IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Heart Disease Prediction Using Machine Learning

Dr. V. Balu [1], K. Varun Reddy[2], N. Reddy Kiran[3]

[1]*ASSISTANT PROFESSOR, Dept. of Computer Science, SCSVMV University, Kanchipuram.*

[2]*Final Year Student, Dept. Of Computer Science, SCSVMV University, Kanchipuram*

Abstract. Heart disease is a major global health issue and is primarily caused by coronary heart disease, which is characterized by plaque build-up in the arteries that supply blood to the heart. Detecting heart disease early is crucial for preventing its progression, and machine learning techniques can aid in this prediction. In this paper, a novel approach that combines Genetic Algorithm and Particle Swarm Optimization (PSO) with neural network algorithms is proposed to improve heart disease prediction. The Cleveland heart disease dataset is augmented with ECG (Electrocardiogram) images to create a hybrid model. Experimental results demonstrate that the proposed method achieves high classification accuracy in predicting heart disease in the Cleveland heart disease dataset with ECG images, surpassing other machine learning methods and existing studies.

Keywords - Machine Learning, Logistic Regression, Support Vector Machines, Naive Byers, Random Forest.

INTRODUCTION

Heart disease prediction is a critical application of machine learning in the healthcare domain. With the widespread use of machine learning in various business applications, including e-commerce and others, the prediction of heart disease using patient datasets has gained significant attention. Our research focuses on developing a machine learning-based approach to predict the likelihood of heart disease occurrence in patients by analyzing their data. This approach involves processing patient datasets and incorporating additional data to enhance prediction accuracy. By leveraging machine learning techniques, we aim to create an innovative solution for heart disease prediction.

The primary focus of this article is to predict heart disease using various classification algorithms, including Genetic Algorithm, PSO Algorithm, and Neural Network Algorithm. Genetic Algorithm and PSO Algorithm are utilized for feature extraction from the datasets, followed by the application of a neural network model to build the prediction model. By incorporating these techniques, we aim to develop an innovative approach for heart disease prediction.

Early detection of heart disease is crucial for improving patient survival rates, as common symptoms such as shortness of breath, swollen feet, and body weakness may not always be evident. To enhance detection, machine learning (ML)-based predictive models have gained support from clinicians as they can aid in clinical decision-making and reduce mortality rates. ML has been utilized to analyze electronic health records and uncover hidden relationships in the data, enabling clinicians to detect heart disease risk and provide appropriate treatments and recommendations. While numerous heart disease datasets and predictive models have been

developed, achieving high prediction performance, and identifying relevant risk factors remains challenging for researchers and scientists. Given the increasing prevalence and mortality rate of heart disease, accurate and timely diagnosis is crucial. To address this, ML research studies have focused on developing data preprocessing and transformation techniques to improve classification performance.

EXISTING SYSTEM AND ITS DRAWBACKS

The diagnostic process often involves a method of elimination or gathering information to narrow down the probability of candidate conditions to negligible levels. This method typically comprises four steps:

- The doctor collects all relevant information about the patient's symptoms and creates a comprehensive symptoms list.
- The doctor generates a list of all possible causes of the symptoms based on the available information.
- The doctor prioritizes the list, placing the most dangerous potential cause of the symptoms at the top.
- The doctor systematically rules out or treats the possible causes, starting with the most urgent and dangerous conditions. The term "rule out" refers to the use of appropriate tests or scientific methods to confirm or eliminate a diagnosis. If a diagnosis is ruled out, it is removed from the list, and tests with distinct results are used to determine the correct diagnosis based on the doctor's knowledge and experience. This method is straightforward to implement and can aid in accurate diagnosis and treatment decisions.

PROBLEM STATEMENT

Medical decision-making can be complex and challenging, particularly in cases involving rare diseases or when there are multiple alternative factors to consider, such as stress or incomplete information. Standard algorithms may need to analyze numerous variables, including the patient's medical history, family records, prevailing conditions, and other hidden factors. Differential diagnosis methods can be employed to identify the presence of a specific condition when multiple alternatives are possible, and these methods may also incorporate candidate alternatives. Misdiagnosis can result from factors such as lack of knowledge on the part of doctors or incomplete information, which may contribute to errors in the decision-making process.

