A novel proposed Adaptive Weight Bi-Directional Long Short-Term Memory (awbi-LSTM) classifier based cerebrovascular stroke risk level prediction models (#91966)

First submission

Guidance from your Editor

Please submit by **5 Jan 2024** for the benefit of the authors (and your token reward) .

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

If this article is published your review will be made public. You can choose whether to sign your review. If uploading a PDF please remove any identifiable information (if you want to remain anonymous).

Files
Download and review all files from the materials page.

5 Figure file(s)
1 Box file(s)
4 Table file(s)
1 Raw data file(s)

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

A novel proposed Adaptive Weight Bi-Directional Long Short-Term Memory (awbi-LSTM) classifier based cerebrovascular stroke risk level prediction models

S Thumilvannan [1], **R Balamanigandan** [Corresp. 2]

[1] Computer Science and Engineering, Saveetha University, Chennai, Tamilnadu, India

[2] Department of Computer Science and Engineering, Saveetha School of Engineering, Chennai, India

Corresponding Author: R Balamanigandan
Email address: balamanigandanr.sse@saveetha.com

Background: To prevent difficulties from developing, people with diabetes need access to healthcare services for the rest of their lives. Large amounts of data are produced by their disease management operations in a variety of areas, from medical to administrative. Even for highly desirable applications like the forecasting of cardiovascular disease, the primary cause of excess mortality in diabetes, difficulties in acquiring and interpreting these data prevent its subsequent use in an institutional context. Stroke care and diagnosis have been improved as a result of the detrimental effects it has on society. Caretakers can develop patient management by effectively mining and storing the patients' medical records according to an increasing synergy among technology and medical diagnosis. Therefore, it is essential to look at the relationships between these risk factors in the records of patients and understand how each one contributes significantly to stroke prediction. **Methods:** This study does a thorough analysis of the numerous stroke risk variables found in electronic medical data.Hence, a novel proposed Adaptive Weight Bi-Directional Long Short-Term Memory (AWBi-LSTM) classifier based stroke risk level prediction model for IoT data is proposed in this paper. Here, to efficiently train the classifier, the missing data are removed by Hybrid Genetic with Kmeans Algorithm (HKGA) and the data are aggregated. Then, to reduce the dataset size, the features are reduced with Independent Component Analysis (ICA). After the correlated features are identified using the T-test-based Uniform Distribution- gradient search rule based elephant herding optimization for cluster analysis (GSRBEHO) (T-test-UD- GSRBEHO). Next, to classify the risk levels accurately, the fuzzy rule-based decisions are created with the T-test-UDEHOA correlated features. The feature values obtained from the fuzzy logic are given to the AWBi-LSTM classifier, which predicts and classifies the risk level of heart disease and diabetes. After the risk level is predicted, the data is securely stored in the database. Here, for secure storage, MD5- Elliptic Curve Cryptography (MD5-ECC) technique is utilized.

Results: The efficiency of the proposed risk prediction model is assessed on the Stroke prediction dataset. By obtaining an accuracy of 99.6%, the research outcomes demonstrated that the suggested model outperforms current techniques.

A novel proposed Adaptive Weight Bi-Directional Long Short-Term Memory (awbi-LSTM) classifier-based cerebrovascular stroke risk level prediction models

Thumilvannan[1], and Dr R Balamanigandan[2*]

[1*]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India

stvvannan@gmail.com

[2]Professor, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai-602105.

balamanigandanr.sse@saveetha.com

Corresponding Author:

Dr R Balamanigandan [2*]

Email address: balamanigandanr.sse@saveetha.com

ABSTRACT

Background: To prevent difficulties from developing, people with diabetes need access to healthcare services for the rest of their lives. Large amounts of data are produced by their disease management operations in a variety of areas, from medical to administrative. Even for highly desirable applications like the forecasting of cardiovascular disease, the primary cause of excess mortality in diabetes, difficulties in acquiring and interpreting these data prevent its subsequent use in an institutional context. Stroke care and diagnosis have been improved as a result of the detrimental effects it has on society. Caretakers can develop patient management by effectively mining and storing the patients' medical records according to an increasing synergy among technology and medical diagnosis. Therefore, it is essential to look at the relationships between these risk factors in the records of patients and understand how each one contributes significantly to heart stroke prediction.

Methods: This study does a thorough analysis of the numerous stroke risk variables found in electronic medical data.Hence, a novel proposed Adaptive Weight Bi-Directional Long Short-Term Memory (AWBi-LSTM) classifierbased stroke risk level prediction model for IoT data is proposed in this paper. Here, to efficiently train the classifier, the missing data are removed by Hybrid Genetic with Kmeans Algorithm (HKGA) and the data are aggregated. Then, to reduce the dataset size, the features are reduced with Independent Component Analysis (ICA). After the correlated features are identified using the T-test-based Uniform Distribution- gradient search rule based elephant herding optimization for cluster analysis (GSRBEHO) (T-test-UD-GSRBEHO). Next, to classify the risk levels accurately, the fuzzy rule-based decisions are created with the T-test-UDEHOA correlated features.The feature values obtained from the fuzzy

logic are given to the AWBi-LSTM classifier, which predicts and classifies the risk level of heart disease and diabetes. After the risk level is predicted, the data is securely stored in the database. ~~Here,~~ for secure storage, MD5- Elliptic Curve Cryptography (MD5-ECC) technique is utilized.
Results: The efficiency of the proposed risk prediction model is assessed on the Stroke prediction dataset. By obtaining an accuracy of 99.6%, the research outcomes demonstrated that the suggested model outperforms current techniques.

KEYWORDS

Internet of Things (IoT), Stroke Prediction, Improved Restricted Boltzmann Machine (IRBM), correlated features, Elephant Herd Optimization Algorithm (EHOA), Correlated Feature (CF).

1. INTRODUCTION

In recent years, the technologies enabling the Internet of Things and its applications have advanced significantly. This has made it possible for a significant number of objects to be linked to one another through the Internet in production, home automation, and health *(Gubbi et al., 2013)*. Numerous applications in the field of intelligent health aim to enhance the treatment and standard of life for people with chronic diseases. As a result of IoT, the importance of mobile health services increases since they are crucial for monitoring and managing patients with chronic conditions like diabetes and cardiovascular disease *(Yuehong et al., 2016; Guariguata et al., 2014)*. Findings show that patient data is especially useful in the field of smart health, and more specifically in the area of patient monitoring. In order to successfully implement an IoT application in this industry, one must have ensured the collection of a significant volume of data gathered through assessments of the patients' medical signs. Studies are useful to find patients who require "proactive care" to prevent their conditions from getting worse. Big data, for instance, should make it possible for patients with certain diseases to receive preventative therapy early on (for instance, heart failure, which is caused by diabetes or hypertension) *(Dhillon & Kalra, 2017)*. Some of them disclosed their personal information in exchange for saving lives, which helped patients' health *(Rghioui et al., 2019)*. There are many chronic diseases in existence today, including diabetes, cancer, heart disease, and stroke. ~~This~~ is a deadly illness that has recently ranked on top of the global list of ~~killers~~ and necessitates intensely vigilant surveillance to maintain patients healthy.

A significant risk factor for stroke is diabetes mellitus, which is characterized by chronic hyperglycemia brought on by an absolute or relative insulin deficit. People with diabetes have a two to five times higher chance of having a stroke than those without the disease. The necessity for focused cardiovascular risk reduction measures to stop the development, recurrence, and progression of acute stroke is supported by large clinical trials conducted in adults with diabetes.

According to an earlier estimate *(Benjamin et al., 2017),* a new or recurrent stroke affects 795 000 individuals annually in the US, with one case occurring every 40 seconds on average and in the first year after a stroke, one out of every five victims would die *(Koton et al., 2014)*. The burden of paying for the survivors' rehabilitation and health care falls heavily on their families and the medical field. From 2014 to 2015, stroke-related direct and indirect expenditures reached

approximately 45.5 billion US dollars *(Benjamin et al., 2019)*. To reduce the expense of earlier medications to delay the onset of and reduce the risks of stroke, accurate stroke prediction is essential. Electronic health records and retinal scans are just two examples of the medical data used to construct stroke risk prediction (SRP) algorithms. Methods based on deep learning and conventional machine learning generally correspond to these methods,such as Support Vector Machine (SVM), Decision Tree, and Logistic Regression *(Khosla et al.,2010; Monteiro et al., 2018; Sung et al.,2020)*. The best results for stroke prediction have reportedly been attained by deep neural networks (DNN) *(Cheon, Kim & Lim, 2019)*. It could be difficult to find the volume of reliable data required in a practical situation *(Wang, Casalino & Khullar, 2019)*. The strict privacy protection laws in the medical field make it difficult for hospitals to share stroke data. Small subsets of the complete database of stroke data are hence usually scattered among numerous institutions. In addition, stroke statistics may show extremely imbalanced positive and negative cases.

Machine learning (ML) techniques are typically selected for enhancing patient care because they deliver faster, more accurate outcomes while using less processing power. Due to its innate capacity to integrate data from numerous sources and handle vast amounts of data, deep learning (DL) improves the predictive feature *(Nasser et al., 2021)*. But they take longer to learn and evaluate data, have long prediction periods, and use a lot of processing resources for both training and recognition *(Raju et al., 2022)*. Previous models for predicting the risk of developing diabetes and heart disease used known risk factors including age, smoking, hypertension, cholesterol, and diabetes to forecast future risk. To determine if those who have both risk of cardiovascular disease and isolated impaired fasting or isolated impaired glucose tolerance, they did not include those with both as a separate group for analysis *(Kumar et al., 2021)*.

