



M-GWO Algorithm to Predict Risk of Silent Heart Attack of Diabetes Patients - Cardiabetes Model

Malti Nagle^{*}, Prakash Kumar^{*}

Department of Computer Science & Engineering, Jaypee Institute of Information Technology, Noida 201309, India

Corresponding Author Email: 18403002@mail.jiit.ac.in

Copyright: ©2024 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/isi.290428>

ABSTRACT

Received: 12 January 2024

Revised: 7 June 2024

Accepted: 1 August 2024

Available online: 21 August 2024

Keywords:

cardiodibet framework, grey wolf optimization, GEETN, diabetes, cardiac arrest, stack model, continued-value based attributes, cholesterol

Modern healthcare system is innovation presided of next generation. Diabetes has been considered most chronic diseases and source of cardiac arrest disease. In this paper, healthcare framework has been proposed to diagnose risk of cardiac arrest due to diabetes mellitus. Findings of around 997 patients has been taken from various sources of cardio vascular disease and diabetes. Novelty of work is to pre-process dataset using GEETN process missing value with novel imputation method, and also proposed a modified Grey Wolf Optimization (M-GWO) algorithm, applied to the task of selecting an optimal feature subset for classification purposes using different machine learning models. The accumulated comparison is based on outcomes that consist of various algorithm with various algorithm like SVM_RBF, DT, KNN, RF, MLP and proposed model. Accuracy of 99 % MCC (70.35%), and f1 score (99.16 %) that helps in early detection of patients' health condition to reduce the rate of death cases, cardioidibet framework for healthcare systems helps in providing better monitoring, communication and early diagnosis of diabetes and cardiac health of patients. The proposed method identifies the preliminary status of diabetes and cardiac vascular diseases parameters of patient through Normal, Moderate and high-risk further message send for critical cases.

1. INTRODUCTION

In today's scenario, majority of the diseases are on the rise due to sustained stress. Hectic lifestyle is also adding stress in everyone's life resulting in fatal health problems. Stress related issues are giving birth to severe health diseases like Diabetes, hypertension and weaken heart. Now a days, cardio vascular diseases are widely noted in majority of populations [1]. It has also become prevalent in early age of populations. Doctors and practitioners conduct inclusive survey and consolidate the data for heart related diagnosis and study. It has been observed that 40% mortality are because of cardiac arrest and diabetes patients. Morbidity of 60% due to diabetes. Many doctors state that traditional model is taking long time to diagnose. Traditional model sometimes diagnose of patients' cardiac test after the cardiac arrest occurs. Cardiac arrest is epidemic in current days, diabetes adds more complication to such patients. Diabetes mellitus is addressed as chronic disease. Diabetes may establish risk of vital organs and so cardiac arrest. All of these are major challenges in saving patients' lives. The data gathered from survey of cardiac patients consist noticeable facts about their health status. If the symptoms can be diagnosed at early stages by any prediction model, it will be highly beneficial for patients to take timely consultation and cure for the heart diseases (like cardio vascular arrest, Heart coronary diseases, etc.).

In modern healthcare systems, various sensors are installed

in patients' body, that can continuously read data from sensors and keep forwarding this data at cloud to process and analyse [2]. This real time health data is difficult to extract such massive and large dataset. Sometimes this information is in the hidden format. Streaming of such massive data is vital challenge for researchers. Techniques applied on real time data streaming can facilitate researchers attain two insights: novel approach and solicitous extraordinary perception of enormous health related datasets. Prediction and labelling of the diseases received through real-time data communication method is a tedious task [3, 4]. This paper, applied various high precision and high accuracy algorithm to the cardiac related dataset to acquire the unidentified trends observed in heart patients. A dataset of 997 patients consisting of 8 attributes are taken into consideration to examine and find the accuracy for different algorithms. Accuracy and recall kappa values are examined and health status of patient is categorised based on dataset value. The sample dataset has been collected from various cycles using contiguous monitoring sensors data of family members + records are collected from association of peripheral adipose tissues pathology with type 2 diabetes in Asian Indians and "prevented from any alteration. Further, Supervised learning technique is applied to refine the dataset and then accuracy is measured. The dataset labelling has been done in category into three stages (Label 0: normal range of attributes, Label 1: borderline and above and Label 2: high risk) of cardiac arrest due to diabetes mellitus.

2. LITERATURE SURVEY

Stress related health diseases are very prevalent due to improper routine and busy schedule of every human being. It is escalating the chances of critical medical issues in person. Some of them are incurable diseases for example cardiac stroke and diabetes are very commonly observed in patients. This motivated the researchers to understand the hidden pattern of medical datasets. This section enlightens the study of real time medical dataset to acknowledge the heart diseases on the basis of their medical diagnosis. Data mining techniques that are used in previous works are also addressed in this section [4, 5]. The gap identified in existing works motivated to introduce suitable techniques that can classify the dataset to address the issues related to heart diseases. World organization stated that cardiac diseases are the most significant challenge majority of population are facing in recent days [6, 7].

In different paper several machine learning (ML) methods applied on various dataset to check false positive and true positives [8]. It has been observed from various articles that, the gathering of information constitutes a segment of retrospective studies on cardiovascular diseases, employing the recordings from multichannel Magnetocardiography (MCG). The database encompasses individuals with both coronary stenosis and those in good health. Within the sample, there are 16 instances of NSTEMI cardiological infraction and lipid profile with angiography is conducted for the ischemic group [9, 10]. The coronary heart diseases can be predicted using ensemble model which can be helpful for early prediction of critical heart risks [11].

ML algorithm has been proposed in article [1, 12] to classify the diabetes using different supervised learning model. The multilayer perceptron along with decision tree and other algorithm has been used in the work. The prediction shown in research has defined the lowest false positive and lowest false negative rates under curve of 86%.

To detect heart diseases using ECG. Specific signal range of ECG has been captured for observation 20 number of times reading of ECG has been encapsulated based on different probabilities of ECG signals reach to Peak and low waves of patients' ECG [13-15].

Classifier algorithm (Neural Network) as a solution for finding is an ensemble-based model proposed as a new model which deals with cardio vascular related issues, combination of predicted value and probability of multiple predecessors. The successful implementation and testing was obscured for 215 patients' dataset and an accuracy of 97.4% was achieved [16].

3. PROBLEM STATEMENT

As healthcare data is exponentially growing day-by-day in globe. Traditional model not sufficient to spread awareness in population for various new diseases worldwide. Therefore, diagnose critical diseases accurately was hard in the traditional health model. In case of emergency diagnose and provide treatment in minimum time to the patients reside at a remote location also tough for doctors. Traditional medical records were not measured accurately because of human error. In traditional health system diagnose chronic diseases efficiently was time taken process that may increase death ratio in OT. The key challenge is to handle the healthcare data in real time.

In traditional healthcare system, patients visit to hospital for their health issues. In critical cases, patients sometimes could not be able to reach hospitals. If any, healthcare system can access the data from remote location. It will help to save life of critical patients. Problem in traditional healthcare systems are as follows:

Real-time healthcare data is continuously changing, so old model cannot fit with new data. Thus, model has to be altered.

Volume of real-time data is too high, traditional tree algorithms cannot work as tree height and width would increase exponentially.

