IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-7, July 2019 E-ISSN: 2347-2693

Enhanced Heart Disease Prediction Using HCR-PSO Based Data Analytical Model

Janani. S

Department of Computer Science, Rathnavel Subramaniam College of Arts and Science, Tamilnadu, India

*Corresponding Author: janusarguru@gmail.com, Tel.: +91-9789 95707.

DOI: https://doi.org/10.26438/ijcse/v7i7.280286 | Available online at: www.ijcseonline.org

Accepted: 10/Jul/2019, Published: 31/Jul/2019

***Abstract*—** Data Mining is an important aspect of diagnosing and predicting diseases in automatic manner. It involves developing appropriate techniques and algorithms to analyze data sets in medical field. At present, heart disease has excessively increased and heart diseases are becoming the most fatal diseases in several countries. In this paper, heart patient datasets are investigate for building classification models to predict the heart disease. This paper implements feature extraction technique construction and comparative study for improving the accuracy of predicting the heart disease. By the use of HCR-PSO (Highly Co-Related PSO) feature selection technique; a subset from whole normalized heart patient datasets is acquired which have only significant attributes. The study emphasized on finding the effective heart disease prediction construction by using various machine learning algorithms that are KNN(K-Nearest Neighbor), Random forest, SVM(Support Vector Machine), Bayesian network and MLP(Multilayer Perceptron). The research work central point is on finding the efficient classification algorithm for the prediction of heart disease in the early stage based on the accuracy using validation metrics that are Mean Absolute Error(MAE), Relative Squared Error(RSE) and Root Mean Square Error(RMSE).

***Keywords*—** HCR-PSO, Feature extraction, Bayesian and heart disease prediction model.

I. INTRODUCTION

Big Data referred as a huge quantity of data that is high in its speed, complexity and shifting that actually requires additional skills and technical aspects for the purpose to acquire storage, manage and analyze the information. Since Big Data holds important characteristics such as diversity and velocity. It greatly supports healthcare industry. Applying existing analytical technologies to the huge quantity of medical data leads to better understanding of outcome and later can be implemented at the point of care. By using this data, specialist knows his/her patient's status and it is further helpful in taking proper decisions for the treatment. In this paper, "HCR-PSO Based Data Analytical Model" is produced, which is an automated software model helps to determine heart related disease risks at early stage.

Section I explains the brief concept of data mining and importance of retrieving hidden information from the large dataset. In Section II, related works are discussed with special reference to develop disease prediction model. Section III contains detail about the methodology included for discovering heart disease prediction model. Section IV gives the detailed results after applying the suggested model. In Section V, conclusions were made after strongly carried out the research.

BIG DATA HEALTH CARE DATA

Big data in healthcare refers to electronic health data sets which is large and complex and difficult to manage with traditional or common data management methods and traditional software and/or hardware. Some health care data are characterized by a need for timeliness; for example, data generated by wearable or implantable biometric sensors; blood pressure, or heart rate is often required to be collected and analyzed in real-time. Data in healthcare can be categorized as follows

- Clinical Data and Clinical Notes: About 80% of this type data are unstructured documents, images and clinical or transcribed process.
- Structured data (e.g., laboratory data, structured EMR/HER)
- Unstructured data (e.g., post-op notes, diagnostic testing reports, patient discharge summaries, unstructured EMR/HER and medical images such as radiological images and X-ray images)
- Semi-structured data (e.g., copy-paste from other structure source)

DATA MINING

Data mining is the method of realizing models in giant fact sets involving tactics at the fork of machine learning and systems. It is an important method wherever intelligent strategies are applied to extract knowledge pattern. The

information mining is also accomplished exploitation classification, clustering, prediction, association and statistic analysis. Data processing is the study of huge datasets to haul out hidden and antecedently unknown models, relationships and information to facilitate or intricate to realize with ancient applied mathematics tactic. Therefore data processing refers to mining or extracting information from giant amount of facts.

TIME SERIES DATA

Statistic information is obtained at determined quantity from any system. Daily value modification of a market, current associated voltage information of an induction motor and therefore the population distribution consistent with year in a very state can moreover live thought of as a statistic. Such a statistic contains events of interest. One approach is to spot essential, time-ordered structures, known as temporal pattern, that are hidden in equivalent weight and are quality of fascinating events. Statistic analysis is an interest in predicting future values from diversion series information. A statistic example has been given in equation (1).

$$X= x_i, \text{----------} x_n \rightarrow 1$$

In such a statistic, t is time index and N is the total variety of observations. Necessary events area unit fashioned over the time.