PROPOSED METHODOLOGY

Heart disease is a prevalent condition that requires effective prediction methods. In this study, we aim to develop a hybrid model by combining the Cleveland heart disease dataset with ECG images to improve prediction accuracy. We utilize Genetic Algorithm and PSO algorithm to extract relevant features from the dataset efficiently. Subsequently, a neural network algorithm is employed to build the prediction model. The model's accuracy is evaluated using test data. The advantages of this approach include efficient feature extraction using genetic algorithms & PSO algorithms, and accurate prediction using neural network algorithms.

METHODOLOGY

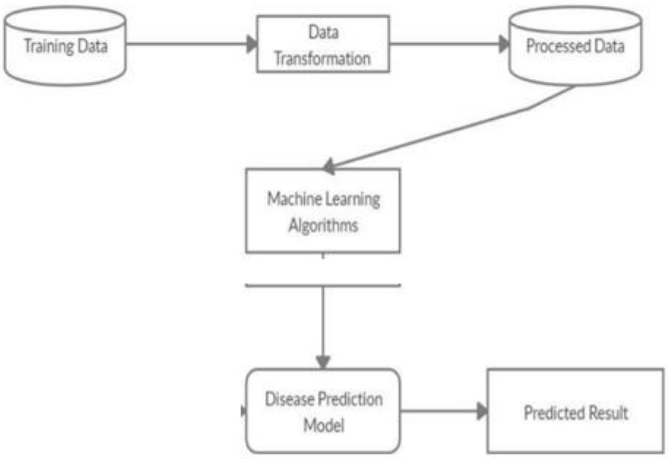
Data Preprocessing: The process of cleaning and formatting data is crucial in ensuring the quality of machine learning models. This is because a machine learning model's accuracy and efficiency heavily depend on the quality of data it is trained on. Therefore, it is essential to identify and address inaccuracies and errors within the dataset before building any models. Cleaning data involves tasks such as removing duplicates, filling in missing values, removing irrelevant data, and correcting data inconsistencies. Once the data has been cleaned, it can be pre-processed to extract useful features, transform the data into a suitable format, and standardize it. This process of cleaning and formatting data is a critical step in ensuring the quality and effectiveness of machine learning models.

Data Cleaning: Ensuring that the dataset is free from errors and missing data is crucial for any machine learning model to function accurately. One common issue that often arises is the presence of NA values, which can significantly reduce the accuracy of our predictions. To address this issue, we can perform various operations, such as removing fields with missing values or substituting them with the mean value of the column. By doing so, we can improve the overall quality of the dataset and increase the accuracy and efficiency of our machine learning model. Therefore, it is necessary to ensure that the dataset is clean and formatted properly before feeding it to a machine learning algorithm.

Feature Scaling: Machine learning algorithms often require feature scaling to ensure accurate predictions. This is because raw data can vary widely in its range of values, and objective functions in some algorithms may not work properly without scaling. For example, classifiers may calculate distance between two points using Euclidean distance, but if one feature has a much larger range of values than the others, it will dominate the distance calculation. To ensure that each feature contributes proportionately to the final calculation, it is necessary to scale the range of values of all features. In our dataset, some fields such as Age have only two values (0, 1) while others such as cholesterol have much higher values. Therefore, scaling is necessary to bring all features closer together in terms of values. This will help to improve the accuracy and efficiency of the machine learning model.

Factorization: In this section, we added a categorical meaning to the values to make them more interpretable by the machine learning algorithm. For instance, we assigned "male" and "female" labels to the gender column instead of using 0 and 1. By doing this, the algorithm can better understand the relationship between different variables and make more accurate predictions. It also helps in preventing the algorithm from treating numerical values as continuous variables, which may lead to incorrect assumptions and predictions. By assigning specific meanings to the values, we are improving the quality and usefulness of the data for the machine learning model.

Support Vector Machine: In our project, we used SVM to classify patients into different risk groups based on the provided parameters. This approach was chosen because it was found to have the highest accuracy when compared to other classification algorithms such as Naïve Bayes and logistic regression. In our experiment, Naïve Bayes had an accuracy of 60%, logistic regression had 61.45%, and SVM had an accuracy of 64.4%. Therefore, we selected SVM as the most efficient algorithm for our web application. It is important to note that when the data is not labeled, unsupervised learning techniques are required to group the data based on their natural clusters.



