A Novel Computational Rough Set Based Feature Extraction For Heart Disease Analysis

Dr.R.Prabha[1], G.A.Senthil[2], Dr.A.Lazha[3], Dr.D.VijendraBabu[4], ,Ms.D.Roopa[5]
{r.praba05@gmail.com[1],, senthilga@mail.com[2], lazhadinesh@gmail.com[3],
drdvijendrababu@gmail.com[4], roopa.cse@sairamit.edu.in[5]}

Associate Professor,,SriSairam Institute of Technology, Tambaram, Chennai[1], Research Scholar, Vels Institute of Science, Technology & Advanced Studies, Chennai[2], MD(S) Reader, Dept of UdalThathuvam, Sri Sairam Siddha Medical College & Research Centre, Chennai[1]., Professor & Vice Principal, Aarupadai Veedu Institute of Technology, Vinayaka Mission's Research Foundation[4], Assistant Professor, Sri Sai Ram Institute of Technology, Tambaram, Chennai[5]

Abstract. Cardiovascular disease is the most difficult disease to diagnose in the medical field. The diagnosis is often contingent on a judgment based on the grouping of vast amounts of clinical and pathological data. As a result of this complication, a number of researchers have based their efforts on determining the most cost-effective and accurate way to predict heart disease.In the case of heart disease, an accurate diagnosis at an early stage is critical, since time is of the essence when heart disease is detected at an inopportune time. Machine learning has evolved in recent years with a plethora of accurate and supporting resources in the medical domain, and it has offered the best support for predicting disease with proper training and research. The main goal of this study is to use a rough computational intelligence approach to find specific heart disease features among a large number of features. The output of the proposed feature selection method outperforms that of conventional feature selection approaches. The rough computation approach's output is evaluated using various heart disease data sets and checked using real-time data sets.

Keywords: Please list your keywords in this section.

1 Introduction

Based on clinical evidence generated by patients, a heart disease prediction system will assist medical professionals in predicting the condition of the heart. When it comes to diagnosing a patient's heart disease, doctors may still make mistakes [1][14]. As a consequence, in order to obtain reliable outcomes, heart disease prediction systems employ machine learning algorithms. The use of adequate therapy and medications is needed for early detection and proper diagnosis of heart disease. Machine learning techniques have the ability to greatly assist in medical diagnosis[2]. For heart disease prediction, various supervised machine learning methods such as decision trees, support vector machines, naive bayes, random forests, and neural networks can be used, and all must be evaluated in terms of output[3][13].

Diabetes, a family history of heart disease, smoking, obesity, high cholesterol, and low cholesterol are all major risk factors for heart disease. The biggest challenge in the health

I3CAC 2021, June 07-08, Chennai, India

DOI 10.4108/eai.7-6-2021.2308575

domain is to identify the key features that cause heart disease. The aim of this study is to define the key characteristics that are used to diagnose heart disease in order to solve the heart attack problem in a safe society [4]. In the healthcare industry, effective and reliable automated heart disease prediction systems can be useful for heart disease prediction. The number of tests a patient must take will be reduced as a result of this automation. As a result, it would save not only money but also time for both doctors and patients [5][6].

The best feature selection approach for finding the main features that cause heart disease is often critical for good heart disease prediction models. Principal component analysis, wavelet transformation, single value decomposition process, linear discriminant analysis, and minimum noise ratio methods have all been used to select key features from large dimensional textile data in previous studies[7][12][15]. To address the shortcomings of current approaches, a rough computational intelligence-based attribute selection algorithm is proposed to identify the key attributes that play a role in predicting heart disease. The rough set soft computing approach is used to determine the value of the attributes. With this detailed introduction Section 2 describes the design and procedure for proposed feature selection approach, Section 3 depicts the comparison results followed by conclusion in Section 4.

