I.J. Modern Education and Computer Science, **2026, 2, 132-146**
Published Online on April 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2026.02.08

A Hybrid CNN-Transformer Model for Multimodal Fake News Detection Using Feature Fusion

Vineela Krishna. Suri*
Department of CSE, Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh, India
Email: vineela.suri@gmail.com
ORCID iD: https://orcid.org/0000-0002-4563-1979
*Corresponding Author

Prasad. GVSNRV
Department of CSE, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India
Email: gutta.prasad1@gmail.com
ORCID iD: https://orcid.org/0000-0003-0679-7324

Received: 01 April, 2025; Revised: 06 September, 2025; Accepted: 31 January, 2026; Published: 08 April, 2026

Abstract: The widespread distribution of fake news poses a critical societal challenge by influencing public opinion and shaping political discourse. Addressing this problem requires models that can capture multimodal cues beyond text alone. This work proposes a lightweight Multimodal Cross-attention Fusion–based Fake News Detection (MCAF-FND) model which combines textual and visual features through cross-attention strategy. The study evaluates MCAF-FND on the Fakeddit benchmark, a large-scale dataset comprising 682,996 multimodal samples collected from social media. Textual features are extracted using DistilBERT, while spatially aware image representations are derived from VGG-19 convolutional layers. The cross-attention module enables semantic alignment between text tokens and image patches, modeling inter-modal dependencies more effectively than conventional fusion strategies. The fused representation is classified using a Multilayer Perceptron(MLP) with softmax, ensuring contributions from both modalities. Experimental results demonstrate that MCAF-FND consistently outperforms unimodal baselines and traditional fusion methods, achieving 93.2% accuracy with strong precision, recall, and F1-score. Cross-attention based visualizations illustrate how the model aligns textual cues with salient visual regions, enhancing interpretability. By combining computational efficiency with robust multimodal reasoning, the proposed approach provides a reliable and extensible solution for automated fake news detection.

Index Terms: Multimodal Data, Fake News Detection, Convolution Neural Networks, Multimodal Fusion, Transformers

1. Introduction

The extensive distribution of fake news has become a critical social challenge, influencing public opinion, political debate, and social cohesion. Social media platforms have amplified this phenomenon, allowing disinformation to reach millions of people in minutes by exploiting multimodal content such as text, images, videos, and audio. Traditional fact checking approaches, although essential, cannot keep up with the massive volume and velocity of online misinformation [1]. This highlights the urgent need for automated detection methods capable of identifying fake news in a reliable and accurate way. An illustrative multimodal example is shown in Figure 1, where perceived trustworthiness is increased through the combination of a textual claim with the corresponding image.

Early research mainly focused on textual analysis, where linguistic cues and semantic inconsistencies were used to classify news articles. While effective in certain contexts, text-only methods struggle to detect disinformation enriched with non-textual elements. Recent studies demonstrate that multimodal approaches, which jointly analyze textual and visual signals, offer significant improvements in accuracy and robustness [2,3,4]. For example, integrating image semantics with textual cues enables models to capture cross-modal inconsistencies—such as mismatched captions and manipulated visuals—that would otherwise be overlooked.

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

Several multimodal datasets have been developed to support this line of research. The LIAR dataset [7], derived from Politifact fact-checks, focuses on short political statements, while more recent corpora such as FANG [9] and Fakeddit [21] provide large-scale multimodal benchmarks combining text, images, and metadata. These resources allow the development of deep learning architectures that leverage pretrained language models, convolutional neural networks, and attention mechanisms to capture semantic and contextual dependencies across modalities. In particular, CNN and transformer-based architectures have shown strong performance in modeling both textual semantics and visual features [5,6,10,11].

