**CHAPTER – 1 INTRODUCTION**

## INTRODUCTION

Heart disease is the Cardiovascular disease that remains the number one cause of death globally and contributes to approximately 30% of all global deaths [21]. The Heart Disease occurs with common symptoms of breath shortness, physical body weakness and swollen feet. According to the European Society of Cardiology, approximately 26 million people of Heart Disease were diagnosed. Most of the people in the United States are suffering from heart disease. Researchers try to come across an efficient technique for the detection of heart disease, as the current diagnosis techniques of heart disease are not much effective in early time identification. The diagnosis and treatment of heart disease is extremely difficult when modern technology and medical expertise are not available. The effective diagnosis and proper treatment can save the lives of many people. Diagnosis of Heart Disease is traditionally done by the analysis of the medical history of the patient, physical examination report and analysis of concerned symptoms by a physician. But the results obtained from this diagnosis method are not accurate in identifying the heart patients. Moreover, it is expensive and computationally difficult to analyze. Thus, to develop a non-invasive diagnosis system based on Machine Learning classification to resolve these issues is proposed. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information. The system will be developed based on various classification algorithms while standard features selection algorithms will be used such as relief, minimal redundancy relevance & local learning for removing irrelevant & redundant features. We propose novel fast conditional mutual information feature selection algorithm, chi square algorithm to solve feature selection problem. These feature selection increased the classification accuracy and reduce the execution time of classification system.

## 1.1 INTRODUCTION TO THE TECHNIQUES

Human heart can be described as a compound body organ contains muscles together with biological nerves. Human heart pumps nearly 5l of blood in the body providing the human body with renewed materials. Medical research literature shows that there is much interest from the scientific researchers in implementing the Machine Learning, including classification algorithms in medical devices. The human heart operation is composite and any failure is risky to human lives. Hence, heart diagnose systems have been a main concern to the scientific researchers in the last decades. It is not possible to use linear systems to perform the heart diagnosis. Advanced heart diagnosis equipment is not always available in every medical center, especially in the rural areas where less support and care. Moreover, it is not possible for many people to travel to regional medical centers where high quality hospital services are affordable. Physician intuition and experience are not always enough to attain high quality medical results. Consequently, medical errors and undesirable results are reasons for a need for a state of the art computer based diagnosis systems, which in turns reduce medical fatal errors, increase patient safety and save lives. In this research work, causes of heart diseases, the complications and the remedies for the diseases have been considered. An intelligent system which can diagnose heart diseases has been implemented. This system will prevent misdiagnosis which is the major error that may occur by medical doctors. The dataset of statlog heart disease has been used to carry out this experiment. The dataset comprises attributes of patients diagnosed for heart diseases. The diagnosis was used to confirm whether heart disease is present or absent in the patient. The datasets were obtained from the UCI Machine Learning. This dataset was divided into training, validation set and testing set, to be fed into the network.

## 1.2 INTRODUCTION TO MACHINE LEARNING

Machine learning is the study of computer algorithms that improve automatically through experience. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks. A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics. It involves computers discovering how they can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks. For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand; on the computer's part, no learning is needed. For more advanced tasks, it can be challenging for a human to manually create the needed algorithms. In practice, it can turn out to be more effective to help the machine develop its own algorithm, rather than having human programmers specify every needed step. The discipline of machine learning employs various approaches to teach computers to accomplish tasks where no fully satisfactory algorithm is available. In cases where vast numbers of potential answers exist, one approach is to label some of the correct answers as valid. This can then be used as training data for the computer to improve the algorithms it uses to determine correct answers.

## 1.3 OBJECTIVE

* + - To predict the Heart Disease more accurately and to reduce the prediction time consumption.
    - Early identification of Heart Stroke with low cost.
    - To find the results effectively and quickly.

## 1.4 PROJECT DESCRIPTION AND FEATURES

Heart disease is one of the complex disease and globally many people suffered from this disease. On time and efficient identification of heart disease plays a key role in healthcare, particularly in the field of cardiology. In this paper we used fisher score and chi square as a feature selection measure. This measure is used to rank the attributes and to prune irrelevant, redundant attributes. After applying feature selection, classification algorithms Support vector machine (SVM) and Logistic Regression will be applied on the datasets. Chi square is a mathematic procedure that transforms a number of correlated attributes into a smaller number of correlated variables called principle components. Simplest way of determining relevant variables is to use chi square technique (χ2). Fisher score is one of the most widely used supervised feature selection methods. The algorithm which we used returns the ranks of the variables based on the fisher’s score in descending order. We can then select the variables as per the case. Chi square technique is used if all the variables are continuous. Assume that a target variable is selected; every parameter is checked to see if the use chi square technique detects the existence of a relationship between the parameter and the target. In this project, we propose an efficient and accurate system to diagnose heart disease and this system is based on machine learning techniques.

## 1.5 SOCIAL IMPACT

It is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. By implementing our project, it is highly helpful in Healthcare, especially in cardiology field and some research laboratories. It is used to predict the heart disease, based on the results patients can consult the doctor for further treatment.

## 1.6 CHALLENGES

* + - This system is suitable even for small datasets.
    - We need patient co-operation to predict the Heart Disease in earlier stages.

## 1.7 LIMITATIONS

Present system has the following limitations

* + - This software is not fit in rural areas because there is not that much of computerized facilities.
    - If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to over fitting. SVM algorithm is not suitable for large data sets.
    - This system can detect only the heart disease of the patients. Other diseases related to heart cannot be detected with this system.

## 1.8 DOCUMENTATION ORGANIZATION

## CHAPTER – 1

It deals with the introduction part which completely tells the introduction to the work, objective, social impact, challenges and limitations.

## CHAPTER – 2

It defines the problem along with the existing works available in the form of Literature Survey and proposed work.

## CHAPTER – 3

It deals with the description of the hardware and software needed to implement the project. It gives the detailed description about the software that are made use in the project.

## CHAPTER – 4

It discusses about the system analysis part. It describes the hardware and software specification required for our work.

## CHAPTER – 5

It explains the architectural design, module description and flowchart

explanation.

## CHAPTER – 6

It contains the snapshots and result of the project.

## CHAPTER – 7

It finally gives the conclusion about the project and the future enhancements that could be made to the project.

**CHAPTER - 2 LITERATURE SURVEY**

## LITERATURE SURVEY

Literature Survey is the work undertaken to study the existing work available on the subject interest. The following are the papers published on the subject matter. The inferences drawn are shown below:

1. Norma Latif Fitriyani ,Muhammad Syafrudin, Ganjar Alifian and Jongtae Rhee, **“HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System “,** Open Access Journal, Volume 8, July 20, 2020.

