



**Manchester
Metropolitan
University**

G, Thippa Reddy and Bhattacharya, Sweta and Maddikunta, Praveen Kumar Reddy and Hakak, Saqib and Khan, Wazir Zada and Bashir, Ali Kashif and Jolfaei, Alireza and Tariq, Usman (2020) Antlion re-sampling based deep neural network model for classification of imbalanced multimodal stroke dataset. Multimedia Tools and Applications. ISSN 1380-7501

Downloaded from: <https://e-space.mmu.ac.uk/626646/>

Version: Accepted Version

Publisher: Springer Science and Business Media LLC

DOI: <https://doi.org/10.1007/s11042-020-09988-y>

Please cite the published version

<https://e-space.mmu.ac.uk>

Antlion re-sampling based deep neural network model for classification of imbalanced multimodal stroke dataset

Thippa Reddy G¹ · Sweta Bhattacharya¹ · Praveen Kumar Reddy Maddikunta¹ · Saqib Hakak² · Wazir Zada Khan³  · Ali Kashif Bashir⁴ · Alireza Jolfaei⁵ · Usman Tariq⁶

Received: 21 March 2020 / Revised: 20 July 2020 / Accepted: 24 September 2020 /

Published online: 09 October 2020

© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Stroke is enlisted as one of the leading causes of death and serious disability affecting millions of human lives across the world with high possibilities of becoming an epidemic in the next few decades. Timely detection and prompt decision making pertinent to this disease, plays a major role which can reduce chances of brain death, paralysis and other resultant outcomes. Machine learning algorithms have been a popular choice for the diagnosis, analysis and predication of this disease but there exists issues related to data quality as they are collected cross-institutional resources. The present study focuses on improving the quality of stroke data implementing a rigorous pre-processing technique. The present study uses a multimodal stroke dataset available in the publicly available Kaggle repository. The missing values in this dataset are replaced with attribute means and LabelEncoder technique is applied to achieve homogeneity. However the dataset considered was observed to be imbalanced which reflect that the results may not represent the actual accuracy and would be biased. In order to overcome this imbalance, resampling technique was used. In case of oversampling, some data points in the minority class are replicated to increase the cardinality value and rebalance the dataset. transformed and oversampled data is further normalized using Standardscalar technique. Antlion optimization (ALO) algorithm is implemented on the deep neural network (DNN) model to select optimal hyperparameters in minimal time consumption. The proposed model consumed only 38.13% of the training time which was also a positive aspect. The experimental results proved the superiority of proposed model.

Keywords Deep neural networks · Antlion optimization · Stroke prediction · Re-sampling · Imbalanced dataset

✉ Wazir Zada Khan
wazirzadakh@jazanu.edu.sa

1 Introduction

The statistical report of WHO (World Health Organization) has identified stroke to be one of the predominant causes of disability in the world wherein an estimated 17 million people succumb to death, being a victim of heart disease and strokes. The primary reasons for individuals getting affected by heart diseases, almost giving it the status of an epidemic are physical inactivity, unhealthy and irregular lifestyle and tobacco smoking. In United States it is the enlisted within the top five reasons of death across advanced aged male and females every year. Naturally, there is a proportional increase in the medical expenses of an estimated 23 billion dollars as per the reports of 2014 [2]. Hence it is extremely important to find suitable and accurate technical solutions to predict the possibilities of this near fatal disease contributing towards control of per-capita cost of medicine and therapeutics.

The present treatments and medical interventions for investigation of ischemic strokes focus on reperfusion of ischemic tissues using intravenous medications and vascular techniques to remove the obstructions in the blood flow [16]. The detection are primarily dependent on neuroimaging and magnetic resonance imaging [39]. Machine learning algorithms have been implemented for early diagnosis personalized treatments, remote monitoring of patients [7] and prompt decision making in acute ischemic strokes and other diseases [19, 36] where time plays an extremely important factor. Cardiac strokes and its relevant areas of research are vast and machine learning algorithms have played significant roles in various spectrum of this disease. As an example, machine learning models have been used in conclusive diagnosis from medical images, estimation of the onset of heart attacks, analysis of cerebral edema, predictions of complications and post treatment results [3, 4, 35, 37, 43]. Also it becomes crucial to highlight that most of the stroke patients face motor deficits after their incidents and machine learning algorithms have also been implemented to predict the possibilities of such outcomes based on analysis of structural medical resonance images (MRI) of the brain and heart. Machine learning and IoT based applications [14, 22, 30] have become extremely predominant in all sectors of human life wherein devices are becoming smart, secure and available in hand held mobile devices. It is obvious that such implementations will become a necessity in the healthcare sector as well.

It is known that prevention is better than cure. The ideal approach thus would be to detect possibilities of stroke early enough, to prevent patients becoming a victim in the first place. This accelerated detection would reduce chances of brain damage and paralysis thereby enabling patients to regain complete mobility and agility to lead a normal life. Although the results of machine learning [9, 13, 27, 42] implementations are promising but there are associated challenges. Firstly, there exist limitations related to sample sizes and quality of data in the datasets [6]. Also, these data are collected from various Institutions and collation of inter-institutional data lead to data imbalances [15, 17, 25, 38]. Hence, the primary motivation for this study was to identify and incorporate the best data pre-processing methodologies and then use the processed data for training of the machine learning model. Deep Neural Networks (DNN) usually require huge amount of time for training. A major amount of time is wasted by machine learning (ML) practitioners and researchers in finding the optimal parameters for DNN [5]. The identification of the best algorithm for selection of optimal hyperparameters in deep neural network acted as the second challenge due to the fact that the existing algorithm used excessive computational time to perform the same job. The identification of the best algorithm acted as a motivational factor to select optimal hyperparameters in the deep neural network ensuring minimal time consumption. The computational complexities involved in solving real-world problems are also quite high which are always not within the scope of conventional methods wherein majority of them are based

on mathematical optimization algorithms. These algorithms depend on various assumptions related to the problem to fit in to a particular method and hence lag flexibility to model the problem close enough to reality. Thus conventional algorithms include limitations which fail to be suitable for a broader spectrum of real-time problem solving. Optimization techniques and algorithms inspired by nature play significant role in solving such practical problems and have been used with the same objective in the present study as well.

Based on the above mentioned motivational factors, the present study focuses on:

1. Eradication of imbalances and heterogeneity in the multimodal stroke dataset collected from the publicly available Kaggle repository using re-sampling method.
2. Replacement of missing values in the dataset by attribute mean and application of LabelEncoder technique in Python for achieving homogeneity
3. re-sampling of the transformed data to eliminate imbalances followed by application of StandardScaler technique for normalization
4. Implementation of Antlion optimization algorithm in the DNN model to ensure optimized choice of hyper-parameters in limited consumption of time.

The success of a deep learning model is highly dependent on the data being used and also the algorithm being implemented. The advantages of the proposed revolve around these two factors primarily. Firstly, the model includes an extremely meticulous pre-processing method which fills all missing values, performs data transformation, re-samples and finally normalizes the data. Secondly the model uses the Antlion optimizer which simulates the hunting characteristics of natural ants. The optimizer includes the basic nature of ants in hunting the prey, involving five steps - random walk, building of traps, trapping of ants in the traps, catching of the prey and finally rebuilding the traps once again. The algorithm is capable of solving constraint problems having diversified search spaces due to its optimized design. It is a gradient free algorithm which visualizes problems as a blackbox and therefore has applicability in solving real-time problems. Considering all these advantages, the output of this framework is expected to yield more accurate prediction results in comparison to the existing machine learning (ML) models for early detection of stroke and heart diseases.

Rest of the manuscript is organized as follows. Related work is presented in Section 2. The preliminaries, background algorithms and the proposed methodology are discussed in Section 3. The experimental results are presented in Section 4. The manuscript is concluded in Section 5.

2 Literature survey

To highlight the dire necessity of solving the problem as mentioned in the paper, an explicit literature review was conducted by exploring related studies in predicting heart disease using machine learning algorithms and state-of-the-art techniques in the area of imbalanced data learning.