II. RELATED WORKS

J.Vijayashree et al [1] imperative to predict the disease at a premature part. The computer aided systems facilitate the doctor as a tool for predicting and identifying cardiopathy. The target of this review is to widespread Heart connected upset and to temporarily use existing call support systems for the prediction and diagnose of cardiopathy supported by data processing and hybrid intelligent techniques .

Dilip Roy Chowdhury [2] represents the employment of artificial neural networks in predicting infant unhealthiness. The projected technique involves training a Multi Layer Perceptron with a BP learning rule to acknowledge a pattern for the designation and prediction of infant diseases. The rear propagation rule was accustomed train the ANN styles and conjointly a similar has been tested for the assorted categories of infant unhealthiness. Regarding this, ninety – four cases of assorted sign and symptoms, parameter is tested during this model. This study exhibits ANN primarily based prediction of infant unhealthiness and improves the identification accuracy of 75 with higher stability.

Niti Guru et al [3] planned a system that uses a neural network for prediction of heart upset, pressure level, and sugar. A set of cardinal records with 13 attributes used for coaching and testing. He urged supervised network for identification of upset and trained exploitation back propagation formula. On the concept of unknown data entered by a doctor the system can notice unknown data from coaching information and generate a listing of attainable unhealthiness from that patient can suffer.

Ms. Ishtake S.H et al [4] was enforced a model cardiovascular disease prediction system b three processing using three classification modeling techniques specifically, Decision. Trees, Naïve Bayes and Neural Network The system extract hidden data from historical cardiovascular disease data. DMX query language and functions are accustomed build and access the models. Five mining goals are outlined supported business intelligence and data exploration. The goals are evaluated against the trained models. All three models would possibly answer difficult queries, each with its own strength with relevance simple model interpretation, access to elaborate information and accuracy.

Mohammad Taha Khan et al [5] bestowed image model for the carcinoma additionally to predict upset using processing techniques. Two decision tree algorithms C4.5 and additionally the C5.0 is employed on these datasets for prediction and performance of every algorithmic rule are compared. Pruning algorithmic program is used to reduce a slip-up and avoids the over-fitting. Pruning a tree is the action to interchange a whole subtree by a leaf. The replacement takes place if the expected error rate within the subtree is bigger than the single leaf. Throughout this study, they started by generating the whole (generally over fitted) classification tree and alter it using pruning merely once.

Sang Hun Han et al [6] designed a framework to gather and store numerous domains of information on the causes of upset, and created giant information. A spread of open supply databases were integrated and migrated onto distributed storage devices. The integrated information was composed of clinical information on vessel diseases, national health and nutrition examination surveys, applied math geographic data, population and housing censuses, meteorologic administration information, and insurance Review and Assessment Service information. The framework was composed of information, speed, analysis, and repair layers, all hold on distributed storage devices. Finally, we tend to plan a framework for a upset prediction system supported lambda design to resolve the issues related to the period of time analyses of massive information. This method will facilitate, predict and diagnose diseases, like cardiovascular diseases.

Kiran et al [7] describe Lambda design and an enormous knowledge technique which will be accustomed support period analyses but, it's the limitation of not having the ability to investigate an outsized volume of knowledge in real time. To deal with this limitation, a way may be utilized that blends knowledge created before in an exceedingly batch layer with knowledge processed in real time. Then, the information may be generated and keep to attain this,

knowledge area unit shaped in batch read in an exceedingly cycle with a batch layer, and identical knowledge area unit shaped in period read via period processing. These two knowledge sets area unit then alloyed and analyzed, facultative the analysis of information reflects period data. Supporting this, Amazon net Services, process massive knowledge, wrote a white book on the mixing of execution and data processing into one network exploitation lambda design.

