JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Towards Accurate and Secure Smart Healthcare: Federated Learning guided ensemble learning-based heart disease prediction

[1]Ashish Das, [2]Koyel Chakraborty, [3]Arindam Sarkar

[1]M-tech Student (CSE), [2]Assistant professor, [3]Assistant professor and Head
[1]Computer Science and Engineering,
[1]Supreme Knowledge foundation group of institutions, Hooghly, India
[3]Department of Computer Science and Electronics,
[3]Ramakrishna Mission Vidyamandira, Belur Marh-711202, India

Abstract: the data collected from all these interconnected devices (healthcare, transportation, etc.) is massive and complex. Machine learning comes in as a powerful tool to analyze and extract meaningful insights from this data. Supervised learning algorithms, like the ones mentioned, are trained on data that's already been labeled (categorized). This training allows the algorithms to learn patterns and then apply those patterns to classify new, unseen data. A form of artificial intelligence called machine learning involves teaching computer systems to make judgments or predictions based on data patterns. It is being used to create prediction models for a variety of medical diseases, which has made it a more and more well-liked instrument in the healthcare industry.

A machine learning technique called ensemble learning combines the forecasts of various models to increase accuracy. In the field of medicine, where it may be used to create more precise prognostic models for diseases like diabetes, breast cancer, and heart disease, this method has been found to be very useful.

I. INTRODUCTION

Healthcare is only one of the numerous domains that machine learning has transformed. It entails teaching computer algorithms to automatically learn from data and then form hypotheses or come to findings based on that learning.

Combining the predictions of various models using ensemble learning is a machine learning strategy that enhances overall accuracy. AdaBoost, LDA, Random Forest, Gaussian Naive Bayes, Bagging, Gradient Boost, and Decision Trees are a few ensemble learning algorithms that have been proven to be successful in medical applications.

II. Decision Tree

A decision tree is a popular machine learning algorithm that is commonly used for classification and regression tasks. In order to achieve homogeneity or the greatest tree depth, it iteratively divides the input data into smaller and smaller subsets according to a set of decision rules.

Decision trees are a common option in many industries since they are simple to understand and visualize. Decision trees have been utilized in the healthcare industry to create predictive models for a number of illnesses, such as diabetes, breast cancer, and heart disease.

III. Gaussian Naïve Bayes

Gaussian Naive Bayes is a popular machine learning algorithm that is widely used in classification tasks, particularly in natural language processing applications. It is predicated on the Bayes theorem and the idea that each feature in a dataset exists independently.

IV. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a popular technique in the field of machine learning and statistical analysis. It is a supervised learning technique used for dimensionality reduction and classification. When the data can be linearly separated, LDA is especially helpful.

Finding a linear combination of features that maximizes the distance between classes is how LDA works. Finding a projection of the data onto a lower-dimensional space that maximizes the ratio of the between-class variance to the within-class variance is how it accomplishes this. It does this by determining the mean and covariance of each class.

2. Decentralized Federated Learning

Decentralized federated learning is an emerging technique that has great potential for improving the healthcare sector. It allows for the training of machine learning models on decentralized data sources without requiring data to be centralized in one location. This is particularly important in healthcare, where privacy and security of patient data are of utmost importance.

In decentralized federated learning, each participant (such as hospitals or clinics) trains their own local model on their own data, and then shares only the model updates with a central server. The server then aggregates the model updates to create a global model, which is sent back to each participant. This process is repeated iteratively until the global model achieves satisfactory accuracy.

2.2 Dataset Description

The heart disease dataset is a collection of medical data from patients with suspected heart disease. It contains 14 different features, including patient age, sex, chest pain type, resting blood pressure, serum cholesterol levels, and electrocardiographic results. The dataset was obtained from the UCI Machine Learning Repository and was originally contributed by Hungarian Institute of Cardiology. The dataset consists of 303 observations, with each observation representing a patient. The target variable indicates whether or not the patient has heart disease, with a value of 0 indicating no heart disease and a value of 1 indicating the presence of heart disease. In each cross-validation, we are checking for most popular ratios like
90% for training data and the remaining 10 % for testing data, 80% for training data and the remaining 20 % for testing data, and 75% for training data and the remaining 25% for testing data. Then selected the best ratio, 90% of the data for training and the remaining 10% for testing.

3. Theoretical framework

Federated Random Forest: A Hybrid Algorithm for Distributed Classification
Inputs:

- Number of trees (T)
- Max depth of trees (D)
- Number of features (F)
- Client data (C_1, C_2, ..., C_n)
- Public test data (T)

Output:
Trained Random Forest model.

Procedure:
Let's assume we have a dataset with N observations and M features, denoted by X = {x_1, x_2, ..., x_N} and the corresponding target variable Y = {y_1, y_2,
..., y_N}. We want to build a model f(x) that can predict the target variable Y based on the features X.

Initialize the model by setting the number of trees T and the number of features to consider at each split m.

