Medical Big Data Analysis using Binary Moth-Flame with Whale Optimization Approach

Saka Uma Maheswara Rao[1]
Department of Computer Science and Systems Engineering
Andhra University College of Engineering (A), Andhra University
Visakhapatnam, India

K Venkata Rao[2]
Department of Computer Science and Systems Engineering
Andhra University College of Engineering (A), Andhra University
Visakhapatnam, India

Prasad Reddy PVGD[3]
Department of Computer Science and Systems Engineering
Andhra University College of Engineering (A), Andhra University
Visakhapatnam, India

***Abstract*—The accurate analysis of medical data is dependent on early disease detection and the value of accuracy is reduced when the medical data quality is poor. However, existing techniques have lower efficiency in handling heterogeneous medical data and the complexity of the features was not enhanced using an optimal feature selection model. The present research work has used the machine learning algorithm effectively for chronic disease prediction such as heart disease, cancer, diabetes, stroke, and arthritis for the frequent communities. The detailed information about the attributes is required to be known as it is significant in analyzing the medical data. The process of selecting the attributes plays an important role in decision-making for medical disease analysis. This research proposes Binary Moth-Flame Optimization (B-MFO) for effective feature selection to achieve higher performance in small and medium datasets. Additionally, the Whale Optimization Algorithm (WOA) is used that showed better performances for LSTM that suited well for the process of classification to predict the time series. The present research work utilizes Spark Streaming layers for data streaming to diagnose using Long Short Term Memory (LSTM) with whale optimization approach which is from the heterogeneous medical data. The proposed B-MFO-WOA method results showed that the proposed method obtained 97.45% accuracy better compared to the existing Modified adaptive neuro-fuzzy inference system of 95.91% of accuracy and B-MFO of 92.43 % accuracy for the models.**

Keywords*—Binary moth-flame optimization; complexity of the features; medical data; long short term memory; spark streaming layers; whale optimization algorithm

I. INTRODUCTION

The healthcare business uses huge data of medical treatment and recordings of every patient. The medical information is recorded and printed with various versions that are converted to digital versions efficiently [1]. Due to the extreme volume of patient details, it is possible for enhancing the quality of health care effectively for saving expenses. All the information present could be used in the multi-health care discipline such that health care illness and surveillance are preventative for management [2]. The Big Data (BD) tools are combined with the machine learning and data mining techniques that showed challenges in areas such as health care, education, transportation, and social media with other networks. The machine learning techniques were used that encompassed the phrase that referred to the large datasets. The regular utilization is progressively performed with little that indicates the basic intricacy and lay the first stone with subsequent ethical and misunderstandings of possible ways [3].

The BD movement is applied to unlock endeavor of large dataset values for making the decision, improving the outcomes, efficiency, and data owners have shown the deliverables. The goals are required to be accomplished, that is, collected, stored, access, managing data with various forms turns the volume with the simple steps [4–7]. Intelligence is applied on data points of high information to improve efficiency. Digital equipment usage increases for model and amount of data increases with unparalleled rate [8–10]. The deep knowledge discovery is computed in big healthcare data which has achieved the best results. However, the selection of the optimum subset of relevant and effective features is used for constructing an accurate model. Thus, the selection of features from a vector of one or zero is used for constructing an accurate model. Apache Spark was deployed in the cloud as it focused to apply on the ML models.

The contribution of the research work is as follows:

- The use of the Whale Optimization Algorithm (WOA) showed better performances for LSTM that suited well for the process of classification to predict the time series and solve optimization problems. To compute simulation of prey search, and prey encircling, humpback whales of bubble-net foraging are mimicked.
- Transfer functions such as U-shaped, V-shaped and S-shaped were applied to convert the continuous value to binary values for the feature selection process using the B-MFO technique.

The organization of the research paper is shown as follows: Section II is the literature review of the existing models and Section III illustrates the proposed method. Section IV shows results and a discussion of the proposed method. The conclusion and future work of the proposed research is given in Section V.

II. Literature Review

Nadimi-Shahraki et al. [11] developed a feature selection technique of B-MFO for the HER medical dataset. The developed model reduced the algorithm performances and therefore, the present research used a binary moth-flame optimization (B-MFO) for the selection of effective features based on the large medical datasets. The features such as S-shaped, U-shaped, and V-shaped transfer functions were used which converted continuous to binary values. The B-MFO technique was applied to use a U-shaped transfer function for feature selection to improve performance in a large dataset. While considering the other datasets (Pima and Lymphography), the suggested B-MFO achieved less accuracy when compared with existing Binary Particle Swarm Optimization (BPSO) method.

Li et al. [12] utilized an optimization approach for reinforcement learning on the Electronic Health Records (EHR) for the treatment. Reinforcement learning provided an efficient path for providing a decision sequentially. The developed model used reinforcement learning for optimizing the treatment for analyzing the diseases, diabetes, and sepsis, and showed complications. The EHRs data was modeled in an environment that obtained a probability that was used for the RL process. The agents were explored better as the basic model was cooperative for multi-agent reinforcement based on the value decomposition. However, the recommended model was additionally required to be extended through the decomposition model and thus the results obtained were better than the existing benchmark models.

