International Journal of Engineering & Technology, 7 (3.20) (2018) 738-742

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Novel Approach to Predict Disease and Avoid Congestion in Data Mining Using Genetic Algorithm

S. Ramasamy[1]*, Dr. K. Nirmala[2]

[1] *Research Scholar*
Quaid-E-Millath Govt college for WomenChennai, India
[2] *Research Supervisor*
Department of Computer ScienceQuaid-E-Millath Govt college for WomenChennai, India
**Corresponding author Email: s_ramasamy@hotmail.com*

Abstract

Presently days the health segment contains concealed information that can be critical in deciding. It is troublesome for medical professionals to anticipate the disease as it is really an intricate errand that requires experience and information. The objective of the examination is to anticipate conceivable disease from the patient dataset utilizing data mining systems and to organize the patients based on their genuine conditions to lessen the blockage in the system. In this paper we propose proficient acquainted order algorithm utilizing genetic methodology for disease expectation. The principle inspiration for utilizing genetic algorithm as a part of the disclosure of abnormal state forecast rules is that the found rules are exceptionally conceivable, having high prescient precision and of high interestingness values.

Keywords: *Data Mining, Association Rule, Keyword Based Clustering, Genetic algorithm, Classification.*

1. Introduction

In writing, numerous exploration papers are accessible which mostly centered around anticipating diseases from health data sets utilizing data mining systems. The real reason that the data mining has pulled in extraordinary arrangement of consideration in the information business in the late years is because of the wide accessibility of tremendous measures of data and the requirement for transforming such data into helpful information and learning. [1].

A noteworthy test confronting healthcare associations like hospitals, medical focus and so on is the procurement of value administrations at low expenses. Quality administration infers diagnosing patients flawlessly and overseeing medicines that are useful. Poor clinical choices can prompt lamentable results which are in this way unacceptable.[2] Hospitals should likewise minimize the expense of clinical tests. Health care associations must have capacity to investigate data. Treatment records of a large number of patients can be put away and electronic and data mining procedures may help in noting a few essential and basic inquiries identified with health care. Expectation includes utilizing a few variables or fields as a part of the data set to anticipate obscure or future estimations of different variables of interest.[2]. In biomedical field data mining and its methods assumes a vital part for forecast of different diseases. The doctors may not ready to analyze it accurately when the patients experience the ill effects of more than one kind of disease of the same classification. Due to missing fixation or unhealthy practices when expectation of disease classification. The healthcare business gives immense measures of healthcare data and that should be mined to discover shrouded information for profitable basic leadership. Find of shrouded examples and connections frequently go unused . The patient's record is arranged and anticipated on the off chance that they have the side effects of coronary illness and utilizing hazard elements of disease. It is essential to locate the best fit algorithm that has more prominent exactness, less cost, fast and memory usage on grouping on account of coronary illness expectation category[3].

Association rules are mined on a medical data set to enhance coronary illness diagnosis. Every rule speaks to a basic prescient example that portrays a subset of the data set anticipated on a subset of qualities. From a medical viewpoint, association rules relate mixes of twofold target properties (non-appearance/presence of course disease) and subsets of free traits (hazard components and heart muscle health estimations). Association rules have vital points of interest over customary algorithms [4].

Medical diagnosis is viewed as a critical yet confused errand that should be executed precisely and productively. The computerization of this framework would be to a great degree profitable. Unfortunately all doctors don't have skill in each sub claim to fame and in addition there is a deficiency of asset persons at specific spots. Along these lines, a programmed medical diagnosis framework would most likely be exceedingly gainful by uniting every one of them. Fitting PC based information and/or choice emotionally supportive networks can help in accomplishing clinical tests at a decreased expense. Proficient and exact execution of computerized framework needs a near investigation of different systems accessible. This paper plans to dissect the diverse prescient clear data mining methods proposed as of late for the diagnosis of disease.

