Received: May 10, 2022. Revised: June 16, 2022.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Medical Big Data Analysis using LSTM based Co-Learning Model with Whale Optimization Approach

Saka Uma Maheswara Rao[1]* **K Venkata Rao[1]** **Prasad Reddy PVGD[1]**

[1]*Department of Computer Science and Systems Engineering, Andhra University College of Engineering (A), Andhra University, Visakhapatnam, India*
* Corresponding author's Email: saka.mahi@gmail.com

Abstract: The medical and health service has become advanced and the smart health care platform has made the diagnosis more robust for the treatment. 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, the existing approaches failed to deploy the learning model to handle the heterogeneous medical data. The present research work has used the machine learning algorithm effectively for the chronic disease prediction such as heart disease, cancer, diabetes, stroke, and arthritis for the frequent communities and the medical data are available from the disparate sources including a wide variety of information has made difficult to analyse and retrieve. 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. The proposed model was tested for its effectiveness and was validated with various benchmark data which was collected from distinct domains. The present research work utilizes Spark Streaming layers for data streaming to diagnose further based on Long Short Term Memory (LSTM) Co-learning with whale optimization approach is from the heterogeneous medical data. The results obtained by the proposed method analysed the disease abnormality better. The existing model obtained SVM-RBF of 81.30 %, k-fold cross-validation and hyperparameters tuning with Random Forest of 94.9%, CNN of 90%, and the proposed LSTM based Co-learning model with Whale optimization approach obtained 98.6 % which is better compared to the existing models.

Keywords: Information technology, Medical and health service, Long short term memory based deep co-learning with whale optimization approach, Spark streaming layers.

1. Introduction

Information Technology is playing an important role in the field of medical and health services. The health care industries make use of big data as it is the most promising field that analyzes the changes, manages, analyzes, and leverages the data [1]. In the field of medicine, healthcare is complex due to large data considered from the previous software and hardware as it is challenging for analyzing. The quality controls, models, integrates and interprets the data by covering the big data. A large amount of data is present that provides comprehensive knowledge for discovering and analyzing the big data applications [2, 3]. The medical imaging data is important for findings its use and treats to diagnose the medical conditions to view the human body. Various existing technologies were used for analyzing the image as it designates the techniques to produce the image from its internal medical image part [4].

Using various technologies for analyzing the medical image designates the techniques for producing the image from the body's internal part for medical imaging. The disease is detected based on a function of the organ and medical images anatomy which provides significant information [5]. Fast development is required for providing quantity health by an organization to quantify the patients. Health care is provided with many areas for screening and diagnosing, to improve the accuracy of medical

images [6, 7]. The healthcare data and the medical data are complex because the large data taken from the software and hardware was difficult for analyzing. The availability of large health care datasets and progression in Machine Learning (ML) techniques were equipped to diagnose health issues [8]. The present research work was developed with a real-time status prediction system to build the open-source processing of the big data engine. Apache Spark was deployed in the cloud as it focused to apply on the ML models [9]. The attributes were applied for ML models that predict the user's health status which is messaged directly to the user [10]. The contributions of the research are given as follows:

- To develop LSTM based Co-Learning Model with a Whale optimization approach that showed better performances for LSTM that was suited for the process of classification. The process was used in predicting the time series that showed lags for the unknown duration of the model as it is based on a deep learning model.
- To co-learn the model for finding the best soft labels that intend to learn the deep neural networks based on the end-to-end training procedure.
- The Whale optimization approach can improve the population quality and the speed of the algorithm for disease prediction.

The structure of the paper is given as follows: Section 2 explains the existing models involved in disease prediction and section 3 is the proposed method that explains the proposed LSTM-based Co-learning model with Whale optimization approach. Section 4 explains the results. The conclusion and future work of this research work is given in section 5.

