ORIGINAL PAPER

An efficient model for detecting COVID fake news using optimal lightweight convolutional random forest

S. Selva Birunda[1] · R. Kanniga Devi[2] · M. Muthukannan[3]

Abstract

Nowadays, social media networks like Facebook, Twitter, Snapchat, Instagram, LinkedIn, and WhatsApp provide communication and connection on a huge scale. The revolution of social media networking has shared information and improved the digital world. Though these platforms are improved in creating new things, they have a dark side that leads in the wrong direction. The main dark side of this media is spreading false news against the people. The fake spreading has both advantages and disadvantages toward people. In particular, during the pandemic of COVID-19, the false news made people believe and misguided people into unexpected situations. Therefore, it is necessary to restrict false news to not reach a huge audience. A novel approach lightweight convolutional random forest-based honey badger (LCRF-HB) is proposed for the detection of fake news via three stages, namely the pre-processing of data, selecting features, and classifying features. Stop-word deletion, stemming, and tokenization are applied to pre-process input data during the pre-processing stage. Then, the features are minimized and the accuracy is enhanced in the selection stage via the honey badger (HB) optimization algorithm. The selected features are then provided to the classification phase where the lightweight convolutional random forest (LCRF) algorithm is used for classifying whether the news is fake or not. The performance metrics attain an accuracy of 98.7%, precision of 98.3%, specificity of 95.4%, and recall of 97.6%, respectively. The comparative analysis and performance evaluation are performed and enable a good performance rate as compared to other fake detection methods.

Keywords COVID-19 · Lightweight convolutional neural network · Fake news · Random forest · Honey badger algorithm · Selection · Classification

1 Introduction

In today's world, social media become a popular online activity that helps to share and communicate with other people throughout the world. The advancement of online activity leads to the creation of abundant fake news that circulates throughout the world. The advertisements made through social media sometimes provide false information. Social media users increased by 25% due to the COVID-19 lockdown. The increment of social media users was the main reason for the circulation of fake news. The fake news obtained in the tweet was spread very fast as compared to other normal social networks. Daily average use of Twitter was 30% during the lockdown [1]. COVID-19 was spread from Wuhan, China. The average affected range of COVID-19 news was circulated through social media instantly. Social media provide freedom for the user to publish fake news without collecting real data related to COVID-19 [2]. The extensive connectivity of people on social media is the main reason for spreading fake news abundantly. Social network including Facebook, Twitter, etc., creates a lot of fake details that are being shared by people without knowing the real information [3]. The fake information is largely spread due to the increase in traffic on social media. Social media shared

✉ R. Kanniga Devi
  rkannigadevi@gmail.com

  S. Selva Birunda
  sbirunda@gmail.com

  M. Muthukannan
  civilkannan@gmail.com;
  muthukannan.civil@kcgcollege.com

1   Department of Computer Science and Engineering, School of Computing, Kalasalingam Academy of Research and Education, Krishnan Kovil, Tamil Nadu, India

2   School of Computer Science and Engineering, Vellore Institute of Technology Chennai, Chennai, Tamil Nadu, India

3   Department of Civil, KCG College of Technology, Karapakkam, Chennai 600 097, Tamil Nadu, India

a large number of fake news that surrounded all over the world among people during the COVID-19 lockdown. During the COVID-19 pandemic, fake news provides more attention among people than real news [4]. During the pandemic situation of COVID-19, incorrect, useless, and harmful information is shared among people from different countries [5].

The SARS-CoV-2 obtained is an important challenge in our world related to health. In the twenty-first century, the spreading of fake news became more prevalent, particularly in the US presidential election held in the year 2016 [6]. Fake news was characterized as misinformation and disinformation. Misinformation means fake news was shared without knowing the facts. Disinformation explains fake news is spread purposely among people to attract their attention [7]. In recent days, the biggest issue circulating the world was the transfer of COVID-19 fake news. Knowledge of complete information about particular news can able to eradicate the spreading of COVID-19 fake news [8]. Sharing of information is a major phenomenon done through social media. The information transfer among people becomes more complex due to language, but social media eradicates the inconvenience obtained in sharing information around the world. All the information shared through social media is not fact, and it does not provide useful details for people. The social media user could analyze the information thoroughly before sharing it on the internet. Cultural evolution (CE) helps to reduce the complexity obtained in sharing information in the cultural system [9]. Twitter and Instagram are the social media networks that have become more popular in the past few years as they provide an easy way to share information abundantly. This leads to the spreading of rumors which creates negative thoughts among people [10].

1.1 Novelty

Due to the spread of fake news during COVID-19 period, apart from the disease, the stress caused by the pandemic and the fear of spreading the disease have become psychological issues. Therefore, the detection of fake news is essential, and many researchers have employed various techniques to detect fake news of COVID-19.

Novel Approach: A novel approach lightweight convolutional random forest-based honey badger (LCRF-HB) is proposed for fake news detection thereby enhancing the detection accuracy.

Minimize Loss Function: The features are selected by employing the honey badger (HB) optimization algorithm, this HB can have the ability to reduce the loss functions, and the lightweight convolutional random forest (LCRF) algorithm is employed for the classification. The LCRF classifies the features that consume less memory and the performance rate is also improved.

The major contribution of this paper is explained as follows;

- A novel technique is proposed for the detection of COVID-19 fake news via four stages such as pre-processing the data, reducing the features, and selecting and classifying the features.
- A novel lightweight convolutional random forest-based honey badger (LCRF-HB) algorithm is proposed for detecting the fake news of COVID-19 with a higher rate of accuracy.
- The proposed LCRF-HB approach is compared with various approaches for analyzing the effectiveness of the system.

The remaining section of the paper is arranged as follows: In Sect. 2, various surveys are discussed. The proposed methodology of LCRF-HB is explained in Sect. 3. The experimental result is described in Sect. 4. The conclusion is explained in Sect. 5.

