An Efficient Machine Learning Model for Clinical Support to Predict Heart Disease

B.Vara Prasada Rao†, B.Satyanarayana Reddy††, I.Naga Padmaja†††, and K.Ashok Kumar††††
boyapativaraprasad@gmail.com snreddy.beeram@gmail.com
nagapadmaja.indeti@gmail.com ashok.kavuru@gmail.com
†Associate Professor, Department of Computer Science & Engineering, RVR&JC College of Engineering, Guntur, Andhra Pradesh. ORCID ID:0000-0001-7130-3381
††Associate Professor & Head, Department of Computer Science & Engineering, Guntur Engineering College, Guntur, Andhra Pradesh. ORCID ID:0000-0001-9593-7194
†††Assistant Professor, Department of Information Technology, RVR&JC College of Engineering, Guntur, Andhra Pradesh. ORCID ID:0000-0003-4336-2127
††††Assistant Professor, Department of Electronics & Communication Engineering, RVR&JC College of Engineering, Guntur, Andhra Pradesh. ORCID ID:0000-0001-7141-5180

Summary
Early detection can help prevent heart disease, which is one of the most common reasons for death. This paper provides a clinical support model for predicting cardiac disease. The model is built using two publicly available data sets. The admissibility and application of the the model are justified by a sequence of tests. Implementation of the model and testing are also discussed

Keywords:
Efficient Machine Learning Model, Predict Heart Disease, Early Detection

Acronyms:
DBSCAN- Density Based Spatial Clustering of Applications with Noise
SMOTE-ENN-Synthetic minority oversampling technique- Edited nearest neighbor
kNN- kth Nearest Neighbor
XGBoost- Extreme Gradient Boost
ANN- Artificial Neural Network
SVM- Support Vector Machine
DT – Decision Tree
NB- Naïve Bayesian
MLA- Machine Learning Algorithm

1. Introduction

Heart disease is one of the major problems the world is facing right now. Out of all deaths, it accounts for about 30 percent of the share. The overall number of deaths is predicted to rise by 22 million if the problem is not handled. Heart disease is a special case where a plaque on arterial walls obstructs the smooth flowing of blood and causes heart stroke. The factors that contribute to this heart disease include excess consumption of alcohol and tobacco. The advanced identification of heart disease in potential individuals and better diagnosis by using a prediction model can be useful in reducing the death rates. The clinical support system will include a prediction model that clinicians may use to assess risk and determine the best course of action. With the increasing severity of heart diseases, they should be identified at an early stage to prevent deaths. Here the model makes use of some machine learning con- cepts to build an accurate model which predicts the occurrence of heart disease and recommends proper treatment to the subject/patient. The overall aim of the model is to accurately predict with a few tests and attributes. We can consider many more attributes for prediction but our goal is to predict the presence of heart disease with a lower number of variables and faster efficiency.

The objectives of this article are:

- To study the existing systems, identify their drawbacks, propose a new model to overcome the problems and evaluate the measures on the subjects.
- To compare the measures with other machine learning models and identify the best one

Manuscript received June 5, 2022
Manuscript revised June 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.6.31

among them.

- To frame a system that aids in making clinical decisions with the help of new model.

After looking at all the previous research on this topic, no one has looked at incor- porating DBSCAN, SMOTE-ENN, and XGBoost algorithms to build the required model. Therefore, the proposed model is effective for clinical support that uses the DBSCAN algorithm for identifying the outliers and eliminating them, the SMOTE- ENN algorithm for balancing the data, and XGboost for predicting the presence of heart disease.

The proposed model is anticipated to assist the clinicians to diagnose the patients with better efficiency and thereby enhancing the process of decision making in case of heart disease. Thereby early treatment can be made to avoid deaths caused by the delayed diagnosis of heart disease.