This project was developed a heart disease prediction model that were implemented in the clinical decision support system and could be used to help clinicians assess the risk of heart disease and provide appropriate treatments to manage the risk further. In addition, numerous studies had also reported that the implementation of Clinical Decision Support System improved preventive care, clinical decision making and decision quality. The authors proposed an effective HDPM for a Clinical Decision Support System which consists of DBSCAN-based to detect and eliminate the outliers, SMOTE- ENN to balance the training data distribution and XGBoost to predict heart disease. The results showed that LR performed better than the other models by achieving up to 85%, 89%, 81%, and 85 for the accuracy, sensitivity, specificity, and precision.

1. Simanta Shekhar Sarmah,(member,IEEE)**, ”An Efficient IoT-Based Patient Monitoring and Heart Disease Prediction System Using Deep Learning Modified Neural Network“** Open Access Journal, Volume 8, July 6, 2020 .

This paper focused on health monitoring with IOT. The body sensors networks (BSN) were formed of disparate wearable or implantable devices, say accelerometer, cardioverter-defibrillator, and pacemaker, which sense as well as monitor the breathing rates, blood pressures, pulse, together with body temperature of users. Studies conducted on heart failure patients show that around 30 percent of patients had been readmitted as a minimum of once within the timeframe of 90 days. Therefore, the key to augment the Heart Disease performance of the healthcare and decreased the death rate is changing the inert Health Care mode into an invasive one. For 100 and 200 numbers of data, the ANN gives 77.5 and 85 specificity.

1. Gamal G. N. Geweid,(Member, IEEE), and Mahmoud A. Abdallah,(Member, IEEE) , **“A New Automatic Identification Method of Heart Failure Using Improved Support Vector Machine Based on Duality Optimization Technique”,** Open Access Journal, Volume 7, October 4, 2019.

This project specifically focused on functions include an initialization function which were used to acquire an ECG signal using a bio-potential amplifier and then displayed it using ECG instrumentation, a preprocessing function, an analyzing function to analyse ECG signal, and a classifier function to classify the heart failure disease. This paper used feature analysis on heart failure classes and associated medications using SVM TOOLSET based on LIBSVM as presented in to classify several categories of patients with heart failure. The obtained results indicate that the proposed technique produced good results, more efficient and increased the accuracy of HFD detection with an acceptable accuracy of 85.97%

1. SenthilKumar Mohan ,ChandraSegar Thirumalai ,And Gautum Srivastava, (Member, IEEE)**, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”,** Open Access Journal, June 19, 2019.

The authors worked on various techniques in data mining and neural networks has been employed to find out the severity of heart disease among humans. The severity of the disease was classified based on various methods like K-Nearest Neighbor Algorithm, Decision Trees, Genetic algorithm, and Naive Bayes. We introduced neural networks using heart rate time series. The experiment results showed that hybrid method has stronger capability to predict heart disease**.** Thus an enhanced performance level with an accuracy level of 85.7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM).

1. Amin Ul Haq, Amin Ul Haq, Jianping Li, Jalaiuddin Khan, Muhammad Hammad Memon, Shadma Parveen, Mordecai Folain Raji, Wasfi Akbar, Tanvir Ahmad, Sana Ullah, Latipova Shosista, Happy N. Monday, “**Identifying The Predictive Capability Of Machine Learning Classifiers For Designing Heart Disease Detection System**” ‘IEEE Journal,2019.

This project enhanced the diagnosis of heart diseases through invoice based techniques as well as ordinary medical based methods were not reliable. The feature selection algorithm Relief was used for selection of important features and on these selected features, classifiers performances were also computed. Ensemble machine learning techniques (boosting, bagging, stacking) were used to further increase the classifiers performance. Based on model performance evaluation metrics the SVM (RBF) performed excellently on full features achieved accuracy 86%. The results showed that the performance of BPNN based diagnosis system was more effective for heart disease diagnosis.

1. Muhammad Hammad Memon, Amin Ul Haq, Jian Ping Li, Shah Nazir, and Ruinan Sun**, “A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms”,** Hindawi -Mobile Information Systems Volume 2018, ,2 December 2018.

This project designed a diagnostic system for heart disease and used machine learning classifier multilayer perceptron ANN-driven back propagation learning algorithm and feature selection algorithm. In order to evaluate the performance of classifier, various performance evaluation metrics such as classification accuracy, classification error, specificity, sensitivity, Matthews’ correlation coefficient (MCC), and receiver optimistic curves (ROC) were used. All these techniques mostly cause imprecise diagnosis and often delay in the diagnosis results due to human errors. Moreover, it was more expensive and computationally complex and takes time in assessments. Thus the classification system achieved a classification accuracy of 81.10%.

1. Sellappan Palaniappan, Rafiah Awang**, “Intelligent Heart Disease Prediction System Using Data Mining Techniques”,** IEEE Journal, 2008.

The authors developed a prototype Intelligent Heart Disease Prediction System using three data mining modeling techniques, namely, Decision Trees, Naïve Bayes and Neural Network. Intelligent Heart Disease Prediction System can discover and extract hidden knowledge (patterns and relationships) associated with heart disease from a historical heart disease database. It could answer complex queries for diagnosing heart disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. By providing effective treatments, it also helped to reduce treatment costs. To enhance visualization and ease of interpretation, it displayed the results both in tabular and graphical forms. The Result showed that each technique had its unique strength in realizing the objectives of the defined mining goals.

## 2.1 METHODOLOGY USED

The Supervised classification algorithm such as Support vector machine and Logistic Regression is used in the proposed system. A Support vector machine (SVM) model is basically a representation of different classes in a hyper plane in multidimensional space. The hyper plane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyper plane. [Logistic regression](https://www.geeksforgeeks.org/understanding-logistic-regression/) is also a classification algorithm used to find the probability of event success and event failure. It is used when the dependent variable is binary in nature.

## 2.2 MERITS

* + Logistic regression is easier to implement, interpret, and very efficient to train.
  + SVM works relatively well when there is a clear margin of separation between classes.

## 2.3 LIMITATIONS

* If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to over fitting.
* SVM algorithm is not suitable for large data sets.

## 2.4 FUTURE WORK

* In future, it is possible to provide extensions or modifications to the proposed clustering and classification algorithms using intelligent agents to achieve further increased performance.
* Apart from the experimented combination of data mining techniques, further combinations such as artificial intelligence, soft computing and other clustering algorithms can be used to improve the accuracy.

## 2.5 PROBLEM DESCRIPTION

* + - 1. **2.5.1 LIMITATION IN THE EXISTING SYSTEM**
* Existing system takes more time consumption.
* Less Prediction Accuracy.
* Theoretical Limits.

**CHAPTER – 3**

**SYSTEM REQUIREMENTS**

## 3.1 HARDWARE REQUIREMENTS

The system is implemented using the following hardware components.

* + - RAM : 4 GB or above
    - Hard Disk : 1000 GB or more

## 3.2 SOFTWARE REQUIREMENTS

The system is implemented using the following software components

* Operating System : Windows 7 or above
* Language : Python
* IDE : Spyder

## 3.3 SOFTWARE DESCRIPTION

* + 1. **3.3.1 PYTHON**

Python is an interpreted, high-level, general purpose programming language. Created by Guido Van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales.