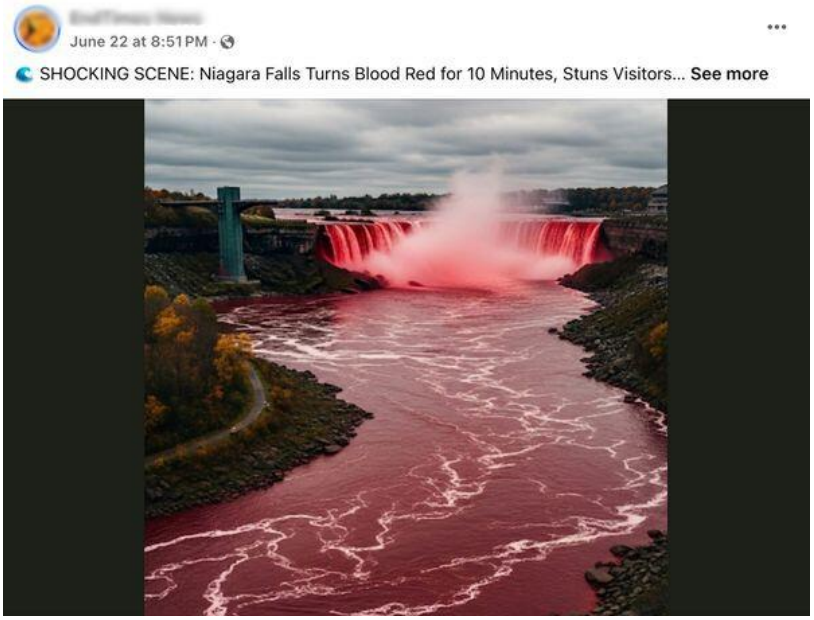


Fig. 1. Multimodal misinformation posted on social media

In conjunction of computer science, data science, and social sciences, multimodal fake news detection has become a crucial research area. By integrating diverse sources of information, researchers aim to design reliable and explainable models that can mitigate the impact of online misinformation. This study contributes to this effort by developing a lightweight yet effective cross-attention fusion framework that aligns textual and visual cues for robust fake news classification.

1.1 Objectives of the proposed work

- Capture semantic and contextual embeddings from textual content using a pre-trained DistilBERT-based textual encoder, facilitating a nuanced understanding of the language.
- Extract visual features using a VGG-19-based visual encoder while preserving spatial structures for effective multimodal fusion.
- Fuse textual and visual cues using a fusion mechanism based on cross-attention that enables the model to capture semantic correspondence between the two modalities.
- Build a unified multimodal pipeline that classifies news as real or fake using an MLP classifier with softmax.
- Evaluate the performance of the model through attention-based heatmaps and various standard metrics.

1.2 Major Contributions

- **Lightweight textual encoder:** DistilBERT is a lighter version of BERT, retains 95% of BERT's performance with fewer parameters and reduced training time, making it efficient for processing textual content.
- **Preservation of spatial structures:** The visual encoder extracts patch-level features using the convolutional layers of VGG-19, preserving spatial structure by maintaining a 7x7 grid and allowing attention to focus on salient regions of the image.
- **Cross-attention based fusion:** This mechanism allows the model to capture semantic correspondence between textual and visual modalities and also learn complex, inter-modal dependencies between them.
- **Unified and computationally efficient multimodal pipeline:** Proposed a lightweight and computationally efficient pipeline, suitable for platforms with limited GPU resources. This pipeline combines textual and visual encoders with cross-attention fusion and MLP classification, providing an efficient end-to-end framework for fake news detection.

- **Evaluation:** The model is evaluated on the Fakeddit dataset, a large and challenging benchmark customized for fake news detection. This evaluation demonstrates the model's ability to comprehend and generalize across manipulated text-image pairs.

The remaining sections are organized as follows: Section 2 reviews existing methodologies to identify false news. Section 3 explores the dataset and elaborates on the preprocessing steps applied. Section 4 presents the overall architecture and detailed operational workflow of the proposed approach for detecting misinformation. Section 5 provides a comprehensive review of the results obtained using various evaluation metrics. Section 6 summarizes the paper and provides directions for future research.

2. Related Works

Most of the existing methods for detecting fake news relied on unimodal data or hybrid methods that merge different types of data. These approaches ignore the effectiveness of modeling different modalities and similarities between multimodal data. An expanding and multidisciplinary subject is research on recognizing fake news in multimodal datasets, which encompasses both textual and non-textual information, such as videos, images, sounds, etc.

