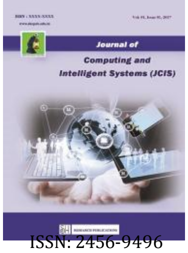
Original Research Article

SACRED HEART RESEARCH PUBLICATIONS

Journal of Computing and Intelligent Systems

Journal homepage: www.shcpub.edu.in



ISSN: 2456-9496

DHL: A Deep Hybrid Learning Framework for COVID-19 Fake News Identification

R Sandrilla

Received on 22 JUN 2022, Accepted on 29 JUL 2022

Abstract — The emergence of false information around has become a serious threat to political, economic, rituals, and communal constancy. Social media is a source of data that is not always accurate, specifically in the case of the COVID-19 outbreak. Fake news is widely spread during the COVID-19 epidemic. Early detection is the right way to handle with this. A deep hybrid framework for identifying the bogus news is proposed in this paper, which employs 1D Convolutional Neural Networks (CNN) with BiLSTM for feature extraction and classic Machine Learning models for classification. The implemented framework model allows one to decrease human exertions for features extraction from a dataset. To the best our knowledge no research was carried on for the classification of detecting the misleading information, based on hybridizing both the strong learning algorithms. Here the researcher introduced a framework as deep hybrid learning approach that which incorporate Deep Learning together with Machine Learning models and enhanced with the much promising results. The suggested framework successfully distinguishes between bogus and true news in the desired fake news dataset given, by achieving a consistent performance of around 99 percent in the accuracy, with very low computational power and less time of about 9 minutes, as compared to the majority of other state-of-the-art concepts of fake news datasets, respectively

Keywords – *Fake News, Real News, Machine Learning, Deep Learning, Hybrid Learning framework.*

. INTRODUCTION

Facebook, Instagram, and Twitter, for example, place a premium on news sharing, interaction, involvement, and collaborative effort. This isn't just for individual sharing, but also for enterprises to promote their brand and catch the attention of customers. Fortunately, the advancement and availability of phone apps have made these platforms widely available and easily accessible. Even so, one of the greatest obstacles is controlling the rational spread of false information or misleading information on commercial networks, which covers a variety of topics such as economics, the surroundings, world affairs, and health. Publishers of fake news and misinformation may be motivated by a variety of factors, including entertainment, misleading public sentiment about an issue, raising the quantity of web users, promoting a prejudiced point, and so on. They are either deceptive or criminal in general. Coronavirus (COVID-19) is a worldwide epidemic caused by Coronavirus2. The outburst occurred during December '19 in the Chinese city of Wuhan, and it is regarded as one of the most contagious and squishy pathogens to have afflicted our planet in recent decades [1]. Since 2021, multitude of coronavirus have erupted and become common in many countries. There were over 485 million confirmed cases, with 6.2 million premature deaths [2]. Since the coronavirus's global spread, social media platforms have played a greater role in informing people and ensuring their protection, efficiency, and engagement with one another. News organizations and medical organizations have used this opportunity to spread virus information.

The Institutes for Disease Control (CDC), World Health Organization (WHO), health agencies, and medical journals have all helped to publish and update information about the virus's spread, prevention, and treatment. However, because this issue is linked to the health and medical aspects of human existence, the spread of information and false news about the pathogen via the Web has incorporated another challenge to combat. Furthermore, this fake news had tragic consequences, such as harming the country's economy,

*** Corresponding author: E-mail:** [1]sandrilla@shctpt.edu

[1]Assistant Professor , Department of Computer Science,
Sacred Heart College (Autonomous), Tirupattur Dt.

lowering people's trust in their governments, promoting specific products to make huge profits, and disseminating incorrect advices and directions for treating and preventing the virus [1]. According to the independent [3], over 480 people died and 280 others were cursed in Iran as a result of false information disperse on social media platforms claiming that drinking alcohol reduces coronavirus infection. As per the Defender [4], a death was reported and his wife was harmed in the United States after taking drugs that Trump declared to be a diagnosis for coronavirus. The term "Fake News" is defined in the literature as "news that is authored to misguide the audience and give them hope that it is factual and reliable" [5].