Python is a multi-paradigm programming language, Object-oriented programming and structured programming are fully supported, and many of its features support functional programming and aspect-oriented programming (including by metaprogramming and meta objects (magic methods)).

Many other paradigms are supported via extensions, including design by contract and logic programming.

Python uses dynamic typing, and a combination of reference counting and a cycle-detecting garbage collector for memory management. It also

features dynamic name resolution (late binding), which binds method and variable names during program execution.

## 3.3.2 FEATURES OF PYTHON

* **Easy to learn** – Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
* **Easy to read** – Python code is more clearly defined and visible to the eyes.
* **Easy to maintain** – Pyhton’s source code is fairly easy to maintain.
* **Interactive Mode** – Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
* **Portable** – Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
* **Databases** – Python provides interfaces to all major commercial databases.
* **Scalable** – Python provides a better structure and support for large programs than shell scripting.

## 3.3.3 ANACONDA

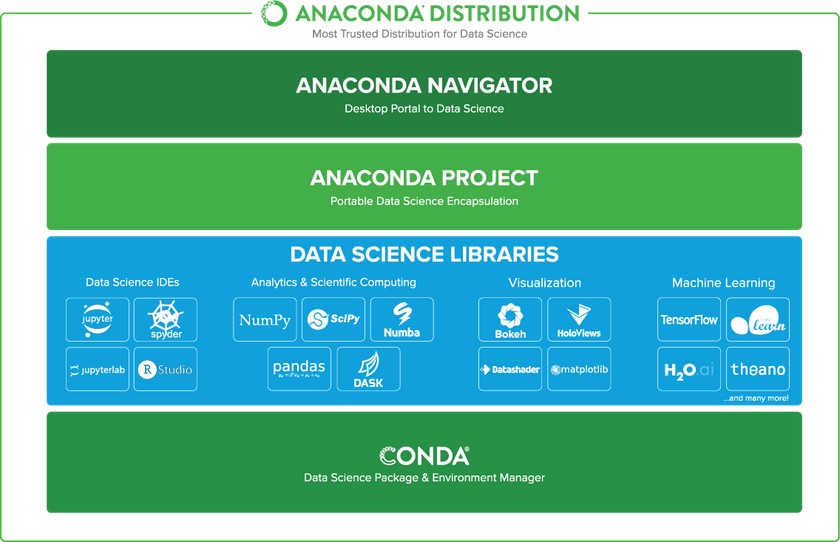
Anaconda is the most popular python data science platform.

## Anaconda Distribution

With over 6 million users, the open source Anaconda Distribution is the fastest and easiest way to do Python and R data science and machine learning on Linux, Windows, and Mac OS X. It's the industry standard for developing, testing, and training on a single machine.

## i) Anaconda Enterprise

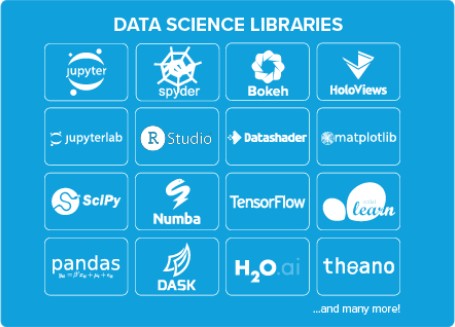
Anaconda Enterprise is an AI/ML enablement platform that empowers organizations to develop, govern, and automate AI/ML and data science from laptop through training to production. It lets organizations scale from individual data scientists to collaborative teams of thousands, and to go from a single server to thousands of nodes for model training and deployment.



## Figure 1. Anaconda Distribution

## iii) Anaconda Data Science Libraries

* Over 1,400 Anaconda-curated and community data science packages
* Develop data science projects using your favourite IDEs, including Jupyter, JupyterLab, Spyder, and RStudio
* Analyse data with scalability and performance with Dask, numpy, pandas, and Numba
* Visualize your data with Matplotlib, Bokeh, Datashader, and Holoviews
* Create machine learning and deep learning models with Scikit-learn, Tensorflow, h20, and Theano



## Figure 2. Data Science Libraries

Conda, the Data Science Package & Environment Manager

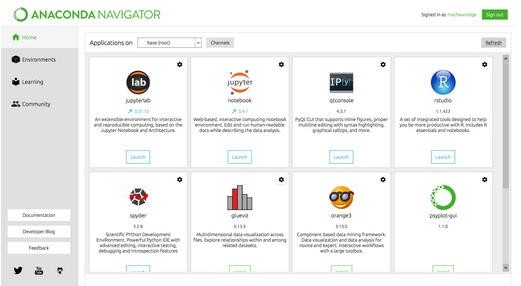
* + Automatically manages all packages, including cross-language dependencies
  + Works across all platforms: Linux, macOS, Windows
  + Create virtual environments
  + Download conda packages from Anaconda, Anaconda Enterprise, Conda Forge, and Anaconda Cloud



## Figure 3. Conda Packages

Anaconda Navigator, the Desktop Portal to Data Science

* + Install and launch applications and editors including Jupyter, RStudio, Visual Studio Code, and Spyder
  + Manage your local environments and data science projects from a graphical interface
  + Connect to Anaconda Cloud or Anaconda Enterprise
  + Access the latest learning and community resources



**Figure 4. Anaconda Navigator**

**iv) Spyder**

Spyder is an open source cross-platform integrated development

environment (IDE) for scientific programming in the Python language. Initially created and developed by Pierre Raybaut in 2009, since 2012 Spyder has been maintained and continuously improved by a team of scientific Python developers and the community. Strongly recommend the free, open-source Spyder Integrated Development Environment

## (IDE) for scientific and engineering programming, due to its integrated editor, interpreter console, and debugging tools. Spyder is included in Anaconda and other distributions.

Spyder is a powerful scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts. It offers a unique combination of the advanced editing, analysis, debugging, and profiling functionality of a comprehensive development tool with the data exploration, interactive execution, deep inspection, and beautiful visualization capabilities of a scientific package.

Beyond its many built-in features, its abilities can be extended even further via its plugin system and API. Furthermore, Spyder can also be used as a PyQt5 extension library, allowing developers to build upon its functionality and embed its components, such as the interactive console, in their own PyQt software.

## v) Editor

Work efficiently in a multi-language editor with a function/class browser, code analysis tools, automatic code completion, horizontal/vertical splitting, and go-to-definition.

## vi) IPython Console

Harness the power of as many IPython consoles as you like within the flexibility of a full GUI interface; run your code by line, cell, or file; and render plots right inline.

## vii) Variable Explorer

Interact with and modify variables on the fly: plot a histogram or time series, edit a date frame or Numpy array, sort a collection, dig into nested objects, and more!