In [12], the authors introduced a novel transfer learning framework called MTL-rtFND, which incorporates multimodal transfer learning techniques, which relied on content for real-time fake news detection. A data preprocessing block, an MFEB, and a CB make up the main parts of this system. The MFEB makes use of a pre-trained multimodal mode, which combines visual and verbal representations, to effectively collect information dependent on context. Using larger-scale multimodal datasets to train this pre-trained model allows for greater contextual information extraction. In order to classify news stories as valid or fraudulent in real time using a deep neural network, the CB receives the multimodal feature vectors produced by the MFEB. The proposed MTL-rtFND model, when validated against some real-world datasets, demonstrated its efficacy in detecting false news. Their experimental findings show some significant increase in accuracy by 6.34 percent in various text embedding strategies compared to other baseline methodologies. Experimental results show that multimodal transfer learning can improve real-time identification of misinformation, which addresses the harmful effects of fake news on the social media platform.

An emotion-driven method to detect false news by capturing emotional signals from multimodal data was developed by [13]. From the input text, the model extracts emotional information, including emotion lexicons and embeddings. They introduced a Vision Transformer (ViT) based architecture with an attention mechanism at multiple levels, improving the classification accuracy by filtering out unnecessary data from pictures and focusing on meaningful visual cues. The emotional features extracted from both text and image content were merged to form a joint representation using a multilayer fusion network. The model was evaluated on various unimodal and multimodal corpora, including Risdal, Pontes, Jruvika, and Fakeddit.

In [14], the authors proposed Adaptive Knowledge-Aware Fake News Detection, or AKA-Fake. In order to restore the most informative knowledge facts, for each news article, the model learns a compact knowledge sub graph from a specific of entities and their contextual neighbors in the knowledge graph using reinforcement learning. They suggested a new heterogeneous graph learning module that uses modality-attentive pooling and topology refinement to obtain reliable correlations between different modes. Three prominent datasets thoroughly assessed the concept and experimental findings demonstrate the dominance of AKA- Fake over other detection models.

A multi-level transformer-based attention framework was designed by [15], to detect false news posted online. They employed a ViT with multi-head self-attention and a multilayer perceptron to retrieve visual features, a textual encoder that uses bi-directional encoder representation from transformers to encode raw text sequences. Finally, a Joint attention-based learning model that uses a visual semantic attention block that retrieves image features and a self-attention block that eliminates redundancy from the data. They assessed the performance using various evaluation metrics and conducted experiments on four prominent datasets: the Jruvika, Twitter, Ponnes and Risdal fake news datasets.

In [16], the authors introduced a new Multimodal fusion model (TLFND), to detect false news, which applies a distance method based on three-level features. TLFND consists of four main components: A multimodal integrated recognition module, a feature matching score module, an image extraction and fusion module, and a two-level text feature extraction module. This methodology precisely integrates images (multi-image fusion) with two levels of textual information (headline and body) found in news stories. Importantly, they were the first to use the Chebyshev distance measure to determine how well these three modalities matched. They also calculated the loss functions for four parts of the model using an adaptive evolutionary method. Their extensive testing on three benchmark datasets confirms the efficacy of the suggested model. The experimental results obtained when evaluated against the PolitiFact dataset found an increase in the F1 score of 6.6%, an increase of 2.9% on the GossipCop dataset and an increase of 2.3% on the Twitter dataset.

The Bootstrapping-based false news detection model (FNDB) was introduced by [17]. Initially, the news articles are analyzed using a multimodal feature extractor that generates multimodal feature representations. Linguistic and visual elements were extracted using XLNet and VGG-19-based feature extractors, respectively. The feature fusion layer employs cross-modal attention mechanism to merge these features in order to use different modal feature

representations to complement each other's information. After identifying the input material using these fused traits, the detector completes the fake news detection process. This architecture enables FNDB to efficiently fuse data from different modalities and extract most of the information from news material. Verification studies conducted against two popular publicly available corpus, Weibo and Gossipcop, indicate that FNDB outperforms the other state-of-the-art approaches by 4.4% and 0.6%, respectively.

