IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Heart Disease Prediction Using Machine Learning

Vedant Bhati, Jatin Sharma
Prof. (Dr.) Shallu Bashambu Prof.(Dr) Bhaskar Kapoor
IT Department
Maharaja Agrasen Institute of Technology, Rohini Sector-22, New Delhi, India

ABSTRACT

In various fields around the world, machine learning is used. There are no exceptions in the healthcare sector. Machine learning can be crucial in determining whether or not there will be locomotor abnormalities, heart ailments, and other conditions [8]. If foreseen far in advance, such information can offer crucial intuitions to doctors, who can then modify their diagnosis and approach per patient [7]. We are attempting to use machine learning algorithms to predict potential heart conditions in humans. In this project, we compare the performance of various classifiers, including Logistic Regression [9]. We also propose an ensemble classifier that performs hybrid classification by combining the best features of both strong and weak classifiers because it can use a large number of training and validation samples. In various fields around the world, machine learning is used [11]. There are no exceptions in the healthcare sector. Machine learning can be crucial in determining whether or not there will be locomotor abnormalities, heart ailments, and other conditions [3]. If foreseen far in advance, such information can offer crucial intuitions to doctors, who can then modify their diagnosis and approach per patient. We are attempting to use machine learning algorithms to predict potential heart conditions in humans. In this project, we compare the performance of various classifiers, including Logistic Regression . We also propose an ensemble classifier that performs hybrid classification by combining the best features of both strong and weak classifiers because it can use a large number of training and validation samples

INTRODUCTION

The World Health Organization estimates that heart disease causes 12 million deaths worldwide each year. One of the leading causes of morbidity and mortality among the global population is heart disease. One of the most crucial topics in the data analysis area is predicted cardiovascular disease[1]. Since a few years ago, the prevalence of cardiovascular disease has been rising quickly throughout the world. Many studies have been carried out in an effort

to identify the most important risk factors for heart disease and to precisely estimate the overall risk. Heart disease is also referred to as a silent killer because it causes a person to pass away without any evident signs. Cardiovascular disease must be detected early. aiding high-risk patients in making decisions regarding lifestyle changes will help to reduce the difficulties. Making choices and predictions from the vast amounts of data generated by the healthcare sector is made easier with the help of machine learning.[3]By evaluating patient data that uses a machine-learning algorithm to categorize whether a patient has heart disease or not, this study hopes to predict future cases of heart disease. Machine learning methods can be extremely helpful in this situation. There is a common set of basic risk factors that determine whether or not someone will ultimately be at risk for heart disease, despite the fact that heart disease can manifest itself in various ways. By gathering information from numerous sources, organising it into categories that make sense, and then performing analysis to get out the desired information based on statistics, we may conclude that this technique is quite adaptable[8]. The main difficulty with heart disease is detecting it. There are tools that can forecast heart disease, but they are either expensive or ineffective at calculating the likelihood of heart disease in a human. The mortality rate and total consequences can be reduced by early identification of heart disorders. Since it takes more intelligence, time, and knowledge, it is not always possible to accurately monitor patients every day, and a doctor cannot consult with a patient for a whole 24 hours. As there is a lot of data available nowadays, we can use a variety of machine learning methods to search for hidden patterns. The underlying patterns may utilised in medical data for health diagnosis

LITERATURE SURVEY

There is ample related work in the fields directly related to this paper. ANN has been introduced to produce the highest accuracy prediction in the medical field . The back propagation multilayer perception (MLP) of ANN is used to predict heart disease. The obtained results are compared with the results of existing models within the same domain and found to be improved[1] . The data of heart disease patients collected from the UCI laboratory is used to discover patterns with NN, DT, Support Vector machines SVM, and Naive Bayes. The results are compared for performance and accuracy with these algorithms. The proposed hybrid method returns results of 86.8% for F - measure, competing with the other existing methods . The classification without segmentation of Convolutional Neural Networks (CNN) is introduced. [10]This method considers the heart cycles with various start positions from the Electrocardiogram (ECG) signals in the training phase. CNN is able to generate features with various positions in the testing phase of the patient . A large amount of data generated by the medical industry has not been used effectively previously. The new approaches presented here decrease the cost and improve the prediction of heart disease in an easy and effective way. The various different research techniques considered in this work for prediction and classification of heart disease using ML and deep learning (DL) techniques are highly accurate in establishing the efficacy of these methods . Prediction of heart disease using data mining techniques has been an ongoing effort for the past two decades. [13] Most of the papers have implemented several data for diagnosis of heart disease such as Decision Tree, Naive Bayes, neural network, kernel density, automatically defined groups, bagging algorithm and support vector machine showing different levels of accuracies (Yan, Zheng et al. 2003; Andreeva 2006; Das,