## viii) Profiler Find and eliminate bottlenecks to unchain your code's performance.

## ix) Debugger Trace each step of your code's execution interactively.

**x) Help** Instantly view any object's docs, and render your own.

## 3.3.4 DATASET

The dataset used for detection is a heart disease data set which can be taken from UC Irvine Machine Learning Repository. This dataset contains information about the transactions age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, target attributes. There are 304 rows and 14 columns.

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Attributes** | **Description** | **Domain of Values** |
| **1** | Age | Age in Years | 29-77 |
| **2** | Sex | Sex | Male (1)  Female (0) |
| **3** | Cp | Chest Pain Type | Typical Angina (0)  Atypical Angina (1)  Non-Anginal (2)  Asymptomatic (3) |
| **4** | Trestbps | Resting Blood Sugar | 94 to 200 mm Hg |
| **5** | Chol | Serum Cholesterol | 126 to 564 mg/dl |
| **6** | Fbs | Fasting Blood Sugar | Greater than 120 mg/dl  True (1)  False(0) |
| **7** | Restecg | Resting ECG Result | Normal (0)  ST-T Wave abnormality (1)  LV Hypertrophy (2) |
| **8** | Thalach | Maximum Heart Rate  Achieved | 71 to 202 |
| **9** | Exang | Exercise Induced Angina | Yes (1)  No (0) |
| **10** | Oldpeak | ST Depression induced by exercise relative in rest | 0 to 6.2 |
| **11** | Slope | Slope of peak exercise ST Segment | Unsloping (0)  Flat (1)  Downsloping (2) |
| **12** | Ca | Number of Major Vessels Coloured by Fluoroscopy | 0 - 3 |
| **13** | Thal | Defect Type | Normal (1)  Fixed Defect (2)  Reversible Defect(3) |
| **14** | Target | Heart Disease | 0 = Absence  1 = Presence |

**Table 1. Attributes Information of Heart Disease Dataset**

**CHAPTER - 4**

**SYSTEM DESIGN**

## ARCHITECTURAL DIAGRAM

Classification

&

Predicton

Check

Missing

Values

Load

Heart Disease Dataset

Feature Selection

**Figure 5. System Design**

In the above system design, the dataset heart disease is loaded as input. The input dataset is taken from UC Irvine(UCI) Machine Learning Repository. The collected input data are subjected to check if there is any missing values. Then the data are split into Train set (80%) and Test set (20%) for decision and Apply fisher score and chi square as feature selection algorithm. Support Vector Machine (SVM) and Logistic Regression algorithms are implemented. Then analyze the performance metrics like Accuracy, Recall, F-Measure and Precision. Finally we predict whether the patient has heart disease or not. Heart Patients count displayed as bar graph in the tkinter window along with classification accuracy comparison chat. Finally an alert dialog box displayed which shows that the result of prediction is done successfully.

## FLOW DIAGRAM

**Figure 6. Flow Diagram**

In Figure 5 working and function of the proposed system is discussed. Initially the dataset has to be loaded and check if there is any missing values. The data has to split into the ratio of 8:2 as train set and test set respectively and then feature selection algorithm, fisher score and chi square has been applied. If the classifier is trained, then predict whether the patient has heart disease or absence of heart disease.

# CHAPTER-5

# SYSTEM IMPLEMENTATION

## 5.1 MODULES

* + - Data selection and loading
    - Checking missing values
    - Feature selection
    - Classification
      * Support Vector Machine (SVM)
      * Logistic Regression
    - Tkinter Window display the accuracy comparison chart for both classifier

## 5.2 MODULE DESCRIPTION

**5.2.1 Heart Disease Dataset**

* + - * The dataset used for prediction is a heart disease data set.
      * This dataset contains information about the transactions age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, target attributes.
      * There are 304 rows and 14 columns.

## Checking missing values

* + - * The collected dataset is subjected to check if there is any missing values.
      * Our dataset is strong and doesn’t have any missing values.

## Feature Selection

We used fisher score and chi square as a feature selection measure. This measure used to rank the attributes and to prune irrelevant, redundant attributes. After applying feature selection, classification using Support vector machine (SVM) and Logistic Regression will be applied on the data sets. Fisher score is one of the most widely used supervised feature selection methods. The algorithm which we used returns the ranks of the variables based on the fisher’s score in descending order. We can then select the variables as per the case. Chi square is a mathematic procedure that transforms a number of correlated attributes into a smaller number of correlated variables called principle components. Simplest way of determining relevant variables is to use chi square technique (χ2). Chi square technique is used if all the variables are continuous. Assume that a target variable is selected; every parameter is checked to see if the use chi square technique detects the existence of a relationship between the parameter and the target.

\chi^{2}=\sum \frac{\left({O}\_{i}-E\_{i}\right)^{2}}{E\_{i}}

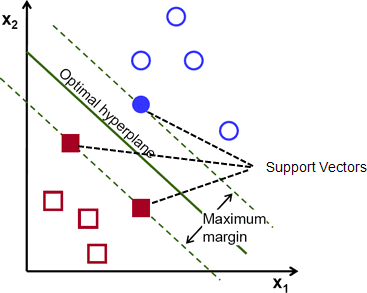
\chi^2 = chi squared

{O}\_i = observed value

E\_{i} = expected value

## Classification

## Support Vector Machine (SVM):

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text.

## Figure 7. Support Vector Machine model

**Types of SVM**

## SVM can be of two types:

* + - **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
    - **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

## Hyperplane:

There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.

The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane.

We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

## Support Vectors:

The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

## Pros of SVM classifiers

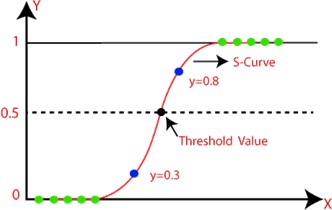
SVM classifiers offers great accuracy and work well with high dimensional space. SVM classifiers basically use a subset of training points hence in result uses very less memory.

## Cons of SVM classifiers

They have high training time hence in practice not suitable for large datasets. Another disadvantage is that SVM classifiers do not work well with overlapping classes

## Logistic Regression:

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression is estimating the parameters of a logistic model (a form of binary regression).

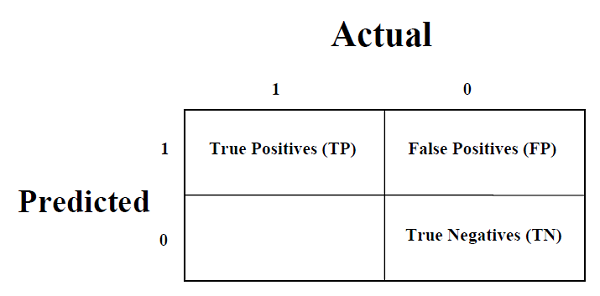


## Figure 8. Logistic Regression model

* + - * + 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.
    1. **Prediction Analysis**

**Confusion Matrix**

It is the easiest way to measure the performance of a classification problem where the output can be of two or more type of classes. A confusion matrix is nothing but a table with two dimensions viz. “Actual” and “Predicted” and furthermore, both the dimensions have “True Positives (TP)”, “True Negatives (TN)”, “False Positives (FP)”, “False Negatives (FN)” as shown below



**Figure 9. Prediction values**

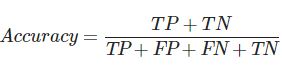
**Explanation of the terms associated with confusion matrix are as follows**

* **True Positives (TP)** − It is the case when both actual class & predicted class of data point is 1.
* **True Negatives (TN)** − It is the case when both actual class & predicted class of data point is 0.
* **False Positives (FP)** − It is the case when actual class of data point is 0 & predicted class of data point is 1.
* **False Negatives (FN)** − It is the case when actual class of data point is 1 & predicted class of data point is 0.