Hereditary calculation has been used as a piece of [4], to dewrinkle the veritable information size to get the perfect subset of credited satisfactory for coronary sickness gauge. Gathering is a managed learning technique to focus models portraying basic information classes or to foresee future examples. Three classifiers for example Decision Tree, Credulous Bayes and Arrangement by methods for bunching have been used to break down the closeness of coronary ailment in patients.Arrangement by means of clustering: Clustering is the way toward gathering comparable components. This method might be utilized as a pre handling venture before sustaining the data to the characterizing model [5].

The credit values should be standardized before clustering to maintain a strategic distance from high esteem characteristics overwhelming the low esteem traits. Further, characterization is performed based on clustering.
Data mining is a urgent stride in disclosure of information from substantial data sets. Lately, Data mining has discovered its critical hold in each field including health care. Mining procedure is more than the data investigation which incorporates arrangement, clustering, association rule mining and expectation. It likewise traverses different orders like Data Warehousing, Statistics, Machine learning and Artificial Intelligence.

2. Literature Review

Medical diagnosis is known not subjective and depends on the accessible data as well as on the experience of the doctor and even on the psycho-physiological state of the doctor. Various studies have shown that the diagnosis of one patient can vary altogether if the patient is tried by various doctors or even by the same doctor at different times.
Himigiri. Danapana in at al[6] This investigation hopes to give an audit of stream frameworks of data divulgence in da-tabases using information mining methods that are being utilized in the present therapeutic research particularly in

Coronary disease Forecast. Number of investigation has been directed to consider the execution of perceptive information mining technique on the equivalent dataset and the outcome reveals that Choice Tree defeats and some time Bayesian portrayal is having relative precision as of decision tree.
Fariba Shadabi in at al[7] Artificially Astute (AI) controlled devices can manage unverifiable and incomplete data sets. Neural network classifiers have been effectively utilized for forecast purposes as a part of numerous mind boggling circumstances.
Research shows that AI-based data mining apparatuses have been likewise effectively utilized as a part of numerous medical situations. This examination propels the comprehension of the utilization of Artificial Intelligence and Data Mining apparatuses to clinical data by exhibiting the capability of these methods in complex clinical circumstances.
Two sorts of information mining calculations named transformative named GA-KM and MPSO-KM group the coronary illness informational index and predict demonstrate precision [8]. This is a cross breed system that joins compel sort particle swarm improvement (MPSO) and K-infers method. The relationship was made in the examination coordinated using C5, Innocent Bayes, K-suggests, Ga-KM and MPSO-KM for surveying the precision of the frameworks. The expe-rimental results showed that precision upgraded while using GA-KM and MPSO-KM [8]. The researchers made class affiliation rules using feature subset decision to anticipate a model for coronary ailment. Affiliation rule chooses relations among quali-ties and portrayal predicts the class in the patient dataset [9]. Feature assurance measures, for instance, hereditary chase de-cides characteristics which contribute towards the gauge of heart dis-facilitates. The researchers [10] executed a half and half structure that utili-zations overall improvement preferred standpoint of hereditary calculation for instatement of neural system loads. The forecast of the coronary illness is based on danger variables, for example, age, family history, diabetes, hypertension, elevated cholesterol, smoking, liquor admission and corpulence [10].
Diagram based methodology for coronary illness forecast was proposed by [11].Their technique is based on most extreme inner circle and weighted association rule mining. Affiliated arrangement for coronary illness forecast was proposed by [12].They utilized Gini list based grouping to anticipate the coronary illness. The analysts [13] utilized the data mining algorithms choice trees, guileless bayes, neural networks, association order and genetic algorithm for foreseeing and dissecting coronary illness from the dataset.
An analysis performed by [14] the analysts on a dataset created a model utilizing neural networks and hybrid keen algorithm, and the outcomes demonstrate that the hybrid insightful method enhanced accuracy of the expectation.
The data mining strategies like artificial neural network procedure is utilized as a part of compelling heart attack expectation framework. In any case the dataset used for estimate of heart sicknesses was pre-arranged and grouped by technique for K-suggests bunching calculation [15]. By then neural system is set up with the picked important precedents. Multi-layer Perceptron Neural Network with Back spread is used for getting ready. The results show that the calculation used is prepared for anticipating the heart maladies even more capably. The desire for heart sicknesses out and out uses 15 attributes, with essential information mining technique like ANN, Clustering and Association Rules, sensitive preparing approaches, etc. The outcome shows that Choice Tree execution is dynamically and multiple times Bayesian game plan is having similar exactness as of decision tree yet other perceptive systems like K-Closest Neighbor, Neural Networks, Characterization dependent on grouping won't perform well [16].
By utilizing the Weighted Cooperative Classifier (WAC), a slight change has been made, rather than considering 5 class marks, just 2 class names are utilized. One for "Coronary illness" and another for "No Coronary illness". The most extreme accuracy (81.51%) has been accomplished. At the point when genetic algorithm is connected, the accuracy of the Choice Tree and Bayesian Order is enhanced by diminishing the real data size. The dataset of 909 patient records with heart diseases has been gathered and 13 characteristics has been utilized for consistency [17]. The patient records have been splitted similarly as 455 records for preparing dataset and 454 records for testing dataset. In the wake of applying genetic algorithm the ascribes has been diminished to 6 and choice tree performs all the more productively with 99.2% accuracy when contrasted and different algorithms [18].

