International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

HEART DISEASE PREDICTION USING NOVEL QUINE MCCLUSKEY BINARY CLASSIFIER (QMBC)

P Nikith[*], P Akshay Patel[*], V S Shiva Kishore[*], Shankar Raj Soni[#]
Department of IT, Guru Nanak Institutions Technical Campus, Hyderabad.
Assistant Professor, Department of IT, Guru Nanak Institutions Technical Campus, Hyderabad.

Abstract: Cardiovascular sickness is the essential justification for mortality around the world, liable for around 33% of all passings. To help clinical experts in rapidly distinguishing and diagnosing patients, various AI and information mining strategies are used to foresee the illness. Numerous analysts have created different models to help the proficiency of these forecasts. Include determination and extraction strategies are used to eliminate superfluous highlights from the dataset, in this way decreasing calculation time and expanding the proficiency of the models. In this review, we present another gathering Quine McCluskey Twofold Classifier (QMBC) procedure for recognizing patients determined to have a few type of coronary illness and the people who are not analyzed. The QMBC model uses a gathering of seven models, including strategic relapse, choice tree, irregular woods, K-closest neighbor, credulous Bayes, support vector machine, and multi-facet perceptron, and performs incredibly well on double class datasets. We utilize highlight choice and component extraction strategies to speed up the expectation interaction. We use Chi-Square and ANOV Ways to deal with distinguish the best 10 elements and make a subset of the dataset. We then apply Head Component Analysis to the subset to identify prime components. We utilize an ensemble of all seven models and the Quine McCluskey strategy to get the Base Boolean articulation for the objective element. The aftereffects of the seven models are viewed as free elements, while the objective property is reliant. We consolidate the extended results of the seven ML models and the objective element to shape a frothing dataset. We apply the group model to the dataset, using the Quine McCluskey least Boolean condition worked with a 80:20 train-to-test proportion. Our proposed QMBC model outperforms all present status of-the-craftsmanship models and recently recommended techniques set forward by different scientists.

Keywords: QMBC, ANOV, ML, strategic relapse, choice tree, irregular woods, K-closest neighbor, credulous Bayes, support vector machine, multi-facet perceptron.

I.INTRODUCTION

The term ‘‘Heart Disease’’ (HD) is used to refer to a variety of pathological disorders that have an impact on the heart and blood vessels. It encompasses a variety of heart-related conditions, including but not limited to vascular diseases and disturbances in heart rhythm. As per the World Health Organization (WHO), it is the deadliest and most devastating disease, taking over 18 million in lives a year. To diagnose it, healthcare professionals rely on a patient’s medical history and various tests, such as blood pressure, blood sugar, and cholesterol tests. Additionally, modern medical procedures like electrocardiograms, exercise stress tests, X-rays, echocardiography, coronary angiography, radionuclide tests, MRI scans, and CT scans can aid in the identification of cardiac conditions. Heart failure is the result of chronic issues that damage or weaken the heart muscles, leading to reduced ejection fraction. It is a condition that can affect both adults and children and cause severe damage to other vital organs in the body. The primary risk factors associated with heart failure are age, ethnicity, family history, hereditary factors, lifestyle choices, and pre-

existing cardiovascular disease (CVD) or genetics. While it affects both men and women equally, women are more likely to develop heart failure later in life [4].

To diagnose diseases at an early stage, ML is becoming an increasingly important tool. It aims to identify patterns hidden in observations and draw conclusions that are consistent with new information. Researchers have investigated the grouping of various techniques to create hybrid models that can outperform standalone models. Typically, these models have two phases. A subset of characteristics is chosen in phase-1 using Feature Selection (FS) and Feature Extraction (FE) techniques. The classifiers used in the phase-2 are then applied to this subset as input Heart disease datasets often contain various attributes, including both relevant and irrelevant as well as duplicate attributes. Relevant attributes are those that have an impact on how the target class is defined, whereas irrelevant do not con tribute to the output class's description. Redundant attributes, on the other hand, introduce noise rather than adding any new information to the target class's definition. Eliminating some traits that not only have an impact on the classification outcomes but also decrease system performance is crucial for improving the classification models. So, the HD diagnosis system requires the use of dimensionality reduction or FS techniques, as the datasets contain irrelevant and redundant features that contribute to noise rather than providing any information about the target class. The chance of overfitting is decreased, the model's capacity to generalize is increased, predictability is improved, and less computation is needed, which results in fewer features to enhance the performance of a model, ensemble techniques have been proven effective.