We can use confusion\_matrix function of sklearn.metrics to compute Confusion Matrix of our classification model.

**Classification Accuracy**

It is most common performance metric for classification algorithms. It may be defined as the number of correct predictions made as a ratio of all predictions made. We can easily calculate it by confusion matrix with the help of following formula



We can use accuracy\_score function of sklearn.metrics to compute accuracy of our classification model.

**Classification Report**

This report consists of the scores of Precisions, Recall, F1 and Support. They are explained as follows

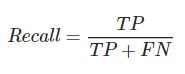
**Precision**

Precision, used in document retrievals, may be defined as the number of correct documents returned by our ML model. We can easily calculate it by confusion matrix with the help of following formula



**Recall or Sensitivity**

Recall may be defined as the number of positives returned by our ML model. We can easily calculate it by confusion matrix with the help of following formula −



**Support**

Support may be defined as the number of samples of the true response that lies in each class of target values.

**F1 Score**

This score will give us the harmonic mean of precision and recall. Mathematically, F1 score is the weighted average of the precision and recall. The best value of F1 would be 1 and worst would be 0. We can calculate F1 score with the help of following formula

***F1 = 2 ∗ (PRECISION ∗ RECALL) / (PRECISION + RECALL)***

F1 score is having equal relative contribution of precision and recall. We can use classification\_report function of sklearn.metrics to get the classification report of our classification model.

## 5.3 Steps Involved in Implementation

**5.3.1 Import Requird Package**

Import Pandas, matplotlib (pyplot), seaborn and warnings libraries.

## 

## 5.3.2 Read the csv file

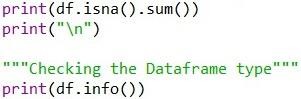
 Select and Load the dataset from the UCI Repository.



**Figure 10. Heart Disease Dataset**

**5.3.3 Check if there is missing values**

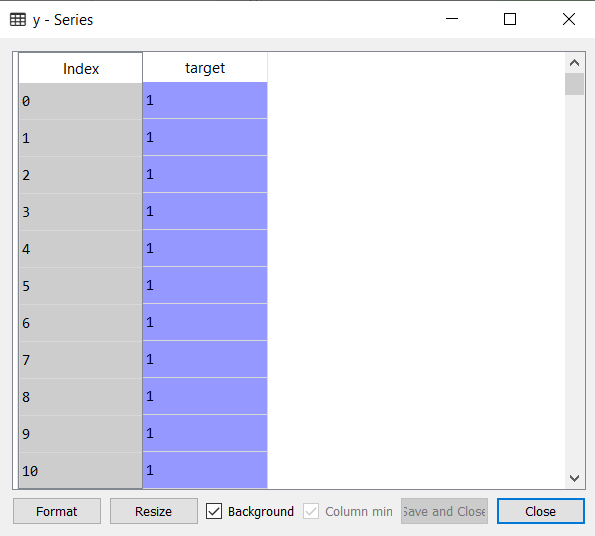
In this step, we have to check if there is any missing values in the dataframe and the display the dataframe along with its type.



## Define X and Y

X stores the input features that we want to consider (Label). Y stores the value of output (Target).



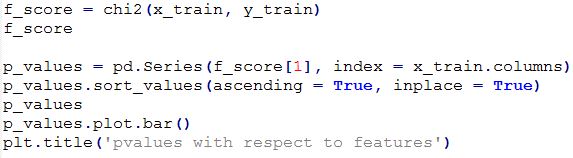


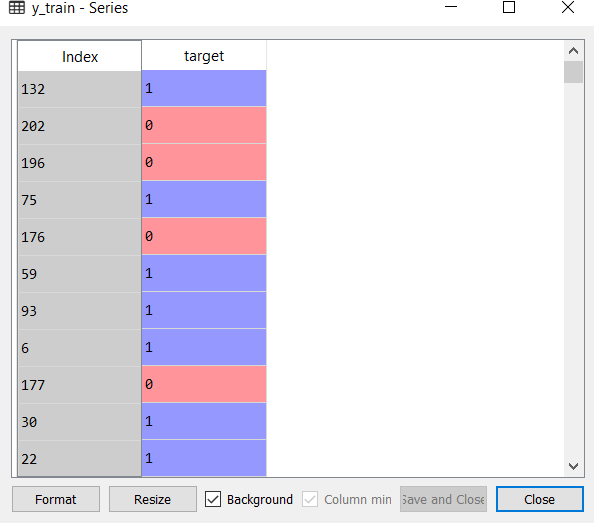
**Figure 11. X Dataframe Figure 12. Y Series**

## 5.3.5 Divide Data into a Training and Testing Dataset and apply feature selection algorithm

Split the data into train and test set in the ratio of 8:2 which was further passed to the svm model to fit, predict and score the model. Then fisher score and chi square algorithm has applied. The same procedure repeated for logistic regression algorithm.

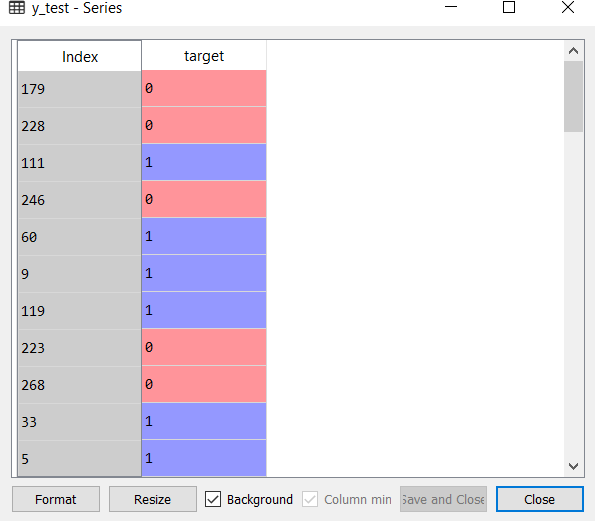






## Figure 13. X Train Figure 14. Y Train

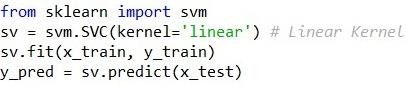
## 



**Figure 15. X Test Figure 16. Y Test**

## Train the model

Create the model. Train our model with 80% of data. Fit the data into that model so that our data has trained according to that model.