3. Research Methodology

In this paper proposes to identify the precise disease based on the user manifestations from the hospital information [19] database by utilizing three algorithms are:

1.Association rule mining Algorithm which is utilized to separate the data from the hospital information database.

2.Keyword based clustering algorithm is utilized to locate the exact disease which is influencing the patient.

3.Genetic algorithm is utilized to organize the patient keeping in mind the end goal to maintain a strategic distance [19] from the clog.

3.1. Association Rule Algorithm

Association rule mining is to discover the association rules met the user-determined least backing and least certainty prerequisite from the exchange database D [20]. The whole mining procedure can be decayed into the accompanying two stages: to start with, locate all successive thing sets, that is, discover all the thing sets had support more noteworthy than the given bolster edge; second, based on the got visit thing sets, produce a comparing solid association rule [21], that is, create the association rules had backing and certainty individually more noteworthy than or equivalent to the given bolster edge and certainty edge [22].

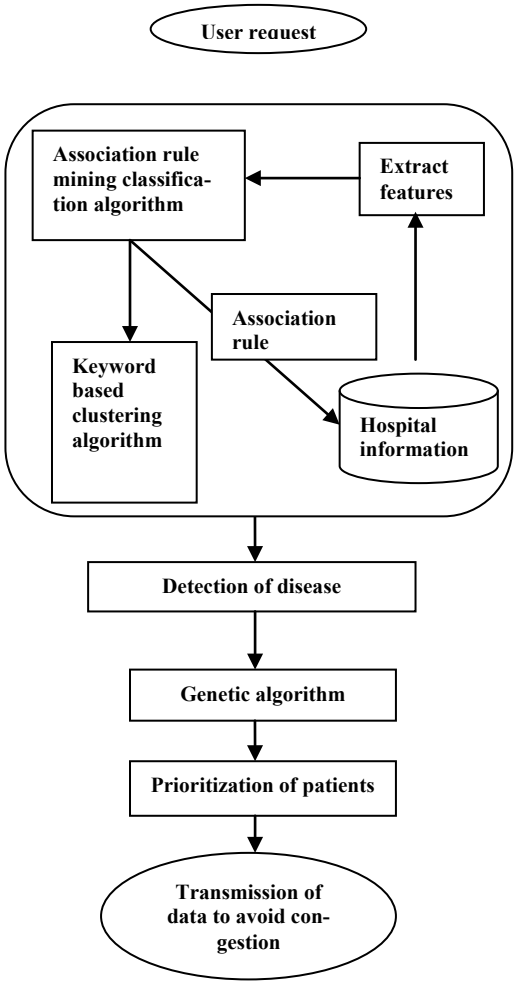

Fig.1. System Architecture

Let I = (i1,i2....im) be a lot of literals, called things. Give D a chance to be a database of exchange, where every exchange T is a lot of things to such an extent that T #I. For a given thing set X #I and a given exchange T, we state that T contains X if and just if X #I.