In the area of predicting heart-diseases using machine learning algorithms, immense work is going on and few of the potential studies include the work of [18] where the authors have proposed an integrated machine learning-based feature selection and risk prediction algorithms for Cardiovascular stroke prediction. The feature selection algorithm is based on conservation mean. On the other hand, the stroke prediction algorithm is based on Margin-based Censored Regression and SVM. For the evaluation and comparison of the proposed stroke prediction approach with other existing approaches (i.e., Cox proportional hazards model) 5-year CHS (Cardiovascular Health Study) cardiovascular disease dataset is used.

The experimental results show that the proposed approach predicted the stroke with 77% accuracy. However, the proposed approach is evaluated only using the CHS dataset, so possibilities of achieving the same performance with lower computational cost using other available healthcare datasets, still need to be verified.

P. Chantamit-o-pas et al. [10] analyzed and compared Support Vector Machine, Naive Bayes and the deep learning (DL) technique for stroke prediction. The data of heart patients is taken from the UCI Machine learning repository; with 899 records and 76 attributes; is used by these techniques to identify risk factors and to predict the disease. Ten attributes related to a stroke risk factor were selected for training the models. The comparison revealed that DL outperformed Naive Bayes and SVM algorithms for stroke prediction.

P. Chantamit-o-pas et al [11]. have investigated two deep learning algorithms - Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) for stroke prediction on healthcare records. The study applied DL algorithms on Electronic Healthcare Records (EHR) of patients with cerebrovascular disease for predicting stroke. The algorithm LSTM-RNN is implemented and then the ability of LSTM-RNN is evaluated to recognize patterns in multi-label classification of cerebrovascular symptoms or stroke. The results proved the efficiency of LSTM-RNN when compared to other state-of-the-art algorithms.

G. Manogaran et al [23]. have proposed a DL-based method called Multiple Kernel Learning with Adaptive Neuro-Fuzzy Inference System (MKL with ANFIS) for diagnosing heart disease. The proposed method has two steps; In the first step, the parameters were chosen by MKL method. In the second step, the heart disease and healthy patients are classified by feeding the data from MKL method to the ANFIS classifier. The authors have used the KEGG Metabolic Reaction Network dataset for evaluating the performance of the proposed model. The results have demonstrated that the proposed model has achieved better results when compared to that of existing methods.

G. Thippa Reddy et al [29]. have proposed a technique for predicting heart disease called the Rule-Based Fuzzy Logic (RBFL) prediction algorithm. The proposed technique works in two steps; in the first step, feature reduction is performed using Locality Preserving Projection (LPP) algorithm to recognize the related attributes that will reduce the number of features and remove unnecessary or noisy information. In the second step, first, the fuzzy rules are produced from the Firefly Bat (FFBAT) algorithm and then these rules are applied to classify the heart disease. The authors have performed experiments to evaluate their proposed technique by using the publicly available UCI datasets, i.e., Cleveland, Hungarian, Switzerland datasets. The results have shown that the proposed RBFL prediction algorithm outperformed the existing approaches by attaining the accuracy of 76.51%.

Similarly, the state-of-the-art approaches in addressing issues relevant to imbalanced data learning include the work of [33], where the authors have proposed a hybrid model to improve the performance of the ensemble of the classifier for imbalanced data. There are two phases in the proposed model. In phase one, the re-sampling of imbalanced data is done by applying the Synthetic Minority Over-sampling Technique (SMOTE) for solving the over-sampling issue and Random under-sampling technique for under-flow. Finally, the pre-processed data is fed to an ensemble of classifiers using the Weka tool. Eight imbalanced datasets were chosen for experimental purposes with imbalance ratios from 2 to 80. The proposed model has shown significant performance compared to ensemble classifiers that were trained using different datasets. Similarly, in [44], the authors proposed a regularized ensemble framework based on deep learning to tackle the issue of imbalanced datasets. The proposed method basically uses undersampling based approach to recover balance among classes and is evaluated on 11 synthetic and real-world datasets (with moderate-high imbalance ratio). The proposed approach has achieved the maximum improvement of about

24.7% compared to benchmark studies. The authors in [47] proposed the algorithm CWsRF (class weights voting) based on random forest to address the imbalance dataset problem in the medical domain. The algorithm consist of three phases i.e. building RF model, building CWsV (where different weights per class and votes are calculated) and classifying votes. For evaluation purposes, five datasets were used mostly related to physiological signals and breast cancer. The authors in [20] presented the notion of swarm fusion to address the imbalance dataset problem. Two swarm optimization algorithms were used in this work where the focus of former one is to rectify exceeding majority data instances and the later one corrects the shortage of minority data instances. The proposed approach was evaluated on 30 public datasets and outperformed the benchmark studies by 13-69%.

In one of the latest research studies [45], the authors have explored the deep belief network (DBN), one of the popular machine learning technique used in classification tasks. As DBN yields inferior output when it comes to imbalanced data classification, the authors used adaptive differential evolution optimization algorithm to improve the DBN network performance. The proposed approach was evaluated on 58 datasets and successfully generated promising results. It is thus evident from all of the above-mentioned studies that imbalance data learning is still an emerging research challenge. The consolidated review of the existing studies is presented in Table 1.

3 Preliminaries and proposed architecture

In this section the Antlion Optimization algorithm and the proposed architecture are discussed.

3.1 Antlion optimization

Several nature inspired algorithms like genetic algorithm, cuckoo search, firefly, BAT, etc. and soft computing techniques have been extensively used for several tasks in machine learning process like dimensionality reduction, choosing the optimal parameters in training the classifiers, prediction algorithms, etc. [8, 26, 31, 34]. Due to advent of Internet of Things, smart cities, advanced medical devices, treatment through remote monitoring[1], etc., huge amount of data is being generated these days [40, 41]. In order to reduce the complexity of the machine learning and deep learning algorithms there exists a need for dimensionality reduction techniques which choose optimal dimensions for training the algorithms [28]. Antlion Optimization (ALO) algorithm is one of the recently developed nature-inspired algorithm, which is based on the characteristics of ants and antlions, resembling their hunting technique [12]. The antlions simply sit under soil pits waiting to catch their prey using a small cone shaped trap as naturally gifted by the nature. The process of catching prey in ants constitute of five stages - firstly the walking of ants following random pattern. Then, building of efficient traps and thereby catching the ants in the traps. This is followed by captivation of preys, consuming, discarding of leftovers and rebuilding of the traps which again commences the next hunting process. The hunting process of antlion is as follows:

The (1) showcases the ant's random walk.

$$R(W) = [0, \text{cums}(2s(w_1) - 1), \dots, \text{cums}(2s(W) - 1)] \quad (1)$$

where cums measures its cumulative sum, the step of a random walk is w , the total iteration is W , and the stochastic function is $s(w)$, as shown in the (2).

$$s(w) = \begin{cases} 1 & \text{if random value} > 0.5 \\ 0 & \text{if random value} \leq 0.5 \end{cases} \quad (2)$$

Table 1 Consolidated Review of the Existing Studies

Reference	Dataset	Methodology	Evaluation Metrics	Research Challenges
[18]	Cardiovascular Health Study (CHS) dataset	Automatic feature selection using conservative mean, Support Vector Machines (SVMs), Margin-based censored regression algorithm	AUC and ROC	Considering other Evaluation metrics, Application on larger dataset
[10]	Heart Disease dataset from UCI machine learning repository	Back propagation learning network (BPN), Deep Learning, Naive Bayes, SVM	Mean Value, Standard Deviation	Comparison with state of the art techniques
[11]	Electronic Health Record, Medical Services, The Ministry of Public Health of Thailand	Recurrent Neural Network (RNN) with Long Short Term Memory (LSTM)	Accuracy, Recall, F1 Score	Consideration of risk factors and real-time lab datasets
[46]	Medical Image Datasets	Review of Machine Learning Techniques in Image Datasets		
[23]	Heart Disease Dataset from private source	Multiple Kernel Learning (MKL) and Adaptive Neuro-Fuzzy Inference System (ANFIS)	Specificity, Sensitivity, MSE	Hybrid Classification model could be implemented and evaluated
[29]	Heart Disease Dataset from UCI machine learning repository	RBFL Prediction Algorithm	Accuracy	Use of real time datasets and hyper-parameter optimizers
[33]	Eight imbalanced dataset from KEEL repository	Re-sampling, SMOTE, Bagging Classifier and Stacking Classifier	AUC, Sensitivity, Specificity	Various other forms of pre-processing not considered
[44]	Capsule endoscopy video of bowel cancer symptoms and synthetic datasets	Ensemble framework of Deep Learning	Accuracy	Applicability on larger datasets
[47]	Four datasets from UCI Machine Learning Repository, One dataset from Zhejiang	CWwRF - Class Weights Random Forest Algorithm	Accuracy, AUC	Implementations on multi-classification problems and ensemble learning using other algorithms could be considered
[20]	30 Public Datasets	Multi-objective Swarm Fusion Algorithm	Accuracy	Implementation on larger real-time datasets

During the computation process, the position of the ants is represented in the form of a matrix defined in (3).

$$M_A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \dots & \dots & \dots & \dots \\ A_{m1} & A_{m2} & \dots & A_{mn} \end{bmatrix} \quad (3)$$

where M_A represents the ant position, $A_{x,y}$ represents x^{th} ant at y^{th} dimension, m represents total number of ants in a search space, and n represents total number of variables.