A considerable increase in performance improvement has also arisen from the inclusion of FS. To improve the ML Models, researchers are continuously exploring new approaches. Ensemble learning is a strategy that has been shown to enhance ML issues. Ensemble learning involves combining predictions from multiple classifiers using a process such as a majority voting. According to research, ensemble classifiers frequently outperform classical classifiers. An ensemble is a type of ML model that produces a final prediction by integrating the predictions from multiple individual models. These models can be of similar or diverse types, and there are numerous techniques available to combine them. Bagging and boosting are the two primary kinds of ensemble models. This research proposes an ML model to predict patients diagnosed with some form of HD and not diagnosed, using LR, DT, RF, KNN, NB, SVC, and MLP models. To enhance the model's performance, we use feature selection and feature extraction methods, including Chi-Square, ANOVA, and PCA techniques, to select and extract critical attributes from the dataset. An integrated strategy is developed by combining FS and FE techniques, reducing the dimensionality and accelerating the model's computation while retaining its best performance. This approach improves the model's efficiency while maintaining its effectiveness, allowing for more accurate predictions. We introduce a new ensemble technique, named QMBC, which aggregates the outputs of multiple individual models to generate a single prediction. Moreover, a variety of assessment measures have been used to rate classifier performance. The effectiveness of the proposed approach has been evaluated using three different datasets, namely the Cleveland HD dataset, the comprehensive HD dataset, and the CVD dataset. The effectiveness of the proposed method has also been compared to techniques that have been stated in the literature; among them are MIFH, RSA and RF model, weighted-average voting (WAVEn), XGBoost with Bayesian optimization, stacking classifier. Below is a list of the research study's achievements.

II. RELATED WORK

The first step of this work involves preprocessing the dataset, followed by applying FS and FE techniques to optimize computation time. Specifically, we utilized the Chi-Square and ANOVA techniques to eliminate irrelevant and redundant attributes and create subsets of features. We then applied the PCA FE technique to extract prime components from these subsets, further improving the efficiency of our model. We introduced a novel ensemble technique called Quine McCluskey Binary Classifier (QMBC) that combines predictions from multiple models to predict patients diagnosed and not diagnosed with some form of HD. In particular, we believe the Quine McCluskey method for predicting diagnosed and not diagnosed patients with HD is a new addition to this field, as no prior research has been conducted in this specific area. The effectiveness of the proposed QMBC is evaluated using 3benchmarkdatasets, including the Clevel and HD dataset, the HD dataset (comprehensive), and the CVD dataset. A detailed comparison of QMBC with current research shows that, in terms of accuracy, precision, recall, specificity, and f1-score, the proposed approach is more efficient than the state-of-the-art models.

As per the World Health Organization (WHO), it is the deadliest and most devastating disease, taking over 18 million in lives a year. To diagnose it, healthcare professionals rely on a patient's medical history and various tests, such as blood pressure, blood sugar, and cholesterol tests. Additionally, modern medical procedures like electrocardiograms, exercise stress tests, X-rays, echocardiography, coronary angiography, radionuclide tests, MRI scans, and CT scans can aid in the identification of cardiac conditions. Heart failure is the result of chronic issues that damage or weaken the heart muscles, leading to reduced ejection fraction. It is a condition that can affect both adults and children and cause severe damage to other vital organs in the body.

This flexibility allows us to leverage the strengths of different models and exploit their complementary nature, ultimately improving the overall performance. By adopting ensemble learning, we aim to maximize the predictive power of our models and provide more robust and reliable predictions for the problem under investigation. This choice is supported by previous studies and empirical evidence that demonstrate the efficiency of ensemble learning in a variety of domains and tasks. Overall, the decision to employ ensemble learning is driven by our pursuit of improved performance, enhanced generalization, and the desire to extract the full potential from the available data. Through this approach, we expect to achieve more accurate and reliable predictions, thereby contributing to advancements in the field of the health section.

