iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** IV **Month of publication:** April 2019

DOI: https://doi.org/10.22214/ijraset.2019.4557

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

Prediction of Cardiovascular Disease – A Review Publication

Mr. Vishwesh Hande[1], Prof. Supriya Kamoji[2]
[1, 2]*Deptt of Computer Engg., Fr. Conceicao Rodrigues College of Engineering, Mumbai, India*

Abstract: ***Cardiovascular diseases have become a leading cause for deaths across the world. A comprehensive approach for the successful prediction and timely intervention by medical specialists is of essence. Data mining and machine learning concepts can be effectively combined in order to develop decision systems that will greatly benefit both patients and professionals. There exist several techniques for the implementation of data mining, and valuable conclusions can be drawn using each of these.***
Index Terms: Cardiovascular; diseases; data mining; machine learning; decision systems

I. INTRODUCTION

A large number of information systems have been developed and deployed in hospitals across the world. These are used to maintain extensive records of parameters essential to the patient, i.e. patient history, blood tests, X-rays, etc. However, most of this information is largely arbitrary and remains untapped, and does not help the medical specialist or the patient in predicting anomalies, with respect to health of the patient.

Therefore, it is essential for systems to be developed in order to tap into the vast resources that already exist, and obtain meaningful and actionable inferences. [1][2]

In order to extract meaningful inferences from repositories of data, data mining is utilized. It effectively combines statistical models and analysis, machine learning techniques and database technology; and is widely used in a variety of domains. These are used to obtain hidden patterns and relationships in large databases. [3]

Data mining, according to Fayyad, is defined as “a nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database”. [4] Giudici defined the same as “a process of selection, exploration and modelling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database”. [5] Data mining, as is machine learning, utilizes two primary approaches: supervised and unsupervised learning.

1) *Supervised Learning:* Supervised learning involves training a data sample from specific data sources, with information of classification of samples embedded within the data. It is a largely efficient methodology of finding solutions to complex problems and has special application in domains that require predictive analysis. [6]
2) *Unsupervised Learning:* The input data set utilized to train the system is largely unlabeled. This method of mining and predictive analysis is utilized in order to efficiently find relationships existing between independent data items. Usually, a self-organizing network is used for implementation of unsupervised learning methodology for data mining. [6] The purpose of this paper is to provide an overview of multiple approaches to build data mining systems, in order to extract inferences and successfully predict the presence of cardiovascular abnormalities in patients, based on a specific set of parameters. This paper first, explains some basic terminologies and then moves into critiquing multiple approaches for the same.

II. ARTIFICIAL NEURAL NETWORKS

Artificial neural networks (ANNs) are mathematical models that describe a function and are associated with a specific learning algorithm. [6]

These networks are based on human neural networks, complete with synaptic connections, in order to emulate decision making abilities and work effectively to arrive at well founded conclusions. In general, artificial neural networks are classified on the basis of the following parameters: (a) interconnection property; (b) application function (optimization, classification, etc); and (c) type of learning (as discussed earlier). [7]

III. GENETIC ALGORITHM

Most data sets that are utilized to train a system have a number of parameters and dimensions associated with them. In order to reach a valid conclusion regarding a specific pattern within the dataset, certain specific features and/or dimensions have greater impact on the eventual result. [2]

It is important to note that in several cases, there exist certain redundant parameters that have null impact on the overall result. The genetic algorithm mimics evolutionary concepts of genetics and natural selection in order to ensure optimum feature selection. It reduces the overall dimensionality of the data and increases the overall efficiency of inference gathering. This saves storage and reduces computation time.

In hybrid systems, an initial population of "chromosomes" is initialized. Chromosomes are the simplest solutions to a specific problem. Selection, crossover and mutation operators are applied to the population, repeatedly until a desirable population stage is reached. [8] Genetic algorithms are widely used in hybrid neuro-fuzzy approaches for predictive analytics and data mining.

IV. GENETIC ALGORITHM BASED RECURRENT

A. Fuzzy Neural Network

The Cleveland dataset (UCI, 1990) was utilized. The dataset includes four independent databases from four different medical institutions.

The data set was classified into training (252 instances) and testing (45 instances).

The performance of the overall system is calculated in terms of the Root Mean Square Error (RMSE), sensitivity, precision, F-score, probability of misclassification error and accuracy of the training set and testing set. The utilization of a GA based RFNN greatly increased the accuracy and precision of the predictive model.

