IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 November 2023, accepted 8 December 2023, date of publication 14 December 2023,
date of current version 5 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3342843

RESEARCH ARTICLE

Combating Fake News on Social Media: A Fusion Approach for Improved Detection and Interpretability

YASMINE KHALID ZAMIL AND NASROLLAH MOGHADDAM CHARKARI
Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran 11366, Iran
Corresponding author: Nasrollah Moghaddam Charkari (charkari@modares.ac.ir)

ABSTRACT The proliferation of fake news on social media prompted research groups to develop statistical and learning methods to combat this menace. Deep learning techniques could not model and improve in terms of adopting multi-transformer topologies, enhancing interpretability, and coping with uncertainty. This article suggests a fusion strategy to create a more reliable fake news detection (FND) model by fusing text and image features. The different combinations of information in single and multi-modalities have been investigated to find optimal conditions. In this paper, we have employed pre-trained models of Electra and XLnet for text feature learning. Furthermore, ELA has been used to highlight the modified image features and EfficientNetB0 for image learning. To enhance the interpretability of the proposed model, the superpixels contributing to its interpretability are identified using the Local Interpretable Model-agnostic Explanations (LIME). Three well-known datasets (Weibo, MediaEval, and CASIA) have been used in this study. The results show that employing ELA and LIME in conjunction with the fusion of text and image features provides a solid and understandable solution to the FND issue in social media compared to other techniques.

INDEX TERMS Social media news, fake news detection, error level analysis, efficientNetB0, LIME.

I. INTRODUCTION

Information has been disseminated to a greater extent than ever before because of the advent of the World Wide Web and the rapid spread of social media platforms like Facebook and Twitter. At the same time, the fast spread of fake news through social media platforms has become a major concern [1], [2], [3]. A fake news platform does its best to use sensationalist and highly negative terms in its content [3], [4]. Since the majority of forms of online news like text, video, and audio are unstructured. Detecting fake news on social networks is difficult as it takes a high level of human expertise to detect and categorize them [5], [6]. Humans might fail to detect misleading news fields [7]. Thus, artificial intelligence (AI) based systems for FND would be needed.

Transformer-based pre-trained language model (PTM), which is also known as self-supervised learning, have recently been adopted in various natural language processing (NLP) and language-modeling problems because of their ability to achieve maximum performance while addressing various problems [8], [9]. A paradigm change occurred with PTM, changing the supervised learning approaches to pretraining ones. The new transformer neural network design for modeling language is based on an encoder–decoder structure [10]. Some popular PTM examples are BERT, Transformer-XL, XLNet, RoBERTa, DistilBERT, BART, MobileBERT, ELECTRA, and BigBird. Nearly every masked language model exhibits generator-like behavior, unlike ELECTRA's discriminator-like behavior. In general, these models require less time and computational resources than others [11], [12], [13].

An ever-expanding pre-trained model has to be trained with a well-selected, linguistically enriched vocabulary on vast volumes of language or domain-specific data [3], [9], [14]. As a result, FND is still faced with the challenges of developing a more effective design to replace or enhance

The associate editor coordinating the review of this manuscript and approving it for publication was Kostas Kolomvatsos.

transformers. The strengths of each language model can be used to overcome the requirements of massive data. Therefore, the benefits of multi-lingualism could be obtained by the combination of the pre-trained LMs to maximize their efficiency [9], [15]. A multimodal framework, a more comprehensive approach to fake news detection, effectively combines texts with audio, images, and other modalities. This hybrid model has shown a higher performance [7], [10], [11]. On the other hand, ensemble fusion is an efficient approach for getting around a PTM's restrictions by combining multiple PTMs into a single framework [16], [17]. For FND tasks, pre-trained word embeddings have been assembled using deep learning (DL). However, the lack of transparency and proper explanation of these methods causes a major challenge [18], [19]. Feature extraction and analysis have a vital role in deep learning models. Proper feature descriptions might capture the qualities of the photos and videos, but a single feature can no longer satisfy the demands of a variety of images and videos. To obtain more information from the images, it is desirable to fuse various feature descriptors. Data fusion refers to the process of combining the results of some classifiers, each of which operates on a different set of feature sets. A single depiction may not completely show the fundamental qualities of the data since it only offers one cue and summarises the information. The performance of a pattern recognition system is expected to improve if more global and local features can be extracted and combined. Essentially, dimensionality reduction strategies save the computational time complexity while mitigating the ''curse of dimensionality.''

This paper makes a significant contribution by proposing an improved framework for FND with multimodal awareness. Accordingly, a multimodal fusion strategy by fusion of text and image features using DL architectures is proposed. Also, Local Interpretable Model-agnostic Explanations (LIME) for the interpretability of the FND model with uncertainty handling is used. To investigate the efficiency of the method, three popular datasets have been used in the experiments that are Weibo Fake News Corpus, Twitter MediaEval Dataset (2015), and CASIA.

The main contribution of this study are as follow:

1) Development of a DL-based fused ensemble multi-modal architecture for FND.
2) Text features are extracted using an ensemble of two PTM's, XLnet and ELECTRA.
3) The proper extraction of image features using ELA and an enhanced EfficientNetB0.
4) Add interpretability and confidence to the proposed model using LIME.

Section II, discusses some related works on fake news. The formulation of the problem and the methodology employed in this research are elaborated in Section III, the results of the experiment are depicted in Section IV. Finally, Section V gives some concluding remarks on this study.

II. RELATED WORK

In the early studies of FND, researchers relied mainly on a single modality. However, since the number of images and videos has increased exponentially in social networks in the last decade, multimodal approaches have become the main focus of studies on fake news detection. On the other hand, statistical studies indicate that the volume of data created and consumed from 2015 to 2023 on social networks has increased from 15.5 to 120 zettabytes. Machine learning has been utilized to tackle these huge amounts of data to detect fake news. Furthermore, DL methods are well-suited for large-scale problem-solving. The following discusses some studies on DL methods for FND. The techniques for identifying fake news highly depend on the trained datasets, limiting the generality of the method. To tackle this problem, exploring the explainability of the methods in detecting fake news helps to understand the reasons behind the performance of the models and improve the transferability. Nevertheless, the interpretable examination of fake news detection is still in its initial phases [20], [21].

Jing et al. [22], introduced a progressive fusion network to detect fake news that incorporates various modalities of media content, including text and images. Consequently, representative information of each modality is captured and fused to find fake news. Wuet et al. [23] presented MUFFLE, a framework aimed at capturing multimodal dynamics and predicting the popularity of false news on social media. Obaid et al. [24] employed a group of deep learners to identify fake news. Mallick et al. [25] introduced a collaborative deep learning model for FND that estimates the level of trust in news based on user feedback and ranks them accordingly. Song et al. [26] used CNN for textual features and the Cross-modal Attention Residual Network (CARN) for visual features. Singh and Sharma [27] proposed a method that empowered the Roberta model for text and CNN for visual feature extraction. The extracted features were used as a vector to detect fake news. Xinyi et al. [28] proposed SAFE (Similarity-Aware Multimodal Fake News Detection) based on textual and visual features to detect fake news. The model computes the cosine similarity between text and image to predict the type of news. Wang et al. [29] introduced a model consisting of three components: extracting text and image features from posts, a fake news detector and an event discriminator. Zhang et al. [30] proposed a model that leverages BERT for textual and VGG-19 for image features to detect fake news on the post. Tanwar and Sharma [31] proposed a method based on an ''encoder, decoder, and fake news detector. They used VGG19, ResNet50, and InceptioenV3 to extract image features and Bi-direction LSTM for textual features. These features were fused to detect fake news on social media. Similarly, Tuan and Minh [32] used a pre-trained BERT and VGG-19 model to extract textual and visual features and fuse them as a vector to detect fake news in the post.