Sousa et al. [13] applied the decision-making technique for big data analysis in healthcare organizations and People management. The decision-making process is based on healthcare on big data analysis to support healthcare decisions and applied some techniques to increase efficiency. The suggested model has the limitation of irrelevant feature selection that degrades the performance of the classification accuracy and showed diversity in terms of performance.

Chelladurai and Pandian [14] developed the blockchain based EHR model for an automation system for healthcare. The model could access the health data from one provider to another which remained a challenge when they accessed the health records. The fragmented model launched with the health models was immutable with the patient log by using the modified Merkle tree data to secure the storage. The health records were updated by exchanging information among distinct providers. Even though, the viewership contracts were developed on peer-to-peer blockchain networks and blockchain using Merkle tree generation and hashing that required an extension to ensure the integrity of the content.

Vidhya and Shanmugalakshmi [15] developed a Modified adaptive neuro-fuzzy inference system (M-ANFIS) to analyze the multi-disease using the Big Data (BD) from health care. The health care domain obtained an influence based on the BD that affects the data sources as they are concerned with healthcare organization as it is famous with the volume, complexity, high dynamism, and heterogeneity. The BD analytical techniques utilize the functions, tools, and platforms for realizing it among distinct domains that were affected by various health organizations. The healthcare applications show possible propitious research directions. The multiple diseases were analyzed by using Modified Adaptive Neuro-Fuzzy Inference System (M-ANFIS). Yet, the increasing of sources like audio, video, image, GPS, and medical sensors are having prioritization and designation for the level of patients at the emergency.

Ahmad et al. [16] developed a hybrid ML model for the prediction of mortality in paralytic ileus patients based on EHR. Various machine learning techniques were used including Support Vector Machine with Radial Basis Function (SVM-RBF) for the classification to find the highest rank order among the extracted features. Yet, the developed model required robust models for improving the accuracy of the model to improve the model's feasibility.

Shi [17] developed a novel hybrid deep learning model architecture for the prediction of acute kidney injury based on the patient's record data that included Ultrasound kidney images. The developed model used Convolutional neural networks (CNN) that has Resnet and VGG was made as a hybrid model. The feature maps were concatenated with both types of models for creating the input. However, the suggested model required a continuous optimized approach using the larger clinical database for the paired datasets.

From the literature works, the major problems with big data analytics are the size of the data sets and the complexity with validating long-term predictions for medical diagnostics and treatment. Both the amount of data used in healthcare organizations and the number of data sources are expanding. Healthcare facilities face issues including inconsistent and inaccuracy in patient data as a result of the high speed and growing size of big data. It also has trouble in organizing the data after extracting and integrating them, and more attention is needed to increase accuracy and reduce errors in clinical judgments and other medical tools Therefore, this paper proposed a Binary Moth-Flame with Whale Optimization technique to deal with such issues and the difficulties of implementing big data analytics for the enhancement of healthcare services.

III. Proposed Method

The proposed approach is developed by accumulating enormous amounts of data related to patient care over time in order to comprehend and predict diseases that demands an aggregated approach. While the structured and unstructured data originating from large data sets are collected from clinical and nonclinical modalities to gain information about the disease states. As a result, this study attempts to assess the value of predictive analytics in the health care system by examining the accuracy and other metrics in the provision of medical care.

The block diagram of the proposed Modified feature selection optimization approach is shown in Fig. 1. The block diagram consists of a health care data block which is undergone the process of pre-processing. The pre-processed data is undergone for the feature selection of the data and obtained the outputs. Initially, the collected data that consisted of laboratory reports, imaging reports, medication reports,

medication, caregiver notes, and mortality for both out and in the hospital, etc. are applied for the Spark Streaming layers. The main aim of using the spark streaming layers is to improve the diagnosis of the disease based on the data streaming. However, the spark streaming model required effective features for the model construction. Therefore, an optimum model was used for selecting the relevant subsets that were required for the model, and thus it helps to process further diagnosing the disease using the data streaming layers. The proposed B-MFO-WOA algorithm is generated by developing an initial solution. The image data is pre-processed and the parameters are optimally selected by using the Whale optimization algorithm. For finding the best solution during the exploration and exploitation phase, the B-MFO algorithm is used for the selection of subset features. The three transfer functions such as S, V, and U are integrated for solving the optimization problem. The S, V and U transfer functions select the best values based on the generation of slopes and the saturation. These integrated functions are required for improving the B-MFO algorithm performances. The transfer functions such as S and V helps to convert the continuous variables to a binary value of 0 and 1. The U-shaped transfer function maps the velocity function that is continuously generated with the probability values and updates the particle positions. The integration of all these three transfer functions into the B-MFO algorithm showed an improvement in the performance of searching it on the binary searching space. The search space values are having the probability value for uploading the particle position. Once finding the best features, they are fed to the LSTM classifier that uses a softmax layer and obtains an output as a final label to predict the named data labels. The score for the label is named based on the weighted average for the probability prediction when the classifier is applied to the data. The probability of tagging is done for the labels as 1 and the weight parameter is either set as 0 or 1 for knowing the relative importance of the classifier compared with others. Thus, the health care diagnosis is performed for the data.

