**A Study On the Determination of Heart Disease Using Machine Learning Algorithm**

**Abstract**

Nowadays, with the rapid improvement of technology, artificial intelligence techniques are also widely utilised. Artificial intelligence methods are commonly used in many fields such as engineering applications, education, defence industry. Machine learning is the Artificial Intelligence (AI) branch that is one of the essential areas of use in the health sector. In this study carried out in the field of the health sector, the data set collected from an open-access website (kaggle) was used. Determine of the accuracy score was carried out using machine learning methods on the data set. It has been used three machine-learning algorithms are Logistic Regression (LR), K Nearest Neighbour algorithm (KNN), and Support Vector Machine (SVM). KNN is the most successful algorithm accuracy score of %99. Moreover, SVM is the second most successful algorithm accuracy score of %96. On the other hand, LR is the gets lowest accuracy score of %81.

**Keywords—** *Machine Learning, supervised and unsupervised learning, Logistic Regression, K Nearest Neighbour algorithm, Support Vector Machine*.

1. **Introduction**

In the last two decades, heart disease rates have dramatically risen to become the leading cause of death. Heart disease, called cardiovascular disease, occurs with a chain of problems that happen due to the accumulation of various substances on the inner wall of the vessels and cause plaque formation. This chain of issues, called atherosclerosis, causes narrowing of the heart arteries, making it difficult for the blood to flow through the vessel (Benjamin & Muntner, 2019, et al.).

Heart diseases that occur this way are among the leading causes of death worldwide; moreover, World Health Organization(who) reported that 17,9 million people die each year from heart disease, which correlates with 31% of all deaths worldwide (WHO, 2017). However, while the number of deaths from heart disease decreases in high-income countries, death occurs in many low- and middle-income countries (Wang, et al., 2021).

Early detection of the disease and appropriate treatment play an important role in preventing deaths and reducing the risk of heart diseases. Therefore, artificial intelligence applications and different classification methods are increasing in detecting heart diseases (Kishor & Chinmay , 2021). The classification methods are used to ensure that correct results are obtained, the time required to detect the disease is minimised, and human-induced errors are prevented.

Machine learning is the artificial intelligence (AI) branch that gives systems the capability to learn and develop from knowledge without being explicitly programmed automatically. Machine learning focuses on developing computer programs that can reach data and use it to learn for themselves (Bonaccorso, 2017).

Learning progress starts with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and create better decisions in the future based on the standards that developers provide. The main goal is to allow the computers to learn automatically without anyone intervention or support and adjust efforts accordingly.

As the definition of machine learning, it learns from the natural phenomenon, natural things. Therefore in this project, The data set gets from an open-access website (kaggle) it occurs 14 columns and 1025 rows. The "target" field points to the presence of heart disease in the patient. It is integer-valued 0 = no disease and 1 = disease. Biological parameters have been used as testing data such as cholesterol, blood pressure, sex, age, etc. According to these, the comparison is made in terms of the accuracy of algorithms. This project has used Logistic Regression (LR), K Nearest Neighbour algorithm (KNN), and Support Vector Machine (SVM).

This project aims to calculate the accuracy score using the three different machine-learning approaches on the heart disease dataset. It will be calculated to conclude which one is best among them.

Section 1 of this paper occurs to the introduction and background to machine learning and heart diseases. Section 2 demonstrates the related work on the previous research. Section 3 introduces the intricate details of the dataset, data pre-processing, data visualisation, data preparation for training, validation and testing, and machine learning model for classification. Section 4 presents the results of each model, along with the overall accuracy of the model. Section 5 concludes this paper with a slight view of the future scope of this paper.

1. **Related work**

The heart is one of the vital organs of the human body and plays a significant role in pumping blood throughout the human body, so it always needs protection. That is why it is one of the biggest reasons researchers work on it; many researchers are working on it. There is always a need to analyse heart-related data, and then the early diagnosis is vital to prevent heart disease. Various fields such as artificial intelligence, machine learning, and data mining contribute to this work.