In [18], the authors proposed a multimodal progressive fusion network (MPFN) for determining the authenticity of news. They introduced BERT-based feature extractor, transformer-based visual feature extractor, and progressive feature fusion process. They designed MLP mixer, as a fusion module for fusing of features between modalities at a fine granularity. The model is accessed on Weibo and Twitter datasets. It has shown better performance over the other advanced art and baseline models. The authors estimated the performance of this model using various evaluation metrics. The model exhibited an accuracy of 83.8% on Weibo and 83.3% on Twitter data sources.

Studies emphasize the importance of multimodal datasets including textual, image and other non-textual information in identifying fake news due to its fast propagation across various social media platforms. Existing models based on multimodal data were unable to effectively identify the distinctive characteristics between the image and text modalities.

3. Dataset Description

Created by Nakamura [21], Fakeddit is a multimodal dataset consisting of more than one millon samples containing text, images, comment and other metadata. It is one of the largest open-access sources of multimodal fake news, allowing researchers to explore multiple options between different modalities for misinformation detection. Fakeddit is sourced from Reddit, a social network that allows registered users to post content which is then voted and commented by other users in the community, collected from 22 different subreddits. The dataset is openly accessible through its official website [19] and can also be downloaded from the GitHub repository [20], which provides both data and tools required for use.

A substantial amount of conventional research on fake news and datasets, including CREDBANK, LIAR, FakeNewsCorpus focus only on text characteristics. In addition, most of the datasets have a relatively small size and variation. Fakeddit consists of more than 1 million samples out of which 682,996 are multimodal samples consisting of both text and images. Table 1 summarizes the distribution of samples in Fakeddit.

Table 1. Distribution of Samples in Fakeddit Dataset

Distribution of Samples	Count
Total	1,063,106
Fake	628,501
True	527,049
Multimodal	682,996
Training set	878,218
Validation set	92,444

Many of the conventional datasets that are available typically categorize their data binarily as true or false. Each sample in the Fakeddit dataset consists of 3 labels as shown in Table 2, which supports 2-way, 3-way and 6-way classification. Figure 2 shows examples from the dataset that depicts 6-way classification labels.

Table 2. Types of Classification in Fakeddit

Classification Type	Class Labels
2-way	• Fake • Real
3-way	• Real • Satire • Fake
6-way	• True • Satire • Misleading • Manipulated • False Connection • Imposter

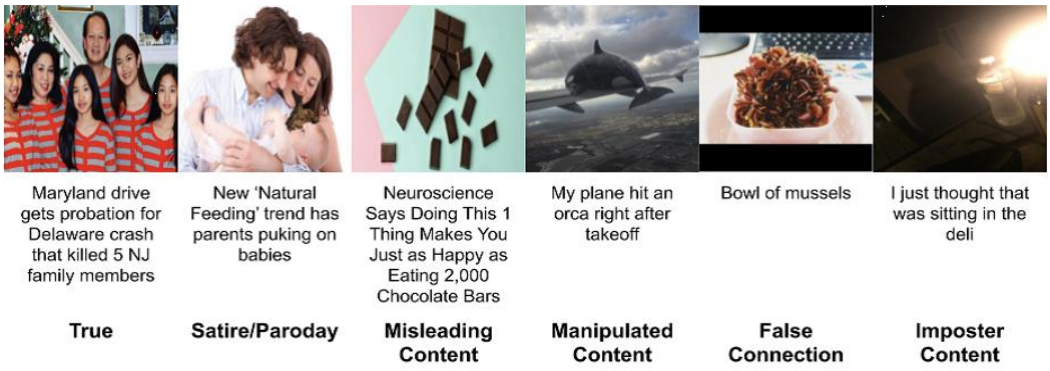


Fig. 2. An example 6-way classification labels

4. Proposed Model

This section provides a detailed description about the proposed Multimodal Cross-attention Fusion based Fake news Detection (MCAF-FND) model that accesses the credibility of multimodal news consisting of both textual and visual content. Although the dataset originally consists of 682,481 multimodal samples, many image URLs were non-functional and unavailable for download at the time of data collection. Preprocessing was then performed separately on both text and image data.

4.1 Preprocessing

- Text Preprocessing: Removed URLs, special characters, and emojis, as DistilBERT cannot effectively process emojis, URLs and special characters may introduce noise into the model. Rows with missing data values were discarded.
- Image Preprocessing: The dataset consists of images associated with news articles. Each image is converted into RGB format to ensure compatibility, and then scaled down to a fixed size of 224x224 pixels and then normalized. The preprocessed images were then passed through VGG-19 and deep features were extracted.