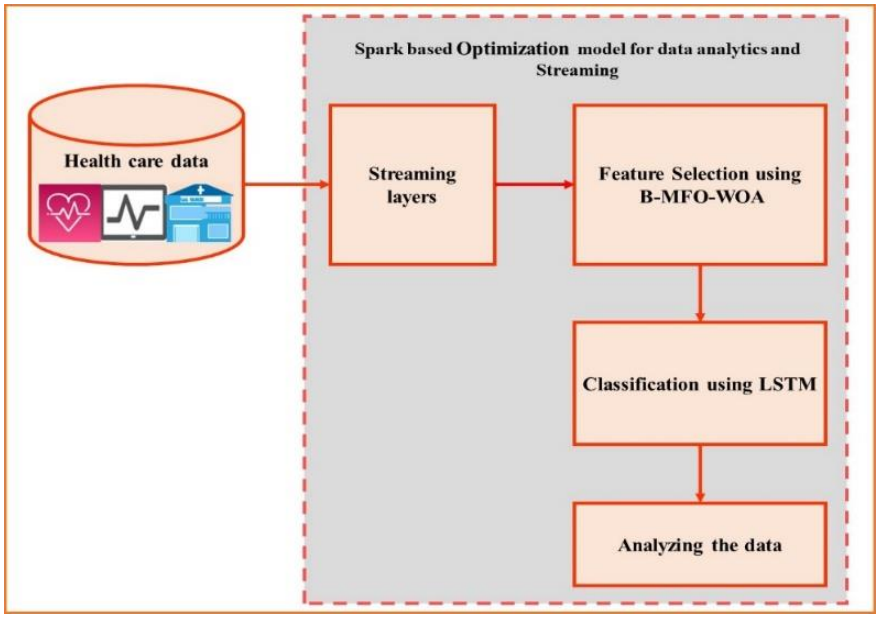

Fig. 1. The B-MFO-WOA and LSTM Model for Medical Data Analysis.

A. Dataset Collection

The health data of the patients are recorded with EHR provides the services for health care in the medical center. The medical centers are registered in detail regarding the patients. The network administrators are registered with the medical centers for participation. The EHR identifies and generates each of the data that is stored in the medical centers. The data is associated total 40,000 number of patients who have stayed in the care units. The units range from the years 2001 and 2012 acquired by Beth Israel Deaconess Medical Center [18]. Initially, the collected data that consisted of laboratory reports, imaging reports, medication reports etc., which are the medical data present in the dataset are applied for the Spark Streaming layers.

B. Feature Selection using B-MFO-WOA

The main aim of using the spark streaming layers is to improve the further diagnosis of the disease based on the data streaming. The stream layers contain polyline, polygon, and point-based features and unlike other feature layers with the services, the data sources have made explicit calls to the data and response for broadcasting the data. In most cases, the features at the irregular intervals are broadcasted and the present research work uses Spark Streaming layers for data streaming to diagnose further.

1) Problem of Feature Selection: The feature process is to select the optimal subset of features to increase the efficiency and relevant features of the model. Accurate data model is constructed and formulated using a feature selection process with a vector having 1 or 0 subsets of features based on the transfer function. It obtains the probability values that change the vector elements that is represented as 0 which can be non-selected and 1 is the selected ones. The feature vector length is the same as the dimensions of the dataset which is used for determining the fitness function which evaluates the subset of

features. This technique reduces the number of features and increases the accuracy of the model. The objective is represented as a fitness function as the CE has shown the error in the classification. N_{sf} and N_{tf} are known as the selected features that are present as a total in the feature dataset. The classification quality significance is given using η and λ (1 − η), as shown in the (1):

$$Fitness = \eta \cdot CE + \lambda \frac{N_{sf}}{N_{tf}} \quad (1)$$

Therefore, the present research uses WOA and B-MFO algorithms for the selection of features.

2) WOA: The WOA uses the exploitation phase for bubble net attacking to perform modeling the bubble net behavior that is having humpback. The two kinds of approaches are designed as follows:

The Shrinking encircling mechanism is performed that achieves the behavior by decreasing the value. The proposed model uses the value to achieve by decreasing a value that represents the fluctuation range when it is decreased. The bubble-net method uses humpback whales to randomly search for prey. Next, in the exploration phase where the prey search is based on the variation in the vector which is used for prey search called exploration. The humpback whales randomly search their positions as per each of their positions. The mutation and evolutionary operations have been included in WOA for formulating and reproducing the behavior of humpback whales that were decided for minimizing the internal parameters and heuristics. This was implemented by the basic WOA version algorithm.

