International Scientific Journal of Engineering and Management (ISJEM)
Volume: 04 Issue: 04 | April – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02758

Machine Learning Based Predictive Analysis of Heart Failure and Type 2 Diabetes

Prof. S.V.Phulari[1], Mahek Bhartiya[2], Aditya Dhas[3] , Pavan Gadave[4], Ayush Gawai[5]

[1]Prof.S.V.Phulari Computer Engineering & PDEA's College Of Engineering
[2]Mahek Bhartiya Computer Engineering & PDEA's College Of Engineering
[3]Aditya Dhas Computer Engineering & PDEA's College Of Engineering
[4]Pavan Gadave Computer Engineering & PDEA's College Of Engineering
[5]Ayush Gawai Computer Engineering & PDEA's College Of Engineering

---

Abstract- heart disease and type 2 diabetes are significant global health challenges, necessitating reliable diagnostic methods for effective management. Traditional approaches often fall short, leading to the exploration of machine learning techniques. In this study, we present a machine learning-based system tailored for predicting both conditions, leveraging extensive datasets. Employing machine learning algorithms, alongside feature selection and cross-validation techniques, our system demonstrates robust performance in identifying individuals at risk. Using Machine learning (ML) we have proposed and built the framework for prediction of heart disease and diabetes. A custom ensemble algorithm with hard and soft voting classifier is built for different cardiovascular datasets yielding accuracies in the range of 95.

***Key Words*:** Heart failure, Cardiovascular diseases, Type-II Diabetes, Machine Learning.

1. Introduction

This In 2019, Nearly 18 million people died of cardiovascular diseases, accounting for 32 percent of all world deaths. Many cardiovascular illnesses can be avoided by addressing risk factors such as unhealthy diet, tobacco use, physical inactivity and harmful consumption of alcohol. It is important to detect heart disease early so that management begins with counselling and medication. Cardiovascular diseases are the most life- threatening diseases. They have gotten quite popular over time and have now reached the healthcare systems of several countries. We aim this project for effective prediction of cardiovascular disease that is heart disease and diabetes using a machine learning framework. This will help doctors as well as patients for early prediction of disease so that they get proper treatment. Other researchers and scientists have approached it with different techniques and methods. Cardiovascular illnesses are the deadliest syndromes in the world, with the greatest fatality rate. These diseases have recently grown exceedingly widespread, putting a strain on countries' healthcare systems. For effective and high-precision prediction of heart disease and diabetes, a machine learning-based system is proposed. Cardiovascular Disease Prediction Using Machine Learning The diagnostic is quite accurate and can be used in the real world to detect cardiovascular illnesses early. Using various datasets and approaches, researchers have suggested multiple algorithms for the prediction of cardiovascular illnesses throughout the last decade. Heart disease, Cleveland, Framingham, and cardiovascular disease are some of the most common datasets used for prediction. These datasets are made up of many attributes that are used to forecast these diseases. Family history, age etc are considered under non-modifiable factors because this cannot be changed. Whereas, Smoking habits, unhealthy lifestyle (having fast food, not maintaining quality standard), blood pressure, and cholesterol can be considered under modifiable risk factors these can be changed and controlled by taking certain precautions and medication.

2. Related Work

Heart Rate and CGM Feature Representation Diabetes Detection From Heart Rate : In the paper by "Hootan rashtian, solmaz shariat torbaghan, salar rahili " titled 'Heart Rate and CGM Feature Representation Diabetes Detection From Heart Rate: Learning Joint Features of Heart Rate and Continuous Glucose Monitors Yields Better Representations 'The paper emphasizes the importance of Used feature Extraction and CAA for classification for Heart disease and Diabetes Detection using classification algorithms are more accurate.

