Best Feature Selection using Modified Whale Optimization Algorithm for Prediction of Heart Disease

M. Geethanjali, P. Madhubala

Abstract: ***Coronary illness is the confusion of heart and blood veins. It is hard for restorative specialists and specialists to foresee precise about coronary illness determination. Information science is the majorobject in primary expectation and takes care of huge information issues nowadays. This examination paper portrays the expectation of coronary illness in restorative area by utilizing information science. The same number of inquiries about done research identified with that issue however the exactness of expectation is yet should have been improved. Thus, this exploration centers on highlight choice methods and calculations where numerous coronary illness datasets are utilized for experimentation investigation and to appearance the precision development. In this work Modified Whale Optimization (MWOA) is utilized for highlight determination reason. By utilizing the Rapid digger as apparatus; Random Forest, (ANN), Decision Tree(DT) and Naive Bayes(NB) calculations are utilized as highlight choice procedures and improvement is appeared in the outcomes by demonstrating the exactness. From the proposed investigation the Artificial Neural Network grouping system is document better outcomes regarding Accuracy, Recall, Precision and F-measure.***

Keywords: Heart Disease Prediction, feature selection, Modified Whale Optimization (MWOA), Random Forest, Artificial Neural Network (ANN), Decision Tree (DT) and Naive Bayes (NB).

I. INTRODUCTION

Coronary illness additionally named as cardiovascular infection is a significant basic state of the heart and blood veins in larger part of passing's. This is the reason for misfortune in view of stroke or coronary failure which is 20 percent of all passings [1]. Various side effects and reasons for coronary illness are chest torment, hurt consume and stomach torment, torment in the arms, weakness, and perspiring. An ongoing report done in 2018 by WHO shows the outcome that 56.9 million passings happened on the planet during the year 2016 is because of heart ailments [2]. In 2008 17.3 million individuals completed in view of heart sicknesses [3]. WHO perceived the capability of information mining that it can anticipate beginning period of coronary illness and can give exact arrangement of the infection. Information mining is fundamentally the revelation of information from colossal measure of crude information. Information mining is otherwise called sub field of information the board [4]. Information mining has classified in primary models named as prescient model and descriptive Model. Prescient model is characterized as a model which is made to anticipate a specific result or result by utilizing prescient demonstrating systems [5]. While illustrative model is characterized as a model made to give a superior comprehend of information without focusing on a variable by utilizing investigation procedures like factor examination and bunch investigation and so on [6].

Information mining has many learning systems that can be helpful to watch gigantic prior accessible information new data. A few instances of the systems are: (DT), (MLP), (NB), K-closest neighbor (K-NN) and (SVM) [7]. Numerous information mining methods are applied on restorative information to find concealed certainties from a lot of information for example grouping, relapse, characterization, and exception and so on. A portion of the savvy models in social insurance things are (CSS) and (DSS). Clinical Decision emotionally supportive networks (CDSS) are use of DSS in medicinal services field which is intended to help specialists and other human services staff for developing and settling on medical choices. Choice emotionally supportive networks (DSS) are the data frameworks utilized in basic leadership exercises for different fields [8]. Computational insight has significant job in the forecast of coronary illness. Ideas that are utilized in calculation knowledge can find the connections between understanding characteristics and various sicknesses [9]. In the contemporary investigations, numerous scientists did their work by utilizing highlight choice method in expectation of coronary illness. Highlight determination is likewise named as factor choice or traits choice. Highlight determination is assorted from dimensionality decrease. Highlight determination centers around lessening the quantity of superfluous characteristics by certain procedures i.e trait subset choice while dimensionality decrease diminishes the quality set by creating new properties from given property set [10].

II. RELATED WORKS

Various creators have spoken to their exploration by investigating different systems which incorporate characterization, affiliation mining, grouping, and choice tress in various wellbeing fields. Different interminable illnesses like asthma, circulatory strain, diabetes cannot be relieved effectively, yet high danger of ailments can be controlled with precise and convenient update information of the patients.

Revised Manuscript Received on January 2, 2020.

M.Geethanjali, Assistant Professor, Department of computer Science, St.Joseph's College of Arts and Science for Women, Hosur, Tamil Nadu, India., E-mail : geethanjalim26@gmail.com

Dr.P.Madhubala, HEAD & Assistant Professor, Department of Computer science, Don Bosco College, Dharmapuri, Tamil Nadu, India. Email: madhubalasivaji@gmail.com

Isler [11] has dissected the pulse changeability to recognize injured person with systolic (CHF) from patients with diastolic CHF. Creators played out the characterization utilizing a multiple layer discernment and the closest relatives. The investigation was executed on a sum of 30 persons: 18 persons having systolic CHF and 12 having diastolic CHF. The most extreme precision is acquired as 96.43% with classifier named as MLF.