TABLE 1. Earlier studies about FND using deep learning models.

Ref.	Year	Multi-transformer	Pre-train image feature extraction	Fine-tuning and topology change	Early fusion	Interpretability - model	Uncertainty handling
[22]	2023	x	Swin Transformer	x	x	x	x
[36]	2022	x	CNN	✓	✓	✓	x
[24]	2022	x	CNN	x	✓	x	x
[38]	2022	x	VGG-19	x	✓	x	x
[39]	2022	x	CNN	✓	✓	x	x
[26]	2021	x	Crossmodal Attention Residual Network	x	✓	x	x
[40]	2021	x	VGG16,Resnet-50	x	✓	x	x
[41]	2021	x	ABM, CNN, RNN	x	✓	x	x
[42]	2021	x	ResNet	✓	✓	x	x
[37]	2021	x	CNN	x	✓	✓	x
[43]	2021	x	ResNet 50	x	✓	x	x
[32]	2021	x	VGG-19	x	✓	x	x
[44]	2020	x	VGG	✓	✓	x	x
[30]	2020	x	VGG-19	x	✓	x	x
[28]	2020	x	Text-CNN	x	✓	x	x
[31]	2020	x	concatenation of visual latent features from three CNN (VGG19, ResNet50, InceptionV3)	x	✓	x	x
[20]	2020	x	VGG19	x	✓	x	x
[45]	2019	x	VGG19	x	✓	x	x
[29]	2018	x	VGG19	✓	✓	x	x
Proposed method	2023	✓	EfficientNetB0	✓	✓	✓	✓

Lu and Li [33] suggested neural network-based GCAN to determine whether a tweet is false or not. Ge et al. [34] proposed a new neural network-based model to identify fake news. The model employs a Gumbel-Max trick in its hierarchical co-attention selection mechanism, facilitating the capture of sentence and word-level information. Fu et al. [35] investigated the phenomenon of feature drift in FND and suggested a new sampling method to explain the cause of feature drift. The authors used the interpretable model to verify the existence of feature drift. To explain multimodal fake news detection, Giri et al. [36], [37] demonstrated a model to detect false information in news articles by leveraging machine learning and deep learning methods for textual and visual components. They also proposed a personalized convolutional neural network with an attention mechanism to identify manipulated images shared through microblogging platforms.

A summary of the literature survey is depicted in Table 1. Each work is presented with various criteria, including using the multi-transformers, pre-training of image feature extraction, finetune, topology change, early fusion, the interpretability model, and uncertainty handling. Differences between the proposed method and other related ones are shown as well.

As is found in Table 1, multi-transformers have not been employed in any of the previous studies. Multi-transformers could solve the image-text gap in FND efforts. It has been shown that multi-transformers improve the performance of language tasks by enabling the model to respond to several components of the input at various degrees of granularity. Multi-transformers allow multimodal models to capture more complicated and delicate interactions between modalities. In addition, none of the studies addressed the problem of handling uncertainty in interpretability. Managing uncertainty is a crucial task since it enables the model to show the uncertainty of prediction. It helps the model avoid those predictions that are erroneous or misleading.

III. PROBLEM FORMULATION AND METHODOLOGY

There are two main methods to identify fake news:

- At the post or tweet level to determine whether a single post is fake or not,
- At the event level, to determine whether a news item with text or image is fake or not.

The problem of FND could be considered as a classification task. A function $F : (I, T) \implies Y$ maps the image and text feature vectors (I, T) to the corresponding labels Y to minimize the classification error. Let $D = \{x_i, y_i\}$ be a dataset of N instances where x_i represents the ith instance with image feature vector I_i and text feature vector T_i. $y_i \in \{0, 1\}$ is the corresponding binary label that shows whether the news is fake or real. Given an image I, text T, trained classifier F, and a target class k, the prediction $F(I, T)$ is done by learning an

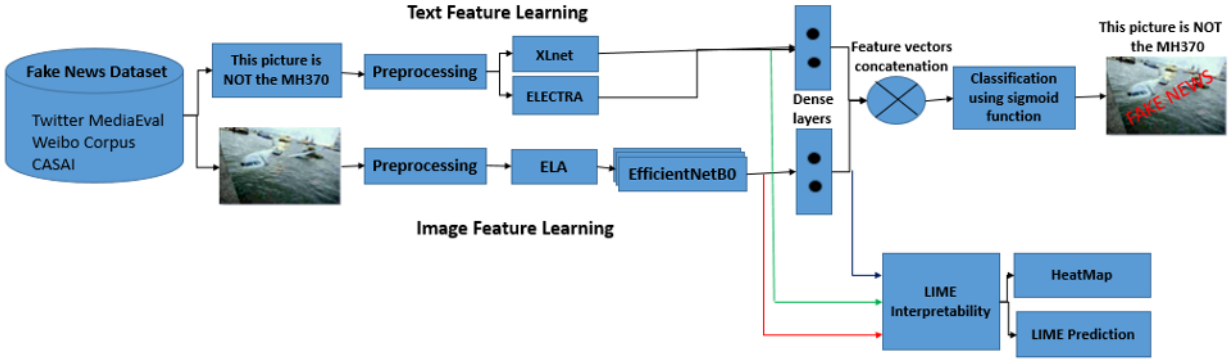

FIGURE 1. Framework of the proposed model for FND.

interpretable model $INt(F(I, T), I, T)$. Mathematically, the aim is to find the optimal function, F^* that minimizes the empirical risk $R(F)$ defined as:

$$F^* = \arg\min(R(F)) = \arg\min\left(\frac{1}{N}\sum_{i=1}^{N} L(y_i, F(I_i, T_i))\right) \tag{1}$$

where $(y_i, F(I_i, T_i))$ denotes the loss function determining the inconsistency between the true label y_i and the predicted label $F(I_i, T_i)$ for the i^{th} instance.

Figure 1 demonstrates the framework of the proposed fusion approach that combines text and image features to deliver a more robust model for FND on social media. This method has been evaluated on three popular datasets: Weibo, CASAI, and MediaEval. After pre-processing, pre-trained models Electra and XLnet are used in the text learning stage. ELA is undertaken to highlight the manipulated image features. Then, the EfficientNetB0 model would be utilized for image learning. A single feature vector is created for each sample by concatenating these feature vectors. Finally, the classification of the news as real or fake would be done using the sigmoid function. The uncertainty handling is done to improve the robustness of the model. Finally, the LIME method is employed to enhance the interpretability of the proposed model.

A. FAKE NEWS DATASET

Multimodal fake news datasets are input into the model in this study, including text and images. To assess how the model could properly distinguish fake news from real ones, we have investigated its ability to properly separate between real and fake news on three publicly available datasets (CASIA, Weibo, and Mediaeval). In this regard, those instances that consist solely of video were eliminated. The CASIA 2.0 dataset comprises 12,613 images with 900 × 600 resolution in BMP, TIFF, or JPG format, 7491 real images, and 5122 fake images [46]. Few of these images were altered by cropping and resizing. Table 2 depicts the CASIA 2.0 dataset used for training and testing after preprocessing in this study.

TABLE 2. CASIA dataset used for training and testing after preprocessing.

Images	NO.
Training through real images	4224
Training through fake images	4296
Testing through real images	827
Testing through fake images	2976

TABLE 3. MediaEval dataset used for training and testing after preprocessing.