2 Literature survey

Al-Ahmad et al. [11] illustrated an evolution-based approach for detecting the fake news of COVID-19. The methods in the approached paper, PSO (particle swarm optimization), GA (genetic algorithm), and SSA (salp swarm algorithm), were used for the reduction of consistent appearance and to select three wrappers of evolutionary classifications for execution. The dataset used for implementation was the Koirala dataset. The result indicated that the approach outperformed the other conventional classifiers and achieved an accuracy of 75.4%. However, the detection method used on other domains needed larger datasets. Paka et al. [12] established fake news detection to avoid the spread of false information. The technique cross-SEAN (stitch semi-supervised neural attention) holds the unlabeled data which have the larger portion. A large-scale CTF dataset was used to eliminate the fake tweets. The metrics applied for predicting the performance rates were accuracy and f1-score, respectively. Finally, the accuracy obtained the rate of 0.95% and the result provided the best performance using the technique for real-time detection of fake tweets. Meanwhile, the image media cannot be extracted.

Abdelminaam et al. [13] elaborated a deep learning technique for the detection of misleading information on Twitter during the COVID-19 pandemic. The techniques are modified LSTM and modified GRU to detect fake news. The dataset used for performance evaluation such as CoAID, Politifact, and gossip cop, respectively. Thus, the result found that the fake news and non-fake detection of tweets from

COVID-19 information has a high accuracy rate. The drawbacks of this introduced method were there are no multi-class stages to combine the context, temporal, and content features. Michail et al. [14] reviewed a novel scheme to detect fake news by utilizing graph convolutional networks (GCN) in social media. The approached methods were used for verifying the profiles, fake news spreading messages, and graphing participants. BuzzFeedNews and LIAR datasets were used. The result obtained is that the fake information can able to extracted from the textual information in social media and shows the best performance achieved by the accuracy of 0.913%, respectively. The challenge of the approached paper was the fusion of multimedia was not improved on detection.

Dong et al. [15] evaluated a two-path deep semi-supervised learning technique for fake news detection in real time. Supervised learning was for analyzing the few amounts of labeled data, and unsupervised learning was for obtaining the huge amount of unlabeled data. The two datasets are PHEME and LIAR. The parameters of metrics are accuracy, precision, f1-score, and recall. As a result, the methods are used to identify fake news from the labeled data. On the other hand, dependency analysis and sentiment analysis on NLP tasks should not take place on detection. Meel and Vishwakarma [16] described self-ensembling for the detection of fake news articles using the convolutional neural network semi-supervised framework method. The methods were used for hiding the stylometric and linguistic information from the unlabeled data. The Kaggle dataset was used for detecting fake news articles. The result executed with 93.4% accuracy acquired the best performance for fake articles of labeled data. Meanwhile, online information multimedia does not analyze the text news for the detection of fake news.

Kaliyar et al. [17] illustrated a deep learning (DL) technique for detecting fake news in social media based on BERT. The technique FakeBERT (Bidirectional Encoder Representations from Transformers) was implemented to bring out the combination of CNN into single-layered parallel blocks. The parameters were FNR (False Negative Rate), FPR (False Positive Rate), accuracy, cross-entropy loss, and confusion matrix utilized to evaluate the performance. Thus, the result found that from the existing method, the approached methods in this paper outperform the high accuracy rate of 95.9%. The drawback is that the binary and the multi-class real-world datasets were not applicable. Madani et al. [18] demonstrated a technique of artificial intelligence for fake news detection during the period of the COVID-19 pandemic. The techniques were used for detecting the new tweets using the process of machine learning (ML), natural language, and deep learning (DL). The parameter metrics evaluated such accuracy, precision, and f1-score. The result showed that the technique performs better-lacking contemplation of new tweet features with an accuracy of 79%. However, the

detection of end-to-end encryption was highly realistic and difficult to detect the manipulation of audio or video.

Khanday et al. [24] analyze the detection of fake news in social media by engaging machine learning (ML) algorithms. Also, Khanday et al. [25] identified a piece of fake news on social media employing ML algorithms. Decision tree (DT) achieves greater efficiency results in detecting fake news systems. Khanday et al. [26] also analyze an LSTM model for propaganda detection on the database of Twitter. The Ensemble approach is explained by Khanday et al. [27] for the identification of fake news on COVID-19 in Online Social Networks (OSN). The AdaBoost attains a greater efficiency of 95.3%; AdaBoost can be employed to enhance the weights of the learning algorithms.

Dixit et al. [28] developed a Levy flight honey badger optimized convolutional neural network to detect fake news, an approach developed for balancing datasets, and poor selection of features. The ISOT dataset is employed for detecting fake news, this approach achieves 95% accuracy, but it does not work well with a large number of datasets. To identify proton-exchange membrane fuel cells, Han and Ghadimi [29] illustrated a convolutional neural network (CNN) and extreme learning machine (ELM) approach with an improved honey badger algorithm (IHBA). By employing the IHBA approach the integration of the CNN and ELM model is enhanced to get the optimal results, this CNN and ELM model attains greater efficiency, and the approach leads to a local optimum.

3 Proposed methodology

The block diagram of the proposed LCRF-HB algorithm based on the detection of fake news about COVID-19 on the Twitter platform is portrayed in Fig. 1. This method consists of three stages such as data pre-processing stage, feature selection stage, and classification stage [19]. The input data are pre-processed by applying stemming, stop-word deletion, and tokenization during the data pre-processing. In the second stage, the features are selected in the feature selection phase using a honey badger algorithm. Finally, a novel light weight convolutional random forest-based honey badger is proposed for identifying fake news. The detailed description of each respective phase is discussed as follows;

3.1 Data pre-processing phase

The data pre-processing can transform the inconsistent, unstructured, unfinished data and variables into the understanding of machine patterns. In this phase, omission of data pre-processing, tokenization, stop-words, and stemming process are executed.