2. Literature review

Lakshmana Rao Namamula and Daniel Chaytor [11] developed an Effective ensemble learning model to analyse large-scale medical data. Because of the enormous amount of storage, and computing resources, appropriate administration is demanded on big data resources. The big data resources effectively need an analytical tool. The robust algorithms and tools were used for obtaining a large dataset for analyzing the ML domain. The accuracy has improved as the present research explains the ensemble learning strategy where the ML algorithms such as Edge Detection Instance preference (EDIP) and Extreme Gradient Boosting (XGboost) were combined aggregated the precision and the voting technique was utilized. The classification of data is performed that obtained outputs but failed to help the users as it limited the values.

Wei Tan [12] developed Multimodal medical image fusion algorithm in the era of big data. The developed multimodal medical imaging technique was proposed for overcoming medical diagnostic problems. The developed model was working based on the boundary measured pulse which was coupled with a neural network applied for various applications. The energy attribute fusion strategy was applied to the non-subsampled shearlet transform model. The developed model validated the improvement in performances that controlled distinct diseases like Alzheimer's disease, glioma, and metastatic bronchogenic carcinoma images which contained more images with 100 pairs. However, the deep learning model was used widely and still needed to focus on multimodal medical image fusion.

Joo-Chang Kim and Kyungyong Chung [13] developed a multi-modal stacked denoising autoencoder to handle the data which was missing under healthcare big data. The developed model uses a stacked denoising auto-encoder for the estimation of missing data. The auto-encoders were the neural networks that generates the output value similar to that of the input value. The present research used a stacked denoising auto-encoder which was applied and configured under the settings for multi-modal operations. The set of data was based on the learning and a label was set with the original data. The auto-encoder input was set and noises were added as an input with random zero numbers. However, the data was missed and that was validated using the multi-modal stacked denoising autoencoder for evaluating healthcare machine learning models.

Vidhya and Shanmugalakshmi [14] developed a Modified adaptive neuro-fuzzy inference system (M-ANFIS) for analyzing the multi-disease healthcare Big Data (BD). 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.

Sivaparthipan [15] performed a statistical assessment healthcare system for diagnosing diabetics based on big data. The present research develops a model for performing statistical

assessment evaluated based on the Hadoop framework. The results are evaluated in terms of accuracy and f-measure that obtains better values when compared to the existing models. The developed model performed a statistical assessment for the health care system for analyzing the big data for diabetic analysis. However, the accuracy and F-measures were evaluated by the Hadoop framework resulting in higher performances compared to the existing models.

Hager Ahmed [16] performed heart disease identification from the social posts, and patients obtained solutions using machine learning techniques. This improved the results for the proposed method and the optimal machine learning algorithm achieves a higher accuracy for the prediction of heart disease. The feature selection algorithms utilize the univariate feature selection and the relief feature selection algorithm was used for the selection of important features. The machine learning algorithms such as Support Vector Machine, Logistic Regression, Random Forest Classifiers, and Decision tree were used for performing the classification of selected features. Yet, the developed model was overwhelmed with the historic data and there was a continuous flow for data streaming generated health care services was a challenging task to process, store, and analyze the machine learning based models.

Fahad Shabbir Ahmad [17] developed a hybrid ML model for the prediction of mortality in paralytic ileus patients based on EHR. There were various machine learning techniques that 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.

Sophia Shi [18] 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 were made as a hybrid model. The feature maps were concatenated with both type of models for creating the input. However, the developed model required a continuous optimized approach using the larger clinical database for the paired datasets were required.

The existing research concludes that the accurate analysis of medical data was dependent on early disease detection and the value of accuracy was reduced when the medical data quality was incomplete. The LSTM with a Co-learning model with Whale Optimization showed improvement in the performances that predicted the times series better for the time lags to unknown model's duration. The proposed research can handle heterogeneous datasets of chronic disease that include heart disease, cancer, diabetes, stroke, and arthritis for the frequent communities and the medical data. The model was intended to co-learn the soft labels and DNN through the process of end to end training procedure.

3. Proposed method

This section includes the workflow for the proposed method with the implementation steps. The health data constitute information of patients who have undergone medical services in the hospital.

3.1 Dataset

The health data of the patients are recorded with EHR provides the services for health care in the medical centre. The medical centres are registered in detail regarding the patients. The network administrators are registered with the medical centres for participation. The EHR identifies and generates each of the data that is stored in the medical centres. 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.