Sudhakar, K. what's more, Manimekalai, D.M. [12] has utilized both arrangement displaying strategies, and affiliation order method to anticipate the hazard to have a cardiovascular breakdown. For successful coronary illness forecast K-implies bunching with the choice tree strategy were placed. Another time the Cleveland Clinic Foundation Heart Infection dataset with 13 traits was utilized. The greatest expectation precision determined was 83.9% subsequent to testing various blends for the centroid.

Seyedamin et al. [13]. Scientists deals with various AI method and thinks about their outcomes in term of precision. Different AI methods are utilized in this investigation on little informational collection and contrasted the outcome and one another. SVM is prepared on restorative coronary illness dataset bringing about a classifier. To improve precision previously mentioned methods are applied namely Bagging, Boosting and Stacking. Utilizing Stacking procedure SVM, MLP classifier has greatest exactness 84.15% higher than different methods.

S. Prakash et al. [14] has proposed a coronary illness forecast which presented Optimality Criterion highlight determination (OCFS) for the extrapolation and capably analyze the coronary illness. Scientist improves their technique for harsh set element determination on data entropy (RFSIE). In this examination they contrast the OCFS and RFS-IE in word of computational period, expectation worth, and mistake level utilizing distinctive kind of informational indexes. OCFS strategy be able to take least performance time when contrasted with another one technique.

III. PROPOSED SYSTEM

The future methodology used to finish this exploration is begun by taking an open source UCI informational index. Subsequent to checking the dataset, following stage is preprocessing and information discretization as Data cleaning, Data Alteration, Data Reduction, Binning and Select Attributes. Subsequent to applying every one of these systems on downloaded dataset, the principle procedure highlight determination is applied. Presently, the following calculations are applied on the information i.e Random Forest, ANN, Decision Tree (DT) and Naive Bayes (NB).

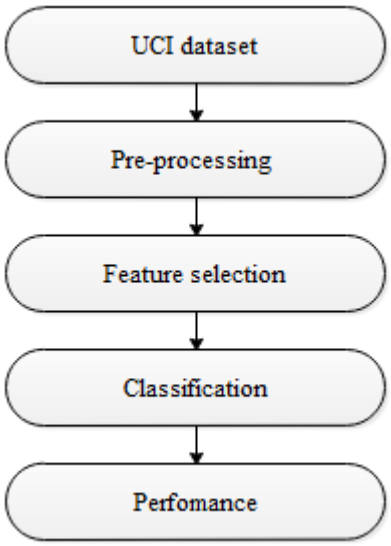

Fig.1.Flow chart of the proposed system.

A. UCI Dataset

The data of UCI archive is utilized in the database. There are complete 13 characteristics and 270 cases in this dataset. One of open online source dataset is UCI which is related with huge numbers of sicknesses and spreads a huge wellspring of databases, area speculations and information originators which are used by the specialists.

B. Preprocessing & Discretization

Future processing of information is displayed in a comprehensible introduction by transforming crude information into fathomable setting for intention.

C. Data Cleaning

Information cleaning is a procedure wherein information is cleaned by evacuating missing information, copy information and settling information irregularities. Therefore information quality is improved bringing about helpfulness of information.

D. Data Transformation

Change of information or data starting with one organization then onto the next arrangement is known as information change. It generally done when expected the source position to change over into needed organization for particular reason.

1. Data Reduction

Change of integer or string advanced data into a remedied arranged and disentangled structure tentatively or experimentally. The principle idea of information decrease is to diminish innumerable measures of information into valuable data.

2. Binning

Binning partitions gatherings and sum of nonstop qualities in toward little receptacles by utilizing equivalent recurrence or equivalent profundity binning procedures.

E. Feature selection

Highlight determination is additionally indicated as factor choice, Attribute choice or inconstant subset choice for typical development which hinders the way toward picking a subset of relevant highlights Feature choice. In this proposed framework (WOA) Algorithm is utilized for include determination.

1. Modified Whale Optimization Algorithm (MWOA) Feature Selection

In this segment, WOA is altered by modifying the control restriction and implanting other looking through techniques. In MWOA, three changes are projected and talked about in detail as pursues.