Fig. 1 Proposed workflow based on the identification of COVID-19 fake news
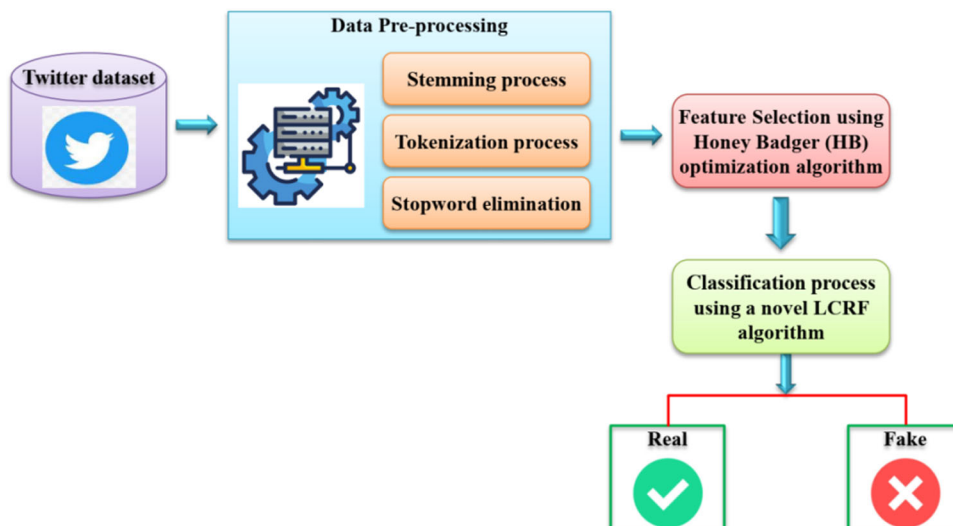
Stemming process: The stemming process objective is to gain basic words that contain similar meanings to different words. Therefore, adjectives, adverbs, nouns, and verbs are transformed into source form. The words consultative, consultant, consulting, and consultants come from the source work consult.

Tokenization: Tokenization is the segmentation of small segments from the original text called tokens. The text data's punctuation is eliminated by way of tokenization. The number filters are applied for the removal of number terms from the specific sentence. Moreover, the textual data are converted into lower cases by using case convertors. Finally, the N-char filters are used for removing a few characters.

Stop-words deletion: Stop-words are not crucial to use these words often in the sentence. But it is used to complete and merge the sentence. In English, there are more than 500 stop words used in the sentence, namely pronouns, prepositions, and conjunctions, e.g., on, am, under, against a, once, too, any, etc. Therefore, processing time and space are saved by the stop-word deletion.

3.2 Feature selection phase

After the data pre-processing phase, the feature selection phase is employed. Feature selection, in other words, referred to as attribute selection selects the appropriate features from the dataset to attain accurate classification performances. In this paper, a honey badger algorithm (HBA) is utilized for selecting the features. A detailed description of the HB algorithm is mentioned as follows.

3.2.1 Honey Badger (HB) algorithm

The behavior of honey badger's foraging is imitated by the utilization of HBA. The honey badger is utilized to either follow or dig and smell the honeyguide bird to locate the food source [22]. In HBA, first case is the digging mode and the second case is the honey mode. In the prior mode, the honey badger utilized its smelling ability to approximate the prey location. The HBA is split into two stages; the honey phase and the digging phase. The mathematical formulation of HBA is described in the below section;

The candidate solution population in the HBA is calculated and expressed in the below equation;

$$
\begin{bmatrix}
z_{11} & z_{12} & z_{13} & \cdots & z_{1F} \\
z_{21} & z_{22} & z_{23} & \cdots & z_{2F} \\
\cdots & \cdots & \cdots & \cdots \\
z_{o1} & z_{o2} & z_{o3} & \cdots & z_{eF}
\end{bmatrix}
\tag{1}
$$

Stage 1: Initialization

The total numbers of the honey badgers are initialized and it is expressed in the below equation;

$$
z_k = \text{LB}_k + \gamma_1 \times (\text{UB}_k - \text{LB}_k)
\tag{2}
$$

From Eq. (2), the honey badger kth position of the individual is depicted as z_k, the upper bound is depicted as UB_k, the lower bound is depicted as LB_k, and the random number is indicated by γ_1.

Step 2: Solution Representation

The objective of using the solution representation process is to reduce the total chosen features as well as the error percentage acquired between the original and predicted density for distinguishing fake news data from real ones. If the dataset contains D features, the decision variable will be assumed $1+D$ for feature selection and bandwidth identification. Each variable in the dataset ranges between 0 and 1. The corresponding features based on fake news are chosen from the dataset when the variable value is more than 0.5, and on the

contrary, if the variable is less than 0.5 then features are not chosen.

Step 3: Fitness evaluation

The obtained solution representations are changed into binary values [0, 1][0, 1] by the HB algorithm to represent the feature. The solution vector dimension is indicated as '1,' and the solution vector dimension y_j^{Dm} with value '0' indicates no features selected from the data. The mathematical formulation of converting solution representation to binary values is depicted by,

$$y_j^{\mathrm{Dm}} = \begin{cases} 1, & y_j^{\mathrm{Dm}} \geq 0.5 \\ 0, & y_j^{\mathrm{Dm}} < 0.5 \end{cases} \quad (3)$$

Subsequently, fitness is determined by,

$$F_{\mathrm{ITNESS}} = \varpi_1 \cdot \left(\begin{array}{c} 1 - \mathrm{Accuracy(LCRF)} + \\ \varpi_2 \cdot \left| \dfrac{\text{No. of features selected}}{\text{Total no. of features}} \right| \end{array} \right) \quad (4)$$

$$\mathrm{Accuracy\ (LCRF)} = \frac{E_n}{E_n + C_n} \quad (5)$$

From the above equations, the error rate weight $\varpi_1 = [0, 1]$, feature selection weight $\varpi_2 = 1 - \varpi_1$; the term Accuracy (LCRF) represents the accuracy rate of the LCRF classification model; E_n signifies incorrectly classified samples; and C_n implies correctly classified samples.

Step 4: Defining the intensity (I).