Another point of discussion is health issues of patients suffers from diabetes type 2. Diabetes type -2 patients have the high risk of silent heart attack and hyperglycemia. Which requires instant treatment. If such, patients can be diagnosed and treated from remote location could be helpful to save their lives. The possible approach is explored to overcome all problems in traditional model.

4. DATASET

The dataset considered for the analysis comprises of diabetes and cardiac related records based on the cholesterol HDL and LDL long with insulin values. Data is collected from association of peripheral adipose tissues pathology with type 2 diabetes in Asian Indians. Total 997 sample records including Asian Indians lab and from different cycles of family members and "without any alteration [17]. It has total various attributes (12 input, 1 output) related to prescribe heart and diabetes diagnosis based on blood sample taken from patients. Table 1 consists of the attributes and its specifications.

Table 1. Attributes of health patient

Attributes	Specifications
Gender	The gender specification of patient
Phenotype	The normal glucose tolerance test a genotype of patient based on blood sample
Triglycerides	One kind of fat. It gets reserved in body from calories that are not burned in routine lifecycle and are stored in body as extra calories.
Cholesterol	It is a fat like waxy substance that gets accumulated in all cells and required to generate healthy cells and hormones in blood. But if it is in excess then this Attributes creates problem in blood flow rate and gives rise to the heart diseases. Sometimes it form clot in blood vein that leads to heart stroke
HDL	Lipoproteins with high-density. It is cholesterol good for heart because it does not remain in blood for long time. HDL sends by body parts to liver and liver eliminate it from body.
LDL	It known as Low- density lipoproteins. It is also called 'bad' cholesterol it gets accumulated around the arteries.
VLDL	it is known as Very low-density lipoproteins. It contains high number of triglycerides and is also considered as one of the reasons to buildup cholesterol around the arteries walls.
Insulin	Insulin is a form of hormone that let the body to use glucose available in food to produce energy. is
...	----- etc.
Outcome	This attribute used to check the health status of one's heart based on the attributes collected from blood strains.

5. PROPOSED METHODOLOGY

Proposed model dinocarid consist of three phases:

GEETN (Pre-processing technique) is a novel approach,

M-GWO (Modified-Grey Wolf Optimization) important feature selection,

Stacking (Train and test Model). To accomplish the diagnosis of Diabetes + cardiac profile of patients both the model contributed successfully to produce efficient outcomes.

Figure 1 shows the proposed model where the process consists various steps. The process of filter the data is based on statistics and list is sorted towards the leaf node splitting of dataset is completed by arguments values.

5.1 Pre-processing technique (GEETN)

This proposed novel technique (GEETN) is to pre-process the raw data consisting of findings of cardio vascular and diabetes. First attributes are labelled based on reading of attributes - (Label 0: normal range of attributes, Label 1:

borderline and above and Label 2: high risk) of cardiac arrest due to diabetes mellitus. At first all the missing values were identified in raw healthcare data. Imputation is conducted with the standard values of attributes based on the standard healthcare defined by medical research center.

[$LDL = TC - HDL - 0.2 * Triglycerides$] is one of the of calculation is applied. Another example blood glucose level ranges below 80 considered as normal and label = 0, moderate range 80 from 180 label = 1, and high glucose above 180 label = 2. In similar manner all attributes have been calculated using (Eqs. (1)-(2)). Outcomes is labeled using Eq. (3).

Pre – processing – algorithm (GEETN) - The missing and null value of dataset has been changed on the basis of standard value of attributes defined by doctor. Second challenge is to process real time health data streaming. Figure 2 shows how to handle the null value is also matter of concern. Because any value in health data is specific health reading Why..... ? – Health attributes can be replaced with any random value and not with mean average value. It cant be 0 either.

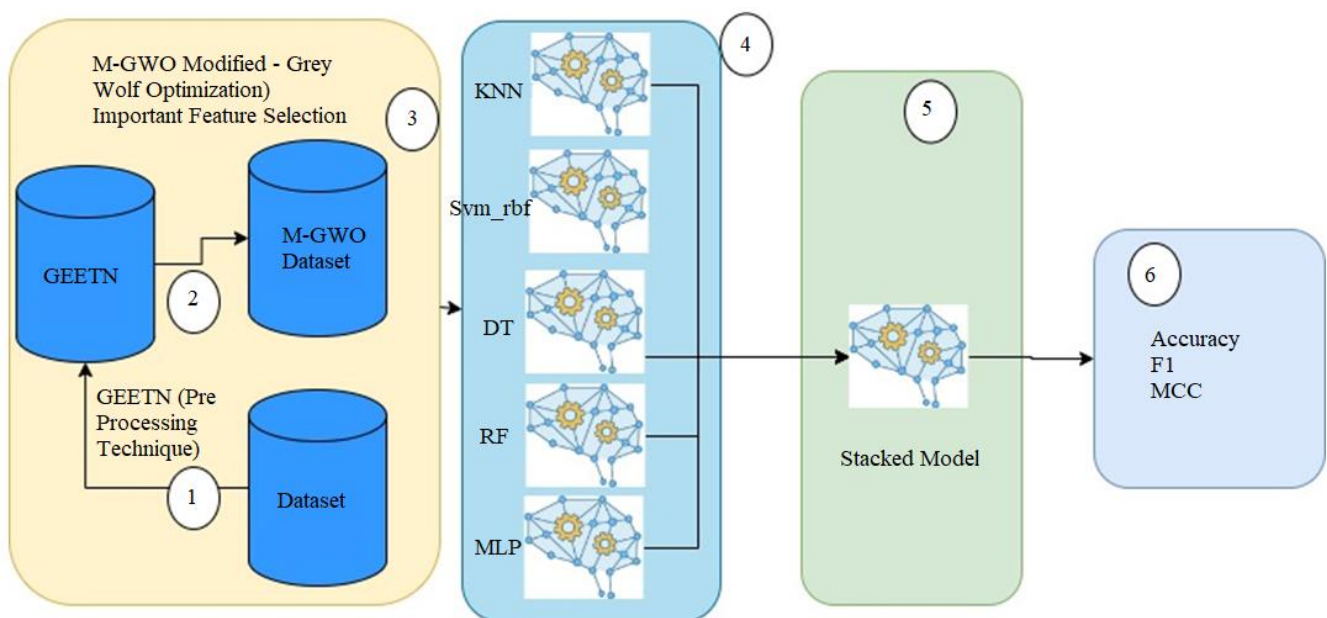


Figure 1. Proposed model – flow chart representation of model

What....? If any all essential attribute is missing for calculation of health status.

Cholest erol	HDL	LDL	VLDL	Sr-cr	HOMA- B	HOMA- R
200	46	76	36	0.74	49.34	0.46
169	45	78	28	0.9	93.41	2.37
259	42	96	28	0.9	108.93	2.56
196	38	98	30	1.67	123.1	2.46
158	29	58	26	2	86.73	2.8
256	34	100	36	0.9	88.62	2.42
206	44	98	23	1.1	56.4	1.23
230	34	144	52		166.07	2.25
226	39	78	28	0.9	124.28	1.69
261	40	80	49.44		348.71	8.85
150	30			1.09	4.18	0.06
192	40	78	26	1.07	604.47	3.37

cannot replace by any random value or 0

Figure 2. Categorical dataset

The following steps is applied to calculate GEETN.