The database includes demographic information which is having vital signs of measurements residing between (1 data point per hour). This tests the

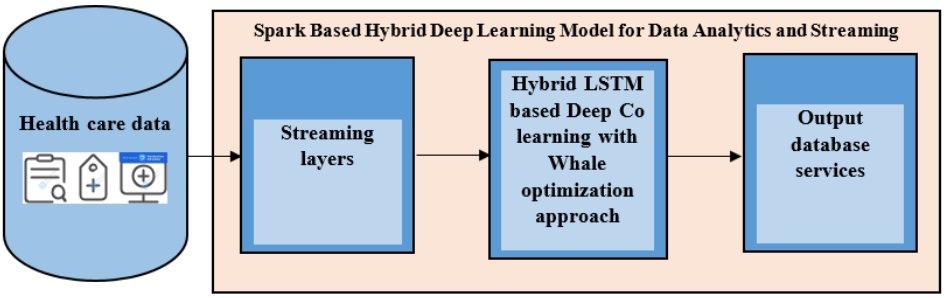

Figure. 1 Block diagram of the proposed method

laboratory results, imaging reports, medication, caregiver notes, and mortality for both out and in the hospital [19]. The block diagram of the proposed model is shown in Fig. 1.

3.2 Streaming layers

The stream layers are known as the feature layers having a stream of services to the data sources. The real-time datasets are having live observations and include the location changes, attributes, or both. The stream layers are having polyline, point, or polygon-like features, and these unlike feature layers are with services for their data source. This has made explicit calls on the data, and actively listens to the stream of data. 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. The data is received from distinct sources which are transformed into minimum batches for obtaining high-speed streaming.

3.3 Long short term memory

The obtained features are now fed for the LSTM to exhibit the performances when the big datasets are handled. The larger datasets were used that increased the memory usage resulting in computational complexity. The motivation for overcoming the drawback is that the proposed model employs the firefly optimization approach with LSTM model obtained hyper parameters. The hyper parameters like some epochs, and hidden layers are needed for optimizing better performances compared to the LSTM model. The prediction is performed for the higher rate of diagnosis which determines the global best function. The expression for the proposed method in terms of the fitness function is mathematically expressed as shown in the below equation for finding the G_{best} value. 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 deep Co-learning model predicts labels for the phrases and the labels are predicted for the other phrases as it cannot detect the classifier. The Architecture Long Short Term Memory (LSTM) model is shown in Fig. 2.

The output from the LSTM cell is denoted as h_t, c_t is the memory cell value, LSTM cell output from

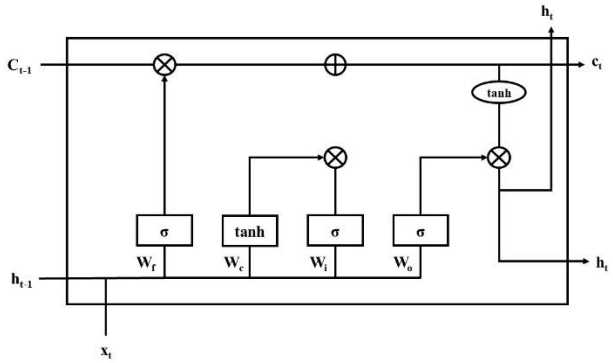

Figure. 2 Long short term memory (LSTM) model architecture

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 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 Eq. (1).

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

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 Eq. (2).

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \quad (2)$$

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 Eq. (3).

$$f_t = \sigma(W_i.[h_{t-1}, x_t] + b_f) \quad (3)$$

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 Eq. (4).