Intensity is defined as concentrating the prey's strength and distance among the kth honey badger I_k representing the prey's smell intensity. The motion becomes vice versa when the smell intensity of prey is higher.

$$I_k = \gamma_2 \times \frac{\delta}{4\pi f_k^2}$$
$$\delta = (z_k - z_{k+1})^2$$
$$f_k = z_{\mathrm{PREY}} - z_k \quad (6)$$

From the above equation, the concentration strength is depicted as δ the distance between kth badger and prey is depicted as f_k.

Step 5: Density Factor Updation

The time-varying randomization is controlled by using the density factor to make a consistent transition from the exploitation and exploration phases. Then, it is expressed as;

$$\gamma = E \times \mathrm{EXP}\left(\frac{-v}{v_{\mathrm{MAX}}}\right), \quad (7)$$

From the above equation, the maximum numbers of iterations are denoted by v_{MAX}, and the constant represented by E.

Step 6: Escape with local optima

The HBA utilizes the flag to assist higher opportunities to scan the search spaces formally.

Step 7: Update the agent's position

The updation of HBA is split into honey and the digging phase.

Step 7-1: Digging stage

In the digging stage, the honey badger acted for cardioid shape and they are expressed in the below equation;

$$z_{\mathrm{NEW}} = z_{\mathrm{PREY}} + H \times \lambda \times K \times z_{\mathrm{PREY}}$$
$$+ H \times \gamma_3 \times \chi \times f_k \times \left| \cos(2\Pi\gamma_4) \times [1 - \cos(2\Pi\gamma_5)] \right| \quad (8)$$

The prey position is represented by z_{PREY}, and the various random numbers are represented by γ_1, γ_2 and γ_3.

$$H = \begin{cases} 1 & \text{if } \gamma_6 \leq 0.5 \\ -1 & \text{else} \end{cases} \quad (9)$$

f_k represents a time-varying factor with search influence.

Step 7-2: Honey stage

The honeyguide birds are followed by honey badger to reach the beehive as expressed in the below equation.

$$z_{\mathrm{NEW}} = z_{\mathrm{PREY}} + H \times \gamma_7 \times \chi \times f_j \quad (10)$$

The random number is represented by γ_7, the prey location is indicated by z_{PREY}, and the honey badger's new position is indicated by z_{NEW}.

3.3 Classification using lightweight convolutional random forest (LCRF) algorithm

The classification phase plays a major role in detecting fake news. In this paper, a novel lightweight convolutional random forest (LCRF) algorithm is employed for the optimal classification. A detailed description of each technique involved in classification is discussed below.

3.3.1 Architecture of LCNN model

The lightweight convolutional neural network (LCNN) is a networking design that consists of two factors: minimum computing unit node-set Nd and connecting dual edge node-set Ed[20]. Hence, the LCNN is represented as follows:

$$\mathrm{LCNN} = (\mathrm{Nd, Ed}) \quad (11)$$

The minimum computing node set is formulated as $\mathrm{Nd} = \{\mathrm{Nd}_i | i = 1, .., n\}$. The computing unit node is utilized for a single convolutional operation or for adding numerous convolutional operations. They were establishing the convolutional operation diversity. The connecting deal edge node

set is represented as Ed $= \{Ed_{i,j} | 1 < i < j \leq n\}$. The Ed is noted that they cannot connect themselves, and there are no loops in an unaligned structure of the network. The values i, j in the set of edge nodes Ed need to meet the requirements $P \cup Q = N$; the value of the set i is captured in the value of the set Nd which is represented as N. To establish the LCNN, the minimum node located in a set of computing units Nd should reorder the group a to several β set nodes in the equal-length path.

$$Nd = \{Nd_{Ka} | K = 1, .., a; i = 1, ...b\}, \ (n = b \times a) \quad (12)$$

The groups of different nodes are not adjacent to each other, the full case of edges are connecting small computing unit nodes that are rejected with a value of a definite probability of rejection connectivity. Hence, the edge node set Ed is represented as follows:

$$Ed = \{Ed_{ij} | i \in \{(1, 1), (1, 2), ...(a, b)\}; .., a; i = 1, ...b\}, \quad (13)$$

$$j \in \{(i + 1, 1), (i + 1, b)\} \cup \beta * d_{\text{rop}}\{(i + 2, 1), ..., (a, b)\}\} \quad (14)$$

The above equation β denotes the value of disconnection probability.

3.3.2 Random forest

Random forest is based on the concept of an ensemble classifier which classifies the decision tree containing a higher number of votes at the final result of classification. In the classification of clustering analysis and regressive analysis problems in machine learning, the RF algorithm has provided excellent performance to establish a suitable model with some minor adjustments in the hyperparameters setting [21]. The different samplings of bootstrap were generated individually from every decision tree in the random forest. The errors in the classifying decision tree are determined by the different tree's abilities in the classification. Furthermore, the majority votes of the decision tree maximized the random forest classification accuracy rate.

Steps for random forest generation:

Bagging: Bagging is the first step for generating a random forest. The algorithm of RF samplings is extracting randomly in $2/3$ times from the initial data training set $Td = \{(a_1, b_1), (a_2, b_2)(a_i, b_i)\}$ to develop training subsets that are formed by a decision tree. The different samplings of bootstrap sets are captured via bagging to establish each decision tree in a random forest. OOB is an out-of-bag-data error rate that is employed to measure the classification ability of the random forest. OOB evaluation is more efficient compared with cross-validation. The Gini index and the OOB error method

are involved in the estimation of features, whereas the DT can compute the error rate while the missed classification probability Mgi is estimated by the Gini approach.

$$Mgi = \sum_{cl=1}^{cl} P_{cl}(1 - P_{cl}) = 1 - \sum_{cl=1}^{cl} P_{cl}^2 \quad (15)$$

From Eq. (5), the numbers of classes are denoted as cl, M indicates the node, and the probability class is represented as p_{cl} According to the node, M the random forest feature importance X_j was calculated by means. This scenario is formulated as follows:

$$Mvi_{Mj}^{GI} = Mgi - Lgi - Rgi \quad (16)$$

Lgi depicts the left node, and Rgi depicts the right node of the Gini index.