(1) All value of attributes has been calculated for its normal range of health parameters.

(2) Health dataset is pre-processed using the normal range of each attribute. Then the outcomes are calculated according to the value of each attribute.

(3) The proposed method then identify as the null value in healthcare dataset and the healthcare dataset based on the novel approach.

(4) Imputation method for categorical columns is applied.

(5) Impute the missing value using GEETN.

(6) The missing value is first detected and replaced by calculating values based on other parameters.

(7) GEETN is applied on attributes those are correlated with other attributes. For example, cholesterol value predicted by blood strain is correlated with (LDL, HDL, VLDL). Likewise, Hb1ac.

(8) Imputation is evaluated on the basis of health value of different attributes.

(9) Dependent variable is calculated by using following equation:

$$(output)^n = \sum_{(i=0)}^n (input_attributes) \quad (1)$$

(10) Further the output (dependent variable) is labelled according to the range of abnormality in diagnosis. Equation to label dependent variable:

TN= Total number of labels considered in range of abnormality

y= average of TN

z= $\sum_y^{TN} z$

$$|label_{outcome}: x| = \begin{cases} x = 0, & x = 0 \\ x = 1, & x > 0, x \leq y \\ x = 2, & x \geq z, x \leq z \end{cases} \quad (2)$$

(11) Output is the new leaf node attribute from all observations of input attributes that has been defined for normal health of patients.

$$(output)^n = \sum_{i=0}^n (input_attributes) \text{ For normal healthy readings of patient.} \quad (3)$$

5.2 Grey wolf optimization

Typically, grey wolves exhibit a preference for living in packs, with an average group size ranging from 5 to 12 individuals. Within the social structure of a grey wolf pack, there are stringent rules governing dominance and hierarchy. The pack consists of the following roles:

- I. Alphas, serving as leaders, are responsible for decision-making, and their directives are followed by the entire pack.
- II. Betas, subordinate to alphas, assist in decision and other group happenings. Betas can be either gender, and they often emerge as strong contenders for assuming the alpha role.
- III. Omegas play the role of scapegoats and are required to update all dominant wolves. They also considered as last in line to partake in meals.
- IV. Alphas and betas may have authority over deltas, but deltas exert control over omegas. This group comprises scouts, sentinels, elders, hunters, and caretakers. Scouts

vigilantly watch over territorial borders, warning the pack of potential threats. Sentinels are dedicated to safeguarding and ensuring the safety of the pack. Elders, having previously held alpha or beta roles, bring experience and wisdom. Hunters collaborate with alphas and betas in capturing prey to sustain the pack. Caretakers, on the other hand, attend to the needs of weak, ill, and injured pack members.

The framework of Grey Wolf Optimization model, the most optimal solution is denoted as alpha (α). The subsequent two best solutions are referred to as (β) and (δ) delta [18]. The remaining solutions for another candidate are collectively designated as (ω) omega. Hunting [2] pattern is provided by these four entities— α , β , δ , and ω —each following one of the three aforementioned candidates [4]. To simulate the encircling pack's behavior necessary to hunt the prey, the mathematical modeling of this behavior is represented by the following equations, namely Eqs. (4)-(7).

$$\vec{P}(t+1) = \vec{P}\rho(t) + \vec{W} \cdot \vec{Z} \quad (4)$$

In Eq. (5), \vec{D} is represented with iteration number t, \vec{W} and \vec{Y} are vectors of coefficients. $\vec{P}\rho$ denotes the prey position, while \vec{P} signifies the grey wolf position [6].

$$\vec{W} = |\vec{Y} \cdot \vec{P}\rho(t) - \vec{P}(t)| \quad (5)$$

The $\vec{W} \cdot \vec{Y}$ vector are calculated as in Eqs. (6) and (7):

$$\vec{x} = 2x \cdot \vec{r1} - x \quad (6)$$

$$\vec{Y} = 2r2 \quad (7)$$

The process involves linearly decreasing the parameter 'a' range 2 to 0 throughout iterations, with 'r1' - 'r2' [18] representing random vectors within the range [0, 1]. Generally, the alpha guides the hunting, with occasional participation from beta and delta. Wolves' positions are updated by expression as in Eq. (8).

$$\vec{P}(t+1) = \frac{\vec{P}_1 + \vec{P}_2 + \vec{P}_3}{3} \quad (8)$$

$\vec{P}_1, \vec{P}_2, \vec{P}_3$ are define as in Eqs. (9)-(11), respectively.

$$\vec{P}_1 = |\vec{P}_\alpha - \vec{W}_1 \cdot \vec{Z}_1| \quad (9)$$

$$\vec{P}_2 = |\vec{P}_\beta - \vec{W}_2 \cdot \vec{Z}_1| \quad (10)$$

$$\vec{P}_3 = |\vec{P}_\gamma - \vec{W}_3 \cdot \vec{Z}_1| \quad (11)$$

where $\vec{P}_\alpha, \vec{P}_\beta, \vec{P}_\gamma$ best 3 solution found in first phase with t iteration of swarm, $\vec{W}_1, \vec{W}_2, \vec{W}_3$ are mentioned in Eq. (3), and $\vec{Z}_\alpha, \vec{Z}_\beta, \vec{Z}_\delta$ Eqs (12)–(14) defined, respectively.

$$\vec{Z}_\alpha = |\vec{Y}_1 \cdot \vec{P}_\alpha - \vec{P}| \quad (12)$$

$$\vec{Z}_\beta = |\vec{Y}_2 \cdot \vec{P}_\beta - \vec{P}| \quad (13)$$

$$\vec{Z}_\gamma = |\vec{Y}_3 \cdot \vec{P}_\gamma - \vec{P}| \quad (14)$$

where $\vec{V}, \vec{Y}_2, \vec{Y}_3$ are define as in Eq. (5).

A concluding note regarding the GWO pertains to the adjustment of the parameter 'x'. The parameter 'a' undergoes a linear update during each iteration, transitioning from 2 to 0 in accordance with Eq. (15).

$$\alpha = 2 - t \frac{2}{Maxiter} \quad (15)$$

Algorithm 1: Grey Wolf Optimization (GWO)

Input:

n: Total count of G-W in the group,
N-Iter: Iteration counts to optimize.

Output:

P α : Best grey_wolf location,
f(P α): Optimal fit_ess cost.

Initialization:

1. Initialize n grey_wolf positions from random population.
2. Find Alpha, Beta, and Delta:
3. Find the Alpha, Beta, and Delta solutions grounded on their appropriateness values.
4. Optimization Loop:
 - While ending criteria are not met, do the following:
 - a. Update Positions:
 - For each wolf i in the pack, update the wolf's current position calculated using Eq. (5).
 - b. Update Parameters:
 - i. Update the parameters a, A, and C.
 - c. Calculate Position:
 - ii. individual wolves positions evaluate.
 - d. Update
 - iii. Update $\alpha, \beta,$ and $\delta.$

End Optimization Loop

Note: The algorithm iteratively updates the positions of the grey wolves, adjusts parameters, evaluates fitness, and updates the alpha, beta, and delta solutions until the stopping criteria are met.