The significant issue of illuminating Large-scale worldwide enhancement (LSGO) with meta-heuristic calculations (Mas) is that most of them unite rashly toward neighborhood optima because of quick decrease of decent variety, and the first WOA is no special case. In the past investigations, the Lévy Flight (LF) process is broadly utilized in MAs to keep the arrangement from nearby optima and quicken the union speediness in light of its productive worldwide hunt capacity. Hence, a LF is utilized in MWOA to get away from the nearby optima by advancing the populace decent variety.

The LF is a sort of non-Gaussian arbitrary procedure with step length following a Lévy appropriation. A straightforward power-law vision of the Lévy conveyance is:

$$L(s)\sim|s|^{-1-\beta}, 0<\beta\leq 2 \quad (1)$$

Where

β - Index,

s -Step distance of the LF. Using Mantegna's algorithm to determine

$$s=\mu/|\vartheta|^{1/\beta} \quad (2)$$

Where, μandϑfollow normal distribution, that is

$$\mu\sim N\left(0,\sigma_\mu^2\right), \vartheta\sim N\left(0,\sigma_\mu^2\right) \quad (3)$$

$$\sigma_\mu=\left[\frac{\tau(1+\beta).\sin(\pi.\beta/2)}{\tau\left(\frac{1+\beta}{2}\right).\beta.2^{\frac{(\beta-1)}{2}}}\right]^{1/\beta} \quad (4)$$

$$\sigma_\vartheta=1 \quad (5)$$

To avoiding the step size the Lévy flight jumping out of the scheme domain is accepted. It'sremains by:

$$Levy=random\left(size(D)\right)\oplus L(\beta)\sim\frac{0.01\mu}{|v|^{\frac{1}{\beta}}(X_i-X^*)} \quad (6)$$

Where the scale of the problem is $size(D)$ and $\oplus$ means entry-wise multiplications, X_i is the i^{th} solution vector, Because of the limitless fluctuation of Lévy circulation, the LF implements the extended separation development sometimes for advancing the investigation capacity, while the short separation development is achieved for upgrading the misuse capacity. Clearly, this legitimacy can guarantee that MAs bounce out of nearby optima. In MWOA, the contracting encompassing instrument is supplanted by a Lévy trip so as to investigate the pursuit space all the more proficiently. The new location is refreshed by the accompanying principle.

$$Y(t+1)=Y(t)+\frac{1}{sqrt\,(t)}.sign(rand-0.5)\oplus Levy \quad (7)$$

Where $1/sqrt(t)$a constraint is related to the present iteration number t and $sqrt(\,)$represents the square root task. In this is regard, a huge range of finding movement can be performed during the initial level while a lesser one is employed in the later period. $Sign(rand-0.5)$signifies a sign function with values of (-1, 0, 1). Examination phase of MWOA is shortened.

$$Y(t+1)=\begin{cases}Y(t)+\frac{1}{sqrt\,(t)}.sign(rand-0.5)\oplus Levy & if\ p<0.5\\ D'.e^{bl}\cos(2\pi l)+Y^*(t) & if\ p\geq 0.5\end{cases} \quad (8)$$

F. Classification Algorithms

In preprocessed dataset the following grouping calculations are then applied

RF: Tree based technique is a RF that is utilized for together arrangement and relapse examination. Developed various trees and a mean forecast would be a yield for grouping.

ANN: A neural system depends on the possibility of organic neural systems; it is performed on the PC to carry out certain responsibilities like bunching, design rearrangement and characterization. ANN is a nonlinear measurable information model since complex connections among data sources and yields are demonstrated. The structure of ANN is influenced by the progression of data since this structure changes and learns dependent on the navigating information and yield in the neural system.

DT: DT is an approach to show the information. It utilizes a tree-like diagram as prescient model. The objective of DTs is to make a model to anticipate an outcome or a worth dependent on input factors. The outcomes are a significant order and broadly utilized for basic leadership. This method is a well-known instrument in AI that help to locate the proper procedure to arrive at the fantastic resolutions since it very well may be changed to a lot of significant standards by coordinating the root hubs to the leaf hubs.

NB: This procedure is a probabilistic classifier that utilizations Bayes hypothesis. The NB hypothesis is the accompanying

$$P(C|X)=P(X|C)\times\frac{P(C)}{P(X)}$$

Where

X -Information

C- Denote the class.

- Consistent or the equivalent for every one of the classes.

NB functions admirably on huge informational index, realizing that depends relying on the prerequisite that characteristics esteem are restrictively free which is unreasonable.