Images	No.	Tweet	No.
Training through real images	176	Training through a real tweet	5008
Training through fake images	185	Training through a fake tweet	7032
Testing through real images	17	Testing through a real tweet	1217
Testing through fake images	33	Testing through a fake tweet	2564

The task of verifying multimedia usage is to automatically differentiate between real and fake news on Twitter, specifically multimedia content. Each data has a tweet, an associated image or video, and some details about the user. This study only retained text-based materials that have attached images. The Twitter MediaEval Dataset [47] was curated as part of the '2015 Verifying Multimedia Use Challenge' conducted by MediaEval. The task involved identifying whether the multimedia items accompanying a tweet reflect reality in the way purported by the tweet. The dataset consists of 17,000 distinct tweets with corresponding photos gathered from some well-known events or news items [48]. The dataset is split into a training set (consisting of 9,000 tweets) and a test set (2,000 tweets), each with its unique set of 6,000 genuine and 9,000 fake news tweets, respectively. The MediaEval dataset for training and testing after preprocessing is shown in Table 3.

The Weibo dataset is based on data from a microblogging platform and the credible Chinese news source Xinhua News Agency. The false content and photographs were taken between 2012 and 2016. All posts were translated into English as they were originally in Chinese. Duplicate images were eliminated to maintain the quality of the dataset while ensuring the final dataset was completely multimodal by excluding posts without images. Table 4 shows the statistical data for the Weibo dataset.

TABLE 4. Weibo dataset used for training and testing after preprocessing.

Tweet	No.
Training through a real tweet	3783
Training through a fake tweet	3749
Testing through a real tweet	3783
Testing through a fake tweet	996

B. DATASET PREPROCESSING

A post is a multimodal content composed of text and an attached image. In the preprocessing stage, an augmentation technique was employed for the image, and then the RGB image was changed to an image using error-level analysis. The image dataset is augmented using various augmentation techniques like standardization, shifts, rotation, and brightness changes [49]. It has been shown that using any image processing filter improves the generalization ability and speeds up the convergence of DL networks [50]. All photos were resized to 128×128 pixels. The ELA image was utilized to highlight the compression features in the image to extract visual features [51]. Several steps have been made to clean the text such as punctuation removal, removal of numbers, spelling correction, lowering the text by converting the text into the same case, and stopping word removal. Text normalization aimed to sanitize the text by substituting each word with its appropriate canonical form. The Chinese dataset was also translated into the English language. After the preprocessing stage, the cleaned texts were converted into vectors. Next, the sentences were tokenized in the text data, any punctuation marks and stop words, and the standardized text dimensions were in uniform size.

C. FEATURE LEARNING

Once the textual and visual feature representations are acquired, the feature fusion technique combines them to generate a shared representation. This section of the proposed architecture is composed of two stages of fusion. The first fusion of the text models is to capture the best contextual feature representations of the words, which employs different text features such as Electra features, XLnet features, and the fusion (Electra +XLnet) features. The second fusion of multimodal fusion works on the fusion of text (F_{T_i}) and image (F_{I_i}) features. In the fusion of text models, after obtaining the output vectors from the (Electra and XLnet) models, the combinations of transformers will be investigated to capture the rich contextual features to enhance the representation. Generally, information in visual form is conceptually acquired and comprehended by the human brain far quicker than the information in textual form [52]. ELA was undertaken to highlight the manipulated image features, and then the EfficientNetB0 model was utilized for image learning. The output in all cases from the last three layers is as follows:

$$F_{(I_i)} = \varnothing_f(W_i F_{B0}) \tag{2}$$

where $\varnothing_f$ represents the activation function, W_i the weights of each layer in EfficientNetB0, and F_B0 is the layer's output.

Assuming the tweet sentence $T = \{W_1, W_2, \ldots, W_n\}$, where W denotes the word in the tweet T and n is the number of words in the tweet sentence, to get a tokenizer for each tweet sentence

$$t_n = \text{idx}(T) \tag{3}$$

where *idx* represents the (Electra and XLnet) models, t_n is the concatenation of word embeddings of a text. The tokenizing of the tweet text is passed to (Electra and XLnet) models to generate the learning features as follows:

$$F_{T_i} = \emptyset_f(W_i F_{EX}) \tag{4}$$

where $\emptyset_{fe}$ represents an activation function, W_i is the weight of the last dense layer, and F_{EX} is the output from any of the transformer (Electra, and XLnet) stacked layers. Consequently, the obtained two feature vectors (F_{T_i}, F_{I_i}) through different modalities were fused.

D. MULTIMODAL CLASSIFICATION

To perform the final prediction, F_{T_i} and F_{I_i} are concatenated and fed to the sigmoid layer, leading to non-linearity in the model. It permits the DL model to learn more complex decision boundaries. The fake news predictor is defined as follows:

$$l_i = \text{sigmoid}(W[F_{T_i} * F_{I_i}] + b) \tag{5}$$

where W represents the parameters of the sigmoid layer, F_{I_i} it a feature of EfficientNetB0, F_{T_i} it can be the feature of Electra, the feature of XLnet, or fusion (Electra+XLnet) features, b is the bias term, and l_i is a list containing two elements $l_i = [l_0, l_1]$.

We adopted binary cross-entropy to define the loss function $CE(\emptyset_{fe})$ for each news by learning ($\emptyset_{fe}$) through back-propagation. Let l_0 and l_1 represent the probability of a given news being real (0) or fake (1).

$$CE(\emptyset_{sm}, \emptyset_{fe}) = -l_i \log(l_1) - (1 - l_i) \log(l_0) \tag{6}$$

E. INTERPRETABILITY MODEL

The relevance of each pixel is quantified as its magnitude. Adjusting function values in a single data sample, LIME observes the impact on performance, similar to human observation of model performance. LIME separates an interpretable representation from the original feature space of the model to ensure comprehensibility [53]. This model computes the interpretability of text, image, and multimodal models. It is defined as a model explanation $g \in G$ that belongs to the class of potentially interpretable models [40], where g acts on the presence or absence of interpretable components over the $0, 1^{d'}$ domain. However, not all g in G might be simple to interpret. $\Omega(g)$ is a measure to show the confidence of $g's$. To determine the proximity between an instance z and x, $\pi_x(z)$ is used as a proximity measure. LIME produces explanations using:

$$\xi(x) = \arg\min L(f, g, \pi_x) + \Omega(g) \tag{7}$$

TABLE 5. Hyper- parameter settings for EfficientNetB0.

Hyperparameter	Values
Optimizer	Adam
Learning rate	10^{-6}
No. of dense layers	2
Dropout	0.5
Batch Size	32
Epochs	10
Total parameters	4,049,571
Trainable Parameters	4,007,548
Non- Trainable Parameters	42,023

$L(f, g, \pi_x)$ is a measure to show how the inaccurate g is in the approximating f in the vicinity defined by $\pi_x(z)$. We should minimize (f, g, π_x) while having $\Omega(g)$ be low enough to be interpretable by humans. To confirm the interpretability and local accuracy, $L(f, g, \pi_x)$ must be minimized while ensuring that $\Omega(g)$ is sufficiently small for humans to understand.

The LIME methodology applies to the text or pictorial data. In the present research, we assess the interpretability of models for text, image, and multimodal data. Furthermore, the LIME methodology is implemented by including the uncertainty management techniques and generating a heatmap that visually depicts the significance or contribution of various characteristics or areas towards the classification outcome determined by the model.

IV. EXPERIMENTAL SETUP

Pretrained XLnet and Electra models for the textual aspect have been used to handle multimodal posts. The XLnet module contains 16 attention heads and 24 layers, with a hidden unit dimension of 768 per token and 340M parameters. The Electra module, on the other hand, has 4 attention heads, 12 layers, and a hidden unit dimension of 256 per token. During the training phase, we have extracted the textual content of each model alone, as well as the fusion (Electra and XLnet) models. For the image model, we used the pre-trained EfficientNetB0 image model.

For automated algorithms, hyperparameter optimization finds the ideal set of hyperparameters. Because the performance of the model relies on appropriate selecting the hyperparameters, optimization is essential. We have improved several parameters, including scale, number of neighbors, distance metrics, and kernel functions. The network achieved its highest level of accuracy by the 10th epoch, using a batch size of 32 and a learning rate of 10^{-6}, while utilizing the Adam optimizer. The stopping criterion for the training process was set to achieve the maximum accuracy metric. Table 5 shows the hyperparameter values for the proposed multimodal approach.