Cheryl Ann Alexander associate degreed Lidong Wang et al [8] describe an Acute infarct (heart attack) deadliest diseases patients face and therefore the key to disorder management is to gauge massive legion datasets, compare and mine for data which will be accustomed predict, prevent, manage and treat chronic diseases like heart attacks, massive knowledge analytics, acknowledged within the company world for its valuable use in dominant, contrastive and managing massive datasets may be applied with abundant success to the prediction, prevention, management and treatment of disorder, data processing, visual image and Hadoop area unit technologies or tools of massive knowledge in mining the voluminous datasets for data. The aim of this literature review was to spot usage of massive knowledge analytics in coronary failure prediction and bar, the employment of technologies applicable to massive knowledge, privacy issues for the patient, challenges and future trends similarly as suggestions for additional use of those technologies. The results can guide suppliers, tending organizations, nurses, and different treatment suppliers in exploitation massive knowledge technologies to predict and manage coronary failure and tailored medical treatment may be developed exploitation these technologies

Vinitha S [11] describes huge knowledge progress in medical specialty and health care communities, correct study of medical knowledge advantages, early sickness recognition, patient care and community services. If the standard of medical knowledge is incomplete then the exactitude of study is reduced. Moreover, totally different regions exhibit distinctive appearances of sure regional diseases, which leads to weakening the prediction of sickness outbreaks. Within the projected system, it provides machine learning algorithms for effective prediction of varied disease occurrences in disease-frequent societies. It experiment the altered estimate models over real-life hospital knowledge collected to beat the problem of incomplete knowledge. It uses a latent issue model to build the missing knowledge. It experiment on a regional chronic eudemonia of cerebral infarct. Mistreatment structured and unstructured knowledge from hospital Machine Learning call Tree formula and Map cut back formula is used.

Kelvin KF Tsoi1 et al [12] describes the study on call trees and their behavior in inward to the conclusion. Tree node cacophonic supported relevant feature choice may be a key step to call tree learning. At constant, nowadays their major shortcoming: the algorithmic nodes partitioning ends up in geometric reduction of information amount within the leaf nodes that causes associate degree excessive model complexness and knowledge over fitting. During this paper, the author bestowed a unique design referred to as a choice Stream.

III. METHODOLOGY

A. BIG DATA ANALYTICAL MODEL

Big data analytics support the concept of artificial intelligence and lie at the heart of many new digital health platforms and precision health tools. Ideally, utilization of big data analytic tools in cardiovascular care will translate this into better care and outcomes at a lower cost.

The potential for more powerful predictive models is an appealing application of big data analytics. Historically, prediction models have relied on a limited number of specified variables manually entered to estimate a 'risk score'. Such models generally lack precision: they perform 'reasonably well' at the population level, but not at the individual patient level. And despite the existence of dozens of risk models related to cardiovascular conditions, few are utilized to make therapeutic decisions.

Big data analytics may yield more powerful prediction of outcomes ranging from mortality to patient-reported outcomes to resource utilization, and thus it could be more clinically actionable. Machine learning, for example, evaluates patterns associated with an outcome directly from the data, rather than from a pre-specified set of variables. A full range of associations and interactions among the data are assessed. Whereas traditional statistical models are 'one and done', machine learning uses a training process whereby the model is iteratively given varied data sets to explore many combinations of predictive features to optimize prediction.

Phenol-mapping, or deep phenol-typing, is another promising big data application. Current disease classifications, or phenotypes, are imprecise and heterogeneous. Big data analytics can identify similar patient clusters, creating multiple phenotypes within each disease entity. In theory, more refined phenol-mapping of disease states and trajectories should help inform more tailored-health decisions

Big data methods can support the combination of multiple data sources from large patient populations to better estimate the potential benefits of therapies such as ICD's for individual patients. Indeed, big data analytic methods are central to the success of precision health, given the growing interest in incorporating '-omic' data, which vastly increases the size and complexity of datasets. Such datasets require

advanced analytic platforms and methods that are the hallmarks of big data analytics.