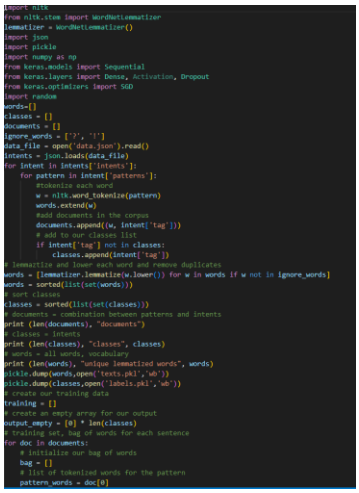
1. Architecture

IMPLEMENTATION

To apply the heart complaint vaticination model, follow the way below.

- Install the necessary tools and import the applicable libraries in your Python terrain.
- Launch the operation Python train in any source law editor.
- Upload the ECG and heart dataset into the operation.
- Train the machine literacy model to read and reuse the dataset using colorful algorithms.
- Compare the performance of different algorithms and select the one with the loftiest delicacy as the stylish model.
- further train the named high-delicacy algorithm to optimize its performance.
- Run the web operation and open it in a cyber surfed.
- Allow druggies to input their data as per the operation's conditions.
- Submit the data and let the operation use the pre-trained model to dissect and prognosticate the affair.
- Display the prognosticated affair in a new window on the webpage.

WORKING



```
import nltk
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
import json
import pickle
import numpy as np
from keras.models import Sequential
from keras.layers import Dense, Activation, Dropout
from keras.optimizers import SGD
import random
words=[]
classes = []
documents = []
ignore_words = ['?', '!']
data_file = open('data.json').read()
intents = json.loads(data_file)
for intent in intents['intents']:
    for pattern in intent['patterns']:
        #tokenize each word
        w = nltk.word_tokenize(pattern)
        words.extend(w)
        #add documents in the corpus
        documents.append((w, intent['tag']))
        # add to our classes list
        if intent['tag'] not in classes:
            classes.append(intent['tag'])
# lemmatize and lower each word and remove duplicates
words = [lemmatizer.lemmatize(w.lower()) for w in words if w not in ignore_words]
words = sorted(list(set(words)))
# sort classes
classes = sorted(list(set(classes)))
# documents = combination between patterns and intents
print (len(documents), "documents")
# classes = intents
print (len(classes), "classes", classes)
# words = all words, vocabulary
print (len(words), "unique lemmatized words", words)
pickle.dump(words,open('texts.pkl','wb'))
pickle.dump(classes,open('labels.pkl','wb'))
# create our training data
training = []
# create an empty array for our output
output_empty = [0] * len(classes)
# training set, bag of words for each sentence
for doc in documents:
    # initialize our bag of words
    bag = []
    # list of tokenized words for the pattern
    pattern_words = doc[0]
```