To evaluate the performance of the proposed model in detecting fake news, we have compared it with the following state-of-the-art models.

- EANN [29]: For fake news detection, EANN acquires event-invariant multimodal features of each post by utilizing an adversarial network. By combining the extracted textual and visual features, the network removes components that are specific to each event from the post features.
- MVAE [45]: MVAE uses a variational autoencoder with separate encoder and decoder networks for both text and image modalities to create a unified multimodal representation that can be used for detecting fake news. The auto-encoder and classifier are trained together in a joint training process.
- AMFB [41]: Used text and visual features to detect fake news.

To measure the efficiency of the proposed FND model, we enable uncertainty handling to compare the LIME results with the model results. The LIME library of Python as an interpretability model has also been employed. The experiments have been conducted in Python. These techniques offer an understanding of the model's decision-making process, ultimately improving its effectiveness in identifying and classifying fake news on social media.

A. EXPERIMENTAL RESULTS AND ANALYSIS

The different combinations of datasets in single and multiple modalities have been investigated, as shown in Table 6.

The explanation of a few of the combinations of models from Table 6 is provided here so that the remainder may be understood using the same terminology.

- $Multimodal_{m_2}$: This multimodal utilizes only XLnet transformers for text and standard EfficientNetB0 for image learning. The few images in the MediaEval dataset necessitated augmentation for this model's images.
- $Multimodal_{m_3}$: This multimodal utilizes only Electra transformers for text and standard EfficientNetB0 for image learning. The few images in the MediaEval dataset necessitated augmentation for this model's images.
- $Multimodal_{m_4}$: This multimodal utilizes Electra+XLnet transformers for text and standard EfficientNetB0 for image learning. The few images in the MediaEval dataset necessitated augmentation for this model's images.
- $Multimodal_{w_1}$: This multimodal utilizes XLnet transformers for webio dataset text and standard EfficientNetB0 for image learning without finetuning.
- $Multimodal_{w_2}$: This multimodal utilizes Electra transformers for webio dataset text and standard EfficientNetB0 for image learning without finetuning.
- $Multimodal_{w_3}$: This multimodal utilizes Electra+XLnet transformers for webio dataset text and the standard EfficientNetB0 for image learning without finetuning.
- $Multimodal_{w_4}$: This multimodal utilizes XLnet transformers for webio dataset text and standard EfficientNetB0 for image learning with finetuning.

TABLE 6. Different combinations of datasets in single and multiple modalities.

Model	Media-Eval	Weibo	CASIA	Text model			Image model					Multimodal
				Elctra	Xlnet	Elctra-Xlnet	-Aug.img	Aug.img	-F.T	F.T	F.T.S	
Visual	x	x	✓	x	x	x	x	x	x	x	✓	x
$Visual_{m_1}$	✓	x	x	x	x	x	✓	x	x	x	x	x
$Visual_{m_2}$	✓	x	x	x	x	x	x	✓	x	x	x	x
$Visual_{w_1}$	x	✓	x	x	x	x	x	x	✓	x	x	x
$Visual_{w_2}$	x	✓	x	x	x	x	x	x	x	✓	x	x
$Visual_{w_2}$	x	✓	x	x	x	x	x	x	x	x	✓	x
$Text_{m_1}$	✓	x	x	✓	x	x	x	x	x	x	x	x
$Text_{m_2}$	✓	x	x	x	✓	x	x	x	x	x	x	x
$Text_{m_3}$	✓	x	x	x	x	✓	x	x	x	x	x	x
$Text_{w_1}$	x	✓	x	✓	x	x	x	x	x	x	x	x
$Text_{w_2}$	x	✓	x	x	✓	x	x	x	x	x	x	x
$Text_{w_3}$	x	✓	x	x	x	✓	x	x	x	x	x	x
$Multimodal_{m_2}$	✓	x	x	✓	x	x	x	✓	x	x	x	✓
$Multimodal_{m_3}$	✓	x	x	x	✓	x	x	✓	x	x	x	✓
$Multimodal_{m_4}$	✓	x	x	x	x	✓	x	✓	x	x	x	✓
$Multimodal_{w_1}$	x	✓	x	✓	x	x	x	x	✓	x	x	✓
$Multimodal_{w_2}$	x	✓	x	x	✓	x	x	x	✓	x	x	✓
$Multimodal_{w_3}$	x	✓	x	x	x	✓	x	x	✓	x	x	✓
$Multimodal_{w_4}$	x	✓	x	✓	x	x	x	x	x	✓	x	✓
$Multimodal_{w_5}$	x	✓	x	x	✓	x	x	x	x	✓	x	✓
$Multimodal_{w_6}$	x	✓	x	x	x	✓	x	x	x	✓	x	✓
$Multimodal_{w_7}$	x	✓	x	✓	x	x	x	x	x	x	✓	✓
$Multimodal_{w_8}$	x	✓	x	x	✓	x	x	x	x	x	✓	✓
$Multimodal_{w_9}$	x	✓	x	x	x	✓	x	x	x	x	✓	✓

Where **XLnet+Electra** fusion XLnet and Electra, **-Aug.img**: without augmentation image, **Aug.img:** with augmentation image, **-F.T**: standard EfficientNetB0 image model, **F.T**: fine-tuning EfficientNetB0 image model, **F.T.S**: fine-tuning EfficientNetB0 with setting parameter (size image 128*128),(256 vector size).

TABLE 7. Comparing the experimental results among the multiple combinations of the proposed model.

Dataset	Method	Accuracy	Precision	Recall	F-score
MediaEval	$Multimodal_{m_2}$	93.09%	93.25%	95.84%	94.52%
	$Multimodal_{m_3}$	93.26%	93.82%	95.47%	94.63%
	Multimodal$_{m_4}$	**93.89%**	93.33%	97.13%	95.19%
Weibo	$Multimodal_{w_1}$	81.26%	79.69%	81.32%	82.14%
	$Multimodal_{w_2}$	77.11%	72.44%	83.71%	77.66%
	$Multimodal_{w_3}$	81.55%	81.90%	78.56%	80.19%
	$Multimodal_{w_4}$	78.27%	74.78%	82.46%	78.43%
	$Multimodal_{w_5}$	74.90%	71.18%	79.42%	75.07%
	$Multimodal_{w_6}$	79.69%	76.03%	83.53%	79.60%
	$Multimodal_{w_7}$	83.03%	86.99%	79.15%	82.89%
	$Multimodal_{w_8}$	79.78%	81.46%	79.24%	80.33%
	Multimodal$_{w_9}$	**85.19** %	83.24%	89.61%	86.30%

- $Multimodal_{w_5}$: This multimodal utilizes Electra transformers for webio dataset text and standard EfficientNetB0 for image learning with finetuning.
- $Multimodal_{w_6}$: This multimodal utilizes Electra+ XLnet transformers for the webio dataset text and standard EfficientNetB0 for image learning with finetuning.
- $Multimodal_{w_7}$: This multimodal utilizes XLnet transformers for webio dataset text and standard EfficientNetB0 for image learning with finetuning+setting parameters (image size 128*128 in pixels) and vector size EfficientNetB0 256 from a standard.
- $Multimodal_{w_8}$: This multimodal utilizes Electra transformers for webio dataset text and standard EfficientNetB0 for image learning with finetuning+setting parameters (image size 128*128 in pixels) and vector size EfficientNetB0 256 from a standard.

TABLE 8. Comparing the experimental results among the multiple combinations of the proposed model.