Cardiovascular disease is the primary reason for mortality worldwide, responsible for around a third of all deaths. To assist medical professionals in quickly identifying and diagnosing patients, numerous machine learning and data mining techniques are utilized to predict the disease. Many researchers have developed various models to boost the efficiency of these predictions. Feature selection and extraction techniques are utilized to remove unnecessary features from the dataset, thereby reducing computation time and increasing the efficiency of the models. Additionally, modern medical procedures like electrocardiograms, exercise stress tests, X-rays, echocardiography, coronary angiography, radionuclide tests, MRI scans, and CT scans can aid in the identification of cardiac conditions. Heart failure is the result of chronic issues that damage or weaken the heart muscles, leading to reduced ejection fraction. It is a condition that can affect both adults and children and cause severe damage to other vital organs in the body.

III.LITERATURE SURVEY

A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, computerized electrocardiogram (ECG) interpretation plays a critical role in the clinical ECG workflow1. Widely available digital ECG data and the algorithmic paradigm of deep learning2 present an opportunity to substantially improve the accuracy and scalability of automated ECG analysis. However, a comprehensive evaluation of an end-to-end deep learning approach for ECG analysis across a wide variety of diagnostic classes has not been previously reported. Here, we develop a deep neural network (DNN) to classify 12 rhythm classes using 91,232 single-lead ECGs from 53,549 patients who used a single-lead ambulatory ECG monitoring device. When validated against an independent test dataset annotated by a consensus committee of board-certified practicing cardiologists, the DNN achieved an average area under the receiver operating characteristic curve (ROC) of 0.97. The average F_1 score, which is the harmonic mean of the positive predictive value and sensitivity, for the DNN (0.837) exceeded that of average cardiologists (0.780). With specificity fixed at the average specificity achieved by cardiologists, the sensitivity of the DNN exceeded the average cardiologist sensitivity for all rhythm classes. These findings demonstrate that an end-to-end deep learning approach can classify a broad range of distinct arrhythmias from single-lead ECGs with high diagnostic performance similar to that of cardiologists. If confirmed in clinical settings, this approach could reduce the rate of misdiagnosed computerized ECG interpretations and improve the efficiency of expert human ECG interpretation by accurately triaging or prioritizing the most urgent conditions.

C. G. D. S. E. Silva, G. C. Buginga, E. A. D. S. E. Silva, R. Arena, C. R. Rouleau, S. Aggarwal, S. B. Wilton, L. Austford, T. Hauer, and J. Myers, objective To develop a prediction model for survival of patients with coronary artery disease (CAD) using health conditions beyond cardiovascular risk factors, including maximal exercise capacity, through the application of machine learning (ML) techniques. Methods Analysis of data from a retrospective cohort linking clinical, administrative, and vital status databases from 1995 to 2016 was performed. Inclusion criteria were age 18 years or older, diagnosis of CAD, referral to a cardiac rehabilitation program, and available baseline exercise test results. Primary outcome was death from any cause. Feature selection was performed using supervised and unsupervised ML techniques. The final prognostic model used the survival tree (ST) algorithm. Results From the cohort of 13,362 patients (60±11 years; 2400 [18%] women), 1577 died during a median follow-up of 8 years (interquartile range, 4 to 13 years), with an estimated survival of 67% up to 21 years. Feature selection revealed age and peak metabolic equivalents (METs) as the features with the greatest importance for mortality prediction. Using these 2 features, the ST generated a long-term prediction with a C-index of 0.729 by splitting patients in 8 clusters with different survival probabilities (P<.001). The ST Root node was split by peak METs of 6.15 or less or more than 6.15, and each patient's subgroup was further split by age or other peak METs cut points. Conclusion Applying ML techniques, age and maximal exercise capacity accurately predict mortality in patients with CAD and outperform variables commonly used for decision-making in clinical practice. A novel and simple prognostic model was established, and maximal exercise capacity was further suggested to be one of the most powerful predictors of mortality in CAD.

M. Ozcan and S. Peker, heart disease remains the leading cause of death, such that nearly one-third of all deaths worldwide are estimated to be caused by heart-related conditions. Advancing applications of classification-based machine learning to medicine facilitates earlier detection. In this study, the Classification and Regression Tree (CART) algorithm, a supervised machine learning method, has been employed to predict heart disease and extract decision rules in clarifying relationships between input and output variables. In addition, the study's findings rank the

features influencing heart disease based on importance. When considering all performance parameters, the 87% accuracy of the prediction validates the model's reliability. On the other hand, extracted decision rules reported in the study can simplify the use of clinical purposes without needing additional knowledge. Overall, the proposed algorithm can support not only healthcare professionals but patients who are subjected to cost and time constraints in the diagnosis and treatment processes of heart disease.