Academics from all over the world are fascinated by the classification of news as fake or real. As a result, it is critical to develop a system for detecting fake news. This article, proposed a framework for detecting fake news which uses CNN to extract features before categorizing fake news using Machine Learning approach.

In this regard few important contributions are discussed in detail are stated as follows:

- An innovative deep hybrid learning approach for automatically detecting untruthful news is presented. The proposed methodology performs existing state-of-the-art strategies for spam detection substantially.
- Section II of this paper outlines previous work in the fields of fake news identification.
- Section III describes the dataset utilized to perform the classification.
- Section IV demonstrates the Methods and Methodology.
- Section V displays Results and Discussion.
- Section VI concludes with a conclusion and future enhancements of the work.

II. Related Work

Fake news is nothing new. Emphasizing the importance of fake news, scientists have been dealing with different techniques to provide a quick and instant solution for spam detection in recent decades.

J A Nasir et al. [6] developed a innovative method of hybrid deep learning algorithm. As part of their research, they developed a model for detecting fake news. Their model used CNN to extract features and LSTM to classify them. On two datasets, it significantly outperformed nonhybrid baseline methods in sensing fake news. Fake news is nothing new. To distinguish between rumors circulated on social media and those pertaining to covid, M Al-Sarem et al. [16] created a hybrid deep learning-based model. Long Short-Term Memory and Concatenated Parallel Convolutional Neural Networks (LSTM–PCNN) are the foundations of the proposed model. The new model outperformed the older ones in terms of recall, precision, and F1-score. Hossain et al. [17], created a COVIDLIES, a benchmark dataset known as COVID-19 misconceptions. They categorized each tweet in the dataset as Agree, Disagree, or express No Stance. A dataset on COVID-19 that contains 10,700 posts and articles of true and false news was created and annotated by Patwa et al. [18]. To categorize articles as real or fake, four ML benchmarks were implemented to a compiled dataset: DT, LR, Gradient Boost, and SVM. With the test dataset, SVM produced great results. In [19] the researchers had a deep discussion on LIAR fake news dataset, which misled the public. In their study they concluded that comparing to most of the machine learning techniques naive bayes algorithm-based model perform good in level of accuracy. Meanwhile the author in the study [20] suggested in their paper that false information could be detected using Bi-directional Long Short-Term Memory and Recurrent Neural Network (RNN) models. After a complete survey on deep hybrid learning approach, the researcher came to a conclusion that there are numerous researches were arrived on finding the true news based on many machine learning algorithms and deep learning algorithms. These machine learning and deep learning algorithms stand high on their way in performing and producing the results good. Whereas, combing both the learning algorithms would lead to enhanced performance has been found from the state-of-art. So, the researcher hybridized the ML with DL to accurately detect the bogus news which seriously Misleads the public. In this research work, we used CNN with BiLSTM to extract features from text data, reducing the amount of human effort. We used various machine learning frameworks and compared their outcomes for the classification tasks with the Standalone Deep neural Network (DNN)

III. Dataset

In this to examine and interpret the results the researcher has used two sets of information gathered and named as Dataset I and Dataset II. The description of both the datasets are discussed in detail below.

Dataset 1: This dataset consists of COVID-19 False news which mislead the people and caused unembellished disturbance among People. This news is directly collected from the internet under the link referred in the research gate as COVID-19 dataset. In this dataset by itself the labels were given as heading, marking with booloen characters as o0 and 1. The Boolean integer 0 indicates the fake news and 1 symbolizes a True news

Dataset II: The second set of information about the COVID-19 false news was collected from Webhose.io. Since the researcher collected from the host the headings were manually labelled. In this dataset we found three set of news content that were posted that is completely positive news which is labelled as true news, then false news, finally indicating both false and true which is labelled as partially true. We considered the partial as well as the false news as 0 and true as 1.