Dataset	Method	Accuracy	Precision	Recall	F-score
MediaEval	$Visual_{m_1}$ (without augmentation)	81.07%	81.61%	89.87%	85.54%
	Visual$_{m_2}$ (with augmentation)	**89.11**%	92.08%	90.29%	91.18%
Weibo	$Visual_{w_1}$ (without finetuning)	57.63%	56.29%	48.74%	52.24%
	$Visual_{w_2}$ (with fine tuning)	66.80%	93.60%	66.76%	77.93%
	Visual$_{w_3}$ (with fine tuning and setting parameter)	**80.13**%	81.40%	81.05%	81.22%
CASIA	Visual model (where we didn't have text in this dataset)	**96.11**%	98.13%	96.83%	97.42%

- $Multimodal_{w_9}$: This multimodal utilizes Electra+ XLnet transformers for webio dataset text and standard EfficientNetB0 for image learning with finetuning+setting parameters (image size 128*128 in pixels) and vector size EfficientNetB0 256 from a standard.

Table 7 shows the experimental results among the multiple components of the proposed model.

The result indicates that Multimodal$_{w_9}$ on the Weibo dataset and Multimodal$_{m_4}$ on the MediaEval datasets provide 85.19% and 93.89% in accuracy, respectively. On the other hand, the accuracy of Multimodal$_{w_i}$ ($\forall i = 1, 2, 3, 4, 5, 6$) is lower than that of the textual model due to the complement of the extracted visual features, which is not evident in the Twitter dataset. This suggests the possibility of noisy images

TABLE 9. Comparison between the proposed model and the earlier studies on the same datasets.

Dataset	Model	Accuracy	Fake			Real		
			Precision	Recall	F1-Score	Precision	Recall	F1-Score
Mediaeval(Tweeter)	Textual	80.01%	78%	67%	72%	76%	85%	80%
	Visual	89.11%	37%	72%	49%	91%	70%	79%
	EANN [29]	64.8%	81%	49.8%	61.7%	58.4%	75.9%	66%
	MVAE [45]	74.5%	80.01%	71.9%	75.8%	68.9%	77.7%	73%
	AMFB [41]	88.3%	89%	95%	92%	87%	76%	81%
	Proposed model	**93.89%**	89%	82%	80.1%	85%	91%	93%
Weibo(Chinese)	Textual	80.69%	80%	81%	81%	83%	82%	83%
	Visual	80.13%	78%	77%	78%	79%	80%	79%
	EANN [29]	79.5 %	82.7%	69.7%	75.6%	75.2%	86.3%	80.4%
	MVAE [45]	82.4 %	85.4%	76.9%	80.9%	80.2%	87.5%	83.7%
	AMFB [41]	83.2%	82%	86%	84%	85%	81%	83%
	Proposed model	**85.19%**	92%	86 84%	91.6%	92%	93.5%	

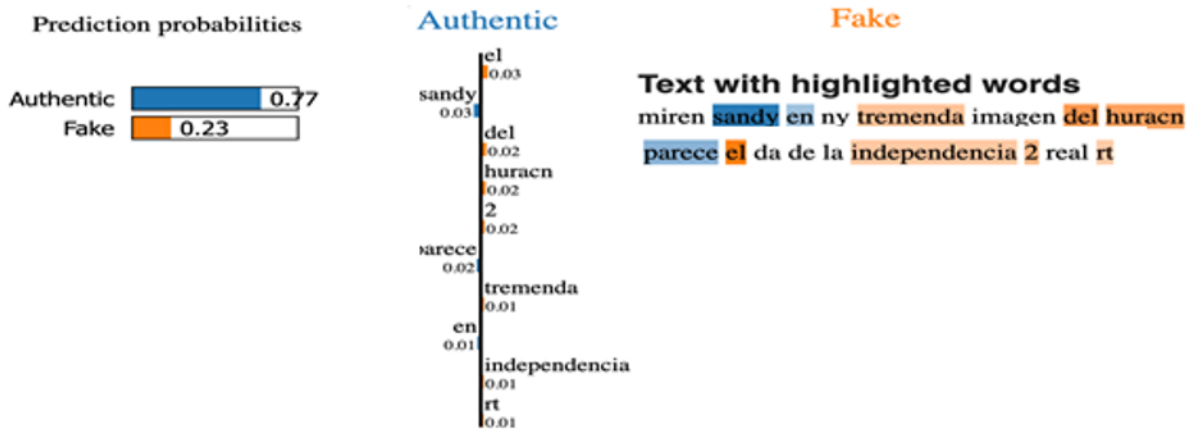

FIGURE 2. Explains the text model from the MediaEval dataset.

in the Weibo dataset, which might not be informative for detecting fake news. Multimodal features with noisy images perform poorly, as discussed in [14] and [32]. Therefore, multimodal$_{w9}$ was used in the final model to evaluate and compare it with the earlier studies.

Table 8 compares the experimental results among the multiple components of the visual model for the proposed model.

Refer to Table 8 the visual model uses the CASIA dataset containing only images with finetuning+setting parameters. Therefore, the visual model shows the best accuracy of 96.11% compared to the others.

Table 9 presents the comparative analysis of the state-of-the-art and the proposed models on three different datasets. Training and validation methodologies were kept similar for all models.

As seen in Table 9; it is found that the proposed model performs better than other models in F1 score and accuracy. The above metrics show that the proposed model better generalizes to the classification of fake and real news articles. In general, the proposed models work better than individual models, but this is not guaranteed. Combining models with incremental accuracy growth has produced a better model. Furthermore, incorporating images into text improves the performance of the model. Similar patterns are observed in the Weibo dataset when comparing the model performance to the Twitter dataset. As found, the proposed model achieves better results with the finetuning parameters compared to MVAE.

The proposed model demonstrates a higher recall value, a positive indication for fake image classification. It is worth mentioning that the performance of the multimodal model is improved when fusing the text transformer models (Electra and XLnet) with the image model (EfficientNetB0), compared to fusing only the model (Electra with EfficientNetB0) or (XLnet with EfficientNetB0). In all datasets, the majority of multimodal approaches resulted in improved accuracy compared to single-model ones. Lastly, in addition to the improved accuracy, the presented framework achieves balanced scores between the real and fake classes, indicating that the model is not biased towards one particular class, preventing systematic prejudice.

B. INTERPRETABILITY MODEL

The textual and visual features of a document are used at this stage. The following are justifications for employing multimodal data: First, different modalities display the material in various ways. Second, the validity of the data is detected using information obtained from many modalities that complement each other. Various sources use different terms depending on their expertise. However, the lack of transparency and a proper explanation of these methods presents a major challenge. There is a need for explainable ML-based fake news detection methods to obtain the trust