Big data analysis can guide policies to address a certain patient segment by specific interventions. The success of the policy is critically dependent on the quality of the underlying research and the quality (effectiveness) of the interventions. For many interventions (for instance in the social/mental health domain) universally accepted methods for validating success are still lacking. There is several challenges solution regarding Big Data and population heart disease such as:

- Data protection regulation makes it difficult to analyze data from different heart disease providers and services in combination
- A significant part of the population health records is unstructured heart disease text
- There are interoperability, data quality and data integration limitations

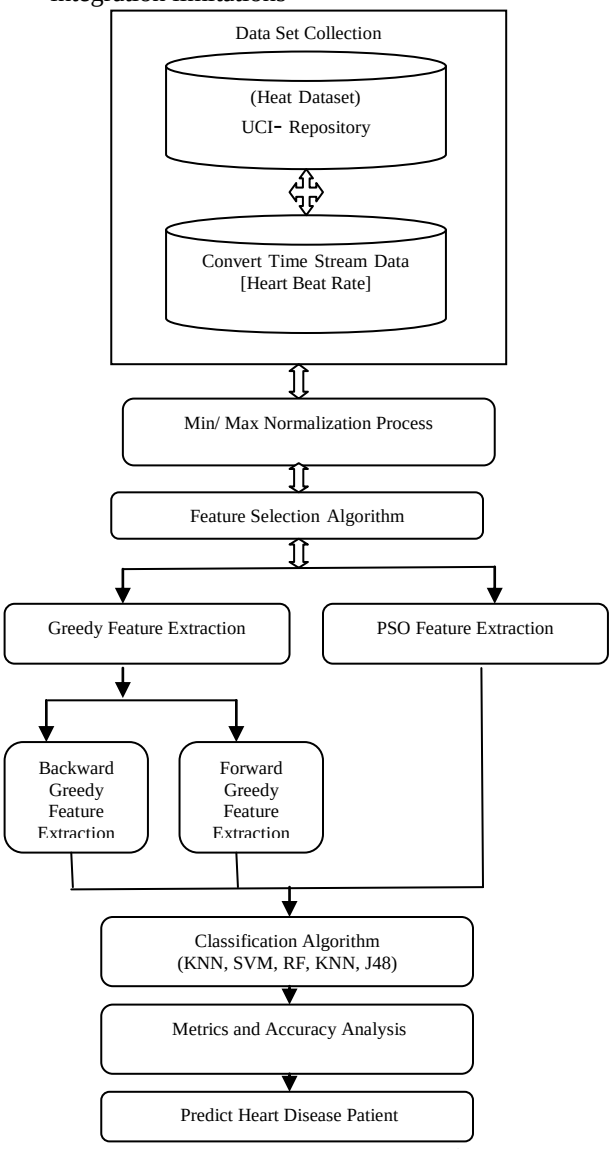

Figure 4.1. Architecture Diagram Of Proposed System

B. NORMALIZATION MODEL

Data transformation such as Normalization is a data preprocessing tool used in data mining system. An attribute of a dataset is normalized by scaling its values so that they fall within a small-specified range, such as 0.0 to 1.0. Normalization is particularly useful for classification algorithms involving neural networks, or distance measurements such as nearest neighbor classification and clustering. There are many methods for data normalization includes min-max normalization, z-score normalization and normalization by decimal scaling.

Min-max normalization performs a linear transformation on the original heart dataset. Min-max normalization maps a value d of P to d′ in the range [new_min(p), new_max(p)].

Min max normalization preserves the relationship among the original heart dataset values**.** The table 4.1 describes a sample normalized heat disease dataset model details shown as follows,

Table 4.1 Normalized Heart Dataset

Attribute	Original Values	Normalized dataset
Age	70.0	50.0
chest pain type	1.0	0.0
resting blood pressure	130.0	150.0
maximum heart rate achieved	109.0	78.0
exercise induced angina	0.0	1.0

C. GREEDY FEATURE EXTRACTION MODEL

Feature selection is one of the dimension reduction techniques which have been used to allow a better understanding of data and improve the performance of other learning tasks. Although the selection of relevant features has been extensively studied in supervised learning, feature selection with the absence of class labels is still a challenging task.

This paper develops a novel method for unsupervised feature selection, which efficiently selects features in a greedy manner. The paper first defines an effective criterion for unsupervised feature selection which measures the reconstruction error of the data matrix based on the selected subset of features. The paper then presents a novel algorithm for greedily minimizing the reconstruction error based on the features selected so far. The greedy algorithm is based on an efficient recursive formula for calculating the reconstruction error.

The greedy algorithm selects iteration the most representative feature among the remaining features, and

then eliminates the effect of the selected features from the data matrix. This step makes it less likely for the algorithm to select features that are similar to previously selected features, which accordingly reduces the redundancy between the selected features. In addition, the use of the recursive criterion makes the algorithm computationally feasible and memory efficient compared to the state of the art methods for unsupervised (forward and backward) feature selection.