1. Designing Disease Prediction Model Using Machine Learning Approach: In this paper the authors Dhiraj Dahiwade, Prof. Gajanan Patle used general disease prediction system based on machine learning algorithm for minimal labeled training data, several data augmentation techniques are applied.
2. Diabetes & Heart Disease Prediction Using Machine Learning: In this the authors Ayman Mir, Sudhir N. Dhage used Support vector machine and random forest algorithms for prediction, Support vector machine and random forest methods are works better on huge datasets.
3. Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare: The

ISSN: 2583-6129
DOI: 10.55041/ISJEM02758

authors Bhavesh Dhandel, Kartik Bamble, Sahil Chavan, Tabassum Maktum proposed the idea of using multi-CNN model is that almost datasets identifiers, namely color, shape and imprint and Three CNNs model are developed based on those three mains identifier and a classification rule to combine the results of CNNs model are applied.

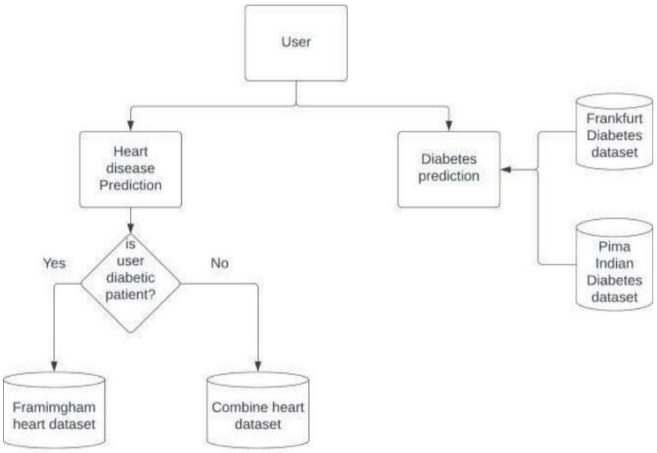
4. Heart Disease Prediction using Hybrid machine Learning Model: In these papers the authors M. Kavitha, G. Gnaneswar, R. Dinesh,Y. R. Sai and R. S. Suraj,proposed study used the Cleveland heart disease dataset, and data mining techniques such as regression and classification are used. Machine learning techniques Random Forest and Decision Tree are applied.

2. Related Work

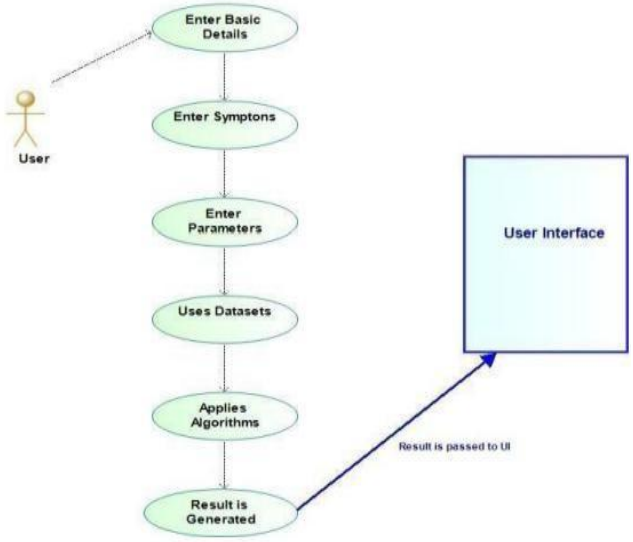
There will be an end-to-end application for early detection of the chosen and stated problem statement. The system will follow classical machine learning tasks like Data Exploration, Data Cleaning, Feature Engineering, Model Building and Model Testing. We will be trying out different machine learning algorithms and selecting the best fit for the above case. The system will have following steps:

- **Data Collection:** Data collection is a process of gathering information, analyzing and restructuring it according to the project requirement. For the proposed hybrid system following datasets are used: – Pima Diabetes dataset – Combine Heart Disease dataset – Frankfurt Diabetes dataset– Framingham Heart Disease dataset
- **Data Preparation:** Data preparation is a process of exploring, structuring, cleaning, enriching and shaping the raw data which is one of the important step in building a ML software. It often involves reformatting data,making corrections to data and the combining of data sets to enrich data.
- **Data Exploration:** Data exploration or visualization is a process of understanding the data by visually representing it in the form of graphs, pie charts, histograms, etc. It helps in providing insights for outliers present in data.
- **Data Mining:** Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs.
- **Model Training:** The process of training an ML model involves providing an ML algorithm (that is, the learning algorithm) with training data to learn from. The term ML model refers to the model artifact that is created by the training process. The training data must contain the correct answer, which is known as a target or target attribute. The learning algorithm finds patterns in the training data that map the input data attributes to the target (the answer that you want to predict), and it outputs an ML model that captures these patterns.
- **Model evaluation metrics:** Evaluation metrics assess the performance of machine learning models by estimating their generalization accuracy on unseen data. Key metrics include Confusion Matrix, accuracy, recall, precision, and sensitivity

- **System Architecture**



- **Use Case Diagram**



- A use case diagram illustrates the potential interactions between users and a system in a

graphical format. It displays different use cases and user types associated with the system and is often complemented by additional diagrams.

3. PROJECT MODULES:

Heart Disease Prediction:

- Data Loading: Acquiring data is a fundamental step in machine learning model creation. This involves extracting, transforming, and loading data from various trusted sources such as Kaggle and the UCI repository. For our project, we gathered data from both Framingham and Combined heart disease datasets.
- Data Validation: Validating the data involves ensuring the correctness and consistency of the parameters. Analyzing trends, collecting data from credible sources, and presenting it for review before model construction is essential.
- Data Preprocessing: Raw data obtained from diverse sources often requires preprocessing to enhance model accuracy. Addressing missing values is a crucial step; techniques like KNN imputation and mode replacement were applied to rectify missing values in the Framingham dataset. Additionally, standard deviation, missing value counts, and exploratory data analysis were conducted.
- KNN imputer: The Framingham dataset contained numerous missing values, which were addressed using the KNN imputer. Imputation, the process of rectifying null values using various methods, is crucial. Optimal imputation involves predicting missing values based on other data points. The KNN imputer accomplishes this by averaging samples from the training set around the missing points. Its effectiveness lies in selecting the appropriate number of neighbors to impute missing values. Through hyperparameter tuning, the distance between missing values and sample data points is configured using the Euclidean distance measure. This approach has demonstrated its efficacy in predicting missing values without relying on plagiarized content.
- Data Balancing: Imbalanced datasets can lead to biased predictions, hence the need for data balancing techniques. We evaluated methods like RandomOverSampler and SMOTE-Tomek, with RandomOverSampler demonstrating superior performance. This technique involves duplicating minority class instances to balance the dataset.

4. Model Making:

We have trained our model on different algorithms and have then used the concert custom and simple algorithm choosing the models with the best accuracies.

– Algorithms

1. Logistic Regression: Logistic regression is a well-known and commonly used classification concept for creating models. It is a form of a supervised machine learning algorithm in which the probability of the predicted class is calculated. The value lies between 0 and 1. If the likelihood is greater than 50 percentage, the class is classed, and vice versa, the probability of the class is calculated using the sigmoid function.
2. Decision Tree: Both regression and classification challenges can be tackled using a decision tree. The tree has various parameters: Internal Nodes (Features), Branches (Decision Making), Leaf Node (Outcomes). The main challenge for decision trees is tree formation. The Information gain (IG), as well as entropy for every column, is found out. If a column has the highest IG, it becomes the parent node. Then the dataset is again divided. We find the entropy and IG again. These 4 processes are repeated until the outcomes are reached.
3. Random Forest: When bagging operation is performed on a tree it is known as random forest technique. A single training algorithm is used on different subgroups of training examples in the bagging technique, and subset sampling is done with re placement. A single algorithm is trained all the subsets and then predictions are made for each subset and these predictions are aggregated. A random forest consists of a collection of decision trees. The handling of missing values is easy as compared to other models. We get accurate results with very low variance. Cross-validation can be done using the holdout and K fold method.
4. Support Vector Machine: The Support Vector Machine (SVM) is one of the most often used algorithms for classification problems. Lines are drawn using the support vector as a reference. To separate the data, a hyperplane is drawn. The hyperplane is chosen with the greatest possible margin. The distance between two parallel SV lines is known as the margin. It is a linear no probabilistic classifier. A hyperplane is an n-1-dimensional plane that distributes data into n dimensions optimally.
5. K-Nearest Neighbor: For both regression and classification, K nearest neighbours is used. The distance between testing data points and predicted points is calculated by KNN in order to forecast the class of test data. The closest K points to the test data are then selected. Once the points have been chosen, the algorithm