Automatic disease detection is performed using the fitness function for achieving a better classification measure which maximizes the accuracy. The positions for the current solution are updated. The prey is encircled with the phase that performs the process of whale hunting which has started encircled prey position. The whale's best position is found and is considered to be the finest whale. The best whale is towards the other whale which moves once the position is updated. The best solution is determined based on the distances among y^{th} whale where the prey shows the best solution. The distance among the y^{th} whale and the prey calculate the best solution which is ranging between [-1, 1].

3) B-MFO: For finding the best solution during the exploration and exploitation phase, the B-MFO algorithm is used for the optimization of subset features. The three transfer functions such as S, V, and U are integrated for solving the optimization problem. The transfer function of V and S shaped techniques are used in the present research work to convert the MFO function into a binary function. These transfer function names were adapted with the multiple alterable parameters which solved the problem of feature selection. Each of the categories has four versions for transferring the functions and twelve versions had introduced 3 categories for the transfer functions. The datasets such as heart disease, cancer, diabetes, stroke, and arthritis are evaluated for the frequent communities. Additionally, the B-MFO is compared with the best results which are known for its binary metaheuristic optimization approach.

4) B-MFO Variants: The transfer function of S-shaped in B-MFO: The S-shaped or sigmoid function is the transfer function used is named as S_2[100]. The model is introduced originally to develop the binary PSO (BPSO).

$$TF_s(v_i^d(t+1) = \frac{1}{\left(1+exp^{-v_i^d(t)}\right)} \quad (2)$$

From (2), $v_i^d(t)$ is the i^{th} search agent's that is operating at a velocity having the dimension d at the t^{th} iteration. The TFs are converting the probability value of velocity to its next position represented as $x_i^d(t+1)$. The expression is obtained based on the velocity probability value. Here, r is a random value which is ranging from 0 and 1 in (3).

$$x_i^d(t+1) = \begin{cases} 0 & if\ r < TF_s\ (v_i^d(t+1)) \\ 1 & If\ r \geq TF_s(v_i^d(t+1)) \end{cases} \quad (3)$$

From the above expressions, the positions of the search agents are computed based on the current and previous positions. The binary metaheuristic algorithm called BPSO and BGSA is used for transferring the functions. It is used for calculating the probability value that can change the position. The applied transfer function updates the position for each search agent that calculates the probability value. Each of the variants is S-shaped transfer function that showed a slope having S-transfer function. Probability value changes to a positive value as increases in the transfer function. A higher probability function is achieved using S-shaped functions. The S4 provides the lowest value which has affected the position and updates the search agents to find the optimum solution.

C. The Transfer Function of V-Shaped in B-MFO

The V-shaped function is a hyperbolic function that is named with V2 for developing BGSA that has the position to update which is shown in (4).

$$TF_v\left(v_i^d(t+1)\right) = \left|tanh(v_i^d(t))\right| \quad (4)$$

Where t is iteration, d is dimension, velocity of i^{th} search agent is denoted as $v_i^d(t)$. S-shaped function differs from the V-shaped function. The new rules of the updated function are given in (5).

$$x_i^d(t+1) = \begin{cases} -(x_i^d(t) & If\ r < TF_v(v_i^d(t+1)) \\ x_i^d(t) & If\ \ r \geq TF_v(v_i^d(t+1) \end{cases} \quad (5)$$

From the above Equations $x_i^d(t)$ has i^{th} search agent having the position and $-x_i^d(t)$ is the complement value of $x_i^d(t)$. The random value range of 0 and 1 is denoted as r. In case the velocity is low then TF_v encourages the search agents for staying in the positions else the velocity is high. Also, $x_i^d(t)$ has three variants having V-shaped function which is represented as V_1 , V_3 and V_4 are introduced. The higher probability is provided by V_1 than V_2, V_3, and V_4 for the same velocity that affects search agent update and finds an optimum solution.

D. The Transfer Function of U-Shaped in B-MFP

The α and β are two control parameters in the Transfer function of U-shaped that define the slope of U-shaped function width. The U-shaped function is given in (6) and (7).

$$TF_u\left(v_i^d(t+1)\right) = \alpha\left|\left(v_i^d(t)\right)^\beta\right| \quad (6)$$

$\alpha = 1, \beta = 1.5,2,3,4$

$$x_i^d(t+1) = \begin{cases} -(x_i^d(t) & If\ r < TF_u(v_i^d(t+1)) \\ x_i^d(t) & If\ \ r \geq TF_u(v_i^d(t+1) \end{cases} \quad (7)$$

Where t is iteration, d is a dimension, velocity of i^{th} search agent is $v_i^d(t)$ and r of the uniform random number is in the range of 0 and 1. The transfer function of the U-shaped is applied with two conditions. The lower and upper bounds are limited by 1 in (8) and (9).

$$lim_{v_i \to \infty}\, U(v_i^d(t) = 1 \quad (8)$$

$$lim_{v_i \to -\infty}\, U(v_i^d(t) = 1 \quad (9)$$

The variants obtained from the U-Shaped Transfer function is named as U_1, U_2, U_3 and U_4 which were used with the control parameters. The initial iterations were explored for the whole search space of important step that was explored with exploitation with the final iterations. The exploitation step is important for finding a better solution.