M. M. Nishat, F. Faisal, I. J. Ratul, A. Al-Monsur, A. M. Ar-Rafi, S. M. Nasrullah, M. T. Reza, and M.R.H.Khan, heart failure is a chronic cardiac condition characterized by reduced supply of blood to the body due to impaired contractile properties of the muscles of the heart. Like any other cardiac disorder, heart failure is a serious ailment limiting the activities and curtailing the lifespan of the patient, most often resulting in death sooner or later. Detection of survival of patients with heart failure is the path to effective intervention and good prognosis in terms of both treatment and quality of life of the patient. Machine learning techniques can be critical in this regard since they can be used to predict the survival of patients with heart failure in advance, allowing patients to receive appropriate treatment. Hence, six supervised machine learning algorithms have been studied and applied to analyze a dataset of 299 individuals from the UCI Machine Learning Repository and predict their survivability from heart failure. Three distinct approaches have been followed using Decision Tree Classifier, Logistic Regression, Gaussian Naïve Bayes, and Random Forest Classifier, K-Nearest Neighbors, and Support Vector Machine algorithms. Data scaling has been performed as a preprocessing step utilizing the standard and min–max scaling method. However, grid search cross-validation and random search cross-validation techniques have been employed to optimize the hyper parameters. Additionally, the synthetic minority oversampling technique and edited nearest neighbor (SMOTE-ENN) data resampling technique are utilized, and the performances of all the approaches have been compared extensively. The experimental results clearly indicate that Random Forest Classifier (RFC) surpasses all other approaches with a test accuracy of 90% when used in combination with SMOTE-ENN and standard scaling technique. Therefore, this comprehensive investigation portrays a vivid visualization of the applicability and compatibility of different machine learning algorithms in such an imbalanced dataset and presents the role of the SMOTE-ENN algorithm and hyper parameter optimization for enhancing the performances of the machine learning algorithms.

P. Ghosh, S. Azam, M. Jonkman, A. Karim, F. M. J. M. Shamrat, E. Ignatious, S. Shultana, A. R. Beeravolu, and F. De Boe, cardiovascular diseases (CVD) are among the most common serious illnesses affecting human health. CVDs may be prevented or mitigated by early diagnosis, and this may reduce mortality rates. Identifying risk factors using machine learning models is a promising approach. We would like to propose a model that incorporates different methods to achieve effective prediction of heart disease. For our proposed model to be successful, we have used efficient Data Collection, Data Pre-processing and Data Transformation methods to create accurate information for the training model. We have used a combined dataset (Cleveland, Long Beach VA, Switzerland, Hungarian and Stat log). Suitable features are selected by using the Relief, and Least Absolute Shrinkage and Selection Operator (LASSO) techniques. New hybrid classifiers like Decision Tree Bagging Method (DTBM), Random Forest Bagging Method (RFBM), K-Nearest Neighbors Bagging Method (KNNBM), AdaBoost Boosting Method (ABBM), and Gradient Boosting Boosting Method (GBBM) are developed by integrating the traditional classifiers with bagging and boosting methods, which are used in the training process. We have also instrumented some machine learning algorithms to calculate the Accuracy (ACC), Sensitivity (SEN), Error Rate, Precision (PRE) and F1 Score (F1) of our model, along with the Negative Predictive Value (NPR), False Positive Rate (FPR), and False Negative Rate (FNR). The results are shown separately to provide comparisons. Based on the result analysis, we can conclude that our proposed model produced the highest accuracy while using RFBM and Relief feature selection methods (99.05%).