D. PSO FEATURE EXTRACTION MODEL

Feature selection is the process of identifying statistically most relevant features to improve the predictive capabilities of the classifiers. To find the best feature subsets, the population based approaches like Particle Swarm Optimization (PSO) and genetic algorithms are being widely employed. However, it is a general observation that not having right set of particles in the swarm may result in sub-optimal solutions, affecting the accuracies of classifiers. To address this issue, propose a novel tunable swarm size approach to reconfigure the particles in a standard PSO, based on the heart data sets, in real time.

The feature selection algorithms have been widely used in many application areas such as genomic analysis, text classification], information retrieval, and bio informatics. Feature selection is an optimization problem which aims to determine an optimal subset of d features out of n features in the input data (d << n). It maximizes the classification or prediction accuracy. Performing an exhaustive search to find an optimal subset of d features out of all possible 2n candidate feature subsets, based on some evaluation criterion, is computationally infeasible, and it becomes an NP-hard problem with the increasing n value.

E. CLASSIFICATION ANALYTICAL MODEL

Machine learning indicates how computers can learn or improve their performance using data. Computer programs do automatically learn to identify patterns and make intelligent based decisions on data. Machine learning is a fast growing discipline. Here, using classic problems in machine learning that are highly related to data mining.

- Supervised classification learning:
  - ✓ Supervised classification learning model consist of all data is labeled and algorithm learn to predict the output from training dataset.
  - ✓ E.g.: Support Vector Machine (SVM), Random Forest, Naive Bayes.
- Unsupervised classification learning
  - ✓ Unsupervised classification learning is used for clustering based algorithm. In this learning all the data is unlabelled and algorithm finds the essential structure from the input dataset. We can use clustering to discover classes within the dataset.
  - ✓ E.g. K-means, KNN. Neural Networks
- Semi-supervised classification Learning
  - ✓ Semi-supervised learning is a combination of supervised learning and unsupervised learning. In Semi-supervised learning some data is labeled and some data is not labeled. In this approach, labeled training dataset are used to learn class models and unlabelled training dataset are used to define the boundaries between classes.

IV. RESULTS AND DISCUSSIONS

A. DATASET DESCRIPTION

Preparing the database - for obtaining the result, this paper uses Heart patient data sets from ILPD (Indian Heart Patient) Data Set (table 5.1). Totally, heart dataset has 583 samples which holds 10 independent variables and one dependent variable. Independent Variables are: Age, Gender, Total Bilirubin, Direct Bilirubin, Total Proteins, Albumin, SGPT (serum glutamic-pyruvic transaminase), SGOT (serum glutamic oxaloacetic transaminase), Alkaline Phosphatase and one dependent variable is Classing (class) [36].

Table 5.1 Dataset Attribute

Attributes Type	Description	Gender Categorical
Age	age given in years	Real number
Sex	sex (Value 1 : male; Value 0 : female)	String
Cp	chest pain type(1: typical angina ; 2: atypical angina 3: non-anginal pain ;4: asymptomatic)	Real number
Trestbps	resting blood pressure (in mm Hg on admission to the hospital)	Real number
Chol	Cholestoral(Serum cholestoral) in mg/dl	Real number
Fbs	Fasting blood sugar in mg/dl (>120) Value 1 = true; Value 0 = false)	Real number
Restecg	Resting electrocardiographic results	Real number
Thalach	Heart rate achieved at maximum	Integer
Exang	Exercise induced angina (Value 1 : yes; Value 0 : no)	Integer
Oldpeak	ST depression originated by	Integer

	exercise relative to rest	
Slope	Slope of the peak exercise ST segment (Value 1: upsloping ; Value 2: flat ; Value 3: downsloping)	Integer
Ca	Major vessels (0-3) colored by flouroscopy	Integer
Thal	Result of thalium stress test (Value 3 = normal; Value 6 = fixed defect; Value 7 = reversable defect)	Integer
Num	status of heart disease (angiographic status) Value 0: < 50% diameter narrowing Value 1: > 50% diameter narrowing	Binary