Ants update their position by taking a random walk. To check if the ants are moving within the search space min-max normalization is applied, which is shown in the (4).

$$A_x^w = \frac{(A_x^w - m_{r_x}) * (n_x - k_x^w)}{n_x^w - m_{r_x}} + k_x \quad (4)$$

where m_{r_x} is denoted as minimum random walk of the x^{th} ant, k_x^w represents the minimum random walk of the x^{th} ant at w^{th} iteration, and n_x^w denotes the maximum random walk of the x^{th} ant at w^{th} iteration.

3.2 Sliding ants towards antlion

A roulette wheel is being used to model the hunting behavior of the antlions. During this process, the best antlion is chosen on the basis of its fitness value. The antlions simply sit under soil pits waiting to catch their prey using a small cone shaped trap. The sliding of ants towards antlion is shown in the (5), (6)

$$k^w = \frac{k^w}{I} \quad (5)$$

$$n^w = \frac{n^w}{I} \quad (6)$$

where k^w indicates the minimum random walk of all ants at w^{th} iteration, n^w represents the maximum random walk of all ants at w^{th} iteration, and I is denoted as the ratio which is shown in (7)

$$I = 10^z \frac{w}{W} \quad (7)$$

where w is denoted as the current iteration, W denotes the maximum number of iterations, z is a constant.

3.3 Trapping in antlion's traps

The random walk of the ant changes dynamically through the selected antlion trap. The change of the random path of the ants to the location of the antlion is represented using (8) and (9).

$$k_x^w = AL_y^w + k^w \quad (8)$$

$$n_x^w = AL_y^w + n^w \quad (9)$$

where k and n are vectors around a selected antlion. AL_y^w represents the position of the selected y^{th} antlion at w^{th} iteration.

3.4 Hooking the prey

This is the final phase of the hunting process, during which the antlion kills the prey and consumes. Update the position of the antlion with the ant if the fitness function of the ant is greater than the antlion using (10).

$$AL_y^w = A_x^w \left\{ \text{if } FF(A_x^w) > FF(AL_y^w) \right. \quad (10)$$

A_x^w denotes the position of x^{th} ant at w^{th} iteration.

3.5 Elitism

The best solution achieved during optimization is considered to be an elite solution. The random walking of the ant is affected by the movement of the antlion and the best antlion (elite). The position of the ant is therefore taken as the average random walk of the antlion and elite represented in (11).

$$A_x^w = \frac{P_A^w + P_E^w}{2} \quad (11)$$

where P_A^w denotes the random walk of ant around the selected antlion at w^{th} iteration, P_E^w denotes the random walk of ant around the elite antlion at w^{th} iteration, and A_x^w denotes the position of x^{th} ant at w^{th} iteration.

Algorithm 1 Conventional ALO Algorithm [24].

```
1 Generate the population of both ant and antlion.
2 Calculate the fitness values of both ant and antlion.
3 The best antlion is considered to be the elite.
4 while if end criterion does not meet do
5   for every ant do
6     | Select an antlion via Roulette wheel
7     | Update  $k$  and  $n$  using (5) and (6)
8     | Create a random walk using (1) and normalize using (4)
9     | Update the position of ant using (11)
10  end
11  Calculate the fitness function of all ants in the search space
12  if  $FF(A_x^w) > FF(AL_y^w)$  then
13    | Replace the position of antlion with ant using (10)
14  end
15  if Fitness Function of antlion is better than elite then
16    | Update elite
17  end
18 end
19 return elite
```

3.6 Proposed architecture

The proposed architecture is presented in Fig. 1. The proposed methodology kicks off by loading the multimodal stroke dataset collected from the publicly available data repository of Kaggle collected from a private source [32]. The inclusion of the multimodal attributes

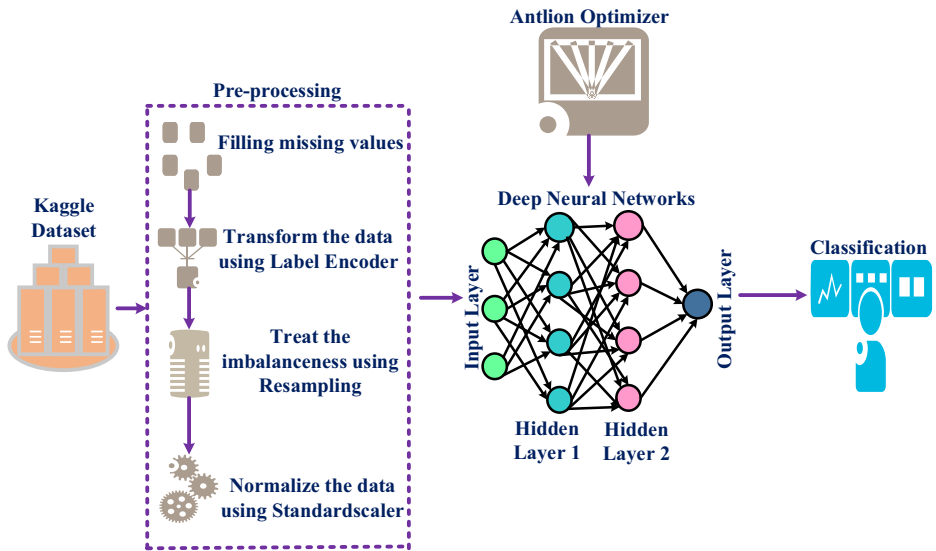


Fig. 1 Proposed Architecture

helps to predict and make considerable progress in understanding the research problem. It is obvious that the dataset is quite raw to be directly used for the training of machine learning algorithms. In this case, a rigorous pre-processing has to be performed to fill in the gaps pertaining to missing values, imbalances, heterogeneity in the existing raw dataset. As the first step, the missing values in the dataset are replaced by attribute mean. The next step involves transformation of the data values for all the attributes to numeric form using LabelEncoder technique available in Python is used. The transformed data is then subjected to re-sampling method to combat the imbalances existing in the dataset. Finally in the pre-processing stage, StandardScaler is used to normalize the data to a range of 0 to 1. This pre-processed data is fed to the deep neural network (DNN) model for further analysis. The objective of achieving accurate predictions from the DNN model depends on the choice of optimal hyper-parameters namely - selection of accurate number of layers in the DNN, selection of the optimized number of neurons for each layer in the DNN, use of appropriate activation function for optimization of the layers, use of appropriate optimization algorithm for the network, number of epochs and finding the most optimized learning rate. In the proposed study Antlion optimization algorithm is used to choose the optimal hyper-parameters. The Antlion algorithm efficiently explores the search domain using random selection of agents and the concept of arbitrary walking found typically in an ant colony. These unique feature of the algorithm help in achieving optimal hyperparameters ensuring reduced time complexity and better prediction results.

