Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7766153

A NOVEL WEIGHTED OPTIMIZATION ALGORITHM TO CLASSIFY THE HEART DISEASE USING MACHINE LEARNING

P. Suganya
Ph.D. Scholar, Bharthiar University, Coimbatore-641046, TamilNadu, India

C.P. Sumathi
Associate Professor & Head, Dept. of Computer Science, Srimathi Devkunvar Nanalal Bhatt Vaishnav College for Women, Chennai, India.

Abstract: Heart disease is difficult to detect due to several risk factors, including high blood pressure, cholesterol, and an abnormal pulse rate. Accurate and timely identification of human heart disease can be very helpful in preventing heart failure in its early stages and will improve the patient's survival. Manual approaches for the identification of heart disease are biased and prone to inter examiner variability. Therefore, detecting heart disease early by utilizing the affluence of high-resolution intensive care records has become a challenging problem. That is why many researchers are trying their best to design a predictive model that can save many lives using data mining. Even though, some Machine Learning (ML) based models are also available, which can reduce the mortality rate, but accuracy is not up to date. According to the recommended study, using a Modified Weighted Empirical Score Optimization (MWESO) with Logistic Regression (LR) algorithm this research identified and predicted human heart disease. Machine learning (ML) algorithms like K-Nearest Neighbourhood (KNN), Support Vector Machine (SVM), Logistic Regression (LR) and Naïve Bayes (NB) have been applied to the heart disease dataset to predict the disease. At first, the LR model was trained. After training, sum of two features decision was combined using a weighted sum optimization. The weights have been assigned to each attribute's decision probability hence that each attribute's effect varies in the summation of weighted empirical score that gave the optimized prediction from the final decision score. The datasets were acquired from the heart diseases repositories from Kaggle. The comparative study has proven that the proposed MWESO algorithm with LR is the most suitable model due to its superior prediction capability to other Machine Learning with an accuracy of 90.7% on heart ailments dataset.

Keywords: Heart disease, Prediction, Modified Weighted Empirical Score Optimization, Machine Learning, Classification.

I. INTRODUCTION

A disease in the human body is an unnatural medical condition. It affects negatively the human body organism's functional state. It is generally associated with few signs of illness in the patient body. According to the World Health Organization (WHO), in the last 15 years, an estimated 17 million people die each year from cardiovascular disease, particularly heart attacks and strokes. Heart disease refers to a series of conditions that include the heart, vessels,

muscles, valves, or internal electrical pathways responsible for muscle contraction. According to the centers for Disease Control and Prevention (CDC), heart disease is one of the leading causes of death in India, the UK, the US, Canada, and Australia. Manually, detecting heart disease needs doing several tests.

II. OBJECTIVE

The primary purpose of this study is to classify patients with heart disease using medical records. The classification model in general can predict the severity stage of the patients with heart disease. This research work has used different ML algorithms to classify heart disease.

III. METHODOLOGY

The workflow of the system has been implemented in different stages including Pre-processing of the dataset, proposed NWEO algorithm, and classification and performance evaluation. Heart disease is diagnosed with the help of Kaggle datasets. Moreover, it is divided into training and testing set.

IV. DATA PREPROCESSING

Pre-processing data means the changes which are made on data before it is fed as an input to the algorithm. Data obtained from many sources is described as raw data, not suitable for analysis. In order to obtain better results, it is necessary to remove outliers, noise, and irregularities from the data, known as data cleaning.

Data Cleaning: The data that needs to be analysed using algorithms of machine learning may be noisy, inconsistent and incomplete. It also deals with the missing values for attributes of interest as it changes the proper average value for the attribute. Likewise, invalid attribute values are cleared and filled manually with its mean value. Data is cleaned up by manipulating missing values, smoothing out noisy data and removing outliers. The dataset utilized to predict heart disease based on 14 variables is shown in Figure 2.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figure 2: Recommended dataset based on randomizing the rows

V. PROPOSED MWEO ALGORITHM

The proposed algorithm specifies decision probability of (D_P) for each feature of heart dataset to predict the test data. From this decision score, the prediction was made for the test

data. Different weights have been assigned to each feature D_P so that each features effect varies in the summation of weights. If one of the features have a higher rate for the right decision, we assigned a bigger weight to that and a comparatively smaller weight to the other features. So, if one of the features has a weight of .65, then the other will have (1-.65) =.35. Here used a loop to check which weights provided the best accuracy for the summation of weighted empirical score model and selected them. The sum of the weights used in the empirical model should be 1 for scaling. Then selected the weights that have the best result for the summation of empirical optimized model, the equation for the weighted sum is as follows equation 1.

$$D_W = \sum_{P=1}^{N} W_P * D_P \quad (1)$$

Where N is the number of the features used for summation of empirical optimized model, then have used two features for every combination of empirical optimized model, so N=13. D_W is the weighted sum, which is the new decision score of the weighted empirical score optimization model. Based on this score, the final decision was given. W_p represents the weighted probability that has been assigned to feature with the decision score, and D_p is the decision probability score for any individual attributes.