The following table 5.2 describes the feature extraction attributes

Table 5.2 Feature Selection Dataset

Attributes Type	Description	Gender Categorical
Trestbps	resting blood pressure (in mm Hg on admission to the hospital)	Real number
Chol	Cholestoral(Serum cholestoral) in mg/dl	Real number
Fbs	Fasting blood sugar in mg/dl (>120) Value 1 = true; Value 0 = false)	Real number
Restecg	Resting electrocardiographic results	Real number
Thalach	Heart rate achieved at maximum	Integer
Exang	Exercise induced angina (Value 1 : yes; Value 0 : no)	Integer
Oldpeak	ST depression originated by exercise relative to rest	Integer
Slope	Slope of the peak exercise ST segment (Value 1: upsloping ; Value 2: flat ; Value 3: downsloping)	Integer

B. PERFORMANCES METRICS ANALYSIS

Table 5.2 describes an evaluation metrics for Heart disease prediction model. The table contains Mean Absolute error, Root Relative square Error, Root Relative Square Error and Accuracy values of SVM, KNN, RF, J.48 and MLP classification algorithm details are shown

Table 2. Performances Analysis

Evaluation Criteria	SVM	KNN	RF	J.48	MLP
Mean Absolute Error (MAE)	0.345	0.323	0.337	0.317	0.322
Root Mean Square Error (RMSE)	0.452	0.386	0.432	0.382	0.393
Relative Absolute Error (RAE)	0.523	0.482	0.502	0.485	0.492
Root Relative Square Error (RRSE)	0.478	0.428	0.438	0.402	0.418
Accuracy	0.882	0.892	0.874	0.934	0.926

Figure 2 describes an evaluation metrics for Heart disease prediction model. The table contains Mean Absolute error and Root Relative square Error of SVM, KNN, RF, J.48 and MLP classification algorithm details are shown

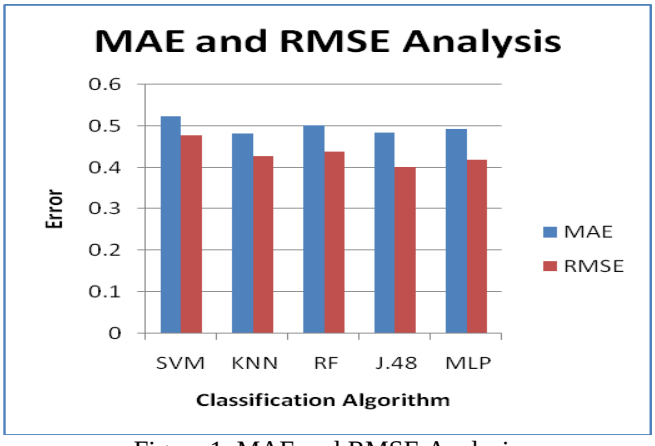

Figure 1. MAE and RMSE Analysis

Fig 5.2 describes an evaluation metrics for Heart disease prediction model. The table contains Root Relative square Error and Root Relative Square Error of SVM, KNN, RF, J.48 and MLP classification algorithm details are shown

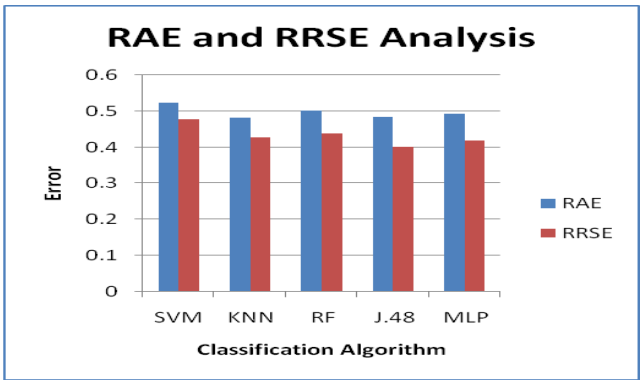

Figure 2. RAE and RRSE Analysis

Figure 3 describes an evaluation metrics for Heart disease prediction model. The table contains accuracy values of SVM, KNN, RF, J.48 and MLP classification algorithm details are shown