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c_t} \quad (4)$$

Where '*' denotes the dot product. Input and forget gate controls update 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. The bias b_0 and the weight matrix is denoted as W_0.

$$o_t = \sigma(W_0.[h_{t-1}, x_t] + b_0) \quad (5)$$

The output h_t for the LSTM cell is calculated as shown in Eq. (6).

$$h_t = o_t * \tanh(c_t) \quad (6)$$

LSTM model update, reset, read and keep long time information easily based on memory cell and control gates. The LSTM model sharing mechanism of internal parameters controls the output dimensions based on weight matrix dimensions' settings. The deep co-learning algorithm predicts the labels based on the phrases as it was not detected and the class probabilities are detected based on the data. Each of the data is trained for two co-learning models as iteratively trained as follows:

The phrase labels are used for prediction based on the labels from other phrases that failed to detect the class probabilities. The data is trained for each of the co-learning models are fed with the data having two models applied for training. The data are assigned with a score that leverages the prediction probabilities where the co-learning classifier is applied to the data. 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 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. The dataset is used to train LSTM models which is described as M_1 and M_2 for the deep neural network architecture. The models are applied on the corresponding labels which are picked for each of the model. The co-learning models M_1 and M_2 predicts the phrases from each of the labels for few phrases. The Co-learning model is applied finally to identify the entities and at each of the iteration when it receives the highest scores. The LSTM based co-learning model performs 2 functions to obtain the loss subject during training for the target labels. The second type is to perform the training loss once the target reaches the soft labels. The soft label is designed for learning and subjecting it to learning the labels at the output. The existing solution obtained is compared and used for raw soft prediction which results in distillation and leveraging of soft labels.

3.4 Feature selection using whale optimization approach

The proposed algorithm is generated by developing an initial solution. The image data is pre-processed and the parameters of the CNN are optimally selected by using the Whale optimization algorithm. The parameters from the LSTM algorithm are randomly initialized with the number of whales. The random value for the search space generated as indicated in Eq. (7):

$$E(u) = (e_1, e_2, \ldots e_h) \quad (7)$$

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

At the exploitation phase, the bubble net attacking is performed for modelling 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. This value is achieved by decreasing a value that represents the fluctuation range that is decreased.

The whale optimization algorithm is mentioned in the Algorithm-1.

Step1: The population of whales are initialized

Step2: Each search agent fitness function is evaluated

X_{best} = searches for the best search agent

Step3: 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 by eq. (1)

else:

the random agent $X_{rand}$is selected

the current agent is updated by using the Eq. (7)

else:

search agent is updated by Eq. (5)

end-for

If the search agents reach beyond the search space it amends

The fitness function for each of the search agents is calculated

The better solution is updated as X_{best}

$t = t + 1$

end-while

Step4: return X_{best}

The bubble-net method uses humpback whales to randomly search 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].

3.5 Output at the storage layers

The abnormalities are generated in the EHR once after the prediction and the data is diagnosed and stored within the server processed for further monitoring and processing. Co-learning improves the deep learning generalization that softens the labels based on privileged information. There are two objectives developed for learning the model and the soft labels are also designed respectively. The Co-learning model is proposed which achieves the goal alternatively by minimizing the two objectives. The Co-linear algorithm is used for the real world image classification as it has lower quality labels that are iterating with the following steps:

1. The fine-tuning is performed for pre-training the first iteration for embedding the function.
2. The selection of reference embeddings.
3. The dense layer is used for learning the model.
4. The multiple runs for the algorithm are performed iteratively which improves the label quality. Likewise, the experts learn from the classification problems and become consistent in labelling the task.

The LSTM model is a learning model that is used for various applications for memory and works with huge databases. The proposed research work uses LSTM with a Whale optimizer that consists of 3 distinct blocks such as input gate (I.G), forget gate (F.G), cell input (C.I) and output gate (O.G). Generally, LSTM is a memory-based neural network that remembers the values after every iteration.

Table 1. Parameter settings for LSTM, CNN, and DNN

Model	Specifications	Total number
LSTM	Hidden layers	32
	Learning Rate	0.0012
	Maximum length	500
	Batch size	128
CNN	Convolutional kernel (number)	128,128,128
	Learning Rate	0.001
	Mini Batch	64
	Drop out	0.5
DNN	Hidden layer	3
	Hidden nodes	3
	Drop out	0.5
	Learning rate	0.0011

The LSTM model was trained using the following parameters such as learning rate of 0.0012, and the decay rate was 0.9. During the training process, the data set was divided into several batches, sized at 128, to speed up the training rate. The number of hidden layers were considered to be 32 having maximum length of 500. The number of convolutional kernel was considered to be as 128,128,128, learning rate of 0.001, the size of mini-batch was 64, dropout of 0.5 shown in Table 1. Among them, the hyper-parameters in the experiments are selected by means of cross-validation, whose referential evaluation index is accuracy.