The process of building a weighted empirical score optimization model is illustrated below. The thirteen features were used to create an empirical optimized model. A weighted score level optimization model was developed by merging those as a single weighted empirical feature score, and it worked better than the individual features scores.

Algorithm for MWESO algorithm

Input: W_p represents weights with the decision score, D_p represents individual attributes for decision score after training, N represents number of separated features

Output: D_W represents new decision score of the weighted empirical score optimization model

Step 1: W_1 = 1

Step 2: for p = 0 to 20 do

Step 3: D_p = 0

Step 4: $\sum_{p=1}^{N} W_p$

Step 5: $W_1 = W_1$ -0.05

Step 6: W_2 = 1 - W_1

Step 7: for p = 0 to N do

$D_W = \sum_{p=1}^{N} W_p * D_p$

End

End

Step 8: Select the weights (W1, W2) that gave the highest decision score.

Step 9: Final weighted empirical score optimization using selected weights.

After selecting weights, the weighted empirical optimization rule was applied at the last step in the algorithm. A weight was assigned to each individual features decision score in a weighted sum. To select the weights, we used a loop for using various values of weights and the weights that have the highest decision score were selected. The individual result was combined, but the outcome of the 13 features in LR algorithm is not taken equally. Rather than weights were assigned that decided the effect of any features in the weighted empirical score optimization model. Train the above algorithm with LR model and this step is implemented at the decision level.

VI. RESULT AND DISCUSSION

In this research work, proposed MWESO with LR, KNN, LR, SVM and NB classifier algorithms are applied to the heart diseases datasets acquired from Kaggle repository respectively. The dataset used in this research is splitting into 75% and 25%, which 75% of original data is considered as training dataset and 25% as testing dataset. Training dataset is used to train a model and testing dataset to check the performance of the trained model. A notable improvement was found in the result of using weighted empirical optimization models. The key reason behind the proposed model's improvement is that if there is a miss classification in first weight score, there is a probability that another weight score may classify that particular data correctly. So, after summation of weighted empirical score, there is a chance that we get the correct result for that specific case. This concept assists in giving the weighted empirical optimization model a better efficiency. The model's performance can be interpreted from the value of these parameters. The first True Negative (TN) classifier predicted "no heart disease" and identified patients who are not affected by sepsis. The second False Positive (FP) classifier predicted individuals who are not affected by "heart disease". The third False Negative (FN) classifier accurately recognized patients with heart disease while predicting "no heart disease". The fourth True Positive (TP) is a classifier that predicted "heart disease" and identified those who had it. To illustrate the classifiers' performance, a confusion matrix has been used for the proposed MWESO classification model with LR, KNN, RF, SVM and NB classifier. It summarizes the results of the predictions of a model.

Table 1: The Performance metrics

Machine Learning Models	TP	TN	FP	FN
Proposed MWESO with LR	36	33	3	4
KNN	33	28	8	7
NB	28	34	6	8
LR	32	28	2	14
SVM	29	30	8	9

Performance Parameter

Classification model performance is measured with the term of accuracy, precision, recall, sensitivity and specificity.

Accuracy is the measure of the percentage of correctly classified objects.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} * 100 \quad (2)$$

Precision is also referred to as the false-positive rate. From precision, we get the number of correctly classified observations as positive to the total classified positive observations.

$$\text{Precision} = \frac{TP}{TP+} \quad (3)$$

Recall is often referred to as a truly positive rate. It is the ratio of total positive assumptions and the total amount of positive class attributes.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

Sensitivity: It determines how much of a classifier to identify positive labels.

$$Sensitivity = \frac{TP}{TP+FN} \quad (5)$$

Specificity: It is assessed what proportion of patients to identify negative labels.

$$\text{Specificity} = \frac{TP}{FP+FN} \quad (6)$$

Table 2: Performance metrics based on proposed model with various classification models

Model	Accuracy	Precision	Recall	Sensitivity	Specificity	Kappa
Proposed MWESO with LR classifier	0.907	0.9231	0.9000	0.9000	0.9167	0.9565
KNN classifier	0.8026	0.8049	0.8250	0.8250	0.7778	0.7104
NB classifier	0.8158	0.8235	0.7778	0.7778	0.8500	0.7124
LR classifier	0.7895	0.9412	0.6957	0.6957	0.9333	0.7035
SVM classifier	0.7763	0.7838	0.7632	0.7632	0.7895	0.7026