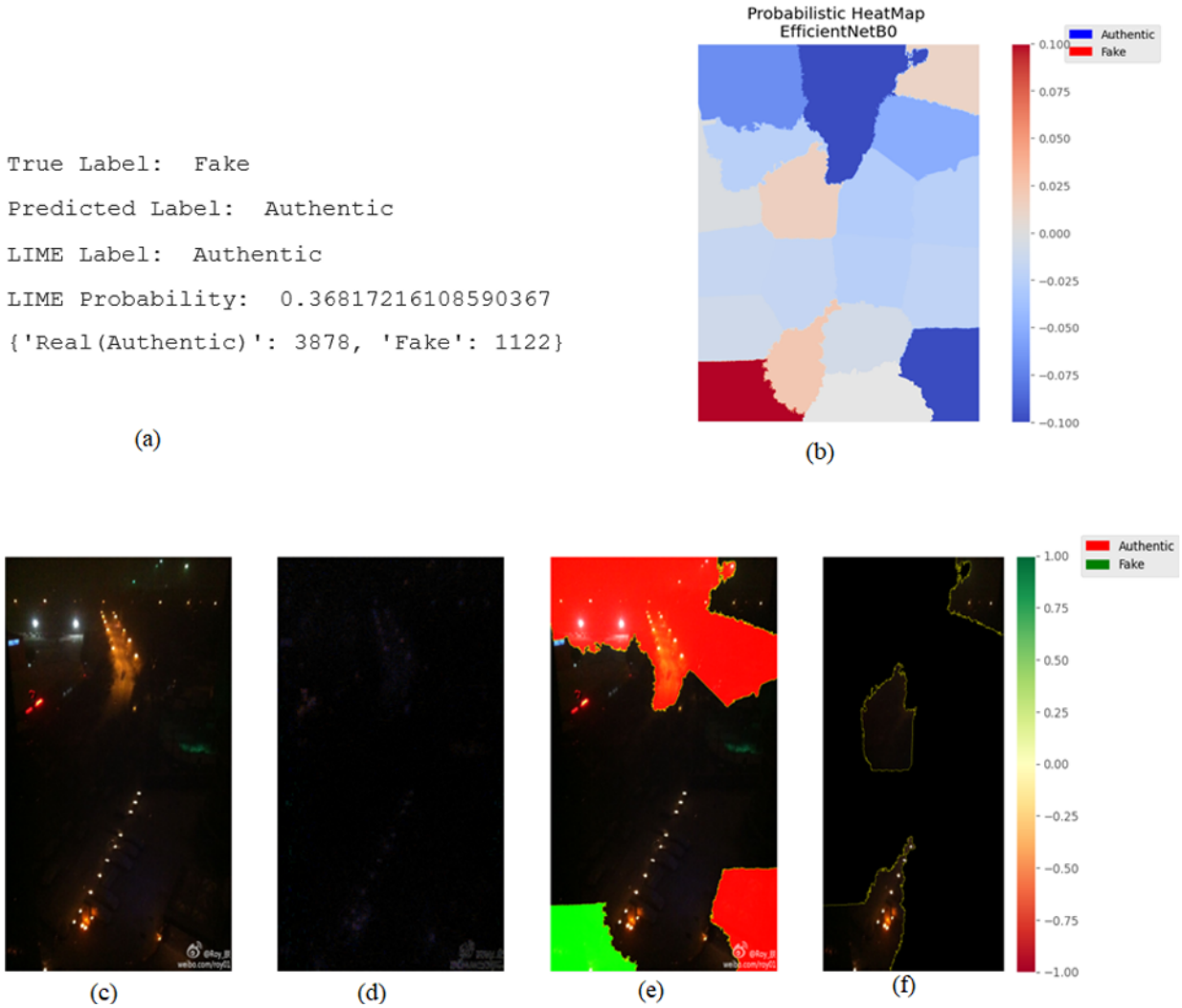

FIGURE 3. Explanation of an image model from the Weibo dataset (a) image model and probability (b) heatmap image (c) original image (d) ELA image (e) LIME image, (f) superpixels area for ELA image.

of various communities, such as journalism and security. Decision-makers in real-world scenarios require a clear understanding of the reasons behind the system's outputs. Therefore, to explain black box models, the LIME library of Python is used [17]. Using text and 3-D image data (a piece of news) and a predictor function, an explanation for the news article would be obtained. In addition, the uncertainty associated with each prediction is found. The uncertainty value can then be used to assess the overall confidence of the model's predictions and to identify instances where the model might be particularly uncertain or unreliable.

The black boxes, Electra, XLnet, and EfficientNetB0, are the models utilized for fake news detection in this study. Hence, to explain the EfficientNetB0 model, the Python LIME library is used. It takes text or image data (news) and a predictor function. On the basis of the predictor function, it returns an explanation for the news article. In the proposed model, LIME has been employed. This algorithm allowed us to determine the final classification decision based on the votes received from LIME explanations. In our work, we develop the LIME output as a heatmap.

The input and output of the LIME model are as follows: **Input:** the data frame for news (text, image), number of generated samples as input to the fusion explain function. **Output:** possible results, which are the true label related to the original dataset, the predicted label for fake news detection, the predicted label for the LIME model. Here, the LIME probability represents the probability of LIME in the predicted type news. The false news detection model makes a conclusion based on the following images: the original image from the dataset, the ELA image, the heatmap image, the LIME image, and the image with superpixel area. Since it is a probabilistic ML model, the LIME package in Python has been used to explain the text. Figure 2 displays how all of the top text attributes in the model are utilized to predict whether the supplied text article is fake or true, with their

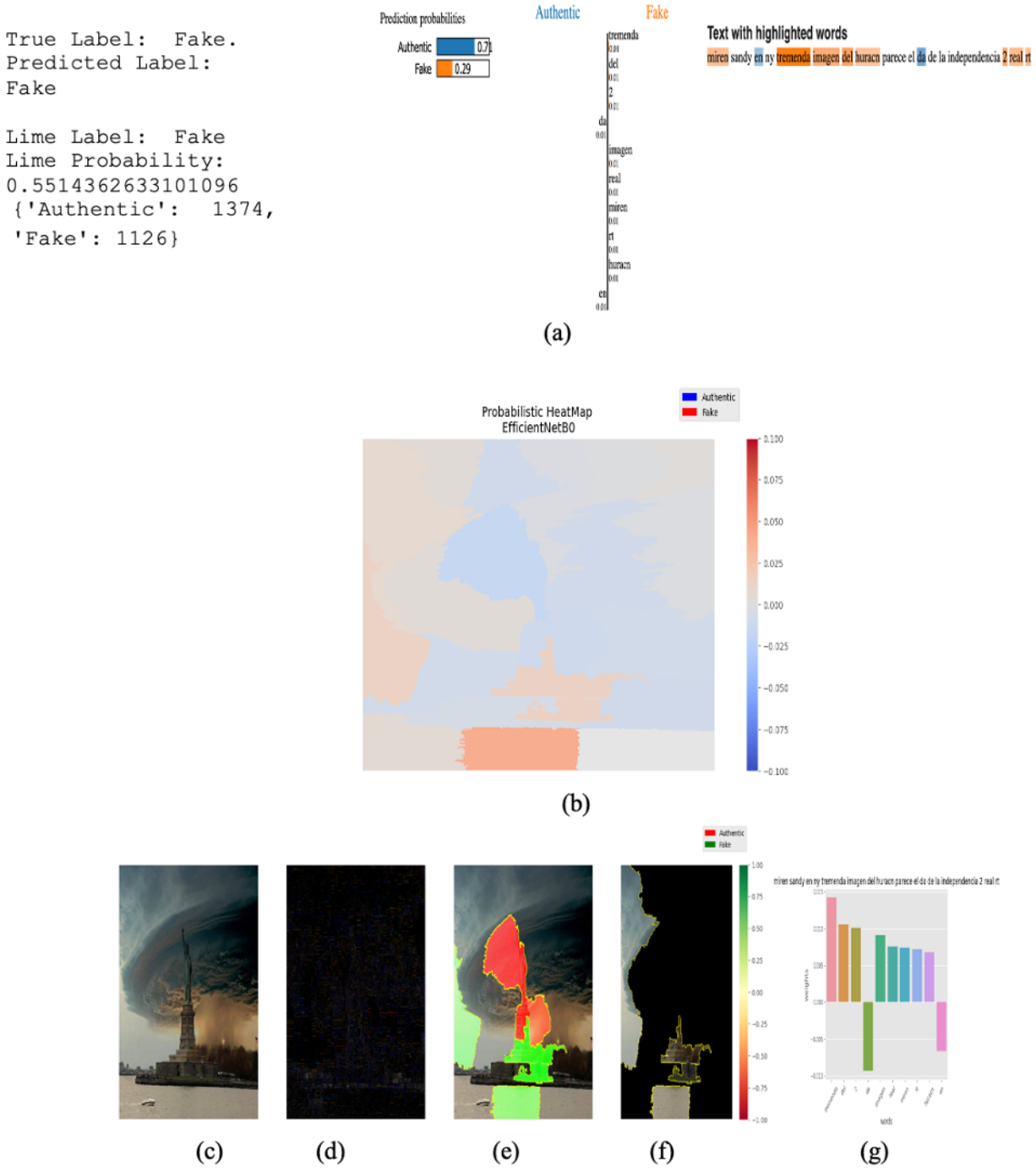

FIGURE 4. LIME multimodal Fake News Detection Model (a) the multimodal and LIME probability (b) heatmap image (c) original image (d) ELA image (e) LIME image, (f) superpixels area for ELA image, (g) superpixels text).

unique uncertainty weighting values. The likelihood of text being fake is 0.77, and real is 0.23.