Modified Grey Wolf Optimization (M-GWO):

Wolf locations are continually adjusted to any point in space using the GWO technique. Certain unique issues, such feature selection, have solutions that can only be found in the ranges from {0,1} binary digits, by which a unique version of the GWO is needed. This proposed model proposes novel M-GWO for the job of feature selection. Each wolf is drawn to the top three best solutions by the wolves updating equation [2], which is a function of three location vectors, $x\alpha, x\beta,$ and $x\delta.$ The pool of solutions in the M-GWO is always in binary form, with every solution located on a hypercube's corner.

$$P_i^{t+1} = Crossover(P1, P2, P3) \quad (16)$$

where Crossover [19] considered as preferred cross over among solutions - x; y; z and P1; P2; P3 vectors of binary, indicating impact on movement of wolf towards alpha, delta, and beta grey_wolves in sequence, are represented by P1, P3, and P2. The calculations for these vectors are determined through Eqs. (17), (20), and (23), respectively.

$$P_i^d = \begin{cases} 1 & \text{if } (P_\alpha^d + bstep_\alpha^d) \geq 1 \\ 0 & \text{Otherwise} \end{cases} \quad (17)$$

Here, P_α^d denotes the alpha wolf's position vectors in

dimension d, and $bstep_\alpha^d$ is represented as binary step for d dimension, which can be computed according to the formula presented in Eq. (18).

$$bstep_\alpha^d = \begin{cases} 1 & \text{if } cstep_\alpha^d \geq rand \\ 0 & \text{Otherwise} \end{cases} \quad (18)$$

In this context, "rand" represents a randomly generated number from a uniform distribution. Meanwhile, $cstep_\alpha^d$ stands for the continuous-valued step size [20] associated with d dimension. This step size may be determined by employing a sigmoidal function [2], as outlined in Eq. (19).

$$cstep_\alpha^d = \frac{1}{1 + e^{-10(A_1^d(D_\alpha^d - 0.5))}} \quad (19)$$

where W_1^d are evaluated using Eqs. (3), and (9) in the d dimension.

$$P_2^d = \begin{cases} 1 & \text{if } x_\beta^d \geq 1 \\ 0 & \text{Otherwise} \end{cases} \quad (20)$$

In this scenario, P_2^d signifies the beta wolf's position vectors of wolf in d dimension, and a binary $bstep_\alpha^d$ step in mentioned in d dimension, the computation is specified in Eq. (21).

$$bstep_\beta^d = \begin{cases} 1 & \text{if } cstep_\beta^d \geq rand \\ 0 & \text{Otherwise} \end{cases} \quad (21)$$

In this context, "rand" denotes a randomly generated number from a uniform distribution. Additionally, $cstep_\alpha^d$ represents the continuous-valued step size [2] associated with dimension d. The calculation of this step size involves the utilization of a sigmoidal_function, as described in Eq. (22).

$$cstep_\beta^d = \frac{1}{1 + e^{-10(A_1^d(D_\beta^d - 0.5))}} \quad (22)$$

where $W1, Z_d,$ and d for β are calculated using Eqs. (3), (9) in the (d) dimension.

$$P_3^d = \begin{cases} 1 & \text{if } (P_\delta^d + bstep_\delta^d) \geq 1 \\ 0 & \text{Otherwise} \end{cases} \quad (23)$$

In this context, P_δ^d designates the delta wolf's position vectors in dimension (d), while $bstep_\beta^d$ represents a binary step for d dimension. The computation of this binary step is outlined in Eq. (24).

$$bstep_\delta^d = \begin{cases} 1 & \text{if } cstep_\delta^d \geq rand \\ 0 & \text{Otherwise} \end{cases} \quad (24)$$

The modified version is executed using two distinct methods. In the initial method, steps followed for top three best solutions are labeled and binarized. A stochastic crossover is then employed among these 3 fundamental moves for determining the grey wolf's next expected position. The next approach involves using a sigmoidal function to transform the continuously updated position. The resulting values are stochastically thresholder to ascertain the updated position of the grey wolf. Both methods for the modified Grey Wolf

Optimization (M-GWO) are applied within the field of feature selection, with the objective of identifying a subset of all features that maximizes classification accuracy while minimizing the selected features. Figure 3 presents the proposed stacked Model.

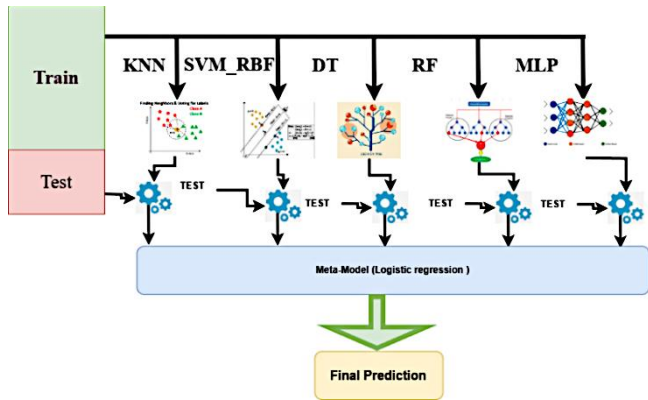


Figure 3. Proposed stacked model

Algorithm 2: Proposed stacked model

Input:

- (D_Train) is dataset, attributes (X) and outcome label (y)
- Bs modls (BM1, BM2, ..., BMn)
- Me_ta_modl (MetaM)

Output:

- Tst_Data (D_tst) predictn rslt

1. Train data split label: multiple folds to split dataset (FM1, FM2,.....FMn) to apply cross validation.
2. initialize stacke data_set: Generate matrix (SM) with empty value, load the predicted result from Bs_modls into SM. Iteration for each FMi fold in FM fold: FMi fold value split to prepare training_set (FMi_training), a validation_set (FMi_value).
3. Predicted values store in Bs_modls: base model from 0 to BMi. BMi trained from FMi_training FMi is used to predict FMi_value. Add the predicted result to SM.
4. MetaM is trained based on stacke data_set: input attributes are stored SM with corresponding true labels from FMi_value outcome variable.
5. SM is used to train the me_ta-modl MetaM. Create predictions from me_ta-modl: For each base model BMi: Train BMi on the entire training dataset. Prediction based on BMi. Matrix TM gets appended based on prediction.
6. Gets final value of predicted results and update me_ta-modl: Labeling is done for from predicted results of BM dataset.
7. Output: Test dataset is prepared from Final prediction (D_test).

KNN:

KNN is acknowledged as forthright and logical machine learning technique. Nearest data points are compared. New

data points are assigned to nearest neighbours. Where "k" is total neighbours needs to considered for calculation of data points. KNN is simple to use and understand. The KNN [21] is beneficial for small size of dataset, while apply is on bigdata may be disappointing.

Random Forest: RM is a well-organized Machine Learning technique. It is applied for both problems – classification and regression. Tree is generated from randomly selected training data; group of decision tree are produced from dataset. afterwards, final prediction is produced by complete results of several decision tree.

Decision Tree (DT): An efficient ML approach based on numerous input features, model is constructed by selected tree and potential outcomes. Best suitable tree is selected among all iteration. In case of overfitting DT is complicated.

SVM (Support vector machines): The aim of SVM is technique in which margin and hyperplane is used to classify dataset. Margin is used to decide boundary by taking support-vectors for hyperplane.