2 DESIGN AND IMPLEMENTATION

The theory, design, implementation, and creation of biologically and linguistically motivated computational paradigms is known as computational intelligence (CI). Rough Set, Neural Networks, Fuzzy Systems, and Evolutionary Computation have traditionally been the three four pillars of CI. Successful intelligent systems, such as medical and cognitive developmental systems, rely heavily on CI. The aim of this study is to use a rough set approach to create a good feature selection model.

2.1 Rough set attribute dependency

Rough set theory is a valuable data mining method. The definition of a simple rough collection has been expanded in several different ways in recent years. Rough set theory can be generalised in three ways: set-theoretic structure with non-equivalence binary relations; granule-based description with coverings; and subsystem-based definition with other subsystems[8]. The definition of rough set theory is depicted in Figure 1.

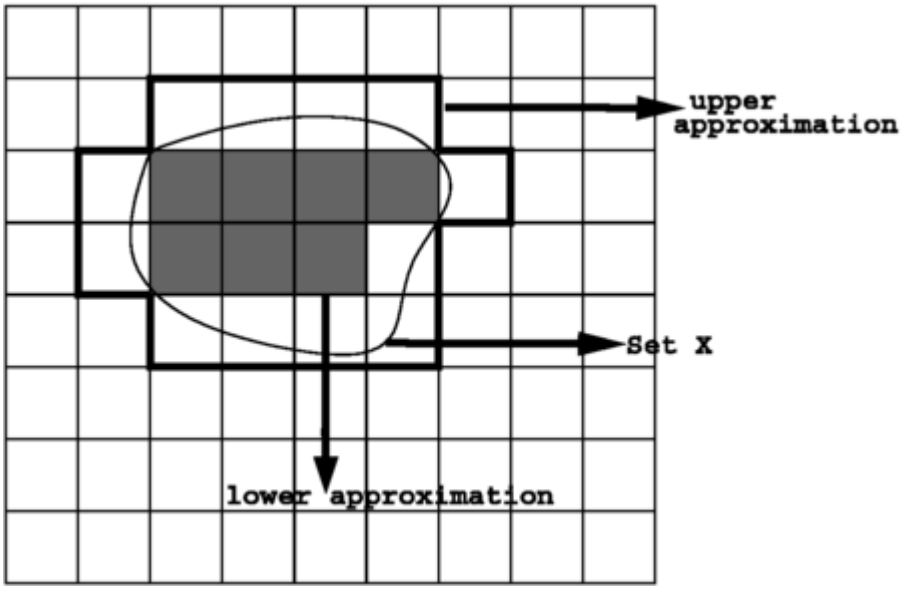

Figure 1. Rule generation using rough set

In a rough collection, attribute dependency can be described as the use of a combination of objects to determine the values of attributes. This determines how dependent the two attributes are on each other. Any attribute causes indistinguishability between the two rough sets. In rough sets, the principle of indiscernibility relation describes the relationship between a set of attributes[9]. The same indistinguishable artefacts may be described multiple times. The following is a list of attributes A that are dependent on attributes B:

$A \Rightarrow B$

Where the values of A's attributes are independently determined by the values of B's attributes. If there is a functional dependency relationship between them, A is entirely dependent on the values of B. Both A and B are subsets of each other, as seen by the functional dependence. If B is absolutely dependent on C, then with a degree of k, it can be written as $k(0 <= k <= 1)$ denoted by the positive region,

$A \Rightarrow_k B$

The degree to which *x* is dependent on A is determined by the value of k. If k=1, we can say that B is fully dependent on A, and if k<1, we can say that B is partially dependent on A. The degree of dependence is expressed by the k, which expresses the blocks of I/B with respect to A[10][11].