The random value for the search space is generated as indicated in (10):

$$E(u) = (e_1, e_2, \ldots e_h) \quad (10)$$

From the above equation (7), E is known as the whales' original population, the interconnected layers with the numbers are represented as h for the process of optimization.

E. B-MFO-WOA

Begin
The population of whales are initialized
Each search agent's fitness function is evaluated
X_{best} = searches for the best search agent
while ($t < maximum\ number\ of\ iterations$)
for each of the search agents:
The positions are Updated as $\alpha, A, C, l\ and\ p$
$if(p < 0.5)$:
$if(|A| < 1)$:
The current agent is updated
else:
The random population of moth is initialized and the objective function is calculated
The set of flames from the same moth is created
The positions of the moths are updated
The flame size has to be changed
End
Return with the best solution

F. Classification

The obtained features are now fed for the LSTM to exhibit the performances. Apache Spark was deployed in the cloud as it focused to apply on the Deep learning LSTM model. The prediction is performed for the higher rate of diagnosis which determines the global best function. The hyperparameters are randomly selected and are passed for the LSTM training. At each iteration, the calculation of parameters is performed. The iteration is stopped when the fitness function is matched The output from the LSTM cell is denoted as h_t, c_t is the memory cell value, LSTM cell output from the previous moment is represented as h_{t-1}. The input data for the LSTM cell is represented as x_t operating at the time t. The process of calculating the LSTM unit is explained in the following steps:

LSTM unit calculation process is explained in steps.

$\widetilde{c_t}$ is known as the candidate memory which is calculated and the bias is represented as b_c. The weight matrix is represented as W_c which is as shown in (11).

$$\widetilde{c_t} = tanh(W_c.[h_{t-1}, x_t] + b_c) \quad (11)$$

The input gate i_t is the current input data that updates the memory cell's state value and controls the input gate. The bias is represented as b_i and the weight matrix is represented as W_i, The sigmoid function is denoted as σ which is shown in (12).

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \quad (12)$$

f_t is the forget gate which calculates the memory state value obtained based on the historic data that updates and controls the forget gate. The bias is represented as b_f and the weight matrix is represented as W_f, as given in (13).

$$f_t = \sigma(W_i.[h_{t-1}, x_t] + b_f) \quad (13)$$

The current moment memory cell c_t is evaluated and the value for the last LSTM unit is denoted as c_{t-1}, as given in (14).

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c_t} \quad (14)$$

Where '*' denotes the dot product. Input and forget gate control updates the memory cell based on the state value for the last cell and the candidate value.

Where, o_t is known as the output gate which calculates the memory cell state value as the output is controlled by the output gate as shown in (15).. The bias b_0 and the weight matrix is denoted as W_0.

$$o_t = \sigma(W_0.[h_{t-1}, x_t] + b_0) \quad (15)$$

The output h_t for the LSTM cell is calculated as shown in (16).

$$h_t = o_t * tanh(c_t) \quad (16)$$

LSTM model update, reset, read and keep long time information easily based on memory cell and control gates. The classifier uses a softmax layer for obtaining an output at the final labels for detecting the named data labels. The score for the label is named based on the weighted average for the probability prediction of disease when the classifier is applied

to the data. The probability of tagging is done for the labels as 1 and the weight parameter is either set as 0 for non-diseased labels or 1 for the diseased labels that knew the relative importance of the classifier when compared with others.

IV. Results and Discussion

The proposed model is operating with Python API libraries that are interfaced with the Local Server running in Windows PC 10 pro, 16 GB NVIDIA Geo-force GPU with i9 CPU operating at 2.5GHz.

A. Performance Metrics and Evaluation

The proposed method results are evaluated in terms of performance metrics for the optimized LSTM based model with the Whale Optimizing approach. Indication must be used to guide the use of diagnostic tests in health care settings. Unfortunately, many order tests without taking into account the supporting data. Therefore, in this research, Sensitivity and specificity are crucial test accuracy indicators that enable medical professionals to decide whether a diagnostic tool is appropriate. Healthcare professionals should use diagnostic tests with the appropriate level of assurance in the accuracy, specificity, sensitivity, Area Under Curve (AUC) and Receiver Operating Characteristics (ROC). The mathematical expression for the performances is given in (17–21):

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100 \quad (17)$$

$$Sensitivity\ or\ Recall = \frac{TP}{TP+FN} \times 100 \quad (18)$$

$$Specificity = \frac{TN}{TN+FP} \times 100 \quad (19)$$

$$AUC = y = f(x) \times 100 \quad (20)$$

where $x = a$ and $y = b$

$$ROC = TPR = \frac{TP}{TP+FN} \times 100 \quad (21)$$

From the above Eq. (17-21), TP is known as True Positive, TN is True Negative, TP is True Positive, TN is True Negative. Table I shows the analysis of different algorithms having data size with 5 GB with feature selection.