After preprocessing, only 124,530 labeled samples, each consisting of title, image, and metadata, were available for experimental analysis. The proposed MCAF-FND model, as shown in Figure 3, is mainly composed of 4 modules: a DistilBERT-based textual encoder, a VGG-19-based visual encoder, a Cross-Attention Fusion module, and a Multi-Layer Perceptron(MLP) Classifier.

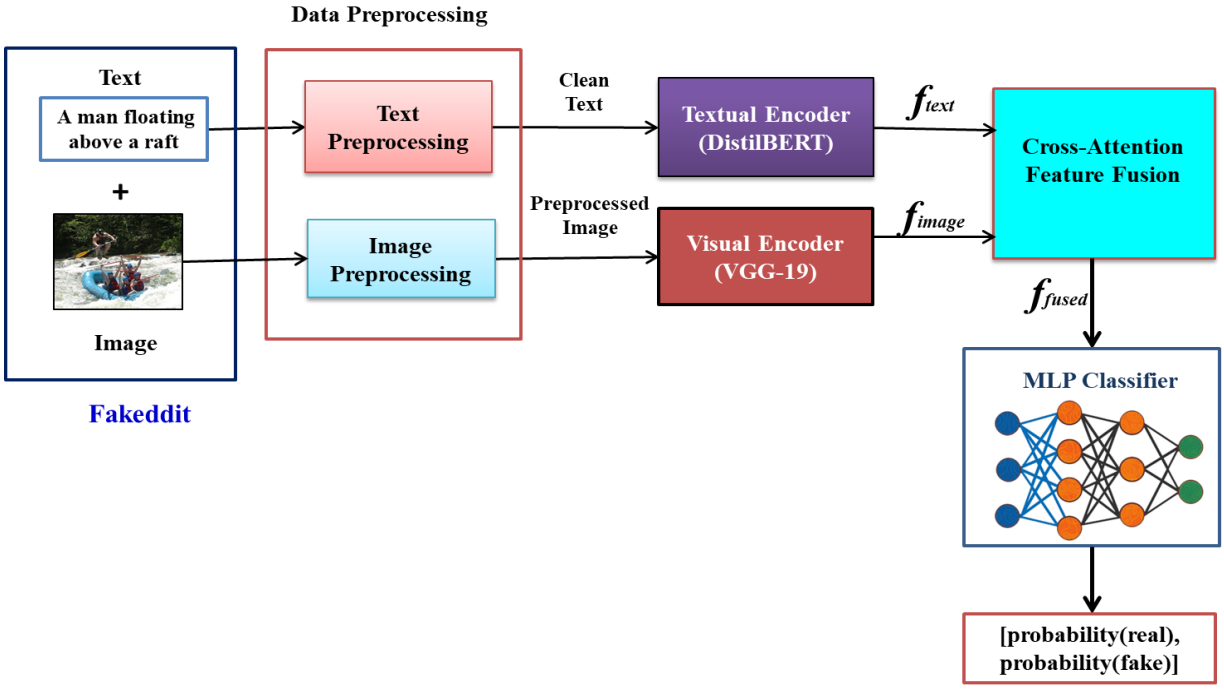


Fig. 3. Proposed MCAF-FND Classifier

4.2 Textual Encoder

Language is the primary representation of news events and provides vital indicators for determining the veracity of news. Most existing methods commonly use recurrent neural networks to extract a contextual representation of textual content and its superficial properties. However, these methods only recover a limited amount of factual information, making it hard to detect the semantic elements of fake news. To extract both contextual and semantic features from text, we use DistilBERT, a lightweight version of BERT that is designed to be computationally effective and significantly faster than BERT. As depicted in Figure 4, DistilBERT consists of multiple transformer layers with a feed-forward neural network and multi-head self-attention mechanism present in each layer. The textual encoder uses a pre-trained DistilBERT model to turn text into numerical representations called contextual embeddings that hold semantic information about the text. [CLS] token attends to all the other tokens present in the text and contains global sematic information. By the time the input text reaches the last hidden layer, the [CLS] vector contains the global context aggregation of the entire text. Figure 5 illustrates the process of extracting [CLS] vector from text using DistilBERT and Algorithm 1 presents a comprehensive step-by-step procedure.