Hence, a novel framework has been proposed an Adaptive Weight Bi-Directional Long Short-Term Memory (AWBi-LSTM) classifier based stroke risk level prediction model for IoT data is proposed in this paper. Here, to efficiently train the classifier, the missing data are removed by HGKA algorithm, and the data are aggregated. Then, to reduce the dataset size, the features are reduced with Independent Component Analysis (ICA). After the correlated features are identified using the T-test-based Uniform Distribution- gradient search rule based elephant herding optimization for cluster analysis (GSRBEHO) (T-test-UD- GSRBEHO). The feature values obtained from the fuzzy logic are given to the AWBi-LSTM classifier, which predicts and classifies the risk level of heart disease and diabetes. After the risk level is predicted, the data is securely stored in the database. Here, for secure storage, MD5- Elliptic Curve Cryptography (MD5-ECC) technique is utilized which obtains better accuracy.

The structure of this research is systematized as follows: Section 2 analyses the various prior works associated with the suggested method. Section 3 discusses the suggested technique. Section 4 analyses the efficiency of the suggested methodologies. Finally, section 5 ends the research with a conclusion.

2. LITERATURE REVIEW

Data mining techniques help to predict heart disease and diabetes in patients using medical records. The latest researches on heart disease prediction using machine learning and deep learning techniques are surveyed in this section.

Hossen et al., (2021) performed a survey that is broken up into three categories: deep learning models for CVD prediction, machine learning models for CVD, and classification and data mining methodologies. The dataset used for prediction and classification, the tools utilized for each group of these methods, and the outcome metrics for reporting accuracy are also compiled and reported in this study.

According to this, *Ahuja, Sharma, & Ali (2019)* utilized SVM, MLP, Random Forest, Logistic Regression, and Decision Tree among other techniques. The PIMA dataset can be used to forecast patients' diabetes more precisely. Significant results for Naive Bayes were obtained by another investigation that employed the PIMA dataset *Pranto et al., (2020)*. The stacking method was used by *Kuchi et al. (2019)* to achieve a 95.4% accuracy. The diagnosis of diabetes needs more research, as stated by *Kavakiotis et al., (2017)*. By combining several classifiers, the accuracy of diabetes disease prediction can be increased. An essential component of medical care is an accurate disease diagnosis. Numerous researchers have produced effective, but inaccurate, diagnostic tools for cardiac disease.

Khan (2020) took the accuracy problem of cardiac illness into consideration and created an IoT-based structure for enhancing the accuracy rate. In this structure, multiple risk factors for heart attacks, including blood pressure and electrocardiogram, are evaluated utilizing a heart monitor and smart watch. Additionally, a better Deep CNN is used to forecast heart attacks accurately utilizing collected data. IOT framework has an accuracy rate of over 95%.

Pan et al., (2020) introduced an improved deep learning and CNN-based method for successful treatment of heart disease via the internet of things. The combination stated above aims to raise heart disease prognosis rates. The efficiency of the model is calculated utilizing every disease-related attribute and its minimization. Additionally, the suggested combination is put into practice via IoMT, and outcomes are assessed utilizing accuracy and processing speed, andthe model yields improved outcomes.

Ahmed et al., (2020) demonstrated a method for forecasting cardiac disease in real time using data streams that included patients' present medical condition. Finding the best machine learning (ML) techniques for heart disease prediction is the research's secondary goal. In order to increase accuracy, ML algorithm parameters are also adjusted. According to the findings, random forest has a greater accuracy rate than other ML techniques.

Yu et al., (2020) created stroke prediction method using each person's bio-signals. Most stroke detection techniques take visual data rather than bio-signals into consideration. In addition, the predicting system incorporates deep learning and random forest algorithms for choosing the best features and performing the prediction task, accordingly. Findings showed that the LSTM system obtains 93.8% accuracy, whereas the random forest-based system obtains 90.4% accuracy.

For managing the multimodality in the stroke dataset, *Bhattacharya et al., (2020)* built a model using antlions and DNNs. The Antlion technique is taken into consideration in this framework to

optimize the hyper parameter DNN. Additionally, the parameter-tuned DNN is used to forecast the data from strokes. When outcomes are compared to training time, it is found that the training time for that model is 38.13.

Ali et al., (2020) suggested an innovative medical system for estimating the probability of a heart attack. The feature fusion and ensemble deep learning algorithms are included in this framework. The feature fusion approach can be thought of as fusing attribute information from electronic records and sensor data. Additionally, the data gathering strategy eliminates irrelevant data. For even better outcomes, the algorithm is further developed via ensemble deep learning. The value of an intelligent medical system for forecasting heart attacks is demonstrated by simulation findings.

Heart disease and stroke are the second leading causes of death *Moghadas et al., (2020)*. If the condition was not identified in time, it got worse. Therefore, created IoT and Fog based system for accurate diagnosis taking the detection rate of heart disease into consideration as a potential problem. Additionally, ECG signals are considered for the accurate and prompt detection of cardiac illness, and k-NN is used to validate the previously described framework.

Yu et al., (2020) demonstrated utilizing the NIHSS the effects of stroke severity on elderly persons older than 65. For determining how severe a stroke will be for elderly people, the C4.5 algorithm is taken into consideration. In addition, thirteen rather than the eighteen elements of the stroke scale are included in the assessment. which shows C4.5 has a 91.11% accuracy rate.

Selvi & Muthulakshmi (2020) presented an optimal ANN (OANN) model for identifying heart disease. DBMIR and TLBO-ANN are two of the methods that make up the OANN model. Tuning the ANN's parameters requires the usage of TLBO. The OANN is implemented using the Apache Spark framework, which functions in both online and offline modes. OANN outperforms competitors thanks to parameter modifying and DBMIR.

Yahyaie et al., (2019) examined the effectiveness of an IoT model for accurately predicting cardiac illness. The ECG signal is considered in this research while assessing the model's efficacy. Utilizing a cloud-based internet application, a total of 271 people's data are gathered. Ninety features for heart disease are included in the gathered dataset. Additionally, the IoT model is trained utilizing a NN approach, and it is stated that this model achieves an acceptable level of accuracy. Smart health products, IoT, IoMT, and intelligent ML approaches like ANN, DNN, CNN, etc. may greatly enhance healthcare systems.

From this literature survey, it is clear that the existing methods have some limitations like lesser accuracy, time consuming etc. To address these limitations, a new deep learning-based approach is proposed to improve the performance of heart disease prediction.

3. PROPOSED METHODOLOGY

In this paper a novel proposed Adaptive Weight Bi-Directional Long Short-Term Memory (AWBi-LSTM) classifier based stroke risk level prediction model for IoT data is proposed. The proposed flow diagram is shown in *Figure 1*.

[Figure 1 about here]

3.1. Input Stroke Prediction Dataset

On the stroke prediction dataset, the suggested risk prediction algorithm's effectiveness is assessed. (https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset)
This dataset will determine if a person is likely to suffer from a stroke using 11 input characteristics including age,gender, profession,marital status, BMI,hypertension inclinations, glucose, chest discomfort, blood pressure, current diseases, and smoking status. This dataset contains more than 5,000 samples. The Kaggle Stroke Prediction Dataset can be found here.

3.2. Data-Preprocessing

Initially, the input data in the dataset I is preprocessed to enhance the working efficacy of the classifier. In the proposed technique, preprocessing is done by removing the missing values and aggregating the data.

a. Missing Data Removal using Hybrid Genetic with Kmeans Algorithm (HKGA)

One of the popular clustering techniques is the K-means approach, which has been applied in a variety of scientific and technological domains. According to the initial center vectors, one of the main issues with the k-means method is that it may result in empty clusters. The evolution concepts of natural selection and genetics form the foundation of genetic algorithms (GAs), which are adaptable heuristic search algorithms. The empty cluster problem is effectively solved by the hybrid k-means technique presented in this research, which is also utilized to cluster the data objects.
The following are the main issues with the K-means algorithm:

- Based on the original center vectors, it might yield empty clusters.
- Could converge to not optimal values.
- With a decent amount of computation work, it is impossible to find global solutions to huge problems.

This study introduces the hybrid genetic algorithm (HKGA), which effectively addresses these disadvantages.

Phase 1: K-Means Algorithm

Step 1: K initial cluster centres are selectedarbitrarily from the n observations .
Step 2: A point is allotted to cluster iff

(1)

Step 3: New cluster centres are computed as follows:

(2)

Where is the number of elements that belong to cluster
Step 4: If then terminate, otherwise continue from step 2.
obtain an initial center for each selected cluster following this procedure.

Phase 2: Genetic Algorithm

Step 1: Population initialization

Each individual represents a row-matrix of where is the number of observations, and each gene contains the integer [1, K] that denotes the cluster to which this observation belongs.For example, suppose there are ten observations that need to be allocated to four clusters k = 4.

Step 2: Evaluation

Determine the desired objective function, and then search for acceptable cluster classifications that minimize the fitness function. The K clusters' clustering fitness function given by

(3)

Step 3: Selection

The purpose of selection is to focus GA search on interesting areas of the search field. In this work, roulette wheel selection used, where individuals from each generation are chosen based on a probability value to survive into the following generation. Based on the following formula, the likelihood of variable selection relates to its fitness value in the population:

; (4)

Where , selection probability of a string in a population and

(5)

Step 4: crossover operator

The crossover is performed on each individual in this stage using a modified uniform crossover, where the offspring is created by selecting the person with a probability.