The DNN models are used in the research that is having 3 hidden layers with 3 hidden nodes. We used the method of early stopping and dropout of 0.5 in the model to avoid the overfitting. When these test values were attributed to the trained DNN with learning rate of 0.0011 have been observed to be close to the experimental values. The Table 1 shows the parameter settings for LSTM based co-learning model, existing DNN and CNN models.

4. 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.

4.1 Performance metrics and evaluation

The performances for the proposed method results are evaluated in terms of performances for the optimized LSTM based model with the Whale Optimizing approach. The mathematical expression for the performances is given in Eqs. (8) to (12):

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100 \quad (8)$$

$$Sensitivity\ or\ Recall = \frac{TP}{TP+FN} \times 100 \quad (9)$$

$$Specificity = \frac{TN}{TN+FP} \times 100 \quad (10)$$

$$AUC = y = f(x) \times 100 \quad (11)$$
$$where\ x = a\ and\ y = b$$

$$ROC = TPR = \frac{TP}{TP+FN} \times 100 \quad (12)$$

From the above Eqs. (8) to (12), TP is known as True Positive, TN is True Negative, TP is True Positive, TN is True Negative.

4.2 Quantitative analysis

Table 2 has different kinds of 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 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 analysed with respect to the healthy patients with the highest percentage of healthy patients. 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 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 performances when compared to the proposed method. The value of accuracy for the proposed LSTM based co-learning model for the data size with 5GB memory is obtained as 93.4 %, sensitivity is obtained as 91.24%, and specificity is obtained as 90.21%.

Table 3 shows the various performances for distinct algorithms such as CNN, DNN, LSTM, and LSTM based Co-learning model for the data size which is greater than 5 GB. The accuracy percentage for the proposed LSTM based Co-learning model is 92.24%, sensitivity is 90.04%, and specificity is 89.11%. Whereas, the existing models obtained accuracy for CNN as 83%, DNN as 86%, and LSTM as 89%. The value of sensitivity is obtained as 82.25%, 83.24%, and 87.8% for the CNN, DNN, and LSTM models. The specificity value is obtained as 80.11% for CNN, DNN as 81.45%, LSTM as 85.21%.

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 value of accuracy for the proposed model for the data size with 5GB memory is obtained as 92.24 %, sensitivity is obtained as 90.04 %, and specificity is obtained as 89.11 %.

Table 4 and 5 shows the various performances for distinct algorithms such as CNN, DNN, LSTM, and LSTM based Co-learning model for the data size which is greater than 5 GB for with and without feature selection algorithm. The accuracy percentage

Table 2. The different algorithms having data size with 5 GB with feature selection

Algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	ROC (%)
CNN	84	83.25	82.11	83.1	81.23
DNN	88	84.24	83.45	86.45	82.12
LSTM	91	89.8	88.21	90.21	87.24
LSTM based Co-learning model	93.4	91.24	90.21	92.45	90.00

Table 3. Different algorithms evaluating performances for distinct data size with 5 GB without feature selection

Algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	ROC (%)
CNN	83	82.25	80.11	78.25	81.02
DNN	86	83.24	81.45	79.65	82.45
LSTM	89	87.8	85.21	85.24	86.21
LSTM based Co-learning model	92.24	90.04	89.11	90.78	89.99

Table 4. 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

Table 5. Performance metrics obtained by distinct algorithm having data size greater than 5 GB without feature selection algorithm

Algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	ROC (%)
CNN	86	85.14	84.45	83.42	84.24
DNN	90	88.16	87.98	83.57	87.45
LSTM	91	90.23	92.11	85.21	91.21
LSTM based Co-learning model	96.21	94.47	95.211	92.7	95.09

for the proposed LSTM based Co-learning model is 92.24%, sensitivity is 90.04%, and specificity is 89.11%. Whereas, the existing models obtained accuracy for CNN as 83%, DNN as 86%, LSTM as 89%. The value of sensitivity is obtained as 82.25%, 83.24%, and 87.8% for the CNN, DNN, and LSTM model. The specificity value is obtained as 80.11% for CNN, DNN as 81.45%, and LSTM as 85. 21%. As the data size is increasing the value of performance measures decreases the reason is that the data is not distributed properly which biases the CNN to be sensitive for particular class instances. The results obtained in terms of accuracy for with feature selection and without feature for 5db and < 5db are shown in Fig. 3 and 4 and also a graph of the AUC and ROC curve is shown in Fig. 5.