TABLE I. The Different Algorithms Having Data Size with 5 GB with Feature Selection

Algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	ROC (%)
CNN	86	85.25	83.11	84.1	82.23
DNN	90	86.24	84.45	87.45	83.12
LSTM	92	90.8	89.21	91.21	88.24
LSTM based Co-learning model	95.4	92.24	91.21	93.45	91.00
Proposed method (B-MFO-WOA)	99.21	95.45	93.48	95.78	96.87

The health analysis was performed on the patients classified as healthy and unhealthy patients. The results inferred that the percentage for each of the patients is analyzed with respect to the healthy patients with the highest percentage. The existing algorithms used for results analysis are Convolution Neural Network (CNN), Deep Neural Network (DNN), Long Short Term Memory (LSTM), LSTM based Co-learning model. The large training data was needed but failed to encode the position and orientation of the object by using the CNN model. The DNN model was hardware-dependent and showed unexplained behavior in the network when the data were fed. Similarly, the LSTMs showed complexity in the model due to large data set training that needed memory to train. Thus, the existing models showed lower values of performance when compared to the proposed method. Table II shows the evaluation of different clusters that are obtained for different diseases. The present research depicts the number of patients with a particular disease carried out with distinct clusters, patients with various diseases. The existing models such as DNN, CNN, LSTM, LSTM based Co-learning model were used for the evaluation of results in terms of accuracy, sensitivity, specificity, AUC, and ROC. The CNN model obtained 84% of accuracy CNN, 82.95% of sensitivity 81.11% of specificity, AUC of 79.25%, and ROC of 83.02%. The DNN model obtained 87% of accuracy, a sensitivity of 84.24%, specificity of 82.45 %, AUC of 80.65%, and ROC of 84.45%. Also, LSTM model obtained 90% of accuracy, 88.8% of sensitivity, 86.21% of specificity, 86.24% of AUC, and ROC of 88.21%. The existing LSTM based Co-learning model obtained 93.4%, 91.04% of sensitivity, 90.11 % of specificity, AUC of 91.78%, and ROC of 90.99%. The proposed method obtained better accuracy of 95.24%, sensitivity of 92.45%, specificity of 90.4%, AUC of 93.45%, 93.25% of ROC. Fig. 2 illustrates the results obtained for the proposed method with feature selection algorithms.

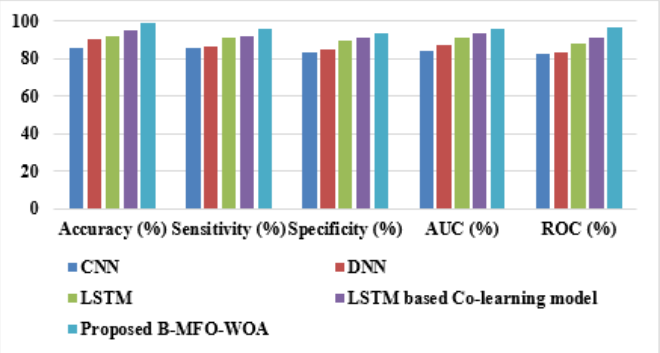

Fig. 2. Results Obtained for the Proposed Method with Feature Selection Algorithm.

TABLE II. Different Algorithms Evaluating Performances for Distinct Data Size with 5 GB Without Feature Selection

Algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	ROC (%)
CNN	84	82.95	81.11	79.25	83.02
DNN	87	84.24	82.45	80.65	84.45
LSTM	90	88.8	86.21	86.24	88.21
LSTM based Co-learning model	93.4	91.04	90.11	91.78	90.99
Proposed method	95.24	92.45	90.4	93.45	93.25

The existing algorithms used for results analysis are Convolution Neural Network (CNN), Deep Neural Network (DNN), Long Short-Term Memory (LSTM), LSTM based Co-learning model. Table III shows the evaluation of performance metrics for different algorithms having greater than 5 GB data size with feature selection algorithm. From the Table III, it clearly shows that the feature selection algorithm with more than 5 GB data size obtained Accuracy of 88%, Sensitivity of 86.25%, specificity of 90.12%, ROC of 82.02%, AUC of 79.25%. The DNN model obtained 92% of accuracy, 89.24% of sensitivity, 91.45% of specificity, AUC of 80.65%, 83.45% of ROC, LSTM of 93%, 92.8% of Sensitivity, AUC of 86.24%, ROC of 87.21. The LSTM based Co-learning model obtained accuracy of 98.6%, sensitivity of 98.21%, specificity of 97.21%, AUC of 91.75%, and ROC of 90.99%. The proposed method obtained 99.32% of accuracy, sensitivity of 98.98%, specificity of 98.78%, AUC of 95%, ROC of 92.56%.