A.K.Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, the prediction of cardiac disease helps practitioners make more accurate decisions regarding patients' health. Therefore, the use of machine learning (ML) is a solution to reduce and understand the symptoms related to heart disease. The aim of this work is the proposal of a dimensionality reduction method and finding features of heart disease by applying a feature selection technique. The information used for this analysis was obtained from the UCI Machine Learning Repository called Heart Disease. The dataset contains 74 features and a label that we validated by six ML classifiers. Chi-square and principal component analysis (CHI-PCA) with random forests (RF) had the highest accuracy, with 98.7% for Cleveland, 99.0% for Hungarian, and 99.4% for Cleveland-Hungarian (CH) datasets. From the analysis, ChiSqSelector derived features of anatomical and physiological relevance, such as cholesterol, highest heart rate, chest pain, features related to ST depression, and heart vessels. The experimental results proved that the combination of chi-square with PCA obtains greater performance in most classifiers. The usage of PCA directly from the raw data computed lower results and would require greater dimensionality to improve the results.

IV.PROPOSED SYSTEM

In this paper, we propose a secure verifiable semantic. Searching scheme that treats matching between queries and documents as an optimal matching task. We treat the document words as “suppliers,” the query words as “consumers,” and the semantic information as “product,” and design the minimum word transportation cost (MWTC) as the similarity metric between queries and documents.

We assume that the data owner is trusted, and the data users are authorized by the data owner. The communication channels between the owner and users are secure on existing security protocols such as SSL, TLS. With regard to the cloud server, our scheme resists a more challenging security model which is beyond the “semi-honest server” used in other secure semantic searching schemes. In our model, the dishonest cloud server attempts to return wrong/forged search results and learn sensitive information, but would not maliciously delete or tamper with the outsourced documents. Therefore, our secure semantic scheme should guarantee the verifiability, and confidentiality under such a security model.

Data Collection

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.

There are several techniques to collect the data, like web scraping, manual interventions and etc.

Heart Disease dataset taken from Kaggle (https://www.kaggle.com/ronitf/heart-disease-uci

Dataset

The dataset consists of 303 individual data. There are 14 columns in the dataset, which are described below.

1. ***Age***: displays the age of the individual.
2. ***Sex***: displays the gender of the individual using the following format:
   1 = male
3. 0 = female
4. ***Chest-pain type(*cp*)***: displays the type of chest-pain experienced by the individual using the following format:
   1 = typical angina
   2 = atypical angina

3 = non — anginal pain
4 = asymptotic

5. ***Resting Blood Pressure(*trestbps*)***: displays the resting blood pressure value of an individual in mmHg (unit)
6. ***Serum Cholestrol(*chol*)***: displays the serum cholesterol in mg/dl (unit)
7. ***Fasting Blood Sugar(*fbs*)***: compares the fasting blood sugar value of an individual with 120mg/dl. If fasting blood sugar > 120mg/dl then: 1 (true)
   Else: 0 (false)

Data Preparation

Wrangle data and prepare it for training. Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, and data type conversions, etc.)

Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data

Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis

Split into training and evaluation sets.

Model Selection:

We used Decision Tree Classifier machine learning algorithm, We got a accuracy of 96.7% on test set so we implemented this algorithm.

Decision Tree Classification Algorithm

Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm**,** which stands for Classification and Regression Tree algorithm**.** A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

Below diagram explains the general structure of a decision tree:

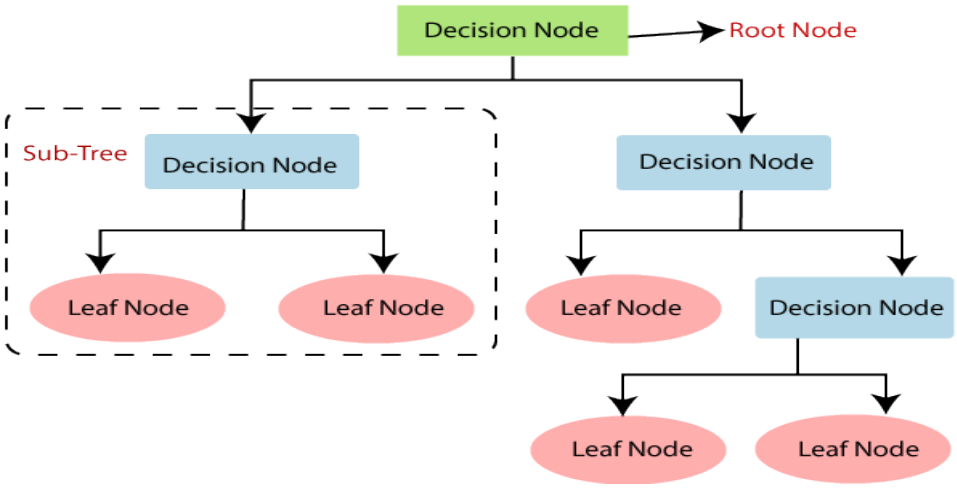

Figure 1. structure of a decision tree.