IV. METHODS AND METHODOLOGY

In this section, the proposed structure of our current proposal for detecting fake news is clearly discussed. We used basic preprocessing, then to extract the useful and relevant data alone we used hybrid based neural networks for feature extraction, in order to reduce the processing time and result the outcome in efficient accuracy. Finally, the traditional machine learning models are castoff to to categorize the data. To the knowledge of the researcher the hybrid method of feature extraction with the hybrid-based classification is a new technique proposed so far. Whereas this method has been applied in other areas, it has not yet been used in nlp.

A. Data Preprocessing

This is the phase of cleaning and preparing the data for further processing where a set of processing steps are applied to correctly predict the forecasting results. The steps followed are stated below:

- *Parsing of text.* This is the initial step taken in Preprocessing the data. To make it simpler the text is parsed using the method called Tokenization, where this function breaks down the text by dividing it within the datasets for further analysis steps.
- *Cleaning of data.* The next process of preprocessing done in our work is data cleaning Here we tried to extract the English Alphabets,
- numerals and alphanumeric combinations using the regular Expression. This step mainly focusses to remove the noisy and irrelevant data from the given set of text data.
- *Tagging of Parts of Speech (PoS).* This is a technique which is used to categories the words by giving the useful information that relates the word used in speech. Finally, it provides with the meaningful phrases from the given text.
- *Remove Stop Words.* This method is used to remove very commonly occurring words or with little information that which provides unnecessary noise. Stop words techniques are particularly used in text Classification Models. This method is done by not missing out the meaning the text with the provided piece of text.
- *Stemming.* This process we used as a replacement method to substitute every term by its core in removing duplication in the text. Furthermore, this step will reduce to the root form of the given text after the process of stemming.

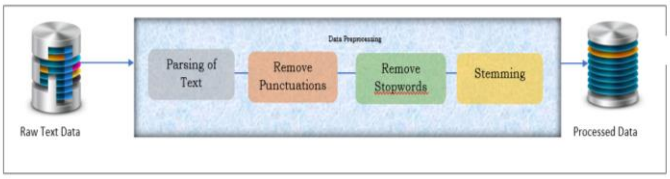


Fig. 2. Workflow of Data Processing

B. Cross-Validation

The concept behind n-Fold Cross-Validation [21] is to divide the data provided with n no of segments. Cross Validation uses statistical method which divides the data into pairs by evaluating and comparing the learning algorithm. This technique uses two different portions for training and testing to predict the level of accuracy while performing in reality.

C. Training and Testing Process

We use a neural network in this architecture to extract features from the dataset. Following the extraction of those features, we compare the performance of five machine learning algorithms for fake detection tasks.

1) Neural Network based Feature Extraction

- Neural Network: Neural network (NN) are the sub modules of Machine Learning models, also known as Artificial Neural Network (ANN) [7]. A Neural Network is a streamlined replica of the Human brain. The ANN is made up of artificial neurons, much like a neuron in the human brain. The artificial neurons resemble nodes. ANN includes components such as loss functions, Activation functions, stable optimizers in addition to artificial neurons. A neural network is simply a weighted directed graph with additional features.

- CNN (Convolutional Neural Network): CNN [8] is commonly known as Convolutional Neural Network, which is most popular network of deep neural networks [9] It uses matrices to perform convolution operations to generate features. CNN works roughly as a human vision. It is normally applied to analyze the visual image. CNN is used for two set of functions like extraction and classification. In our study we used for extracted the best features from the dataset. In case of extraction, it uses multiple layers of Max Pooling with suitable activation function to enhance the performance. The subsequent features have been used effectively in sentiment classification, language processing, classification techniques, and other NLP tasks.