2. Survey of Literature

Several research works have described the development of cardiac disease diag- nostics based on models of machine learning with the goal of developing a better prediction model. Some of them are: Long et al.(2015) [5], Nahato et al.(2015) [6], Verma et al.(2016) [7], Dwivedi et al.(2018) [2], Haq et al.(2018) [3], Latha and Jeeva (2019) [4] and Ali et al.(2019) [1]. The findings revealed that the suggested model surpass other designs and earlier results, as well as the references therein.

An attempt is made in this paper to develop a new model to the data sets to achieve accuracy, sensitivity and precision to the extent possible. The remainder of the article is laid out as follows: System analysis is induced in Section 3. The approach to modeling and documenting software is discussed in Section 4. Algorithm and its implementation are given in Section 5. Different testing types used to test the model fitted to the data are given in Section 6. The paper is ended up with Summary and conclusions in Section 7.

3. System Analysis

A software requirement specification is a detailed description of how a framework should behave when it is built. It includes a huge number of use cases that show all of the product's interactions with customers. In addition to use cases, the software re- quirement specifications include non-useful requirements. Non-practical requirements are those that impose constraints on the plan or usage, such as operational efficiency requirements, quality benchmarks, or structure imperatives, among others.

- In business terminology, business necessities describe what must be commu- nicated or accomplished in order to provide some motivation.
- The features of a framework or object are described in product requirements (It could be one of several options for meeting a variety of corporate needs).
- The activities carried out by the generating company are depicted in the pro- cess prerequisites. Process requirements, for example, may include clear tech- niques that must be pursued and requirements with which the organization must comply.

The requirements for both the item and the procedure are inextricably linked. The exercises that will be performed to fulfill an item requirement are frequently deter- mined by procedure prerequisites.For example, a most extreme advancement cost requirement (a procedure prerequisite) may be forced to help achieve a most extreme deals value requirement (an item prerequisite), and a requirement that the item is viable (a Product prerequisite) is frequently tended to by forcing necessities to pursue specific development styles. In general, the numerous sorts of outputs are:

- External Outputs are those that have a destination outside of the organization.
- Internal outputs are the user's primary

interface with the computer and have a destination within the organization.

- Operational outputs that are only used by the computer branch.
- Interface outputs that allow the customer to communicate straight with the design.
- Understanding user's preferences, expertise level and his business requirements.

Non-functional requirements that aren't addressed can result in systems that don't meet users' needs.The non-functional requirements are as follows:

- Usable
- Serviceable
- Manageable
- Recoverable
- Secured
- Integrity of Data
- Capacity
- Scalable
- Reliable
- Maintainable

4. Unified Modeling Language

Unified Modeling Language (UML) is a cutting-edge method to software modelling and documentation, to put it simply. It is, in fact, one of the most extensively used methodologies for modelling business processes. It is built on the foundation of software component diagrams. Using visual representations, we can better understand potential faults or errors in software or business processes. The elements resemble components that can be connected in a variety of ways to construct a complete UML diagram. As a result, understanding the various schematics is vital in order to apply what you've learnt in real-world scenarios. The best method to understand a complex system is to draw diagrams or representations of it.

We design UML diagrams to help us understand the system more precisely and simply. To represent all features of the system, a single diagram is insufficient. UML defines a variety of diagrams to address the majority of a system's characteristics. The creation of UML was prompted by the misunderstanding surrounding software development and documentation. As a general-purpose modelling language, UML has primarily been used in the field of software engineering.

They provide a more consistent approach to workflow modelling as well as a more comprehensive collection of features to improve readability and efficiency. The easiest method to uncover use cases is to look at the actors and determine what the system will allow them to do. Because all of a design's requirements are unlikely to be met by a single-use instance, it's common to have a group of these.

This use case collection details all of the possible applications for the system. A communication channel is provided by an association. Use cases, actors, classes, and interfaces can all communicate. Associative relationships are the most general of all relationships and, as a result, the conceptually weakest. The relationship between two objects is referred to be an association when they are generally treated independently. They provide a more consistent approach to work flow modeling as well as a more comprehensive collection of features to improve readability and efficiency. The best method to uncover use cases is to look at the actors and describe what they'll be able to accomplish using the system. Because all of a design's requirements are unlikely to be met by a single-use instance, it's common to have a group of them.

