IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 July 2024, accepted 10 August 2024, date of publication 13 August 2024, date of current version 22 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3443092

RESEARCH ARTICLE

Improving Data Fusion for Fake News Detection: A Hybrid Fusion Approach for Unimodal and Multimodal Data

SUHAIB KH. HAMED[1], MOHD JUZAIDDIN AB AZIZ[1], AND MOHD RIDZWAN YAAKUB[2]

[1]Center for Software Technology and Management (SOFTAM), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor 43600, Malaysia

[2]Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor 43600, Malaysia

Corresponding author: Suhaib Kh. Hamed (p105401@siswa.ukm.edu.my)

This work was supported by Universiti Kebangsaan Malaysia (UKM) under Geran Skim Insentif Penerbitan under Grant GP-K007915.

ABSTRACT The proliferation of fake news, exacerbated by social media and modern technology, presents significant challenges across sectors such as health, the economy, politics, and national stability. This study addresses the limitations of current multimodal fake news detection models, which often struggle to effectively integrate heterogeneous modalities like text and images. We propose a hybrid data fusion approach (HF-TIM) that combines the early fusion of multimodal data with the late fusion of unimodal data, leveraging the strengths of both techniques to enhance detection accuracy. Our HF-TIM approach employs a Softmax classifier for early fusion and integrates it with unimodal features extracted from BERT and VGG-19 classifiers through a neural network-based meta-learning classifier. This approach captures the complementary and unique properties of each modality, resulting in a more comprehensive and robust fake news detection model. Experimental results demonstrate that the HF-TIM method significantly improves classification accuracy across various fake news categories by effectively addressing the complex interrelationships between text and images. Our fine-grained detection model, based on the HF-TIM method, achieved a detection accuracy of 93.4%, outperforming state-of-the-art models in related studies. The proposed hybrid fusion HF-TIM approach offers an innovative and effective solution for multimodal fake news detection, with potential applications extending to other domains.

INDEX TERMS Multimodal data, fake news detection, text and image, data fusion, meta-learning.

I. INTRODUCTION

The spread of fake news poses significant challenges across various sectors, including health, the economy, politics, and national stability [1]. 'Fake news' refers to news items published with misleading information intended to deceive readers for malicious purposes [2]. The primary goal of fake news is often to manipulate public opinion, mislead people, or achieve specific outcomes such as political gain, financial profit, or social disruption [3]. Social media and modern technology have facilitated the rapid dissemination of fake news [4], [5], predominantly in multimedia formats [6]. Literature shows that there are several types of multimodal fake news, including Satire, Misleading Content, Imposter Content, False Connection, and Manipulated Content [7]. In news posts, both images and text typically provide information about the same subject or concept. In ambiguous situations, extracting information from both modalities can be advantageous [8]. Despite the widespread sharing of images in news articles on social networks, their potential for verifying the authenticity of news on social media platforms has not been fully explored. Therefore, it is essential to fuse all types of features to enable a supervised deep-learning classifier to assess the credibility of news articles [9]. However, due to the heterogeneity of data from different modalities, effectively integrating this diverse information remains a significant challenge and a critical area for research breakthroughs [10].

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

Integrating information from various sources or modalities, known as data fusion, plays a vital role in improving the efficacy of multimodal fake news detection through deep learning models [11], [12]. By merging different types of data, such as text and images, data fusion enhances the representation and comprehension of complex issues. This integration of diverse data sources offers a more comprehensive viewpoint, allowing for the identification of significant patterns and relationships that might not be apparent when examining individual modalities separately. Such a holistic approach boosts the accuracy, robustness, and generalization capabilities of multimodal fake news detection models. Moreover, data fusion is highly impactful across various domains, including biomedicine [13], healthcare [14], environmental monitoring [15], and sentiment analysis [16], [17]. The importance of multimodal fusion lies in leveraging the unique strengths of each individual modality while simultaneously addressing their inherent limitations. In a world where data comes in various forms and formats, merging information from different modalities allows us to overcome the limitations of each individual data source. This comprehensive approach enables us to capture details, patterns, and relationships that might remain hidden when considering data sources in isolation.

The core concept of multimodal fusion is that various modalities offer complementary and valuable insights for identifying multimodal fake news. In the realm of fake news detection, it is challenging to retain the distinct characteristics of each modality while fusing pertinent information across them. Moreover, fusing data from different modalities can sometimes introduce noise, which can degrade the model's performance. For instance, redundant information can arise when combining text and image data that repeat or overlap, making the model's learning process less efficient [10]. Contradictory information is another issue, where different modalities provide conflicting details, such as a text description stating that a political figure gave a speech, while an accompanying image depicts a different event, confusing the model and leading to incorrect predictions [18]. Additionally, irrelevant data from different modalities might not contribute to the fake news detection task, such as background elements in an image unrelated to the news content, introducing unnecessary noise [19]. Misalignment noise also poses a problem, as temporal or spatial misalignment between modalities can occur, making it challenging to correlate data accurately, such as when an image is taken at a different time than the text was written, leading to inconsistencies [20]. Consequently, it is crucial to simultaneously consider both the original and integrated text and image data. Current multimodal fake news detection techniques frequently fail to meet these criteria [21].

Most data fusion methods employ a single fusion strategy or model, lacking fine-grained modal interactions [10]. Typically, these methods generate a joint representation by merely concatenating a text vector with an image vector, ignoring the dependencies between them [22]. Researchers often focus on simpler methods due to computational limitations. Early fusion techniques, such as concatenating feature vectors, are easier to implement but may overlook complex dependencies [23]. Moreover, aligning text and image data temporally or spatially presents challenges. Misalignments can introduce noise, complicating the capture of accurate dependencies [24]. Additionally, developing models capable of simultaneously processing and integrating multimodal information is difficult. Many models are specialized for either text or image data, but not both [21].

The successful integration of marginal and joint representations from diverse modalities is key to multimodal fusion. 'Marginal representation' involves transforming unimodal input data to reveal hidden useful elements. Conversely, 'joint representation' includes features that encapsulate latent factors derived from multiple modalities, thereby encoding information that can be complementary, redundant, or cooperative [25]. One primary challenge of multimodal fusion is determining the best way to combine and utilize various types of data. So far, feature-level fusion has been predominantly explored in previous fake news detection studies [26]. Despite these efforts, most multimodal studies have not adequately emphasized the unique characteristics of each modality, relying primarily on early fusion [19], [27], [28], [29]. There is a need to benefit from the advantages of both early and late fusion and combine them to produce a hybrid fusion method.

The motivation behind this research is that the field of fake news detection faces a significant challenge due to the lack of an advanced data fusion approach that combines the advantages of early and late fusion. Such a model should preserve the unique features of each modality (text and image) while effectively fusing the relevant multimodal features across different modalities. This fusion process is crucial for a comprehensive understanding of the individual and combined feature characteristics, as well as their semantic interrelations. Achieving this level of fusion is essential for accurately analyzing and identifying the nuanced attributes of fake news and increased detection accuracy. All the aforementioned challenges, as illustrated in **Fig 1**, reduce the effectiveness of multimodal data fusion, negatively affecting fine-grained multimodal fake news detection models and leading to poor detection accuracy. This study aims to answer the following research questions:

1) How can the integration of textual and visual data improve the accuracy of fake news detection models?
2) What are the limitations of current multimodal fake news detection methods, and how can a hybrid fusion approach address these limitations?
3) How can a data fusion method that combines early and late fusion techniques be optimized to integrate heterogeneous modalities in multimodal fake news detection models?

These research questions guide our study and provide a focused framework for investigating the efficacy of the

proposed method. Our approach aims to address the challenges of integrating unimodal and multimodal data, optimizing computational efficiency, and capturing complex dependencies between text and images. By answering these questions, we aim to demonstrate the effectiveness of the proposed method in improving the accuracy, scalability, and robustness of fake news detection systems. This study proposes a hybrid multimodal data fusion HF-TIM approach, which combines early data fusion prediction results for multimodal features, obtained using a Softmax classifier-based model, with late data fusion prediction results for unimodal features generated by BERT and VGG-19 classifiers. Hybrid data fusion involves combining the early fusion of multimodal data with the late fusion of unimodal data. The key contributions of this study include:

- **Proposed a Hybrid Fusion Method (HF-TIM):** Combining the early fusion of multimodal data with the late fusion of unimodal data to leverage the strengths of both techniques for the heterogeneity of data from different modalities.
- **Enhanced Detection Accuracy:** Achieved high detection accuracy, outperforming state-of-the-art models in related studies.

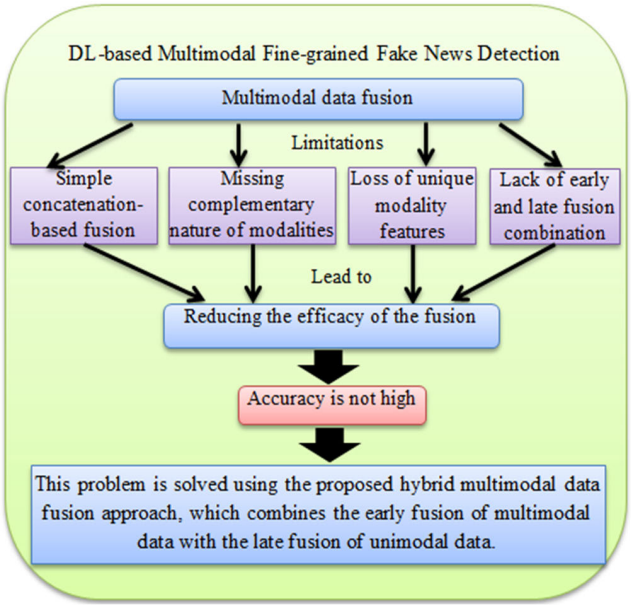


FIGURE 1. Multimodal feature representation problems and the proposed method for solving them.

The structure of this paper is as follows: Section II provides an overview of data fusion methods, including their advantages and disadvantages, while Section III investigates related studies in the field. Section IV outlines the methodology, describing the dataset used and detailing the components of our proposed fake news detection model. This section also introduces the hybrid data fusion HF-TIM approach for enhancing data fusion methods and outlines the evaluation metrics and baseline models for comparison. Section V presents the results of our experimental findings, and Section VI discusses these results and compares them with baseline studies. Finally, Section VII concludes the research, offering insights and directions for future work to further enhance multimodal fake news detection models.

II. MULTIMODAL DATA FUSION METHODS

Data fusion is a critical research domain in multimodal studies, involving the integration of data from various unimodal sources into a cohesive, multimodal representation. This technique has garnered significant attention due to its efficiency in processing multimodal data [30]. By combining different modalities, multimodal fusion extracts rich features [11], offering a more comprehensive understanding by capturing relationships between images and text. This enhances the contextual interpretation of events and can generate additional information, thereby improving result accuracy [31]. The primary challenge lies in effectively fusing and refining information from diverse modalities, each contributing unique aspects to the overall task. During the analysis of fusion features, it is essential to filter out noise and extract pertinent information [18]. As illustrated in **Fig. 2**, multimodal fusion methods include:

- **Early or Feature-Level Fusion:** This method integrates inputs from various modalities into one feature vector before feeding it into a learning model. Techniques such as concatenation, pooling, or gated units can achieve this. There are two types of early fusion: Type I, which combines original features, and Type II, which merges features extracted by another neural network. This approach captures and utilizes correlations between modalities at an early stage, facilitating comprehensive analysis when modalities are interdependent [32]. However, it may fail to identify intermodal relationships evident at higher abstraction levels since it does not explicitly learn marginal representations [25].
- **Joint Fusion or Intermediate Fusion:** Also known as intermediate fusion, this method integrates learned feature representations from the intermediate layers of neural networks with features from other modalities, serving as input to a final model. Unlike early fusion, joint fusion involves feeding the loss back to the neural network during training, which progressively enhances feature representation with each iteration. This technique, specifically Type I joint fusion, extracts and integrates feature representations from all modalities, thereby improving their overall integration and effectiveness [33].
- **Late or Decision-Level Fusion:** This method uses predictions from various models to arrive at a final decision, typically achieved through an aggregation function such as averaging, weighted voting, majority voting, or stacking. Each modality trains individual models, and their predictions are combined. Late fusion is beneficial when one modality dominates or when all unimodal models perform well. It allows for effective learning of good marginal representations, with each

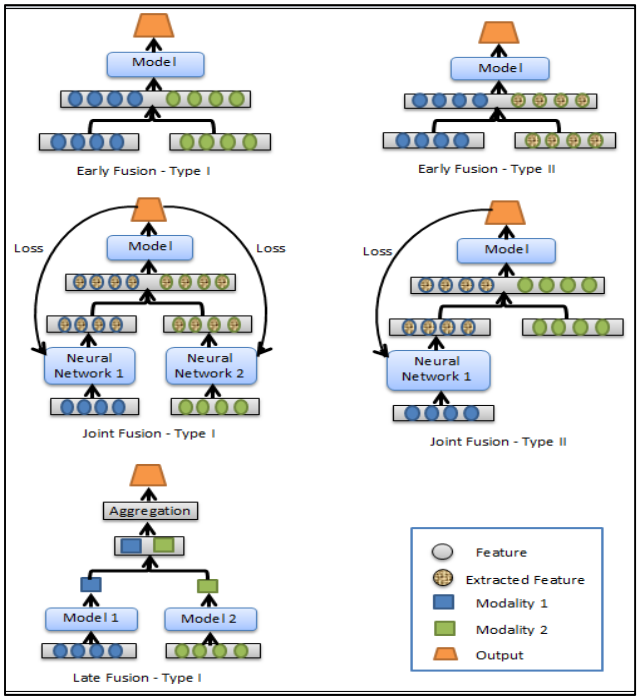


FIGURE 2. The types of multimodal data fusion.

model tailored to its specific modality. While late fusion enhances performance by handling errors from multiple models independently, improvements are realized only when models complement each other [34], [35], [36]. This approach is popular for its simplicity, especially when data sources vary significantly in sampling rate, dimensionality, and measurement units [37], [38]. Using multiple information modalities provides complementary insights and enhances the accuracy of the overall decision-making process. To establish the most effective fusion approach, it is crucial to determine the appropriate level for implementing the fusion strategy and how to effectively combine the information. For visual-textual classification tasks, fusion can be implemented at three levels: early (feature-level), intermediate (joint-level), and late (decision-level). These levels are defined based on the type of information available in a specific field. Table 1 provides an overview of these fusion techniques, detailing their advantages and disadvantages [10], [20], [25], [26], [32], [35], [39].

III. RELATED WORKS

Research on multimodal fake news detection has predominantly adopted an early data fusion approach. Most of these studies [19], [27], [28], [29], [40], [41], [42], [43], [44], [45] employed various methods, such as concatenation, maximum, and average, to directly fuse extracted text and image features into a multimodal vector. However, some studies [24], [46] utilized progressive fusion to represent shared features between modalities (text and image), capturing their interrelationships. Others [21], [22], [47] applied an attention mechanism to highlight relevant information across text and image modalities. Despite these efforts, most studies have

TABLE 1. Advantages and disadvantages of data fusion methods.

Fusion Approach	Advantages	Disadvantages
Early- or feature-level fusion	• Can exploit early-stage correlations between features from different modalities, improving task performance. • Demands just one learning phase for the unified feature vector. • Implementation is relatively simple; only one model needs to be trained. • Fused data contains more information than data from a single modality, often resulting in superior performance. • Tends to outperform unimodal approaches.	• Fails to encompass the complementary aspects of various modalities. • Creates high-dimensional feature vectors that might include redundant information. • Requires uniform formatting of all features before fusion. • Cannot handle time synchronicity across modalities, complicating the retrieval of features from interconnected modalities. • Difficulty increases with the number of modalities due to challenges in learning cross-correlations among heterogeneous features. • Converting data sources into a fixed representation is challenging and time-consuming. • May remove significant data from modalities before fusion, leading to a loss of potentially useful information. • Requires highly engineered and preprocessed features to align or semantically match different modalities. • Fusion using low-level features might be irrelevant to the task, reducing the efficacy of the fusion. • Sensitive to varying sampling rates across modalities. • Joint representation modeling may obscure useful representations of individual modalities. • Higher-level abstract features, relevant to one modality, may be obscured in a joint representation. • Modalities can have different relationships with one another, complicating joint modeling. • Lacks interaction within modalities, potentially

TABLE 1. *(Continued.)* **Advantages and disadvantages of data fusion methods.**

Fusion Approach	Advantages	Disadvantages
		leading to data sparsity issues.
Intermediate - or joint-level fusion	• Simulates interactions between features from several modalities. • Improves the learning of feature representations from each modality. • Combines inputs at various abstraction levels. • Requires training only one model.	• This may result in overfitting; careful design is required to simulate relationships effectively between modalities. • Ineffective at performing multimodal classification when parts of multimodal information are missing.
Late- or decision-level fusion	• Can predict outcomes even when not all modalities are available. • Does not require large quantities of training data. • Allows the use of the most suitable methods for each modality without needing data format conversion. • Offers more flexibility and is scalable, supporting graceful degradation or enhancement during the fusion process. • Sub-models' errors can be uncorrelated, providing complementary effects.	• Cannot capture relationships between several modalities. • Ignores low-level interactions between modalities. • Difficult to effectively ensemble all classifiers. • May miss local interactions between modalities.

not adequately emphasized the unique characteristics of each modality.

Simple vector concatenation, commonly used in early fusion, may overlook significant inter-modal relationships, potentially introducing redundancy and noise into the feature vector. If multimodal and single-modality features convey overlapping information, this redundancy could detract from the mode's value and complicate the learning process by adding noise. Table 2 provides a critical analysis of these studies, highlighting the limitations that impacted their detection accuracy results. Notably, only four studies presented a fine-grained classification of fake news: Kalra et al. [28], Segura-Bedmar and Alonso-Bartolome [19], Wang et al. [27], and Liu et al. [29], while the rest focused on binary classifications.

This analysis underscores the common limitation of early fusion methods: the loss of modality-specific features and the failure to account for inter-feature correlations across different modalities. These limitations highlight the need for more advanced data fusion techniques that can better capture

TABLE 2. **A critical analysis of fake news detection studies regarding data fusion and their limitations.**

Study	Data fusion Approach	Fusion method	Acc.-based result	Limitation
Wang, et al. [40]	Early Fusion	Concatenation	71.5% and 82.7%	Loss of Modality-Specific Features
Singhal, et al. [41]	Early Fusion	Concatenation	77.77% and 89.23%	Loss of Modality-Specific Features
Zhang, et al. [42]	Early Fusion	Concatenation	83% and 84.2%	Loss of Modality-Specific Features
Giachanou, et al. [43]	Early Fusion	Concatenation	92.5%, 62.2% and 82.9%	Loss of Modality-Specific Features
Song, et al. [21]	Early Fusion	Concatenation	92%	Loss of Modality-Specific Features
Wu, et al. [24]	Early Fusion	Concatenation	80.9% and 89.9%	Loss of Modality-Specific Features
Kumari and Ekbal [47]	Early Fusion	Element-wise multiplication	88.3% and 88.2%	Loss of Modality-Specific Features
Segura-Bedmar and Alonso-Bartolome [19]	Early Fusion	Concatenation	87%	Loss of Modality-Specific Features
Palani, et al. [44]	Early Fusion	Average	92% and 93%	Loss of Modality-Specific features and Ignoring inter-feature correlations across modalities
[22]	Early Fusion	Element-wise sum	88.5%	Loss of Modality-Specific Features
Uppada and Patel [45]	Early Fusion	Maximum	91.94%	Loss of inter-feature correlations across modalities

TABLE 2. *(Continued.)* **A critical analysis of fake news detection studies regarding data fusion and their limitations.**

Kalra, et al. [28]	Early Fusion	Concatenation	60.83%	Loss of Modality-Specific Features
Liu, et al. [29]	Early Fusion	Concatenation	88.3% and 90.57%	Loss of Modality-Specific Features
Jing, et al. [46]	Early Fusion	Concatenation	83.3% and 83.8%	Loss of Modality-Specific Features
Wang, et al. [27]	Early Fusion	Concatenation	89.82%	Loss of Modality-Specific Features

and integrate the unique aspects of each modality, leading to improved accuracy and robustness in fake news detection models.

IV. METHODOLOGY

This section provides a detailed description of the Fakeddit dataset used in our experiments and explains the proposed hybrid data fusion (HF-TIM) approach. It also introduces the multimodal MFND-HF-TIM model for fake news detection, which leverages the HF-TIM approach. Additionally, we outline the evaluation metrics used to assess the performance of our proposed models and discuss the baseline models used for comparative analysis.

A. DATASET

The core dataset for our study is 'Fakeddit', a state-of-the-art, large-scale multimodal dataset tailored for fine-grained fake news detection. Created by Nakamura et al. [7], Fakeddit was compiled from Reddit posts ranging from March 19, 2008, to October 24, 2019. This dataset encompasses over a million posts across various domains, featuring multiple attributes such as users, images, comments, domains, and additional metadata. Fakeddit offers three classification schemes for each post: binary, three-way, and six-way. The six-way labeled dataset is imbalanced and includes 682,461 examples comprising images and their captions. After downloading the dataset images, the total dataset comprised 124,530 labeled examples, each consisting of an image and a title. This was after removing records that lacked an image link or had non-functional links. The dataset was then divided into 80% for training, 10% for validation, and 10% for testing. Table 3 outlines the characteristics of the six categories within the Fakeddit dataset.

B. PROPOSED HYBRID DATA FUSION (HF-TIM) METHOD

We propose a hybrid multimodal data fusion approach that combines early data fusion prediction results for multimodal features, obtained using a Softmax classifier-based model, with late data fusion prediction results for unimodal features generated by BERT and VGG-19 classifiers. Hybrid data fusion involves integrating the early fusion of multimodal data with the late fusion of unimodal data, as detailed in Algorithm 1. **Fig 3.** illustrates the process of the proposed HF-TIM approach.

TABLE 3. **Characteristics of the six categories of the Fakeddit dataset.**

Label	Category	Description	Relation between Image and caption
True	Real News	Accurate information with real images and precise captions	Has relation
Satire	Fake News	Material that uses sarcasm or false information to comment on current events	Has relation
Misleading Content	Fake News	Information deliberately altered to deceive	Has relation
Imposter Content	Fake News	Content generated by bots	Has relation
False Connection	Fake News	A mismatch between image content and textual descriptions	No relation
Manipulated Content	Fake News	Material manually altered through image editing or similar techniques	Has relation

The proposed approach consists of a stacking ensemble with two primary levels, Level-0 and Level-1, as illustrated in **Fig 4**.

- In Level 0:
  1) Implement the three base models: BERT-Model 1, VGG-19-Model 2 (unimodal), and Softmax-Model 3 (multimodal).
  2) During stacking training, merge the output probabilities of each base model's validation set.
  3) During the stacking test, combine the output probabilities of each base model's testing set.
- In Level 1:
  1) Use stacking validation to train and optimize the meta-learner based on Softmax.
  2) Use stacking testing to evaluate the meta-learner and make the final prediction results.

Through this hybrid fusion strategy, our model combines the advantages of feature-level and decision-level fusion methods to better integrate the modalities of image and text.

1) STACKING DL MODELS

Stacking is a heterogeneous ensemble method that combines base learners (Level-0) with a meta-classifier (Level-1). The proposed stacking fusion model, which possesses a powerful

Algorithm 1 Hybrid Fusion for Multimodal Fake News Detection

Input Textual data T, Image data I
Output Final prediction of multimodal fake news P_F
Step1 Early Fusion of Multimodal Data
Extract textual features F_T using BERT from T
Extract image features F_I using VGG-19 from I
Merge F_T and F_I using concatenation method
Fuse multimodal features through multiple layers
Use a Softmax classifier-based model to provide predictions P_M
Step2 Late Fusion of Predictions of Unimodal and Multimodal Models
Use P_M generated by the Softmax classifier for multimodal data from Step 1
Use predictions P_T generated by BERT classifier for textual data T
Use predictions P_I generated by VGG-19 classifier for image data I
Merge P_M, P_T, and P_I using a meta-learning classifier-based NN with Softmax
Produce the final prediction P_F of multimodal fake news

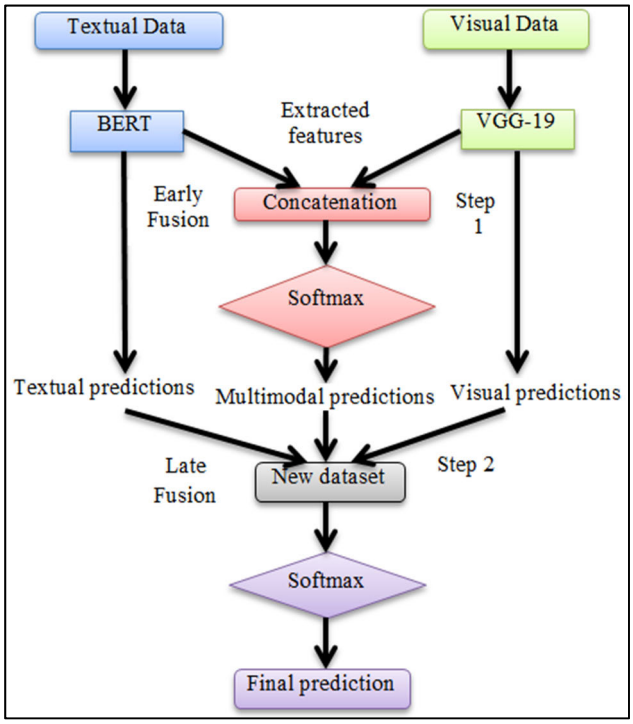


FIGURE 3. HF-TIM process diagram.

generalization capability, integrates BERT, VGG-19, and Softmax as the base learners at Level-0 and utilizes the meta-learning classifier at Level-1 to construct a new prediction model. This stacking fusion method enhances overall prediction accuracy by generalizing the output from multiple models. This section clarifies the base classifiers used to build the stacking prediction model. We employed three deep-learning-based classifiers, detailed in the following sections.

- **Base model 1 (BERT classifier):** The first base classifier at Level-0, which deals with text modality and

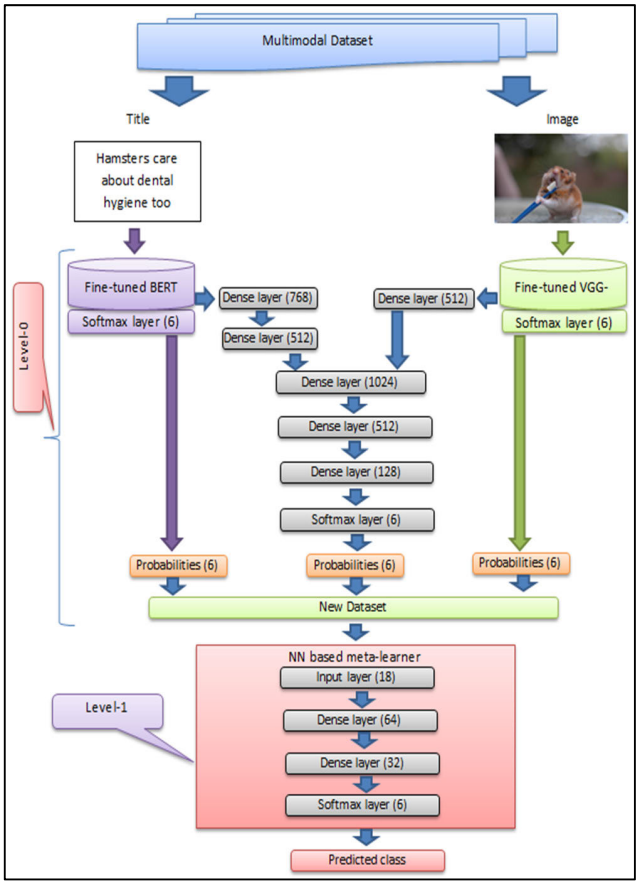


FIGURE 4. MFND-HF-TIM model architecture.

classifies textual news, is BERT. BERT is a state-of-the-art pre-trained language model known for its ability to capture deep semantic and contextual information from text. It has been widely adopted in natural language processing (NLP) tasks due to its high performance and effectiveness in understanding complex language patterns [48], [49]. Studies have shown that BERT significantly outperforms traditional NLP models in various tasks such as sentiment analysis, question answering, and text classification [50]. The BERT model will be fine-tuned. The fusion of features (probabilities) from the BERT model with probabilities from other modalities falls within the category of late fusion methods.

- **Base model 2 (VGG-19 classifier):** The second base classifier at Level-0, which deals with image modality and classifies visual news, is VGG-19. VGG-19 is a deep convolutional neural network (CNN) known for its excellent performance in image classification and feature extraction. It is pre-trained on the ImageNet dataset, which allows it to learn rich and diverse visual features. VGG-19 has been demonstrated to achieve high accuracy in various image recognition tasks and is commonly used in research and industry [51], [52]. Compared to traditional CNNs, VGG-19 provides superior performance in capturing intricate visual details [53]. The

VGG-19 model will be fine-tuned. The fusion of features (probabilities) from the visual news classification model with probabilities from other modalities falls within the category of late fusion methods.

- **Base model 3 (Softmax classifier):** The third base classifier at Level-0, which deals with multimodalities (text and image) and classifies multimodal news, is the Softmax-based model. The Softmax classifier is a widely used activation function in neural networks for classification tasks. It converts raw scores into probabilities, making it suitable for combining features from multiple modalities. The use of a Softmax classifier in early fusion allows for effective integration of textual and visual features, leveraging the complementary information from both modalities [54]. This multimodal model will fuse the text features extracted from BERT and the image features extracted from VGG-19 based on a progressive multimodal fusion. It combines text and image data through concatenation, progressively refining the combined representation through dense layers, culminating in a classification task. This fusion falls under the early fusion method. The features (probabilities) generated from the multimodal news classification model will then be fused with the probabilities generated from other modalities.
- **Stacking prediction models:** Stacking can significantly improve predictive performance by combining the strengths of different models while compensating for their individual weaknesses [23]. We employ three base classifiers, $\boldsymbol{C_1}$, $\boldsymbol{C_2}$ and $\boldsymbol{C_3}$, representing BERT, VGG-19, and Softmax, respectively. These classifiers, referred to as Level-0 learners or generalizers, operate within the Level-0 space, which consists of distinct learning sets for each modality. These learning sets are further divided into subsets: the training set is used to train the learners, while the validation and testing sets are employed to make predictions at Level-0. Each classifier outputs a probability distribution over six classes for each sample. For instance, for a sample $\boldsymbol{x_i}$:

$$\boldsymbol{C_1}(\boldsymbol{x_i}) = [\boldsymbol{p_{i1}}, \boldsymbol{p_{i2}}, \boldsymbol{p_{i3}}, \boldsymbol{p_{i4}}, \boldsymbol{p_{i5}}, \boldsymbol{p_{i6}}] \tag{1}$$

where $\boldsymbol{p_{ij}}$ represents the probability that $\boldsymbol{x_i}$ belongs to class j according to the classifier $\boldsymbol{C_1}$. or each sample, we obtain three vectors, each containing six probabilities. These vectors are concatenated to form a single feature vector for each sample. Therefore, for a sample $\boldsymbol{x_i}$, the new feature vector $\boldsymbol{x'_i}$ will be:

$$\begin{aligned}\boldsymbol{x'_i} &= [\boldsymbol{C_1}(\boldsymbol{x_i}), \boldsymbol{C_2}(\boldsymbol{x_i}), \boldsymbol{C_3}(\boldsymbol{x_i})] \\ &= \begin{bmatrix} \boldsymbol{p_{i11}}, \boldsymbol{p_{i12}}, \boldsymbol{p_{i13}}, \boldsymbol{p_{i14}}, \boldsymbol{p_{i15}}, \boldsymbol{p_{i16}}, \boldsymbol{p_{i21}}, \boldsymbol{p_{i22}}, \\ \boldsymbol{p_{i23}}, \boldsymbol{p_{i24}}, \boldsymbol{p_{i25}}, \boldsymbol{p_{i26}}, \boldsymbol{p_{i31}}, \boldsymbol{p_{i32}}, \boldsymbol{p_{i33}}, \boldsymbol{p_{i34}}, \\ \boldsymbol{p_{i35}}, \boldsymbol{p_{i36}} \end{bmatrix}\end{aligned} \tag{2}$$

The new dataset (validation and testing sets) will have n samples, with each sample represented by an 18-dimensional feature vector (since each of the three classifiers contributes six probabilities). The Level-0 predictions form the Level-1 learning sets, where $\boldsymbol{C_i}$ are the individual models trained independently on each modality. Two sets of predictions are generated using three different integrated base models with high prediction accuracy as base learners. These sets of predictions, along with the labels of the original sets which consider high-level features, are then fed into the second level, employing a meta-learner, chosen to train and validate the stacking model to achieve the final stacking model prediction results. The prediction from this final learner is tested on an unseen third subset of data. This hybrid fusion mechanism, based on the meta-learning classifier, serves as an additional deep-learning algorithm.

2) META-LEARNING CLASSIFIER

The goal of stacking is to combine the strengths of various base models by feeding their predictions into a meta-model, which learns to weigh and integrate these predictions to generate the final outcome. This approach often results in higher performance than using a single model alone. Meta-learning classifiers are designed to optimize the learning process by using information from multiple models [55]. In our approach, a neural network-based meta-learning classifier is used for late fusion, which combines predictions from the BERT, VGG-19, and Softmax classifiers. This approach ensures that the unique properties of each modality are retained and effectively integrated, improving the overall performance of the fake news detection model. The meta-learning classifier is trained using the predictions provided by the base models on the validation set. These predictions serve as high-level features for the meta-model. In this work, a neural network (NN) based on a Softmax classifier is utilized as the meta-learner. The architecture of the meta-learning classifier is as follows:

- **Input Layer:** An 18-dimensional input layer designed to accommodate the predictions in the new dataset as a feature vector.
- **Hidden Layers:** The ReLU activation function is used for each hidden layer. Two fully connected (dense) layers with dimensions of 64 and 32 serve as the first and second hidden layers, respectively. These layers perform the main computations and transformations on the input data, learning complex patterns and representations. The first layer acts as a feature extractor, transforming the raw input predictions into a set of higher-level features. These features are more informative and can be more easily processed by subsequent layers. The second layer refines the features learned in the first layer by combining and re-weighting them in a more compact form.
- **Output Layer:** A 6-dimensional output layer, utilizing Softmax as a classifier, is employed to produce the final predictions of the model.

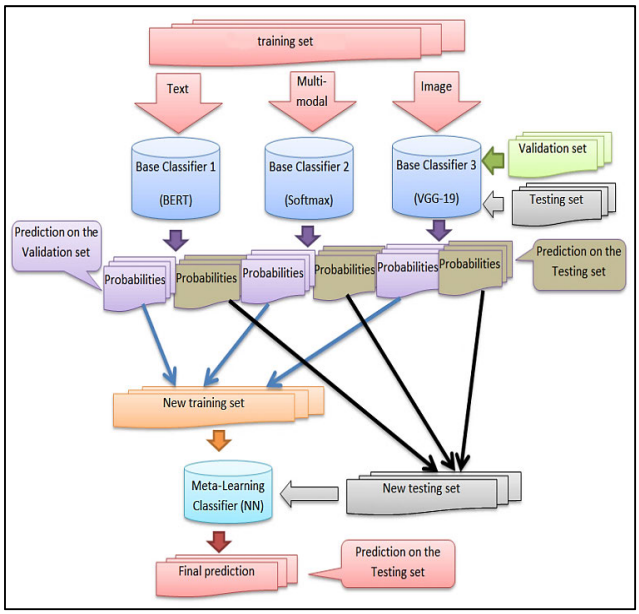


FIGURE 5. The process of creating the new training and test sets.

C. PROPOSED MULTIMODAL FAKE NEWS DETECTION (MFND-HF-TIM) MODEL

This section provides a detailed description of the components of the proposed MFND-HF-TIM model, as illustrated in **Fig 4**. This model employs the hybrid data fusion approach (HF-TIM) to enhance data fusion, aiming to evaluate the effectiveness of the HF-TIM approach and its impact on detection performance.

1) DATA PREPROCESSING PHASE

In this phase, we outline the process of creating new datasets for training and testing the meta-learning classifier. After training the base models (BERT, VGG-19, and the Softmax classifier-based multimodal model), these models are used to make predictions on the validation and test sets. The predictions from these base classifiers on the validation and test sets are then collected to form new datasets. Specifically, a new training dataset is generated from the predictions on the validation set, and a new test dataset is created from the predictions on the test set, as illustrated in **Fig 5**. These new datasets are subsequently used to train and test the meta-learning classifier.

2) MULTIMODAL DATA FUSION PHASE

In this phase, we apply the HF-TIM method to effectively integrate multimodal data. The detailed procedure for implementing the HF-TIM method is outlined in section IV-B. This approach combines the early fusion of multimodal data with the late fusion of unimodal data, leveraging the strengths of both techniques. The primary focus is on enhancing feature fusion, as understanding the interrelationships between unimodal and multimodal features contributes to developing a robust model with high detection accuracy.

TABLE 4. Hyperparameters of the proposed multimodal fake news detection model (MFND-HF-TIM).

Hyperparameters	Value
Optimizer	Adam
Loss function	Cross-entropy
Learning rate	0.0001
Batch size	64
Epochs	20
Training strategy	Early stopping

3) MFND-HF-TIM MODEL CONFIGURATION

To identify the optimal architecture and appropriate hyperparameters for the proposed MFND-HF-TIM model, we conducted a series of experiments comparing different hyperparameter values. The selected hyperparameters are based on the best-conducted experiment, which significantly enhanced the performance of the MFND-HF-TIM model. The most successful experiment established both the model's architecture and the hyperparameters, detailed in Table 4.

D. EVALUATION METRICS

To evaluate the performance of the proposed models, we use several key metrics: precision, recall, F1-score, and accuracy. For the fine-grained classification of news into six classes, precision, recall, F1-score, and accuracy are calculated for each class individually. Additionally, macro-average and weighted-average metrics are computed to assess the overall model performance. Macro-average metrics are particularly useful for imbalanced datasets as they treat each class equally. Weighted-average metrics, on the other hand, assign weights to each class based on their representation in the dataset. For our evaluation, we focus on macro-average F1-Score and accuracy to compare the performance of baseline models. Higher values of these metrics indicate superior model performance.

E. BASELINE MODELS

To evaluate the effectiveness of the proposed hybrid data fusion HF-TIM method, we apply it to the multimodal fake news detection model, MFND-HF-TIM, and analyze its impact on the model's performance. Additionally, we compare MFND-HF-TIM against various baseline models in the realm of fake news detection:

- **Segura-Bedmar and Alonso-Bartolome [19] Model:** Utilizes a CNN algorithm to extract and concatenate image and text features into a multimodal vector, which undergoes early fusion for classification.
- **Kalra et al. [28] Model:** Employs DistilBERT for text feature extraction and VGG-16 for image feature extraction. These features are combined through early fusion using concatenation.
- **Wang et al. [27] Model:** Encodes text and image captions with BERT's tokenizer and extracts global and entity image features via ResNet and Faster R-CNN.

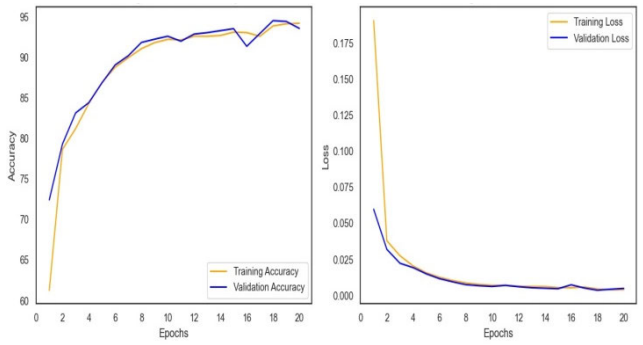


FIGURE 6. The accuracy and loss of training and validation of the MFND-HF-TIM model.

These features are concatenated in an early fusion method to create a multimodal embedding, processed by a multimodal transformer.

- **Liu et al. [29] Model:** Integrates image and text information by generating image captions and merging them with text. ResNet and Faster R-CNN extract image features, while BERT encodes text. The concatenated features are analyzed by a multimodal transformer with a self-attention mechanism to capture cross-modal interactions for accurate classification.

V. EXPERIMENTAL RESULTS

The experiments for the proposed multimodal fake news detection models were conducted using Python version 3.10 within the Spyder editor version 5.15. Our experimental setup included a PC equipped with an NVIDIA GeForce GTX 1060 GPU, featuring 6 GB of VRAM, to efficiently train and test the models. To evaluate the effectiveness of the proposed method in enhancing multimodal data fusion, experiments were performed using the fine-grained multimodal Fakeddit dataset and assessed across various metrics. Additionally, the results of the MFND-HF-TIM model were evaluated to compare the proposed hybrid data fusion method with baseline models.

A. EXPERIMENT: EVALUATION OF THE MFND-HF-TIM MODEL

The stacking model, which combines the base models at Level 0 and the meta-learning classifier at Level 1, was implemented and evaluated as a single model: MFND-HF-TIM. This proposed multimodal detection model, utilizing the hybrid data fusion (HF-TIM) method, achieved excellent results. The macro-average and weighted-average metrics presented in Table 6 demonstrate the model's strong performance, with an accuracy of 93.4%, as shown in **Fig. 6**. This enhancement positively impacted the multimodal fake news detection model (MFND-HF-TIM), as evidenced by the high precision and recall values for the news classes detailed in Table 5.

Evaluating the MFND-HF-TIM model based on the F1-score across various classes reveals insightful performance

TABLE 5. Evaluation of the MFND-HF-TIM model based on precision, recall, and F1-score for each class.

Class	Precision	Recall	F1-score
True	0.977	0.954	0.965
Satire	0.945	0.928	0.936
Misleading content	0.952	0.929	0.940
Manipulated content	0.878	0.851	0.864
False connection	0.893	0.918	0.905
Imposter content	0.826	0.835	0.830

TABLE 6. Evaluation of the MFND-HF-TIM model based on accuracy, macro-average, and weighted-average.

Measure	Precision	Recall	F1-score	Accuracy
Accuracy	-	-		0.934
Macro-average	0.911	0.902	0.907	-
Weighted-average	0.940	0.931	0.935	-

metrics. The model exhibits outstanding performance in identifying true content, achieving an impressive F1-score of 0.965. This high score indicates balanced precision and recall, showcasing the model's ability to accurately and reliably classify true content. Similarly, the model performs exceptionally well in detecting satire and misleading content, with F1-scores of 0.936 and 0.940, respectively. These scores demonstrate the model's effectiveness in distinguishing these types of content, highlighting its robustness in handling the subtle nuances that differentiate satire and misleading information.

However, the model shows relatively moderate performance in classifying manipulated content, as evidenced by an F1-score of 0.864. While this score still indicates good performance, it suggests that the model encounters more challenges in accurately identifying manipulated content, possibly due to its complex and deceptive nature. For false connections, the model achieves a commendable F1-score of 0.905, indicating a strong ability to detect content that misleads by connecting unrelated pieces of information. The classification of imposter content presents the most significant challenge for the model, with an F1-score of 0.830. This lower score suggests a need for further refinement and improvement in this area.

Overall, the application of the proposed hybrid fusion HF-TIM method, which combines the early fusion of multimodal features with the late fusion of unimodal features, significantly improved classification results across all categories. This improvement underscores the importance of hybrid fusion, which leverages the strengths of three heterogeneous models: multimodal, uni-textual, and uni-visual, through the high-level features generated by these models. This approach addresses any weaknesses that may exist in multimodal models.

Specifically, the 'False Connection' category, characterized by the lack of a relationship between the text and the image, highlights that the early fusion method alone may

not be sufficient to achieve high classification results. This necessitates leveraging the results of the late fusion of unimodal data to fully understand the nature of the features in this category and to cover it comprehensively. The use of the hybrid fusion method (HF-TIM) positively impacted the effectiveness of the multimodal fake news detection model (MFND-HF-TIM), demonstrating the benefits of utilizing the hybrid fusion method in complex classification tasks involving heterogeneous modalities and different categories in terms of the relationship between multimodal features such as text and images.

B. COMPARISON WITH BASELINE MODELS

To evaluate the performance of our proposed MFND-HF-TIM model against existing baseline models, we conducted a series of experiments to highlight the advantages of our hybrid data fusion (HF-TIM) method. This approach significantly enhances multimodal data fusion, thereby improving the accuracy of fake news detection. We benchmarked the performance of the MFND-HF-TIM model against various baseline models, as illustrated in Table 7. All the baseline models, including those by Kalra et al. [28], Segura-Bedmar and Alonso-Bartolome [19], Wang et al. [27], and Liu et al. [29], utilized the early data fusion method to combine text and image modalities. The detection accuracies achieved by these studies were 60.3%, 87%, 89.8%, and 90.5%, respectively, as shown in Table 8.

Our proposed MFND-HF-TIM model significantly outperforms the current state-of-the-art models, achieving an accuracy of 93.4%. This represents a 3.6% improvement over the model by Wang et al. [27] and a 2.9% improvement over the model by Liu et al. [29]. By comparing the HF-TIM method with existing approaches, we highlight its superior performance in key areas such as accuracy, multimodal data handling, scalability, and robustness to noise. This comparison underscores the enhanced performance of our hybrid data fusion method (HF-TIM) in improving the accuracy of fake news detection models. By leveraging the strengths of both early and late fusion techniques, our approach sets a new benchmark in the field of multimodal fake news detection.

VI. DISCUSSION

This study aims to enhance data fusion methods in fake news detection by introducing the HF-TIM approach. This innovative method integrates early and late fusion, combining both homogeneous and heterogeneous modalities, and merging multimodal and unimodal data to address the specific needs of the research task and the unique characteristics of the dataset. The hybrid fusion method utilizes early fusion of multimodal data, implemented with a Softmax classifier-based model, alongside late fusion of unimodal data generated by BERT and VGG-19, to better align with our research objectives. Our experiments investigate the proposed HF-TIM method and assess how these data fusion enhancements impact the performance of the multimodal fake news detection model (MFND-HF-TIM).

TABLE 7. Data fusion methods used in baseline models.

Model	Modality	Data fusion approach	Fusion method
Segura-Bedmar and Alonso-Bartolome [19]	Text and image	Early fusion	Concatenation
Kalra, et al. [28]	Text and image	Early fusion	Concatenation
Wang, et al. [27]	Text and image	Early fusion	Concatenation
Liu, et al. [29]	Text and image	Early fusion	Concatenation
MFND-HF-TIM	Text and image	Hybrid fusion	Stacking based on Meta-learning

TABLE 8. Comparison with fake news detection baseline models using the benchmarking dataset Fakeddit.

Model	Accuracy	Precision (Macro-average)	Recall (Macro-average)	F1-score (Macro-average)
Segura-Bedmar and Alonso-Bartolome [19]	87.0	76.0	70.0	72.0
Kalra, et al. [28]	60.3	-	-	-
Wang, et al. [27]	89.8	-	-	-
Liu, et al. [29]	90.5	-	-	-
MFND-HF-TIM	93.4	91.1	90.2	90.7

Compared to baseline models, our proposed HF-TIM approach for improving feature fusion enabled the MFND-HF-TIM model to outperform these benchmarks. Given the specific characteristics of certain dataset categories, such as the 'False Connection' category, where the text does not describe the images and no common relationship exists, it is crucial to analyze the unimodal features of each modality in addition to the multimodal features. Similar limitations are observed in studies by Wang et al. [27], Kalra et al. [28], Liu et al. [29], and Segura-Bedmar and Alonso-Bartolome [19], which rely solely on the early fusion of multimodal data and fail to account for subtle differences within the dataset categories that are vital for enhancing data fusion. The proposed MFND-HF-TIM model excels by considering both unimodal and multimodal features, thereby addressing all the nuances of the categories in our task.

Thus, the HF-TIM approach offers a flexible framework that can adapt to heterogeneous multimodal data, regardless of whether these multimodal features are interrelated. This adaptability highlights the broader significance and potential impact of our contribution to the field of analyzing, fusing, and classifying heterogeneous multimodal features in deep learning models. Implementing the HF-TIM method in real-world scenarios involves addressing several challenges, including computational complexity, data processing requirements, system integration, and real-time processing

capabilities. Parallel processing techniques, such as task distribution across multiple processors or cores, can significantly reduce computation time. Hardware accelerators like GPUs and TPUs enhance the performance of deep learning models by handling neural network computations efficiently. Distributed computing frameworks like Apache Spark or Hadoop improve scalability and data handling by processing large datasets across multiple nodes in a cluster. Cloud computing platforms, such as AWS, Google Cloud, and Azure, provide on-demand computational resources and storage, enabling scalability according to workload demands. Additionally, designing the HF-TIM method with a modular architecture allows for independent scaling and optimization of each component, ensuring efficient resource utilization and easier maintenance. By adopting these strategies, the HF-TIM method can be effectively deployed, offering robust and scalable solutions for multimodal fake news detection.

VII. CONCLUSION AND FUTURE WORK

This paper introduces HF-TIM, a hybrid data fusion approach that combines the early fusion of multimodal data with the late fusion of textual and image modalities. The HF-TIM method is designed to improve feature fusion, by understanding unimodal features on the one hand and their interrelationships on the other hand, contributing to the development of a robust model with high detection accuracy. By examining various fake news categories, we demonstrate how HF-TIM enhances classification accuracy by leveraging the presence or absence of relationships between text and images. Our approach integrates multimodal and unimodal features using a neural network-based meta-learning classifier, effectively combining features through several dense layers and classifying them with a Softmax classifier. The scalability of the HF-TIM approach is a crucial factor for its practical application in real-world scenarios. The method is designed to handle large-scale data by leveraging advanced deep-learning techniques and efficient data processing pipelines. The use of BERT and VGG-19 models ensures that the method can process vast amounts of textual and visual data effectively. This hybrid fusion approach combines the strengths of early and late fusion techniques, providing a robust and scalable solution for multimodal fake news detection.

The applicability of the HF-TIM method extends across various domains. In healthcare, integrating textual data such as patient records and medical literature with visual data like medical images can enhance diagnostic accuracy and patient care. In autonomous driving systems, the HF-TIM method can enhance the perception and decision-making capabilities of vehicles by fusing data from various sensors, such as cameras, LiDAR, and radar. In security and surveillance, combining video footage with textual data like incident reports and sensor readings can enhance threat detection and response strategies. The entertainment industry can benefit by integrating textual scripts with visual storyboard images during the production process, resulting in more cohesive and engaging media content. Additionally, in social media analysis, the HF-TIM method can enhance sentiment analysis, trend detection, and user behavior analysis by processing both text and images, providing a holistic view of social media content.

Despite its strengths, this study has limitations, particularly in the area of hyperparameter tuning. Future research should incorporate advanced methods such as Grid search, random search, and Bayesian optimization to identify optimal hyperparameters. Additionally, we intend to explore other fusion techniques, such as dynamic fusion methods and graph-based fusion approaches, which may offer superior integration of multimodal data by capturing more complex dependencies and interactions, thereby increasing accuracy and robustness. Expanding the diversity of datasets for training and evaluation is also essential to enhance the model's generalizability and performance across various scenarios and domains. To further validate the generalizability of our approach, we plan to apply HF-TIM to additional multimodal datasets. We also aim to extend our HF-TIM approach to include additional modalities such as audio, video, and metadata. By integrating audio and video data, we can capture dynamic and temporal information, further enhancing the detection capabilities of our model.

REFERENCES

[1] S. A. Alameri and M. Mohd, "Comparison of fake news detection using machine learning and deep learning techniques," in *Proc. 3rd Int. Cyber Resilience Conf. (CRC)*, Jan. 2021, pp. 1–6.

[2] M. R. Islam, S. Liu, X. Wang, and G. Xu, "Deep learning for misinformation detection on online social networks: A survey and new perspectives," *Social Netw. Anal. Mining*, vol. 10, no. 1, p. 82, Dec. 2020, doi: 10.1007/s13278-020-00696-x.

[3] S. A. Alkhodair, S. H. H. Ding, B. C. M. Fung, and J. Liu, "Detecting breaking news rumors of emerging topics in social media," *Inf. Process. Manage.*, vol. 57, no. 2, Mar. 2020, Art. no. 102018.

[4] F. F. Said, R. S. Somasuntharam, M. R. Yaakub, and T. Sarmidi, "Impact of Google searches and social media on digital assets' volatility," *Humanities Social Sci. Commun.*, vol. 10, no. 1, pp. 1–17, Nov. 2023.

[5] J. A. Khan, T. Ullah, A. A. Khan, A. Yasin, M. A. Akbar, and K. Aurangzeb, "Can end-user feedback in social media be trusted for software evolution: Exploring and analyzing fake reviews," *Concurrency Comput., Pract. Exper.*, vol. 36, no. 10, p. e7990, May 2024.

[6] G. Arya, M. K. Hasan, A. Bagwari, N. Safie, S. Islam, F. R. A. Ahmed, A. De, M. A. Khan, and T. M. Ghazal, "Multimodal hate speech detection in memes using contrastive language-image pre-training," *IEEE Access*, vol. 12, pp. 22359–22375, 2024.

[7] K. Nakamura, S. Levy, and W. Wang, "R/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," 2019, *arXiv:1911.03854*.

[8] I. Gallo, A. Calefati, and S. Nawaz, "Multimodal classification fusion in real-world scenarios," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 5, Nov. 2017, pp. 36–41.

[9] A. Azri, C. Favre, N. Harbi, J. Darmont, and C. Nous, "Rumor classification through a multimodal fusion framework and ensemble learning," *Inf. Syst. Frontiers*, vol. 25, no. 5, pp. 1795–1810, Oct. 2023.

[10] Y. Wang, Y. Gu, Y. Yin, Y. Han, H. Zhang, S. Wang, C. Li, and D. Quan, "Multimodal transformer augmented fusion for speech emotion recognition," *Frontiers Neurorobotics*, vol. 17, May 2023, Art. no. 1181598.

[11] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, "Deep multimodal fusion for semantic image segmentation: A survey," *Image Vis. Comput.*, vol. 105, Jan. 2021, Art. no. 104042.

[12] M. F. Mridha, A. J. Keya, Md. A. Hamid, M. M. Monowar, and M. S. Rahman, "A comprehensive review on fake news detection with deep learning," *IEEE Access*, vol. 9, pp. 156151–156170, 2021.

[13] M.-Y. Cao, S. Zainudin, and K. M. Daud, "Protein features fusion using attributed network embedding for predicting protein–protein interaction," *BMC Genomics*, vol. 25, no. 1, p. 466, May 2024.

[14] A. AlMohimeed, R. M. A. Saad, S. Mostafa, N. M. El-Rashidy, S. Farrag, A. Gaballah, M. A. Elaziz, S. El-Sappagh, and H. Saleh, "Explainable artificial intelligence of multi-level stacking ensemble for detection of Alzheimer's disease based on particle swarm optimization and the sub-scores of cognitive biomarkers," *IEEE Access*, vol. 11, pp. 123173–123193, 2023.

[15] A. Rahman, M. E. H. Chowdhury, A. Khandakar, S. Kiranyaz, K. S. Zaman, M. B. I. Reaz, M. T. Islam, M. Ezeddin, and M. A. Kadir, "Multimodal EEG and keystroke dynamics based biometric system using machine learning algorithms," *IEEE Access*, vol. 9, pp. 94625–94643, 2021.

[16] M. M. Aziz, A. A. Bakar, and M. R. Yaakub, "CoreNLP dependency parsing and pattern identification for enhanced opinion mining in aspect-based sentiment analysis," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 36, no. 4, Apr. 2024, Art. no. 102035.