6. RESULT AND DISCUSSION

Proposed algorithm introduced advance healthcare system that can successfully predict diabetes patient’s health status. It also helps to predict heart health and cardiac arrest chances at early stage which is very critical to predict in traditional healthcare system. It was observed by doctors that due high diabetes patient died due to silent cardiac arrest. Proposed framework cardiobibet is capable to predict the risk of cardiac arrest at early stage. The diabetes detection health system generated using prediction of real time health data stream consist of diabetes and heart related attributes as it is combinational dataset, where attributes are dependent and missing value calculation is key challenge. Therefore, it requires pre-processing algorithm by applying imputation methods. The proposed work contributed GEETN pre-processing algorithm to prepare training dataset after pre-processing. Figure 4 depicts the classification of patients based on risk of cardiac arrest due to diabetes mellitus. Results are categorised in three distinguished states: Norma, Moderate and At High Risk. It has been observed that out 997, 23 patients are below the risk of cardiac arrest. Mortality is 91 out of 997 and morbidity cases are 883.

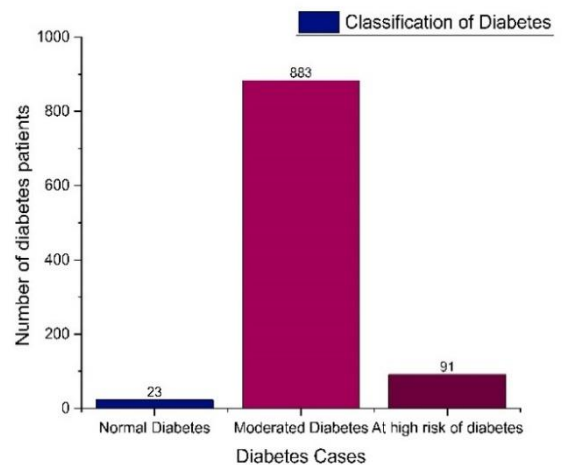


Figure 4. The results categorised the health status of patients

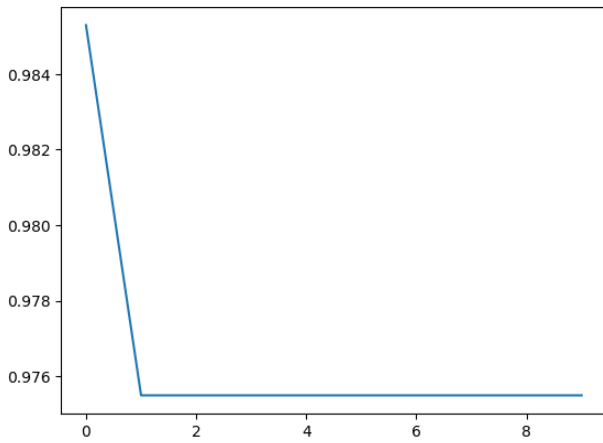


Figure 5. Modified Grey Wolf Optimization (M-GWO) feature selection

In this study presents a Modified Grey_Wolf Optimization (M-GWO) algorithm, applied to the task of selecting an optimal feature subset for classification purposes. This technique is applied on the preprocessing dataset obtained from technique. After 10 iteration we gotten features are selected Gender, Phenotype, NEFA, hsCRP, Triglycerides, Cholesterol, LDL, VLDL, Sr-cr, HOMA-R. as shown in Figure 5.

In Figure 6 depicts the count of patients have cholesterol different categories Normal diabetes, Moderated diabetes, at high risk of diabetes of M-GWO dataset.

In Figure 7 depicts the skewness of patient's data have Gender, Phenotype, hsCRP, Triglycerides, Cholesterol, HOMA-R different categories Normal diabetes, Moderated

diabetes, at high risk of diabetes of M-GWO dataset.

In Figure 8 depicts the correlation of patient's data having features Gender, Phenotype, NEFA, hsCRP, Triglycerides, Cholesterol, LDL, VLDL, Sr-cr, HOMA-R different categories Normal diabetes, Moderated diabetes, at high risk of diabetes of M-GWO dataset.

In Figure 9 depicts the importance of patient's data having features Gender, Phenotype, NEFA, hsCRP, Triglycerides, Cholesterol, LDL, VLDL, Sr-cr, HOMA-R different categories Normal diabetes, Moderated diabetes, at high risk of diabetes of M-GWO dataset. Higher the value denotes more important features. Based on observation Figure 9 shows the priority of features for investigation of risk of cardiac arrest due to diabetes mellitus.