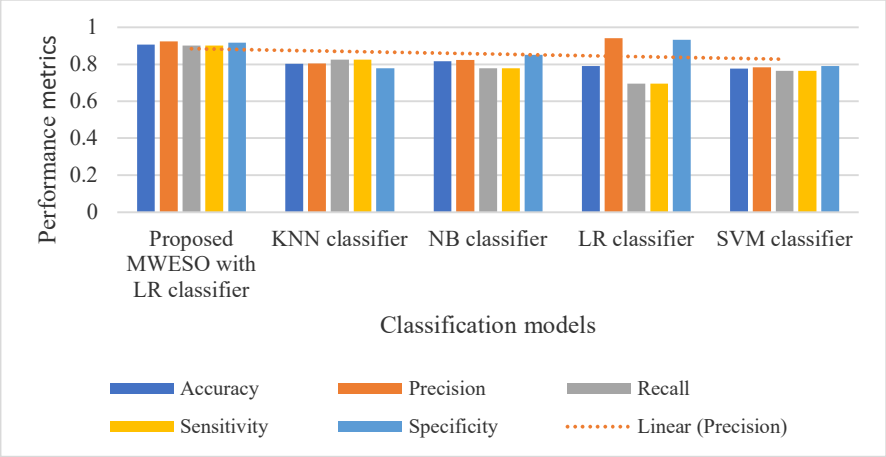

Figure 3: Graphical representation of performance evaluation based on various models

Table 2 and figure.3 shows the performance of proposed MWESO classification model with LR, KNN, LR, SVM and NB classifier. After the process of building a weighted empirical score optimization model based on merging a single weighted empirical feature score which result as performance parameters increased (LR) for proposed MWESO with LR classifier. It has the highest accuracy of 90.7% among the other classifier models.

VII. CONCLUSION

In light of the recent rise in malicious software and hacking attempts, the implementation of an intelligent security system has become essential. AI approaches are more adaptable and resilient when compared to contemporary cyber security solutions; as a result, they increase security execution and better guard systems from an expanding variety of sophisticated cyber threats. Despite the profound shift that AI has brought about in the field of cyber security, associated frameworks are not yet prepared to totally transform and, as a result, adapt to changes in their state. Although there are numerous advantages to using AI methods for cyber security, it is important to keep in mind that AI is not the sole solution to security problems.The intelligent security system will be rendered ineffective when it is attacked by a human adversary with the intention of circumventing it. This doesn't imply we shouldn't use AIapproaches; nevertheless, we should be aware of the constraints they have. AI requires ongoing engagement and training from human beings. This combined strategy has shown success on numerous occasions, and it collaborates with security experts in an effective manner.

REFERENCES

[1] Alsmadi, I. (2019). Cyber Security Management. The NICE Cyber Security Framework, 243–251. https://doi.org/10.1007/978-3-030-02360-7_10

[2] Behavioural science in cyber security. (n.d.). Cyber Security: Law and Guidance. https://doi.org/10.5040/9781526505897.chapter-027

[3] Bockus, N. F. (2015). Cyber in space: 2035. Advances in Information Security, 39–57. https://doi.org/10.1007/978-3-319-23585-1_4

[4] Bradbury, R. (2021). Educating for cyber (security). The Oxford Handbook of Cyber

Security, 394–408. https://doi.org/10.1093/oxfordhb/9780198800682.013.24

[5] CRUZ LOBATO, L. U. Í. S. A. (n.d.). Unraveling the cyber security market: The struggles among cyber security companies and the production of Cyber (in) security. https://doi.org/10.17771/pucrio.acad.27784

[6] Cyber security 2019. (2019). 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security). https://doi.org/10.1109/cybersecpods.2019.8885065

[7] Cyber security 2020 cover page. (2020). 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security). https://doi.org/10.1109/cybersecurity49315.2020.9138881

[8] Cyber security evolution. (2012). Cyber Security Policy Guidebook, 15–38. https://doi.org/10.1002/9781118241530.ch2

[9] Cyber Security Objectives. (2012). Cyber Security Policy Guidebook, 39–67. https://doi.org/10.1002/9781118241530.ch3

[10] Dunn Cavelty, M. (2018). 27. cyber-security. Contemporary Security Studies, 410–426.https://doi.org/10.1093/hepl/9780198804109.003.0027

[11] Gostev, A. (2012). Cyber-threat evolution: The Year Ahead. Computer Fraud &Security, 2012(3), 9–12. https://doi.org/10.1016/s1361-3723(12)70052-0

[12] Goyal, D., & Rajput, R. S. (2020). Cloud computing and security. The Evolution ofBusiness in the Cyber Age, 293–319. https://doi.org/10.1201/9780429276484-12

[13] Raiu, C. (2012). Cyber-threat evolution: The past year. Computer Fraud & Security,2012(3), 5–8. https://doi.org/10.1016/s1361-3723(12)70051-9

[14] Security and Trust in Cyber Space. (n.d.). Cyber Law and Cyber Security in Developingand Emerging Economies. https://doi.org/10.4337/9781849803380.00005