2.2 Rough set based feature selection procedure

Rough set theory's definition of reducts and core is used to evaluate the significant attributes. Finding redundant values or redundant attributes of a set becomes easier thanks to this indiscernibility relationship. Reductions are the various set approximation subsets of attributes that occur in minimal. A core is defined as the intersection of all the reductions to a set or a system considered, and it is the set of all the conditional attributes of set approximations that exist as a set. For instance, if *A* is a set of attributes, and *r* is a subset of*e*, then the diagram looks like this:

$A = (U, r, d)$ &$P_R(d) = P_C(d)$

where, core (*e*) defines all the conditional attributes, and *r*(*e*) defines all the set of reducts of attribute *e* .One method for calculating these reducts or conditional attributes is to use

decision tables, which are dynamically generated. The attributes are listed in two ways in these decision tables: important and frequently. When attributes are replicated often, they are given the status of majority or significant, and the collection of attributes that tend to be common to the original sets in decision tables is given priority. The rough computational intelligence based attribute selection algorithm is proposed based on the rough set theory principles core and reduce.

Significant attributes can be described as the removal of unnecessary data from a decision table or information table without affecting other data in the table. As a result, the value of attributes is used to generalise the reduction of redundant attributes. Prior to determining the value of attributes, they must first be assessed. As a result, dispensability and indispensability are included in this assessment. A very similar interval [0,1] can also be used to achieve this.

By removing attributes from the attribute collection, the process of acquiring significant attributes in a decision table can be completed.For a set considered as $\beta(r, e)$ let the attribute be in a set. And when the attribute a is removed from the set $\beta(r, e)$ then it can be given as,

$\beta((r \Leftrightarrow a, e))$

Then by the above conditions and processes the significance of attributes can be given as by normalizing the basic difference in between the coefficient and the set obtained after removing the attribute i.e.$\beta(r, e)$ and $\beta((r \Leftrightarrow a, e))$ is given as below:

$\alpha(r,e)(a) = (\beta(r <=> a, e))/(\beta(r, e))$

Thus here, The coefficient $\alpha(a)$ is termed as error of classification. This error of classification in general occurs when the attribute is removed from the set considered. And so the significance of the attributes can be protracted by the other remaining attributes of a set and can be given as,

$\alpha(r,e)(x)=(\beta(r, e))<=> (\beta(r<=>x,e))/(\beta(r, e))$

Here, $\alpha(x)$ is given as the coefficient obtained from the extension of an attribute significance. Also x is considered as subset of r i.e. x is a reduct of the set of attributes in r. The attribute of any subsets x and r is deliberated as the reduct of r and so after removing the attribute this can be given as,

$\alpha(r,e)(x) = (\beta(r, e)) <=> \beta(x, e)/(\beta(r, e))$

Thus, $\alpha(r, e)$ is defined as the reduct approximation or error of reduct approximation which depicts the significance of attributes of x relatively in the set r. Via a classification method, the minimum error of approximation leads to an improvement in accuracy in a series. On heart disease data sets, the proposed Rough Computational Intelligence based Attribute Selection method is used to find the most important features that cause heart diseases in the health domain.

3 EVALUATION RESULTS

Using rough set techniques, a rough computational intelligence-based attribute selection algorithm is built in Matlab to find the significant attributes among the various conditional attributes. The attributes will be listed using the attribute significant value derived from the computational programme. The significance value of the Hungarian data set is shown in Table 1.

Table 1. Significance Value of Hungarian Data Set

Feature	**Value**
Age	0.6673
Sex	0.6216
Cp	0.6537
trestbps	0.6690
chol	0.6554
fbs	0.7795
restecg	0.6775
thalach	0.6486
Exang	0.9887
Oldpeak	0.6690
Slope	0.6673
Ca	0.6809
Thal	0.6758

The graphical representation of each heart disease feature's significance value with respect to Hungarian data sets is shown in Figure 2. The graph depicts the importance of each trait in the diagnosis of heart disease. Exang (exercise-induced angina) and fbs (fasting blood sugar) contribute significantly more than other factors. The attribute significant values of the Cleveland data set are shown in Table 2.