Step 5: mutation operator

The following mutation operator implementation for each person is used: first, randomly select two columns from ith individual; next, create two new columns.

Step 6: The best solutions identified so far throughout the process, as opposed to GA keeping the best solutions found among the current population.

b. Data Aggregation

After the missing data are removed, the data are aggregated by taking mean (α), median (β), standard deviation (sd) , and variance (ν) to standardize the dataset.

$$\alpha = \frac{\sum_{x=1}^{b} m_x}{b} \quad (6)$$

$$\beta = median(m_x) \quad (7)$$

$$sd = \sqrt{\frac{\sum_{x=1}^{b} (m_x - \alpha)^2}{b}} \quad (8)$$

$$\nu^2 = \frac{\sum_{x=1}^{b} (m_x - \alpha)^2}{b-1} \quad (9)$$

Thus, the preprocessed dataset (Y) is given as,

$$Y = \{K_1, K_2, ..., K_B\} or\ K_\upsilon, \upsilon = 1,2,..., B \quad (10)$$

Where, K_B represents the preprocessed B^{th} patient data.

3.3. Independent Component Analysis (ICA) for Feature Reduction

Independent Component Analysis (ICA) is the unsupervised feature extraction technique, which has been applied on many applications. It transforms the original data by using a transformation function. The model of the ICA is defined as,

(11)

Where, Y – Transformed data. s - Scalar matrix. X – Original data.

Here, the original data is transformed into transformation data by using tanh transformation function as a scalar function. The non-linearity among the data will be maximized and orthogonally for each data vector is achieved using this tanh transformation function. Selecting the number of Independent components is one of the important problem in ICA. The components which having greater than the 0.1 of average in the newly transformed data set.

3.4. Feature Correlation using T-test-UD- GSRBEHO

After the feature reduction, the correlated features are identified using the T-test-based Uniform Distribution- gradient search rule based elephant herding optimization for cluster analysis (UD-GSRBEHO) (T-test-UD- GSRBEHO).

Initially, the obtained features $\{t_r\}$ undergo a T-test, and the T-test process is given as,

$$\tau_r = \frac{\bar{t}_r - \bar{t}_{r+1}}{\sqrt{\delta^2\left(\left(d_r + d_{r+1}\right)/d_r \times d_{r+1}\right)}} \tag{12}$$

Where, τ is the T-value for the feature r, δ depicts the pooled standard errors of t_r, t_{r+1}, and d_r, d_{r+1} depicts the total number of data under the features t_r, t_{r+1}. $\bar{t}_r, \bar{t}_{r+1}$ depicts the mean values of the features t_r, t_{r+1}.

After the τ is calculated for all samples, the correlation between features is determined with the spearman correlation coefficient.

$$\lambda_r = 1 - \frac{6\sum \tau_r^2}{l\left(l^2 - 1\right)} \tag{13}$$

Where, l depicts the total number of features. From the correlated features, the non-zero values are combined with the reduced features $\{t_1, t_2, ..., t_l\}$. By doing this, a valid feature set is obtained. From the obtained feature sets, the optimal feature set is selected using the Uniform Distribution- Gradient Search Rule Based Elephant Herding Optimization for cluster analysis (UD-GSRBEHO) as follows,

Initialization: The obtained feature sets are the initial clan with a fixed number of elephants which is written as,

$$U = \{[u_1], [u_2], ..., [u_d]\} or \lfloor u_\varphi \rfloor, \varphi = 1,2,..., d \tag{14}$$

Where, U depicts the elephant population and $\lfloor u_\varphi \rfloor$ depicts the φ^{th} elephant clan with the same number of elephants as other clans. In each generation, the female elephants live with their clan, but the male elephants tend to move from the clan and live far away from the clan. Each clan of elephants is led by the matriarch.

Clan Updating Operator: The fitness of each elephant in a clan is estimated. The elephant with the best fitness is considered as the matriarch, meanwhile, the other elephants (ω) in the clan update their position according to the matriarch. Here, fitness is considered as the new position of ω in the clan $\lfloor u_\varphi \rfloor$ is given as $N_{[u_\varphi],\omega}^{R+1}$ which is evaluated as,

$$N_{[u_\varphi],\omega}^{R+1} = N_{[u_\varphi],\omega}^{R} + \Omega\left(N_{[u_\varphi]}^{*} - N_{[u_\varphi],\omega}^{R}\right)\gamma \tag{15}$$

Here, R signifies the iteration, $N_{[u_\varphi]}^{*}$ depicts the best solution of clan, Ω signifies the algorithm parameter, which indicates the influence of matriarch in the group, and γ signifies the random number obtained from uniform distribution as,

$$\gamma = \frac{1}{N_{[u_\varphi]+1}^{R} - N_{[u_\varphi]}^{R}} \tag{16}$$

The position of the best solution in each clan is updated with respect to the following equation,

$$N_{[u_\varphi]}^{R+1} = \Delta . N_{[u_\varphi]}^{c} \tag{17}$$

Where, Δ depicts the second algorithm parameter, which controls the influence of the clan center $N_{[u_\varphi]}^{c}$. The clan center is mathematically represented as,

$$N_{[u_\varphi],\Re}^{c} = \frac{1}{\eta_{[u_\varphi]}} . \sum_{\omega=1}^{\eta[u_\varphi]} N_{[u_\varphi],\omega,\Re} \tag{18}$$

Where, $\Re$ represents the dimension of $N_{[u_\varphi],\omega}$, and $\eta_{[u_\varphi]}$ indicates the total number of elephants in the clan $\lfloor u_\varphi \rfloor$.

Clan Separating Operator: The male elephants separate from the clan, which can be modeled by the separating operator. The separation is the removal of the worst elephants from the clans in each iteration as,

$$N'_{[u_\varphi]} = N_{mnm} + \left(N_{mxm} - N_{mnm} + 1\right) \tag{19}$$

Where, N_{mxm}, N_{mnm} depicts the upper and lower bound of the elephant in the clan $\lfloor u_\varphi \rfloor$. $N'_{[u_\varphi]}$ indicates the worst elephant position in the clan $\lfloor u_\varphi \rfloor$ which gets removed from the clan.

Gradient Search based Optimization (GBO)

The gradient approach is used to resolve the population-based technique known as GBO. In GBO, Newton's algorithm determines the search direction. In order to further explore the search space, two primary operators and a collection of vectors are modified. The worst-positioned agents are only arbitrarilychanged by (10), according to the research of EHO. The lack of a variation mechanism in this type of method results in an inadequate exploitation capability and a sluggish convergence. The best-positioned agents are also (8) altered. This would decrease population variety while being worthless once the population has settled into a local optimum. Additionally, EHO's exploitation potential is only moderately strong, which raises the probability of encountering a local optimum *(Khalilpourazari et al., 2021)*. A improved answer can be obtained by integrating with GBO since the search direction can be directed throughout the

iteration to prevent being stuck in a local optimum. The local escape operator (LEO) in GBO can increase population diversity and prevent overly long periods of stagnation. The gradient data can be fully utilized by the suggested method in this situation, increasing the program's search efficiency *(Hassan et al., 2021)*.

Gradient Search Rule (GSR)

To regulate the vector search's direction, Newton's technique yielded the gradient search rule (GSR). A number of vectors are included in order to maintain equilibrium among exploration and exploitation throughout the iterations and speed up convergence:

(20)

(21)

(22)

where and are taken as 1.2 and 0.2, correspondingly, m and M signify the current and the maximum number of iterations, correspondingly, and rand means a random number among [0,1]. The value of α changes during the iterations and might be utilized to control the rate of convergence. The technique can enhance the variety and fast converge to the region where it hopes to discover the best answer because early in the iteration, the value of α is high. The value falls as the loop progresses. As a result, the program can more effectively utilize the studied regions. Based on this, the GSR expression is as follows:

(23)

where and signify positions of the worst and the best agents, and ε is a small number in the range of [0,0.1]. The suggested GSR's capacity for a arbitrary search improves GBO's capacity for exploration and its capacity to depart from the local optimum. is determined with the following equation:

(24)

(25)

(26)

where represents N random numbers among [0,1] and step is the step size. means the global optimal agent, and represents the mth dimension of the nth agent. r1,r2,r3,r4 are different integers arbitrarily selected from [1, N].

For a local search, a motion parameter called DM is also set in order to enhance the exploitation capability. The expression is displayed as follows

(27)

signifies a random number among [0,1], and ρ2 is the parameter that controls the step size and is denoted as follows:

(28)

Finally, the current location of the search agent () can be improved by GSR and DM shown as follow:

(29)

The following is another way of expressing (29) into the context of 14 and 18:

(30)

where and is a newly generated variable defined by the average of and . Based on Newton's method, is expressed by:

(31)

where is definite by (15), and and signify the current worst and best agents, individually. After replacing the current vector in (21) with , a new vector can be attained with the following expression.

(32)

According to 21 and 23, the new solution can be denoted as:

(33)

(34)

where and are random numbers among [0,1].

Local Escaping Operator (LEO)

The method is tuned using the local escaping operator (LEO), which increases the probability of obtaining the ideal solution by allowing the program to move away from local optima.
The LEO introduces a solution that performs better and is expressed as:

(35)

end
is a predetermined threshold, here is a random number among [-1,1], and is a random number that conforms to the standard normal distribution. are respectively represented by:

(36)

(37)

(38)

where is a binary parameter of 0 or 1, and is a random number among [0,1]. When , otherwise,
In conclusion, the resulting solution is stated as follows:

(39)

where is a randomly selected solution from the population, is a binary parameter of 0 or 1, and is a random number among [0,1].
When otherwise, xrand is the newly generated solution in the following manner.