By default, the toolbox's association tool is uni-directional and depicted as a single arrow at one end of a diagram. Who or what is receiving the message is indicated by the arrow at the conclusion of the message. Two model elements have a relationship in which a modification to one has an influence on the other is known as dependency.

A dependence connection on a class diagram often shows that the client's operations invoke the supplier's operations. The work flow in this case begins from importing the data set by the developer and then replacing missing values with mean value of corresponding column, model building, validating that model by generating a confu- sion matrix and finally predicting the test sample class label. The usage of transitions is employed to depict the control flow from one task to the next.

These diagrams help comprehend the designer's viewpoints and can effectively ex- plain the goals behind the system's design. There are eight different types of UML Diagrams in use, each with its style of communicating the design.

The various UML diagrams are:

(1) Use case diagram
(2) Activity diagram
(3) Sequence diagram
(4) Class diagram

Use case Diagram. An actor graph, a set of use cases separated by a system boundary, communication (participation) linkages between actors and users, and gen- eralisation between use cases make up a use instance diagram. The use case design specifies the system's behaviour from both outside (actors) as well as inside (users) perspectives (use case). The system does not involve actors. Actors are the people or things who interact with the system (provide input or receive output). Through-out the analysis phase, Use case diagrams as necessary to define design requirements and exemplify how the system should function. Use-case diagrams might help you to clarify the behaviour of the design as it is constructed throughout the design process. The use cases are all of the system's possible applications. A use case is a sequence of activities or moments, usually seven in number, that explain how a role (known as an actor) interacts with a design to accomplish a goal. The developer and the end-user are the actors in the use case diagram.The best method to uncover use cases is to look at the actors and explain what they will be able to do with the system. Because a single-use case is unlikely to meet all of a system's needs, it's typical to have a collection of them. The entire use case collection specifies.

A prototypical connection between a base use case and an inclusion use case is an include connection. The behaviour from the inclusion use case is utilized in the basic use case using an included connection. An extended connection is a standardised communication that describes how one use case's feature work can be combined with that of another. The Extend archetype is used to represent interdependence between uses cases. Some of the different use cases include importing data sets, preprocessing, model development, validation, and prediction.

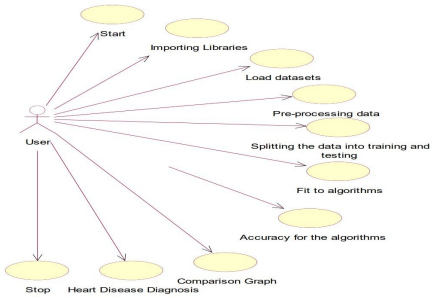

Fig. 1 Use case diagram

Activity Diagram

The states of a state machine are activities that reflect the execution of operations and the completion triggers the transitions of those operations is known as an activity diagram. The purpose of an activity diagram's is to depict flows and what's happening within a use case or between many classes. Activities, transitions between activities, decision points, and synchronization bars are all included in activity diagrams.

The performance of behavior in the workflow is represented by an activity. In the UML, transitions are represented by directed arrows, decision points are represented by diamonds, and synchronisation is represented by rectangles with rounded corners. As indicated in the diagram, 8 bars are drawn horizontally or vertically as thick horizontal or vertical bars. The activity icon is a rectangle with rounded ends that has a name and an action component.