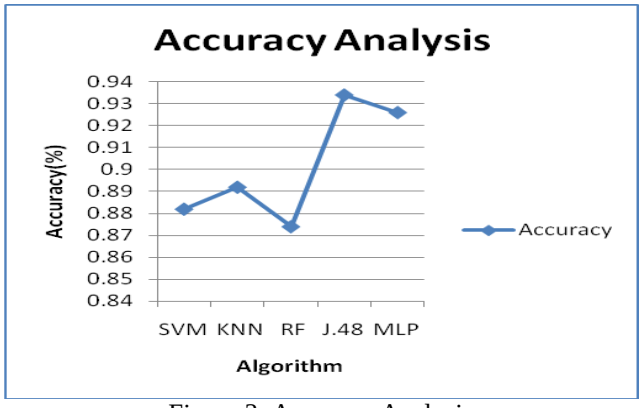

Figure 3. Accuracy Analysis

V. CONCLUSION

The proposed technique produces an enhanced concept over the heart disease prediction within novel data mining techniques; SVM, RF, KNN, MLP and j48 the weighted association classifier. The PSO is an enhance approach for classification with weighted association rule with Weighted Support and Confidence Framework to extract Association rule from big data warehouse. J48 classification is the technique to cluster the attributes from the patient record. The SVM clustering and J48 with weighted association classifier can enhance the classification performance and accuracy of the heart disease diagnosis.

These all metrics is being verified by the experts and professional doctors of heart specialist. The proposed technique PSO is producing 93.46% accuracy for centroid selection and classification, PSO is producing 93.40% accuracy for classification with weighted support and confidence. The overall accuracy of the system is 94.40%

In future work, we will look for the more enhancements to get better result over the heart disease prediction by increasing the metrics and doctors suggestion within different types of medical term and also provide the first-aid suggestion in unavailability of heart specialist or experts.

REFERENCES

[1]. J.Vijayashree and N.Ch.Sriman Narayana Iyengar, "Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques: A Review ", International Journal of BioScience and Bio Technology, Vol.8, No.4 (2016), pp. 139-148.

[2]. Dilip Roy Chowdhury, Mridula Chatterj ee & R. K. Samanta, An Artificial Neural Network Model for Neonatal Disease Diagnosis, International Journal of Artificial Intelligence and Expert Systems (IJAE), Volume (2): Issue (3), 2011.

[3]. Milan Kumari, Sunila Godara, Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction, IJCST Vol. 2, Iss ue2, June 2011.

[4]. Ishtake S.H ,Prof. Sanap S.A., "Intelligent Heart Disease Prediction System Using ata Mining Techniques", International J. of Healthcare & Biomedical Research,2013.

[5]. Mohammad Taha Khan, Dr. Shamimul Qamar and Laurent F. Massin, A Prototype of Cancer/Heart Disease Prediction Model Using Data Mining, International Journal of Applied Engineering Research, 2012.

[6]. Sang Hun Han,ID, Kyoung Ok Kim, Eun Jong Cha, Kyung Ah Kim and Ho Sun Shon, System Framework for Cardiovascular Disease Prediction Based on Big Data Technology, Symmetry 2017, 9, 293.

[7]. Kiran, M.; Murphy, P.; Monga, I.; Dugan, J.; Baveja, S.S. Lambda architecture for cost-effective batch and speed big data processing. In Proceedings of the 2015 IEEE International Conference on Big Data,Santa Clara, CA, USA, 29 October–1 November 2015; pp. 2785–2792.

[8]. Cheryl Ann Alexander and Lidong Wang," Big Data Analytics in Heart Attack Prediction", JNurs Care, an open access journal, Volume 6 , Issue 2, ISSN:2167-1168,2017.

[9]. Ms. S.Suguna, Sakthi Sakunthala,N, S.S anjana,S.S.Sanjhan, "A Survey On Prediction of Heart Diseases Using Big Data Algorithms", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 6, Issue 3, March 2017,ISSN:2278–1323.

[10]. Saranya P and Satheeskumar B "A Survey on Feature Selection of Heart Disease Using Data Mining Techniques", International Journal of Computer Science and Mobile Computing, Vol.5 Issue.5, May- 2016, pg. 713-719.

[11]. Vinitha S, Sweetlin S, Vinusha H and Sajini S. "Disease Prediction Using Machine Learning Over Big Data", Computer Science & Engineering: An International Journal (CSEIJ), Vol.8, No.1, February 2018.

[12]. Kelvin KF Tsoi1, Yong-Hong Kuo and Helen M. Men, "Dmitry Ignatov and Andrey Ignatov. Decision Stream: Cultivating Deep Decision Trees", 3 Sep 2017 IEEE".