Turkoglu et al. 2009; Sitar-Taut, Zdrenghea et al. 2009; Raj Kumar and Reena 2010; Srinivas Rani et al. 2010) on multiple databases of patients from around the world mining techniques

Implementation Study

Heart disease is even being emphasised as a silent killer that causes a person to pass away without showing any outward signs. Growing concern about the illness and its effects is a result of the disease's nature [4]. Thus, efforts to foresee the potential occurrence of this fatal disease in the past continue. a group of datasets For the foundation of our heart disease prediction system, we first gather a dataset. We divided the dataset into training and testing data after it was collected. The learning of the prediction model takes place on the training dataset, and the evaluation of the prediction model occurs on the testing dataset [6]. 30% of the data are utilised for testing in this project, while 70% are used for training. The information gathered for this project is UCI Heart Disease. The dataset has 76 properties, of which the system uses 14 for its operation. Selection of attributes The choice of acceptable attributes for the prediction system is included in attribute or feature selection [7]. This is done to make the system more effective. For the prediction, a number of patient characteristics are used, including gender, the nature of the patient's chest discomfort, fasting blood pressure, serum cholesterol, and exang. Pre-processing of Data The pre-processing of data is a critical stage in the development of a machine learning model.[9] Data that isn't initially clean or in the model's required format can lead to inaccurate results. Pre-processing involves transforming data into the format we need. It is used to handle the dataset's noise, duplication, and missing values. Activities like importing datasets, partitioning datasets, attribute scaling, etc. are all part of data pre-processing. Preprocessing the data is necessary to increase the model's accuracy. Balancing of Data Imbalanced datasets can be balanced in two ways. They are Under Sampling and Over Sampling (a) Under Sampling: In Under Sampling, dataset balance is done by the reduction of the size of the ample class. This process is considered when the amount of data is adequate. (b) Over Sampling: In Over Sampling, dataset balance is done by increasing the size of the scarce samples. This process is considered when the amount of data is inadequate. Prediction of Disease For classification, a variety of machine learning algorithms are employed, including SVM, Naive Bayes, Decision Trees, Random Trees, Logistic Regression, Ada-boost, and Xg-boost. Algorithms are compared, and the one that predicts heart disease with the best degree of accuracy is chosen. This is a multivariate type of dataset which means providing or involving a variety of separate mathematical or statistical variables, multivariate numerical data analysis. It is composed of 14 attributes which are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, oldpeak — ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels and Thalassemia. This database includes 76 attributes, but all published studies relate to the use of a subset of 14 of them. The Cleveland database is the only one used by ML researchers to date. One of the major tasks on this dataset is to predict based on the given attributes of a patient that whether that particular person has heart disease or not and other is the experimental task to diagnose and find out various insights from this dataset which could help in understanding the problem more.

PROPOSED WORK AND ALGORITHM

The collection of data and selection of the most crucial attributes is the first step in the system's operation. The relevant data is then preprocessed into the format needed[1]. After that, the data is split into training and testing data. The algorithms are used, and the training data is used to train the model[3]. By testing the system with test data, the correctness of the system is determined. The modules listed below are used to implement this system.

1. Collection of Dataset
2. Selection of attributes
3. Data Pre-Processing
4. Balancing of Data
5. Disease Prediction

Implementation

Data Collection:

Begin by acquiring a diverse dataset of animal images, specifically focusing on those afflicted with heart disease. Categorize the images as "normal" and "diseased." Ensure the dataset is comprehensive, accurately labeled, and representative of various heart disease. Sufficiently populate each category to promote effective model training. We took data from UCI Heart Study.

Data Augmentation:

Introduce data augmentation techniques to enrich the dataset. Apply random rotations, zooming, horizontal or vertical flipping, and shifting to generate new images. Data augmentation enhances the model's ability to generalize to unseen data, reducing overfitting.

Model Selection:

Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class or not. It is a kind of statistical algorithm

Model Evaluation:

Assess the model's performance using metrics like accuracy, precision, recall, and F1 score. Employ a dataset split into training and testing sets to gauge the model's generalization capabilities. Consider incorporating optimization techniques such as the Adam optimizer for robust performance.