IV. RESULT AND DISCUSSION

The examinations are executed utilizing Python 3.7.3 idle on a computer with Intel Core i5 CPU 2.2 GHz with 8.00 GB RAM.

The content order network has generally utilized estimation exactness (the bit of records named positive that genuinely are certain), review (the segment of non-negative reports that are named positive). For a given class c_i, the general recipe for computing the exactness and review is given in the conditions (9) &(10).

$$Precision = \frac{|Documents\ \ correctly\ \ classified\ \ to\ the\ class\ c|}{|Total\ \ documents\ \ classified\ \ to\ class\ c|} \quad (9)$$

$$Recall = \frac{|Documents\ \ correctly\ \ classified\ \ to\ the\ class\ |}{|Total\ \ documents\ \ in\ class\ \ c|} \quad (10)$$

Accuracy is the calculateof statistical variability and a description of random errors. The total accuracy of text grouping results for determining the intrusion is specified in the Eq. (11).

$$Accuracy = \frac{|Total\ \ correctly\ \ classified\ \ documents\ |}{|Total\ \ number\ \ of\ documents\ |} \quad (11)$$

F-measure is the measure of accuracy test and it considers each precision and recall of the test in order to evaluate the score. The common formula for F-measure is given in the Eq. (12).

$$F - Measure = \frac{2*Precision\ *Recall}{Precision\ +Recall} \quad (12)$$

In this trial look into, reenacted UCI information is utilized for contrasting the exhibition assessment of existing philosophies and the proposed methodology as far as exactness, F-measure, accuracy, and review. The use of systems uncovers the aftereffects of every one of the four applying calculations NB, DT, ANN, and RF. By applying various information emulating learning methods and leading trials on the given dataset, we presume that the cross-breed determination model has the most elevated precision. It upgrades the presentation of coronary illness forecast model. The correlation of Accuracy, Recall, Precision and F-proportion of arrangement strategies is appeared in Table 1.

Table 1: Comparison of proposed results

Classification Technique	Accuracy (%)	Recall (%)	Precision (%)	F-measure (%)
WOA-NB	74.24	57.23	36.49	47.01
MWOA-NB	84	59.47	40.12	43.82
WOA-DT	67.4	58.13	47.51	47.49
MWOA-DT	87	60.25	45.85	49.07
WOA-ANN	75	65.82	65.24	65.17
MWOA-ANN	90.04	89.20	93.58	91.17
WOA-RF	74.24	74.74	74.20	74
MWOA-RF	84	81.16	80.85	80.83

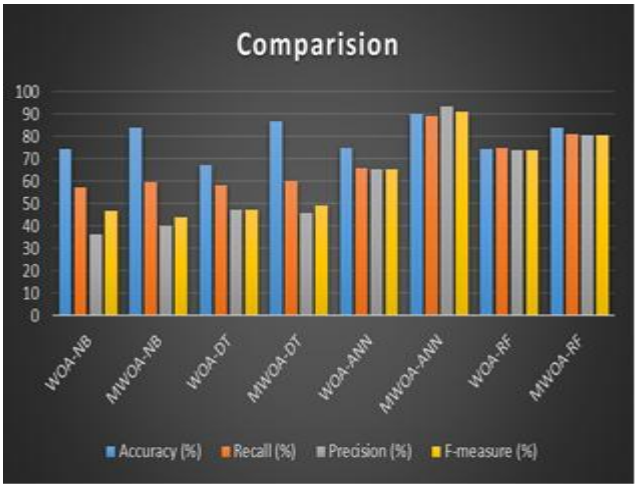

Fig.2. Comparison of proposed results.

From the table.1.And the Fig.2. Shows that the compared to WOA algorithm the proposed MWOA algorithm provides much better results for all classification techniques. In that specifically, compared to other classification algorithms the proposed MWOA technique is working better for ANN algorithm.

V. CONCLUSION

The primary aspiration of this paper is to improve exactness in forecast of coronary illness by utilizing highlight determination strategies. Various information mining systems for example NB, DT, ANN and RF. These are submitted independently in Rapid excavator on a UCI coronary illness date set and contrasted outputs and the previous inquires about. This examination accomplishes the objective which was according to desire and precision has been developed from past referenced qualities in writing audit. Exactness is significant for information mining in restorative industry. Various calculations can be applied for recognizing various kinds of maladies. This thusly makes the framework shrewd. Progressively combinational models are built to foresee the coronary illness which can help specialists in expectation of various kinds of heart ailments at a beginning time. The contrasted with WOA calculation the proposed MWOA calculation gives much better outcomes to all arrangement systems. In that explicitly, contrasted with other arrangement calculations the proposed MWOA strategy is working better for ANN calculation.