evaluates the probability of test points belonging to each of K training point classes, and then chooses the class with the greatest probability. As the complete data is stored in the model for training, the KNN technique is also known as the Lazy learner algorithm.

6. Gaussian NB: The basic algorithm which works on probabilistic function is known as Naive Bayes. It gives an effective classification. The Conditional probability of an event A is calculated considering event B has already happened. When the predictor value is continuous Gaussian NB classifier is used.
7. XGBoost: Boosting is a technique that involves several trees where the model making starts from the weaker decision and keeps on building the model such that the final prediction is a weighted sum of all the weaker decision-makers. The weights are assigned to the performance of individual trees. XGBoost is extreme gradient boosting. It regularizes data better than a normal gradient boosting tree. The objective function is the sum of loss function evaluated overall prediction and regular ization function for all predictors.
8. Custom Ensemble Algorithm: We have created our own cus tom ensemble model for predicting cardiovascular diseases. It gives an excellent accuracy on both the datasets (Framingham and Combine). Custom Ensemble algorithm is built by creating different machine learning algorithms and then using a Voting Classifier to get the predictions. Every model predicts its own class and then a Voting Classifier either (hard or soft) is used for taking max voting's of class predicted.
9. Hard Voting Classifier - In hard voting, every model gives its predicted class or votes, the class with a majority of votes wins.
10. Soft Voting Classifier: In soft voting, every model in the ensemble technique gives the probability for the class predicted. These probabilities are weighted and summed up. Then the class with the highest sum wins the vote.
11. Model Saving: The custom ensemble model is saved in the form of a pickle file which is further used for deployment.

5. Diabetes Prediction:

Dataset preprocessing: There were no null values present but there were values as 0 which is not acceptable as parameters like BMI can't be zero for living human. So, the zero values in parameters glucose, BMI, Blood pressure, Skin thickness, and insulin are equated with mean value of respective parameter. After above steps outliers rejection is performed on dataset using inter quartile method. After outlier rejection diabetes dataset size is reduced to 1536 rows whereas PID dataset is reduced to 619 rows.

– Algorithms:

- Random forest Classifier: Random Forest classifier comes un- der in supervised classification algorithm. As the name implies, it generates a forest with a lot of trees. The number of trees in the forest has a direct relationship with the algorithm's ac- curacy. When splitting a node, instead of looking for the most essential characteristic, it looks for the easiest feature from a random subset of features. Random forest classifier uses ensemble learning techniques. It handles missing values. It is kind of decision tree where bagging operation is performed.
- K nearest neighbour: KNN is supervised learning algorithm which is employed to for classification and regression problems. KNN works by finding the distances between a question and every one the examples within the data, selecting the required number examples (K) closest to the query, then votes for the foremost frequent label (in the case of classification) or averages the labels (in the case of regression). It then selects K points that are closest to the test data. Once the points are selected, the algorithm calculates the probability of the test points belonging to the classes of K training points, and the class with the highest probability is selected. KNN algorithm is also known as the Lazy learner algorithm as the entire data is held in the model for training.
- Support vector machine: For classification, a support vector machine is utilized. It can categorize both linear and non- linear data. To move the original training data into higher dimensions, SVM utilizes a non-linear mapping. It seeks for a linear optimum separation hyperplane in this dimension (hyperplanes are decision boundaries that are used to classify the data points). It finds hyperplanes using support vectors and margins.
- Decision Tree Classifier: It is supervised learning algorithm. It is built on a tree-like structure that explains the categorization process based on input features. Both regression and classification issues may be solved using the Decision Tree. The tree has various parameters: Internal Nodes (Features), Branches (Decision Making), Leaf Node(Outcomes). The main challenge for decision trees is tree formation. The Information gain(IG), as well as entropy for every column, is found out. If a column has the highest IG it becomes the parent node. Then the dataset is again divided. We find the entropy

and IG again. These 4 processes are repeated until the outcomes are reached. LGBM classifier: LightGBM is a decision tree-based gradient boosting framework that improves model efficiency and mem- ory utilisation.It employs two innovative techniques: Gradient-based One Side Sampling and Exclusive Feature Bundling (EFB), which ad- dress the shortcomings of the histogram-based approach utilised in all GBDT frameworks. LightGBM Algorithm's characteristics are formed by the two methodologies of GOSS and EFB explained below. They collaborate to make the model run smoothly and give it an advantage over competing GBDT frame- works.

- Cat Boost classifier: Unlike many other machine learning algorithms, CatBoost can directly handle categorical features without the need for preprocessing such as one-hot encoding. It internally converts categorical variables into numerical values using various encoding techniques. Gradient-boosted decision trees underpin CatBoost. A series of decision trees is built sequentially during training. In comparison to prior trees, each new tree is built with less loss. The initial parameters determine how many trees are created. Stacking model: Stacking frequently analyses heterogeneous weak learners, learns them in parallel, and then combines them by training a meta-learner to output a prediction based on the multiple weak learner's predictions, whereas bagging and boosting employed homogeneous weak learners for ensemble. A meta learner takes the predictions as features and the ground truth values in the data as the target, and it tries to figure out how to combine the input predictions to generate a better output prediction.

7. RESULTS:

We utilized two datasets to predict cardiovascular diseases, evaluating models with various machine learning algorithms. Evaluation metrics included Accuracy, Precision, Recall, F1 Score, and Support, calculated using a confusion matrix: Confusion Matrix: This visualizes classifier performance based on four parameters:
True Positive (TP): Model predicts a heart disease case correctly. True Negative (TN): Model predicts a non-disease case correctly.

False Positive (FP): Model predicts disease incorrectly (Type 1 error).

False Negative (FN): Model predicts non-disease incorrectly (Type 2 error).

Accuracy: Overall correctness of the model.
Accuracy = (TP + TN) / (TP + TN + FP + FN)
Precision: Identifies correct positive class instances.
Precision = TP / (TP + FP)

Recall: Percentage of actual positives accurately detected.
Recall = TP / (TP + FN)

F1 Score: Balances precision and recall.
F1 Score = 2*((precision*recall) / (precision + recall))

Support: Number of true response data points within the target class. These metrics provide comprehensive insights into model performance.

6. CONCLUSIONS

As the prevalence of heart disease continues to escalate the imperative to develop robust forecasting systems becomes increasingly evident. Our approach involved crafting tailored ensemble models for both datasets, delivering optimal accuracy. For the combined datasets, our bespoke ensemble model, comprising Logistic Regression, Decision Tree, Random Forest, Gaussian NB, and XGBoost classifiers, achieved an impressive 97% accuracy using a soft voting classifier. Meanwhile, for the Framingham datasets, custom ensemble models, incorporating decision tree, random forest, and XGBoost classifier with hard voting, yielded a commendable 95% accuracy. Given the alarming rise in diabetes cases, early detection becomes paramount for the effective treatment. Our proposed method boasts over 90% accuracy in diabetes prediction, facilitated by the utilization of diverse ensemble algorithm across two distinct datasets