In their study, Palaniappan and Awang (2008) developed an Intelligent Heart Disease Prediction System IHDPS prototype using machine learning methods: Decision Trees, Naïve Bayes, and Neural Network. The probability of having heart disease is estimated by taking into account the characteristics of people with heart disease, such as age, gender, blood pressure and blood sugar. According to the results obtained, decision trees provided convenience in terms of reading and interpreting the result. At the same time, only decision trees detailed access to patient characteristics. The Naïve Bayes method performed better than the decision tree method, as it was able to identify all essential medical predictions.

In the study conducted by Das and Sengur (2009), the 9.1.3 version of the Statistical Analysis System (SAS) software was used to detect heart diseases. New models were created by associating the values found in the system, which is based on a neural network ensemble method, with more than one previous model. Three separate neural network models were used to predict heart diseases in a fully automated manner. In the data obtained in the database, 89.01% classification accuracy, 80.95% specificity and 95.91% sensitivity values were obtained.

Shao et al. (2014) proposed a hybrid intelligent modelling scheme that effectively classifies heart disease in their study. A data set created by actual patients was used to show the development of the model. The proposed hybrid modelling consists of logistic regression (LR), multivariable adaptive regression splines (MARS), Artificial Neural Network (ANN) and Rough Cluster (RS) techniques. Firstly, LR, MARS and RS techniques were used to reduce the explanatory variable set in the study. In the next step, the remaining variables are assigned as inputs for the use of the ANN method. The modelling results showed that the recommended hybrid schemes efficiently classify heart disease and outperform the typical, single-stage ANN method.

In their study, Singh et al. (2016) discussed various machine learning methods using the Cleveland heart diseases dataset from the University of California Irvine (UCI). Different characteristics (age, gender, type of chest pain, blood pressure, blood sugar, etc.) that cause heart disease were used in the study. It has been tried to predict heart diseases by using different machine learning methods such as Aprior, FP-Growth, Naive Bayes, ZeroR, OneR, J48 and k-NN. It has been observed that the K-NN method gives better results than the others.

Furthermore, Priyanka and Kumar (2017) compared Naïve Bayes & Decision Tree techniques, one of the data mining methods, to increase the prediction accuracy of heart disease in their study. For the prediction system, 13 features of heart disease were considered. As a result of the study, it was observed that the decision tree gave more precise results in the prediction of heart disease. According to this study Decision, Tree accuracy score is (%98.03,%98.21,%90) and the Naïve Bayes accuracy score is (%82.35,%82.14,%70).

Finally, Kohali et al. (2018) researched classification algorithms, each with its advantage on three separate disease databases (Heart, Breast cancer, Diabetes). According to this research, the prediction accuracy of our offered method reaches 98.57% using AdaBoost classifier for Breast Cancer detection, 87.1% in heart disease detection using Logistic Regression, and 85.71% in Diabetes prediction using Support Vector Machine (linear kernel).

1. **Materials and Methods**
2. ***Selection of Attribute***

The 'Public Health Dataset' data set of the Kaggle database, an open-source internet site, was used in the study to detect heart disease. The data set used consists of 1025 data, 499 of which are not heart disease patients, and 526 are heart disease patients. There are 14 attributes in the received data set. The "target" field points to the presence of heart disease in the patient. It is number valued 0 = no disease and 1 = disease. Below Table.1 show those attributes.