3.2. Apriori Algorithm

The Apriori calculation works iteratively. It first finds the mastermind ment of tremendous 1-thing sets, and after that course of action of 2-itemsets, and so forth. The amount of breadth over the trade database is a similar number of as the length of the maximal thing set. Apriori depends on the going with conviction: The fundamental anyway serious observation prompts the period of a more diminutive contender set using the game plan of significant thing sets found in the past accentuation

The Apriori algorithm presented is given as follows:

```
Apriori()
L1 ={large 1-itemsets}
k = 2
while Lk_1–/ do
begin
Ck ¼ apriori genðLk_1Þ
for all transactions t in D do
begin
Ct ¼ subsetðCk; tÞ
for all candiate c 2 Ct do
c.count = c.count+1
end
Lk ¼ fc 2 Ckjc:count P minsupg
k = k + 1
end
```

Apriori first sweeps the exchange databases D with a specific end goal to check the backing of everything i in I, and decides the arrangement of large 1-itemsets. At that point, iteration is performed for each of the calculation of the arrangement of 2-itemsets, 3-itemsets, et cetera. The kth iteration comprises of two stages:

- Generate the candidate set Ck from the set of large (k-1)-itemsets,
- Scan the database in order to compute the support of each candidate itemset in Ck.

3.2.1. Keyword Based Clustering Algorithm

Watchword based document bunching makes a group by the catchphrases of each report. Accept that C is a course of action of bunches that is finally made by the grouping calculation. If n is the amount of groups in C, at that point C is a plan of bunches C1,C2,C3...Cn.

C = { C1,C2,C3....Cn}

Each group is initialised by record d that isn't doled out to the present bunches, and d is a seed file of . Exactly when another bunch is made, expansion and decline steps are reiterated until the point that it accomplishes an enduring state from the start state [23].

3.2.2 Cluster Initialization

The underlying advance of the grouping calculation is a creation and initia-lisation of another bunch. A chronicle D is picked that does not have a place with some other group, and it is alloted to another bunch Ci that is a fundamental state of bunch [24].

Ci= {0}

As of now, a record D that is the main archive in the new group is known as a seed report [25]w.

3.2.3 Expansion of Cluster

In the initialisation adventure of the bunch, another group Ci , a hidden condition of group C, is developed as the seed file, and the watchword set I is initialised by the catchphrase Kc set of the seed record. In the developing step of the bunch, the group is reached out by adding progressively related chronicles to the group, that fuse the catchphrases of the seed report as the related records of the seed record. The bunch improvement is performed by the emphasis of catchphrase expansion and group advancement. More reports are added to a bunch by the similarity appraisal between the watchword set and the record. If another report is added to a group, at that point the catchphrases in the extra record are similarly added to the watchword set of the bunch.

3.2.4 Cluster Reduction

This movement is to make a total bunch by clearing the records that are not related to the group. For the group Ci reports of a low likeness to the bunch are cleared, that are not related to a group Ci through the similarity figuring with the bunch. In the end, the bunch C is finished that includes the related records in the wake of isolating the non-related chronicles. If a group C is finished the accompanying bunch is made through a similar method. Grouping is finished if all of the reports are bunched or no more bunches are made.

3.2.5 Entropy Based Genetic Algorithm

Hereditary calculations are enlisting strategies worked in comparability with the method of headway. It almost takes after the ordi-

nary strategy of recuperation, age, inheritance improvement. Hereditary calculations are routinely used for issues that can't be enlightened capably with standard systems. Hereditary calculations are useful for looking particularly wide spaces and streamlining issues. Each plan made in Genetic calculations is known as a chromosome (individual).Each chromosome is included characteristics, which are the individual parts (alleles) that addresses the issue. The social event of chromosomes is known as a people. The inside portrayal of the chromosomes is known as its genotype.