(40)

Finally, the best position of the clan (optimal feature sets) is updated $N^{*}_{[\mu_{\varphi}]}$ removing the $N'_{[\mu_{\varphi}]}$.

The optimal feature set is given as,

$$O = \{[\kappa_1], [, \kappa_2], ..., [\kappa_{tt}]\} or [\kappa_x], x = 1,2....., tt \quad (41)$$

Where, $[\kappa_{tt}]$ depicts the tt^{th} optimally selected feature set, and O represents the dataset after optimal feature set selection.

The pseudocode for UDEHOA is given as follows,

Input: Feature set $\{[u_1],[u_2],.....,[u_d]\}$

Output: selected feature sets

Begin

Initialize $\{[u_1],[u_2],.....,[u_d]\}$, population size, Maximum iteration $R_{\max}$

Set $R=1$

While $(R \le R_{\max})$ **do**

Compute Fitness of elephants

Determine clan updating operator $N^{R+1}_{[u_\varphi],\omega}$ with $\gamma = \dfrac{1}{N^{R}_{[u_\varphi]+1} - N^{R}_{[u_\varphi]}}$

Determine clan separating operator $N'_{[u_\varphi]}$

Evaluate fitness of $N^{R+1}_{[u_\varphi],\omega}$

If fitness of $N^{R+1}_{[u_\varphi],\omega}$ higher **Then**

Update clan position $N^{*}_{[u_\varphi]}$

for n=1:N do

fori=1: dim do

Arbitrarily selects r1,r2,r3,r4 in the range of [1,N]

Estimate GSR and DM based on (14) and (18)

Calculate ' '

Calculate

End for

If rand<pr then

Generate

End if

Calculate and update the fitness according to each position

End for

Else

$R = R+1$

End If

End While

Return optimal feature set $N^{*}_{[u_\varphi]}$

End

3.5. Decision Making

After the correlated feature sets are selected, the selected feature sets $\{[\kappa_x]\}$ are given to the Fuzzy logic, which fuzzifies the crisp inputs, generates decision-making rules, and fuzzily gives crisp feature values.

Initially, in the fuzzy logic, the membership function is assigned to fuzzify the input feature set. Here, to fuzzify $\{[\kappa_x]\}$ trapezoidal membership function is used, which is denoted as,

$$\nabla([\kappa_x], w, xx, Dia-cls, Hea-cls) = \max\left(\min\left(\frac{[\kappa_x]-w}{xx-w}, 1, \frac{Hea-cls-[\kappa_x]}{Hea-cls-Dia-cls}\right), 0\right) \quad (42)$$

Where, $\nabla(\ )$ depicts the trapezoidal membership function. $w, xx, dia-cls, z$ are the input parameters such as heart and diabetic feature values, diabetic and heart class. Then, the decision-making is performed with the rules such as,

$$normal = \{1 \ \ if\ dia-cls = 0\ \&\ Hea-cls = 0 \quad (43)$$

$$Dia-risk = \begin{cases} 2 & if\ dia-cls = 1\ \&\ Hea-cls = 0\ \&\ \alpha(dia) > xx \\ 3 & if\ dia-cls = 1\ \&\ Hea-cls = 0\ \&\ \alpha(0) < xx < \alpha(dia) \end{cases} \quad (44)$$

$$Hea-risk = \begin{cases} 5 & if\ dia-cls = 0\ \&\ Hea-cls = 1\ \&\ \alpha(Hea) > w \\ 6 & if\ dia-cls = 0\ \&\ Hea-cls = 1\ \&\ \alpha(0) < w < \alpha(Hea) \end{cases} \quad (45)$$

Where, 1, 2, 3, 4, 5, and 6 are the decision rules for the normal patient, low risk of diabetes, high risk of diabetes, low risk of heart disease, and high risk of heart disease respectively. $Dia-risk, Hea-risk$ depicts the diabetic and heart risk respectively, $Dia-cls, Hea-cls$ depicts the diabetic and heart-disease classes.The data aggregation means the value of heart disease and diabetes is given as $\alpha(Hea)$ and $\alpha(Dia)$ respectively. Similarly, the decision rules for low and high heat and diabetic risk of a patient were also determined. The diabetic patient has high risk of stroke. Finally, the crisp value of the feature can be obtained with defuzzification. For all the feature sets, the crisp values are given as,

$$V = \{g_1, g_2, ..., g_{cv}\} or\ g_{\square}, \square = 1,2,....,cv \quad (46)$$

Here, g_{cv} is crisp value of $[\kappa_{tt}]$ after applying fuzzy rules, and V depicts the defuzzified feature values. The feature values obtained from the fuzzy logic are given to the AWBi-LSTM classifier, which predicts and classifies the heart disease and diabetes.

3.6. Adaptive Weight Bi-Directional Long Short-Term Memory (AWBi-LSTM) for Classification and Risk Prediction

One aspect that a Recurrent Neural Network (RNN) network is different from the feed-forward network is that the neurons in hidden layers get the feedback, which involves from the prior state to the current state. Theoretically, RNN can learn the features of any length of time series. But, experiments show that the performance achieved with the RNN network can be limited owing to vanishing gradient or gradient explosion. To deal with the gradient problems that the RNN network experiences, Long Short Term Memory (LSTM) network is developed by presenting a core element known as the memory unit.

The LSTM includes specialized components known as memory blocks present in the recurrent hidden layer. The memory blocks includes memory cells having self-connections, which store the temporal state of the network along with specialized multiplicative modules known as gates for the information flow control. Every memory block in the actual model includes an input gate and an output gate. The input gate regulates the flow pertaining to input activations into the memory cell. The output gate regulates the output flow associations of the cell activations into the remaining part of the network. Subsequently, the forget gate was included in the memory block. It determines the amount of the memory cell that should be removed in a current memory cell. This deals with the setback of LSTM models stopping them from performing the processing of persistent input streams, which is not divided into subsequences. The forget gate carries out the scaling of the internal state of the cell prior to sending it as the input to the cell using the self-recurrent connection of the cell, thus achieving an adaptive forget or reset of the cell's memory. Moreover, the recent LSTM structure has keyhole connections running from its internal cells to the gates present in the same cell for learning the exact timing of the outputs. The final gate represented as o, whose name is given following the output gate, regulates the amount of information used for computing the output activation of the memory unit and also flows into the remaining part of the network (Yao *et al.* 2014; Zhang *et al.* 2016).

[Figure 2 about here]

With an LSTM network, an input sequence $x = (x_1, .., x_T)$ is mapped on to an output sequence $y = (y_1, .., y_T)$ by estimating the network unit activations applying the following equations in an iterative manner from t = 1 to T (See *Figure 2*). In the LSTM, W terms represent weight matrices (e.g. W_{ix} indicates the matrix of weights from the input gate to the input), W_{ic}, W_{fc}, and W_{oc} stand for the diagonal weight matrices for peephole connections, and the b terms specifies the bias vectors (b_i refers to the input gate bias vector), σ signifies the logistic sigmoid function, and i, f, o, and c notates the input gate, forget gate, output gate, and cell activation vectors correspondingly, each one of which hold equal size as the cell output activation vector m, indicates the element-wise product of the vectors, g and h refer to the cell input and cell output activation functions, and stands for the network output activation function, softmax.

In the LSTM classifier, weights can be considered as the connection strength. Weight is accountable for the degree of effect that will be put on the output when a modification in the input is seen. A lesser weight value will not change the input, and on the other hand, a bigger weight value will modify the output drastically. Every component includes weights corresponding to all of its input from the earlier layer, in addition to the input from the earlier time step. Associative memory applying fast weights is a short-term memory technique, which considerably enhances the memory capability and time scale of RNNs.

Bi-LSTM extends LSTM; it is helpful in discovering the associations between datasets. Two LSTM networks, one exhibiting a forward direction and another in the backward direction, are linked to the same output layer to select the features optimally. In this research work, Rand Index (RI) is regarded as the fitness function for optimally selecting the features from the dataset. The same sequence of data is used for training both of them. Three gates exist, which are known as

input, forget, and output gate, in an LSTM unit. These gates operate on the basis of the expressions (47-52),

(47)

(48)

(49)

(50)

(51)

(52)

Here, w_i, w_f, and w_o refer to the weights of LSTM, and b_i, b_f, and b_o indicate the biases. i_t stands for the input gate, f_t signifies the forget gate, and o_t represents the output gate. x_t signifies the input vector and h_t stands for the output vector. c_t refers to the cell state and t implies the candidate of the cell state. In the case of the forward LSTM, it can be expressed as . In accordance, the backward LSTM is with . Both and constitute the output of Bi-LSTM at a time,

(53)

Especially, the optimization of the Bi-LSTM (i.e, weight values) is performed dynamically. Therefore, the fitness function can be changed and can assess the fitness score of every Bi-LSTM from the respective training process in the same weight creation process. It implies that the fitness scores assessed in multiple generations cannot be compared with one another. In the AWBi-LSTM algorithm, the mutation parameter is used for generating new weights according to the mean value of a feature. The selection technique of AWBi-LSTM is denoted as , and it is ranked based on their fitness function, the highest mean weight values (μ) are chosen as the top feature .