|  |  |
| --- | --- |
| Attributes | Description |
| Age | Patient’ s age 29 to 77 |
| Sex | Gender of patient 0=female, 1= male |
| Cp (Chest pain) | 1=typical angina, 2=atypical angina, 3=non-anginal pain |
| Trestbps (Resting blood pressure) | in mm Hg on admission to hospital, values from 94 to 200 |
| Chol(cholesterol) | Serum cholesterol in mg/dl, values from 126 to 564) |
| Fbs | Fasting blood sugar > 120 (mg/dl): 1=true 0=false |
| Restecg | 0=normal, 1=having ST-T abnormality, 2= probable left ventricular hypertrophy |
| Thalach | Maximum heart rate achieved 71 to 202 |
| Exang | Exercise-induced angina: 1=yes, 0=no |
| Oldpeak | ST depression induced by exercise relative to rest |
| Slope | Peak exercise slope segment: 1=up sloping 2=flat 3=down sloping |
| Ca | Number of significant vessels (0-3) |
| Thal | thal: 0 = normal; 1 = fixed defect; 2 = reversable defect |
| Targets | 1= heart disease, 0 = not heart disease |

1. ***Data pre-processing***

Data pre-processing meaning is the process of cleaning raw data and converting it into a readable format. Data mining is also significant because it is challenging to work with raw data, and accurate results are not obtained (García, et al., 2015). Before implementing machine learning or data mining algorithms, the quality of the data should be checked, for example, whether there are any null values. In this study, it has been inspected, and there are not any null values. Also, it has been checked data type; there are 13 integers and one float number.

1. ***Data Visualization***

Data visualisation is the process of converting large data sets into graphs, charts, and other visuals. The visual description of the resulting data performs it straightforward to identify and share real-time trends, outliers, and new insights about the details represented in the data (Engebretsen & Kennedy, 2020).

Firstly, data balancing is crucial for accurate results; the below graph shows that this project target class Fig.1 represent target classes 0 no heart disease and 1 heart disease patient. As seen in this dataset balancing graph, it can seem that both the target classes are nearly equal.

Chart, bar chart

Description automatically generated

*Fig.1 Target Class View*

Secondly, it can see all data in Fig.2 by Histograms, which helps predict where values are concentrated, their maximum and minimum points break between data, and whether there are exceptional values. They are also helpful in that they give a general impression of the probability distribution.

Chart, box and whisker chart

Description automatically generated

*Fig.2 Attributes Histogram*

Finally, it can see in Fig.3 Heatmap, which helps cross-examine multivariate data by placing variables in rows and columns and colouring cells in the table. Heatmaps can be used to show variance among multiple variables, reveal any design, show whether any variable is like each other, and detect any correlation between them.

Chart, treemap chart

Description automatically generated

*Fig.3 Cross-correlation values heat map*

1. **Machine Learning**

Machine learning involves computers discovering how to perform tasks without being explicitly programmed. It covers computers that learn from data provided to perform specific tasks (Surden, 2014). For simple tasks assigned to computers, it is possible to program algorithms that instruct the machine how to complete all the steps required to solve the problem at hand; There is no need to learn on the computer side. In more advanced tasks, it can be challenging to create the necessary human algorithms manually.

In practice, it may be more effective to help the machine develop its algorithm rather than have human programmers determine every essential step. There are three types of machine learning algorithms.

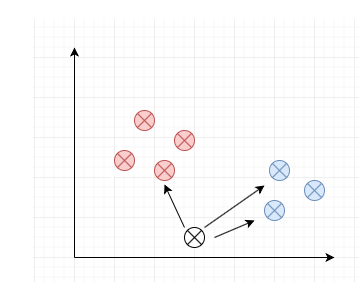
***Supervised learning***: Generates a function that maps inputs to target outputs. In this type of learning, algorithms use labelled data to make predictions based on what they have learned. The data to be used in training is known in advance. With this information, the system learns and interprets the new data. Algorithms use labelled data to make predictions based on what they have learned. The data to be used in training is known in advance. With this information, the system learns and interprets the new data. They are used in classification and regression problems.

***Unsupervised learning***: Models a set of inputs. It provides inferences about the data by using the distances of the data samples from each other and the neighbourhood relations.

***Reinforcement learning*** is a type of machine learning that aims to train the agent (robot, vehicle) with the reactions it receives from the environment without training data. Every action creates an effect in the background, and the setting gives feedback in the form of rewards that guide the learning algorithm.