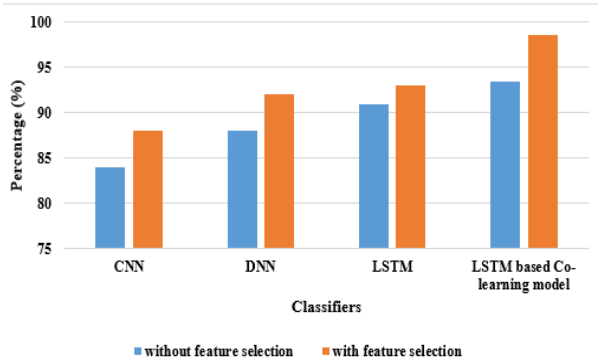

Figure. 3 Results obtained in terms of accuracy for with feature selection and without feature selection (5dB)

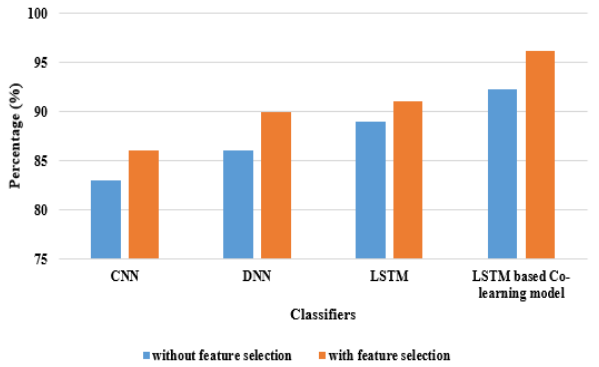

Figure. 4 Results obtained in terms of accuracy for with feature selection and without feature selection (<5dB)

4.3 Comparative analysis

Table 6 shows the comparative analysis of the existing and the proposed model. The comparative analysis performed based on the EHR taken from Beth Israel Deaconess Medical Centre. In [13], the deep learning model was widely required to be focused on multimodal medical image fusion. The data were missed and required to evaluate using multi modal stacked denoising approach for medical image fusion thus obtained an accuracy of 94.21%. In [17], 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%. In [18], the developed CNN model was overloaded with thehistoric data as it was continuous for data streaming it was challenging for storing, processing, and analyzing thus obtained an accuracy of 90 % and Sensitivity of 90%. However, the proposed LSTM based co-learning model with the Whale optimization approach obtained better values of accuracy of

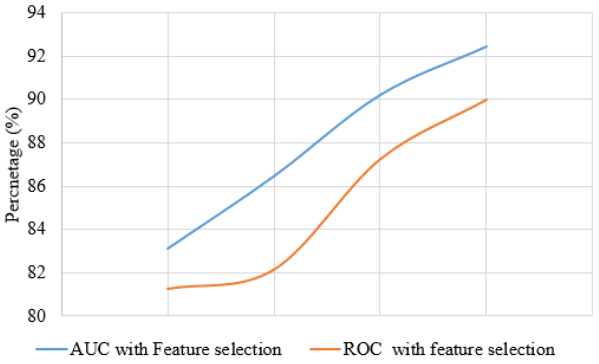

Figure. 5 AUC and ROC curve

Table 6. Comparative analysis

Method	Dataset	Accuracy (%)	Sensitivity (%)	Specificity (%)
k-fold cross-validation and hyper parameters tuning with Random Forest [13]	EHR from Beth Israel Deaconess Medical Centre	94.21	-	-
SVM-RBF [17]		81.30	35.59	91
Convolutional neural networks [18]		90	90	-
LSTM based Co-learning model with Whale optimization approach		98.6	98.21	97.21

98.6% and recall of 98.21% that overcome the problem of data complexities for heterogeneous data.