**Figure 17. Prediction values**

## 5.3.7 Prediction Analysis

The final result will get generated based on the overall classification and

prediction. The performance of this proposed approach is evaluated using some of the

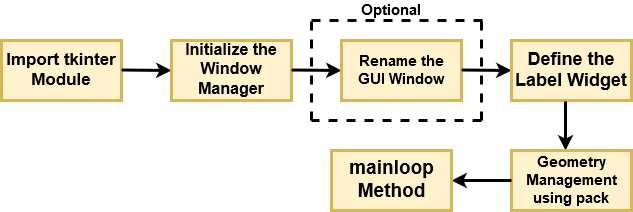
measures

* Precision
* Recall
* F1 Measure
* Support



**5.3.8** **TKinter**

TKinter is a library written in Python that is widely used to create GUI applications. It is very easy to build GUI using Tkinter and the process is even faster. Tkinter has several widgets that can be used while developing GUI. These include buttons, radio buttons, checkboxes, etc. The Heart Disease Count of the patient is visualized as Bar graph in Tkinter window along with Pie Graph. Finally the accuracy comparison bar graph chart will be displayed in the Tkinter Window. It is commonly comes bundled with Python, using Tk and is Python's standard GUI framework. It is famous for its simplicity and graphical user interface. It is open-source and available under the Python License.



**Figure 18. TKinter WorkFlow Diagram**

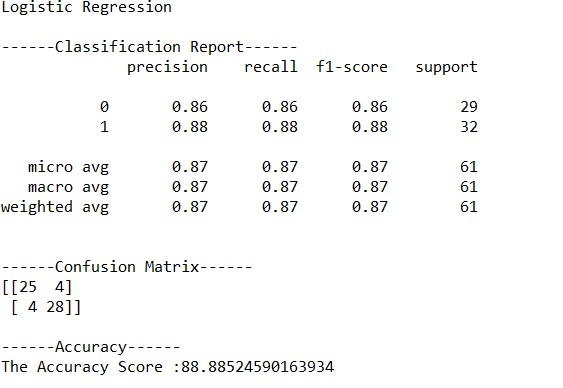
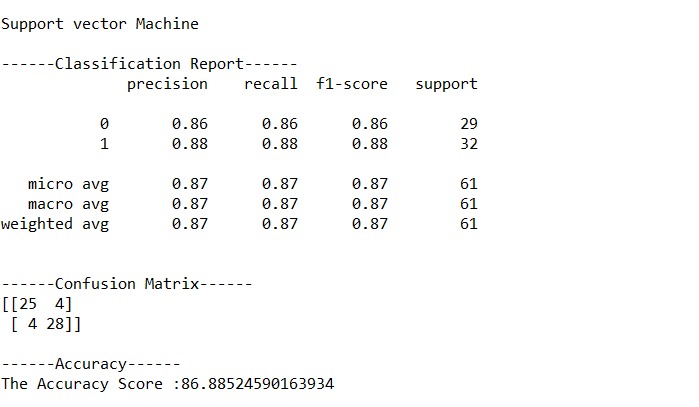
**CHAPTER - 6**

**RESULTS AND DISCUSSION**

**RESULTS AND DISCUSSION**

This system is able to predict whether the patient has heart disease or not using Support Vector Machine and Logistic Regression Classification Algorithm and gives accuracy of 86.88% and 88.88% respectively. Finally it displays the heart disease patients count and Classification Accuracy comparison chart as a bar graph in the Tkinter window. A alert dialog box displayed at the end as the prediction has done successfully. This system improves accuracy, prevent the patients from heart disease at earlier stages and thus heart disease can be controlled to a certain level.

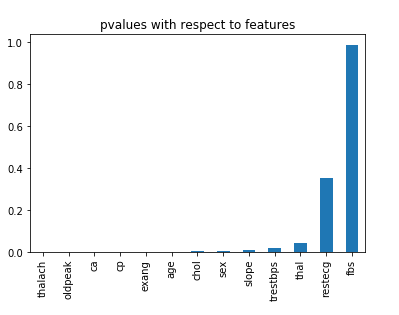
**6.1 SNAPSHOT**

****

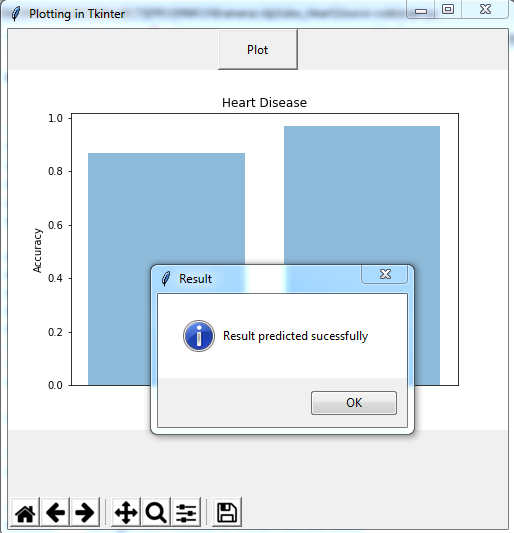
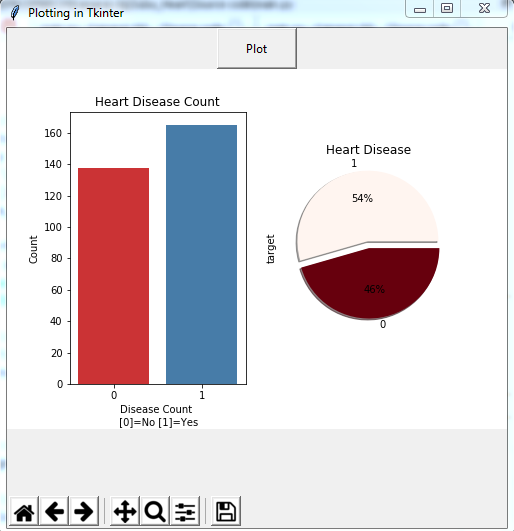
**Figure 19. Classification Report, Figure 20. Classification Report,**

**Confusion Matrix and Confusion Matrix and**

**Accuracy for SVM Accuracy for LR**

****

**Figure 21. Pvalues with respect to features**



**Figure 22. Heart Disease Count Figure 23. Comparison Chart of SVM and LR**

**CHAPTER - 7**

**CONCLUSION AND FUTURE WORK**

**7.1 CONCLUSION**

In this paper we proposed a new feature selection method for heart disease classification. We applied different feature selection methods to rank the attributes which contribute more towards classification of heart disease, which indirectly reduces the no. of diagnosis tests to be taken by a patient. Our experimental results indicate that on an average with Chi Square and feature subset selection provides on the average better classification accuracy and dimensionality reduction. Our proposed method eliminates useless and distortive data. This research will contribute reliable and faster automatic heart disease diagnosis system, where easy diagnosis of heart disease will saves lives. Coronary heart disease can be handle successfully if more research is encourage in this area.