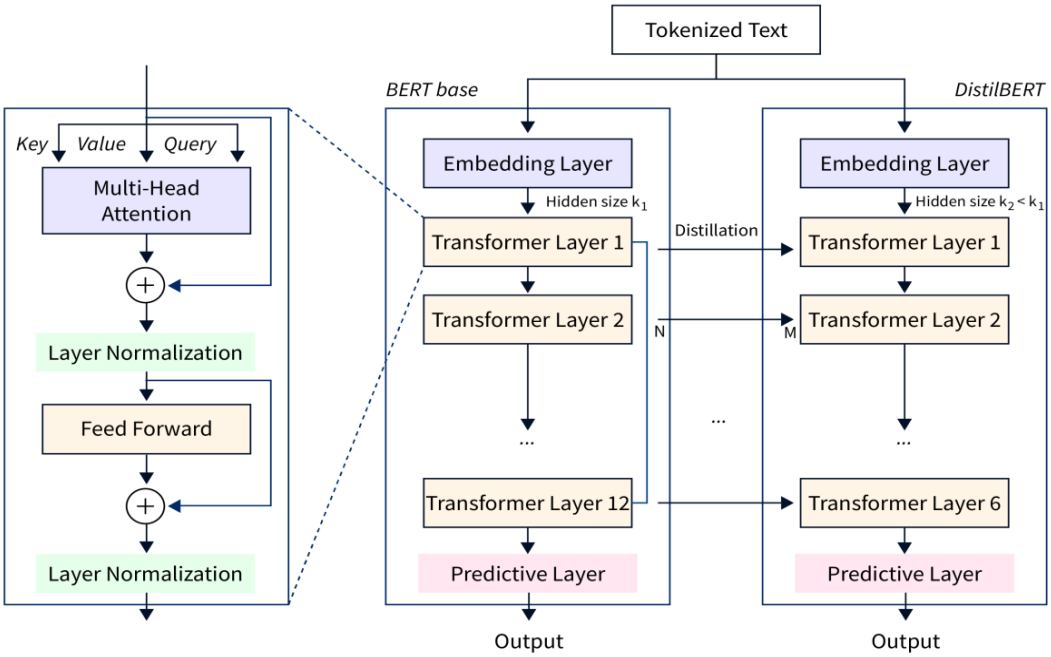


Fig. 4. DistilBERT architecture

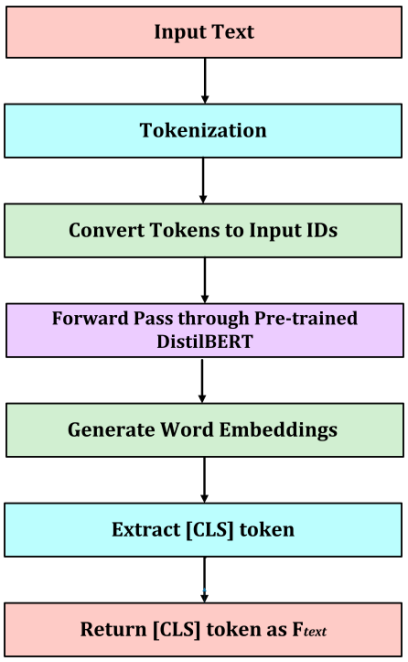


Fig. 5. Textual Encoder

Algorithm 1: Textual Encoder

Input: Text T

Output: Text feature vector F_{text}

1: Load Pre-trained DistilBERT model and its tokenizer

2: Tokenize the input text using tokenizer.tokenize method as shown in Equation 1

$$X = Tokenizer.tokenize(T) \quad (1)$$

where X = x_1, x_2, ..., x_n is the sequence of tokens generated

3: Generate input IDs and attention mask for the input tokens using Equation 2

$$(\{x_{ID1}, x_{ID2}, \dots, x_{IDn}\}, attn_mask) = Tokenizer.encode_plus(X) \quad (2)$$

where x_{IDi} is the sequence of numerical IDs generated and *attn_mask* is the attention mask to identify which tokens are real(1) and which of them are padding(0).