TABLE III. EVALUATION OF PERFORMANCE METRICS FOR DIFFERENT ALGORITHMS HAVING GREATER THAN 5 GB DATA SIZE WITH FEATURE SELECTION ALGORITHM

Algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	ROC (%)
CNN	88	86.25	90.12	79.25	82.02
DNN	92	89.24	91.45	80.65	83.45
LSTM	93	92.8	92.8	86.24	87.21
LSTM based Co-learning model	98.6	98.21	97.21	91.78	90.99
Proposed method	99.32	98.98	98.78	95	92.56

Table III show the results obtained by the proposed method that is evaluated using existing algorithms, such as CNN, DNN, LSTM, and LSTM based co-learning model when the data was greater without feature selection and feature selection algorithm. The existing LSTM based Co-learning model obtained 98.6% of accuracy, sensitivity of 98.21%, specificity of 97.21%, AUC of 91.78 %, and ROC of 90.99%. Similarly, the proposed method obtained 99.32 % of accuracy, 98.98 % of Sensitivity, specificity of 98.78%, AUC of 95%, and ROC of 92.56%.

TABLE IV. PERFORMANCE METRICS OBTAINED BY DISTINCT ALGORITHM HAVING DATA SIZE GREATER THAN 5 GB WITHOUT FEATURE SELECTION ALGORITHM

Algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	ROC (%)
CNN	87	86.14	85.45	85.42	85.24
DNN	91	89.16	88.98	84.57	88.45
LSTM	92	91.23	93.11	86.21	92.21
LSTM based Co-learning model	97.21	94.47	95.211	93.7	96.09
Proposed method	97.45	95.02	96.78	96.87	95.4

Table IV shows the results obtained for different algorithms that are having 5GB greater size without feature selection algorithm. The accuracy of the proposed model without feature selection was obtained as 97.45%, sensitivity of 95.02%, specificity of 96.78%, AUC of 96.87%, and ROC of 95.4%. Fig. 3 illustrates the comparison of results for the proposed method without using the feature selection algorithm.

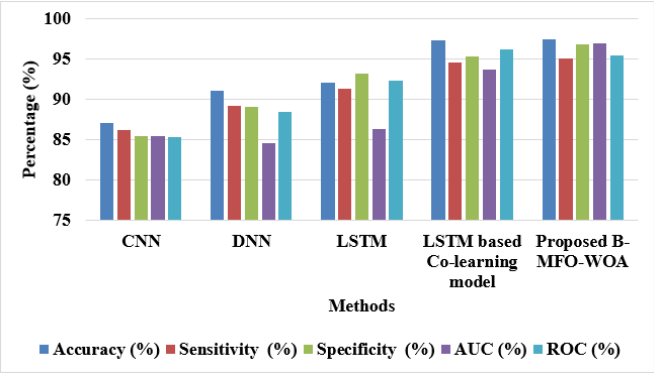

Fig. 3. Comparison of Results for the Proposed Method without using the Feature Selection Algorithm.

B. Comparative Analysis

Table V shows the comparative analysis of the proposed method and existing models evaluated in terms of accuracy, specificity, and sensitivity. The existing B-MFO model needed U-shaped transfer functions for the selection of effective features to overcome the problem of large scale optimization that resulted in Accuracy of 92.43%, Specificity of 92.43%, and sensitivity of 83.51%. Similarly, the Modified adaptive neuro-fuzzy inference system obtained an accuracy of 95.91%, specificity of 98%, and sensitivity of 97.9%. The classification was performed using an SVM-RBF that applied data to the outputs where the classification limited the results for a few of the data values thus obtained an accuracy of 81.30%, sensitivity of 35.59%, and specificity of 91%. The developed CNN model was overloaded with the historic data as it was continuous for data streaming and it was challenging to store, process, and analyse obtained an accuracy of 90% and Sensitivity of 90%. However, the model was required to be extended through the value decomposition model and thus the results obtained were better than the existing benchmark models.

TABLE V. COMPARATIVE ANALYSIS

Method	Dataset	Accuracy (%)	Specificity (%)	Sensitivity (%)
B-MFO [11]	EHR from Beth Israel Deaconess Medical Centre	92.43	96.85	83.51
Modified adaptive neuro-fuzzy inference system [12]		95.91	98	97.9
SVM-RBF [16]		81.30	35.59	91
Convolutional neural networks [17]		90	90	-
Proposed method		97.45	96.78	95.02

V. CONCLUSION

The proposed method showed better performances when operated with B-MFO for the selection of effective features

which were evaluated for large and small medical datasets. The three transfer functions such as S, V, and U-shaped transfer functions are used for the conversion of MFO from the values of continuous to binary values. The WOA is used as an appropriate algorithm to select constrained and unconstrained problems for overcoming the practical applications based on the structural reformation. The combination of the B-MFO-WOA is iteratively executed and is compared with various solutions till an optimum or satisfactory solution is found. The WOA showed better performances for LSTM that suited well for the process of classification to predict the time series. The given time is lagged for an unknown duration of the model as it is based on a deep learning model. The developed model co-learns the best soft labels and deep neural networks based on the training procedure. The Whale optimization approach has the ability for improving the population quality and improves the speed of the algorithm for disease presence prediction. The simulation results showed that the proposed method achieved the objectives by attaining 97.45% accuracy which is better when compared to the existing Modified adaptive neuro-fuzzy inference system of 95.91% of accuracy and B-MFO attained the accuracy of 92.43 %. However, the model showed the complexity in the model due to more features included in a given predictive model which will be analyzed in the future work.