**7.2 FUTURE ENHANCEMENT**

In future, it is possible to provide extensions or modifications to the proposed clustering and classification algorithms using intelligent agents to achieve further increased performance. Apart from the experimented combination of data mining techniques, further combinations such as artificial intelligence, soft computing and other clustering algorithms can be used to improve the accuracy.

**APPENDIX I**

**APPENDIX I**

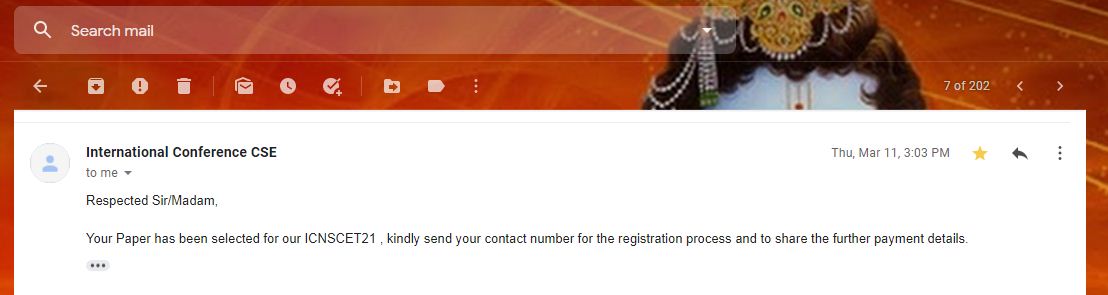
**PAPER PUBLICATION**

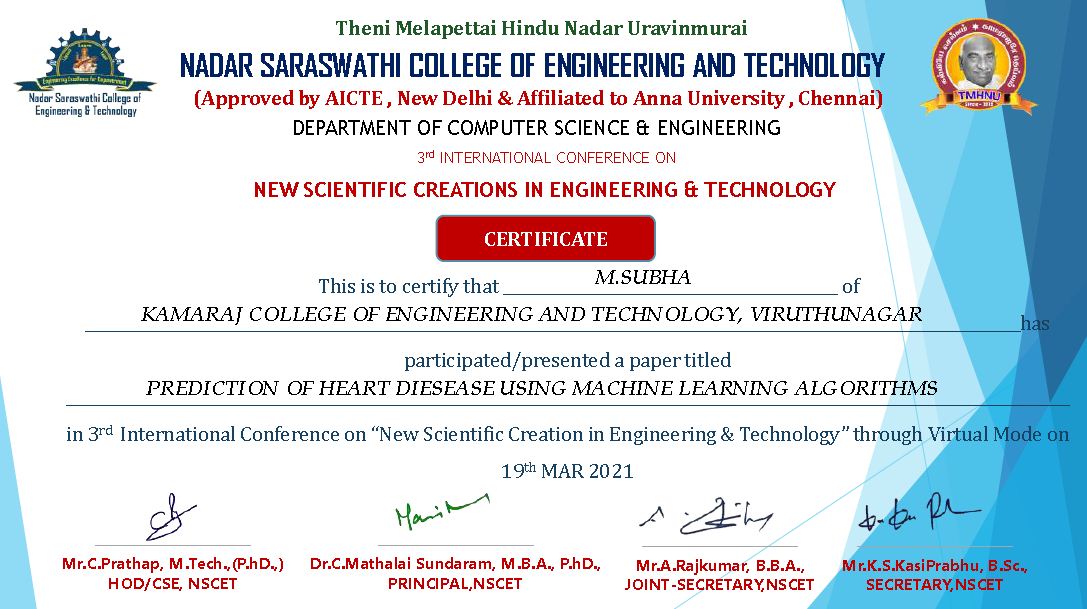
**Conference Name :** 3rd International Conference on New Scientific Creations in Engineering and Technology (ICNSCET-2k21)

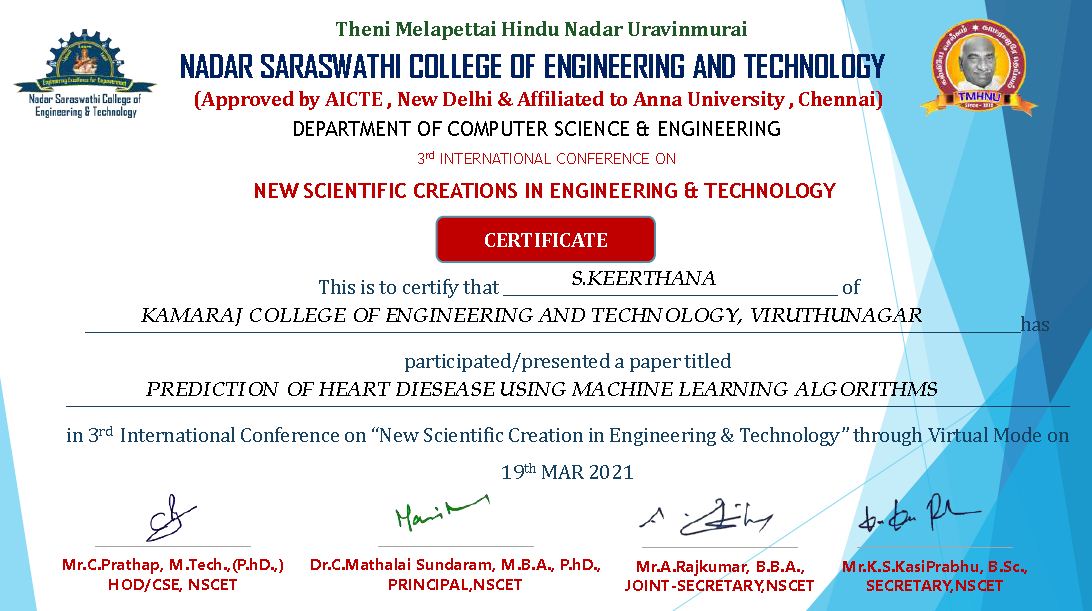
**Held On :** 19-03-2021 on Virtual Mode