- Word Embeddings: Word embedding refers to a class of methods for conveying words and texts that represents the dense vector. Word embedding is a mechanism for representing words as vectors, with similar vectors for words with the same meaning. Each word in the text is represented in dense vectors. Words in an embedding are not associated with distinct snippets of data; instead, they are signified as high-dimensional matrices. The temporal status of a word in the sentence can be determined by before or after they appear.

2) The Feature Extractor's architecture

Neural Network-based feature extractor are used to implement our process, we used the TensorFlow [10] sequential API. This proposed model Fig.1. consists of multiple layers. To begin, we add an embedding layer which is the first layer in the model initiating 100 dimensions as its initial weights. Following that, a dropout layer is used to with the rate of 0.3 to minimize the overfitting effects. Then we used a 1-d convolutional layer. The following layer BiLSTM The filter and the kernel size used in that is 256, and 5 along with sigmoid as the activation function. As a result, we used a pooling layer where the BiLSTM layer captures the pooled features maps. In this sense, max-pooling is effective.

We've also added another dropout layer with the same rate of 0.3. We used sigmoid as an activation function in this layer. Finally, in the final layer, we used the SoftMax function. That is the overall design of the feature extractor.

We used Adam optimizer [11] with a learning rate of 0.001 to compile the model. We experimented with various optimizers and finally chose the Adam optimizer. Moreover, Adam Optimizer outperformed excellently in our case. Furthermore, we used the binary cross-entropy loss function [12] to compute failures. Once compiled and fitted the model with the training phase in each iteration, we used the testing set to predict the outcome of this learning algorithm. The outcome of the second last layer is chosen for the next step and sent to the machine learning algorithm to diagnose false news.

In the proposed framework the researcher compared the standalone DNN model with five different combination of DHL Models.

TABLE I The Proposed Parameter for Hybrid Deep Neural Network

Parameters	**Values**
Size of the Filter	64, 128
Size of the Kernel	256, 5
Max pool	6
Dropout	0.3
Batch	219
Epochs	10
Activation Function used	Sigmoid
Optimizer	Adam

3) Machine Learning Algorithms:

- ***Support Vector Machine:*** A popular Supervised Learning technique is using the Support Vector Machine (SVM) [13] for Classification and Regression. Nevertheless, it is primarily based on Machine Learning algorithms rather added to solve classification problems. Finding the ideal decision boundary for classifying n-dimensional space into groups that can easily accommodate new data points is the aim of the SVM method. The limit of the best course of action is symbolized by a hyperplane. Hyperplanes, extreme points, and vectors are chosen using SVM. The method is known as SVM because support vectors are an exceptional case.
- ***The decision tree (DT)*** is a popular and effective method for facing the challenges in classification and forecasting models [14]. The decision tree's nodes and branches represent feature tests; each node in the tree signifies a trial, and each branch signifies the outcome of the study. A resultant class label appears at the termination of every leaf node.
- ***Ada Boost***: Adaboost [4] is a binary class improving the algorithm. Boosting model is precisely an ensemble approach that creates a sturdy classifier via way of means for combining more than one weak classifier to beautify version performance. Similarly, the AdaBoost classifier works on a similar principle to boosting. It's so-called adaptive boosting.
- ***K-Nearest Neighbors***: K-Nearest Neighbor (KNN) [5] is a popular supervised algorithm most beneficially applied in machine learning methods. Although it can be cast-off to solve machine learning problems, it is often used to solve classification problems. KNN methodology also known as lazy learning technique.
- ***Random Forest***: This has many trees which is used for making decision. A kind of supervised learning algorithm are preferred, which is efficiently used for classification and regression problems.

V. EVALUATION METRICS:

In our study, the four standard machine learning metrics were used. The below mention equations numbered from 1 to 4 describes the formula for calculating the metrics. The accuracy, recall, precision, and F1 scores were used to assess the algorithms. By means of a result, TP denotes truly positive, TN denotes truly negative, FP denotes falsely positive, and FN denotes falsely negative.