The method is open for use with multiple types of data sets, in order to provide a more all encompassing approach to the prediction. [2]

V. CLINICAL DECISION SUPPORT SYSTEM: RISK LEVEL PREDICTION USING DECISION TREE FUZZY RULES [9]

The data set used for the implementation of this approach is the UCI data set obtained from the Data Mining Repository of the institution. Specific parameters contained within the data set are utilized in order to arrive at conclusions for the Clinical Decision Support System (CDSS).

These include the following: age, sex, cholesterol levels, smoking status, hypertension and pre-eclampsia.

In order to account for the variability in results, the soft computing technique of fuzzy logic is adopted in this approach. The fuzzy rules are assigned weights depending upon their relative importance. The data is first pre - processed so that the format of all the data items is uniform. This is important as there are missing values, null points, etc. that exist within the data set. The data set is then classified into explicit categories, with chances of cardiovascular disease existing within the patient, at a risk of 50% or less; or more.

Following this, the weighted fuzzy rules are generated in order to make the system learn effectively. For this purpose, the most relevant attribute in terms of its frequency is mined from the data set. Next, based on the frequency and the weights assigned, the fuzzy rules are generated. Performance analysis is carried out using specificity, sensitivity and accuracy.

Importantly, this method involves the deployment of an actual support system with ability to mimic human decision making abilities and greatly aid medical professionals. However, an important drawback of the decision tree approach is its inability to adapt to larger and more voluminous data sets.

VI. INTELLIGENT HEART DISEASE PREDICTION SYSTEM USING DATA MINING TECHNIQUES [10]

Multiple approaches are adopted for the development of a comprehensive Intelligent Heart Disease Prediction System (IHDPS). The system utilizes the CRISP-DM methodology for the development and implementation. This includes the following: business understanding, data understanding, data preparation, modeling, evaluation and deployment. In order to access the data models' contents, DMX (Data Model eXtension), an SQL based querying language is utilized. Furthermore, data visualizations for trend analysis is also incorporated.

The Cleveland heart disease data set of the UCI repository is utilized for implementation. The training data set consisted of 455 records and the testing data set consisted of 454 records. The records were selected independently to account for bias mitigation. Parameters were set to default values, except "Minimum Support = 1" for decision tree learning and "Minimum Dependency

Probability = 0.005" for Naïve Bayes classifier. Lift charts and classification matrix were incorporated to evaluate the effectiveness of the models implemented.

A total of five mining goals were set for the approaches to satisfy and each technique was evaluated in terms of the number of goals that it could achieve.

A goal of importance, is "To identify characteristics of patients with heart disease". In decreasing order of effectiveness, in terms of goal attainments, Naïve Bayes, followed by Neural Networks and Decision Trees.

This approach is creditable as it provides multiple methodologies for the development of the IHDPS. It conclusively proves the superiority of a specific technique of data mining over the others, by means of explicitly stating what goals can and cannot be achieved. Possible drawbacks are due to the limited utilization and applicability of DMX for data modelling.

VII. CARDIOVASCULAR DISEASE PREDICTION USING GENETIC ALGORITHM AND NEURO-FUZZY SYSTEM [11]

This approach adopts the benefits that genetic algorithm and a hybrid neuro-fuzzy system have to offer. The genetic algorithm is used to reduce the dimensionality of the data set. By means of selection, crossover and mutation operators, the fitness function is used to evaluate the overall strength of a population.

Subsequently, fuzzy logic is applied to a selected set of attributes. In this methodology, the UCI Cleveland data set is used. The data set is known to possess ground truth and hence makes it applicable to the multi-layer perceptron network. Selected attributes from the data set, are assigned multiple values owing to variability.

Upon research, a valid conclusion to ensure maximum success in terms of precision and accuracy of results, is the utilization of a back propagation network. This will allow for error correction and comparison. This method works well with larger data sets owing to its ability for feature selection and reduced dimensionality. This greatly increases efficiency, reducing storage and computation costs.

VIII. EVALUATION CRITERION

In this section, we will discuss the criteria utilized for evaluation of the different methodologies used for cardiovascular abnormality prediction. In most methodologies that we have reviewed and surveyed, the common summarization and evaluation techniques used are the Root Mean Square Error (RMSE) and the evaluation metrics. [10] The evaluation metrics widely used for the purpose of performance analysis are: specificity, sensitivity and accuracy.