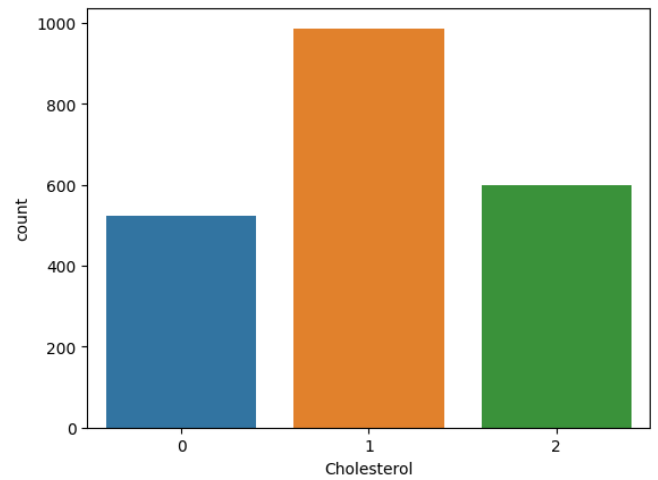


Figure 6. The results categorised the health status of patients

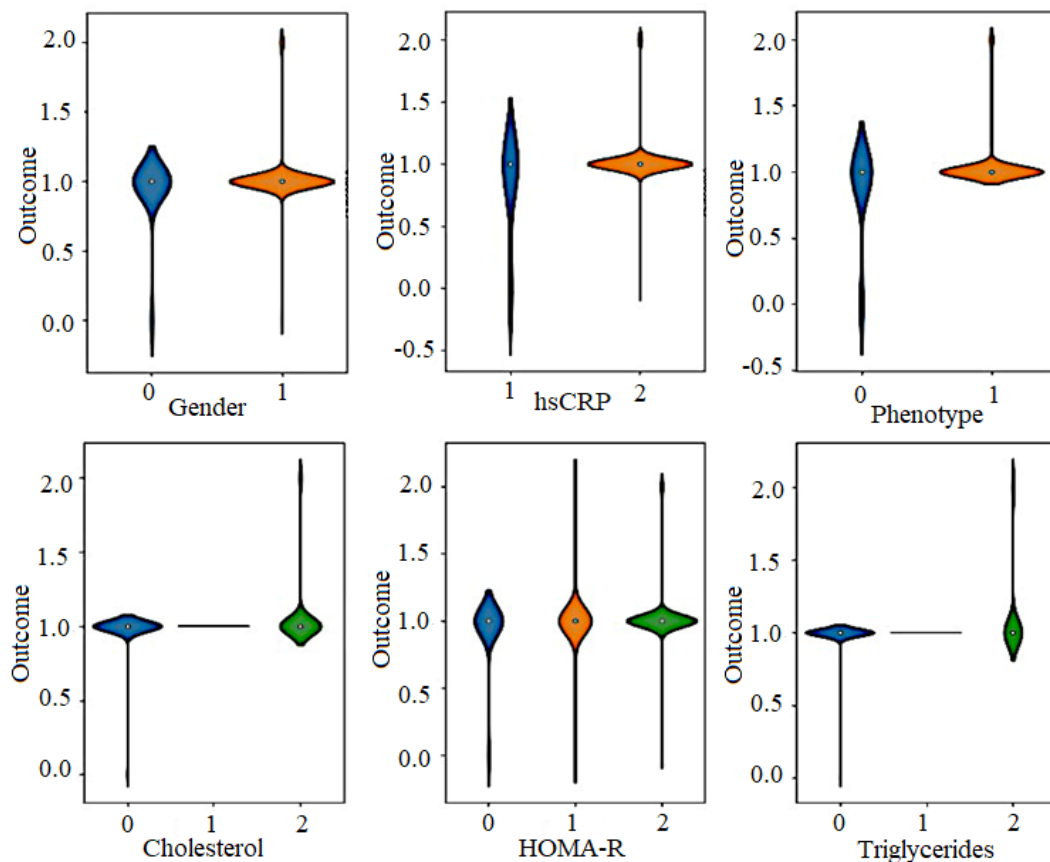


Figure 7. Skewness of patients data



Figure 8. Correlation among different features

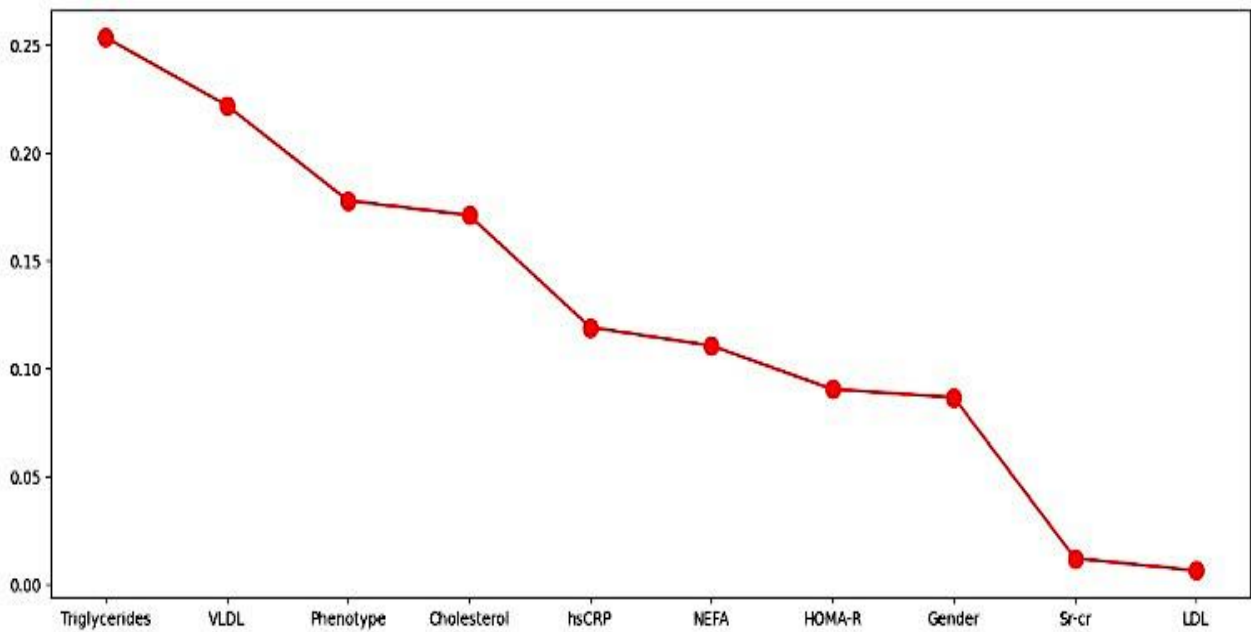


Figure 9. The results categorized the health status of patients

Accuracy Measurement: Accuracy serves as an assessment of the overall accuracy of the model, gauging the ratio of correctly predicted instances to the total instances.

Matthews Correlation Coefficient (MCC): The MCC depicts true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [14]. Its consider ranges from -1 to 1, where 1 denoting a flawless prediction, 0 indicating no improvement over random chance, and -1 signaling complete discordance between prediction and observation.

Harmonic Mean of Precision and Recall (F1 score): The F1 is the harmonic mean of recall and precision [10] striking a

balance between the classifier's ability to avoid mislabeling negative samples and its capability to identify all positive samples.

Table 2 and Figure 10 shows overall training performance pf various algorithm such as KNN SVM_RBF, DT, RF, MLP and proposed model. accuracy of 96.65% MCC (36.76%), and f1 score (95.02 %) that helps in early detection of patients' health condition to reduce the rate of death cases, cardioidibet healthcare systems helps in providing better monitoring, communication and early diagnosis of diabetes and cardiac health of patients.

Table 3 and Figure 11 shows overall training performance of various algorithm such as KNN SVM_RBF, DT, RF, MLP and proposed model. accuracy of 99% MCC (70.35%), and f1 score (99.16%) that helps in early detection of patients' health condition to reduce the rate of death cases, cardioidibet healthcare systems helps in providing better monitoring, communication and early diagnosis of diabetes and cardiac health of patients.

Table 2. Overall training performance of various algorithm

	Accuracy	MCC	F1
knn	0.950948	0.606191	0.957273
Svm_rbf	0.960981	0.231563	0.942816
dt	0.963211	0.333375	0.945750
rf	0.965440	0.402098	0.951409
mlp	0.958751	0.000000	0.938561
Proposed model	0.964326	0.367657	0.950252

Table 3. Overall testing performance of various algorithm

	Accuracy	MCC	F1
knn	0.96	0.438086	0.972921
Svm_rbf	0.99	0.000000	0.985025
dt	0.98	-0.005051	0.980000
rf	0.98	0.572471	0.986565
mlp	0.98	-0.010101	0.980000
Proposed model	0.99	0.703526	0.991641

The pre-processed data were further used for prediction. Comparison outcome was illustrated by comparing various

prediction algorithms various algorithm such as KNN SVM_RBF, DT, RF, MLP and proposed model helps on the cardiac related dataset [20-22]. The pre-processing of is conducted by comparing the data for the healthy range of cardiac blood sample's readings. Outcomes of cardioidibet suggested that proposed classifier gives better performance to filter the dataset and provides the health status of cardiac patient with more accuracy. Handling real time data streaming of the missing value for health data is also the challenge that could overcome by GEETN method. Table 3 shown description of proposed algorithm with different parameters - accuracy of proposed method is proposed accuracy of 99% MCC (70.35%), and f1 score (99.16%) that helps is also higher than another traditional algorithm. Overall performance of proposed method is outperform as compared to another traditional algorithm. The graphical representation of all parameters shown in Figure 10 is well defined. Along with performance the proposed method also categories the number of patients according to their health status Normal, Moderate and High risk of their health status on the basis of prediction. Figure 3 shows the counts the health status of patients by checking the risk factor evaluated by attributes of diabetes and cardiac symptoms reading. Proposed model in developing cardioidibet supports to provide outcomes of real time healthcare data, following parameters have been produced to analyse the outcomes of diabetes + cardiac data. Matrix consists of - accuracy, MCC, f1 score are calculated using label_outcome as output attribute and patient's medical readings input attributes. Figure 10 represents the visualization of performance of cardioidibet healthcare system.