REFERENCES

[1] A. K. Gárate-Escamilla, A. H. E. Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," Inf. Med. Unlocked, vol. 19, p. 100330, January 2020.

[2] J. E. Dalton, M. B. Rothberg, N. V. Dawson, N. I. Krieger, D. A. Zidar, and A. T. Perzynski, "Failure of Traditional Risk Factors to Adequately Predict Cardiovascular Events in Older Populations," Journal of the American Geriatrics Society, vol. 68, no. 4, pp. 754–761, April 2020.

[3] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, "MIFH: A machine intelligence framework for heart disease diagnosis," IEEE Access, vol. 8, pp. 14659–14674, December 2019.

[4] G. Magesh and P. Swarnalatha, "Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction," Evolutionary Intelligence, vol. 14, no. 2, pp. 583–593, June 2021.

[5] D. Swain, P. Ballal, V. Dolase, B. Dash, and Santhappan, "An Efficient Heart Disease Prediction System Using Machine Learning," in Proc. of ICMLIP 2019, Machine Learning and Information Processing, Advances in Intelligent Systems and Computing, vol. 1101, D. Swain, P. Pattnaik, and P. Gupta, Eds. Singapore: Springer, 2020, pp. 39–50.

[6] S. Sajeev, A. Maeder, S. Champion, A. Beleigoli, C. Ton, X. Kong, and M. Shu, "Deep Learning to Improve Heart Disease Risk Prediction," in Proc. of Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting: 1st International Workshop, MLMECH 2019, H. Liao, S. Balocco, G. Wang et al., Eds. Heidelberg: Springer, 2019, pp. 96–103.

[7] R. Venkatesh, C. Balasubramanian, and M. Kaliappan, "Development of Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique," Journal of Medical Systems, vol. 43, no. 8, p. 272, July 2019.

[8] R. T. Selvi and I. Muthulakshmi, "An optimal artificial neural network based big data application for heart disease diagnosis and classification model," Journal of Ambient Intelligence and Humanized Computing, vol. 12, pp. 6129–6139, 2021.

[9] H. Das, B. Naik, H. S. Behera, S. Jaiswal, P. Mahato, and M. Rout, "Biomedical data analysis using neuro-fuzzy model with post-feature reduction," Journal of King Saud University-Computer and Information Sciences, 2020.

[10] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," IEEE Access, vol. 5, pp. 8869–8879, 2017.

[11] M. H. Nadimi-Shahraki, M. Banaie-Dezfouli, H. Zamani, S. Taghian, and S. Mirjalili, "B-MFO: a binary moth-flame optimization for feature selection from medical datasets," Computers, vol. 10, no. 11, p. 136, October 2021.

[12] T. Li, Z. Wang, W. Lu, Q. Zhang, and D. Li, "Electronic health records based reinforcement learning for treatment optimizing," Inf. Syst., vol. 104, p. 101878, February 2022.

[13] M. J. Sousa, A. M. Pesqueira, C. Lemos, M. Sousa, and Á. Rocha, "Decision-making based on big data analytics for people management in healthcare organizations," Journal of Medical Systems, vol. 43, no. 9, pp. 1–10, 2019.

[14] U. Chelladurai and S. Pandian, "A novel blockchain based electronic health record automation system for healthcare," J. Ambient Intell. Hum. Comput., vol. 13, no. 1, pp. 693–703, 2022.

[15] K. Vidhya and R. Shanmugalakshmi, "Modified adaptive neuro-fuzzy inference system (M-ANFIS) based multi-disease analysis of healthcare Big Data," The Journal of Supercomputing, vol. 76, no. 11, pp. 8657–8678, 2020.

[16] F. S. Ahmad, L. Ali, H. A. Khattak, T. Hameed, I. Wajahat, S. Kadry, and S. A. C. Bukhari, "A hybrid machine learning framework to predict mortality in paralytic ileus patients using electronic health records (EHRs)," J. Ambient Intell. Hum. Comput., vol. 12, no. 3, pp. 3283–3293, 2021.

[17] S. Shi, "A novel hybrid deep learning architecture for predicting acute kidney injury using patient record data and ultrasound kidney images," Applied Artificial Intelligence, vol. 35, no. 15, pp. 1329–1345, September 2021.

[18] D. Kalra, "Electronic Health Record Standards," Yearbook of Medical Informatics, vol. 15, no. 01, pp. 136–144, 2006.