Fig 1. Prediction code

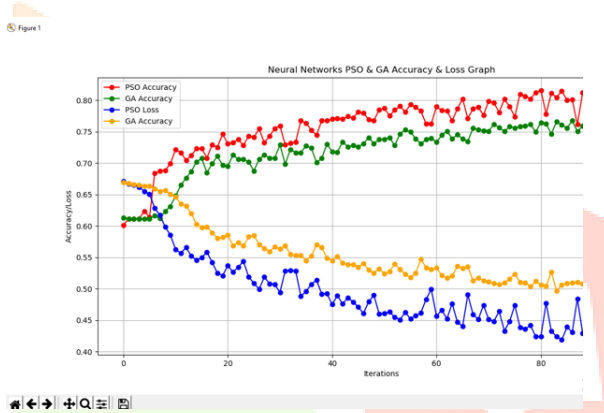


Fig 2. Neural Networks PSO & GA Accuracy & Loss Graph

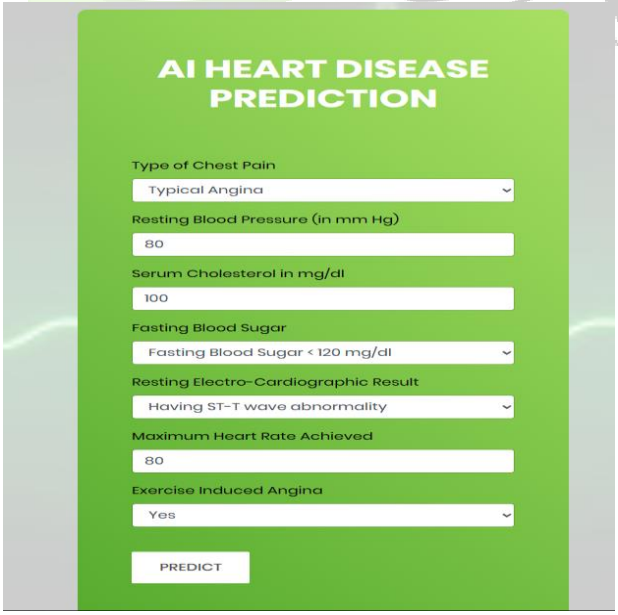


Fig 3. Output 1

CONCLUSION

Given the rising number of losses attributed to heart complaints, it has come imperative to develop a robust and accurate system for prognosticating this condition. The primary provocation of this study was to identify the most effective machine literacy (ML) algorithm for heart complaint vaticination. To achieve this, the study compared the delicacy scores of four generally used ML algorithms- Decision Tree, Logistic Retrogression, Random Forest, and Naive Bayes- using a dataset from the UCI machine literacy depository. Grounded on the findings, the Random Forest algorithm surfaced as the most effective, achieving a delicacy score of 90.16 for prognosticating heart complaints. In unborn exploration, the work could be further enhanced by developing a web operation grounded on the Random Forest algorithm and employing a larger dataset for more accurate results. This could greatly help healthcare professionals in prognosticating heart complaints more effectively and efficiently, potentially saving lives and mollifying the impact of this wide health concern.

FUTURE ENHANCEMENT

The system being developed has a wide range of implicit benefits and features, including.

- robotization of complaint opinion, which can streamline the process and reduce mortal error.
- Relinquishment of a paperless approach, contributing to environmental sustainability.
- bettered effectiveness and delicacy in prognosticating and managing conditions for the benefit of cases.
- Effective operation of information related to conditions, leading to better data association and availability.
- perpetration of a chatbot, which can enhance stoner experience and give precious support in navigating the system.

These features punctuate the significant eventuality and compass of the system in delivering effective healthcare results.

REFERENCES

[1] Mihir Patel, Rahul Patange, Chaitanya Patil, Prof. Anuradha Kapoor, “Predicting Heart Disease Using Machine Learning Algorithms,” International Research Journal of Engineering and Technology (IRJET) Volume: 09 Issue: 04, Apr 2022.

[2] J. S. Rumsfeld, K. E. Joynt, and T. M. Maddox, “Big data analytics to improve cardiovascular care: promise and challenges”, Nature Reviews Cardiology, Vol.13, No.6, pp.350, 2016.

[3] M. Tabrez Quasim, S. Alhuwaimel, A. Shaikh, Y. Asiri, K. Rajab et al., "An improved machine learning technique with effective heart disease prediction system," Computers, Materials & Continua, vol. 69, no.3, pp. 4169–4181, 2021.

[4] A. Lakshmana Rao, Y. Swathi, P. Sri Sai Sundareswar, “Machine Learning Techniques for Heart Disease Prediction,” International Journal Of Scientific & Technology Research Volume 8, Issue 11, November 2019.

[5] Fahd Saleh Alotaibi,” Implementation of Machine Learning Model to Predict Heart Failure Disease”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.

[6] Nagaraj M Lutimath, Chethan C,Basavaraj S Pol.,’Prediction Of Heart Disease using Machine Learning’, International journal Of Recent Technology and Engineering,8,(2S10), pp 474-477, 2019.

[7] Jafar Alzubi, Anand Nayyar, Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview", Journal of Physics: Conference Series, 2018

[8] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, ―Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field,‖ in Machine Learning Paradigms, 2019, pp. 71–99.

[9] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications", Journal of Intelligent Learning Systems and Applications, Vol.9, No.01, pp.1, 2017.