Why use Decision Trees?

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

- Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
- The logic behind the decision tree can be easily understood because it shows a tree-like structure.

Decision Tree Terminologies

Root Node: Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.

Leaf Node: Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.

V.TECHNIQUE USED OR ALGORITHM USED

ANOVA

ANOVA is a statistical technique that measures the significance of variations among categories or groups of data. The F-test score is used to determine the degree of variance in the target variable that can be attributed to the variance in a particular feature. The following describes how the ANOVA Algorithm 3 works:

By contrasting the variance of the feature's values among several classes of the target variable, find the F-test score for each characteristic in the dataset. Arrange the characteristics in descending order according to their F-test results. To generate the ultimate feature subset, pick the top k characteristics with the greatest F-test outcomes. In order to exclude attributes that do not strongly correlate to the target variable, a threshold may need to be set on the F-test result. 5) A fresh dataset created by the ANOVA technique, which only includes the chosen features, can be utilized to create an ML model. ANOVA can improve precision and effectiveness by focusing on the most crucial features.

Quine McCluskey Binary Classifier QMBC model

Quine McCluskey Binary Classifier (QMBC) technique for identifying patients diagnosed with some form of heart disease and those who are not diagnosed. The QMBC model utilizes an ensemble of seven models, including logistic regression, decision tree, random forest, K-nearest neighbor, naive bayes, support vector machine, and multilayer perceptron, and performs exceptionally well on binary class datasets. We employ feature selection and feature extraction techniques to accelerate the prediction process. We utilize Chi-Square and ANOVA approaches to identify the top 10 features and create a subset of the dataset. The results of the seven models are considered independent features, while the target attribute is dependent. We combine the projected outcomes of the seven ML models and the target feature to form a foaming dataset. We apply the ensemble model to the dataset, utilizing the Quine McCluskey minimum Boolean equation built with an 80:20 train-to-test ratio. Our proposed QMBC model surpasses all current state-of-the-art models and previously suggested methods put forward by various researchers.

VI.RESULTS AND DISCUSSION

We have conducted experiments on our collected dataset and extensive results have demonstrated that our model outperforms all other existing models. In the future, we will investigate more tasks under this framework, such as event summarization and event attribute mining in social media.

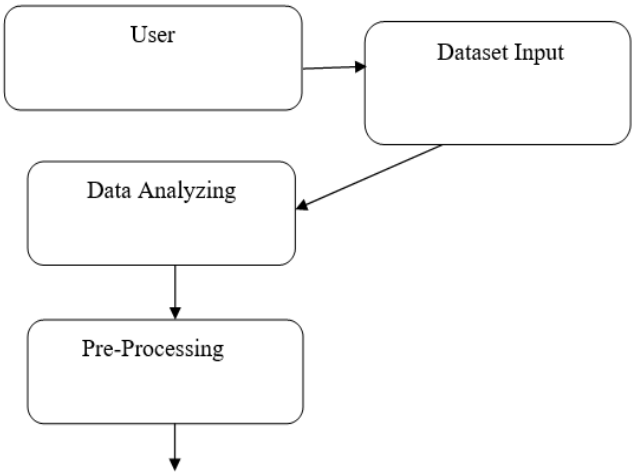

Figure 2. Basic flow of the work.

A data flow diagram (DFD) is a graphical representation of the "flow" of data through an information system, modeling its process aspects. Often they are a preliminary step used to create an overview of the system which can later be elaborated. DFDs can also be used for the visualization of data processing (structured design).

A DFD shows what kinds of data will be input to and output from the system, where the data will come from and go to, and where the data will be stored. Data user can also a login. Data user can have a search keyword and have a requested files. Data owner can also have a login. Data owner can also have an uploaded file and data user request. Cloud server can also a login. Cloud server can also have a data owner information data users information data .Cloud server can also have a keys. Cloud server can also have a file attacker details. All information is gather and store at a database.