1. Initialize an empty list to store the trees
2. For each client c in C_1, C_2, ..., C_n:
    a. Sample a fraction of the client's data, D_c, uniformly at random.
    b. Train a decision tree with depth D on D_c using a subset of F randomly selected features.
    c. Append the trained tree to the list of trees.
3. For each test example in T:
    a. Aggregate the predictions from all trees in the forest by taking the majority vote.
    b. Return the overall predicted label.

Training a local Random Forest model:

1. Initialize an empty tree.
2. If the stopping criteria are met (e.g., maximum depth is reached or minimum number of samples is reached), return the current node as a leaf with the majority class label.
3. Select the feature that provides the highest information gain or Gini impurity reduction as the splitting criterion.

4. Split the data based on the selected feature and its threshold value.
5. Recursively repeat steps 2-4 for each child node until the stopping criteria are met.

Aggregating local models:

1. For each Random Forest model from each device, traverse the model to predict the class label for each sample in the testing set.
2. Assign each sample to the class label that is predicted by the majority of the models.
   Return the final aggregated model

3.1 Equations

- PPV (Positive Predictive Value) or Precision: It measures the proportion of true positives (TP) among the total predicted positives (TP+FP).

PPV (**Positive Predictive Value**)**or Precision** = TP/TP+FP

- Recall (Sensitivity, True Positive Rate, or Hit Rate): It measures the proportion of true positives (TP) among the total actual positives (TP+FN).

Recall (**Sensitivity**, **True Positive Rate**, **or Hit Rate**) = TP/TP+FN

- Specificity (True Negative Rate or Selectivity): It measures the proportion of true negatives (TN) among the total actual negatives (TN+FP).

Specificity (**True Negative Rate or Selectivity**) = TP/TP+FP

- FPR (False Positive Rate): It measures the proportion of false positives (FP) among the total actual negatives (TN+FP)

False Positive Rate (**FPR**)= FP/FP+TN

4. RESEARCH METHODOLOGY

This thesis puts forth a strategy that integrates block chain technology with federated learning to handle privacy, integrity, and ownership issues while training Gradient Boost, GNB, Random Forest, and Bagging with IoT data from many suppliers. In order to ensure that data analysts can only access data through communication with the correct data providers on the block chain, the suggested approach involves each provider's data before recording it on a distributed ledger. The paper develops secure protocols for four fundamental Gradient Boost, GNB, RF, and BAGG training operations to enable secure training on data with the aid of federated learning. This is done while ensuring that data providers cannot access each other's data and that the model parameters of the data analyst are kept a secret from the data providers throughout the training process.

4.1 **Performance Evaluation**

This section discusses the evaluation of Gradient Boost, Gaussian Naive Bayes (GNB), Bagging (BAGG), Random Forest (RF) in terms of its accuracy and efficiency using real-world datasets. We begin by describing the experiment settings, and then present the experimental results to demonstrate its effectiveness and efficiency.

To demonstrate that does not sacrifice the accuracy of the classifiers, we conducted experiments using the standard Gradient Boost, Gaussian Naive Bayes (GNB), Bagging (BAGG), Random Forest (RF) implementation with federated learning in python with TensorFlow, named Federated_Gradient Boost, Federated_GNB, Federated_Bagging, Federated_RF. Since our focus is on securely training classifiers, we used the default parameters and did not adjust the training parameters. Table VI presents the precision and recall results.

4.2 **Security Analysis**

In this section, we provide a security analysis under the known background model. We adopt two security definitions: secure federated learning computation [21] and differential privacy computation [22], which are commonly used in the literature to ensure secure and private protocols in the presence of honest-but-curious adversaries. Our security proof is based on the ideas of these two definitions, and we refer the interested reader to [21] for a detailed discussion on secure two-party computation and to [22] for modular sequential composition.

5. RESULTS AND DISCUSSION

5.1 Results of Descriptive Statics of Study Variables

Table 5.1: Descriptive Statics

Model	Split Ratio	Precision	Recall	F1_Score	Accuracy	Model
Random Forest	Train 90%, Test 10%	0.9999	0.9999	0.911320755	0.9999	Random Forest
Random Forest	Train 80%, Test 20%	0.9999	0.9619047619	0.9805825243	0.9804878049	Random Forest
Random Forest	Train75%, Test 25%	0.9999	0.9848484848	0.9923664122	0.9922178988	Random Forest
Gaussian Naive Bayes	Train 90%, Test 10%	0.9074074074	0.9848484848	0.9158878505	0.9126213592	Gaussian Naive Bayes
Gaussian Naive Bayes	Train 80%, Test 20%	0.7863247863	0.8761904762	0.8288288288	0.8146341463	Gaussian Naive Bayes

5.2 Kernel Density Plot

In a kernel density plot, a smooth curve is drawn over a histogram of the data, where each observation in the data set is represented by a small "kernel" or function. The height of the curve at any point represents the estimated probability density of observing a value in that range.