The steps involved in the proposed model are given below:

1. Load the Stoke Dataset with modalities from Kaggle
2. Pre-processing
 - (a) Fill the missing values by Attribute Mean
 - (b) Transform the data by LabelEncoder
 - (c) Treat the imbalances in dataset using re-sampling method

- (d) Normalize the data using StandardScaler
3. Use Antlion optimization algorithm for selecting optimal hyperparameters for Deep Neural Networks.
- (a) Initialize the population of both ant and antlion.
- (b) During optimization the fitness values of each ant are saved in the form of (12).

$$M_{FA} = \begin{bmatrix} F([A_{11} & A_{12} & \dots & A_{1n}]) \\ F([A_{21} & A_{22} & \dots & A_{2n}]) \\ \dots & \dots & \dots & \dots \\ F([A_{m1} & A_{m2} & \dots & A_{mn}]) \end{bmatrix} \quad (12)$$

where M_{FA} represents the fitness of each ant and F is the objective function. (13), (14) represents the position and fitness of the antlions.

$$M_{AL} = \begin{bmatrix} AL_{11} & AL_{12} & \dots & AL_{1n} \\ AL_{21} & AL_{22} & \dots & AL_{2n} \\ \dots & \dots & \dots & \dots \\ AL_{m1} & AL_{m2} & \dots & AL_{mn} \end{bmatrix} \quad (13)$$

$$M_{FAL} = \begin{bmatrix} F([AL_{11} & AL_{12} & \dots & AL_{1n}]) \\ F([AL_{21} & AL_{22} & \dots & AL_{2n}]) \\ \dots & \dots & \dots & \dots \\ F([AL_{m1} & AL_{m2} & \dots & AL_{mn}]) \end{bmatrix} \quad (14)$$

- (c) Create a random walk of ant using (1) and normalize the random walk with in the search space using (4).
- (d) The sliding of ants towards antlion is shown in the (5), (6).
- (e) Trapping ant in antlion's traps: The change of the random path of the ants to the location of the antlion is represented using (8) and (9).
- (f) Update the position of ant towards antlion and elite using (11) and catch the prey
- (g) Update the position of the antlion with the ant if the fitness function of the ant is greater than the antlion using (10).
- (h) Calculate the Fitness Function using

$$Minimize : FF = \frac{M_e}{M} \quad (15)$$

M_e :Number of misclassified samples

M :Total number of testing samples

4. Train the pre-processed data using Deep Neural Networks.
5. Evaluate the performance of the DNN model using Precision, Recall, F1-Score, Accuracy, Specificity, Sensitivity measures
6. Compare the proposed model with DNN, Naïve Bayes, Decision Tree, Random Forest, SVM and XGBoost Machine Learning Algorithms.

The next section discusses the experimentation results of the proposed work. Also the proposed model is compared with other models to prove the efficiency of the proposed model.

4 Results and discussion

The experimentation is carried out in a personal laptop with Windows 10 Operating System having a RAM of 8GB, Hard Disk of 500 GB and Python 3.7 is used as the programming language for implementation. The dataset used in this work is a “Multimodal Healthcare Dataset Stroke Data [32]” collected from the renowned Kaggle Repository.

The dataset has 43400 records and 12 multimodal attributes of patients namely, ID, Gender, Hypertension, Whether the patient suffers from heart disease or not, if the patient is ever married, type of work done by the patient, residence type, average glucose level, body mass index, smoking status, and stroke (if the patient has ever had a stroke or not).

4.1 Measures used for evaluating the model

The following are the measures used to evaluate the proposed model.

1. Accuracy: Accuracy refers to the ability of the classifier to predict the class label or attribute accurately for new data values. It derived by computing the ratio of correct predictions to the total number of predictions made for the input values.

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{total number of predictions made}} \quad (16)$$

2. Precision: In machine learning and data mining, False Positives are instances wherein the model inaccurately labels a negative case as a positive one. The value of precision is derived by computing the ratio of True Positives to the total number of True Positives and False Positives.

$$Precision = \frac{\text{True positives}}{\text{True positives} + \text{False Negatives}} \quad (17)$$

3. Recall: The recall value helps us to calculate the total number of actual Positive cases that the model identifies and labels as True Positive. The value of recall is derived by computing the ratio of true positive to the total number of True Positive and False Negative

$$Recall = \frac{\text{True positives}}{\text{True positives} + \text{False Negatives}} \quad (18)$$

4. F1 Score: The F1 score is used to measure the accuracy of the test conducted and is derived by computing the harmonic mean of the precision and recall values.

$$F1 = 2 * \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} \quad (19)$$

5. Sensitivity: Sensitivity denotes the proportion of actual positives that are predicted as True Positives. It is derived the computing the ratio of False Positive to the total number of False Positive and True Negative.

$$Sensitivity = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}} \quad (20)$$

6. Specificity: Specificity denotes the proportion of actual negatives that are predicted as True Negatives. It is derived by computing the ratio of True Positives to the total number of False Negative and True Positive.

$$Specificity = \frac{\text{True Positive}}{\text{False Negative} + \text{True Positive}} \quad (21)$$

4.2 Preprocessing

The first attribute in the stroke dataset, i.e., ID of the patient does not have any significant effect on the possibility of a patient having suffered with stroke or not. and is thus eliminated. The resultant dataset now has 11 attributes to be considered for further processing.

Several instances in the dataset have missing values which result in negative effect on the classification accuracy. These missing values in the dataset have thus been replaced with attribute mean. The dataset considered also has severe imbalances and some of the attributes have categorical values. It is a known fact that Machine learning algorithms fail to process categorical data. Hence the categorical attributes in the stroke dataset need to be transformed to numerical values. For this purpose, a LabelEncoder is used to convert all the attributes into to numerical form and the imbalance is treated using re-sampling.

As can be observed from Fig. 2, number of instances having the value of class label (stroke) as “0” are 42617 and the number of instances with class label “1” are mere 783. If machine learning algorithms are implemented on this imbalanced dataset, biased results are likely to be generated invariably. re-sampling technique has two variants. In the first variant, number of attributes which have higher presence of particular values are reduced to match with the instances of values having lesser presence. This technique is called “Under-sampling”. The other technique is “Over-sampling” wherein the instances with lesser presence of particular values will be randomly duplicated to match the other instances. In the present work, over-sampling method is used. Using over-sampling, instances with values of class label “1” have been increased to match with that of instances with values of class label “0”. After re-sampling, the number of instances with values of class label “1” and class label “0” both become balanced with 42617 instances each. The results of re-sampling are depicted in Fig. 3.

The resultant balanced dataset is then normalized using StandardScaler method available in Python. This method converts all the values in the dataset to a range of 0 to 1. The purpose of normalization is to make every attribute important by giving each attribute an equal weightage. The next subsection discusses the results of experimentation on this pre-processed data.

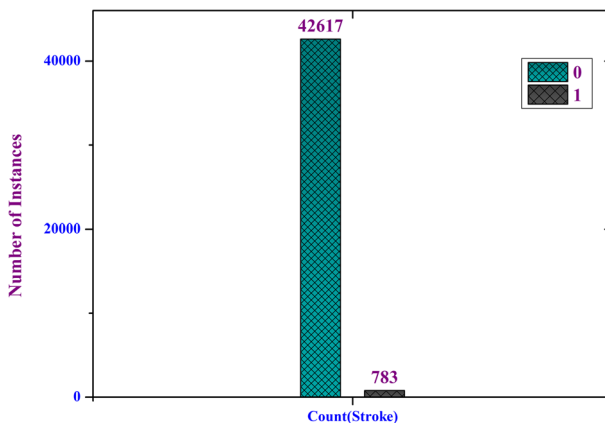


Fig. 2 Distribution of instances of class label (Imbalanced Data)

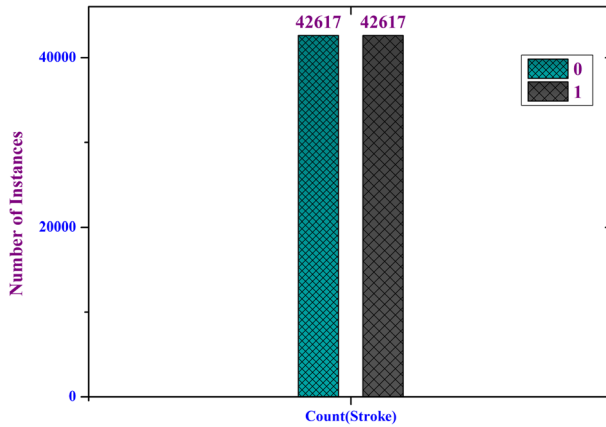


Fig. 3 Distribution of instances of class label (Balanced Data using re-sampling)

4.3 Performance evaluation of the proposed model

In this section the experimentation results are discussed. At the outset, the stroke dataset without oversampling is classified by Deep Neural Networks (DNN). Then the dataset after re-sampling is experimented using Deep Neural Networks. Antlion optimization algorithm is used in this study to choose the optimal hyperparameters namely the activation functions at each layer, optimal number of layers in the DNN model, number of neurons to be used in each layer, optimization function for the DNN, learning rate of the DNN network and the number of epochs to train the DNN network. The results achieved by DNN integrated with Antlion Optimization (ALO) algorithm with and without re-sampling are depicted in Fig. 4.