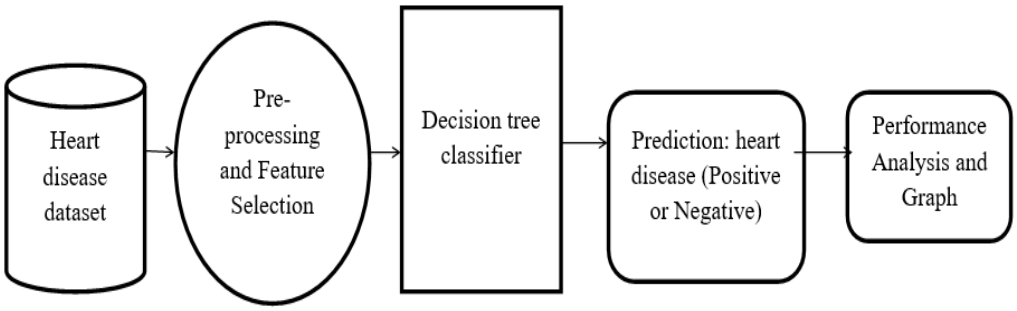

Figure 3. System architecture.

In this project data owner has a register all details and then login. Data owner can be an upload a document. Data owner can have a send request to the data user. Data user can search a query with uploaded document. The file has also a download it will show an encryption format. Data user also a send a request to the cloud server. Cloud server can a login. It will accept a key approve. Cloud server can also see all the data information's. Cloud server can also see all the user information. Cloud server can see all the stored information. Cloud server can approve a key request from the user. Then data owner has get the request data owner can send a secret key to the user. Then user can also download a file. If the user has given wrong keys it gets warning the user has a block permanently. The file it gets an attacks.

VII.CONCLUSION

In this paper, we present the privacy-preserving distributed extremely randomized trees algorithm for learning without privacy concerns in the healthcare domain. We have evaluated our proposed algorithm extensively using two popular structured healthcare datasets and two mental health datasets associated with the Norwegian Introducing Mental health through Adaptive Technology (INTROMAT) project. Our approach outperforms the state of the art in distributed tree-based models by up to 11.2% in terms of F1-score, 11.8% in terms of ACC, and 0.232 in terms of MCC for the Depression augmented dataset, and by up to 12.9% in terms of F1-score, 13.2% in terms of ACC, and 0.261 in terms of MCC for the Psykose augmented dataset. Moreover, we present the implementation of our technique on Amazon's AWS cloud, as a proof of concept, to evaluate the latency and scalability of our framework. The proposed algorithm has linear overhead with respect to the number of parties and can also handle datasets with missing values. We demonstrated our framework's efficiency in terms of prediction performance, scalability, and overheads, as well as privacy. The proposed framework provides the possibility of developing high-quality and accurate machine learning models without privacy concerns and is expected to contribute to a better healthcare system in the long term. As future work, we plan to explore the possibility of extending the proposed framework to settings where the parties do not follow the honest-but curious security model, which is beyond the scope of this work.

As future work, we plan to explore the possibility of extending the proposed framework to settings where the parties do not follow the honest-but curious security model, which is beyond the scope of this work.