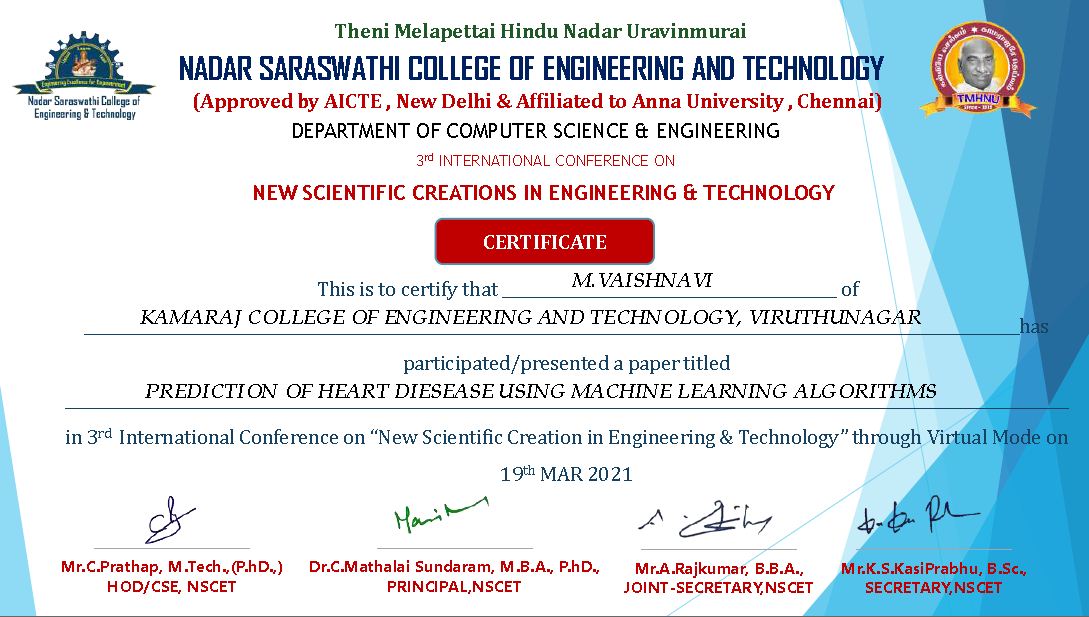
**Organized By :** Nadar Saraswathi College of Engineering and Technology, Theni .

Department of Computer Science and Engineering .









**REFERENCES**

**REFERENCES**

1. Norma Latif Fitriyani ,Muhammad Syafrudin, Ganjar Alifian and Jongtae Rhee, **“HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System “,** Open Access Journal, Volume 8, July 20, 2020.
2. Simanta Shekhar Sarmah,(member,IEEE)**, ”An Efficient IoT-Based Patient Monitoring and Heart Disease Prediction System Using Deep Learning Modified Neural Network“** Open Access Journal, Volume 8, July 6, 2020.
3. Gamal G. N. Geweid,(Member, IEEE), and Mahmoud A. Abdallah,(Member, IEEE) , **“A New Automatic Identification Method of Heart Failure Using Improved Support Vector Machine Based on Duality Optimization Technique”,** Open Access Journal, Volume 7, October 4, 2019.
4. SenthilKumar Mohan ,ChandraSegar Thirumalai ,And Gautum Srivastava, (Member, IEEE)**, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”,** Open Access Journal, June 19, 2019.
5. Amin Ul Haq, Amin Ul Haq, Jianping Li, Jalaiuddin Khan, Muhammad Hammad Memon, Shadma Parveen, Mordecai Folain Raji, Wasfi Akbar, Tanvir Ahmad, Sana Ullah, Latipova Shosista, Happy N. Monday, “**Identifying The Predictive Capability Of Machine Learning Classifiers For Designing Heart Disease Detection System**” ‘IEEE Journal,2019.
6. Muhammad Hammad Memon, Amin Ul Haq, Jian Ping Li, Shah Nazir, and Ruinan Sun**, “A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms”,** Hindawi -Mobile Information Systems Volume 2018, ,2 December 2018.
7. Sellappan Palaniappan, Rafiah Awang**, “Intelligent Heart Disease Prediction System Using Data Mining Techniques”,** IEEE Journal, 2008.

**[8]** Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Prof. Kailas Devadkar, “**Prediction of Heart Disease Using Machine Learning**”, Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018).

**[9]** A. U. Haq, J. P. Li, M. H. Memon, J. Khan, A. Malik, T. Ahmad, A. Ali, S. Nazir, I. Ahad, and M. Shahid, **‘‘Feature selection based on L1-norm support vector machine and effective recognition system for Parkinson’s disease using voice recordings,’’** IEEE Access, vol. 7, pp.37718–37734, 2019.

**[10]** A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, **‘‘Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson’s disease symptom severity,’’** J. Roy. Soc. Interface, vol. 8, no. 59, pp. 842–855, 2011.

**[11]** E. O. Olaniyi, O. K. Oyedotun, and K. Adnan, **‘‘Heart diseases diagnosis using neural networks arbitration,’’** Int. J. Intell. Syst. Appl., vol. 7, no. 12, pp. 72, 2015.

**[12]** F. Fleuret, **‘‘Fast binary feature selection with conditional mutual information,’’** J. Mach. Learn. Res., vol. 5, pp. 1531–1555, Nov. 2004.

**[13]** J. Lopez-Sendon, ‘**‘The heart failure epidemic,’’** Medicographia, vol. 33, no. 4, pp. 363–369, 2011.

**[14]** J. H. Gennari, P. Langley, and D. Fisher, ‘**‘Models of incremental concept formation,’’** Artif. Intell., vol. 40, nos. 1–3, pp. 11–61, Sep. 1989.

**[15]**  L. Zhu, J. Shen, L. Xie, and Z. Cheng, ‘**‘Unsupervised topic hypergraph hashing for efficient mobile image retrieval,’’** IEEE Trans. Cybern., vol. 47, no. 11, pp. 3941–3954, Nov. 2017

**[16]** L. A. Allen, L. W. Stevenson, K. L. Grady, N. E. Goldstein, D. D. Matlock,R. M. Arnold, N. R. Cook, G. M. Felker, G. S. Francis, P. J. Hauptman,E. P. Havranek, H. M. Krumholz, D. Mancini, B. Riegel, and J. A. Spertus,‘**‘Decision making in advanced heart failure: A scientific statement from the American heart association,’’** Circulation, vol. 125, no. 15,pp. 1928–1952, 2012.

**[17]**  A. Heidenreich, J. G. Trogdon, O. A. Khavjou, J. Butler, K. Dracup, M. D. Ezekowitz, E. A. Finkelstein, Y. Hong, S. C. Johnston, A. Khera, D. M. Lloyd-Jones, S. A. Nelson, G. Nichol, D. Orenstein, P. W. F. Wilson, and Y. J. Woo, **‘‘Forecasting the future of cardiovascular disease in the united states: A policy statement from the American heart association,’’** Circulation, vol. 123, no. 8, pp. 933–944, 2011.

**[18]** R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, **‘‘Relief-based feature selection: Introduction and review,’’** J. Biomed.Informat., vol. 85, pp. 189–203, Sep. 2018.

**[19]** S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, **‘‘Innovative artificial neural networks-based decision support system for heart diseases diagnosis,’’** J. Intell. Learn. Syst. Appl., vol. 5, no. 3, 2013, Art. no. 35396.

**[20]** S. Nazir, S. Shahzad, S. Mahfooz, and M. Nazir, **‘‘Fuzzy logic based decision support system for component security evaluation,’’** Int. Arab J. Inf. Technol., vol. 15, no. 2, pp. 224–231, 2018.

**[21]** World Health Organization. (2017). Cardiovascular Diseases *(CVDs)*.[Online]. Available: <https://www.who.int/health-topics/cardiovasculardiseases/>