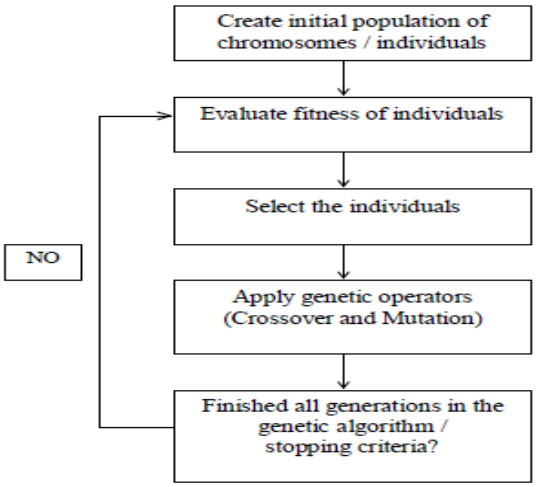

Fig. 2: Flow chart of genetic algorithm

The functions of genetic operators are as follows:
1) Selection: selection manages the probabilistic survival of the fittest in that, more fit chromosomes are survived.
2) Crossover: This operation is performed by selecting an irregular quality along the length of the chromosomes and swapping every one of the qualities after that point. Different sorts of crossover administrators are a) single point b) two point c) uniform d) half uniform e) lessened surrogate crossover f) mix crossover g) divided crossover
3) Mutation: mutation adjusts the new arrangements in order to include stochasticity in the quest for better arrangement. The most well-known strategy method for actualizing mutations is to choose a bit aimlessly and flip (change) its quality.
There are 2 sorts of mutations use in genetic network programming 1) transforming the judgment hub 2) changing the estimation of the judgment hub. In acquainted grouping properties and their qualities are taken as judgment hubs and class values as handling hubs.

3.2.6 Entrophy Measures

Entropy is a usually utilized measure as a part of information hypothesis. Initially it is utilized to portray the impurity of a self-assertive gathering of illustrations. In our usage entropy is utilized to gauge the homogeneity of the illustrations that a rule matches. Given an accumulation S, containing the cases that a specific rule R matches, let Pi be the extent of case in S having a place with class i, then the entropy Entropy(R) identified with this rule is characterized as:

$$Entropy(R) = -\sum_{i=1}^{n} (p_i \log_2 (p_i))$$

where n is the number of target classifications. While an individual consists of a number of rules, the entropy measure of an individual is calculated by averaging the entropy of each rule:

$$Entropy(individual) = \frac{\sum_{i=1}^{N_R} Entropy(R_i)}{N_R}$$

where N_R is number of rules in the individual

4. Performance Analysis

The performance of the algorithm is evaluated using the measures like accuracy, Time computation, efficiency defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

Table 1: The comparison results on the prediction accuracy and standard deviation

	Our GA approach	Decision trees	Decision trees with boosting	Neural networks	Naive Bayes
Run 1	90.77	87.69	89.23	89.23	66.15
Run 2	89.23	84.62	86.15	86.15	78.46
Run 3	89.23	89.23	90.77	90.77	84.62
Run 4	92.31	90.77	90.77	89.23	81.54
Run 5	86.15	81.54	81.54	84.62	75.38
Run 6	89.23	87.69	87.69	87.69	80.00
Run 7	84.62	81.54	84.62	84.62	73.85
Run 8	87.69	86.15	87.69	86.15	83.08
Run 9	90.77	86.15	89.23	87.69	76.92
Run 10	86.76	88.24	91.18	86.76	75.00
Average	88.68	86.36	87.89	87.29	77.50
Standard deviation	2.37	3.06	3.08	2.03	5.36

Table 3: Comparison of parameters between Non GA and GA

Parameters	Decision trees	Decision tree With boosting	Neural networks	Navie bayes	Entrophy based GA
Accuracy	67.56	60.98	56.09	53.66	80.04
Time computation	0.08	0.87	0.01	0.15	0.03
Efficiency	44.93	45.44	43.34	39.60	49.52