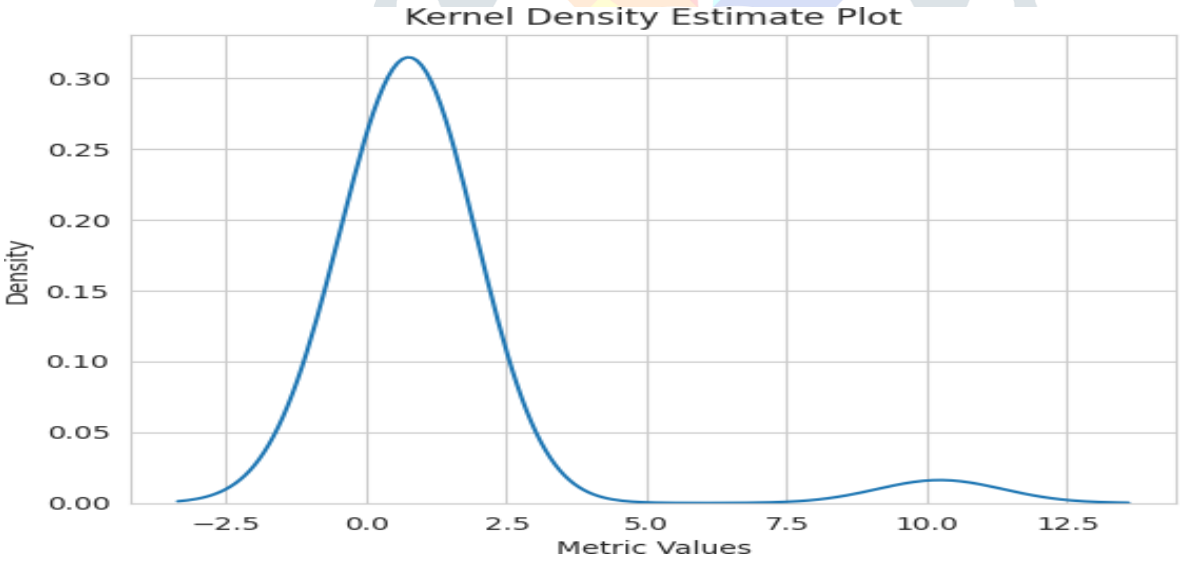

Figure5-2: Kernel Density Plot of metrics for the Best Model

5.3 Pie Plot

Pie chart could be used to display the accuracy rates of different models or algorithms used in a machine learning project. Each slice of the pie chart would represent the accuracy rate of a particular model, with the largest slice indicating the model with the best accuracy. Including a small note or label indicating which model has the best accuracy can make the information even more clear and understandable for the audience.

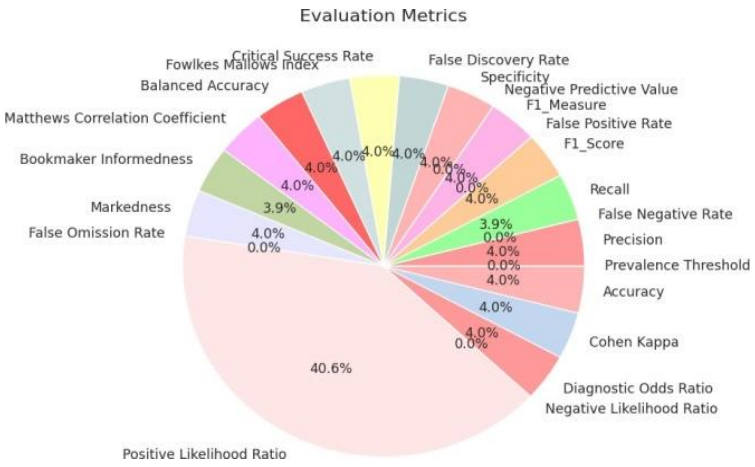

Figure 5-3: Pie plot of metrics for the Best Model

6. ACKNOWLEDGMENT

Acknowledge the source of the heart disease data used for training and testing your models. If the data is publicly available, mention the source and reference. If it was provided by a specific institution, thank them for sharing the data.

We would like to thank Supreme knowledge foundation group of instituions for sharing the heart disease data used in this study.

REFERENCES

[1] Q. Li et al., "A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection," in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 4, pp. 3347-3366, 1 April 2023, doi:https://doi.org/10.1109/TKDE.2021.3124599.

[2] Amelia Jiménez-Sánchez, Mickael Tardy, Miguel A. González Ballester, Diana Mateus, Gemma Piella, Memory-aware curriculum federated learning for breast cancer classification,Computer Methods and Programs in Biomedicine,Volume 229,2023, 107318,ISSN 0169-2607,https://doi.org/10.1016/j.cmpb.2022.107318.

[3] Yaqoob, Mateen & Nazir, Muhammad & Qureshi, Sajida & Al-Rasheed, Amal. (2023). Hybrid Classifier-Based Federated Learning in Health Service Providers for Cardiovascular Disease Prediction. Applied Sciences. 13. 1911. 0.3390/app13031911. DOI: https://doi.org/10.3390/app13031911

[4] Shaheen, Momina & Farooq, Shoaib & Umer, Tariq & Kim, Byung-Seo. (2022). Applications of Federated Learning; Taxonomy, Challenges, and Research Trends. Electronics.11. 670. 10.3390/electronics11040670. doi: https://doi.org/10.3390/electronics11040670