It is observed from the Fig. 4 that the Specificity of DNN-Antlion applied on imbalanced dataset is almost negligible due to the lower number of records with values of class label

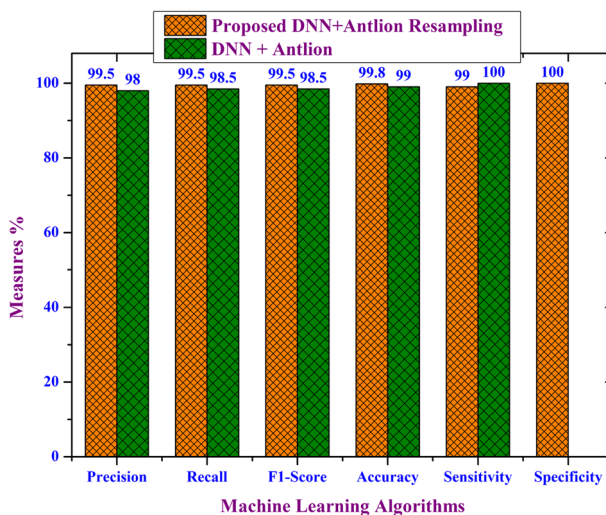


Fig. 4 Performance evaluation of DNN + Antlion based Models

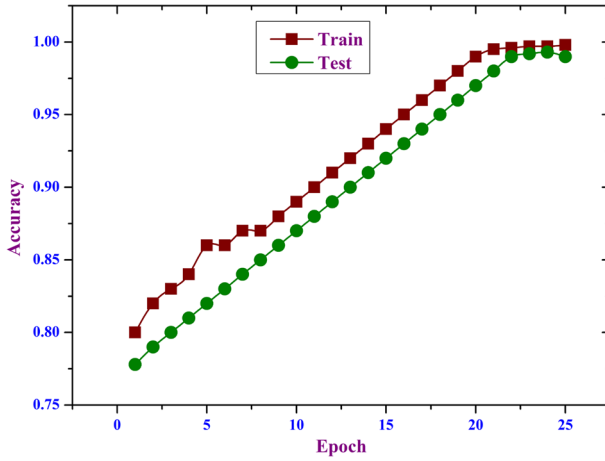


Fig. 5 Training Vs Testing Accuracy of the Proposed Model

“1”. When the dataset is resampled and balanced, the Specificity of DNN-ALO enhances to 100 percent. The figure also highlights the fact that DNN-ALO on balanced data yields better results considering other performance metrics.

The training and testing accuracy, error rates of DNN on resampled dataset are depicted in Figs. 5 and 6. These figures reveal that the training and testing accuracy of the proposed model gradually increases after each epoch in contrast to the error rate which gradually decrease after each epoch.

The dataset without re-sampling and with re-sampling are then experimented with other popular machine learning (ML) algorithms like DNN, Naïve Bayes, Decision Tree, Random Forest, SVM and XGBoost classifiers. The results of these experimentation are depicted in Figs. 7, 8, 9, 10, 11, 12. The Figs. 7–12 depict the capability of re-sampling to almost

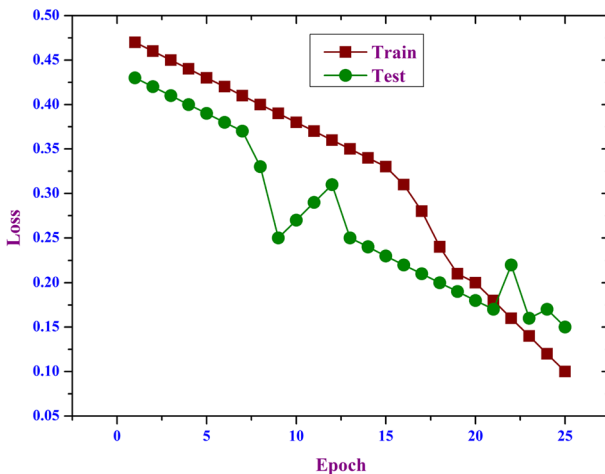


Fig. 6 Training Vs. Testing Error Rate of the Proposed Model

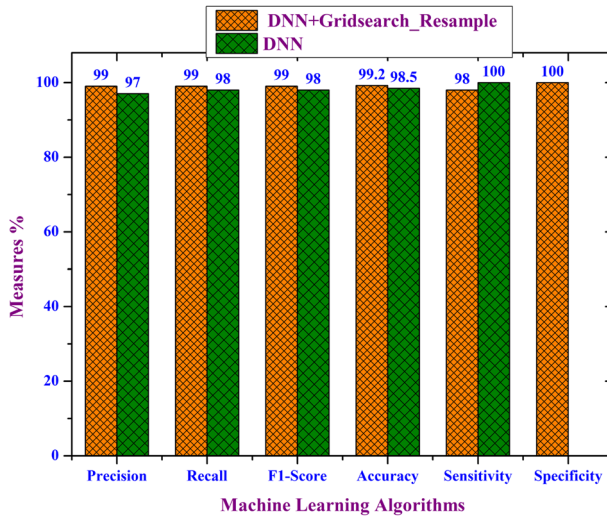


Fig. 7 Performance evaluation of DNN based Models

eliminate biased results thereby improving specificity of these machine learning (ML) algorithms. From Fig. 13 it is clear that the proposed model has a better convergent rate when compared to the existing models Artificial Bee Colony (ABC), Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Gravitational Search Algorithm (GSA). Figure 14 represents the comparative analysis of time consumption (in seconds) for experiments conducted using GridSearch and Antlion Optimizer (ALO) algorithm for selecting optimal hyperparameters. It is evident from the pie chart that the Antlion optimizer algorithm selects the best hyperparameters in a considerably minimal amount of time in comparison to the GridSearch algorithm. The ability of the Antlion algorithm to efficiently explore the search domain

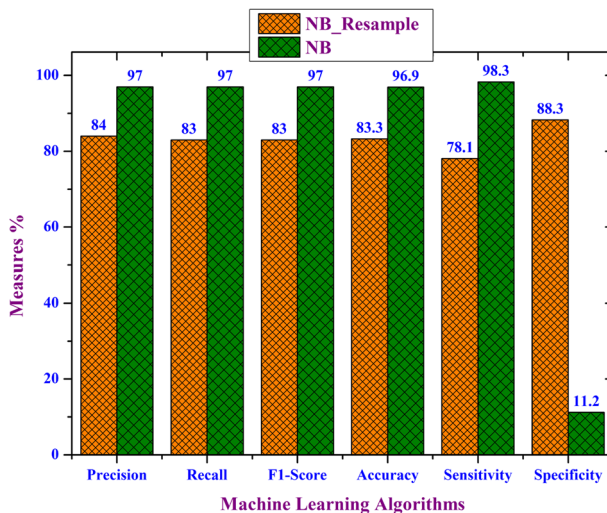


Fig. 8 Performance evaluation of Naïve Bayes based Models

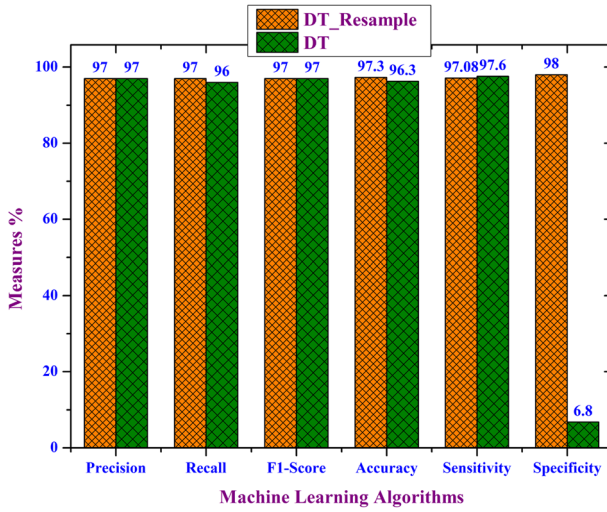


Fig. 9 Performance evaluation of Decision Tree based Models

using random selection of agents incorporating the concept of arbitrary walking, found typically in an ant colony, have contributed towards reduction in the time complexity. The proposed model is a nonlinear method that performs arbitrary operations, since from a stochastic viewpoint it is not feasible to perform a complex analysis. However, an understanding of this complexity can be obtained through Big O notation. The computational capacity is calculated using $O(n * Training\ Time)$ to find the optimal solution for the proposed model which is shown in Fig. 15.