3.3.3 Constructing the decision tree

The random forest technique maximizes the decision tree diversity through the construction of various training subsets to enhance their efficiency. Finally, each decision tree model received each classification result.

$$B = \{b_1(a), b_2(a), ...b_i(a)\} \quad (17)$$

From the above equation, B the systematic classification model consists of each model in the decision tree. The result of the terminal categorization voting is determined through a combination of categorization models.

$$Td(a) = \max \arg \sum_{i=1}^{i} F(b_I(a) = BF) \quad (18)$$

From the above equation, classifying the single decision tree model is represented as $b_I(a)$, BF is the fault blocking class, and the combined classification model is denoted as Td(a). The indicator function is denoted as $F(.)$ Fig. 2 illustrating COVID-19 fake news detection by using the proposed LCRF-HB approach.

4 Experimental results and discussion

The proposed light weight convolutional random forest-based honey badger (LCRF-HB) algorithm for analyzing the COVID-19 fake news by using Twitter tweets. The remaining sub-sections explain the exactness of COVID-19 fake news prediction.
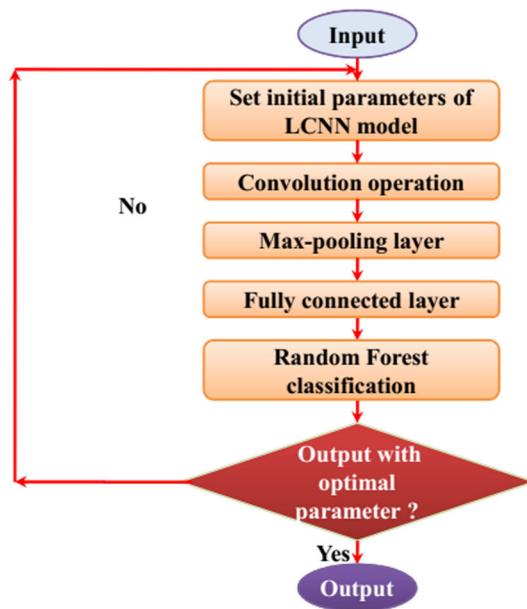
Fig. 2 LCRF-HB approach for the COVID-19 fake news detection

Table 1 Hyperparameter configuration

Parameters	Values
Information gain	< 0.1
Number of trees	100
Population size	50
Total number of iterations	1000
Number of runs	30
Decay rate	0.000001
Batch size	32
Learning rate	0.0001
Number of epochs	50

4.1 Hyperparameter configuration

To identify the optimal parameter values of the LCRF-HB in achieving better performance a hyperparameter configuration is engaged. The LCRF-HB algorithm hyperparameter configuration is tabulated in Table 1.

4.2 Dataset description

The Twitter dataset [23] is employed for analyzing COVID-19 fake news from tweets, and the data are gathered through public accounts. The COVID-19-related information is determined from the COVID-19-related tags. The fake news data were collected from the duration of December 2019 to June 2020. Fake data were collected from Google and Twitter, and these data were filtered through their specific browser.

Table 2 Features of the twitter dataset

Features	Descriptions
Reply	How people respond to a tweet is known as a reply
Date	The tweets are represented in the form of MM.DD.YYYY and when tweets are updated on Twitter is the date
Popularity	Tweet's popularity depends upon the sum of the RRL number
Stance	The source's attitude to quoting the original tweets is represented in the stance
Sentiment	The sentiment is cross-validated and labeled on the tweet's emotion
Veracity	Veracity is defined as the degree to trusted and accurate information

These collected data were gathered to form the Twitter dataset which is in the form of a.csv file. After this, the collected data were taken into the processing stage. On Twitter, the COVID-19 tags were gathered and stored, and each rumor sentiment was labeled by carefully analyzing the sentiment of the rumor's content and context. In addition, metadata for each tweet are gained including reply or retweet comment content, numbers like retweet and like, and date of publication. These data were stored individually; also, the Google data were gathered by employing the HTTP servers. The data are taken from the results page of Google, and the results are absolute from their relative paths by URL and date. Each rumor record has an authenticity label and its content, replay record, date, reply website, etc., not all source websites contain date information. The Twitter dataset features are explained in Table 2.

4.3 Performance analysis

Various approaches like particle swarm optimization (PSO) algorithm, cross-stitch semi-supervised neural attention model (cross-SEAN), modified long short-term memory (Modified LSTM), convolutional neural network-based long short-term memory (C-LSTM), and proposed light weight convolutional random forest-based honey badger (LCRF-HB) algorithm are utilized for the evaluating the performance. The accuracy analysis performed by utilizing various approaches such as PSO, modified LSTM, C-LSTM, cross-SEAN, and proposed LCRF-HB approaches is illustrated in Fig. 3. As compared to other existing approaches, the proposed LCRF-HB method acquired a greater accuracy of 98.7%. The COVID-19 fake news is detected with higher accuracy in the LCRF-HB than other approaches.