Algorithm 3.1. Adaptive Weight Bi-Directional Long Short-Term Memory (AWBi-LSTM)

Input: Total number of samples in the dataset N, the number of mutations n_m, the batch size m, dataset D, and initial weight ,

Output: Best chosen features from the dataset

Start w =
Initialize model parameter
for i = 1 to m/(Nn_m)
param⟵w save model parameters
for j = 1 to N
for k = 1 to n_m
M(param) assign parameters to the model
obtain a set D as input x_i of AWBi-LSTM;
switch(k)
case1: $loss_{square}$, $param_{square}$⟵M(x_i, square, param)
case2: $loss_{abs}$, $param_{abs}$⟵M(x_i, abs, param)
case3: $loss_{huber}$, $param_{huber}$⟵M(x_i, huber, param)
end switch

if k = 1 to n_m
$loss_{min} \leftarrow \min(loss_{square}, loss_{abs}, loss_{huber})$
$param_{new} \leftarrow (loss_{min}, param_{square}, param_{abs}, param_{huber})$
$w \leftarrow param_{new}$
end for
end for
End

3.7. Data Security using MD5-ECC

After the risk level is predicted, the data is securely stored in the database. Here, for secure storage, MD5-ECC technique is utilized. Elliptic curves cryptography (ECC) algorithm, which is secure ones despite requiring very little computation and a very small key size compared to other techniques with more computation and a larger key size. The complexity and difficulty of this algorithm, however, increases the probability of implementation errors and reduces the system's security. Therefore, MD5-ECC is suggested to increase the security level of ECC. In the ECC, only the public and private types of keys are produced; however, MD5-ECC added a third type of key known as the secret key by using the MD5 hash function to increase the security of the system. The use of the MD5 is intended to increase the complexity of the ECC. The complexity of the algorithms rises as the attackers attempt to assault the data. The produced secret key is used for decryption as well as encryption. Thus, the MD5-ECC, whose mathematical description is shown here,

(54)

here, a and b means the integers. In the suggested work, '3' different types of keys must be established.

Step 1:regarded point as the curve's base point. Create the public key A using the equation (20).

(55)

here, K implies the private key that has been chosen inside the range of (1 to n − 1).

Step 2: Create a new secret key by appending the salt value to the public key and using the MD5 hash algorithm to create a hash value from this value.The new key is developed as

(56)

hereindicates a salt value that is arbitrarily chosen.

Step 3: Use the secret key along with the public key, which is a point on the curve, to encrypt the data. The secret key is combined with the encryption algorithm in the suggested MD5-ECC. The encrypted data consists of '2' ciphertexts, which are mathematically denoted as,

(57)

(58)

wherein E1 and E2 indicates the encrypted text 1 and encrypted text 2, R implies the random number, which is on the gamut[1, n − 1]together with D indicates the data. Since the decryption, the original data has been obtained.

Step 4: By performing the encryption's reverse operation, the data can be decrypted. The secret key is subtracted from the decryption equation during decryption, which is formally represented as

(59)

Hence, with the ECC cryptography technique, the medical results are stored on the medical database securely.

4. RESULTS AND DISCUSSION

Here, the performance of the suggestedmethod is evaluated. The suggested technique is employed in the working platform of MATLAB. The performance analysis, as well as the comparative analysis, is was done to prove the effectiveness of the work.

The proficiency of the proposed risk prediction model is estimated on the Stroke prediction dataset (https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset)

This dataset will determine if a person is likely to have a stroke using 11 input characteristics including age,gender, marital status, hypertension inclinations, profession, glucose,BMI, blood pressure, current diseases, chest discomfort, and smoking status. This dataset contains more than 5,000 samples. Here is a link to the Kaggle Stroke Prediction Dataset.

csv dataset: The heart stroke prediction dataset from Kaggle was used as a csv dataset. The dataset contains of 11 parameters such as age, id, gender, worktype, hypertension,residence type, avg level of glucose, heart disease, body mass index (bmi), smoking behaviour, marital status, and stroke.

Description of the dataset: There are 11 attributes in the dataset, and each one identifies whether the data is categorized or numerical.

id: This component displays a person's distinctive identifier. Information that can be computed.

age: This trait serves as a proxy for the person's age. details on classifications

Gender: This attribute reveals the person's gender. Information that is obtainable.

Hypertension: This characteristic shows if the person has high blood pressure or not. details regarding the classifications.

Work type: This characteristic describes a person's employment. details regarding the classifications.

Residence type: This characteristic reflects the person's current state of affairs.

Heart disease: This characteristic suggests that the person may have heart disease. Information that is calculable.

Average glucose level: This statistic shows the average level of a person's blood sugar. Information that is calculable.

Bmi: The acronym Bmi stands for "numerical data." The BMI (body mass index) of a person is referred to in this attribute.

ever married: details from the group. This attribute denotes a person's marital status.

Smoking status: Statistical data broken down by category. This trait reveals a person's smoking status.

stroke: This characteristic tells whether or not someone suffered a stroke. The complete attribute dash represents the choice class, while the answer class represents the remaining attributes.
The input dataset is alienated into train and test datasets, with the training model's dataset comprising 80% of the whole amount. The collection of data utilized to develop a machine learning model is known as a training dataset. The efficiency of the trained model is demonstrated using the test datasets.

4.1. Performance Analysis of HKGA

The suggestedHKGAperformance is analyzed with the existing methods, such as K-means, Gaussian Mixture Model (GMM) algorithm, and K-Nearest Neighbor (KNN) based on the time consumed for clustering.

[Table 1 about here]

The clustering time of the suggestedand the presenttechnique are illustrated in *table 1*.The proposed HKGAmethod takes a clustering time of 1.181 sec. But, the existingmethods consumed high time for clustering. Thepartial derivative of the Hamiltonian in the conventionalK-means showed improvement in the clustering time.The above analysis delivered that the proposedmethod yields lesser time for clustering than the existing methods.

4.2. Performance Analysis of ICA

The suggested ICA performance is compared with the prevailing techniquelike SS-PCA, PCA, Linear Discriminative Analysis (LDA), and Gaussian Discriminative Analysis (GDA) based on the metrics, such as Peak Signal-to-Noise Ratio (PSNR), Mean Square Error (MSE) and R-Square.

[Table 2 about here]

The performance of the suggested ICA along with present approaches is assessed in *table 2* regarding the quality metrics, PSNR and MSE. The better performance of the feature reduction technique is represented by the higher PSNR and lower MSE values. The PSNR value attained by the proposed technique is 2.7 dB higher than the prevailing SS-PCA, PCA, LDA and GDA technique. When compared with the conventional frameworks, the proposed ICA has obtained a low error value of 0.01010. The process of shell sorting has improved the performance of the present PCA. Thus, the suggested ICA is efficient in reducing the features.
R-Square is a statistical measure that must be high to exhibit better performance values shown in *table 2*. The suggested technique exhibited an R-Square value of 0.810, whereas the existing models had lower R-Square values of 0.756 (SS-PCA) 0.653 (PCA), 0.374 (LDA), and 0.175(GDA). Thus, the analyses show that the suggested strategy is significantly superior to others.

4.3. Performance Analysis of AWBi-LSTM Classifier

The suggested AWBi-LSTM classifier model performance is analyzed with the present methods, such as RBM, Convolution Neural Network (CNN), Deep Neural Network (DNN),andRecurrent Neural Networks (RNN). The proposed technique is compared with the present one based on the

quality metrics like sensitivity, specificity, precision, accuracy, F-Measure, False Positive Rate (FPR), False Recognition Rate (FRR), False Negative Rate (FNR), Net present value (NPV), Mathew Correlation Coefficient(MCC), and confusion matrix.

[Table 3 about here]

[Figure 3. About here]

In *figure 3*, the performance of the suggested and present approaches are analyzed according to Sensitivity, Specificity, accuracy, and precision values shown in *table 3*. The proposed method achieved a sensitivity of 98.81%, whereas the existing PLD-SSL-RBM, RBM, CNN, DNN, and RNN have 98.42%, 88.25%, 87.28%, 85.83%, and 85.71%, respectively, Likewise, the suggestedtechnique has specificity, accuracy, and precision of 99.73%, 99.55%, and 98.42%, respectively, which is higher than the existing methods.Semi-Supervised Learning and Power Lognormal Distribution have enhanced the performance of the classifier to a greater extent.Overall, the performance analysis reveals that the proposed method accurately classified the risk classes.

[Table 4 about here]

Table 4 exhibits the performance of the suggestedAWBi-LSTMaccording to F-Measure, MCC, and NPV. The values of F-Measure, NPV, and MCC of the proposed method are 98.89%, 99.84%, and98.52%, whereas the existing methods provide comparatively lower performance. In this performance comparison, the proposed AWBi-LSTM method proffered a better performance than all other existing techniques.

[Figure 4. about here]

Figure 4 illustrates the analysis of the suggested method with the present methods according to F-measure, NPV and MCC are the values that contribute to the false prediction. The proposed model attained higher F-measure, NPV and MCC values. Hence, it is stated that the proposed method achieved greater performance and classified the classes accurately.

[Figure 5. about here]

The classification model's behaviour utilizing the confusion matrix is shown in *Figure 5*. By contrasting the predicted class with the actual class, the confusion matrix is utilized to assess the model's accuracy. The percentage of occurrences that are successfully classified is calculated using a classifier's accuracy. The confusion matrix clearly shows that the proposed AWBi-LSTMprovides better accuracy. Thus, the suggestedframework is more efficient in classifying the stages.

