IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 December 2022, accepted 19 January 2023, date of publication 27 January 2023, date of current version 1 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3240373

RESEARCH ARTICLE

Multimodal Arabic Rumors Detection

RASHA M. ALBALAWI[1], **AMANI T. JAMAL**[1], **ALAA O. KHADIDOS**[2], **AND AREEJ M. ALHOTHALI**[1]

[1]Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2]Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Corresponding author: Rasha M. Albalawi (rsalemalbalawi@stu.kau.edu.sa)

ABSTRACT Recently, the use of social media platforms has increased with ease of use and fast accessibility, making such platforms a place of rumor proliferation owing to the lack of posting constraints and content authentication. Therefore, there is a need to leverage artificial intelligence techniques to detect rumors on social media platforms to prevent their adverse effects on society and individuals. Most existing works that detect rumors in Arabic target the textual features of the tweet content. Nevertheless, tweets contain different types of content, such as (text, images, videos, and URLs), and the visual features of tweets play an essential role in rumor diffusion. This study proposes an Arabic rumor detection model to detect rumors on Twitter using textual and visual image features through two types of multimodal fusion: early and late fusion. In addition, we leveraged the transfer learning of the pre-trained language and vision models. Different experiments were conducted to select the best textual and visual feature extractors for building a multimodal model. MARBERTv2 was used as a textual feature extractor, whereas the ensemble of VGG-19 and ResNet50 was used as a visual feature extractor to build the multimodal model. Subsequently, the language and vision models of the single models were used as a baseline to compare their results with those of multimodal models. Finally, the experimental results demonstrate the effectiveness of textual features in rumor detection tasks compared to multimodal models.

INDEX TERMS Arabic NLP, artificial intelligence, deep learning, multimodal fusion, rumor detection, transfer learning.

I. INTRODUCTION

In recent years, there has been growing use of social media platforms owing to their ease of use and fast access. It has become a method of rapid communication with the world and a medium for sharing information and news sources. In addition to the positive effects of social media platforms, the lack of posting constraints, content authentication, and ease of use makes these platforms a place for rapid rumor proliferation.

Rumor is defined as ''unverified and instrumentally relevant information statements in circulation that arise in contexts of ambiguity, danger, or potential threat'' [1], [2]. Another definition proposed by [3] defined rumors as ''an item of circulating information whose veracity status is yet to be verified at the time of posting''. Rumors, Fake news, and misinformation are often used interchangeably in the literature; the study's authors [4] distinguished between concepts related to fake news based on three characteristics: authenticities, intention, and whether the information is news. The difference between rumors, fake news, and misinformation is that fake news is false with a negative intention to mislead people. In contrast, misinformation is false information with unknown intentions and is news or non-news. Simultaneously, rumors are not necessarily false information, and their intentions are unknown and may be news or non-news. However, finding a unified definition of rumors in the literature remains a big challenge because there is a degree of uncertainty in using rumor terminology.

In this paper, a rumor is defined as a statement whose authenticity is still unverified at the time of spreading, has an unknown intention, and could be news or non-news based on the definition by [4]. The problem is that the social media structure makes the proliferation of rumors faster and can reach large numbers of people in a short time; this rapid spread of rumors harms society and affects public opinion. In addition, uncontrolled online rumor flooding causes unnecessary panic and changes public perception [5].

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

Moreover, as in the recent covid-19 pandemic, many rumors related to covid-19 symptoms and vaccines have spread on Twitter, which can affect people's decisions and negatively impact their health. One of the most significant current efforts of the Saudi government to defeat the spread of rumors is the Anti-Rumors Authority.[1] It is a fact-checking website established in 2012 to clarify the truth through official sources and debunk rumors on social media. These efforts were made to address the problem of spreading rumors to ensure they do not harm society negatively by clarifying the truth regarding them. Concurrently, tracking rumors circulating on social media from this massive amount of information to clarify their truth is a significant challenge for human effort and time. Therefore, there is a need to use Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques in the context of rumor detection to reduce human effort in tracking rumors and to prevent their negative impacts from spreading.

Detecting Arabic rumors is a big challenge that requires research and investigation due to its rich vocabulary and many dialects. The shortcoming of current Arabic studies is that they focus on detecting rumors from textual features without considering visual features. Most tweets have different content types (text, images, videos, and URLs), and image features are essential for indicating rumors. The study [6] demonstrated that images play an important role in news propagation, and tweets with images attract more attention than tweets with only text. In addition, users tend to believe in the information attached to an image because the image increases the credibility of the tweet from their perspective [7]. In contrast, attached images can be manipulated to attract the audience's attention and cause the proliferation of rumors regardless of their intention.

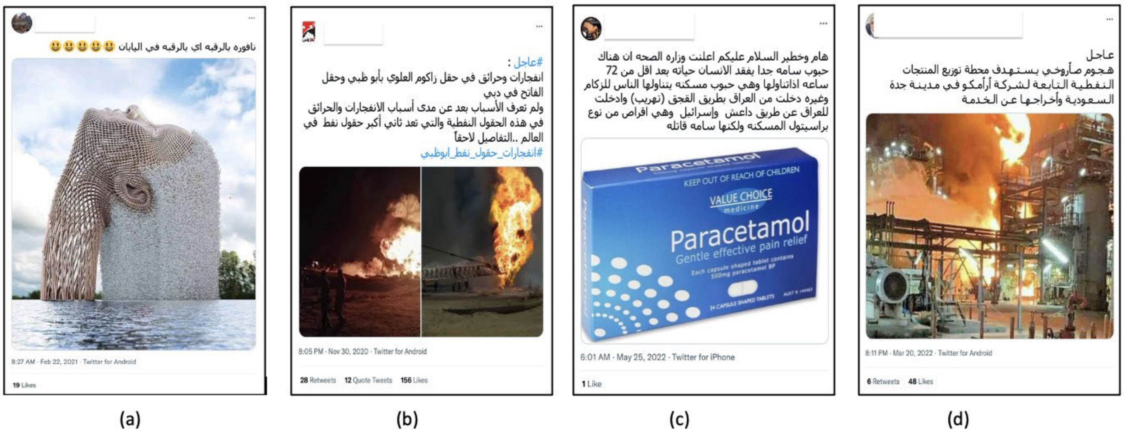
Many studies have been conducted to detect rumors in English by considering the different types of rumor features. A previous study [8] employed text features from content-based features to detect rumors using a convolutional neural network (CNN). Another study incorporated an attention mechanism with a recurrent neural network (RNN) to learn the latent representation of sequential microblog posts and detect rumors at an early stage [9]. Another study extracted text content-based features with user features, such as follower count, retweet count, age of tweets, and friend count [10]. Another study incorporated two types of content-based features: text and image visual, and statistical features [6]. In addition, many studies have proposed multimodal representation learning for rumor detection tasks using neural networks to fuse the text and visual features of tweet posts [11], [12], [13], [14]. The latter methods demonstrate their success in terms of performance compared with unimodal approaches. In contrast, in Arabic content, the works focused only on rumor text without considering the image attached to the text.

A recent study [14] investigated the transfer learning of the pre-trained language model Bidirectional Encoder Representations from Transformers (BERT) and the pre-trained vision-based model VGG-19 to extract textual and visual features, respectively. The transfer learning technique transfers the knowledge of pre-trained models on a specific task to another by sharing the learned low-level features, which helps improve the performance of these models in another task. In addition, it is used as a feature extractor by transferring its knowledge and fine-tuning it to a different but related task, thereby reducing the effort required to build and train new models. In addition, ensemble learning is used to improve the classifier performance by combining two or more classifiers to provide a robust predictive model.

To the best of our knowledge, no study has investigated rumor detection techniques based on tweet content with textual and visual features in Arabic tweets. Therefore, this study investigates the detection of Arabic rumors on Twitter using the transfer learning of pre-trained models to identify whether they are rumors by considering it as a classification problem. The goal was to investigate pre-trained models that classify tweets as rumors or non-rumors based on their visual and textual features.

In this study, we used the Arafacts dataset and extracted the total number of 1726 tweets that met our specifications [15]. In addition, considering the small number of extracted tweets from the previous dataset, we created a multimedia dataset containing rumor tweets and their visual features. The total number of tweets in both datasets was 4025, labeled as rumor and non-rumor. Fig.1 shows a sample of rumor tweets; in Fig.1, (a) a rumor is propagated on Twitter about a head-shaped water fountain in Japan, attached to the image, while the truth is that it is not a real fountain but a digital sculpture picture. Fig.1 (b) shows a rumor propagated on Twitter about pictures of explosions and fires in two oil fields in the United Arab Emirates (UAE). At the same time, the truth is that both images are old, and both are old events related to oilfield explosions outside the UAE. Fig.1 (c) represents a rumor about the Saudi Ministry of Health warning regarding toxic pills smuggled with paracetamol. Simultaneously, the Saudi Ministry of Health denied this rumor. Fig.1 (d) shows a rumor and pictures claiming a fire at the Aramco facility in Jeddah after missiles from the Houthi group were targeted. The truth is that the picture was taken from an old event outside of Saudi Arabia. However, the attached image took advantage of the Houthis bombed by Aramco during the Jeddah event to propagate false images. Fig. 2 shows a sample of the non-rumor tweets in Fig.2, where (a) shows a picture attached to a tweet text about the condition of Ukrainians inside the tunnels. Fig.2 (b) shows the tweet attached to the image of the remnants of the Houthi drones that were declared intercepted by the Saudi Ministry of Defense. Fig.2 (c) shows a rare picture of the Jamarat Mina ritual 142 years ago. Fig.2, (d) shows a tweet and pictures of Elon Musk's announcement of the Neuralink chip.

[1] http://norumors.net



(a)
نافوره بالرقبه اي بالرقبه في اليابان 😃😃😃😃😃
8:27 AM · Feb 22, 2021 · Twitter for Android
19 Likes

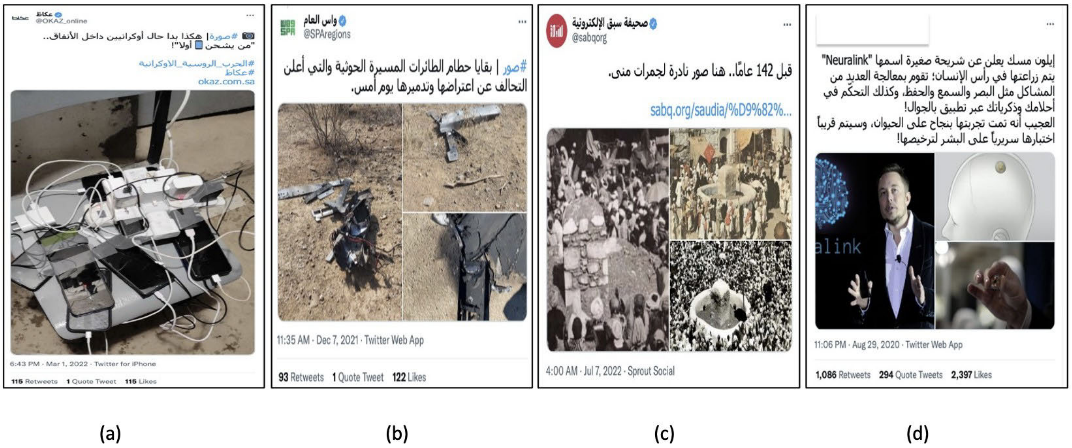
(b)
#عاجل :
انفجارات وحرائق في حقل زاكوم العلوي بأبو ظبي وحقل الفاتح في دبي
ولم تعرف الأسباب بعد عن مدى أسباب الانفجارات والحرائق في هذه الحقول النفطية والتي تعد ثاني أكبر حقول نفط في العالم ..التفاصيل لاحقاً
#انفجارات_حقول_نفط_ابوظبي
8:05 PM · Nov 30, 2020 · Twitter for Android
28 Retweets 12 Quote Tweets 156 Likes

(c)
هام وخطير السلام عليكم اعلنت وزاره الصحه ان هناك حبوب سامه جدا يفقد الانسان حياته بعد اقل من 72 ساعه اذاتناولها وهي حبوب مسكنه يتناولها الناس للزكام وغيره دخلت من العراق بطريق القجق (تهريب) وادخلت للعراق عن طريق داعش وإسرائيل وهي اقراص من نوع براسيتول المسكنه ولكنها سامه قاتله
6:01 AM · May 25, 2022 · Twitter for iPhone
1 Like

(d)
عاجل
هجوم صأروخي يستهدف محطة توزيع المنتجات النفطية التابعة لشركة أرأمكو في مدينة جدة السعودية وأخراجها عن الخدمة
8:11 PM · Mar 20, 2022 · Twitter for Android
6 Retweets 48 Likes

Paracetamol
KEEP OUT OF REACH OF CHILDREN
VALUE CHOICE
medicine
Paracetamol
Gentle effective pain relief
Each capsule shaped tablet contains 500mg paracetamol BP
24 CAPSULE SHAPED TABLETS

FIGURE 1. Rumors tweets example.



(a)
عكاظ
@OKAZ_online
📷 #صورة| هكذا بدا حال أوكرانيين داخل الأنفاق..
"من يشحن 📱 أولا"!
#الحرب_الروسية_الاوكرانية
#عكاظ
okaz.com.sa
6:43 PM · Mar 1, 2022 · Twitter for iPhone
115 Retweets 1 Quote Tweet 115 Likes

(b)
واس العام
@SPAregions
#صور | بقايا حطام الطائرات المسيرة الحوثية والتي أعلن التحالف عن اعتراضها وتدميرها يوم أمس.
11:35 AM · Dec 7, 2021 · Twitter Web App
93 Retweets 1 Quote Tweet 122 Likes

(c)
صحيفة سبق الإلكترونية
@sabqorg
قبل 142 عامًا.. هنا صور نادرة لجمرات منى.
sabq.org/saudia/%D9%82%...
4:00 AM · Jul 7, 2022 · Sprout Social

(d)
إيلون مسك يعلن عن شريحة صغيرة اسمها "Neuralink" يتم زراعتها في رأس الإنسان؛ تقوم بمعالجة العديد من المشاكل مثل البصر والسمع والحفظ، وكذلك التحكم في أحلامك وذكرياتك عبر تطبيق بالجوال!
العجيب أنه تمت تجربتها بنجاح على الحيوان، وسيتم قريباً اختبارها سريرياً على البشر لترخيصها!
11:06 PM · Aug 29, 2020 · Twitter Web App
1,086 Retweets 294 Quote Tweets 2,397 Likes

FIGURE 2. Not rumors tweets example.

The main contributions of this paper are:

1) To the best of our knowledge, no research has investigated how to detect rumors based on textual and visual features in detection models used for the Arabic language.
2) Constructing a multimedia dataset with both textual and visual features containing 4025 tweets labeled as rumors and non-rumors.
3) Investigate different pre-trained models to extract textual and visual features of tweet content to build a multimodal model for Arabic rumor detection.
4) Investigating two types of Multimodal features fusion; early and late fusion.
5) Evaluating the multimodal detection model on the proposed dataset.

II. RELATED WORK

In recent years, many studies have been conducted to detect rumors [16], fake news [17], and misinformation [18] on social networks. This field has attracted the research community to develop models that can effectively detect rumors. Rumors can diffuse through social media using different modalities such as text, images, or videos. Simultaneously, it can be spread by users whose social context indicators help detect rumors, such as checking the credibility of those users through their account characteristics. The rumor features used by the detection systems can be divided into content and context-based features [19]. Content-based features represent features extracted from text or visual content such as images or videos [19]. In addition, context-based features represent social interactions between users and others through following, retweeting, liking, commenting, and tagging, and analyzing these features can be important indicators for detecting rumors [19].

In addition to user features, propagation features are extracted through a network of rumor diffusions, such as statistics of the tree structure of message propagation and temporal features that extract the rumor diffusion period [19]. Studies on rumor detection in Arabic content have focused only on detecting rumors using a unimodal method that detects rumors from textual features. In contrast, many studies on English-language rumor detection systems have considered extracting features from different modalities, such as using the image, text, and social features as input to the classification model. This section covers the work done in Arabic using unimodal approaches and in English using multimodal approaches.

A. RUMOR DETECTION USING UNIMODAL APPROACHES

This section summarizes rumor detection approaches in the Arabic language, which can be categorized into studies that use machine learning and deep learning approaches. Machine learning approaches involve supervised or unsupervised learning. Supervised learning algorithms require extracting useful features to train the classifier to differentiate between the rumor and non-rumor classes.

Study [20] proposed a semi-supervised learning algorithm for detecting rumors in Arabic tweets using a machine learning-based technique. The features were extracted from the user and content-based features. The model was then trained using supervised gaussian naive bayes (GNB) and semi-supervised expectation-maximization models. The results show that the semi-supervised expectation maximization model outperforms the GNB with an f1-score of 78.6%.

Covid-19 pandemic-related fake news on Twitter has been reported in previous studies [21], [22], [23], [24]. Most of these studies used popular machine learning algorithms such as support vector machine (SVM), naive bayes (NB), logistic regression (LR), extreme gradient boosting (XGBoost), and random forest (RF).

On the other hand, another study [25] detected cancer treatment-related rumors on social media using tweet text.

Another corpus was collected from YouTube comments on rumors of the death of famous Arab celebrities and used three machine learning algorithms: SVM, multinomial NB (MNB), and decision tree (DT) to classify rumor and non-rumor comments [26].

Another study has examined extracting linguistics features from the text: emotions, linguistics, polarity, and part of speech [27]. Three classifiers, NB, RF, and SVM, were used to train the classifier to detect Arabic fake news using the extracted features. The real news articles focused on pilgrimage news during a specific period, whereas fake news articles were collected through crowdsourcing. The results showed that the extracted textual features were dominant for fake news detection, and the best classifier was RF, with 79% accuracy.

The authors in [28] extracted two types of features from the text, content-based and topic-based, in addition to extracting user-based features. The XGBoost algorithm was used, and the results indicated that the proposed model achieved an accuracy of 97%.

A comparative study was conducted to detect covid-19 rumors using the textual features of tweets from different machine learning and deep learning methods [29]. This study compared the performance of different machine learning methods, SVM, stochastic gradient descent (SGD), LR, k-nearest neighbors (KNN), NB, RF, XGBoost, and DT, using various feature representations and examined the use of ensemble learning. The deep learning methods used were RNN, GRU, LSTM, bidirectional RNN (Bi-RNN), bidirectional GRU (Bi-GRU), and Bi-LSTM and examined the use of seven optimizers. The study concluded that ensemble learning enhances machine-learning algorithms for predicting rumors. At the same time, the best performance was achieved by LSTM and Bi-LSTM with the RMSprop optimizer among all other deep learning algorithms.

Extracting handcrafted features is difficult and time-consuming; however, deep learning techniques overcome these limitations, proving their ability to learn feature representations better than traditional machine learning methods. Experiments were conducted to detect covid-19 misinformation using different deep learning algorithms, namely, CNN, RNN, Bi-LSTM, convolutional recurrent neural network (CRNN) [22], and hybrid deep learning algorithm long short-term memory-parallel convolutional neural network (LSTM-PCNN) [30]. Another study examined the detection of general rumors in Arabic tweets using the CNN-LSTM approach [31]. In [32], the authors presented a comparative study using neural networks and transformers for Arabic fake news detection. The study concluded that transformer-based models outperformed neural-based ones. While the study [33] detected fake news using eight BERT transformer-based models, two were multilingual, and the remaining were BERT models for the Arabic language.

Table 1 summarizes studies that have detected rumors in Arabic content. From Table 1, we indicate the research gap: no paper has detected rumors from visual and textual features in Arabic content.

B. RUMOR DETECTION USING MULTIMODAL APPROACHES

Multimodal learning aims to associate the different features from multiple modalities. This learning process allows the model to capture important information regarding phenomena [34]. In addition, multimodal learning uses the ability of a neural network to learn the representation of feature data and fuses different modalities [34]. Data fusion combines information from multiple models to predict the output of a regression or classification model [34].

Multimodal fusion has proven successful in various applications, such as visual question answering [35], image captioning [36], and multimedia event detection [37]. Multimodal fusion offers three benefits. First, it can provide a robust prediction because it extracts features of the same phenomenon from different modalities. Second, different entities complement each other because a single mode cannot provide sufficient information. Third, a multimodal model can operate in a single mode without using other modalities [34].

The rumor features were extracted from different modalities using image, text, and social features in [11]. They proposed a multimodal model based on an RNN with an attention mechanism (att-RNN). Using LSTM, they fused textual and social features, whereas visual features were extracted using a CNN and fused with a joint representation of textual and social features. They experimented with an att-RNN model using two multimedia datasets. Their results showed that the multimodal model was better at detecting rumors than the unimodal model.

Text-CNN was used to extract textual features, whereas the VGG-19 model was used to extract visual features of the rumor posts [12]. The proposed multimodal model uses textual and visual features through a self-attention fusion. To predict upcoming rumors, they implemented latent topic memory to store the rumors' semantic information. The proposed model was trained and tested on two multimedia datasets, and the results showed that the proposed multimodal fusion network was more effective than unimodal models.

In [13], the study addressed the problem of previous works [11] and [12] in that the models cannot be generalized to identify rumors on a new event because it is dependent on a specific event of the dataset. Accordingly, they proposed event features to detect fake news on social media, allowing a detection system to capture new rumors. The proposed Event Adversarial Neural Network (EANN) consisted of three components. The first multimodal feature extractor is responsible for extracting the textual features of text using the text-CNN model. In contrast, the visual features of the attached images were extracted using a pre-trained VGG-19. The second is a fake news detector, used to determine whether a specific post is fake or not. The third is the event discriminator, which removes event-specific features to identify transferable ones. Finally, the EANN model was trained and tested on two multimedia datasets, and the results showed that it outperformed the aforementioned models.

The authors of [14] proposed SpotFake, a multimodal model for fake news detection. SpotFake consists of two models: the pre-trained BERT model to extract textual features and the VGG-19 model to extract visual features. The two feature vectors are then concatenated by fusing them to obtain a news representation. The model is trained and tested using two multimedia datasets. The proposed multimodal SpotFake model outperformed the unimodal models and other models proposed in [11] and [13].

In contrast to the earlier results of fusing textual and visual features for rumor detection tasks, the author in [52] had different results and argued that [14] used accuracy as the primary metric and ignored the imbalanced nature of the dataset. Although [11] did not address an imbalanced dataset, they reported an f1-score for each class. Furthermore, they explained that concatenating different features resulted in noisy representations and that the model could not be generalized to new feature combinations. Thus, the multimodal model cannot outperform the unimodal models of text and images.

III. DATASET DESCRIPTION

In this study, we used the Arafacts dataset, the first Arabic dataset that uses the fact-checking website as the source for extracting claims [15]. Arafacts is the first Arabic dataset that contains tweets with multimedia content for rumor detection tasks. The number of claims in the dataset is 6,222, of which 4141 are video or image claims. The dataset uses four classes to classify claims (false, partly false, sarcasm, and true). Sarcasm claims were excluded from the dataset, and the remaining claims were retained. Tweets with multimedia content as an image were extracted without considering video claims, and tweets with images without text were excluded. The number of extracted tweets from this dataset was 1726 since many URL links of tweets were deleted by users or could not be found because of account suspensions or deletions. The classes of the extracted tweets were unbalanced; Table 2 lists the dataset statistics. Given the lower number of extracted tweets and unbalanced classes, we collected our dataset using a Python scraper that scraped targeted tweets from Twitter.

A. DATA COLLECTION

1) NON-RUMORS DATA COLLECTION

Non-rumor data were collected from Twitter accounts of trusted Saudi government news agencies, the Saudi Press Agency (SPA),[2] Okaz,[3] and Sabq.[4] The collected tweets belonged to general domains, including politics, economics, culture, and health. Additionally, two fact-checking websites were used. First, the Fatabyyano[5] website was established in 2016 to debunk fake news and rumors by collaborating with Facebook as a third-party fact-checker. The second is Misbar,[6] an Arabic fact-checking website that debunks rumors and false news in online media.

2) RUMORS DATA COLLECTION

The rumor data were collected from two fact-checking websites: The no-Rumor website, Anti-Rumors Authority, established by the Saudi government in 2012 to debunk rumors, and the second is Misbar.

B. DATA ANNOTATION

The role of visual content in detecting fake news has been explored in a previous study [38]. Visual content can be classified into three categories: manipulated media, irrelevant media, such as past events reposted with a new event, or a false claim made about a real visual (unmanipulated) but published along with it. All three types fall under the definition of fake news, irrespective of the truthfulness of the textual or visual content because text and images together provide false information [38]. Therefore, the class in the

[2] https://twitter.com/SPAregions
[3] https://twitter.com/okazonline
[4] https://twitter.com/sabqorg
[5] https://fatabyyano.net/fatabyyano-team/
[6] https://misbar.com/

TABLE 1. Summary of different studies to detect rumors in Arabic content (Features of rumor: U: User, T: Text, V: Visual).

Ref	Task	Classifier	U	T	V	Accuracy
[20]	Detecting rumors in general	ML methods: GNB and semi-supervised expectation-maximization	✓	✓		78.6%
[21]	Fake news related to the Covid-19 pandemic	ML methods: NB, LR, SVM, MP, RFB, XGBoost		✓		93.3%
[22]	Covid-19 misinformation	ML methods: SVM, NB, XGBoost, RF, SGD. DL methods: CNN, RNN Bi-LSTM, and CRNN		✓		86.8% 85%
[23]	Rumors related to covid-19 and the source of rumors	ML methods: LR, SVM, NB		✓		86%
[24]	Covid-19 Arabic Tweets	ML methods: SVM Transformer-based methods: AraBERT		✓		85% 83.9%
[25]	Health-related rumors	ML methods: LR, SVM, BNB, SGD, KNN, DT, RF, Ada, Bag		✓		85%
[26]	Rumors related to the death of famous Arab celebrities	ML methods: SVM, NB, DT		✓		95%
[27]	Arabic fake news related to pilgrimage	ML methods: NB, RF, SVM		✓		79%
[28]	Detecting rumors in general	ML methods: XGBoost	✓	✓		97%
[29]	Detecting covid-19 rumors	ML methods: LR, SGD, NB, SVM, KNN, DT, RF, XG- Boost DL methods: RNN, bidirectional RNN (Bi-RNN), GRU, bidirectional GRU (Bi-GRU), LSTM, and Bi-LSTM		✓		80%
[30]	Covid-19 Arabic rumors	Hybrid Deep Learning: (LSTM–PCNN)		✓		86%
[31]	Detecting rumors in general	ML methods: KNN, XGBoost, GB DL methods: LSTM, Bi-LSTM, CNN-LSTM		✓		95.9%
[32]	Arabic fake news	DL methods: (GRU, CNN, RNN) Transformer-based methods: AraBERT v1, AraBERT v02, AraBERT v2, ArElectra, QARiB, Arbert, Marbert		✓		83% 97%
[33]	Arabic fake news	Transformer-based methods: GigaBert-base, RobertaBase, Arabert, Arabic-Bert, ArBert, MARBert, Araelectra, QaribBert—base		✓		98.8%

Arafacts dataset (partly false) changed to false. Additionally, when collecting rumor data from Misbar, the data category is either false or misleading. It was false when the images were manipulated. It is misleading when the image is associated with an irrelevant event or if it is real but published with a false tweet. Both the false and misleading categories were treated as false claims (rumors). In contrast, the collected non-rumor data were labeled true (non-rumor).

IV. METHODOLOGY

In this study, we propose a multimodal model that uses a transfer learning technique to detect whether a tweet is a rumor or not. In the first step, the model receives pairs of inputs: the tweet text and the associated image. The proposed model is divided into three sub-models. The first was a pre-trained BERT model used to extract contextual features from the text. The second is an ensemble of pre-trained vision models, VGG-19 [39] and ResNet50 [40], to extract visual features from the image. The third is a multimodal model that concatenates the extracted features of the text and image to represent the rumor vector, feeds it to the classifier, and provides the classification result. Fig. 3 shows the proposed model.

A. DATA PREPROCESSING

In the first step, the model receives pairs of inputs: the tweet text and the associated image. Before feeding it as input to the model, we preprocessed and cleaned the tweets by removing punctuation marks, Arabic diacritics, non-Arabic words, Emojis, Arabic and English numbers, URLs, mentions symbol @, user's mention, and multiple white spaces. Finally, we normalized the hashtags by removing # symbols and underscores. In addition, we removed hashtags and keywords that represented news agency names to ensure that the model was not biased toward correctly identifying non-rumor tweets. For the same reason, we preprocessed the images in the non-rumor tweets that have news agency logos by cropping these logos. Table 3 lists the removed hashtag keywords, and Fig. 4 shows a sample of the images before and after pre-processing.

TABLE 2. Arafacts dataset statistics.

Label	True	False	Partly false	Total
Number of tweets	55	628	1043	1726

B. MODEL ARCHITECTURE

The proposed model consists of three sub-models: the pre-trained language model, the pre-trained vision model, and the multimodal model used to fuse the two representations of rumor, textual and visual features. This section explains the role of each model.

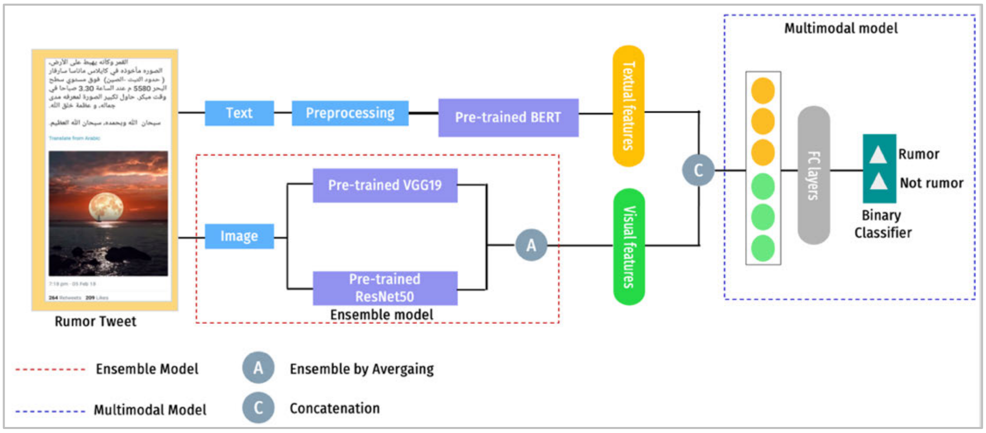
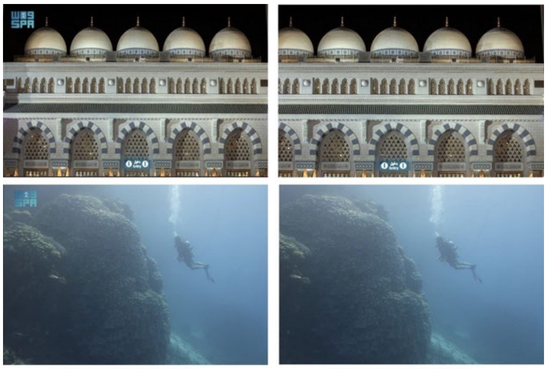


FIGURE 3. The proposed model.


Images before cropping logo's

Images After cropping logo's

FIGURE 4. Example of images before and after preprocessing.

TABLE 3. Hashtags keywords that were removed from the dataset.

Arabic Hashtags keywords	English Hashtags keywords
عكاظ	Okaz
أن تكون أولا	to be the first
تطبيق عكاظ	Okaz application
واس	SPA
صور واس	SPA photos
تقارير واس	SPA reports
واس عام	SPA general
سبق	Sabaq

The left column lists the Arabic hashtags, and the right column lists the translations into English.

1) LANGUAGE MODEL (TEXTUAL FEATURES EXTRACTOR)

BERT is a language model for NLP tasks based on the transformer architecture developed by Google [41]. BERT is unlike sequence-to-sequence models, which can read text sequentially, either from left to right or right to left, making it fail to understand contextual information. BERT is bidirectional and can capture contextualized information by learning the context of a word based on all the surrounding words in both directions. In addition, using an attention mechanism, BERT overcomes the inability of sequence models to capture long-context dependencies between text inputs. The large scale of the application of NLP fields proves the BERT transformers' success versus directional language models. In addition, the Text-to-Text Transfer Transformer (T5) model is a language model that is used to generate natural language [47]. In our experiments, we used different models of the Arabic checkpoints of BERT and T5 to extract textual information from rumor text. The versions used can be summarized as follows:

1) AraBERT [42]: it is a pre-trained language model for the Arabic language, and it was evaluated using three tasks: named entity recognition, sentiment analysis, and question answering. This study used different versions of AraBERT, such as AraBERTv2, AraBERTv01, and AraBERTv02, which were pre-trained on Modern Standard Arabic (MSA). AraBERTv0.2-Twitter is pre-trained on Dialects Arabic (DA) and tweets, which overcomes the limitations of the previous version.
2) ARBERT and MARBERT [43]: ARBERT is a pre-trained language model for MSA that uses the same architecture as the BERT base. In contrast, MARBERT was pre-trained on both MSA and DA. MARBERTv2 was trained on the same MSA dataset used by ARBERT, in addition to the Arabic News dataset with a larger sequence length. These models were evaluated using NLP tasks, such as social meaning, sentiment analysis, dialect identification, named entity recognition, and topic classification. The versions used in this study were ARBERT, MARBERT, and MARBERTv2.
3) QARiB [44]: This is a pre-trained transformer-based model of MSA and DA. It was trained on a large number of tweets collected using the Twitter API and a large

number of text sentences. The model was evaluated using NLP tasks such as emotion detection, named entity recognition, offensive language detection, Arabic dialect identification, and sentiment analysis. In this study, we used the QARiB and QARiB far.

4) Arabic Bert [45]: it is a pre-trained transformer-based model on the Arabic version of the OSCAR corpus and other Arabic texts. It was trained on MSA and DA and used downstream hate speech detection tasks.
5) arabert Covid-19 and mbert Covid-19 [46]: It is a fine-tuned version of AraBERTv2 and mBERT on 1.5 million of covid-19 fake news tweets that have multidialectal Arabic and were trained on fake news detection tasks.
6) Ara-DialectBERT[7]: it is a fine-tuned Camelbert MSA eighth model trained in Arabic hotel reviews from the Booking website and was trained in both MSA and DA languages.
7) AraT5 [47]: it is an Arabic text-to-text transformer model that uses a T5Base encoder-decoder architecture [48]. There are a variety of models for AraT5: AraT5 MSA, which was trained on MSA data; AraT5 Tweet, which was trained on Twitter data; and AraT5, which was trained on both MSA and Twitter data. The models were evaluated using the Arabic natural-language-generation ARGEN benchmark. ARGEN has seven tasks: code-switched text, translation, text summarization, machine translation, transliteration, question generation, news title generation, and paraphrasing.

We fine-tuned the pre-trained checkpoints of the abovementioned models by adding a simple classification layer that conducts binary classification to classify tweets into rumors or non-rumors.

2) VISION MODEL (VISUAL FEATURES EXTRACTOR)

The human brain tends to believe the rumor with visual content rather than text. In addition, visual content plays an important role in fast rumor proliferation. ResNet50 [40], InceptionV3 [49] and VGG-19 [39], and such models are trained from scratch on huge, annotated image datasets, ImageNet, and using high GPU capabilities. These models learn shallow feature representations of an image, such as shapes, edges, and blobs.

In this paper, we leverage transfer learning techniques with pre-trained models VGG-19 and ResNet50 to extract the visual features and leveraged ensemble learning techniques to reduce the prediction error variance and ensure the robustness of our model.

3) MULTIMODAL MODEL

Multimodal deep learning combines information from different modalities, such as text, images, videos, and audio [53]. Multimodal deep learning has been inspired by how humans integrate visual and audio information to understand speech [53]. In this paper, the multimodal model fuses two different modalities by concatenating the textual and visual vector representations obtained from the above submodels to learn the rumor representation. Another aspect of multimodal learning is its flexibility to provide different fusion structures, which are used to integrate various modality features, namely, early fusion (feature-based), intermediate fusion (hybrid fusion), and late fusion (decision-based) [50]. The difference between the types of fusion is as follows. Early fusion aims to concatenate each modality's representation after extracting the features and before being input into the multimodal classifier as a single vector of features. In comparison, late fusion fuses the features after classifiers make the decision of different modalities [50].

TABLE 4. The proposed text-based model.

Language Model - Textual Features Extractor
Inputs (input ids, input mask)
Freeze embedding layer of BERT
Batch Normalization layer
Dense (768 units, ReLU)
Batch Normalization layer
Dense (128 units, ReLU)
Output Dense (1 unit, sigmoid)

In comparison, intermediate fusion benefits from both early and late fusion [50]. This study used two types of fusion: early and late, as shown in Fig. 5 and 6.

V. EXPERIMENT

This section describes the experimental setup, implementation details, dataset statistics, hyperparameter tuning, and the evaluation metrics.

A. EXPERIMENT SETUP AND IMPLEMENTATION DETAILS

The experiment was implemented using the Google Colab environment, and the model was trained using the Keras library with TensorFlow as the backend.

1) LANGUAGE MODEL (TEXTUAL FEATURES EXTRACTOR)

In this stage, we applied the same settings to all pre-trained models. First, the text in the tweet is cleaned and preprocessed. It was then tokenized and padded to a fixed length before being fed to the BERT model. Subsequently, all the models were fine-tuned by freezing the embedding layer, and a classification layer was added. For the classification layer, we added a batch normalization layer, followed by a ReLU layer, and this step was repeated twice. The former had 768 neurons, whereas the latter had 128. A batch normalization layer is added to the classification layer to accelerate the computation time and increase the learning speed. Finally, the last layer with a sigmoid activation function was added. The proposed text-based model architecture is presented in Table 4.

[7] https://huggingface.co/MutazYoune/AraDialectBERT

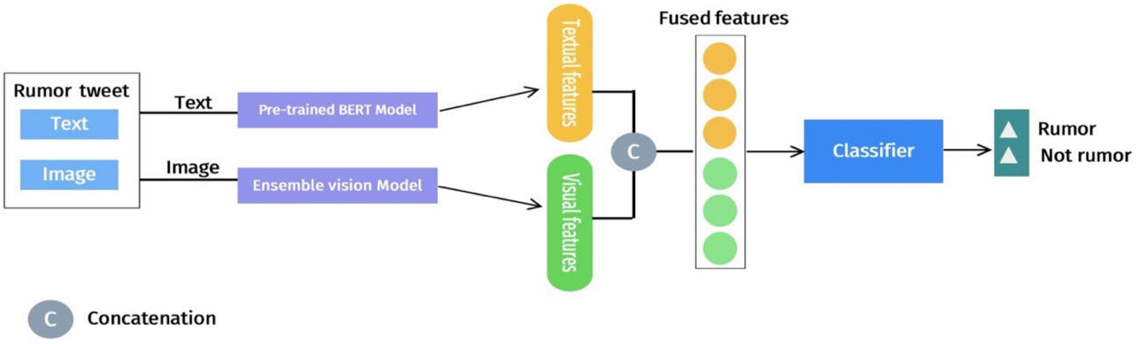


FIGURE 5. Early fusion.

TABLE 5. The proposed vision-based model.

Vision Model - Visual Features Extractor
Input image (224,224,3)
VGG-19 InceptionV3 ResNet50
Global average pooling 2D
Flatten layer
Dense (2048, ReLU)
Dropout - probability 30%
Dense (128, ReLU)
Output Dense (1 unit, sigmoid)

TABLE 6. The proposed ensemble models.

Ensemble Model
Input image (224,224,3)
The proposed vision-based model, except for the sigmoid layer
Average () [model1 layer: Dense (128, ReLU), model2 layer: Dense (128, ReLU)]
Output Dense (1 unit, sigmoid)

2) VISION MODEL (VISUAL FEATURES EXTRACTOR)

In this stage, before passing the images to the pre-trained models, they were prepossessed by cropping images that contained the agencies' logos and were resized to a fixed size (224,224,3). Various experiments have been conducted to select the best vision model. In the first experiment, the VGG-19, ResNet50, and InceptionV3 models were used.

In the second experiment, we leverage ensemble learning of the two fine-tuned models, ResNet50 and VGG-19, to increase the robustness of the predictions of our model by taking the average of the predicted probabilities. VGG-19, ResNet50, and InceptionV3 are fine-tuned by freezing the base layer of the model. A classification layer is then added to the frozen model. A global average pooling layer, followed by a flattened layer and then two fully connected layers, were added: one layer with 2048 neurons and one layer with 128 neurons. A dropout layer with a ratio of 0.3 was added. Finally, the last layer with the sigmoid activation function was added. The proposed architecture of the vision-based model is presented in Table 5. Subsequently, the two best-performing models were used to ensemble their results; the proposed ensemble model is shown in Table 6.

3) MULTIMODAL MODEL

At this stage, two types of fusion were tested: early and late fusion. For early fusion, the last layer of the chosen unimodal model is concatenated. Each was a 128-dimensional vector, and when concatenated, became a 265-dimensional vector passed through a fully connected layer with 128 neurons, followed by a dropout layer with a (0.2) ratio, and then passed to a fully connected layer with 64 neurons, followed by a dropout layer with a (0.2) ratio. Finally, the last layer with a sigmoid activation function is added. The proposed architecture of the early fusion multimodal model is presented in Table 7. In late fusion, each model's classifier decision was taken before concatenating the two vectors. Subsequently, the output of each model was concatenated. A sigmoid layer is then added to the final layer of the model. The proposed architecture of the late fusion multimodal model is shown in Table 8.

B. DATASET STATISTICS

The total number of tweets in the dataset is 4025. The number of tweets extracted from the Arafacts dataset was 1726 tweets. As mentioned in Section III, the remaining 2299 samples were collected. The dataset was split into a training set of 80% and a testing set of 20%. To tune the model's hyperparameters, the dataset was divided into a training set of 70% and 10% for validation. After validating the model performance on the validation set and selecting the best hyperparameters, the training and validation sets were combined to train the entire model. Table 9 presents the statistics of the dataset, and Fig. 7 shows the overview of the dataset.

C. HYPER-PARAMETERS TUNING

In all experiments, the validation set was used to find the optimal hyperparameters for the training set and to help make the model generalize and not overfit, such as the batch size, optimizer, learning rate, number of hidden layers,

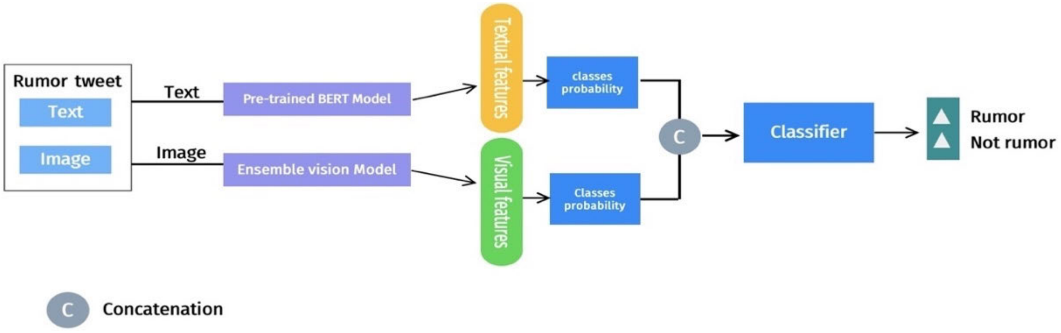


FIGURE 6. Late fusion.

TABLE 7. The proposed multimodal model - early fusion.

Multimodal model architecture - Early Fusion	
Input Text (input ids, input mask)	**Input** Image image (224,224,3)
BERT model except for sigmoid layer **Output**: Dense (1, sigmoid)	Ensemble model except for the sigmoid layer **Output**: Dense (1, sigmoid)
Concatenate (BERT model output, Ensemble model output)	
Dense (128, ReLU) Dropout - probability 20% Dense (64, ReLU) Dropout - probability 20% **Output** Dense (1 unit, sigmoid)	

TABLE 8. The proposed multimodal model - late fusion.

Multimodal model architecture - Late Fusion	
Input Text (input ids, input mask)	**Input** Image image (224,224,3)
BERT model **Output**: Dense (1, sigmoid)	Ensemble model **Output**: Dense (1, sigmoid)
Concatenate (BERT model output, Ensemble model output)	
Output Dense (1 unit, sigmoid)	

TABLE 9. Dataset statistics.

Label	Not Rumor	Rumor	Total
Number of tweets	1793	2232	4025
Training set	1297	1603	2900
Validation set	152	169	321
Testing set	344	460	804

activation function, and number of neurons. For the language model, different learning rates (2e-5, 3e-5, 5e-5) were examined. Various architectures of the vision models were examined to determine the best hyperparameters. For the multimodal model, we examined different optimizers, RMSprop and Adam, with different learning rates in the range (1e-3, ..., 5e-6). Finally, binary cross-entropy was used as the loss function for all models, and early stopping was used to enable all models to generalize and avoid overfitting. Table 10 presents the hyperparameter settings for each model.

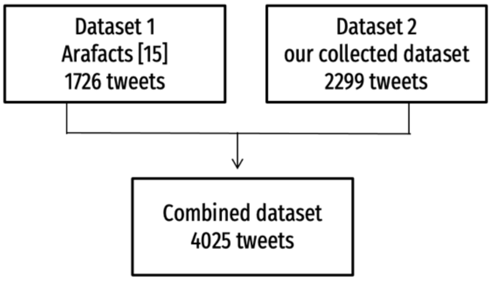


FIGURE 7. Dataset overview.

D. EVALUATION METRICS

The overall performance of our multimodal classification model was measured using the classification metrics of accuracy, F1 score, recall, and precision, which were calculated as follows:

- **Accuracy**: it is described as the ratio of the number of correctly predicted tweets to the total number of

TABLE 10. Hyper-parameters setting.

Model	Batch size	# Of epochs	Optimizer	Learning rate
All language models	64	50	Adam	3e-5
VGG-19 ResNet50 InceptionV3	64	100	Adam	2e-5
Ensemble model	64	100	Adam	3e-5
Multimodal model Early Fusion	64	100	Adam	5e-6
Multimodal model Late Fusion	64	100	Adam	3e-6

TABLE 11. Language models result.

Model	Accuracy	Precision	Recall	F1-score
AraT5 MSA	0.7761	0.8160	0.7460	0.7512
AraT5	0.7910	0.8070	0.8074	0.7910
AraT5 Tweet	0.8097	0.820	0.7919	0.7983
QARiB far	0.8557	0.8557	0.8632	0.8549
mbert Covid-19	0.8595	0.8559	0.8610	0.8576
AraBERTv01	0.8595	0.8596	0.8672	0.8587
AraBERTv2	0.8744	0.8713	0.8726	0.8719
Arabic-Bert	0.8729	0.8712	0.8744	0.8720
MARBERT	0.8868	0.8949	0.8757	0.8819
Ara-DialectBERT	0.8843	0.8813	0.8879	0.8830
ARBERT	0.8856	0.8822	0.8857	0.8837
arabert Covid-19	0.8937	0.8884	0.8918	0.8903
QARiB	0.8955	0.8929	0.8940	0.8934
AraBERTv02	0.8955	0.8924	0.8955	0.8937
AraBERTv0.2-Twitter	0.8968	0.8937	0.8965	0.8949
MARBERTv2	0.8980	0.8947	0.8987	**0.8964**

TABLE 12. Language models result for rumor and not rumor class.

Model	Rumor Class			Non-Rumor class		
	Precision	Recall	F1-score	Precision	Recall	F1-score
AraT5 MSA	0.73	0.95	0.83	0.90	0.54	0.67
AraT5	0.92	0.69	0.79	0.69	0.92	0.79
AraT5-Tweet	0.79	0.92	0.85	0.86	0.67	0.75
QARiB far	0.93	0.81	0.87	0.78	0.92	0.84
mbert Covid-19	0.90	0.85	0.87	0.81	0.87	0.84
AraBERTv01	0.93	0.81	0.87	0.79	0.92	0.85
AraBERTv2	0.89	0.88	0.89	0.85	0.86	0.85
Arabic-Bert	0.90	0.88	0.89	0.85	0.86	0.85
MARBERT	0.86	0.95	0.91	0.93	0.80	0.86
Ara-DialectBERT	0.93	0.86	0.90	0.83	0.91	0.87
Arbert	0.91	0.88	0.90	0.85	0.89	0.87
arabert Covid-19	0.93	0.88	0.90	0.85	0.91	0.88
QARiB	0.91	0.90	0.91	0.87	0.88	0.88
AraBERTv02	0.92	0.90	0.91	0.87	0.90	0.88
AraBERTv0.2-Twitter	0.92	0.90	0.91	0.87	0.90	0.88
MARBERTv2	0.93	0.89	0.91	0.86	0.90	0.88

TABLE 13. Vision models results.

Model	Accuracy	Precision	Recall	F1-score
VGG-19	0.7090	0.7247	0.7243	0.7089
ResNet50	0.7786	0.7935	0.7944	0.7786
InceptionV3	0.5759	0.5542	0.5131	0.4277
Ensemble of VGG-19 and ResNet50	0.7910	0.7917	0.7979	**0.7900**

TABLE 14. Vision models result for rumor and not rumor class.

Model	Rumor Class			Non-Rumor Class		
	Precision	Recall	F1-score	Precision	Recall	F1-score
VGG-19	0.83	0.62	0.71	0.60	0.82	0.69
ResNet50	0.91	0.68	0.78	0.68	0.90	0.78
InceptionV3	0.58	0.95	0.72	0.53	0.08	0.14
Ensemble of VGG-19 and ResNet50	0.87	0.75	0.80	0.72	0.85	0.78

predicted tweets.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Recall:** This was described as the proportion of correctly predicted positive tweets to the total number of positive tweets.

$$Recall = \frac{TP}{TP + FN}$$

- **Precision:** This was described as the proportion of correctly predicted positive tweets out of the total number of predicted positive tweets, either correctly True-Positive (TP) or incorrectly False-positive (FP) tweets.

$$Precision = \frac{TP}{TP + FP}$$

- **F1 score** is computed as the average rate between the recall and precision.

$$F1\ score = \frac{2(Precision \times Recall)}{Precision + Recall}$$

VI. RESULTS

In this section, the performance of the proposed models was reported. The unimodal models of the best language model and ensemble of vision models were used as a baseline to compare their results with those of the multimodal models in terms of the F1 score.

The first experiment was conducted to choose the best pre-trained model for extracting textual features. Sixteen models were used in this study. Table 11 shows the results of training different transformer-based models on the proposed dataset; the MARBERTv2 model obtained the highest F1 score. At the same time, Table 12 shows the result of these models for each class.

The second experiment was conducted to choose the best pre-trained vision model to extract visual features. Then we used ensemble learning of VGG-19 and ResNet50 to increase the robustness of the vision models. In addition, InceptionV3 was excluded because of its poor results and to ensure it did not affect model results. Table 13 presents the results of training the three vision models, InceptionV3, VGG-19, and ResNet50, and the ensemble of VGG-19 and ResNet50 on the proposed dataset. Table 14 presents the models results for each class.

The third experiment was conducted to build a multimodal model from the best language model and the ensemble of vision models. MARBERTv2 is a textual feature extractor, and an ensemble of vision models is used as the visual feature extractor. The results are presented in Table 15 for the two multimodal models, early and late fusion, and Table 16 shows the results of the multimodal model for each class. MARBERTv2 and the ensemble model were used as a baseline to compare the results with those of multimodal models. Table 17 presents the results of the multimodal models compared to those of the baseline models for the f1-score. In addition, Fig. 8 compares multimodal models with early and late fusion with baseline models.

TABLE 15. Multimodal models result.

Model	Accuracy	Precision	Recall	F1-score
Early Fusion	0.8557	0.8536	0.8607	0.8545
Late Fusion	0.8383	0.8373	0.8444	0.8372

TABLE 16. Multimodal model for rumor and not rumors class.

Model	Rumor Class			Not Rumor Class		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Early Fusion	0.88	0.87	0.87	0.83	0.83	0.85
Late Fusion	0.90	0.80	0.85	0.77	0.89	0.82

TABLE 17. Comparis on of multimodal with the baseline model.

Text model	Vision model	Early fusion	Late Fusion
0.8964	0.7900	0.8539	0.8372

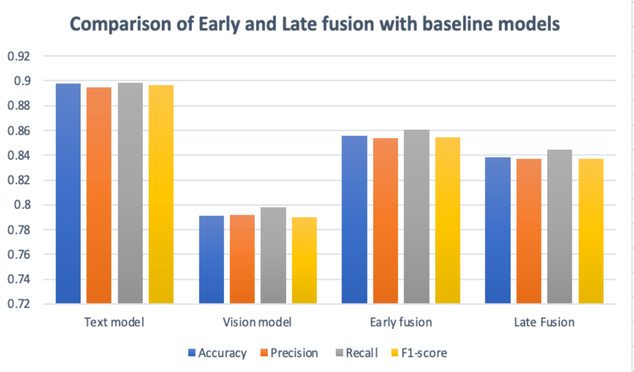


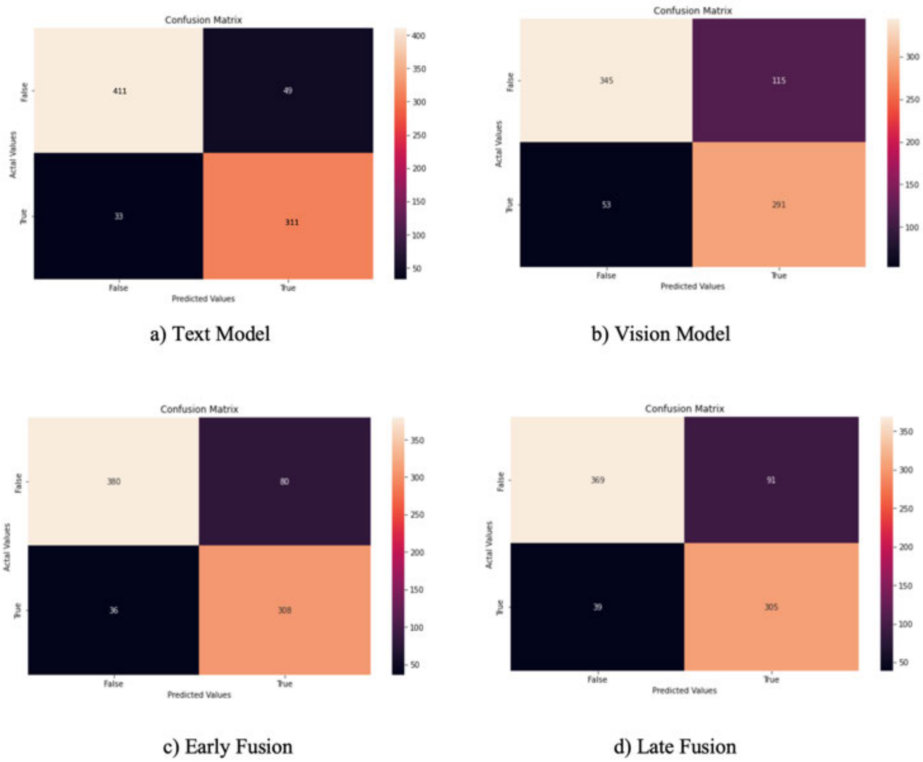
FIGURE 8. Comparison of Early and Late fusion with baseline models.

VII. DISCUSSION

In this work, different experiments were conducted to build an Arabic rumor detection model to detect Twitter rumors using textual and visual features. The first experiment involved selecting the best textual feature extractor. From Table 11, the experimental results show that all language models, except AraT5 MSA, can equally differentiate between the rumor and non-rumor classes. Furthermore, because the number of samples in each class was the same, the recall, precision, and f1-score were also similar. Further research should be conducted to investigate these models' performance using larger datasets. In addition, we found that MARBERTv2 achieved the highest result with an F1 score of 90%. That's because the proposed dataset had different Arabic language varieties, such as MSA and DA; simultaneously, the MARBERTv2 model was pre-trained on MSA and DA and trained on the Arabic news dataset.

The second experiment involved the selection of the best visual feature extractor. Visual features were incorporated into this model to benefit from the ability of neural networks to detect fake and manipulated images. In addition, the ensemble of the VGG-19 and ResNet50 models provides a robust performance compared with each model alone. Table 13 shows that ensemble learning increases the probability of detecting rumors that obtain a 79% f1-score.

The third experiment built a multimodal model by using the best single modality. From Table 17, we can observe that the fusion of text and images is better than that of the single modality for the vision models. Simultaneously, the fusion of text and images cannot outperform the single modality of



a) Text Model

b) Vision Model

c) Early Fusion

d) Late Fusion

FIGURE 9. Confusion matric for each model.

text-based models. Therefore, it is important to highlight that the fusion of textual and visual features is challenging owing to the differences between the feature spaces of text and images. In addition, it is challenging for the neural network to detect rumors tweets when the image is not manipulated but irrelevant to the text or if the text itself is unrelated to the image or was from another event.

Simultaneously, simple concatenation fusion fails to make the model find a correlation between two features. The experimental results of the multimodal models confirmed those obtained in a previous study [52]. Notably, these results may be due to the dataset size and nature of the images. The study [51] concluded that visual features are essential for fake news detection tasks, but their usefulness highly depends on the dataset.

Experiments show that language models can easily identify rumor patterns in tweet texts. We deduce that using neural networks, rumor and non-rumor words are easier to distinguish than visual features. In contrast, the vision models cannot find a pattern in the images of tweets; a possible explanation for the results of vision models may be the diverse types of images that do not have a relation or patterns that the model cannot detect. In addition, the dataset was composed of different domains such as politics, sports, health, and economics, and building a specific domain multimedia dataset may positively affect the results of these models. In addition, the extracted visual features are insufficient to enhance multimodal performance, and incorporating the correlation between the image caption, image source, and image content may help improve the obtained results.

The main finding of the experiments was that textual features are more crucial for detecting rumors than fusing textual and visual features for Arabic tweets.

According to a study [33], False-Negative (FN) and False-Positive (FP) are important for rumor detection tasks. Misclassifying rumors as non-rumors, and vice versa, affects and misleads society by spreading untruthful news. In our task, FN refers to non-rumor tweets (true labels) that are misclassified as rumor tweets, and FP refers to rumor tweets (false labels) that are misclassified as non-rumor tweets. From Fig. 9, which shows the confusion matrix of each model, we can observe that the text-based model shown in Fig. 9(a) is better than the other models because it shows an improvement in the number of FN and FP tweets.

Regarding these results, developing tools for studying the model's behavior and explaining its prediction is essential to analyze why the model misclassified some rumors as non-rumors and vice versa. In addition, it is essential to understand the role of each sub-model in the multimodal model for performance prediction.

VIII. CONCLUSION

Rumor proliferation can harm and mislead both individuals and society. The Arabic rumor detection task focuses on detecting rumors from textual features, and social media content includes different types of content, such as images. This study proposes a multimodal model that uses a fusion of tweets' text and image modalities to detect Arabic rumors on Twitter. The proposed multimodal models could not outperform unimodal text-based models, and it found that textual features are the most critical in Arabic rumor detection tasks. A limitation of this study was the size of the dataset. Future studies should be conducted to enhance these results. First, the dataset could be extended to include additional samples. Second, tools that analyze visual features can help the model differentiate between real and manipulated images, thereby improving the prediction performance of the vision model. Third, the attention mechanism can enhance the fusion of different models and identify the correlations between these features. Fourth, incorporating user features may improve multimodal model predictions. However, developing tools for studying the model's behavior and explaining its prediction is important to analyze why the model misclassified some rumors as non-rumors and vice versa.

REFERENCES

[1] N. DiFonzo and P. Bordia, "Rumor, gossip and urban legends," *Diogenes*, vol. 54, no. 1, pp. 19–35, 2007.

[2] M. Al-Sarem, W. Boulila, M. Al-Harby, J. Qadir, and A. Alsaeedi, "Deep learning-based rumor detection on microblogging platforms: A systematic review," *IEEE Access*, vol. 7, pp. 152788–152812, 2019.

[3] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *ACM Comput. Surv.*, vol. 51, no. 2, pp. 1–36, 2018.

[4] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Comput. Surv.*, vol. 53, no. 5, pp. 1–40, 2020.

[5] Y. Lan, Z. Lian, R. Zeng, D. Zhu, Y. Xia, M. Liu, and P. Zhang, "A statistical model of the impact of online rumors on the information quantity of online public opinion," *Phys. A, Stat. Mech. Appl.*, vol. 541, Mar. 2020, Art. no. 123623.

[6] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 598–608, Mar. 2017.

[7] S. Zannettou, M. Sirivianos, J. Blackburn, and N. Kourtellis, "The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans," *J. Data Inf. Qual.*, vol. 11, no. 3, pp. 1–37, Sep. 2019.

[8] A. Alsaeedi and M. Al-Sarem, "Detecting rumors on social media based on a CNN deep learning technique," *Arabian J. Sci. Eng.*, vol. 45, no. 12, pp. 10813–10844, Dec. 2020.

[9] T. Chen, X. Li, H. Yin, and J. Zhang, "Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*. Cham, Switzerland: Springer, 2018, pp. 40–52.

[10] A. Vijeev, A. Mahapatra, A. Shyamkrishna, and S. Murthy, "A hybrid approach to rumour detection in microblogging platforms," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2018, pp. 337–342.

[11] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 795–816.

[12] J. Chen, Z. Wu, Z. Yang, H. Xie, F. L. Wang, and W. Liu, "Multimodal fusion network with latent topic memory for rumor detection," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2021, pp. 1–6.

[13] 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.

[14] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "SpotFake: A multi-modal framework for fake news detection," in *Proc. IEEE 5th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2019, pp. 39–47.

[15] Z. S. Ali, W. Mansour, T. Elsayed, and A. Al-Ali, "AraFacts: The first large Arabic dataset of naturally occurring claims," in *Proc. 6th Arabic Natural Lang. Process. Workshop*, 2021, pp. 231–236.

[16] A. R. Pathak, A. Mahajan, K. Singh, A. Patil, and A. Nair, "Analysis of techniques for rumor detection in social media," *Proc. Comput. Sci.*, vol. 167, pp. 2286–2296, 2020.

[17] C. K. Hiramath and G. C. Deshpande, "Fake news detection using deep learning techniques," in *Proc. 1st Int. Conf. Adv. Inf. Technol. (ICAIT)*, Jul. 2019, pp. 411–415.

[18] L. Wu, F. Morstatter, K. M. Carley, and H. Liu, "Misinformation in social media: Definition, manipulation, and detection," *ACM SIGKDD Explorations Newslett.*, vol. 21, no. 2, pp. 80–90, Nov. 2019.

[19] J. Cao, J. Guo, X. Li, Z. Jin, H. Guo, and J. Li, "Automatic rumor detection on microblogs: A survey," 2018, *arXiv:1807.03505*.

[20] S. M. Alzanin and A. M. Azmi, "Rumor detection in Arabic tweets using semi-supervised and unsupervised expectation–maximization," *Knowl.-Based Syst.*, vol. 185, Dec. 2019, Art. no. 104945.

[21] A. R. Mahlous and A. Al-Laith, "Fake news detection in Arabic tweets during the COVID-19 pandemic," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, 2021.

[22] S. Alqurashi, B. Hamoui, A. Alashaikh, A. Alhindi, and E. Alanazi, "Eating garlic prevents COVID-19 infection: Detecting misinformation on the Arabic content of Twitter," 2021, *arXiv:2101.05626*.

[23] L. Alsudias and P. Rayson, "COVID-19 and Arabic Twitter: How can Arab world governments and public health organizations learn from social media?" 2020.

[24] H. Mubarak and S. Hassan, "ArCorona: Analyzing Arabic tweets in the early days of coronavirus (COVID-19) pandemic," 2020, *arXiv:2012.01462*.

[25] F. Saeed, M. Al-Sarem, and E. Abdullah, "Detecting health-related rumors on Twitter using machine learning methods," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, pp. 1–9, 2020.

[26] M. Alkhair, K. Meftouh, K. Smaïli, and N. Othman, "An Arabic corpus of fake news: Collection, analysis and classification," in *Proc. Int. Conf. Arabic Lang. Processing*. Cham, Switzerland: Springer, 2019, pp. 292–302.

[27] H. Himdi, G. Weir, F. Assiri, and H. Al-Barhamtoshy, "Arabic fake news detection based on textual analysis," *Arabian J. Sci. Eng.*, vol. 47, pp. 1–17, Feb. 2022.

[28] A. Gumaei, M. S. Al-Rakhami, M. M. Hassan, V. H. C. De Albuquerque, and D. Camacho, "An effective approach for rumor detection of Arabic tweets using eXtreme gradient boosting method," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 1, pp. 1–16, Jan. 2022.

[29] 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.

[30] M. Al-Sarem, A. Alsaeedi, F. Saeed, W. Boulila, and O. AmeerBakhsh, "A novel hybrid deep learning model for detecting COVID-19-related rumors on social media based on LSTM and concatenated parallel CNNs," *Appl. Sci.*, vol. 11, no. 17, p. 7940, Aug. 2021.

[31] S. Alharbi, K. Alyoubi, and F. Alotaibi, "Deep learning based rumor detection for Arabic micro-text," *Int. J. Comput. Sci. Netw. Secur.*, vol. 21, no. 11, pp. 73–80, 2021.

[32] M. Al-Yahya, H. Al-Khalifa, H. Al-Baity, D. AlSaeed, and A. Essam, "Arabic fake news detection: Comparative study of neural networks and transformer-based approaches," *Complexity*, vol. 2021, pp. 1–10, Apr. 2021.

[33] A. B. Nassif, A. Elnagar, O. Elgendy, and Y. Afadar, "Arabic fake news detection based on deep contextualized embedding models," *Neural Comput. Appl.*, vol. 34, pp. 16019–16032, May 2022.

[34] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[35] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1–9.

[36] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[37] Z.-Z. Lan, L. Bao, S.-I. Yu, W. Liu, and A. G. Hauptmann, "Multimedia classification and event detection using double fusion," *Multimedia Tools Appl.*, vol. 71, no. 1, pp. 333–347, Jul. 2014.

[38] J. Cao, P. Qi, Q. Sheng, T. Yang, J. Guo, and J. Li, "Exploring the role of visual content in fake news detection," in *Disinformation, Misinformation, and Fake News in Social Media*. Cham, Switzerland: Springer, 2020, pp. 141–161.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[42] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for Arabic language understanding," in *Proc. LREC Workshop Lang. Resour. Eval. Conf.*, May 2020, p. 9.

[43] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep bidirectional transformers for Arabic," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.* Toronto, ON, Canada: Association for Computational Linguistics, vol. 1, Aug. 2021, pp. 7088–7105. [Online]. Available: https://aclanthology.org/2021.acl-long.551

[44] A. Abdelali, S. Hassan, H. Mubarak, K. Darwish, and Y. Samih, "Pre-training BERT on Arabic tweets: Practical considerations," 2021, *arXiv:2102.10684*.

[45] A. Safaya, M. Abdullatif, and D. Yuret, "KUISAIL at SemEval2020 task 12: BERT-CNN for offensive speech identification in social media," in *Proc. 14th Workshop Semantic Evaluation. Barcelona Int. Committee Comput. Linguistics*, Dec. 2020, pp. 2054–2059.[Online]. Available: https://www.aclweb.org/anthology/2020.semeval-1.271

[46] M. S. Hadj Ameur and H. Aliane, "AraCOVID19-MFH: Arabic COVID-19 multi-label fake news & hate speech detection dataset," *Proc. Comput. Sci.*, vol. 189, pp. 232–241, 2021.

[47] E. M. B. Nagoudi, A. Elmadany, and M. Abdul-Mageed, "AraT5: Text-to-text transformers for Arabic language generation," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*. Dublin, Ireland: Association for Computational Linguistics, vol. 1, May 2022, pp. 628–647. [Online]. Available: https://aclanthology.org/2022.acl-long.47

[48] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[49] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2015, pp. 1–9.

[50] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.

[51] 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.

[52] L. Canas, "A multi-modal fake news classifier using transfer learning," MSc Dissertation, Dept. Comput. Sci., Univ. Sheffield, Sheffield, U.K., 2022.

[53] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. ICML*, 2011, pp. 1–8.

RASHA M. ALBALAWI received the bachelor's degree in computer science from the University of Tabuk, Tabuk, Saudi Arabia, in 2013. She is currently pursuing the master's degree in computer science with the Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. She is also working as a Teaching Assistant at the University of Tabuk. Her research interests include deep learning, natural language processing, and computer vision.

AMANI T. JAMAL received the master's and Ph.D. degrees from Concordia University, Montreal, Canada. She is currently an Associate Professor with the Computer Science Department, Faculty of Computing and Information Technology, King Abdulaziz University. Her current research interests include natural language processing and computer vision related to Arabic text and historical documents. She is also the Co-Founder and a member of the Director Board of the Saudi Artificial Intelligence Association.

ALAA O. KHADIDOS received the B.Sc. degree from King Abdulaziz University, Jeddah, Saudi Arabia, in 2006, the M.Sc. degree from the University of Birmingham, Birmingham, U.K., in 2011, and the Ph.D. degree from the University of Warwick, Coventry, U.K., in 2017, all in computer science. He is currently an Associate Professor with the Faculty of Computing and Information Technology, King Abdulaziz University. His research interests include computer vision, machine learning, optimization, and medical image analysis.

AREEJ M. ALHOTHALI received the master's and Ph.D. degrees in computer science (artificial intelligence) from the University of Waterloo, Canada, in 2017. She is currently an Associate Professor with the Faculty of Computer Science and Information Technology, King Abdulaziz University. Her research interests include the areas of machine learning, deep learning, natural language processing, computer vision, intelligent agent systems, and affective computing.