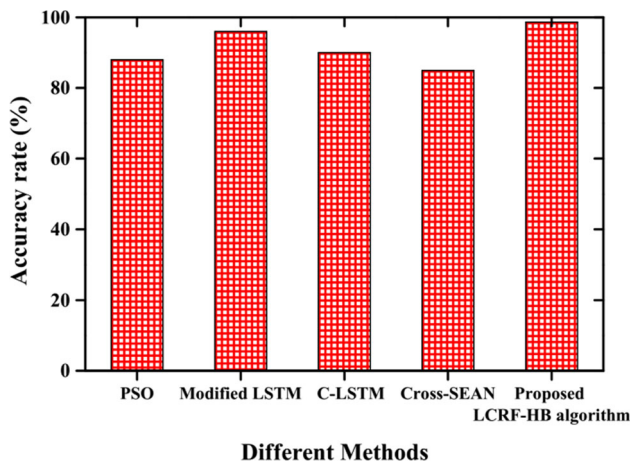Figure 4 illustrates the precision analyses of various
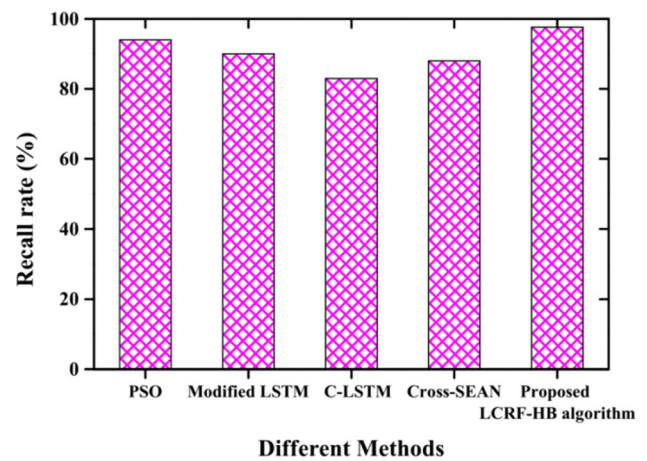
Fig. 3 Comparative analysis of accuracy

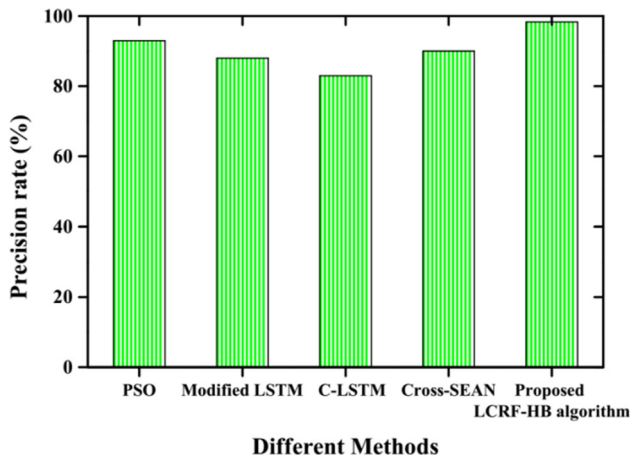

Fig. 5 Comparative analysis of recall
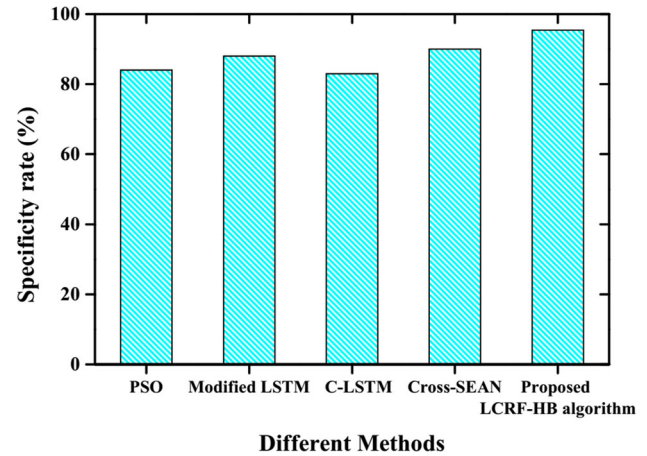

Fig. 4 Comparative analysis of precision

Fig. 6 Comparative analysis of specificity

approaches like PSO, modified LSTM, C-LSTM, cross-SEAN, and the proposed LCRF-HB approach. The precision analysis of the LCRF-HB approach acquired a greater value of 98.3%, and the C-LSTM approach has the lowest precision rate of 88%. From this comparative analysis, the COVID-19 fake news is detected with a high precision rate.

The recall rate of various methods like PSO, Modified LSTM, C-LSTM, and cross-SEAN and the proposed LCRF-HB approach is portrayed in Fig. 5. The proposed LCRF-HB approach acquired a higher recall rate of 97.6% than the other methods. The recall rates of 92%, 85%, and 88% are acquired from the PSO model, modified LSTM, and cross-SEAN.

Figure 6 depicts the specificity analysis by utilizing PSO design, modified LSTM approach, C-LSTM model, cross-SEAN model, and proposed LCRF-HB approach. The specificity rates of 84%, 90%, 80%, 92%, and 95.4% are detected by using the PSO design, modified LSTM approach, C-LSTM model, cross-SEAN model, and proposed LCRF-HB

approach, respectively. The proposed LCRF-HB approach attained a higher sensitivity rate for detecting the fake news of COVID-19 when compared to other methods.

Figure 7 illustrates the performance analysis of the proposed LCRF-HB approach with different parameters like recall, precision, accuracy, and specificity. The proposed LCRF-HB approach attained 98.7% of accuracy, 95.4% of specificity, 98.3% of precision, and 97.6% recall for COVID-19 fake news detection.

Figure 8a, b depicts the performance of training and validation of accuracy and loss. The performance of training data accuracy attains the value of 0.8%, where the validation accuracy of 0.7%. The effectiveness of training loss attains 0.4%, and validation loss of 0.2%.