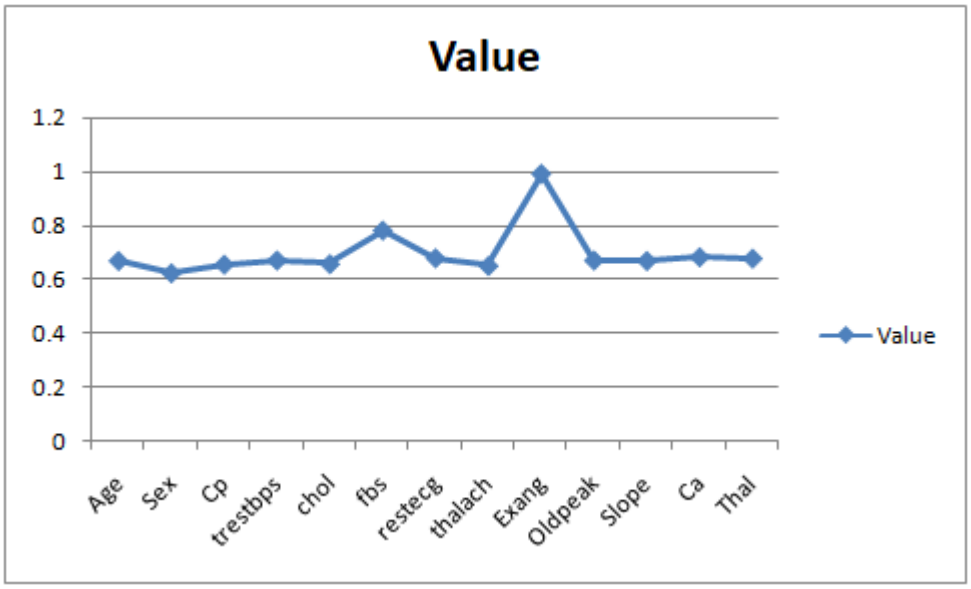

Figure 2. Hungarian Heart disease data set features with significance value.

The graphical representation of each heart disease feature's importance value in relation to Cleveland data sets is shown in Figure 3. The graph depicts the importance of each trait in the diagnosis of heart disease. Fasting blood sugar, exercise-induced angina, ca (number of main vessels), and resteog (resting electrocardiographic results) all play a bigger role in heart disease prediction than other factors.Based on the proposed procedure result analysis on Hungarian and Cleveland data set, Figure 4 shows that, the important attributes, are exang(exercise induced angina) ,fbs(fasting blood sugar),ca (number of major vessels),restecg (resting electrocardiographic results),thal,slope (the slope of the peak exercise ST segment),oldpeak (ST depression induced by exercise relative to rest) and trestbps (resting blood pressure).

Table 2. The attribute significant values of Cleveland data set.

Feature	**Value**
Age	0.5036
Sex	0.4856
Cp	0.4986
trestbps	0.5069
chol	0.4953
fbs	0.5564
restecg	0.5267