In order to calculate these metrics, we need to compute terms like True Positive (TP), True Negative, False Positive (FP) and False Negative (FN). [9]

Sensitivity = TP/(TP + FN) (1)

Specificity = TN/(TN + FP) (2)

Accuracy = TN + TP/(TN + TP + FN + FP) (3)

The RMSE method reflects the difference between the actual result and the predicted results of diagnosis. Additionally, F-score and Probability of Misclassification Error (PME) have been adopted to evaluate performance of methodology. All of these techniques are valid performance evaluation techniques, and can give a general idea to the researcher of the effectiveness of a specific method.

IX. COMPARISON OF METHODS

A number of methods for prediction of cardiovascular disease have been reviewed. A comparison of these methodologies and approaches has been made in the table (Tab. 1) that follows. The bases of comparison are the data sets and selection of parameters, the individual evaluation criteria and the results accrued. We find that different approaches have varying degrees of performance, that we have summarized in the table.

Specifically, with advancement in technology, the evolvement of approaches has resulted in increased levels of accuracy and specificity.

It is understood that while conditions for comparison are varied, an effort has been made to ensure uniformity for the sake of fairness.

Heart Disease prediction methodology	Dataset and Parameters	Performance Analysis Measure	Results
Genetic Algorithm based Recurrent Fuzzy Neural Network	Cleveland UCI Dataset, 1990; 14 parameters used	RMSE, Sensitivity, Specificity, Precision, Accuracy, F-score, PMSE	Achieved excellent prediction for specific instances lacking heart diseases; high sensitivity
CDSS: Risk Level Prediction using Decision Tree Fuzzy Rules	Cleveland, Hungarian and Switzerland UCI Dataset; 6, 8, 11 parameters respectively	Sensitivity, specificity and accuracy	Sensitivity and specificity measures found to be high; Comparative study of 3 variations achieved
Intelligent Heart Disease prediction using Data Mining Techniques	Cleveland UCI Dataset, 1990; 15 attributes; Instances randomly selected	Five “mining goals” as ascertained and tested experimentally based on business intelligence and data exploration	Naïve-Bayes approach exhibited best performance; followed by Decision trees and neural networks; training model for nurses and specialists
Genetic Algorithm and Neuro-Fuzzy System	Cleveland UCI Dataset,1990; 14 selected parameters	Highest level of accuracy expected based on method of feature selection	.Error rate greatly reduced due to genetic algorithm; increased accuracy measures

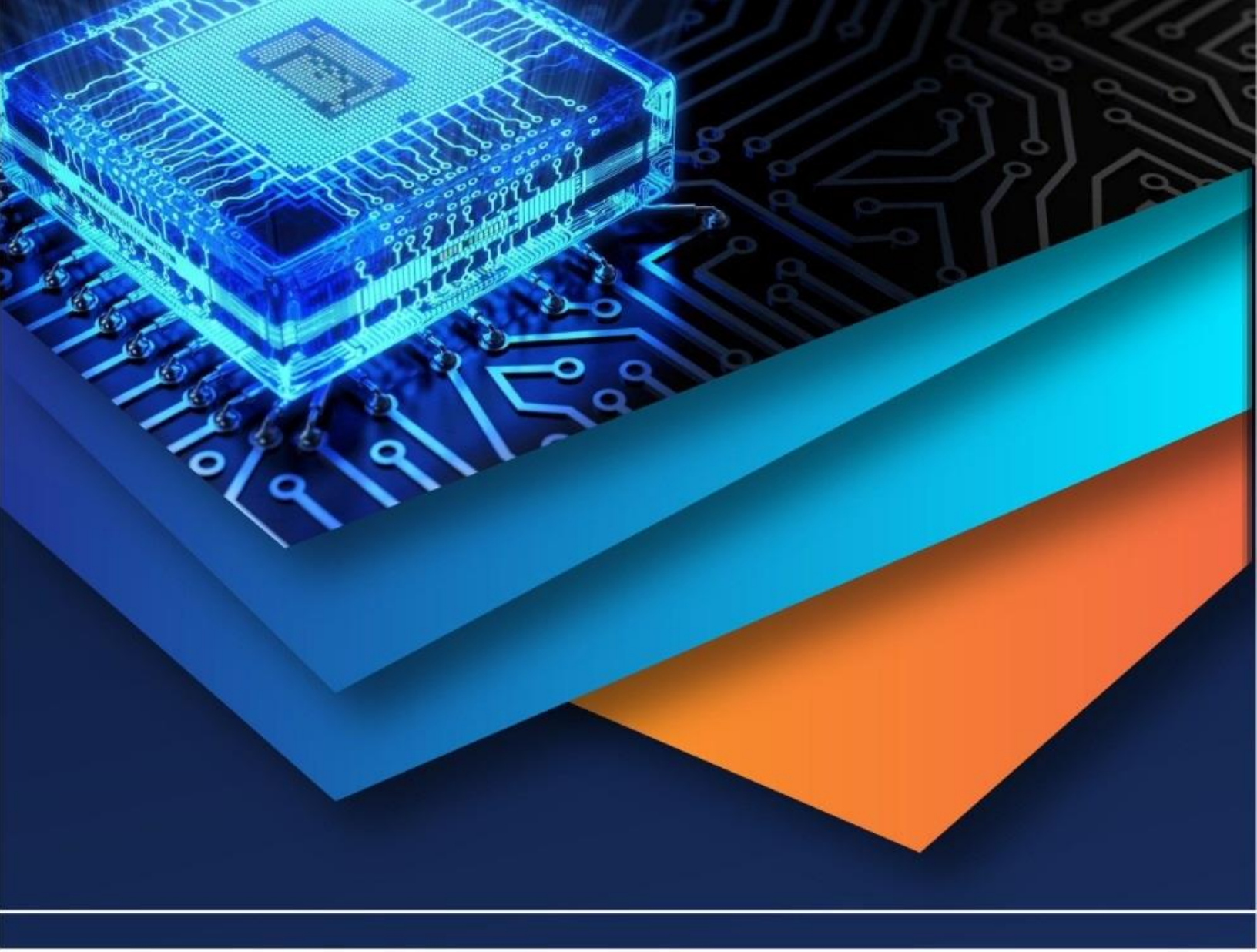
Table 1. Comparison of methodologies with stated parameters