The comparative analysis of the performance of proposed model with other models is depicted in Table 2. We compared our work with the latest work on stroke prediction [21],

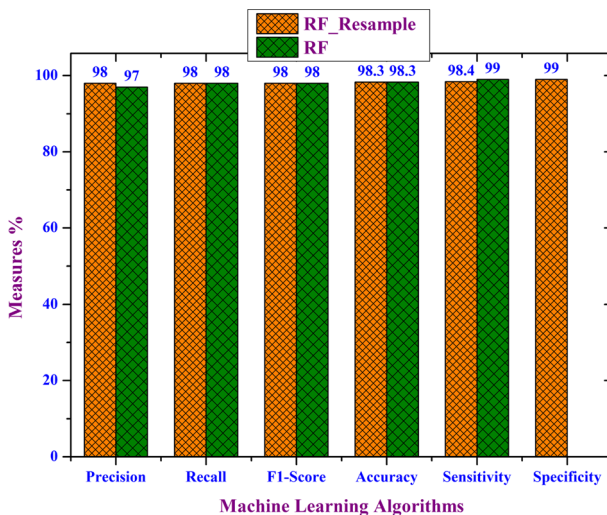


Fig. 10 Performance evaluation of Random Forest based Models

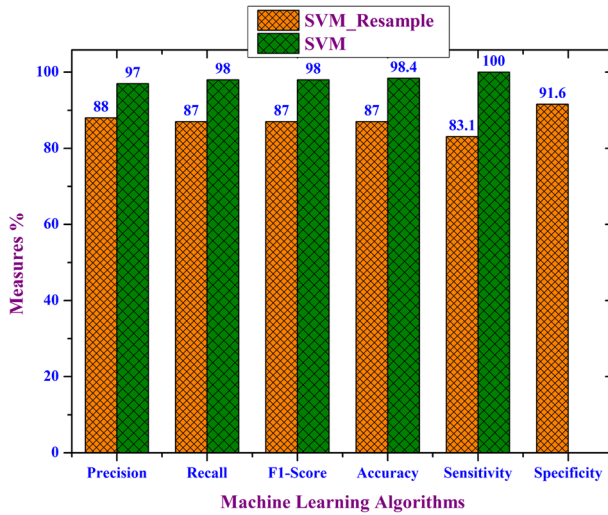


Fig. 11 Performance evaluation of SVM based Models

in which the authors used automated hyperparameter optimization (AutoHPO) using DNN and achieved 33.1% false positive rate, 71.6% accuracy and 67.4% sensitivity. The experimental result analysis shows that the proposed model outperforms [21]. Table 3 explains a comparative analysis of the proposed model vs. DNN with the meta - heuristic optimization models. The proposed Antlion algorithm selects the best hyper-parameters in a considerably minimal amount of time compared to other meta-heuristic algorithms. The ability of the Antlion algorithm to efficiently explore the search domain using random selection of agents incorporating the concept of arbitrary walking, typically found in an ant colony, has contributed to increasing the performance of the system.

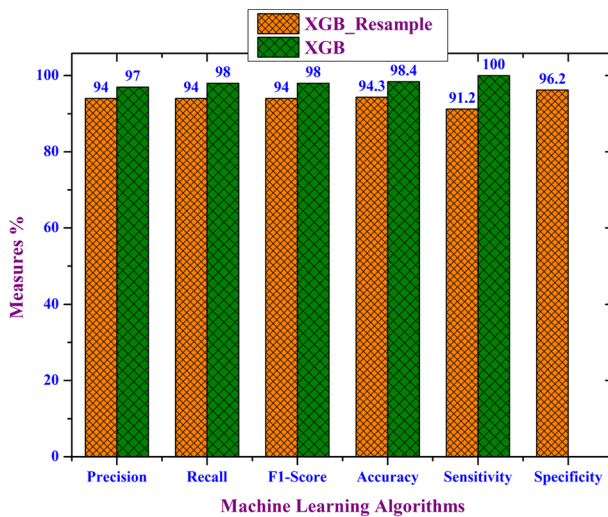


Fig. 12 Performance evaluation of XGBoost based Models

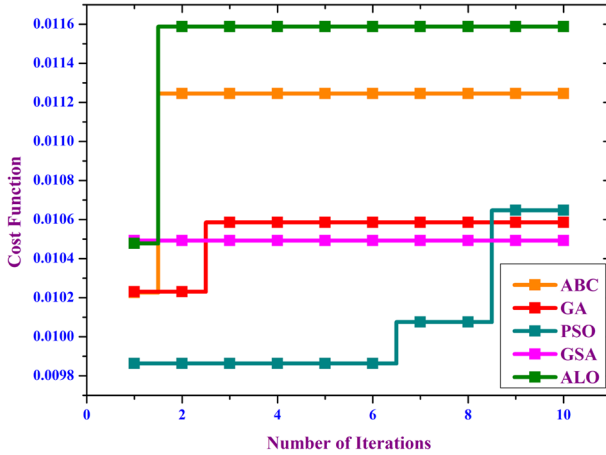


Fig. 13 Convergence Analysis

In order to further analyse the impact of the various factors and attributes contributing towards occurrence of strokes, statistical significance tests - Paired Samples T-Test, Non parametric Chi-Square tests were conducted on the dataset. The results of the Paired samples T-Test revealed all the ten attributes to have significant effect on occurrence of stroke with significance value less than 0.05. The non-parametric Chi-square test revealed almost similar results wherein nine out of ten attributes contributed significantly towards occurrence of stroke and only one attribute - residence type did not have any significance (0.591) on the output. The results of the statistical tables are depicted in Table 4 and Table 5.

From the above discussion, the contribution of the proposed work is summarized below:

1. Re-sampling method is successfully used for balancing the imbalanced stroke dataset with multiple modalities.
2. Rigorous pre-processing is done to eliminate unnecessary attributes, filling missing values, transformation and normalization of the raw data.