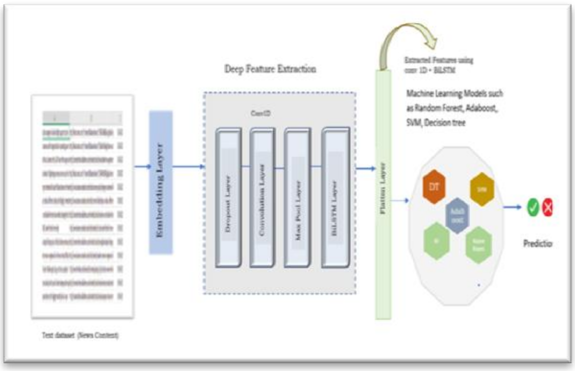


Fig 2. Proposed fake news detection Framework based on DHL

i) Accuracy is a prevalent measured state of metric for performing ML and DL models. It calculates the proportion of predicted positive observations. The formula for calculating accuracy is as follows:

Accuracy = TP + TN / TP + FP + TN + FN. (1)

ii) Ratio of the actual number of positively predicted to the number of actual positives expressed by means of precision.

Precision= TP/ TP + FP. (2)

iii) The recall value, instead of displaying the sum of truly reacted positives minus falsely reacted negatives, which displays the ratio of truly positive to the total number of truly positives and the false negatives.

Recall=TP / TP + FN. (3)

iv) Finally, the F1 score is calculated as the weighted average of the accuracy and recall values. F1 is thought to be most useful when the data from different classes is uneven.

F1_score =2*Precision *Recall /Precision + Recall. (4)

V. RESULTS AND DISCUSSION

The discussion of the findings produced via the machine learning Models employed in the second half of the proposed hybrid model is shown in table

1 using Accuracy, Precision, Recall, and F1 scores. The findings were calculated using the mean of five segments of the K-fold validation. They are calculated individually examined to find the fake data. Because of the current asymmetry of genuine and false information in the training set, fake data is validated separately [15]. Because our dataset contains both genuine and false news, the precision and recall attributes can be confusing at times, as the count of false negatives and false positives can be quite high. As a result, the F1 score and accuracy better reflects an algorithm's effectiveness because it takes into account both recall and precision.

For the dataset 1, all the models considered in our work attained an identical F1 score of above 95 percent. The Random Forest algorithm had the highest F1 score, 95.66 percent, when fake labelled dataset 1 was used.

Furthermore, we obtained the lowest value of 87.60 percent using the Standalone Deep Neural Network (DNN) algorithm. If the dataset contained a larger set of fake data, the efficiency can be enhanced. Despite the fact that precision and recall values can be misleading at times, they are still scrutinized in our research.

Although for the dataset 2 the Accuracy value collected was extremely scoring nearly 99 percent for (Deep hybrid learning_ Random Forest (DHL-RF), the precision value of 99.02 percent achieved by Random Forest was the highest. In both cases, the number of epochs for training the model and generating features is the same (10 epochs), and the other hyper parameter tuning is consistent. The results obtained on the dataset1 and 2 using the deep neural network (DNN) model evaluation metrics of Accuracy, Precision, Recall, F1 Score, and the confusion metrics appear to be low promising. For the training data, the overall training time for 10 epochs was nearly 15 minutes. Whereas, in case of DHL with all algorithms yields better results as an increase in almost all parameters by nearly 7–8%, which is a significant improvement in performance, and the time taken was 9 minutes which shows a promising result compared to DNN. Overall, it appears that the Deep Hybrid

Network model outperformed the standalone DNN model in terms of all metrics and model training time. And the DHL-Random Forest variant outperformed the others. VI.