4.4. Comparative Measurement with Literature Papers

Here, the effectiveness of the recommended approach is compared with traditional approaches like Classification and Regression Tree (CART) (Carrizosa, Molero-Río, & Romero Morales, 2021), Stacked Sparse Auto-Encoder and Particle Swarm Optimization (SSAE-PSO) (Mienye &

Sun, 2021), and Stacking Algorithm (SA) (Abdollahi & Nouri-Moghaddam, 2022) based on precision, accuracy and F-Measure obtained using the Framingham dataset.

Table 5 demonstrates the comparative analysis of the suggested AWBi-LSTM model and the models used in recent studies. From the analysis, it is revealed that the suggested framework was more efficient than other frame working predicting diabetes and heart disease.

[Table 5 about here]

5. CONCLUSION

This study includes, a novel framework termed AWBi-LSTM -based diabetes and stroke disease prediction models for IoT has been proposed. This framework works under the following phases: Pre-processing phase, Feature reduction, Feature Correlation, Decision making, Optimal Feature set Selection, Classification and Risk prediction, and finally, the Encryption stage. Then, The stroke prediction dataset is used for the performance study to compare outcomes with those of current systems and assess how well the system that is suggested performs. From the experimental analysis, the proposed framework achieves an accuracy of 99.65%, precision of 98.64%, and F-measure of 98.89%. The suggested approach required a clustering time of 1 sec less than the current system. Thus, it concluded the suggested approach is better and more efficient than other present techniques. However, the proposed work focused only on diabetes and heart stroke disease risk analysis provides better results.

DECLARATION

Ethics Approval and Consent to Participate: No participation of humans takes place in this implementation process

Human and Animal Rights: No violation of Human and Animal Rights is involved.

Funding: No funding is involved in this work.

Conflict of Interest: Authors and Co Authors have on conflict of interest.

Data Availability Statement: No data is generated during this study.

REFERENCES

1. Gubbi J, Buyya R, Marusic S, Palaniswami M. 2013. Internet of Things (IoT): A vision, architectural elements, and future directions. *Future generation computer systems* 29(7): 1645-1660 DOI: 10.1016/j.future.2013.01.010
2. Yuehong YIN, Zeng Y, Chen X, Fan Y. 2016. The internet of things in healthcare: An overview. *Journal of Industrial Information Integration* 1:3-13 DOI: https://doi.org/10.1016/j.jii.2016.03.004
3. Guariguata L, Whiting DR, Hambleton I, Beagley J, Linnenkamp U, Shaw JE. 2014. Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes research and clinical practice* 103(2):137-149 DOI: https://doi.org/10.1016/j.diabres.2013.11.002

4. Dhillon PK, Kalra S. 2017. Secure multi-factor remote user authentication scheme for Internet of Things environments. *International Journal of Communication Systems* 30 (16):1-20 DOI: https://doi.org/10.1002/dac.3323
5. Rghioui A, Lloret J, Parra L, Sendra S, Oumnad A. 2019. Glucose data classification for diabetic patient monitoring. *Applied Sciences* 9(20):1-15 DOI: 10.3390/app9204459
6. Benjamin EJ, Blaha MJ, Chiuve SE, Cushman M, Das SR, Deo R, De Ferranti SD, Floyd J, Fornage M, Gillespie C, Isasi CR. 2017. Heart disease and stroke statistics-2017 update: a report from the American Heart Association. *circulation* 135(10):e146-e603 DOI: https://doi.org/10.1161/CIR.0000000000000485
7. Koton S, Schneider AL, Rosamond WD, Shahar E, Sang Y, Gottesman RF, Coresh J. 2014. Stroke incidence and mortality trends in US communities, 1987 to 2011. *Jama* 312(3):259-268.
8. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Das SR, Delling FN. 2019. Heart disease and stroke statistics-2019 update: a report from the American Heart Association. *Circulation* 139(10):e56-e528 DOI: https://doi.org/10.1161/CIR.0000000000000659
9. Khosla A, Cao Y, Lin CCY, Chiu HK, Hu J, Lee H. 2010. An integrated machine learning approach to stroke prediction. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* 183-192 DOI: https://doi.org/10.1145/1835804.1835830
10. Monteiro M, Fonseca AC, Freitas AT, e Melo TP, Francisco AP, Ferro JM, Oliveira AL. 2018. Using machine learning to improve the prediction of functional outcome in ischemic stroke patients. *IEEE/ACM transactions on computational biology and bioinformatics* 15(6):1953-1959 DOI: 10.1109/TCBB.2018.2811471
11. Sung SF, Lin CY, Hu YH. 2020. EMR-based phenotyping of ischemic stroke using supervised machine learning and text mining techniques. *IEEE journal of biomedical and health informatics* 24(10):2922-2931 DOI: 10.1109/JBHI.2020.2976931
12. Cheon S, Kim J, Lim J. 2019. The use of deep learning to predict stroke patient mortality. *International journal of environmental research and public health* 16(11):1-12 DOI: 10.3390/ijerph16111876
13. Wang F, Casalino LP, Khullar D. 2019. Deep learning in medicine-promise, progress, and challenges. *JAMA internal medicine* 179(3):293-294 DOI: 10.1001/jamainternmed. 2018.7117
14. Nasser AR, Hasan AM, Humaidi AJ, Alkhayyat A, Alzubaidi L, Fadhel MA, Santamaría J, Duan Y. 2021. Iot and cloud computing in health-care: A new wearable device and cloud-based deep learning algorithm for monitoring of diabetes. *Electronics* 10(21):1-12 DOI: https://doi.org/10.3390/electronics10212719
15. Raju KB, Dara S, Vidyarthi A, Gupta VM, Khan B. 2022. Smart Heart Disease Prediction System with IoT and Fog Computing Sectors Enabled by Cascaded Deep Learning Model. *Computational Intelligence and Neuroscience* DOI: https://doi.org/

10.1155/2022/1070697
16. Kumar D, Verma C, Dahiya S, Singh PK, Raboaca MS, Illés Z, Bakariya B. 2021. Cardiac diagnostic feature and demographic identification (CDF-DI): An iot enabled healthcare framework using machine learning. *Sensors* 21(19):1–30 DOI: https://doi.org/10.3390/s21196584
17. Hossen MA, Tazin T, Khan S, Alam E, Sojib HA, Monirujjaman Khan M, Alsufyani A. 2021. Supervised machine learning-based cardiovascular disease analysis and prediction. *Mathematical Problems in Engineering* 1-10. DOI: 10.1155/2021/1792201.
18. Ahuja R, Sharma SC, Ali M. 2019. A diabetic disease prediction model based on classification algorithms. *Annals of Emerging Technologies in Computing (AETiC)* 3:44–52. DOI: 10.33166/AETiC.2019.03.005.
19. Pranto B, Mehnaz SM, Mahid EB, Sadman IM, Rahman A, Momen S. 2020. Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. *Information* 11(8):1-20. DOI: 10.3390/info11080374.
20. Kuchi A, Hoque MT, Abdelguerfi M, Flanagin MC. 2019. Machine learning applications in detecting sand boils from images. *Array* 3-4:1-15. DOI: 10.1016/j.array.2019.100012.
21. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. 2017. Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal* 15:104-116. DOI: 10.1016/j.csbj.2016.12.005
22. Khan MA. 2020. An IoT framework for heart disease prediction based on MDCNN classifier. *IEEE Access* 8:34717-34727. DOI: 10.1109/ACCESS.2020.2974687
23. Pan Y, Fu M, Cheng B, Tao X, Guo J. 2020. Enhanced Deep Learning Assisted Convolutional Neural Network for Heart Disease Prediction on the Internet of Medical Things Platform. *IEEE Access* 8:189503-189512. DOI: 10.1109/ACCESS.2020.3026214
24. Ahmed H, Younis EM, Hendawi A, Ali, AA. 2020. Heart disease identification from patients' social posts, machine learning solution on Spark. *Future Generation Computer Systems* 111:714-722. DOI: https://doi.org/10.1016/j.future.2019.09.056
25. Yu J, Park S, Kwon SH, Ho CMB, Pyo CS, Lee H. 2020. AI-Based Stroke Disease Prediction System Using Real-Time Electromyography Signals. *Applied Sciences* 10(19), pp.1-19. DOI: 10.3390/app10196791
26. Bhattacharya S, Maddikunta PKR, Hakak S, Khan WZ, Bashir AK, Jolfaei A, Tariq U. 2020. Antlion re-sampling based deep neural network model for classification of imbalanced multimodal stroke dataset. *Multimedia Tools and Applications* 1-25 DOI: https://doi.org/10.1007/s11042-020-09988-y
27. Ali F, El-Sappagh S, Islam SR, Kwak D, Ali A, Imran M, Kwak KS. 2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion* 63:208-222 DOI: https://doi.org/10.1016/j.inffus.2020.06.008
28. Moghadas E, Rezazadeh J, Farahbakhsh R. 2020. An IoT patient monitoring based on fog computing and data mining: Cardiac arrhythmia usecase. *Internet of Things* 11:1-11 DOI:

https://doi.org/10.1016/j.iot.2020.100251
29. Yu J, Park S, Lee H, Pyo CS, Lee YS. 2020. An elderly health monitoring system using machine learning and in-depth analysis techniques on the NIH stroke scale. *Mathematics 8*(7):1-17 DOI:10.3390/math8071115
30. Selvi RT, 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* 1-11 DOI: 10.1007/s12652-020-02181-x
31. Yahyaie M, Tarokh MJ, Mahmoodyar MA. 2019. Use of internet of things to provide a new model for remote heart attack prediction. *Telemedicine and e-Health* 25(6):499-510 DOI: https://doi.org/10.1089/tmj.2018.0076
32. Khalilpourazari S, Doulabi HH, Çiftçioğlu AÖ, Weber GW. 2021. Gradient-based grey wolf optimizer with Gaussian walk: Application in modelling and prediction of the COVID-19 pandemic. *Expert Systems with Applications* 177:1-23 DOI: 10.1016/j.eswa. 2021.114920
33. Hassan MH, Houssein EH, Mahdy MA, Kamel S. 2021. An improved manta ray foraging optimizer for cost-effective emission dispatch problems. *Engineering Applications of Artificial Intelligence* 100:1-20. DOI: https://doi.org/10.1016/j.engappai.2021.104155
34. Carrizosa E, Molero-Río C, Romero Morales D. 2021. Mathematical optimization in classification and regression trees. *Top* 29(1):5-33. DOI: https://doi.org/10.1007/s11750-021-00594-1
35. Mienye ID, Sun Y. 2021. Improved heart disease prediction using particle swarm optimization based stacked sparse autoencoder. *Electronics* 10(19):1-15 DOI: https://doi. org/10.3390/electronics10192347
36. Abdollahi J, Nouri-Moghaddam B. 2022. Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction. *Iran Journal of Computer Science 5*(3):205-220. DOI: https://doi.org/10.1007/s42044-022-00100-1

Table 1(on next page)

Comparative Analysis of the Proposed HKGA Model

1 **Table 1.** Comparative Analysis of the Proposed HKGA Model

Techniques	**Clustering Time(sec)**
Proposed HKGA	1.181
DH-CC-KC	1.239
K-means	1.293
GMM	2.636
KNN	5.174

2

Box 1(on next page)

Performance Analysis of the Proposed ICA

1 **Table 2.** Performance Analysis of the Proposed ICA

Techniques/Metrics	**PSNR(dB)**	**MSE**	**R-Square**
ICA	40.99	0.01010	0.810
SS-PCA	39.87	0.01015	0.756
PCA	38.85	0.01141	0.653
LDA	37.94	0.01268	0.374
GDA	31.25	0.02738	0.175

2

Table 2(on next page)

Performance Analysis of the Proposed AWBi-LSTM Classifier Method

1 **Table 3.** Performance Analysis of the Proposed AWBi-LSTM Classifier Method

Techniques/Metrics	Sensitivity	Specificity	Accuracy	Precision
ProposedAWBi-LSTM	**98.81**	**99.80**	**99.65**	**98.64**
PLD-SSL-RBM	98.42	99.73	99.55	98.42
RBM	88.25	99.61	99.02	97.56
CNN	87.28	99.40	98.52	97.03
DNN	85.83	99.12	98.31	96.31
RNN	85.71	99.03	98.01	96.21

2

Table 3(on next page)

Performance Analysis of Proposed AWBi-LSTM

1 **Table 4.** Performance Analysis of Proposed AWBi-LSTM

Techniques/Metrics	**F-measure (%)**	**NPV (%)**	**MCC (%)**
ProposedAWBi-LSTM	98.89	99.84	98.52
PLD-SSL-RBM	98.42	99.73	98.16
RBM	88.25	98.04	86.29
CNN	87.29	97.88	85.16
DNN	85.83	97.63	83.47
RNN	85.71	97.61	83.33

2

Table 4(on next page)

Comparative Analysis of the Proposed Model and Previous Studies

1 **Table 5.** Comparative Analysis of the Proposed Model and Previous Studies

Techniques/Metrics	**Accuracy (%)**	**Precision (%)**	**F-Measure (%)**
CART	91	92	91
SSAE- PSO	97.3	94.8	97.3
SA	90.24	92	90
PLD-SSL-RBM	99.55	98.42	98.42
Proposed AWBi-LSTM	99.65	98.64	98.89