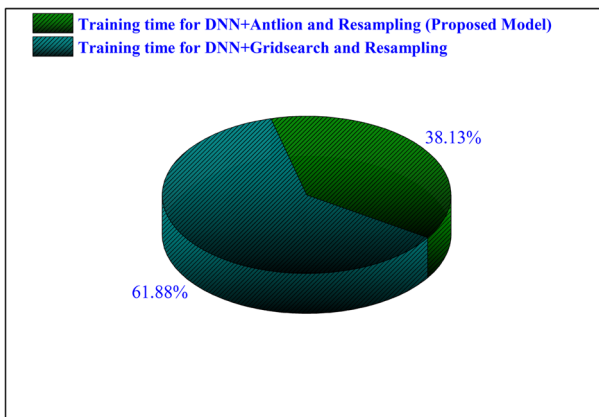


Fig. 14 Training Time Comparison in Seconds

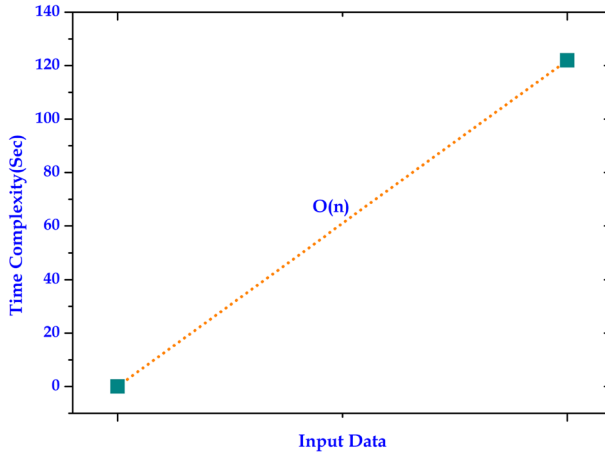


Fig. 15 Asymptotic Analysis of the Proposed Model

Table 2 Comparative Analysis of Proposed Model with other Models

Machine Learning Models	Precision	Recall	F1-Score	Accuracy	Sensitivity	Specificity
Proposed DNN + Antlion re-sampling	99.5	99.5	99.5	99.8	99	100
DNN + Antlion	98	98.5	98.5	99	100	0
DNN + Gridsearch_Resample	99	99	99	99.2	98	100
DNN	97	98	98	98.5	100	0
NB_Resample	84	83	83	83.3	78.1	88.3
NB	97	97	97	96.9	98.3	11.2
DT_Resample	97	97	97	97.3	97.08	98
DT	97	96	97	96.3	97.6	6.7
RF_Resample	98	98	98	98.3	98.4	99
RF	97	98	98	98.3	99.8	0
SVM_Resample	88	87	87	87	83.1	91.6
SVM	97	98	98	98.4	100	0
XGB_Resample	94	94	94	94.3	91.2	96.2
XGB	97	98	98	98.4	100	0

Table 3 Comparative Analysis of the Proposed Model vs. DNN with the Meta - Heuristic Optimization Models

DNN+Metaheuristic	Precision	Recall	F1-Score	Accuracy	Sensitivity	Specificity
Proposed DNN + Antlion re-sampling	99.5	99.5	99.5	99.8	99	100
DNN+ABC re-sampling	99.2	99.1	99	99.2	98.7	99.6
DNN+GA re-sampling	99	98.9	98.9	98	98.3	99.1
DNN+PSO re-sampling	99.1	99	98.9	99.1	98.5	99.56
DNN+GSA re-sampling	98.7	98	98.5	98.5	98.2	99

Table 4 Paired Sample T-Test

Paired Samples Test		Paired Differences							
		Mean Deviation	Std. Error Mean	Std. Deviation	95% Confidence Interval of the Difference	t	df	Sig.(2-tailed)	
					Lower	Upper			
Pair 1	gender - stroke	1.574	.511	.002	1.569	1.579	641.423	43399	.000
Pair 2	age - stroke	42.200	22.499	.108	41.988	42.412	390.740	43399	.000
Pair 3	hypertension - stroke	.076	.311	.001	.073	.078	50.602	43399	.000
Pair 4	heart_disease - stroke	.029	.238	.001	.027	.032	25.822	43399	.000
Pair 5	ever_married - stroke	.626	.488	.002	.621	.630	267.247	43399	.000
Pair 6	work_type - stroke	3.458	1.288	.006	3.446	3.470	559.164	43399	.000
Pair 7	Residence.type - stroke	1.483	.517	.002	1.478	1.488	597.531	43399	.000
Pair 8	avg_glucose_level - stroke	104.46471	43.10145	.20689	104.05919	104.87022	504.920	43399	.000
Pair 9	bmi - stroke	28.58971	7.76850	.03793	28.51535	28.66406	753.662	41937	.000
Pair 10	smoking_status - stroke	1.348	1.077	.005	1.338	1.358	260.816	43399	.000

Table 5 Non parametric Tests – Chi Square test

Test Statistics												
	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	
Chi-Square	23846.488a	12061.475b	28675.974c	35543.875c	3586.419c	40746.736d	.289c	67590.246e	50077.819f	5777.969g	40324.506c	
df	2	103	1	1	1	4	1	225	554	3	1	
Asymp. Sig.	0	0	0	0	0	0	0.591	0	0	0	0	

a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 14466.7.

b. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 417.3.

c. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 21700.0.

d. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 8680.0.

e. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 192.0.

f. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 75.6.

g. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 10850.0.

3. The existing hyperparameter optimization algorithms have their associated challenges pertinent to excessive time consumption. Antlion optimization (ALO) algorithm is used in the present study for selecting optimal hyperparameters in Deep Neural Networks model in minimal time.
4. The performance of the DNN-Antlion in classification of the balanced dataset is compared with the results of DNN -Antlion applied on imbalanced datasets. This has successfully established the fact that balancing of data is the key in building a successful classification model to achieve accurate results.
5. A detailed comparative analysis is done to evaluate performance of the proposed model against other prevalent classification models.

The next section concludes the current work and also discusses about the future extensions of the current work.

5 Conclusions and future work

The present study has focused on development of a rigorous data pre-processing technique that eliminates most of the challenges pertinent to data quality in the existing multimodal stroke dataset collected from the publicly available Kaggle repository. The multimodal data in the dataset helps to increase the prediction accuracy and contributes towards enhanced learning performance. To summarize, the pre-processing is initiated with replacement of the missing values in the dataset with attribute means. The data is then transformed using LabelEncoder and imbalances are treated using re-sampling method followed by normalization using StandardScaler. Antlion optimization algorithm is applied on the dataset which is finally fed into the optimally hyperparameterized DNN model generating extremely accurate results in minimal time. The performance of the model when evaluated against traditional machine learning (ML) methodologies prominently justifies its superiority. The future directions could be creation of an extremely robust large cross institutional dataset which would further optimize the classification and prediction results generated from the machine learning (ML) models.

Acknowledgments The work of Saqib Hakak is supported by the University of Northern British Columbia under FUND 15021 ORG 4460.

References


1. Al-khafajiy M, Baker T, Chalmers C, Asim M, Kolivand H, Fahim M, Waraich A (2019) Remote health monitoring of elderly through wearable sensors. *Multimedia Tools and Applications* 78(17):24681–24706
2. Benjamin EJ, Virani SS, Callaway CW, Chamberlain AM, Chang AR, Cheng S, Chiuve SE, Cushman M, Delling FN, Deo R, et al. (2018) Heart disease and stroke statistics-2018 update: a report from the American Heart Association. *Circulation* 137(12):e67
3. Bentley P, Ganesalingam J, Jones ALC, Mahady K, Epton S, Rinne P, Sharma P, Halse O, Mehta A, Rueckert D (2014) Prediction of stroke thrombolysis outcome using ct brain machine learning. *NeuroImage: Clinical* 4:635–640
4. Chen L, Bentley P, Rueckert D (2017) Fully automatic acute ischemic lesion segmentation in dwi using convolutional neural networks. *NeuroImage: Clinical* 15:633–643
5. Chiroma H, Gital AY, Rana N, Shafi'i MA, Muhammad AN, Umar AY, Abubakar AI (2019) Nature inspired meta-heuristic algorithms for deep learning: Recent progress and novel perspective. In: *Science and Information Conference*, pp 59–70, Springer

6. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639):115
7. Feng L, Ali A, Iqbal M, Bashir AK, Hussain SA, Pack S (2019) Optimal haptic communications over nanonetworks for e-health systems. *IEEE Transactions on Industrial Informatics* 15(5):3016–3027
8. Gadekallu TR, Khare N, Bhattacharya S, Singh S, Maddikunta PKR, Srivastava G (2020) Deep neural networks to predict diabetic retinopathy. *Journal Of Ambient Intelligence and Humanized Computing*
9. Garg S, Kaur K, Kumar N, Rodrigues JJPC (2019) Hybrid deep-learning-based anomaly detection scheme for suspicious flow detection in sdn: A social multimedia perspective. *IEEE Transactions on Multimedia* 21(3):566–578
10. Goyal M et al (2017) Prediction of stroke using deep learning model. In: *International Conference on Neural Information Processing*, pp 774–781, Springer
11. Goyal M et al (2018) Long short-term memory recurrent neural network for stroke prediction. In: *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pp 312–323, Springer
12. Heidari AA, Faris H, Mirjalili S, Aljarah I, Mafarja M (2020) Ant lion optimizer: Theory, literature review, and application in multi-layer perceptron neural networks. In: *Nature-Inspired Optimizers*, pp 23–46, Springer
13. Huang C, Liu B (2019) New studies on dynamic analysis of inertial neural networks involving non-reduced order method. *Neurocomputing* 325:283–287
14. Jindal A, Aujla GS, Kumar N, Prodan R, Obaidat MS (2018) Drums: Demand response management in a smart city using deep learning and svr. In: *2018 IEEE Global Communications Conference (GLOBECOM)*, pp 1–6, IEEE
15. Johnson JM, Khoshgoftaar TM (2019) Survey on deep learning with class imbalance. *Journal of Big Data* 6(1):27
16. Kamal H, Lopez V, Sheth SA (2018) Machine learning in acute ischemic stroke neuroimaging. *Frontiers in neurology* 9:945
17. Kaur H, Pannu HS, Malhi AK (2019) A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR)* 52(4):1–36
18. Khosla A, Cao Y, Lin CC-Y, Chiu H-K, Hu J, Lee H (2010) An integrated machine learning approach to stroke prediction. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp 183–192
19. Kutia S, Chauhdary SH, Iwendi C, Liu L, Yong W, Bashir AK (2019) Socio-technological factors affecting user's adoption of ehealth functionalities: A case study of china and ukraine ehealth systems. *IEEE Access* 7:90777–90788
20. Li J, Fong S, Wong RK, Chu VW (2018) Adaptive multi-objective swarm fusion for imbalanced data classification. *Information Fusion* 39:1–24
21. Liu T, Fan W, Wu C (2019) A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. *Artif Intell Med* 101:101723
22. Maddikunta PKR, Gadekallu TR, Kaluri R, Srivastava G, Parizi RM, Khan MS (2020) Green communication in iot networks using a hybrid optimization algorithm. *Comput Commun*
23. Manogaran G, Varatharajan R, Priyan MK (2018) Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system. *Multimedia tools and applications* 77(4):4379–4399
24. Mirjalili S (2015) The ant lion optimizer. *Advances in engineering software* 83:80–98
25. Patel H, SinghRajput D, ThippaReddy G, Iwendi C, KashifBashir A, Jo O (2020) A review on classification of imbalanced data for wireless sensor networks. *International Journal of Distributed Sensor Networks* 16(4):1550147720916404
26. Pham Q-V, Mirjalili S, Kumar N, Alazab M, Hwang W-J (2020) Whale optimization algorithm with applications to resource allocation in wireless networks. *IEEE Trans Veh Technol* 69(4):4285–4297
27. Qin Z, Li H, Liu Z (2014) Multi-objective comprehensive evaluation approach to a river health system based on fuzzy entropy. *Math Struct Comput Sci*, 24(5). <https://doi.org/10.1017/S0960129513000777>
28. Reddy G, Kumar ReddyM P, Lakshmana K, Kaluri R, SinghRajput D, Srivastava G, Baker T, et al. (2020) Analysis of dimensionality reduction techniques on big data. *IEEE Access* 8:54776–54788
29. Reddy GT, Khare N (2018) Heart disease classification system using optimised fuzzy rule based algorithm. *Int J Biomed Eng Technol* 27(3):183–202
30. Reddy T, RM SP, Parimala M, Chowdhary CL, Hakak S, Khan WZ, et al. (2020) A deep neural networks based model for uninterrupted marine environment monitoring. *Comput Commun*
31. RM SP, Bhattacharya S, Maddikunta PKR, Somayaji SRK, Lakshmana K, Kaluri R, Hussien A, Gadekallu TR (2020) Load balancing of energy cloud using wind driven and firefly algorithms in internet of everything. *Journal of Parallel and Distributed Computing*

32. Stroke prediction (2020 (Accessed on January 22, 2020)). <https://www.kaggle.com/swatis1/stroke-prediction>
33. Salunkhe UR, Mali SN (2016) Classifier ensemble design for imbalanced data classification: a hybrid approach. *Procedia Computer Science* 85:725–732
34. Sattar HA, Cheetar A (2019) A new strategy based on gsabat to solve single objective optimization problem. *International Journal of Swarm Intelligence Research (IJSIR)* 10(3):1–22
35. Scalzo F, Alger JR, Hu X, Saver JL, Dani KA, Muir KW, Demchuk AM, Coutts SB, Luby M, Warach S, et al. (2013) Multi-center prediction of hemorrhagic transformation in acute ischemic stroke using permeability imaging features. *Magnetic Resonance Imaging* 31(6):961–969
36. Sultan S, Javed A, Irtaza A, Dawood H, Dawood H, Bashir AK (2019) A hybrid egocentric video summarization method to improve the healthcare for Alzheimer patients. *Journal of Ambient Intelligence and Humanized Computing* 10(10):4197–4206
37. Takahashi N, Lee Y, Tsai D-Y, Matsuyama E, Kinoshita T, Ishii K (2014) An automated detection method for the mca dot sign of acute stroke in unenhanced ct. *Radiological Physics and Technology* 7(1):79–88
38. Thabtah F, Hammoud S, Kamalov F, Gonsalves A (2020) Data imbalance in classification: Experimental evaluation. *Inf Sci* 513:429–441
39. Thomalla G, Simonsen CZ, Boutitie F, Andersen G, Berthezene Y, Cheng B, Cheripelli B, Cho T-H, Fazekas F, Fiehler J, et al. (2018) Mri-guided thrombolysis for stroke with unknown time of onset. *N Engl J Med* 379(7):611–622
40. Tripathy BK, Mitra A, Ojha J (2008) On rough equalities and rough equivalences of sets. In: *International Conference on Rough Sets and Current Trends in Computing*, pp 92–102, Springer
41. Tripathy BK, Sooraj TR, Mohanty RK (2017) A new approach to interval-valued fuzzy soft sets and its application in decision-making. In: *Advances in Computational Intelligence*, pp 3–10, Springer
42. Wang D, Huang L, Tang L (2017) Dissipativity and synchronization of generalized bam neural networks with multivariate discontinuous activations. *IEEE Transactions on Neural Networks and Learning Systems* 29(8):3815–3827
43. Yu Y, Guo D, Lou M, Liebeskind D, Scalzo F (2017) Prediction of hemorrhagic transformation severity in acute stroke from source perfusion mri. *IEEE Trans Biomed Eng* 65(9):2058–2065
44. Yuan X, Xie L, Abouelenien M (2018) A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data. *Pattern Recogn* 77:160–172
45. Zhang C, Tan KC, Li H, Hong GS (2018) A cost-sensitive deep belief network for imbalanced classification. *IEEE Transactions on Neural Networks and Learning Systems* 30(1):109–122
46. Zerdoumi S, Sabri AQM, Kamsin A, Hashem IAT, Gani A, Hakak S, Chang V (2018) Image pattern recognition in big data: taxonomy and open challenges: survey. *Multimed Tools Appl* 77(8):10091–10121
47. Zhu M, Xia J, Jin X, Yan M, Cai G, Yan J, Ning G (2018) Class weights random forest algorithm for processing class imbalanced medical data. *IEEE Access* 6:4641–4652

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Thippa Reddy G¹ · Sweta Bhattacharya¹ · Praveen Kumar Reddy Maddikunta¹ · Saqib Hakak² · Wazir Zada Khan³  · Ali Kashif Bashir⁴ · Alireza Jolfaei⁵ · Usman Tariq⁶

✉ Saqib Hakak
saqib.hakak@unbc.ca

Thippa Reddy G
thippareddy.g@vit.ac.in

Sweta Bhattacharya
sweta.b@vit.ac.in

Praveen Kumar Reddy Maddikunta
praveenkumarreddy@vit.ac.in

Ali Kashif Bashir
dr.alikashif.b@ieee.org

Alireza Jolfaei
alireza.jolfaei@mq.edu.au

Usman Tariq
u.tariq@psau.edu.sa

¹ School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu-632014, India

² Canadian Institute for Cybersecurity, Faculty of Computer Science, University of New Brunswick, Fredericton, Canada

³ Faculty of CS, IT, Jazan University, Jazan, Saudi Arabia

⁴ Department of Computing and Mathematics, Manchester Metropolitan University Manchester, Manchester, UK

⁵ Department of Computing, Macquarie University, Sydney, Australia

⁶ College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj, 11942, Saudi Arabia