REFERENCES

1. C. G. D. S. E. Silva, G. C. Buginga, E. A. D. S. E. Silva, R. Arena, C. R. Rouleau, S. Aggarwal, S. B. Wilton, L. Austford, T. Hauer, and J. Myers, ‘‘Prediction of mortality in coronary artery disease: Role of machine learning and maximal exercise capacity,’’ Mayo Clinic Proc., vol. 97, no. 8, pp. 1472–1482, Aug. 2022.
2. World Health Organization. (2009). Cardiovascular Diseases (CVDS). [Online]. Available: http://www.who.int/mediacentre/factsheets/fs317/en/ index.html.
3. Ravindra Changala, “Decision Tree Induction Approach for Data Classification Using Peano Count Trees” published in International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), ISSN: 2277 128X, Volume 2, Issue 4, April 2012.
4. M. Ozcan and S. Peker, ‘‘A classification and regression tree algorithm for heart disease modeling and prediction,’’ Healthcare Anal., vol. 3, Nov. 2023, Art. No. 100130.
5. Ravindra Changala, "Brain Tumor Detection and Classification Using Deep Learning Models on MRI Scans", EAI Endorsed Transactions on Pervasive Health and Technology, Volume 10, March 2024.
6. Ravindra Changala, "Sentiment Analysis in Social Media Using Deep Learning Techniques", International Journal of Intelligent Systems and Applications in Engineering, 2024, 12(3), 1588–1597.
7. Ravindra Changala, "Integration of IoT and DNN Model to Support the Precision Crop", International Journal of Intelligent Systems and Applications in Engineering, Vol.12 No.16S (2024).
8. M. M. Nishat, F. Faisal, I. J. Ratul, A. Al-Monsur, A. M. Ar-Rafi, S. M. Nasrullah, M. T. Reza, and M.R.H.Khan,‘‘Acomprehensiveinves tigation of the performances of different machine learning classifiers with SMOTE-ENN oversampling technique and hyperparameter optimization for imbalanced heart failure dataset,’’ Sci. Program., vol. 2022, pp. 1–17, Mar. 2022
9. P. Ghosh, S. Azam, M. Jonkman, A. Karim, F. M. J. M. Shamrat, E. Ignatious, S. Shultana, A. R. Beeravolu, and F. De Boer, ‘‘Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques,’’ IEEE Access, vol. 9, pp. 19304–19326, 2021.
10. Ravindra Changala, Development of Predictive Model for Medical Domains to Predict Chronic Diseases (Diabetes) Using Machine Learning Algorithms And Classification Techniques, ARPN Journal of Engineering and Applied Sciences, Volume 14, Issue 6, 2019.
11. A.K.Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, ‘‘Classifica tion models for heart disease prediction using feature selection and PCA,’’ Informat. Med. Unlocked, vol. 19, 2020, Art. No. 100330.
12. S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, ‘‘Improving heart disease prediction using feature selection approaches,’’ in Proc. 16th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST), Jan. 2019, pp. 619–623.
13. Ravindra Changala, “Evaluation and Analysis of Discovered Patterns Using Pattern Classification Methods in Text Mining” in ARPN Journal of Engineering and Applied Sciences, Volume 13, Issue 11, Pages 3706-3717 with ISSN:1819-6608 in June 2018.
14. J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, ‘‘Heart disease identification method using machine learning classification in e healthcare,’’ IEEE Access, vol. 8, pp. 107562–107582, 2020.
15. M. Ayar, A. Isazadeh, F. S. Gharehchopogh, and M. Seyedi, ‘‘Chaotic based divide-and-conquer feature selection method and its applica tion in cardiac arrhythmia classification,’’ J. Supercomput., vol. 78, pp. 5856–5882, Mar. 2022.
16. S. Shilaskar and A. Ghatol, ‘‘Feature selection for medical diagnosis: Evaluation for cardiovascular diseases,’’ Expert Syst. Appl., vol. 40, no. 10, pp. 4146–4153, Aug. 2013.

17. N. C. Long, P. Meesad, and H. Unger, ‘‘A highly accurate firefly based algorithm for heart disease prediction,’’ Expert Syst. Appl., vol. 42, no. 21, pp. 8221–8231, Nov. 2015.
18. Ravindra Changala “A Survey on Development of Pattern Evolving Model for Discovery of Patterns in Text Mining Using Data Mining Techniques” in Journal of Theoretical and Applied Information Technology, August 2017. Vol.95. No.16, ISSN: 1817-3195, pp.3974-398
19. D. Mienye, Y. Sun, and Z. Wang, ‘‘Improved sparse autoencoder based artificial neural network approach for prediction of heart disease,’’ Infor mat. Med. Unlocked, vol. 18, Jan. 2020, Art. No. 100307.
20. Ravindra Changala, “Classification by Decision Tree Induction Algorithm to Learn Decision Trees from the class-Labeled Training Tuples” published in International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), ISSN: 2277 128X , Volume 2, Issue 4, April 2012.
21. C. B. C. Latha and S. C. Jeeva, ‘‘Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques,’’ Informat. Med. Unlocked, vol. 16, Jan. 2019, Art. No. 100203.
22. Ravindra Changala, “Automated Health Care Management System Using Big Data Technology”, at Journal of Network Communications and Emerging Technologies (JNCET), Volume 6, Issue 4, April (2016), 2016, pp.37-40,ISSN: 2395-5317, ©EverScience Publications.
23. R. K. Sevakula and N. K. Verma, ‘‘Assessing generalization ability of majority vote point classifiers,’’ IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 12, pp. 2985–2997, Dec. 2017.