thalach	0.5069
Exang	0.5300
Oldpeak	0.5217
Slope	0.5267
Ca	0.5316
Thal	0.5250

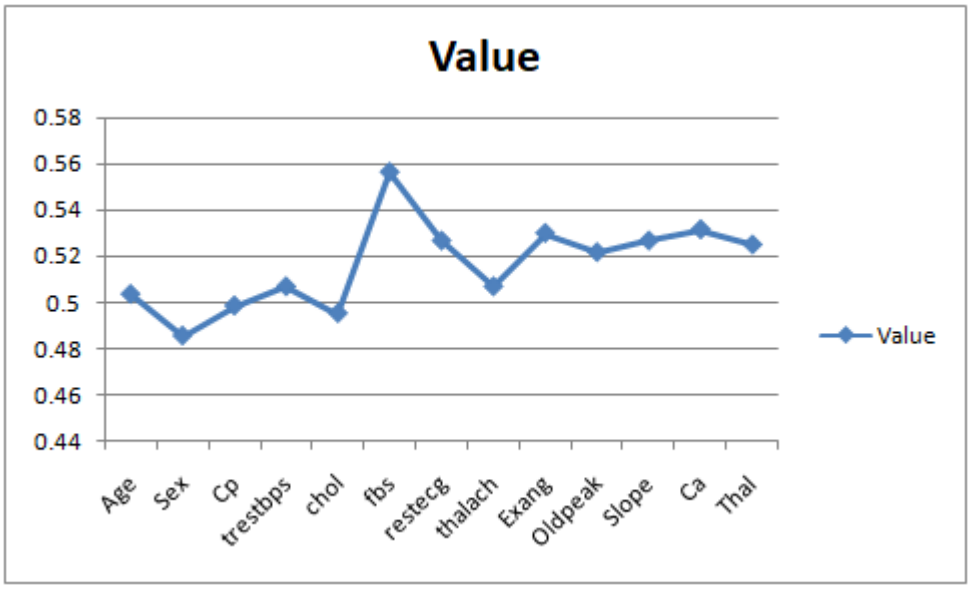

Figure 3. Cleveland Heart disease data set features with significance value

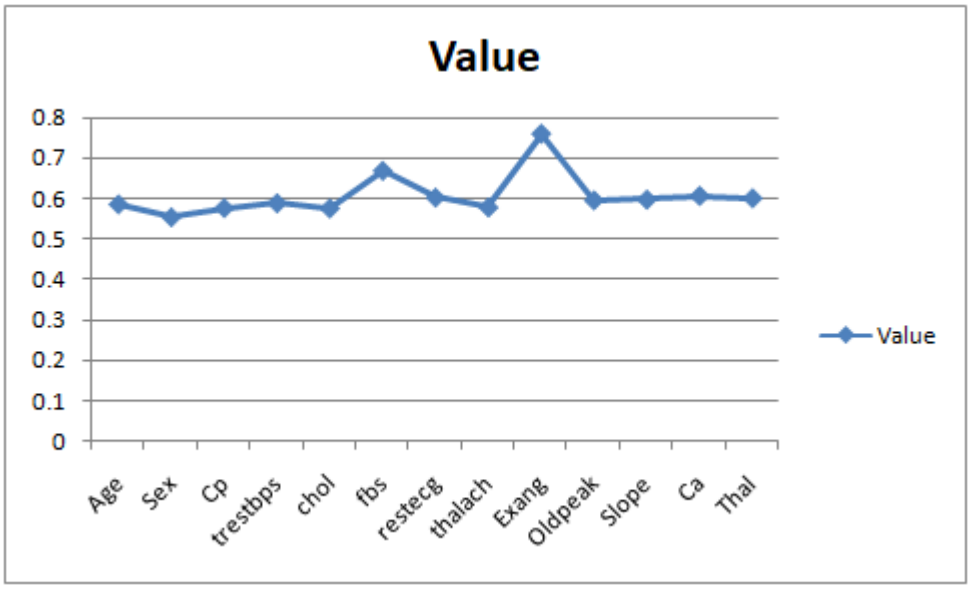

Figure 4. Significance of Features based on both the data sets

Based on the proposed algorithm result analysis on Hungarian and Cleveland data set, Figure 4. shows that, the significant attributes, are exang(exercise induced angina) ,fbs(fasting blood sugar),ca (number of major vessels),restecg (resting electrocardiographic results),thal,slope (the slope of the peak exercise ST segment),oldpeak (ST depression induced by exercise relative to rest) and trestbps (resting blood pressure).To predict heart disease, the significant attribute will be given to the data mining prediction algorithm. The main goal of this project is to solve the problem of dealing with high-dimensional data. In the health domain, the proposed algorithm is used to solve this problem.