Therefore, machine learning algorithms, divided into training and testing, provide feedback in connection with the accuracy of the predictions. In this project, the data was split into training and testing; it was set into a ratio of 70% training data and 30% testing data. And then it has been chosen three classification models are: K-Nearest Neighbor, Logistic Regression, Support Vector Machines (SVM).

1. ***K-Nearest Neighbor:*** The k-nearest neighbor (KNN) algorithm is a supervised learning algorithm that is easy to implement. Although it is used in solving both classification and regression problems, it is mainly used in solving classification problems in industry. Figure.4 shows the KNN sample drawing

******

*Fig.4 KNN chart*

First, the parameter K is determined. This parameter is the number of nearest neighbors to a given point. According to the existing data, the distance of the new data, which will be added to the sample data set, is calculated one by one, with the help of the corresponding distance functions. The K nearest neighbors of the related distances are considered. It is assigned to the class of k neighbors or neighbors according to the attribute values.

The selected class is considered to be the class of the observation value expected to be estimated. In other words, the new data is labelled.

1. ***Logistic Regression:*** Logistic Regression (LR) is a regression method for classification. It is used to classify categorical or numerical data. It works when the dependent variable can only take two different values. For this reason, in logistic regression, the dependent variable contains data encoded 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, error, non-pregnant, etc.). Figure.5 shows the LR sample drawing.

***Chart, scatter chart

Description automatically generated***

*Fig.5 logistic regression chart*

1. ***Support Vector Machines (SVM):*** Support Vector Machine is one of the supervised learning methods generally used in classification problems. In this algorithm, each data item is plotted as a point in the n-dimensional space (where n is the number of features your data has), with the value of each feature being the value of a particular coordinate. It is used to separate data belonging to two classes in the most appropriate way. Figure.6 shows the SVM sample drawing.

***Chart, scatter chart

Description automatically generated***

*Fig.6 Support Vector Machines chart*

**4. Result and Analysis**

In this study, in the analysis part, firstly, it will show the test accuracy score made by default parameters of machine learning algorithms. Secondly, it will show the tuned best accuracy score with GridsearchCV. Finally, it is going to deliver a confusion matrix. A confusion matrix is a table often used to describe the performance of a classification model on a set of test data for which the actual values are known. fig.6 shows the confusion matrix below.

Graphical user interface, application

Description automatically generated

*Fig.6 Confusion matrix*

True Positives (TP): These are instances where true value is 1, and the predicted value is 1.

True Negatives (TN): These are instances where true value is 0, and the predicted value is 0.

False Positives (FP): These are instances where true value is 0, but the predicted value is 1.

False Negatives (FN): These are instances where true value is 1, but our predicted value is 0.

Starting with the K-Nearest Neighbor (KNN), the model was created with default parameters and ran on the unseen test set. The accuracy came out to be 82%, and after that, model tuning was made to find the best parameters with GridsearchCV, and the best parameter came out as 1. And then, it was run with the best parameter; the accuracy distinctly increased to 99%. Fig.7 shows the confusion matrix got from the results.

Chart

Description automatically generated

*Fig.7 KNN confusion matrix*

I was moving on to the second model, which is the Logistic Regression (LR) classifier. The model was created using the default parameters, and the classification was formed based on the unseen test set. The accuracy came out to be 80%, and after conducting GridsearchCV, the optimised parameters based on cross-validation were found, and the accuracy only increased to 1% and to be 81%. Fig.8 shows the confusion matrix got from the results.

***Chart

Description automatically generated***

*Fig.8* LR *confusion matrix*

Moreover, the last model that was built was the Support Vector Machines (SVM). The model was created using the default parameters and conducted predictions on the unseen test set. The accuracy came out to be 89%, and then model tuning was made to find the best parameters with GridsearchCV, and the best parameter came out as {'C': 9, 'kernel': 'rbf'} And then, It was run with the best parameter; the accuracy distinctly increased to %96. Fig.9 shows the confusion matrix got from the results.