Model Deployment:

Save the trained model, encompassing its learned weights and architecture, for future use. This facilitates deployment without the necessity for retraining, streamlining its application for inference tasks.

Algorithm

Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class. It is used for classification algorithms its name is logistic regression. it's referred to as regression because it takes the output of the linear regression function as input and uses a sigmoid function to estimate the probability for the given class. The difference between linear regression and logistic regression is that linear regression output is the continuous value that can be anything while logistic regression predicts the probability that an instance belongs to a given class or not.

Logistic Regression:

It is used for predicting the categorical dependent variable using a given set of independent variables.

- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value.
- It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

Logistic Function (Sigmoid Function):

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1. o The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form.
- The S-form curve is called the Sigmoid function or the logistic function.

- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Type of Logistic Regression:

On the basis of the categories, Logistic Regression can be classified into three types:

1. **Binomial:** In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
2. **Multinomial:** In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
3. **Ordinal:** In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

Conclusion and Future Work

Application of promising technology, such machine learning, to the first prediction of heart problems would have a significant social impact because heart diseases are a leading cause of death in India and around the world. Early detection of cardiac disease can help high-risk patients make decisions about lifestyle modifications that will lessen problems, which can be a significant advancement in the field of medicine. Each year, more people are diagnosed with cardiac illnesses. This calls for an early diagnosis and course of action. The medical community as well as patients may benefit greatly from the use of appropriate technology support in this area. The seven machine learning algorithms employed in this study to gauge performance , applied to the dataset along with Logistic Regression. The dataset, which includes 76 features, contains the expected characteristics that contribute to heart disease in individuals, and 14 significant characteristics are chosen from them to help assess the system. If all the features are taken into account, the creator receives a less efficient system. Attribute selection is carried out to improve efficiency. In this case, n characteristics must be chosen in order to evaluate the model that provides greater accuracy. Several dataset features have virtually equal correlations, so they are eliminated. When all of the dataset's qualities are taken into consideration, the effectiveness drops significantly. A prediction model is created after comparing the accuracy of each of the seven machine learning techniques. So, the objective is to employ a variety of evaluation metrics, such as the confusion matrix, accuracy, precision, recall, and f1-score, which accurately predicts the disease. The extreme gradient boosting classifier has the highest accuracy (81%), when all seven are compared.

References

[1] Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43- 8

[2] Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-8.

[3] Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine, 10(2), 334-43.

[4] Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease prediction system using k-means clustering and Naive Bayes algorithm. International Journal of Computer Science and Information Technologies, 6(1), 637-9.

[5] Bashir S, Qamar U & Javed M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In International Conference on Information Society (i-Society 2014) (pp. 259-64). IEEE. ICCRDA 2020 IOP Conf. Series: Materials Science and Engineering 1022 (2021) 012072 IOP Publishing doi:10.1088/1757- 899X/1022/1/012072 9

[6] S. Palaniappan and R. Awang, "Intelligent heart disease prediction systemusing data mining techniques," pp. 108–115, 2008.

[7] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, "Hybrid intelligent modelling schemes for heart disease classification," Applied Soft Computing,vol. 14, pp. 47–52, 2014.

[8] M. Shouman, T. Turner, and R. Stocker, "Using data mining techniquesin heart disease diagnosis and treatment," pp. 173–177, 2012.;3

[9] P. V. Ankur Makwana, "Identify the patients at high risk of re-admissionin hospital in the next year," International Journal of Science andResearch, vol. 4, pp. 2431–2434, 2015.

[10] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, "Computationalintelligence for heart disease diagnosis: A medical Knowledge driven approach," Expert Systems with Applications, vol. 40, no. 1, pp. 96–104,2013.

[11] Y. Xing, J. Wang, Z. Zhao, and Y. Gao, "Combination data miningmethods with new medical data to predicting outcome of coronary heartdisease," pp. 868–872, 2007.

[12] Combination data mining methods with new medical data to predictingoutcome of coronary heart disease," in Convergence InformationTechnology, 2007. International Conference on. IEEE, 2007, pp. 868–872. [9] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, "Hybrid intelligent modelling, schemes for heart disease classification," Applied Soft Computing,vol. 14, pp. 47–52, 2014.

[13] https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data