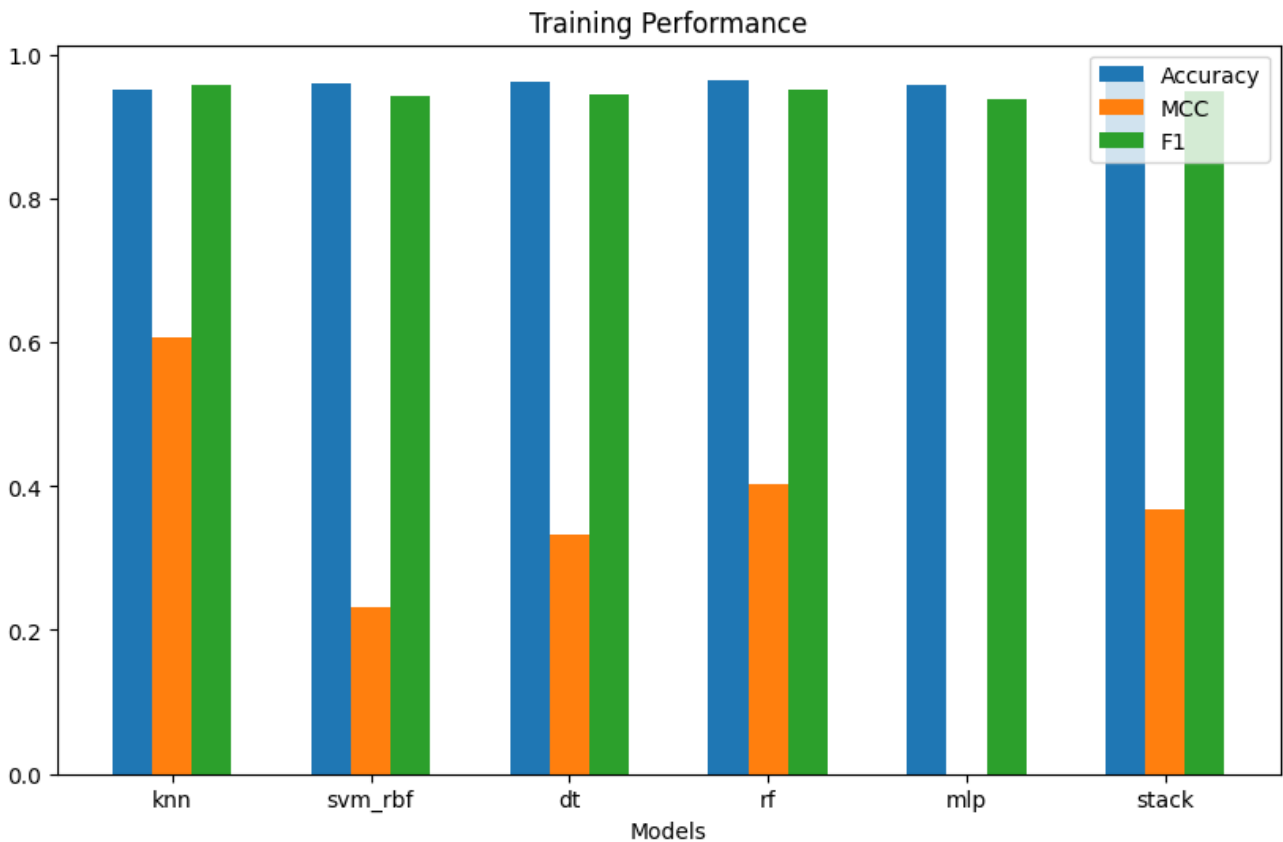


Figure 10. The results categorized the health status of patients

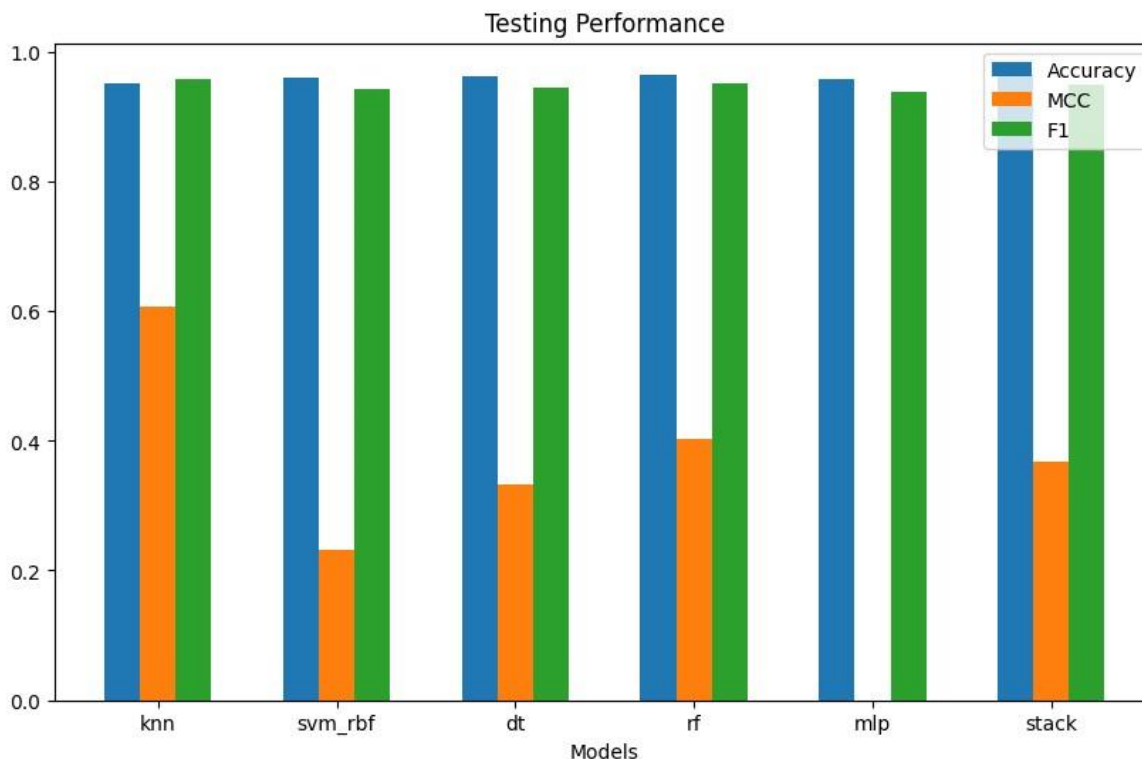


Figure 11. The results categorized the health status of patients

7. CONCLUSIONS

To detect the healthcare real time data (Based on Diabetes and cardiac diseases), algorithm applied on dataset those are best for feature prediction and splitting the dataset to predict the disease accurately. Comparison of traditional algorithm and proposed model applied for prediction of health disease. various algorithm such as KNN SVM RBF, DT, RF, MLP [23] and proposed model. accuracy of 99 % MCC (70.35%), and fl score (99.16 %). Overall performance of proposed model is outshined than as compared to traditional algorithms. features to evaluate risk of cardiac arrest due to diabetes. The graphical representation of all parameters shown in Figure 11 is well defined. Along with the performance, the proposed method also categories the number of patients according to their health status (Normal, Moderate and High risk) using proposed methodology. Novel approach applied in proposed work counts the health status of patients by checking the risk factor evaluated by attributes of diabetes and cardiac symptoms reading. proposed model algorithm gives 99% accuracy helps to predict the health disease at early stage based on patient's data and precautionary measures may be suggested by doctors. The cardioidibet is based on prediction of cardiac + diabetes symptoms the classification of patient's health status can be improved. Framework is successfully predicting cardiac arrest and the impact of diabetes over cardiac arrest. In future, framework can be able to predict the effect of diabetes on other vital organ those are not addressed in current studies like – kidney, eye retina etc. and predict the risk of damage in another vital organ.