Chart

Description automatically generated

*Fig.9 SVM confusion matrix*

Table.2 delivered to test accuracy score and tuned accuracy score of three models.

|  |  |  |
| --- | --- | --- |
| **Model** | **Test Accuracy Score** | **Tuned Accuracy Score** |
| K-Nearest Neighbor (KNN | 84.09 | 99.03 |
| Logistic Regression (LR) | 80.52 | 81.52 |
| Support Vector Machines (SVM). | 88.96 | 96.43 |

Finally, the Support Vector Machines (SVM) was built using the default parameters and conducted predictions on the unseen test set with the highest accuracy score. However, K-Nearest Neighbor (KNN gets the highest score after tuning with GridsearchCV. On the other hand, Logistic Regression (LR) got the lowest accuracy score in both tuned and test scores.

**5. Conclusion**

In conclusion, Machine learning methods have an important place in medicine as in many different fields and play a significant role in diagnosing various diseases. Early and accurate diagnosis of conditions that set human life at risk will prevent the number of deaths caused by these diseases. This project aimed to calculate the accuracy score by the three different machine-learning approaches and predict the diagnosis of heart disease condition

So, it has been calculated to conclude which one is best among them. Accor doing to this project, KNN got the highest accuracy score of %99, and SVM got the second-highest accuracy score of %96, LR got the lowest accuracy score of %81.

In future research, more machine learning approaches should use for the best analysis of heart diseases. Using different machine learning approaches will increase the accuracy score.

# References

Palaniappan, S. & Awang, . R., 2008. Intelligent heart disease prediction system using data mining. *In 2008 IEEE/ACS international conference on computer systems and applications*, March, pp. 108-115.

Benjamin, E. J. & Muntner, P., 2019. Heart disease and stroke statistics—2019 update: a report from the American Heart Association. *Circulation,* 139(10), pp. e56-e528.

Bonaccorso, G., 2017. *Machine learning algorithms.* s.l.:Packt Publishing Ltd..

Das, R., Turkoglu, . I. & Sengur, A., 2009. Effective diagnosis of heart disease through neural networks ensembles. *Expert systems with applications,* 36(4), pp. 7675-7680.

Engebretsen, M. & Kennedy, H., 2020. Data visualization in society. In: s.l.:s.n., p. 441.

García, S., Luengo, J. & Herrera, F., 2015. *Data preprocessing in data mining.* Cham, Switzerland: Springer International Publishing..

Kishor, A. & Chinmay , C., 2021. Artificial intelligence and internet of things based healthcare 4.0 monitoring system. *Wireless Personal Communications,* pp. 1-17.

Kohli, P. S. & Arora, S., 2018. December. Application of machine learning in disease prediction. *In 2018 4th International conference on computing communication and automation (ICCCA)*, pp. 1-4.

Organization, W. H., 2017. *Cardiovascular diseases (CVDs): World Health Organization,* Geneva: s.n.

Priyanka, N. & Kumar, . P. R., 2017. Usage of data mining techniques in predicting the heart diseases—Naïve Bayes & decision tree. *In 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, April, pp. 1-7.

Shao, Y. E., Hou, C. D. & Chiu, C. C., 2014. Hybrid intelligent modeling schemes for heart disease classification. *Applied Soft Computing,* Volume 14, pp. 47-52.

Singh, J., Kamra, A. & Singh, H., 2016. *October. Prediction of heart diseases using associative classification.* s.l., International conference on wireless networks and embedded systems (WECON).

Surden, H., 2014. Machine learning and law. *Wash,* p. 87.

Wang, C. et al., 2021. Wang, C., Sun, Y., Jiang, D., Wang, C. and Liu, S., 2021. Risk‐Attributable Burden of Ischemic Heart Disease in 137 Low‐and Middle‐Income Countries From 2000 to 2019.. *ournal of the American Heart Association,* 10(19), p. p.e021024.