CONCLUSION AND FUTURE WORKS

Our results show that exploiting a 1D convolutional network for feature extraction pursued through conventional machine learning algorithms on classifications generates better outcomes than using only distinct neural network model such as CNN [7]. Furthermore, the extraction of features through the convolutional network along with BiLSTM contribute to better results than classic DNN model. One of the prime and prominent motives for Deep Learning's fame is that it eradicates the need for manual feature engineering on unstructured data, which is enormously difficult and on which almost all traditional machine learning algorithms rely. Understanding the dataset and the capacity to do characteristic engineering on the dataset have resulted in the remaining effectivity and accuracy of the algorithm in typical computing device gaining knowledge of methods. The closing classification or clustering layer of a Deep Learning mannequin pushed with the aid of thoroughly related neural community layers, on the different side, may also lead in over-fitting when fed "fewer" data, or even most of the time, such fashions contain vain use of computational energy and resources, which are now not current in classical desktop studying algorithms. This has been achieved using Deep Hybrid Learning, which acts as a resultant fusion network created by combining Deep Learning and Machine Learning, has achieved this. Thus, sing Deep Hybrid Learning (DHL) to obtain the advantages of each DL and ML, one may want to alleviate the boundaries of each technique and furnish greater correct and much less computationally high-priced solutions. However, the lack of fake news in our dataset slowed the learning experience of the concepts used significantly. As a result, the ample supply of fake news will aid in the effectiveness of research development. We aim to generalize our framework to multimedia dataset in the future because the convolutional network possesses the burden of extracting features from data, which relieves one of the burdens of understanding the data.

REFERENCES

[1] . W.H. Organization, Geneva, Switzerland (2019). https://www.euro.who.int/en/health-topics/health emergencies/coronavirus covid19/news/news/2020/3/who-announces-covid-19-outbreak-a-pan demic.

[2] T. L. Huynh. (2020). "The COVID-19 risk perception: A survey on socio economics and media attention,", in "Economic Bulletin", vol. 40, pp. 758–764.

[3] P. C. Bala et al. (2020). "Automated markerless pose estimation in freely moving macaques," in "Nature Communications", vol. 11, Article ID 4560.

[4] Tu Chengsheng et al (2019) “AdaBoost typical Algorithm and its application research”. “MATEC Web of Conferences 139”, p. 00222.

[5] Gongde Guo et al. (2003). “KNN Model-Based Approach in Classification”. “An International conference of Springer Berlin Heidelberg”, pp. 986–996.

[6] Jamal Nasir et al. (April 2021). “Fake news detection: A hybrid CNN-RNN based deep learning approach”. In: “International Journal of Information Management Data Insights “1, pp. 100007.

[7] Warren S McCulloch and Walter Pitts. “A logical calculus of the ideas immanent in nervous activity”. In: “The bulletin of Mathematical Biophysics” 5.4 (1943), pp. 115–133.

[8] Xiaobo Huang. (2018) “Convolutional Neural Networks in Convolution.”. arXiv: 1810.03946.

[9] Kim, Yoon. (2014). “Convolutional Neural Networks for Sentence Classification”, in “Proceedings of the ‘14 Conference on Empirical Methods in Natural Language Processing. 10.3115/v1/D14-1181.

[10] Martín Abadi et al. (2015). “TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org”.

[11] Diederik P Kingma et al. (2014). “Adam: A method for stochastic optimization”. In: arXiv preprint arXiv:1412.6980.

[12] Zhilu Zhang et al. (2018). “Sabuncu. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels”. arXiv: 1805 . 07836.

[13] Vladimir Vapnik et al. (2018). “Support vector networks”. In: Machine learning 20.3, pp. 273–297.

[14] Yan-Yan Song and LU Ying. (2015). “Decision tree methods: applications for classification and prediction”. In: “Shanghai archives of psychiatry 27.2, p. 130”.

[15] Md Zobaer Hossain et al. (2020). “BanFakeNews: A Dataset for Detecting Fake News in Bangla”. “arXiv: 2004.08789 [cs.CL]”.

[16] Mohammed Al-Sarem et al. (2021). “A Novel Hybrid Deep Learning Model for Detecting COVID-19-Related Rumors on social media Based on LSTM and Concatenated Parallel CNNs”. In: “Applied Sciences” 11.17.