Figure 9 depicts the performance analysis of the various state-of-the-art algorithms such as FakeBERT, GCN, and ConvNet concerning the proposed LCRF-HB approach. The performance of FakeBERT attains 95.9%, GCN attains
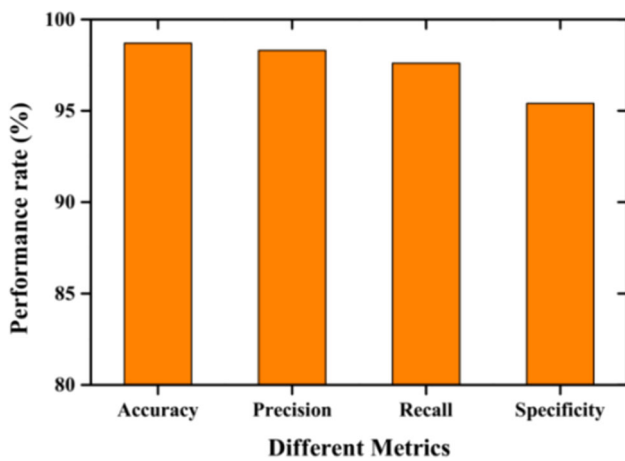
Fig. 7 Performance analysis

Fig. 8 a Efficiency of training and validation accuracy, **b** efficiency of training and validation loss
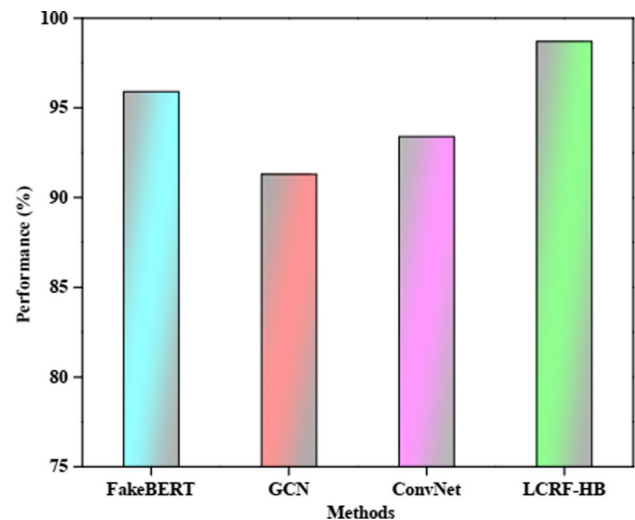

Fig. 9 Performance analysis under various state-of-the-art algorithms

91.3%, and ConvNet attains 93.4%, but the proposed LCRF-HB attains a greater efficiency of 98.7%.

5 Conclusion

The major objective of this paper is to examine COVID-19 fake news detection. The data pre-processing, feature selection, and classification are the various phases proposed to detect fake news. In data pre-processing, stop-word deletion, stemming, and stop-word deletion are utilized for pro-processing the input data. After data pre-processing, the features are selected using the honey badger algorithm. The LCRF approach is utilized for classifying COVID-19 news as real or fake manually. The Twitter dataset is employed for evaluating the performance of the proposed method. Recall, accuracy, specificity, and precision are the metrics employed for predicting the rate of efficiency. The comparative analysis is performed by utilizing various approaches like PSO design, modified LSTM approach, C-LSTM model, cross-SEAN model, and proposed LCRF-HB approach. The proposed LCRF-HB approach attained a higher 98.3% precision, 95.4% specificity, 97.6% recall, and 98.7% accuracy for detecting COVID-19 fake news. The proposed LCRF-HB approach attained a training accuracy of 98.7%. For future research, various techniques will be integrated to enhance the performance as well as COVID-19 fake news detection optimally.
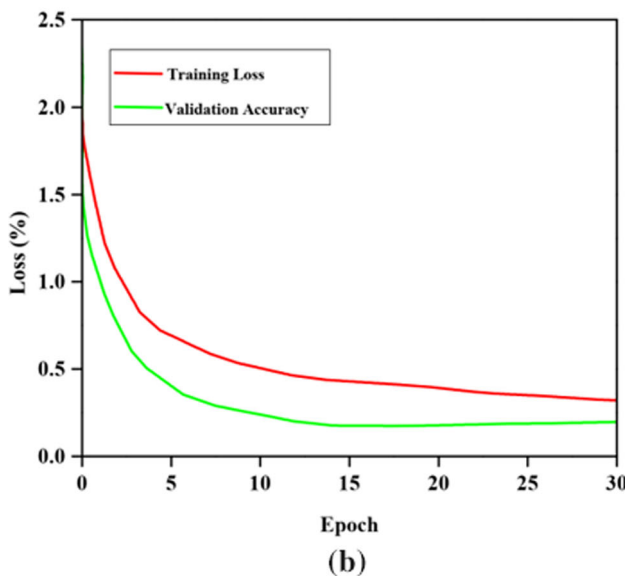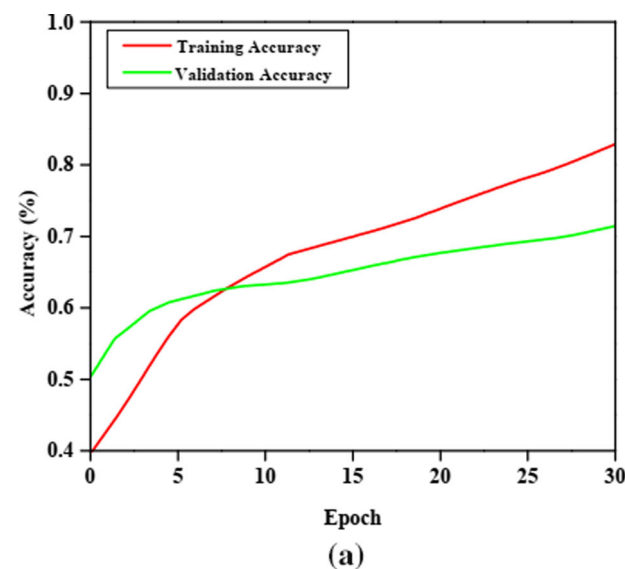
Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Consent for publication Not applicable.

Consent to participate Not applicable.

Informed consent Not applicable.

Human and animal rights This article does not contain any studies with human or animal subjects performed by any of the authors.