Being a black box model of ML, an explanation of the image is done using the LIME library of Python. As an example, as shown in Figure 3, the prediction done by the model about the first image has been explained by the adjoining 5 images shown below. The green portion indicates the superpixels that positively contribute to the prediction of the image, and the red portion shows the superpixels that negatively contribute to the prediction of the image. Next, the same image has also been interpreted into a heatmap to show how each pixel contributes to the prediction. In Figure 3, the number of samples generated is 5000. The LIME predicted 3878 samples of images being real and 1122 samples as fake.

An explanation of the multimodal (news article) has been done using the LIME library of Python. As an example,

as shown in Figure 4, all the top text features of the text model contribute to the prediction as multimodal. Furthermore, in the image explained by the adjoining 4 images shown below, the original image in the post, the ELA image, and the green portion show the superpixels, which positively contribute to the prediction of the image. The red portion, which negatively contributes to the image prediction, and the image shows the superpixel area that contributes to making decisions from all images. Next, the same image is also interpreted as a heat map to show how much each pixel contributes to the prediction.

The LIME model for multimodal fake news detection is shown in Figure 4. From 2500 created samples, the important decision-making samples are 1374 genuine and 1126 false. The LIME label, which represents the expected class, is determined by the LIME probability value. After calculating the LIME probability, a threshold is established. The label is ''Fake'' if the LIME probability is greater than threshold t1, and ''Real'' otherwise. The threshold-based method offers a binary decision-making framework for the detection of fake news using the LIME model, enabling the classification of news pieces as genuine or fraudulent based on the LIME probability. Research indicates that multimodality-based fake news detection models outperform conventional single-modality algorithms in terms of performance. It confirms those studies (Khattar et al. [45], Song et al. [26]) that performing fake news detection using cross-modal information might improve its performance.

C. ERROR ANALYSIS OF THE PROPOSED METHOD

There are some remaining problems for wrongly detecting fake posts. Firstly, in the fake news detection model, when the image is high resolution with short associated text, a misclassification error could occur. Secondly, when the image is detached from its context, it lacks a meaningful correlation with the accompanying text. In such scenarios, the model exhibits diminished performance.

V. CONCLUSION

This paper introduces a hybrid fusion approach to improve the detection and interpretability of fake news with uncertainty handling. The existing methods ignore the trust fake news detection models that provide an accurate fake news detection model. In addition, employing multi-transformers (fusion multi-transformers) shows a higher performance than single transformers for textual feature extraction, which leads to better features that help detect fake news in the post. Based on our findings in the Weibo dataset, we discussed how the existence of noisy images leads to inferior results. The main contributions of this paper are;

1) Using a deep learning-based stacked ensemble multi-modal architecture for FND is successfully presented and tested.
2) An ensemble of two pre-trained text models, XLnet and ELECTRA are proposed and successfully tested.
3) An improved novel image feature extraction model based on EfficientNetB0 is proposed for the detection of fake news with an additional layer to improve the discrimination between real and fake news.
4) LIME was used to add interpretability and confidence to the proposed model.

In the future, we plan to propose a deep model to train the model on the LIME image to increase its efficiency in distinguishing between fake and real images.

REFERENCES

[1] M. Aldwairi and A. Alwahedi, ''Detecting fake news in social media networks,'' *Proc. Comput. Sci.*, vol. 141, pp. 215–222, Jan. 2018.
[2] O. D. Apuke and B. Omar, ''Fake news and COVID-19: Modelling the predictors of fake news sharing among social media users,'' *Telematics Informat.*, vol. 56, Jan. 2021, Art. no. 101475.
[3] P. Meel and D. K. Vishwakarma, ''Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities,'' *Expert Syst. Appl.*, vol. 153, Sep. 2020, Art. no. 112986.
[4] B. Collins, D. T. Hoang, N. T. Nguyen, and D. Hwang, ''Trends in combating fake news on social media—A survey,'' *J. Inf. Telecommun.*, vol. 5, no. 2, pp. 247–266, 2021.
[5] T. Khan, A. Michalas, and A. Akhunzada, ''Fake news outbreak 2021: Can we stop the viral spread?'' *J. Netw. Comput. Appl.*, vol. 190, Sep. 2021, Art. no. 103112.
[6] S. I. Manzoor, J. Singla, and Nikita, ''Fake news detection using machine learning approaches: A systematic review,'' in *Proc. 3rd Int. Conf. Trends Electron. Informat. (ICOEI)*, Apr. 2019, pp. 230–234.
[7] S. R. Sahoo and B. B. Gupta, ''Multiple features based approach for automatic fake news detection on social networks using deep learning,'' *Appl. Soft Comput.*, vol. 100, Mar. 2021, Art. no. 106983.
[8] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth, ''Recent advances in natural language processing via large pre-trained language models: A survey,'' *ACM Comput. Surv.*, vol. 56, no. 2, pp. 1–40, Feb. 2024.
[9] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song, ''A survey of controllable text generation using transformer-based pre-trained language models,'' *ACM Comput. Surv.*, vol. 56, no. 3, pp. 1–37, Mar. 2024.
[10] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, ''Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408820.
[11] S. Gundapu and R. Mamidi, ''Transformer based automatic COVID-19 fake news detection system,'' 2021, *arXiv:2101.00180*.
[12] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, ''MPNet: Masked and permuted pre-training for language understanding,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 16857–16867.
[13] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. Le, ''Rethinking pre-training and self-training,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 3833–3845.
[14] Z. Zhang et al., ''CPM: A large-scale generative Chinese pre-trained language model,'' *AI Open*, vol. 2, pp. 93–99, 2021.
[15] M. Zaib, Q. Z. Sheng, and W. Emma Zhang, ''A short survey of pre-trained language models for conversational AI–A new age in NLP,'' in *Proc. Australas. Comput. Sci. Week Multiconference*, Feb. 2020, pp. 1–4.
[16] G. Amoudi, R. Albalawi, F. Baothman, A. Jamal, H. Alghamdi, and A. Alhothali, ''Arabic rumor detection: A comparative study,'' *Alexandria Eng. J.*, vol. 61, no. 12, pp. 12511–12523, Dec. 2022.
[17] M. Turkoglu, D. Hanbay, and A. Sengur, ''Multi-model LSTM-based convolutional neural networks for detection of apple diseases and pests,'' *J. Ambient Intell. Humanized Comput.*, vol. 13, no. 7, pp. 3335–3345, Jul. 2022.
[18] C. K. Lee, M. Samad, I. Hofer, M. Cannesson, and P. Baldi, ''Development and validation of an interpretable neural network for prediction of postoperative in-hospital mortality,'' *npj Digit. Med.*, vol. 4, no. 1, p. 8, Jan. 2021.
[19] C.-H. Lin and O. Lichtarge, ''Using interpretable deep learning to model cancer dependencies,'' *Bioinformatics*, vol. 37, no. 17, pp. 2675–2681, Sep. 2021.

[20] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty, and P. Kumaraguru, "SpotFake+: A multimodal framework for fake news detection via transfer learning (student abstract)," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 10, pp. 13915–13916.