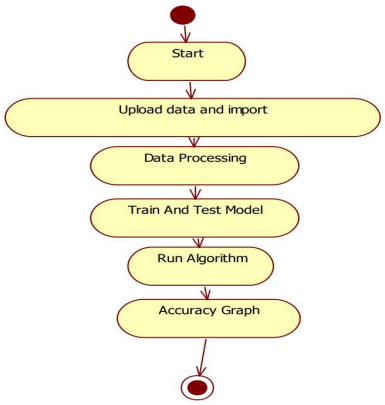

Fig.2 Activity Diagram

Sequence Diagram. A sequence diagram is an interaction diagram that displays how and in what order processes interact with one another. It's a Message Sequence Chart in action. The interactions between items are shown in chronological order on a sequence diagram. It shows the scenario's objects and classes, as well as the messages that are sent between them to carry out the scenario's functionality. In the Logical View of the system under development, sequence diagrams are usually coupled with the use of case realizations. Event diagrams are another name for sequence diagrams

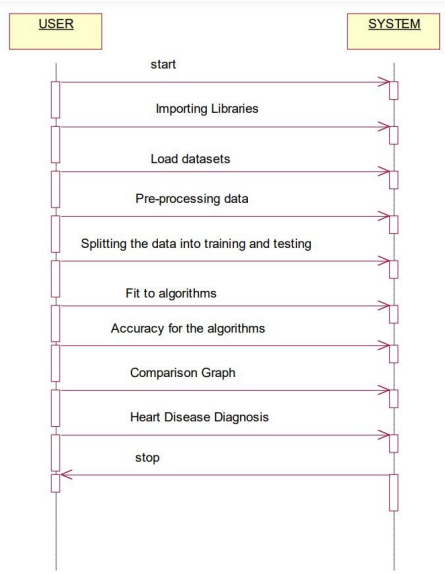

Fig.3 Sequence Diagram

Class Diagram. Icons indicate classes, interfaces, and relationships in class dia- grams. The classes at the top-level of the current model should be depicted, so that you can design one or more class diagrams; these The present model's top level has class diagrams as well. You can also draw one or few class diagrams to represent the classes in each of your model's packages; the package encompassing the classes they represent contains these class diagrams; class diagrams utilise icons to represent logical packages and classes.

Class diagrams are used to show an image or view of one or more of the model's classes.

In the logical view of the model, the primary class diagram is usually a pic- ture of the system's packages. Every package also has its own primary class diagram, which usually shows the "public" package's classes.

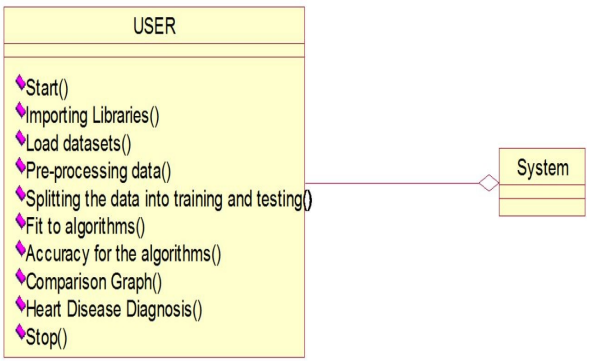

Fig.4. Class Diagram

5. Implementation

Modules.

(1) Importing Libraries
(2) Load data sets
(3) Pre-processing data
- DBSCAN
- SMOTE-ENN for balancing the data

(4) Splitting the data into training and testing
(5) Fit to algorithms
(6) Algorithms
(7) Accuracy for the algorithms
(8) Comparison Graph
(9) Front application (flask)
- Heart Disease Diagnosis

Parameters. Name, Age, Gender, Constrictive pericarditis (CP), The person's max- imum heart achieved, old peak, calcium, thalassemia

Data Sets.

1. Statlog Dataset: The Statlog Heart Disease database at the Uni- versity of California Irvine (UCI) Repository presents dataset I for the study of heart disease. The initial dataset contains 270 subjects with 13 attributes and one output class. 120 subjects were found to be positive (presence of disease) and 150 subjects were found to be negative (absence of disease).

2. Cleveland Dataset: Cleveland Dataset II (Cleveland Heart Disease dataset) was supplied by Dr. Robert Detrano, M.D., to research heart disease. Only 13 of the 303 individuals and 79 raw attributes in the original data set are used, 12 as input classes and one as output class. Study excluded 6 patients' observations due to missing values in data and in the pre-processing stage, the remaining 297 data were employed.