References

1. Apuke, O.D., Omar, B.: Fake news and COVID-19: modelling the predictors of fake news sharing among social media users. Telematics Inform. **56**, 101475 (2021)
2. Naeem, S.B., Bhatti, R., Khan, A.: An exploration of how fake news is taking over social media and putting public health at risk. Health Inform. Libr. J. **38**(2), 143–149 (2021)
3. Schuetz, S.W., Sykes, T.A., Venkatesh, V.: Combating COVID-19 fake news on social media through fact checking: antecedents and consequences. Eur. J. Inf. Syst. **30**(4), 376–388 (2021)
4. Zeng, J., Zhang, Y., Ma, X.: Fake news detection for epidemic emergencies via deep correlations between text and images. Sustain. Cities Soc. **66**, 102652 (2021)
5. Monterrubio, S.M.M., Noain-Sánchez, A., Pérez, E., Crespo, R.G.: Coronavirus fake news detection via MedOSINT check in health care official bulletins with CBR explanation: the way to find the real information source through OSINT, the verifier tool for official journals. Inf. Sci. **574**, 210–237 (2021)
6. Kim, J., Aum, J., Lee, S., Jang, Y., Park, E., Choi, D.: FibVID: comprehensive fake news diffusion dataset during the COVID-19 period. Telemat. Inf. **64**, 101688 (2021)
7. Ceron, W., de Lima-Santos, M.F., Quiles, M.G.: Fake news agenda in the era of COVID-19: identifying trends through fact-checking content. Online Soc. Netw. Media **21**, 100116 (2021)
8. Choraś, M., Demestichas, K., Giełczyk, A., Herrero, A., Ksieniewicz, P., Remoundou, K.U.D., Woźniak, M.: Advanced machine learning techniques for fake news (online disinformation) detection: a systematic mapping study. Appl. Soft Comput. **101**, 107050 (2021)
9. De Oliveira, D.V.B., Albuquerque, U.P.: Cultural evolution and digital media: diffusion of fake news about COVID-19 on Twitter. SN Comput. Sci. **2**(6), 1–12 (2021)
10. Wani, A., Joshi, I., Khandve, S., Wagh, V., Joshi, R.: Evaluating deep learning approaches for covid19 fake news detection. In Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1, pp 153–163. Springer (2021)
11. Al-Ahmad, B., Al-Zoubi, A.M., Abu Khurma, R., Aljarah, I.: An evolutionary fake news detection method for covid-19 pandemic information. Symmetry **13**(6), 1091 (2021)
12. Paka, W.S., Bansal, R., Kaushik, A., Sengupta, S., Chakraborty, T.: Cross-SEAN: a cross-stitch semi-supervised neural attention model for COVID-19 fake news detection. Appl. Soft Comput. **107**, 107393 (2021)
13. Abdelminaam, D.S., Ismail, F.H., Taha, M., Taha, H., Houssein, E.H., Nabil, A.: Coaid-deep: an optimized intelligent framework for automated detecting covid-19 misleading information on Twitter. IEEE Access **9**, 27840–27867 (2021)
14. Michail, D., Kanakaris, N., Varlamis, I.: Detection of fake news campaigns using graph convolutional networks. Int. J. Inf. Manag. Data Insights **2**(2), 100104 (2022)
15. Dong, X., Victor, U., Qian, L.: Two-path deep semisupervised learning for timely fake news detection. IEEE Trans. Comput. Soc. Syst. **7**(6), 1386–1398 (2020)
16. Meel, P., Vishwakarma, D.K.: A temporal ensembling-based semi-supervised ConvNet for the detection of fake news articles. Expert Syst. Appl. **177**, 115002 (2021)
17. Kaliyar, R.K., Goswami, A., Narang, P.: FakeBERT: fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools Appl. **80**(8), 11765–11788 (2021)
18. Madani, Y., Erritali, M., Bouikhalene, B.: Using artificial intelligence techniques for detecting Covid-19 epidemic fake news in Moroccan tweets. Results Phys. **25**, 104266 (2022)
19. Dixit, D.K., Bhagat, A., Dangi, D.: Automating fake news detection using PPCA and levy flight-based LSTM. Soft. Comput.Comput. **26**(22), 12545–12557 (2022)
20. He, Y., Li, T.: A lightweight CNN model and its application in intelligent practical teaching evaluation. In: MATEC Web of Conferences, EDP Sciences, vol. 309, p. 05016 (2020)
21. Zhou, X., Xu, X., Zhang, J., Wang, L., Wang, D., Zhang, P.: Fault diagnosis of silage harvester based on a modified random forest. Inf. Process. Agric. (2022)
22. Hashi, F.A., Houssein, E.H., Hussain, K., Mabrouk, M.S., Mand Al-Atabany, W.: Honey Badger Algorithm: new metaheuristic algorithm for solving optimization problems. Math. Comput. SimulComput. Simul. **192**, 84–110 (2022)
23. https://www.kaggle.com/datasets/goyaladi/twitter-dataset
24. Khanday, A.M.U.D., Khan, Q.R., Rabani, S.T.: Analysing and predicting propaganda on social media using machine learning techniques. In: 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), pp. 122–127. IEEE (2020)
25. Khanday, A.M.U.D., Khan, Q.R., Rabani, S.T.: Identifying propaganda from online social networks during COVID-19 using machine learning techniques. Int. J. Inf. Technol. **13**, 115–122 (2021)
26. Khanday, A.M.U.D., Khan, Q.R., Rabani, S.T., Wani, M.A., ELAffendi, M.: Propaganda identification on twitter platform during COVID-19 pandemic using LSTM. In: International Conference on Cybersecurity, Cybercrimes, and Smart Emerging Technologies, pp. 303–314. Springer, Cham (2022)
27. Khanday, A.M.U.D., Khan, Q.R., Rabani, S.T.: Ensemble approach for detecting COVID-19 propaganda on online social networks. Iraqi J. Sci. 4488–4498 (2022)

28. Dixit, D.K., Bhagat, A., Dangi, D.: An accurate fake news detection approach based on a Levy flight honey badger optimized convolutional neural network model. Concurr. Comput. Pract. Exp. **35**(1), 7382 (2023)

29. Han, E., Ghadimi, N.: Model identification of proton-exchange membrane fuel cells based on a hybrid convolutional neural network and extreme learning machine optimized by improved honey badger algorithm. Sustain. Energy Technol. Assess. **52**, 102005 (2022)