[17] M. I. Marwat, J. A. Khan, D. M. D. Alshehri, M. A. Ali, H. Ali, and M. Assam, "Sentiment analysis of product reviews to identify deceptive rating information in social media: A sentideceptive approach," *KSII Trans. Internet Inf. Syst. (TIIS)*, vol. 16, no. 3, pp. 830–860, May 2022.

[18] S. Zhang, B. Li, and C. Yin, "Cross-modal sentiment sensing with visual-augmented representation and diverse decision fusion," *Sensors*, vol. 22, no. 1, p. 74, Dec. 2021.

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

[20] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," *Multimedia Tools Appl.*, vol. 80, no. 2, pp. 2887–2905, Jan. 2021.

[21] 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.

[22] J. Wang, H. Mao, and H. Li, "FMFN: Fine-grained multimodal fusion networks for fake news detection," *Appl. Sci.*, vol. 12, no. 3, p. 1093, Jan. 2022.

[23] S. D. A. Rihan, M. Anbar, and B. A. Alabsi, "Meta-learner-based approach for detecting attacks on Internet of Things networks," *Sensors*, vol. 23, no. 19, p. 8191, Sep. 2023.

[24] Y. Wu, P. Zhan, Y. Zhang, L. Wang, and Z. Xu, "Multimodal fusion with co-attention networks for fake news detection," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, vol. 2021. Switzerland: SENSORS/MDPI, 2021, pp. 2560–2569.

[25] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: A review," *Briefings Bioinf.*, vol. 23, no. 2, Mar. 2022, Art. no. bbab569.

[26] N. J. Shoumy, L.-M. Ang, K. P. Seng, D. M. M. Rahaman, and T. Zia, "Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals," *J. Netw. Comput. Appl.*, vol. 149, Jan. 2020, Art. no. 102447.

[27] B. Wang, Y. Feng, X.-C. Xiong, Y.-H. Wang, and B.-H. Qiang, "Multi-modal transformer using two-level visual features for fake news detection," *Int. J. Speech Technol.*, vol. 53, no. 9, pp. 10429–10443, May 2023.

[28] S. Kalra, C. H. S. Kumar, Y. Sharma, and G. S. Chauhan, "Multimodal fake news detection on fakeddit dataset using transformer-based architectures," in *Proc. 4th Int. Conf.*, 2022, pp. 281–292.

[29] P. Liu, W. Qian, D. Xu, B. Ren, and J. Cao, "Multi-modal fake news detection via bridging the gap between modals," *Entropy*, vol. 25, no. 4, p. 614, Apr. 2023.

[30] C. Che, H. Wang, X. Ni, and R. Lin, "Hybrid multimodal fusion with deep learning for rolling bearing fault diagnosis," *Measurement*, vol. 173, Mar. 2021, Art. no. 108655.

[31] M. Person, M. Jensen, A. O. Smith, and H. Gutierrez, "Multimodal fusion object detection system for autonomous vehicles," *J. Dyn. Syst., Meas., Control*, vol. 141, no. 7, pp. 1–24, Jul. 2019.

[32] J. Njoku, A. Caliwag, W. Lim, S. Kim, H.-J. Hwang, and J.-W. Jeong, "Deep learning based data fusion methods for multimodal emotion recognition," *J. Korean Inst. Commun. Inf. Sci.*, vol. 47, no. 1, pp. 79–87, Jan. 2022.

[33] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *npj Digit. Med.*, vol. 3, no. 1, pp. 1–9, Oct. 2020.

[34] G. Chandrasekaran, T. N. Nguyen, and J. Hemanth D., "Multimodal sentimental analysis for social media applications: A comprehensive review," *WIREs Data Mining Knowl. Discovery*, vol. 11, no. 5, p. e1415, Sep. 2021.

[35] V. A. Torres Caceres, K. Duffaut, A. Yazidi, F. Westad, and Y. B. Johansen, "Automated well log depth matching: Late fusion multimodal deep learning," *Geophys. Prospecting*, vol. 72, no. 1, pp. 155–182, Jan. 2024.

[36] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.

[37] B. Ding, T. Zhang, G. Liu, L. Kong, and Y. Geng, "Late fusion for acoustic scene classification using swarm intelligence," *Appl. Acoust.*, vol. 192, Apr. 2022, Art. no. 108698.

[38] T. L. V. Zyl, "Late meta-learning fusion using representation learning for time series forecasting," in *Proc. 26th Int. Conf. Inf. Fusion (FUSION)*, Jun. 2023, pp. 1–8.

[39] I. K. S. Al-Tameemi, M.-R. Feizi-Derakhshi, S. Pashazadeh, and M. J. Asadpour, "A comprehensive review of visual-textual sentiment analysis from social media networks," 2022, *arXiv:2207.02160*.

[40] 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.

[41] 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.

[42] 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)*, Jul. 2020, pp. 1–8.

[43] A. Giachanou, G. Zhang, and P. Rosso, "Multimodal fake news detection with textual, visual and semantic information," in *Proc. 23rd Int. Conf.*, 2020, pp. 30–38.

[44] B. Palani, S. Elango, and V. Viswanathan K, "CB-fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and BERT," *Multimedia Tools Appl.*, vol. 81, no. 4, pp. 5587–5620, Feb. 2022.

[45] S. K. Uppada and P. Patel, "An image and text-based multimodal model for detecting fake news in OSN's," *J. Intell. Inf. Syst.*, vol. 61, no. 2, pp. 367–393, Oct. 2023.

[46] 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.

[47] 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.

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

[49] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[50] G. Wang, C. Li, W. Wang, Y. Zhang, D. Shen, X. Zhang, R. Henao, and L. Carin, "Joint embedding of words and labels for text classification," 2018, *arXiv:1805.04174*.

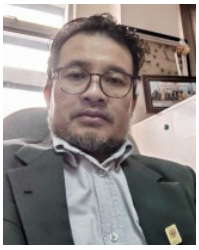
[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[53] 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.

[54] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[55] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, Sep. 2022.

SUHAIB KH. HAMED received the master's degree from the Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology (FTSM), Universiti Kebangsaan Malaysia (UKM), in 2016. He is currently pursuing the Ph.D. degree with the Center for Software Technology and Management (SOFTAM), FTSM, UKM. In administration, he was the Head of the IT Department, Baghdad Electricity Distribution, from 2019 to 2020. He is a Researcher with the Sentiment Analysis Laboratory, CAIT. His expertise is deep learning, supervised learning, classification, feature extraction, data fusion, natural language processing, image processing, sentiment analysis, and ontology.



MOHD JUZAIDDIN AB AZIZ started his career as a Tutor, in 1997. He has been a Professor of computer science with the Faculty of Technology and Information Science, Universiti Kebangsaan Malaysia, since 2017. Throughout the career, he has contributed ideas and experts in various fields, such as writing papers to obtain Malaysia Super Corridor (MSC) status for UKM, a coordinator of the Bachelor of Computer Science Program (Executive), the Head of the Computer Science Program, the Head of the Master's Program, the Deputy Dean (Academic), the Deputy Dean (Undergraduate), the Director of the Information Technology Center, and the Chief Information Officer of UKM. Despite taking a break from the administration with UKM, since May 2021, he is still being appointed as a Subject Matter Expert with the Ministry of Higher Education Malaysia, the ICT Technical Committee, and the ICT Steering Committee of the Ministry. Research and teaching on language processing and computer science is a field that is very close to him. To date, he has successfully produced more than 15 Doctor of Philosophy students and 30 master's students with more than 90 publications. He received the UKM Outstanding Staff Award three times, in 2003, 2013, and 2021.

MOHD RIDZWAN YAAKUB received the Ph.D. degree from Queensland University of Technology (QUT), Australia. He is currently a Senior Lecturer with the Center for Artificial Intelligence and Technology (CAIT), Faculty of Information Science and Technology (FTSM), National University of Malaysia (UKM). He is the Head Researcher of the Sentiment Analysis Laboratory, CAIT. In administration, he is the Deputy Dean of Industry and Community Partnerships Affairs. His expertise is sentiment analysis/opinion mining, feature selection, feature extraction, ontology, and data mining.