8. REFERENCES

1. M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th Interna- tional Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.
2. N. Ahmad, Shafiullah, A. Algethami, H. Fatima and S. M. H. Akhter, "Comparative study of Optimum Medical Diagnosis of Human Heart Disease using Machine Learning Technique with and without Sequential Feature Selec- tion," in IEEE Access, doi: 10.1109/ACCESS.2022.3153047.
3. Riyaz L., Butt M.A., Zaman M., Ayob O. (2022) Heart Disease Prediction Using Machine

Learning Techniques: A Quantitative Review. In: Khanna A., Gupta D., Bhattacharyya S., Hassanien A.E., Anand S., Jaiswal A. (eds) International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, vol 1394. Springer, Singapore. https://doi.org/10.1007/978-981- 16-3071-2-8

4. D. P. Yadav, P. Saini and P. Mittal, "Feature Optimization Based Heart Dis- ease Prediction using Machine Learning," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), 2021, pp. 1-5, doi: 10.1109/ISCON52037.2021.9702410
5. . A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim and A. W. Muzaffar, "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," in IEEE Access, vol. 9, pp. 106575-106588, 2021, doi: 10.1109/ACCESS.2021.3098688.
6. Garg, Apurv Sharma, Bhartendu Khan, Rizwan. (2021). Heart disease pre- diction using machine learning techniques. IOP Conference Series: Materials Science and Engineering. 1022. 012046. 10.1088/1757-899X/1022/1/012046.
7. Rajdhan, Apurb Agarwal, Avi Sai, Milan Ghuli, Poonam. (2020). Heart Dis- ease Prediction using Machine Learning. International Journal of Engineering Research and. V9. 10.17577/IJERTV9IS040614.
8. Combine heart disease dataset: https://www.kaggle.com/fedesoriano/heart-failureprediction
9. Framingham Dataset: https://framinghamheartstudy.org/
10. Mitushi Soni, Dr. Sunita Varma, "Diabetes Prediction using Machine Learn- ing Techniques", International Journal of Engineering Research Technology (IJERT), IJERTV9IS090496,Vol. 9 Issue 09, September-2020, ISSN: 2278-0181
11. Aishwarya Mujumdar, V Vaidehi, "Diabetes Prediction using Machine Learn- ing Algorithms", INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING 2019, ICRTAC 2019
12. Jingyu Xue, Fanchao Min, Fengying Ma, "Research on Diabetes Prediction Method Based on Machine Learning" Publication: Journal of Physics: Conference Series, Volume 1684, Issue 1, article id. 012062 (2020). Pub Date: Novem- ber 2020 ,DOI: 10.1088/1742-6596/1684/1/012062
13. Quan Zou, Kaiyang Qu, Yamei Luo, Ying Ju, "Predicting Diabetes Melli- tus With Machine Learning Techniques" Front. Genet., 06 November 2018 — https://doi.org/10.3389/fgene.2018.00515
14. Md. Ashraful Alam, Dola Das, Eklas Husain, Mahmuddal Hasan "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers", March 2022, DOI:10.48550/arXiv.2203.04921
15. S. Ananya, J. Aravinth, R. Karthika, " Diabetes Prediction us- ing Machine Learning Algorithms with Feature Selection and Dimensionality Reduction" Published in: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), DOI: 10.1109/ICACCS51430.2021.9441935
16. Tawfik Beghriche, Mohamed Djerioui, Youcef Brik, Bilal Attallah, and Samir Brahim Belhaouari, "An Efficient Prediction System for Diabetes Dis- ease Based on Deep Neural Network", Volume 2021 — Article ID 6053824 — https://doi.org/10.1155/2021/6053824.