ACKNOWLEDGMENT

This research was supported by Dr.P.MADHUBALA, Research Supervisor. Her guidance helped me in all time of research and writing of this paper. I would like to express my sense of gratitude to ST. JOSEPH'S COLLEGE OF ARTS AND SCIENCE FOR WOMEN, Hosur for their support and encouragement. And I also like to thank PERIYAR UNIVERSITY, Salem for providing me the opportunity to carry out the research work.

REFERENCE

1. World Congress of Cardiology Scientific Sessions 2016 Volume 11, Issue 2, Supplement, Pages e1-e203, June 2016
2. World Health Organization. The world health report 2000: health systems: improving performance. World Health Organization, 2000.
3. ARCHANA, BADE, AHER DIPALI, and SMITA KULKARNI PROF. "International Journal On Recent and Innovation Trends In Computing and Communication." 2277-4804.
4. Silwattananusarn, Tipawan, and KulthidaTuamsuk. "Data mining and its applications for knowledge management: A literature review from 2007 to 2012." arXiv preprint arXiv: 1210.2872 (2012)
5. Leventhal, Barry. "An introduction to data mining and other techniques for advanced analytics." Journal of Direct, Data and Digital Marketing Practice 12, no. 2 (2010): 137-153.
6. Leventhal, Barry. "An introduction to data mining and other techniques for advanced analytics." Journal of Direct, Data and Digital Marketing Practice 12, no. 2 (2010): 137-153.
7. 22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth 2017

8. McKhann, Guy, David Drachman, Marshall Folstein, Robert Katzman, Donald Price, and Emanuel M. Stadlan. "Clinical diagnosis of Alzheimer's disease Report of the NINCDS- ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease." Neurology 34, no. 7 (1984): 939-939.
9. Nahar, Jesmin, Tasadduq Imam, Kevin S. Tickle, and Yi-Ping Phoebe Chen. "Association rule mining to detect factors which contribute to heart disease in males and females." Expert Systems with Applications 40, no. 4 (2013): 1086-1093.
10. El Mountassir, Mahjoub, SlahYaacoubi, José Ragot, Gilles Mourot, and Didier Maquin. "Feature selection techniques for identifying the most relevant damage indices in SHM using Guided Waves." In 8th European Workshop On Structural Health Monitoring, EWSHM 2016. 2016.
11. Isler, Y.: Discrimination of systolic and diastolic dysfunctions using multi-layer perceptron in heart rate variability analysis. Comput. Biol. Med. 76, 113–119 (2016).
12. Sudhakar, K., Manimekalai, D.M.: Study of heart disease prediction using data mining. Int. J. Adv. Res. Comput. Sci. Softw. Eng. 4(1) (2014).
13. Nahar, Jesmin, Tasadduq Imam, Kevin S. Tickle, and Yi-Ping Phoebe Chen. "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach." Expert Systems with Applications 40, no. 1 (2013): 96-104.
14. Long, Nguyen Cong, PhayungMeesad, and Herwig Unger. "A highly accurate firefly-based algorithm for heart disease prediction." Expert Systems with Applications 42, no. 21 (2015): 8221-8231.

AUTHORS PROFILE

M.Geethanjali was born in salem, Tamil Nadu, India, in the year 1983. She has working as an Assitant Professor, Deparment of Computer Science, St.Joseph's College of Arts and Science for Women, Hosur, Tamil Nadu, India. She received the Master of Science (M.Sc) in Computer Science from Periyar University, Salem, Tamil Nadu, India, in the year 2006 and Master of Philosophy(M.Phil.) of Computer Science from Periyar University ,Salem,TN,India, in the year 2007.Currently she is pursuing her Ph.D in Computer Science in Periyar University, Salem, Tamil Nadu, India under the guidance of Dr.P.Mahubala.

Dr. P.Madhubala Head &Asst Professor at the Department of computer Science in Don Bosco College Dharmapuri, Tamil Nadu, India. She completed her Ph.D. in computer Science at Mother Terasa University, Kodaikanal, Tamil Nadu, India, in the year 2017. She has 13 years teaching experience and 5 years research experience. Her area of interest in research are cloud computing, networks, datamining. she has published more than 25 papers at various national and international journals.