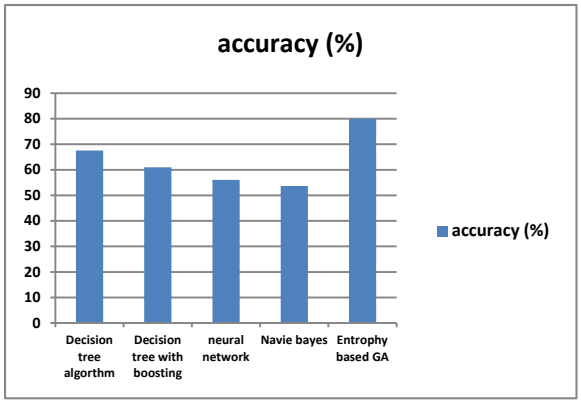

Fig. 5: Comparison of accuracy in (%)

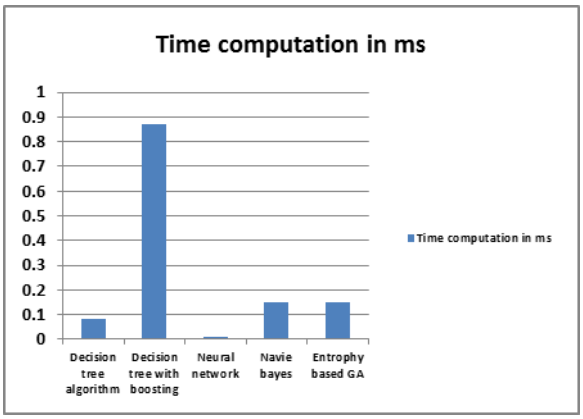

Fig. 6: Comparison of computational time

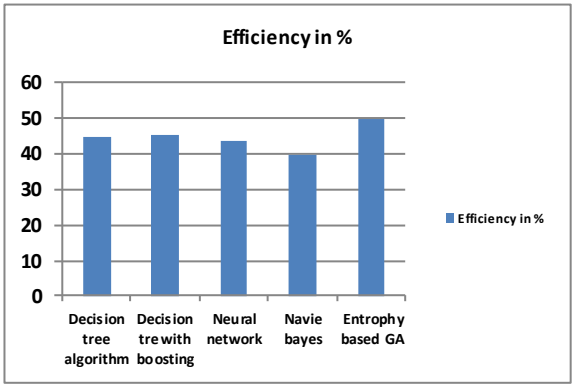

Fig.7: Comparison of efficiency in (%)

5. Conclusion

Data mining is the way toward breaking down a data from various planned and gives valuable information is based on that the results of foreseeing the diseases for a patient from the tremendous volume of data presents in the hospital information database. In this paper utilizing the association rule mining algorithm for concentrate the coordinated elements from the hospital information database and keyword based clustering algorithm is utilized to locate the exact disease which is influenced by the patient. The proposed proficient acquainted characterization algorithm utilizing entrophy genetic methodology for disease forecast brought about having high prescient accuracy and of high effectiveness values.

References

[1] S. Vijayarani* and S. Sudha "An Efficient Clustering Algorithm for Predicting Diseases from Hemogram Blood Test Samples" Indian Journal of Science and Technology, Vol 8(17), August 2015.

[2] Shakeel PM, Baskar S, Dhulipala VS, Mishra S, Jaber MM., "Maintaining security and privacy in health care system using learning based Deep-Q-Networks", Journal of medical systems, 2018 Oct 1;42(10):186.https://doi.org/10.1007/s10916-018-1045-z

[3] G. Purusothaman* and P. Krishnakumari "A Survey of Data Mining Techniques on Risk Prediction: Heart Disease" Indian Journal of Science and Technology, Vol 8(12), DOI: 10.17485/ijst/2015/v8i12/58385, June 2015.

[4] Jyoti Soni Ujma Ansari Dipesh Sharma "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" international Journal of Computer Applications, Volume 17– No.8, March 2011

[5] Shakeel PM, Baskar S, Dhulipala VS, Jaber MM., "Cloud based framework for diagnosis of diabetes mellitus using K-means clustering", Health information science and systems, 2018 Dec 1;6(1):16.https://doi.org/10.1007/s13755-018-0054-0

[6] Himigiri. Danapana, M. Sumender Roy,‖ Effective Data Mining Association Rules for Heart Disease Prediction System‖ IJCST Vol. 2,Issue 4, Oct . - Dec. 2011.