The attributes are: Age, Sex, Cp, Trestbps, Chol, Fbs, Restecg, Thalach, Exang, old peak, Slope, Ca, Thal(input classes); Target(output class).

Sequence of steps followed in the design:

Step 1: Importing essential libraries

Step 2: Importing and understanding our data set

i: Verifying it as a 'data frame' object in pandas

ii: Shape of data set

iii: Printing out a few columns

iv: Description

v: Let's understand our columns better

vi: Analysing the target variable

vii: Exploratory Data Analysis (EDA)

viii: Analysing the 'Chest Pain Type' feature

ix: Analysing the FBS feature

x: Analysing the restecg feature

xi: Analysing the 'exang' feature

xii: Analysing the Slope feature

xiii: Analysing the 'ca' feature

Step 3: Train Test split

Step 4: Model Fitting

- Logistic Regression
- Naive Bayes
- Support Vector Machine
- K Nearest Neighbours:
- Decision Tree
- Random Forest

Step 5: Output final score

6. Testing

Testing is the process of running a software to look for errors. Our software must be error-free in order to function properly. If the testing is completed successfully, the software will be free of all errors. Types of testing are:

(1) White Box Testing
(2) Black Box Testing
(3) Unit testing
(4) Integration Testing
(5) Alpha Testing
(6) Beta Testing
(7) Performance Testing and so on

7. Summary & Conclusions

We proposed a useful model for predicting cardiac disease to increase prediction accuracy. Two freely available cardiac disease datasets were used to develop a broad forecasting model. In an assessment study, we compared our suggested design to existing categorization models as well as the outcomes of earlier investigations. We further created and implemented the proposed model into the clinical support system to accurately and efficiently diagnose the subjects'/patients' heart disease state. All of the communicated diagnosis data was then saved in MongoDB, a database that can efficiently respond quickly to rapidly growing medical data. As a result, the proposed model is likely to assist clinicians recognize patients more accurately and efficiently, as well as improve clinical decision-making in cardiac disease.

References

[1] Ali.L, Niamat.A, Khan.J.A, Golilarz.N.A, Xingzhong.X, Noor.A, Nour.R and Bukhari.S.A.C An optimized stacked support vector machines based expert system for the effective prediction of heart failure, IEEE Access, 7:54007-54014, 2019,doi: 10.1109/ACCESS.2019.2909969.

[2] Dwivedi.A.K, Performance evaluation of different machine learning techniques for prediction of heart disease, Neural Comput. Appl., 29(10):685-693, May. 2018, doi: 10.1007/s00521-016-2604-1.

[3] Haq.A.U, Li.J.P, Memon.M.H., Nazir.S and Sun.R A hybrid intelligent system frame- work for the prediction of heart disease using machine learning algorithms, Mobile Inf. Syst., 2018:1-21, Dec.2018,doi: 10.1155/2018/3860146.

[4] Latha .C.B.C and Jeeva.S.C, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Inform. Med. Unlocked., 16(100203): Jan.2019,doi: 10.1016/j.imu.2019.100203.

[5] Long.N.C, Meesad.P and Unger.H, A highly accurate firefly based algorithm for heart disease prediction, Expert Syst. Appl., 42(21):8221-8231, Nov.2015,doi: 10.1016/j.eswa.2015.06.024.

[6] Nahato.K.B, Harichandran.K.N and Arputharaj.K, Knowledge mining from clinical datasets using rough sets and backpropagation neural network, Comput. Math. Meth- ods Med., 2015:1-13, Mar.2015,doi: 10.1155/2015/460189.

[7] Verma.L, Srivastava.S and Negi.P.C, A hybrid data mining model to predict coro- nary artery disease cases using non-invasive clinical data, J. Med. Syst., 40(7):178, Jul.2016,doi: 10.1007/s10916-016-0536-z.