Al-augby. Heart Diseases Diagnosis System Using MultiM ethodsMachineLearning. PhDthesis, 112023.
[19] Nandakumar Pandiyan and R Subhashini. Heart disease prediction using convolutional neural network with elephant herding optimization. Computer Systems Science and Engineering, 48, 12 2023.
[20] Maryam Poornajaf and Sajad Yosefi. Improvement of the performance of machine learning algorithms in predicting breast cancer. Frontiers in Health Informatics, 12:132, 03 2023.
[21] Press Information Bureau, Government of India. Cabinet approves Establishment of World Class 'Residential Schools' (' '). Press Release, 2024. Accessed on April 12, 2024.
[22] Sumaia Rahman and Suraiya jabin. Diabetes prediction using machine learning algorithm. International Journal of Scientific and Engineering Research, 15, 04 2024.
[23] Indira Rustempasic and Mehmet Can. Diagnosis of parkinson's disease using fuzzy c-means clustering and pattern recognition. SOUTHEAST EUROPE JOURNAL OF SOFT COMPUTING, 2, 03 2013.
[24] Shilpa Sharma, Mandeep Kaur, and Savita Gupta. A Comparison of Machine Learning Approaches for Forecasting Heart Disease with PCA Dimensionality Reduction, pages 333–347. 09 2023.
[25] Yogendra Singh and Mahendra Tiwari. Revolutionizing diabetes disease prediction through novel machine learning techniques. Nano, 05 2023.
[26] P. Smith. Pima indians diabetes database. Kaggle, 2022. [Online; accessed April 10, 2024].
[27] Tarigoppula Sriram, M. Rao, G Narayana, T. Vital, and Kaladhar SVGK Dowluru. Intelligent parkinson disease prediction using machine learning algorithms. IJEIT, 3:212–215, 07 2013.
[28] Mangasarian Olvi Street Nick Wolberg, William and W. Street. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository, 1995. DOI: https://doi.org/10.24432/C5DW2B.
[29] World Diabetes Day. World Diabetes Day - Facts and Figures. Webpage, 2024. Accessed on April 12, 2024.
[30] Linlin Yuan, Yao Liu, and hsuan-Ming Feng. Parkinson disease prediction using machine learning-based features from speech signal. Service Oriented Computing and Applications, pages 1–7, 06 2023.
[31] Muhammad Zarar and Yulin Wang. Early stage diabetes prediction by approach using machine learning techniques, 07 2023.