X. CONCLUSION

Diagnostics in the medical field require a wholesome understanding of past cases, patient history and extensive medical literature. Often, a lot of information, in terms of patient history is made available to the doctor, but the relationship between them remain seemingly hidden. Therefore, in order to maximise the potential the information and information systems possess, comprehensive predictive analytics and mining systems need to be set in place. This paper attempts at weighing some of these approaches and providing a ready reckoner, of sorts to further researchers, in terms of referencing existing work. The power that data has, is put to best use to predict an ailment with far reaching consequences.

REFERENCES

[1] Feixiang Huan, Shengyong Wang, Chien-Chung Chan, Predicting Disease by using Data Mining Based on Healthcare Information Systems, 2012 IEEE Conference on Granular Computing
[2] Kaan Uyar, Ihmet Ilhan, Diagnosis of Heart Disease using Genetic Algorithm based trained recurrent fuzzy neural networks, Procedia Computer Science 120 (2017) 588-593
[3] T. Porter and B. Green, "Identifying Diabetic Patients: A Data Mining Approach," Americas Conference on Information Systems, 2009.
[4] U. Fayyad, et al., Knowledge Discovery and Data Mining: Towards a Unifying Framework, KDD-96, 1996
[5] P. Guidici, Applied Data Mining: Statistical Methods for Business and Industry
[6] R. Sathya, Annamma Abraham, Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification, International Journal of Advanced Research in Artificial Intelligence, Vol.2, No.2, 2013
[7] Kenji Suzuki, Artificial Neural Networks – Methodological Advances and Biomedical Applications
[8] Lingaraj, Haldurai. (2016). A Study on Genetic Algorithm and its Applications. International Journal of Computer Sciences and Engineering. 4. 139-143.
[9] PK Anooj, 2012, Clinical Decision Support System: Risk Level Prediction of Heart Disease using Weighted Fuzzy Rules, Journal of King Saud University, Vol. 24 Issue 1
[10] Sellappan Palaniappan, et al. 2008, Intelligent Heart Disease Prediction using Data Mining Techniques
[11] Sneha Nikam, et al. Cardiovascular Disease Prediction using Data Mining Techniques, IJLTET
[12] Salvatore R. Mangamo, "A Genetic Algorithm White Paper"
[13] PK Anooj, Clinical Decision Support System: Risk Level Prediction of Heart Disease Using Decision Tree Fuzzy Rules, Asian Transactions on Computers, Volume 02, Issue 04



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY