Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7778381

COMPARISON OF MACHINE LEARNING ALGORITHMS FOR ECG INTERPRETATION: RANDOM FOREST, K-NEAREST NEIGHBORS, AND LOGISTIC REGRESSION

Dr. Sreelekha Menon
Associate Professor, SCMS School of Engineering and Technology, Ernakulam

Ms. Surya K A
Assistant Professor, SCMS School of Engineering and Technology, Ernakulam'

Ms. Reshma R
Assistant Professor, SCMS School of Engineering and Technology, Ernakulam

ABSTRACT

Our research intends to extensively investigate the capabilities of three distinct machine learning algorithms, namely Random Forest, K-Nearest Neighbours, and Logistic Regression. Our key goal is to assess their ability to foresee outcomes and comprehend them easily. We intend to acquire significant insights into the potential of these three strategies and their applicability in diverse fields by undertaking a complete analysis. We used 188 ECG interpretations from 19566 patients, which included various sorts of data. We evaluated the data with various quantities of data ranging from 165 to 520 training sets. Algorithms based on K-Nearest Neighbour (KNN), Logistic Regression (LR), and Random Forest each achieved 98%, 94%, and 99% accuracy, respectively.

KEYWORDS: Random Forest, K-Nearest Neighbors, Logistic regression, Electrocardiogram

INTRODUCTION

Recently, the field of machine learning has advanced in an impressively large way. Numerous businesses and academic disciplines are seeing changes as a result. With so much data being generated every day, having reliable forecasting models in place is becoming increasingly crucial. This is where machine learning approaches come into play because they give us powerful tools to better understand complex data and base judgments on it. Our study focuses on the K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest machine learning algorithms. These algorithms are very adaptable and efficient in a variety of tasks, from classification issues to regression issues. We can identify the algorithm that performs best in a particular situation by comparing its advantages and disadvantages. When deciding how to approach various datasets as well as issue domains, experts will be better equipped with this information.

The use of machine learning techniques for predictive analysis as well as choice-making in medical applications has grown with the development of electronic medical records and the immense quantity of data created in the healthcare industry. Analyzing patient electrocardiogram (ECG) data is one use of these methods that has a lot of promise. An electrocardiogram, or ECG, is a vital diagnostic tool that records the electrical signals of the heart and is frequently used to investigate abnormalities in the heart's electrical activity and assess heart health. Our main goal is to analyze ECG data in order to use these algorithms to identify and categorize cardiac problems. Our aim is to enhance the precision and reliability of heart illness diagnosis and prognosis by evaluating these algorithms on actual patient ECG data. Ultimately. By introducing more accurate and effective diagnostic tools, we anticipate that this research will completely transform patient care in cardiology.

1. Background and Motivation:
   Due to their enormous impact on morbidity and death rates, cardiovascular conditions continue to be a serious worldwide health issue. Quick and accurate cardiac abnormality detection is crucial for delivering appropriate medical therapies and improving patient outcomes. Due to human error, traditional methods of ECG analysis can be tedious and susceptible to mistakes. Automating and improving ECG interpretation with the aid of machine learning can result in speedier and more accurate diagnoses.
2. Objectives:
   The main goal of the study is to evaluate how well Random Forest, K-Nearest Neighbors, and Logistic Regression perform in categorizing patient ECG data into different cardiac diseases. We want to assess how well each algorithm can identify common cardiac problems in terms of reliability, specificity, sensitivity, and other important parameters. We also intend to look at how well the algorithms can handle other ECG data types, including regular, irregular, and noisy recordings. Additionally, we will evaluate the models' computational effectiveness and scalability while dealing with sizable ECG datasets. Finally, in order to gain an understanding of the diagnostic procedure, we shall pinpoint the ECG parameters or features that affect each method the most.
3. Data Set and Methodology:
   We are now working with a large and varied ECG dataset derived from real patients, representing a variety of heart diseases. To guarantee that noise is eliminated, values that are missing are handled, and the way the features are represented is standardised, we will begin implementing preprocessing methods prior to evaluating the data. On the dataset, we will also train, validate, and test 3 machine learning algorithms using the proper hyperparameter modification and cross-validation methods.
4. Significance:
   The results of this study have significant implications for the medical field and cardiac professionals. The diagnostic procedure might be significantly sped up, allowing for the prompt identification of cardiac issues and the development of individualized treatment plans, with the invention of trustworthy machine-learning algorithms for ECG interpretation. Making medical resources more effectively used, could also lessen the stress on healthcare professionals and improve patient care.
5. Ethical Considerations:

   In this study, we take the security and privacy of patient data very seriously. We completely anonymize all patient data we utilize, and we treat it in strict accordance with all applicable data protection laws and ethical standards. We recognize the value of protecting sensitive data, and we aim to make sure that all study participants feel comfortable and secure in the knowledge know their information is being treated with the highest respect and care.

RELATED WORKS

The effectiveness of machine learning algorithms and their applicability for various use cases have drawn more attention in recent years. A study done in 2021 by K. Uma Pavan Kumar, Ongole Gandhi, M. Venkata Reddy, and S. V. N. Srinivasu is noteworthy in this regard. The study examined KNN, Decision Trees, and Random Forest Algorithms' statistical and mathematical components, examining their strengths and weaknesses. Researchers as well as professionals in the discipline of machine learning can use the study's in-depth analysis of these algorithms as a reference.

Additionally, in their 2021 work, Mohammad Savargiv, Behrooz Masoumi, and Mohammad Reza Keyvanpour suggested a clever strategy to improve the effectiveness of the random forest algorithm. The algorithm's flexibility to multiple issue domains and independence from the data domain are increased by the authors' incorporation of learning automata. This advancement is noteworthy because it could increase the algorithm's precision and broaden its application to a variety of use scenarios.

The success rate of Logistic Regression, Random Forest, and KNN models for text classification using BBC News text classification was studied in a different research project by Kanish Shah, Henil Patel, Devanshi Sanghvi, and Manan Shah in their 2020 publication. The authors highlighted each algorithm's applicability for various text categorization use cases by providing a thorough review of its advantages and disadvantages.

Similarly to this, Ratna Astuti Nugrahaeni & Kusprasapta Mutijarsa examined the efficacy of KNN, SVM, and Random Forest algorithms in 2016's article "Comparative analysis of machine learning KNN, SVM, and random forests algorithm for facial expression classification." The authors gave a thorough examination of the efficacy of each method, noting its flaws and applicability to various emotions classification use cases.

We conducted in-depth research on numerous online platforms, including analytics vidhya, Datacom, comparable web, and Google Scholar, to learn more about these investigations and their findings. These platforms were useful tools for identifying pertinent scholarly material and evaluating study results.

METHODOLOGY

Data Collection and Preprocessing:

It is crucial to obtain a trustworthy ECG dataset from trusted sources or collaborate with healthcare organizations while abiding by ethical standards and patient privacy laws in order to produce accurate and trustworthy results. To maintain data quality and consistency, the ECG data must be prepared properly. This entails removing artefacts, baseline wander, and noise, and undertaking to resample as required. Additionally, standardizing the characteristic structure of the ECG data is essential to enabling fair comparisons amongst machine learning models.

Feature Extraction:

In order to provide insightful information about diverse cardiac diseases, it is essential to extract specific characteristics from preprocessed ECG data. Heart rate, QRS length, PR interval, ST segment, and other morphological traits are examples of these aspects. It is essential to consult with medical professionals and take domain knowledge into account to make sure the traits selected have therapeutic significance. With the correct characteristics, we can accurately categorize heart illnesses and provide patients with the best care.

Data Splitting:

The ECG dataset was divided into three separate sets with different goals in accordance with the instructions. Using the first subset, machine learning models will be trained. To improve the hyperparameters and prevent overfitting, utilize the second group. The performance of the final model is evaluated using the third subgroup. This approach ensures the crucial accuracy and dependability of the ECG analysis.

Model Implementations:

Random Forest, K-Nearest Neighbors, and Logistic Regression are just a few of the algorithms we've implemented using the Python sk-learning module. We made a hyperparameter space for each method to investigate the range of hyperparameters and fine-tune the models. The models' performance is optimized by the use of this technique, leading to precise forecasts.

Model Training and Validation:

For our machine learning models to be more effective, it is crucial to train them on the training set using different hyperparameter configurations. We can accurately evaluate the models' performance on the validation set after training them using a range of assessment metrics, including accuracy, precision, recall, and F1-score. We can also get more accurate performance estimates by using k-fold cross-validation, which also helps to lower variance. We can improve our machine-learning models' accuracy and get more precise forecasts by doing this.

Hyperparameter and Tuning:

For our machine learning models, we have been experimenting with various hyperparameters. On our validation set, I utilized grid search, random search, and Bayesian optimization to identify the optimal hyperparameters for each algorithm. The most efficient hyperparameter settings for each model were discovered after further fine-tuning. Although it was a challenging endeavour, we feel the models are now performance-optimized.

Model Comparison and Selection:

Using statistical tests to examine the results of the Random Forest, K-Nearest Neighbors, and Logistic Regression models on the validation set, we have found that there are notable discrepancies between these models. We chose the model that performed best based on evaluation results and retrained it using a combination of datasets for training and validation. We are certain that this strategy will result in extremely accurate and dependable results.

Model Evaluation:

It is essential to evaluate the model on the test set, which consists of fresh data, in order to obtain an accurate assessment of its generalization capacity. By doing so, we'll be able to see how well the model works in actual situations. The key to determining the model's strengths and weaknesses is to evaluate the predictions made on the data set, particularly the degree of sensitivity, specificity, and misdiagnosis rates. We can then focus on those aspects of the model that may need development and make the necessary adjustments to ensure the greatest performance.

Interpretability Analysis:

To better understand the judgments made by the models, use feature significance analysis and interpretability methodologies. The clinical interpretability of the models is improved by analyzing the role of the ECG features in the classification process, offering insightful information about the variables affecting the models' conclusions. The patients would ultimately benefit from more precise and efficient diagnoses and treatment regimens as a result of this strategy.

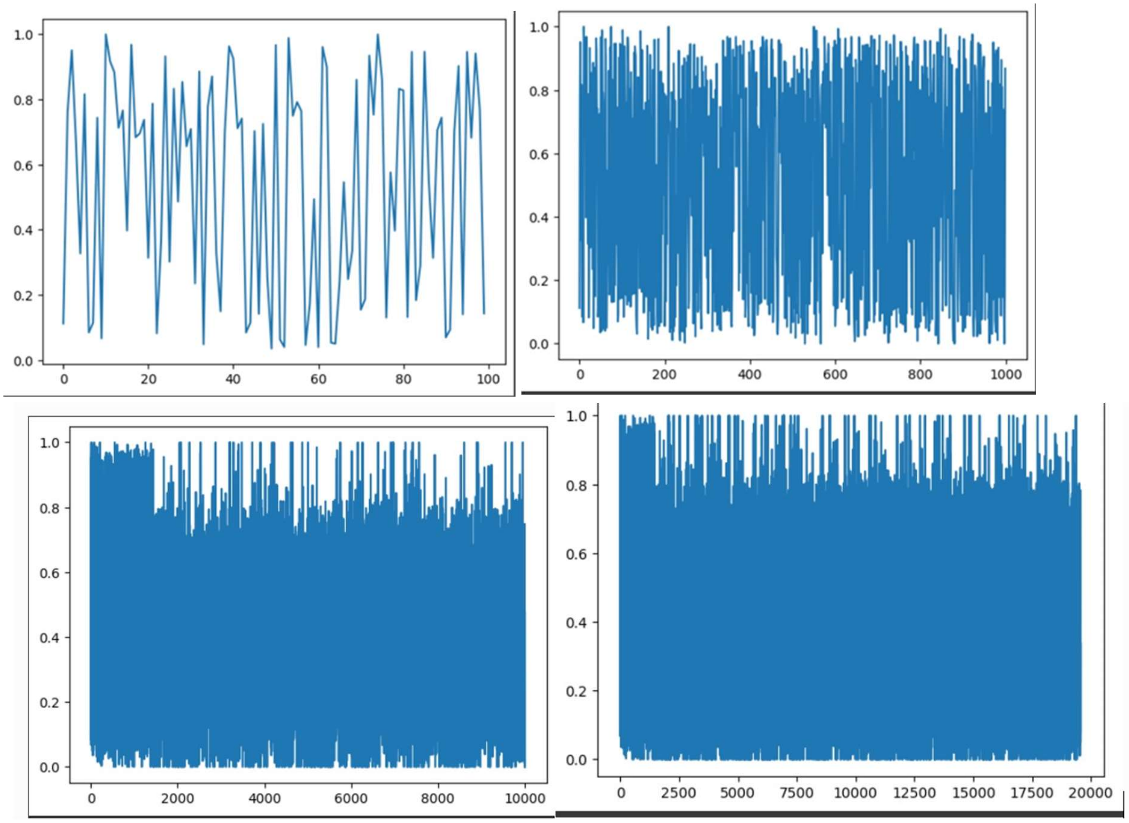
Sensitivity and Robustness Analysis:

To assess the models' robustness in handling noisy or missing ECG data, perform a sensitivity analysis by adding artificial noise or perturbations to the test set.

RESULTS AND CONCLUSIONS

This study aims to evaluate the performance of three distinct machine learning algorithms, namely K-Nearest Neighbour (KNN), Logistic Regression (LR), and Random Forest, for ECG interpretation utilising a dataset of 19566 patients, totalling 188 ECG interpretations. The algorithms were assessed based on their accuracies after the data was divided into training sets with sample counts ranging from 165 to 520. According to the results, the Random Forest algorithm had the highest accuracy (99%), followed by KNN (98%), and LR (94%). These results point to Random Forest as a promising method for reliable ECG interpretation, with potential clinical repercussions for better patient diagnosis and care.

DATA EXPLORATION



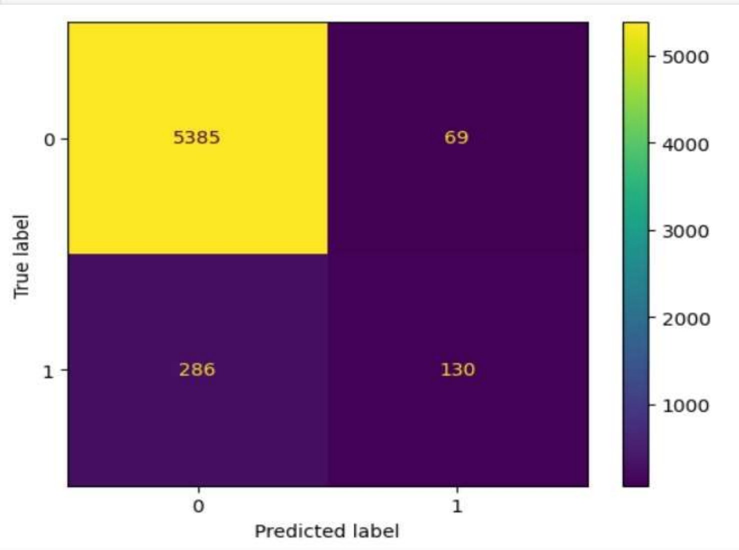
MODEL METRICS OF LOGISTIC REGRESSION

Model Score:

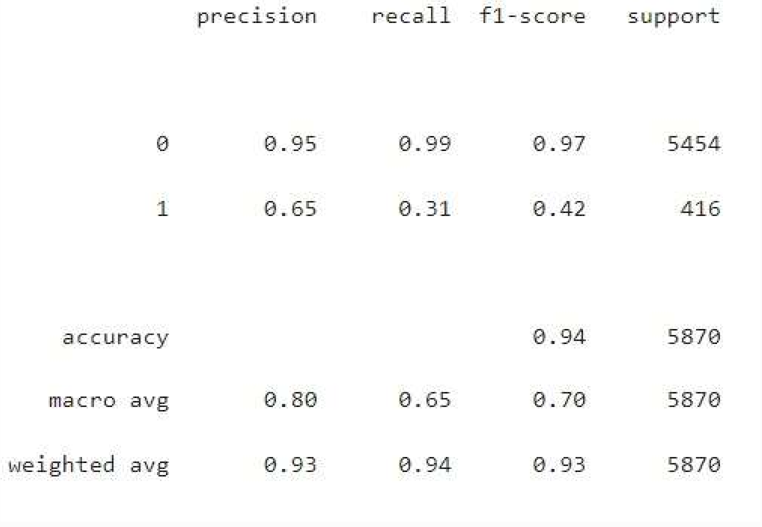
Model Score when using LR – 0.94

A high model score in Python is typically considered to be anything above 0.90. This means that the model is able to make accurate predictions 94% of the time. would be more important

Confusion Matrix



- **CLASSIFICATION REPORT**



```
              precision    recall  f1-score   support


           0       0.95      0.99      0.97      5454

           1       0.65      0.31      0.42       416


    accuracy                           0.94      5870

   macro avg       0.80      0.65      0.70      5870

weighted avg       0.93      0.94      0.93      5870
```

Precision, recall, and the F1 score for at-risk are all extremely high in the image. This indicates that the model is quite effective at foretelling occurrences of at-risk patients. For healthy, the F1 score, recall, and precision are all lower but still reasonable. This indicates that while the model is still doing a respectable job, it is not as good at predicting cases of healthy that are positive. The weighted average of the precision, recall, and F1 scores is 0.93. This is a reliable indicator of the model's overall performance. It shows that the model is effective at forecasting both positive and negative events. With a 0.94 accuracy rating, the model correctly categorized 94% of the data in the evaluation dataset.

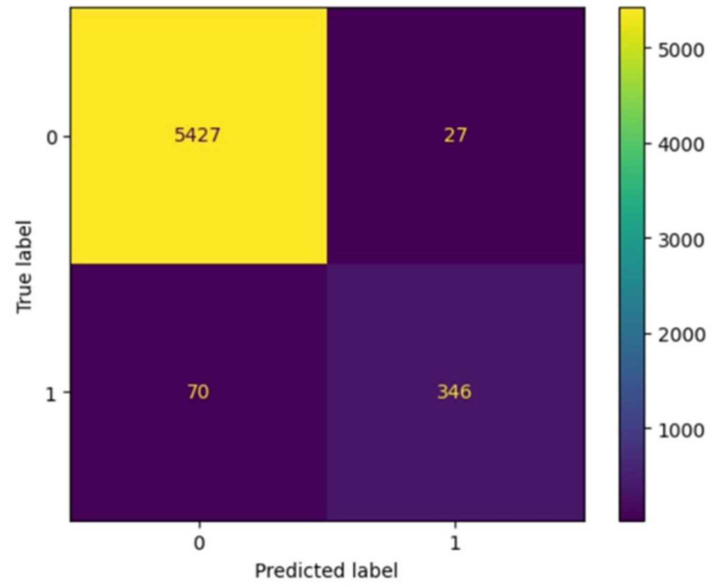
MODEL METRICS OF K- NEAREST NEIGHBOR

Model Score:

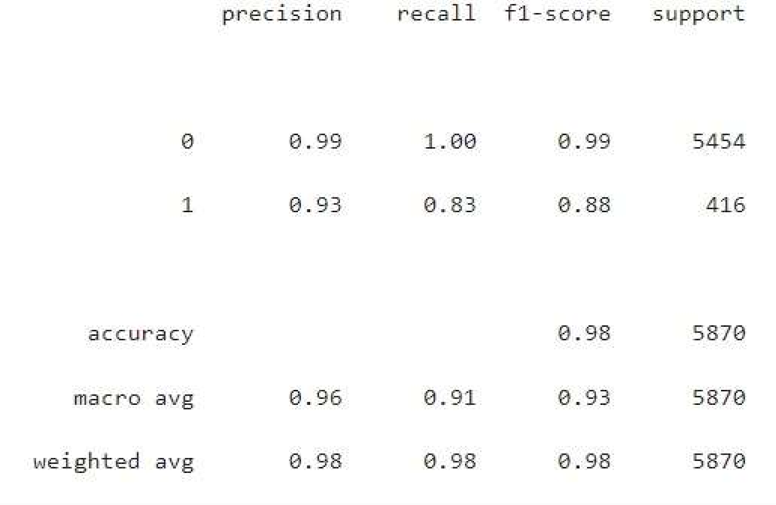
Model Score when using KNN – 0.98

A good model score has different implications depending on the application, though. A better model score may be more significant if the objective is, for instance, to reduce the amount of false positives. A lower model score would be more significant if the objective were to reduce the amount of false negatives.

Confusion Matrix



Classification Report



```
              precision    recall  f1-score   support


           0       0.99      1.00      0.99      5454

           1       0.93      0.83      0.88       416


    accuracy                           0.98      5870

   macro avg       0.96      0.91      0.93      5870

weighted avg       0.98      0.98      0.98      5870
```

The classification report shows that the accuracy rating of the model is 0.98 which means the model correctly categorized 98% of the data in the evaluation dataset.

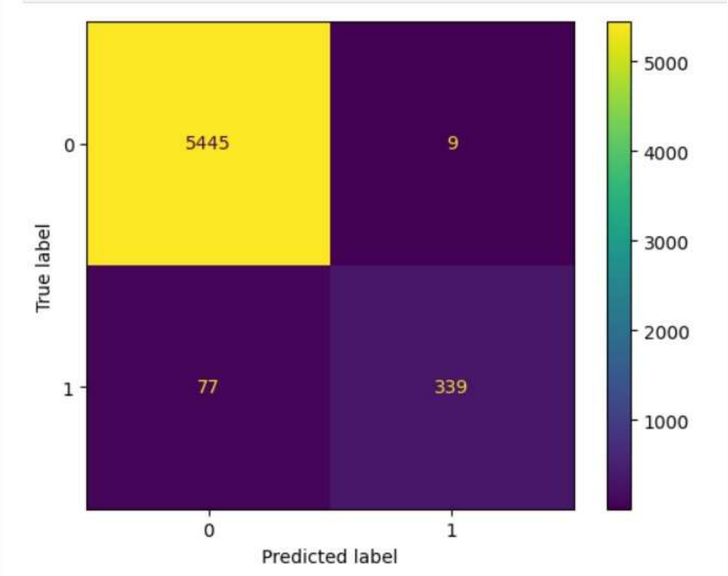
MODEL METRICS OF RANDOM FOREST

Model Score:

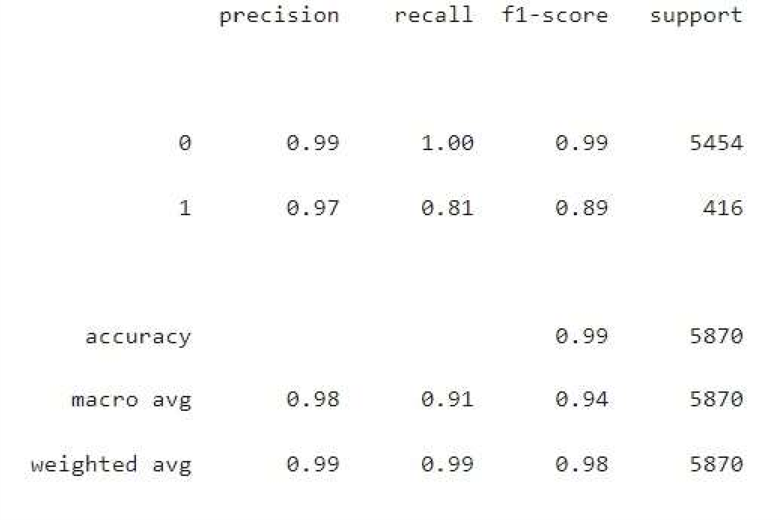
Model Score when using Random Forest : 0.99

An accuracy score of 0.99 often implies good accuracy, but it is essential to precise assessment metrics employed and take other pertinent elements into account before drawing conclusions concerning the model's predictive ability.

Confusion Matrix



Classification Report



```
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      5454
           1       0.97      0.81      0.89       416

    accuracy                           0.99      5870
   macro avg       0.98      0.91      0.94      5870
weighted avg       0.99      0.99      0.98      5870
```

The model is functioning well, according to the results in the image. It excels in identifying positive cases of patients who are in danger. For healthy patients, it could be enhanced by slightly raising the recall and precision.

CONCLUSION

The study evaluated the interpretability and predictive performance of three machine learning algorithms for ECG interpretation in the diagnosis of cardiac disease: Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression (LR).99% accuracy was reached by Random Forest, according to the results, followed by 98% by KNN and 94% by Logistic

Regression. These findings have important ramifications for the medical community, particularly in cardiology, as rapid detection and management of cardiac problems depend on precise and effective ECG interpretation. Medical professionals and researchers can benefit from interpretability analysis's insights into the elements that affect model choices. The work demonstrates the potential of machine learning algorithms for heart illness classification and ECG interpretation, empowering healthcare professionals to make wise decisions, better patient care, and increase diagnostic effectiveness. To increase their relevance and generalizability in actual medical situations, more investigation and validation using larger datasets are required.

SCOPE FOR FUTURE WORK

The data set used can also be used for analysis using other ML algorithms

REFERENCES

- .A New Random Forest Algorithm Based on Learning Automata: Mohammad Savargiv, Behrooz Masoumi, and Mohammad Reza Keyvanpour: Special Issue Interpretation of Machine Learning: Prediction, Representation, Modeling, and Visualization 2021 by Hindawi: Mar 2021
- The random forest algorithm for statistical learning: Matthias Schonlau and Rosie Yuyan Zou: Sage Journals: Mar 2020.
- Medical Image Retrieval via Nearest Neighbor Search on Pre-trained Image Features: Deepak Gupta, Russell Loane, Soumya Gayen, Dina Demner-Fushman: Research Gate, Oct 2022
- A Weighted k -Nearest Neighbours Ensemble With Added Accuracy and Diversity: Naz Gul, Muhammad Aamir, Saeed Aldahmani, Zardad Khan: IEEE Access, Jan 2022
- A Review of the Logistic Regression Model with Emphasis on Medical Research: Ernest Yeboah Boateng, Daniel A. Abaye: Journal of Data Analysis and Information Processing Nov 2019.
- An Improved Algorithm based on KNN and Random Forest: Qin Liu, Jun Liang, Nuihua Nie, Biqing Zeng, Zanbo Zhang. Proceedings of the 3rd International Conference on Computer Science and Application Engineering: Oct 2019
- Usage of KNN, Decision Tree and Random Forest Algorithms in Machine Learning and Performance Analysis with a Comparative Measure: K. Uma Pavan Kumar, Ongole Gandhi, M. Venkata Reddy,vN. Srinivasu: Machine Intelligence and Soft Computing(Book), Jan 2021
- A New Random Forest Algorithm Based on Learning Automata: Mohammad Savargiv, Behrooz Masoumi, Mohammad Reza Keyvanpour: Computational Intelligence and Neuroscience: Mar 2021
- Comparison of machine learning methods for the classification of cardiovascular disease: Rachael Hagan, Charles Gillan, Fiona Mallett: Informatics in Medicine Unlocked, 2021
- Comparative analysis of machine learning KNN, SVM, and random forests algorithm for facial expression classification: Ratna Astuti Nugrahaeni; Kusprasapta Mutijarsa:

International Seminar on Application for Technology of Information and Communication (ISemantic): Aug 2016

- https://www.analyticsvidhya.com/
- https://colab.research.google.com/
- https://bard.google.com/u/5/