4. CONCLUSION

Heart disease diagnosis is complex because it relies on the grouping of vast amounts of clinical and pathological data. A rough computational intelligence-based attribute selection algorithm is proposed to find significant heart disease features among the large number of features, and the algorithm's output is checked with important heart disease data sets Cleaveland and Hungarian. The proposed algorithm was also compared to a variety of feature selection methods used in the prediction of heart disease. Observation reveals that the proposed algorithm outperforms other feature selection approaches in the vast majority of cases. The proposed feature selection technique suggests that exang, fbs, ca, restecg, thal, slope, oldpeak, and trestbps are the most significant features for heart diseases. For the creation of novel heart disease prediction models, a good feature selection algorithm would be more useful.

References

[1] Bashir, Saba & Khan, Zain & Khan, Farhan&Anjum, Aitzaz& Bashir, Khurram. (2019). Improving Heart Disease Prediction Using Feature Selection Approaches. 619-623. 10.1109/IBCAST.2019.8667106.

[2] El-Bialy, Randa& A. Salama, Mostafa&Karam, Omar &Khalifa, M.. (2015). Feature Analysis of Coronary Artery Heart Disease Data Sets. Procedia Computer Science. 65. 459-468. 10.1016/j.procs.2015.09.132.

[3] Gárate Escamilla, Anna &Hajjam, Amir &Andrès, Emmanuel. (2020). Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked. 19. 100330. 10.1016/j.imu.2020.100330.

[4] Güzel Aydın, Seda& Kaya, Turgay&Guler, Hasan. (2016). Heart Rate Variability (HRV) Based Feature Extraction for Congestive Heart Failure. International Journal of Computer and Electrical Engineering. 8. 272-279. 10.17706/IJCEE.2016.8.4.272-279.

[5] LE, HUNG & TRAN, TOAN & Tran, Lang. (2018). AUTOMATIC HEART DISEASE PREDICTION USING FEATURE SELECTION AND DATA MINING TECHNIQUE. Journal of Computer Science and Cybernetics. 34. 33-48. 10.15625/1813-9663/34/1/12665.

[6] Martiana, Entin&Barakbah, Ali &Hermawan, Aditya. (2018). Feature Extraction For Application of Heart Abnormalities Detection Through Iris Based on Mobile Devices. EMITTER International Journal of Engineering Technology. 5. 312. 10.24003/emitter.v5i2.202.

[7] Methaila, Aditya &Kansal, Prince & Arya, Himanshu. (2014). Early Heart Disease Prediction Using Data Mining Techniques. Computer Science & Information Technology. 4. 53-59. 10.5121/csit.2014.4807.

[8] Muthuvel, Marimuthu&Abinaya, M &Hariesh, K &Madhankumar, K &Pavithra, V. (2018). A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach. International Journal of Computer Applications. 181. 975-8887. 10.5120/ijca2018917863.

[9] Raju, V & Devi, G &Sampath, A &Achanta, Sampath&Saikumar, K &Jagan, B.Omkar. (2020). HEART DISEASES CLASSIFICATION AND FEATURE EXTRACTION BY SEGMENTATION AND MACHINE LEARNING MODEL. 24. 8992-9001.

[10] Ramalingam, V V&Dandapath, Ayantan& Raja, M. (2018). Heart disease prediction using machine learning techniques: A survey. International Journal of Engineering & Technology. 7. 684. 10.14419/ijet.v7i2.8.10557.

[11] Ravindran, Kavitha&Kannan, E.. (2016). An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. 1-5. 10.1109/ICETETS.2016.7603000.

[12] Sun, Shuping. (2015). An innovative intelligent system based on automatic diagnostic feature extraction for diagnosing heart diseases. Knowledge-Based Systems. 75. 224-238. 10.1016/j.knosys.2014.12.001.

[13] Yadav, Dhyan& Pal, Saurabh. (2020). Prediction of Heart Disease Using Feature Selection and Random Forest Ensemble Method. International Journal for Pharmaceutical Research Scholars. 12. 56-66. 10.31838/ijpr/2020.12.04.013.