[21] C. Fu, Y. Zheng, Y. Liu, Q. Xuan, and G. Chen, "NES-TL: Network embedding similarity-based transfer learning," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 3, pp. 1607–1618, Jul. 2020.

[22] J. Jing, H. Wu, J. Sun, X. Fang, and H. Zhang, "Multimodal fake news detection via progressive fusion networks," *Inf. Process. Manage.*, vol. 60, no. 1, Jan. 2023, Art. no. 103120.

[23] C.-L. Wu, H.-P. Hsieh, J. Jiang, Y.-C. Yang, C. Shei, and Y.-W. Chen, "MUFFLE: Multi-modal fake news influence estimator on Twitter," *Appl. Sci.*, vol. 12, no. 1, p. 453, Jan. 2022, doi: 10.3390/app12010453.

[24] A. Al Obaid, H. Khotanlou, M. Mansoorizadeh, and D. Zabihzadeh, "Multimodal fake-news recognition using ensemble of deep learners," *Entropy*, vol. 24, no. 9, p. 1242, Sep. 2022.

[25] C. Mallick, S. Mishra, and M. R. Senapati, "A cooperative deep learning model for fake news detection in online social networks," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 4, pp. 4451–4460, Apr. 2023.

[26] C. Song, N. Ning, Y. Zhang, and B. Wu, "A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks," *Inf. Process. Manage.*, vol. 58, no. 1, Jan. 2021, Art. no. 102437.

[27] B. Singh and D. K. Sharma, "Predicting image credibility in fake news over social media using multi-modal approach," *Neural Comput. Appl.*, vol. 34, no. 24, pp. 21503–21517, Dec. 2022, doi: 10.1007/s00521-021-06086-4.

[28] X. Zhou, J. Wu, and R. Zafarani, "SAFE: Similarity-aware multimodal fake," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Cham, Switzerland: Springer, 2020, pp. 354–367, doi: 10.1007/978-3-030-47436-2.

[29] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 849–857, doi: 10.1145/3219819.3219903.

[30] T. Zhang, D. Wang, H. Chen, Z. Zeng, W. Guo, C. Miao, and L. Cui, "BDANN: BERT-based domain adaptation neural network for multi-modal fake news detection," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, vol. 1, Jul. 2020, pp. 1–8.

[31] V. Tanwar and K. Sharma, "Multi-model fake news detection based on concatenation of visual latent features," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Jul. 2020, pp. 1344–1348.

[32] N. M. D. Tuan and P. Q. N. Minh, "Multimodal fusion with BERT and attention mechanism for fake news detection," 2021, *arXiv:2104.11476*.

[33] Y.-J. Lu and C.-T. Li, "GCAN: Graph-aware co-attention networks for explainable fake news detection on social media," Apr. 2020, *arXiv:2004.11648*.

[34] X. Ge, S. Hao, Y. Li, B. Wei, and M. Zhang, "Hierarchical co-attention selection network for interpretable fake news detection," *Big Data Cognit. Comput.*, vol. 6, no. 3, p. 93, Sep. 2022.

[35] C. Fu, X. Pan, X. Liang, S. Yu, X. Xu, and Y. Min, "Feature drift in fake news detection: An interpretable analysis," *Appl. Sci.*, vol. 13, no. 1, p. 592, Jan. 2023.

[36] M. Giri, T. T. Aditya, P. Honnavalli, and S. Eswaran. (2020). *Automated and Interpretable Fake News Detection With Explainable Artificial Intelligence*. [Online]. Available: https://ssrn.com/abstract=4076594

[37] B. Singh and D. K. Sharma, "SiteForge: Detecting and localizing forged images on microblogging platforms using deep convolutional neural network," *Comput. Ind. Eng.*, vol. 162, Dec. 2021, Art. no. 107733.

[38] H. Wang, S. Wang, and Y. Han. *A Hybrid Feature Fusion Method for Fake News Detection on Chinese Social Media*. [Online]. Available: https://ssrn.com/abstract=4083983

[39] I. Segura-Bedmar and S. Alonso-Bartolome, "Multimodal fake news detection," *Information*, vol. 13, no. 6, p. 284, Jun. 2022.

[40] D. K. Sharma and S. Garg, "IFND: A benchmark dataset for fake news detection," *Complex Intell. Syst.*, vol. 9, no. 3, pp. 2843–2863, Jun. 2023.

[41] R. Kumari and A. Ekbal, "AMFB: Attention based multimodal factorized bilinear pooling for multimodal fake news detection," *Expert Syst. Appl.*, vol. 184, Dec. 2021, Art. no. 115412.

[42] L. Ying, H. Yu, J. Wang, Y. Ji, and S. Qian, "Multi-level multi-modal cross-attention network for fake news detection," *IEEE Access*, vol. 9, pp. 132363–132373, 2021, doi: 10.1109/ACCESS.2021.3114093.

[43] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, "Detecting fake news by exploring the consistency of multimodal data," *Inf. Process. Manage.*, vol. 58, no. 5, Sep. 2021, Art. no. 102610, doi: 10.1016/j.ipm.2021.102610.

[44] J. Zeng, Y. Zhang, and X. Ma, "Fake news detection for epidemic emergencies via deep correlations between text and images," *Sustain. Cities Soc.*, vol. 66, Mar. 2021, Art. no. 102652.

[45] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. World Wide Web Conf. (WWW)*, May 2019, pp. 2915–2921.

[46] J. Dong, W. Wang, and T. Tan, "Casia image tampering detection evaluation database," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process.*, Jul. 2013, pp. 422–426.

[47] C. Boididou, K. Andreadou, S. Papadopoulos, D. T. D. Nguyen, G. Boato, M. Riegler, and Y. Kompatsiaris, "Verifying multimedia use at MediaEval 2015," in *Proc. MediaEval*, Sep. 2015, vol. 1436 and 3, no. 3, p. 7.

[48] P. Singh, R. Srivastava, K. P. S. Rana, and V. Kumar, "SEMI-FND: Stacked ensemble based multimodal inferencing framework for faster fake news detection," *Expert Syst. Appl.*, vol. 215, Apr. 2023, Art. no. 119302.

[49] W. Liang, Y. Liang, and J. Jia, "MiAMix: Enhancing image classification through a multi-stage augmented mixied sample data augmentation method," 2023, *arXiv:2308.02804*.

[50] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, pp. 1–74, Mar. 2021.

[51] D. Puspita and B. E. Pranoto, "The attitude of Japanese newspapers in narrating disaster events: Appraisal in critical discourse study," *Stud. English Lang. Educ.*, vol. 8, no. 2, pp. 796–817, May 2021.

[52] A. Szymkowiak, B. Melović, M. Dabić, K. Jeganathan, and G. S. Kundi, "Information technology and gen Z: The role of teachers, the Internet, and technology in the education of young people," *Technol. Soc.*, vol. 65, May 2021, Art. no. 101565.

[53] E. Jing, Y. Liu, Y. Chai, J. Sun, S. Samtani, Y. Jiang, and Y. Qian, "A deep interpretable representation learning method for speech emotion recognition," *Inf. Process. Manage.*, vol. 60, no. 6, Nov. 2023, Art. no. 103501.

YASMINE KHALID ZAMIL received the B.S. and M.Sc. degrees in computer science from the University of Babylon, Iraq, in 2010 and 2019, respectively. She is currently pursuing the Ph.D. degree in computer engineering with Tarbiat Modares University, Iran.

NASROLLAH MOGHADDAM CHARKARI received the B.Sc. degree from Shahid Beheshti University, Tehran, Iran, in 1986, and the M.Sc. and Ph.D. degrees from Yamanashi University, Japan, in 1993 and 1996, respectively. He is currently a Faculty Member of the Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran. His research interests include complex network analysis, computer vision, and knowledge discovery.