[17]] T. Hossain. et al. (2020) “COVID: detecting COVID-19 misinformation on social media,” in “Proceedings of the 1st Workshop on NLP for COVID-19”

[18] P. Patwa, S. Sharma, S. Pykl et al., “Fighting an infodemic: COVID-19 fake news dataset,” arXiv:2011.03327, 2020.

[19] Vasu Agarwal et al. (2019). Analysis of classifiers for fake news detection. Procedia Computer Science, 165:377–383.

[20] Pritika Bahad et al. (2019). “Fake news detection using bi-directional lstm recurrent neural network. “Procedia Computer Science”, 165:74–82.

[21] Payam Refaeilzadeh, et al. (2009) “Cross-Validation”. In: Encyclopedia of Database Systems”. Ed. by Ling Liu “Springer”, pp. 532–538.

TABLEII Fake News detection Results for dataset I

MODEL	ACCURACY	PRECISION	RECALL	F1-SCORE
DHL-Support Vector Machine	93.10	94.04	94.89	95.23
DHL-Decision tree	92.93	97.02	96.85	95.09
DHL-Ada Boost	94.01	96.06	97.41	95.32
DHL-K-Nearest Neighbors	97.04	98.40	97.0	95.20
DHL-Random Forest	**99.41**	**99.02**	**97.77**	**95.66**
Standalone DNN	92.31	93.90	90.91	95.01

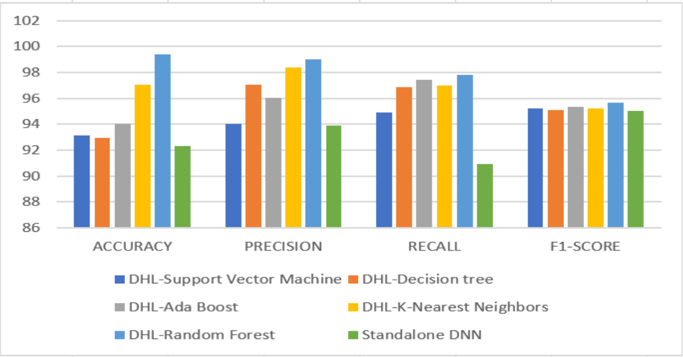


Fig 3. Performance analysis for Dataset I

TABLE III Fake News Dataset Results for Dataset

MODEL	**ACCURACY**	**PRECISION**	**RECALL**	**F1-SCORE**
DHL-Support Vector Machine	94.10	95.4	95.81	93.12
DHL-Decision tree	97.39	94.21	95.81	98.10
DHL-Ada Boost	96.11	97.6	95.47	97.11
DHL-K-Nearest Neighbors	97.42	99.0	97.12	99.22
DHL-Random Forest	99.10	98.99	98.00	97.66
Standalone DNN	89.97	91.02	90.11	94.02

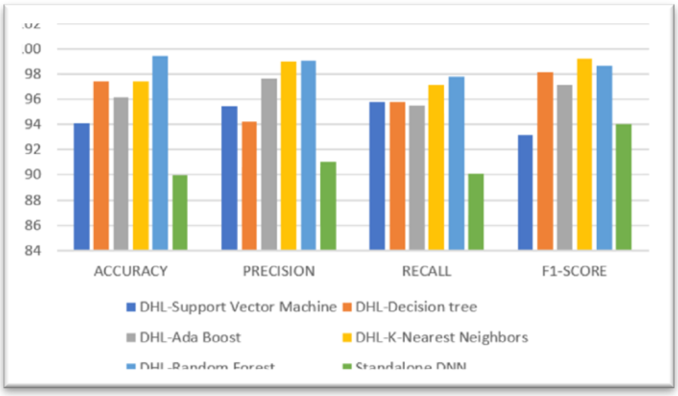


Fig 4. Performance analysis for Dataset I