The existing Random Forest for the classification showed quite slow in creating predictions once features were trained. Similarly, the SVM-RBF showed overlapping in target classes during classification because the number of training data samples were exceeding. At the next, the deep learning model such as CNN was used showed slower because of training process takes more time. Whereas, the proposed LSTM model took longer time to train but because of the Co-learning model with WOA, reduced the complexity of time because it learns the features in parallel to one another during training and improved the learning rate of the model. The proposed LSTM based co-learn the model determined the best soft labels that intend to learn the deep neural networks based on the end-to-end training procedure. Additionally, the Whale optimization approach improved the population quality and the speed of the algorithm for disease prediction.

5. Conclusion

Big Data analytics plays a vital role in predicting and detecting diseases at an early stage. To make better decisions and predictions, the analysis showed great opportunities for predicting the health status. The developed model has consisted of traffic data which gave rise to difficulty and uncertainty in disease prediction. To overcome this problem, big data is used for ensuring that medical services operate accurately with respect to time and to analyse the patient's performance. The results of the proposed method showed that the proposed method obtained better performances when operated with medical services that required accurate time and diagnosis for analyzing patient's history. The proposed LSTM-co-learning with WOA showed better performances for LSTM as it suits for classification process and predicts 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 results obtained by the proposed method showed better classification results compared to the existing SVM-RBF of 81.30%, k-fold cross-validation and hyperparameters tuning with Random Forest of 94.9%. In the future, the complexity of the features has to be enhanced using an optimal feature selection model.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

The paper conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, have been done by 1st author. The supervision and project administration, have been done by 2nd author.

References

[1] A. K. G. Escamilla, A. H. E. Hassani, and E. Andres "Classification models for heart disease prediction using feature selection and PCA", *Informatics in Medicine Unlocked*, p. 100330, 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.

[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, 2019.

[4] G. Magesh and P. Swarnalatha, "Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction", *Evolutionary Intelligence*, pp. 1-11, 2020.

[5] D. Swain, P. Ballal, V. Dolase, B. Dash, and Santhappan, "An Efficient Heart Disease

Prediction System Using Machine Learning", In: *Proc. of Machine Learning and Information Processing*, pp. 39-50, 2020.

[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*, pp. 96-103, 2015.

[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, 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] L. R. Namamula and D. Chaytor, "Effective ensemble learning approach for large-scale medical data analytics", *International Journal of System Assurance Engineering and Management*, pp. 1-8, 2022.

[12] W. Tan, P. Tiwari, H. M. Pandey, C. Moreira, and A. K. Jaiswal, "Multimodal medical image fusion algorithm in the era of big data", *Neural Computing and Applications*, pp. 1-21.

[13] J. C. Kim and K. Chung, "Multi-modal stacked denoising autoencoder for handling missing data in healthcare big data", *IEEE Access*, Vol. 8, pp. 104933-104943, 2020.

[14] 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.

[15] C. B. Sivaparthipan, N. Karthikeyan, and S. Karthik, "Designing statistical assessment healthcare information system for diabetics analysis using big data", *Multimedia Tools and Applications*, Vol. 79, No. 3, pp. 8431-8444, 2020.

[16] H. Ahmed, E. M. Younis, A. Hendawi, and A. A. Ali, "Heart disease identification from patients' social posts, machine learning solution on Spark", *Future Generation Computer Systems*, Vol. 111, pp. 714-722, 2020.

[17] 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)", *Journal of Ambient Intelligence and Humanized Computing*, Vol. 12, No. 3, pp. 3283-3293, 2021.

[18] S. Shi, "A novel hybrid deep learning architecture for predicting acute kidney injury using patient record data and ultrasound kidney images", *Applied Artificial Intelligence*, pp. 1-17, 2021.

[19] D. Kalra, "Electronic health record standards", *Yearbook of Medical Informatics*, Vol. 15, No. 01, pp. 136-144, 2021.