2

3

Figure 1

PROPOSED FLOW DIAGRAM

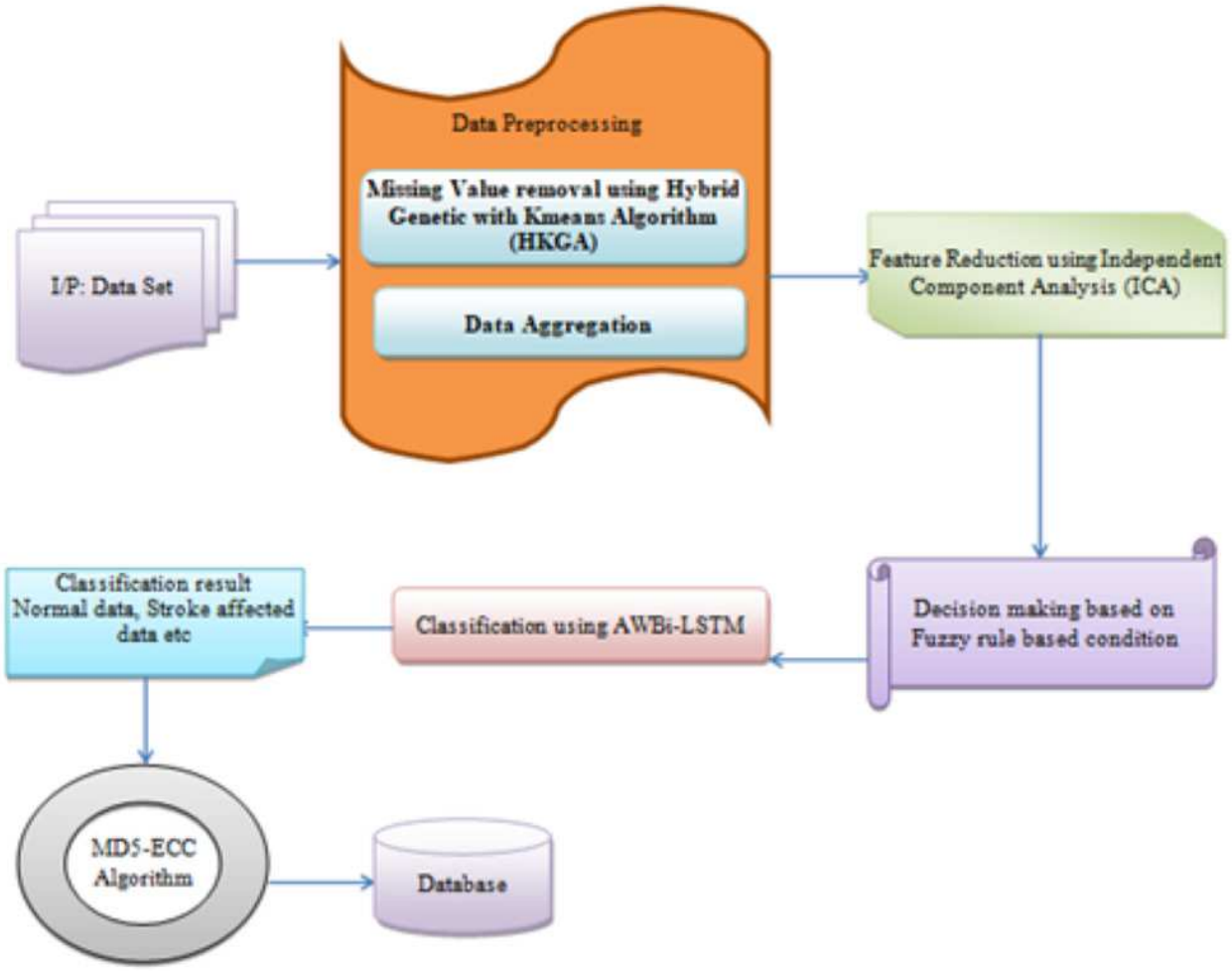

Figure 2

LSTMP RNN ARCHITECTURE

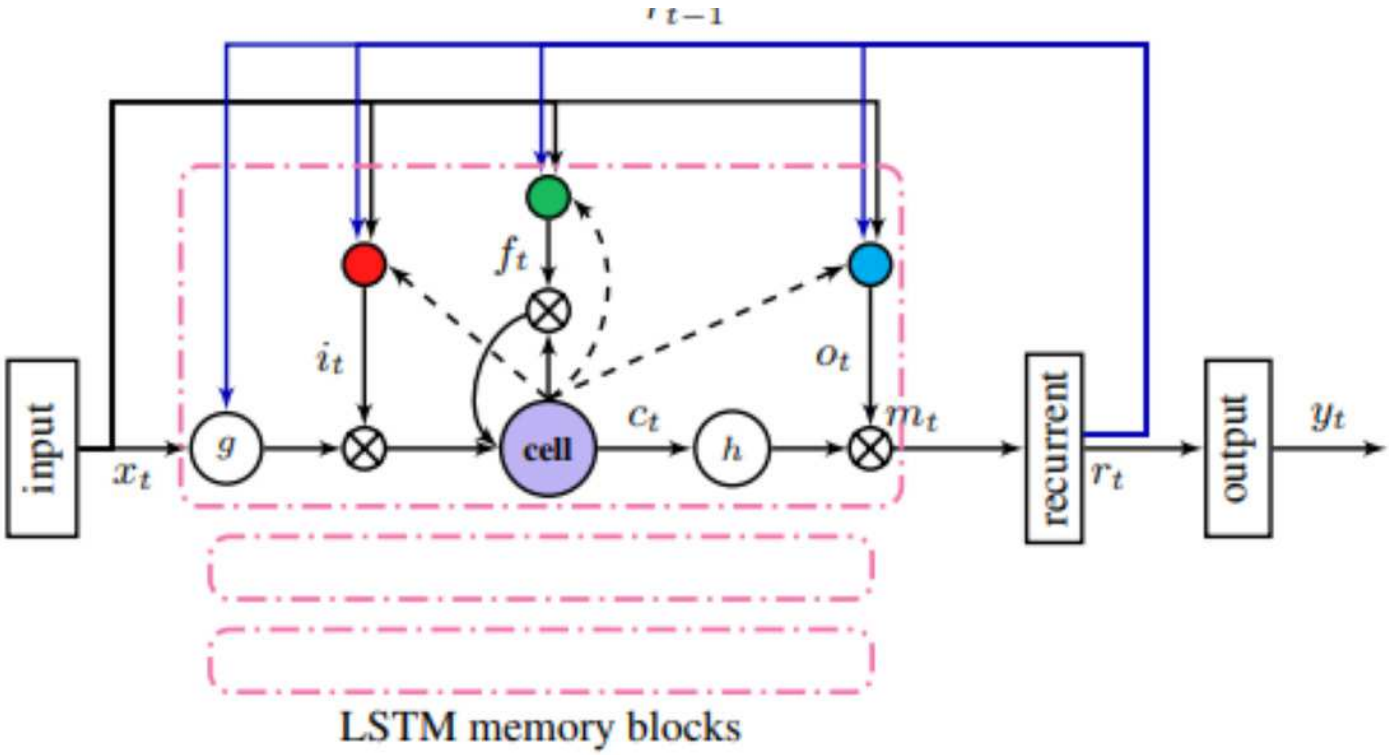

Figure 3

PERFORMANCE ANALYSIS OF THE PROPOSED AWBi-LSTM CLASSIFIER METHOD

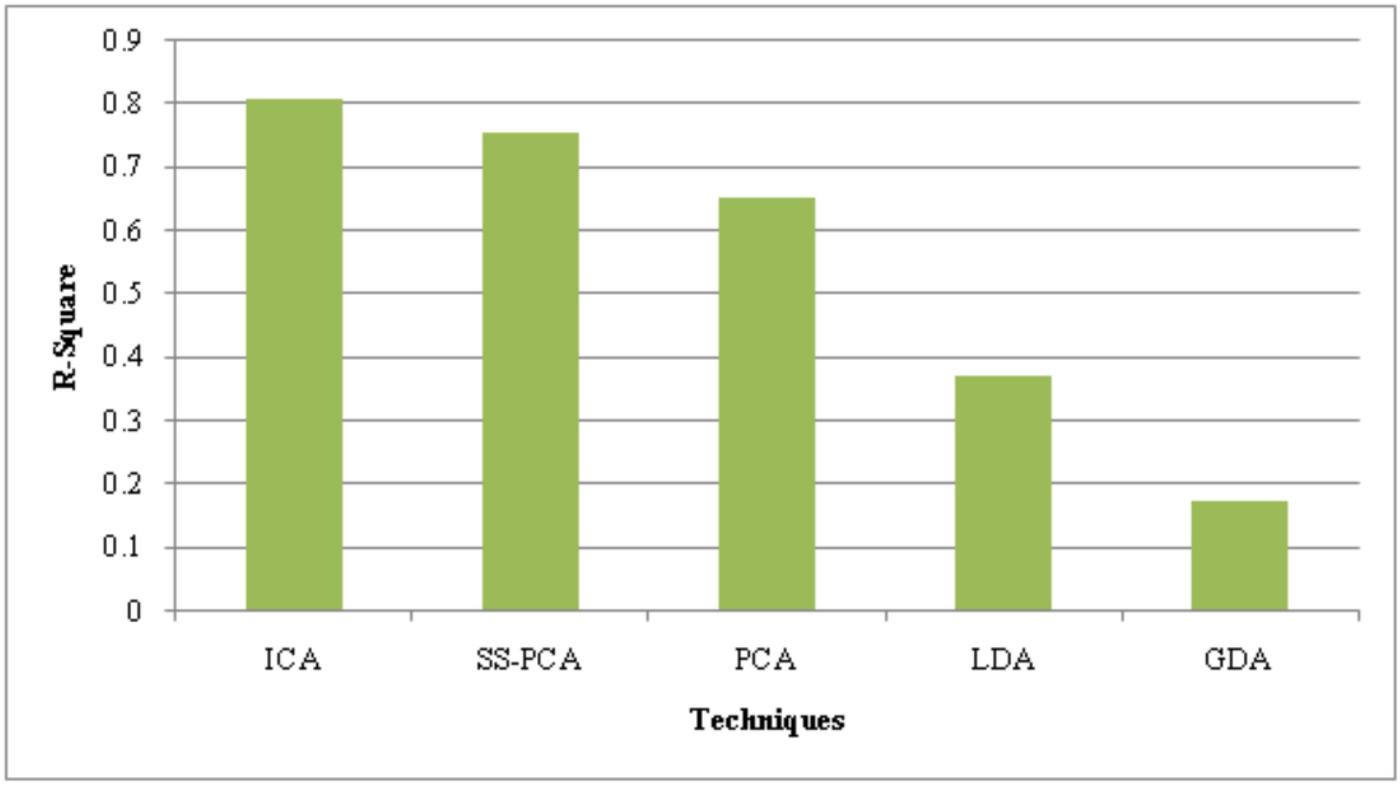

Figure 4

COMPARATIVE ANALYSIS OF THE PROPOSED AWBi-LSTM AND THE EXISTING METHODS

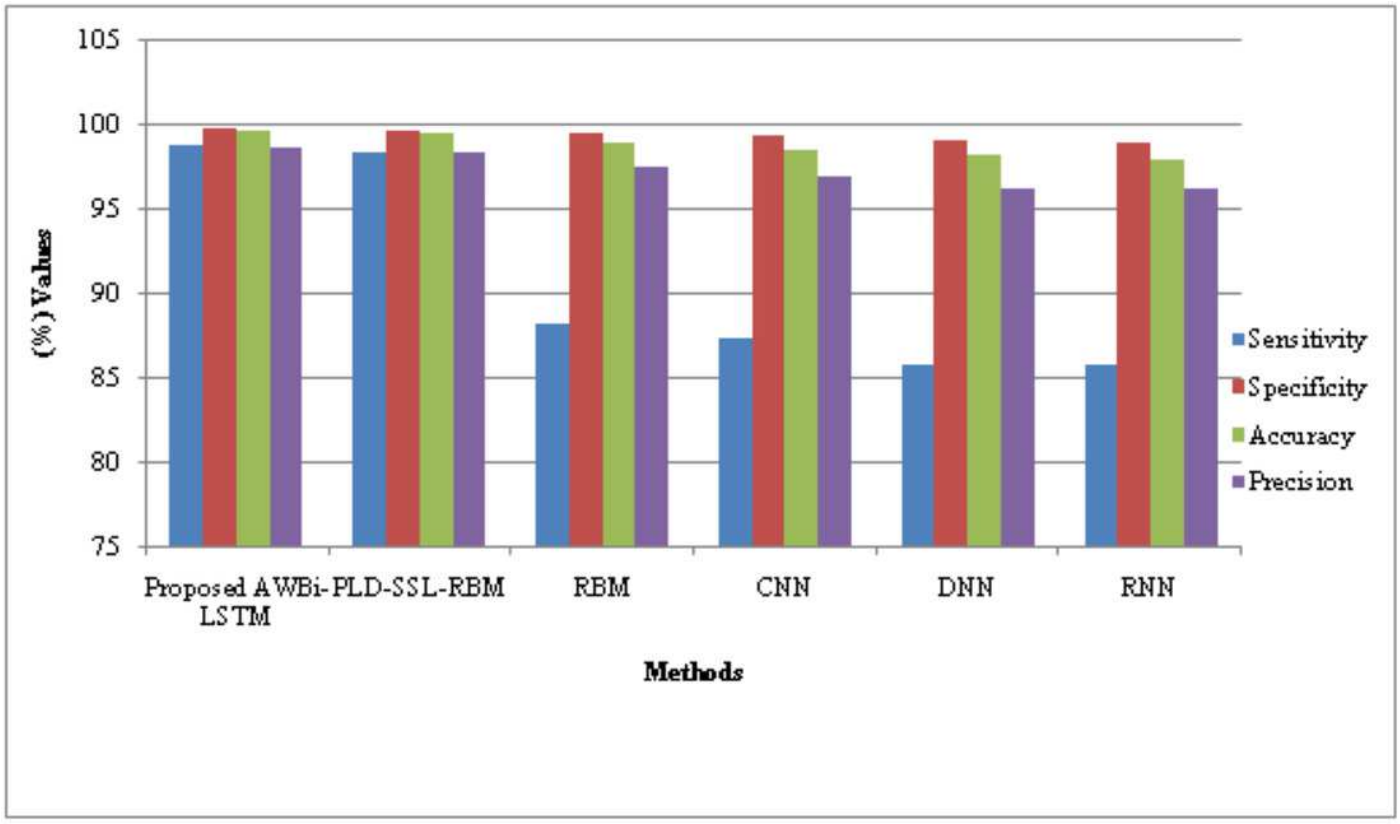

Figure 5

CONFUSION MATRIX FOR THE PROPOSED AWBi-LSTM