[7] Fariba Shadabi and Dharmendra Sharma,‖ Artificial Intelligence and Data Mining Techniques in Medicine – Success Stories‖ International Conference on BioMedical Engineering and Informatics-2008.

[8] Shakeel, P.M., Tolba, A., Al-Makhadmeh, Zafer Al-Makhadmeh, Mustafa Musa Jaber, "Automatic detection of lung cancer from biomedical data set using discrete AdaBoost optimized ensemble learning generalized neural networks", Neural Computing and Applications,2019,pp1-14.https://doi.org/10.1007/s00521-018-03972-2

[9] J. Liu, Y.-T. HSU, and C.-L. Hung, "Development of Evolutionary Data Mining Algorithms and their Applications to Cardiac Disease Diagnosis," in WCCI 2012 IEEE World Congress on Computational Intelligence, 2012, pp. 10–15.

[10] P. Chandra, M. . Jabbar, and B. . Deekshatulu, "Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature Subset Selection," in 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012, pp. 628– 634.

[11] P. Mohamed Shakeel; Tarek E. El. Tobely; Haytham Al-Feel; Gunasekaran Manogaran; S. Baskar., "Neural Network Based Brain Tumor Detection Using Wireless Infrared Imaging Sensor", IEEE Access, 2019, Page(s): 1

[12] S. U. Amin, K. Agarwal, and R. Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors," in Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013), 2013, no. Ict, pp. 1227–1231.

[13] Preeth, S.K.S.L., Dhanalakshmi, R., Kumar, R.,Shakeel PM.An adaptive fuzzy rule based energy efficient clustering and immune-inspired routing protocol for WSN-assisted IoT system.Journal of Ambient Intelligence and Humanized Computing.2018:1–13. https://doi.org/10.1007/s12652-018-1154-z

[14] Zhao, Q., Rezaei, M., Chen, H., Franti, and P.: Keyword clustering for automatic categorization. Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, (2012).

[15] Michael Pucher, F. T. W.: Performance Evaluation of WordNet-based Semantic Relatedness Measures for Word Prediction in Conversational Speech. (2004).

[16] K. Sudhakar, "Study of Heart Disease Prediction using Data Mining," vol. 4, no. 1, pp. 1157–1160, 2014.

[17] R. Chitra and V. Seenivasagam, "REVIEW OF HEART DISEASE PREDICTION SYSTEM USING DATA MINING AND HYBRID INTELLIGENT TECHNIQUES," Journal on Soft Computing (ICTACT), vol. 3, no. 4, pp. 605–609, 2013.

[18] Shanta kumar, B.Patil,Y.S.Kumaraswamy, "Predictive data mining for medical diagnosis of heart disease prediction" IJCSE Vol .17, 2011

[19] M. Anbarasi et. al. "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology Vol. 2(10), 5370-5376 ,2010.

[20] Hnin Wint Khaing, "Data Mining based Fragmentation and Prediction of Medical Data", IEEE, 2011.

[21] MA.Jabbar, Priti Chandra, B.L.Deekshatulu..:Cluster based association rule mining for heart attack prediction,JATIT,vol 32,no2,(Oct 2011)

[22] Ping Ning tan, Steinbach, vipin Kumar. : Introduction to Data Mining, Pearson Education, (2006).

[23] Picek, S., Golub, M.: On the Efficiency of Crossover Operators in Genetic Algorithms with Binary Representation. In: Proceedings of the 11th WSEAS International Conference on Neural Networks (2010)

[24] Aswathy Wilson, Gloria Wilson, Likhiya Joy K " Heart disease prediction using data mining techniques"

[25] Hnin Wint Khaing, "Data Mining based Fragmentation and Prediction of Medical Data", International Conference on Computer Research and Development, ISBN: 978-1-61284-840-2,2011