ACKNOWLEDGMENT

This work would not have been possible without

exceptional support of my supervisor, Prakash Kumar. His inspiration and motivation kept my work on track. I give my sincere thanks to anonymous peer reviewer for insightful comments on my work.

REFERENCES

- [1] Theerthagiri, P., Ruby, A.U., Vidya, J. (2022). Diagnosis and classification of the diabetes using machine learning algorithms. *SN Computer Science*, 4(1): 72. <https://doi.org/10.1007/s42979-022-01485-3>
- [2] Thaiparnit, S., Kritsanasung, S., Chumuang, N. (2019). A classification for patients with heart disease based on hoefding tree. In 2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE), Chonburi, Thailand, pp. 352-357. <https://doi.org/10.1109/JCSSE.2019.8864158>
- [3] Domingos, P., Hulten, G. (2000). Mining high-speed data streams. In *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, MA, pp. 71-80. <https://doi.org/10.1145/347090.347107>
- [4] Hulten, G., Spencer, L., Domingos, P. (2001). Mining time-changing data streams. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 97-106. <https://doi.org/10.1145/502512.502529>
- [5] Butt, U.M., Letchmunan, S., Ali, M., Hassan, F.H., Baqir, A., Sherazi, H.H.R. (2021). Machine learning based diabetes classification and prediction for healthcare applications. *Journal of Healthcare Engineering*, 2021(1): 9930985. <https://doi.org/10.1155/2021/9930985>
- [6] Wang, J., Liu, C., Li, L., Li, W., Yao, L., Li, H., Zhang, H. (2020). A stacking-based model for non-invasive

- detection of coronary heart disease. *IEEE Access*, 8: 37124-37133. <https://doi.org/10.1109/ACCESS.2020.2975377>
- [7] Qi, Z.Y., Zhang, Z.R. (2021). A hybrid cost-sensitive ensemble for heart disease prediction. *BMC Medical Informatics and Decision Making*, 2021: 73. <https://doi.org/10.1186/s12911-021-01436-7>
- [8] Mirmozaffari, M., Alinezhad, A., Gilanpour, A. (2017). Data mining classification algorithms for heart disease prediction. *Int'l Journal of Computing, Communications & Instrumentation Engg. (IJCCIE)*, 4(1): 11-15. <https://doi.org/10.15242/IJCCIE.DIR1116008>
- [9] Jia, S. (2020). A VFDT algorithm optimization and application thereof in data stream classification. In *Journal of Physics: Conference Series*, 1629(1): 012027. <https://doi.org/10.1088/1742-6596/1629/1/012027>
- [10] Bifet, A., Holmes, G., Pfahringer, B., Kranen, P., Kremer, H., Jansen, T., Seidl, T. (2010). Moa: Massive online analysis, a framework for stream classification and clustering. In *Proceedings of the first workshop on applications of pattern analysis*, pp. 44-50.
- [11] Bifet, A., Gavalda, R. (2009). Adaptive learning from evolving data streams. In *Advances in Intelligent Data Analysis VIII: 8th International Symposium on Intelligent Data Analysis, IDA 2009, Lyon, France*, pp. 249-260. https://doi.org/10.1007/978-3-642-03915-7_22
- [12] Nourmohammadi-Khiarak, J., Feizi-Derakhshi, M.R., Behrouzi, K., Mazaheri, S., Zamani-Harghalani, Y., Tayebi, R.M. (2020). New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection. *Health and Technology*, 10(3): 667-678. <https://doi.org/10.1007/s12553-019-00396-3>
- [13] Habib, A.Z.S.B., Tasnim, T., Billah, M.M. (2019). A study on coronary disease prediction using boosting-based ensemble machine learning approaches. In *2019 2nd International Conference on Innovation in Engineering and Technology (ICIET)*, Dhaka, Bangladesh, pp. 1-6. <https://doi.org/10.1109/ICIET48527.2019.9290600>
- [14] Shorewala, V. (2021). Early detection of coronary heart disease using ensemble techniques. *Informatics in Medicine Unlocked*, 26: 100655. <https://doi.org/10.1016/j.imu.2021.100655>
- [15] Sharma, H., Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research (IJSR)*, 5(4): 2094-2097.
- [16] Das, R., Turkoglu, I., Sengur, A. (2009). Diagnosis of valvular heart disease through neural networks ensembles. *Computer Methods and Programs in Biomedicine*, 93(2): 185-191. <https://doi.org/10.1016/j.cmpb.2008.09.005>
- [17] Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 3(22): 41-46.
- [18] Hoeglinger, S., Pears, R. (2007). Use of hoeffding trees in concept based data stream mining. In *2007 Third International Conference on Information and Automation for Sustainability*, Melbourne, VIC, Australia, pp. 57-62. <https://doi.org/10.1109/ICIAFS.2007.4544780>
- [19] Kumar, A., Kaur, P., Sharma, P. (2015). A survey on Hoeffding tree stream data classification algorithms. *CPUH-Research Journal*, 1(2): 28-32.
- [20] Gomathy, V., Padhy, N., Samanta, D., Sivaram, M., Jain, V., Amiri, I.S. (2020). Malicious node detection using heterogeneous cluster based secure routing protocol (HCBS) in wireless adhoc sensor networks. *Journal of Ambient Intelligence and Humanized Computing*, 11(11): 4995-5001. <https://doi.org/10.1007/s12652-020-01797-3>
- [21] Thanga Selvi, R., Muthulakshmi, I. (2020). An optimal artificial neural network based big data application for heart disease diagnosis and classification model. *Journal of Ambient Intelligence and Humanized Computing*, 12: 6129-6139. <https://doi.org/10.1007/s12652-020-02181-x>
- [22] Biswal, A.K., Singh, D., Pattanayak, B.K., Samanta, D., Chaudhry, S.A., Irshad, A. (2021). Adaptive fault-tolerant system and optimal power allocation for smart vehicles in smart cities using controller area network. *Security and Communication Networks*, 2021(1): 2147958. <https://doi.org/10.1155/2021/2147958>
- [23] Guo, A., Pasque, M., Loh, F., Mann, D.L., Payne, P.R. (2020). Heart failure diagnosis, readmission, and mortality prediction using machine learning and artificial intelligence models. *Current Epidemiology Reports*, 7: 212-219. <https://doi.org/10.1007/s40471-020-00259-w>