4: Forward pass tokenized text through DistilBERT using Equation 3

$$\{e_1, e_2, \dots, e_n\} = DistilBERT(\{x_{ID1}, x_{ID2}, \dots, x_{IDn}\}, attn_mask) \quad (3)$$

where output $E = e_i$ represents token x_i's contextual embedding. The DistilBERT model generates a set of contextual embeddings $\{e_1, e_2, \dots, e_n\}$. Each e_i is a high-dimensional vector that captures the contextual representation of token x_i in the full input sequence.

5: Extract [CLS] token representation using Equation 4

$$F_{text} = E.last_hidden_state[:,0,:] \quad (4)$$

6: Return text feature vector F_{text}

4.3 Visual Encoder

The visual encoder is designed to extract visual features using VGG-19, a convolutional neural network, entails running an image through the network and extracting features from one of its intermediary layers. As shown in Figure 6, the CNN model, VGG-19, consists of multiple convolutional, pooling and fully connected layers. The visual encoder extracts spatially-aware image features using only convolutional layers of VGG-19 with no classifier to extract features, applies average pooling to ensure consistency across input and then reshapes the features to patch tokens. Instead of obtaining features from fully connected layers for classification of the original VGG-19 encoder, our encoder extracts intermediate feature maps from the convolutional backbone. Unlike using a single 4096-dimensional global vector obtained from the second-last convolutional layer (FC2), the resulting patch sequences captures spatial structure and enables the attention mechanism to focus on more salient regions of the image (e.g., faces, logos, or text). Algorithm 2 outlines the sequence of steps in detail.

Algorithm 2: Visual Encoder

Input: Image I

Output: Image feature vector F_{image}

1: Load pre-trained VGG-19 convolutional layers

2: Resize the image I to 224 x 224 pixels and normalize pixel values to ImageNet mean and variance

3: Pass the image I through the convolutional layers of VGG-19 and extract features using Equation 5. VGG yields 7x7 spatial map with 512 channels.

$$F_{conv} = VGG19_{features}(I), \qquad F_{conv} \in \mathbb{R}^{512\ X\ 7\ X\ 7} \quad (5)$$

4: To ensure consistency across inputs apply average pooling using Equation 6

$$F_{pool} = Avgpool(F_{conv}) \quad (6)$$

5: Reshape the features into a sequence of patches using Equation 7

$$F_{seq} = Reshape(F_{pool}), \qquad F_{pool} \in \mathbb{R}^{49\ X\ 512} \quad (7)$$

where each of the 49 tokens corresponds to one spacial cell in the 7x7 grid, each token is a 512-dimension descriptor.

6: Return spatial-aware image feature sequence, as shown in Equation 8.

$$F_{image} = F_{seq}$$

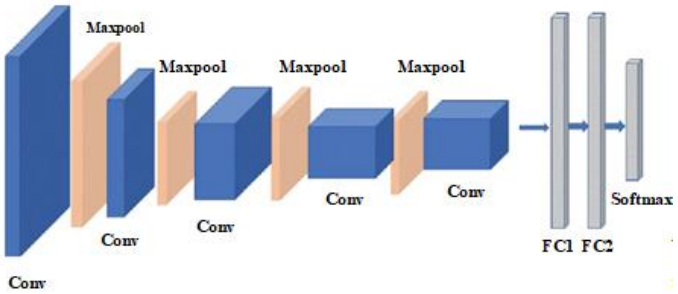


Fig. 6. VGG-19 architecture

4.4. Cross-Attention Fusion

Fusion enables a model to enhance complementary information from different modalities, resulting in better performance. Among various methods available for multimodal fusion, this study utilizes cross-attention based fusion method. Cross-Attention Fusion involves merging features from different modalities in a way that allows each modality to focus on the relevant parts of the other modality. It allows the model to capture semantic correspondence between different modalities and also learns complex, intermodal dependencies between them.

Cross-Attention Fusion module enables text features to interact with relevant parts of the image. For example, the word "explosion" refers to fiery parts of the image. Here, queries (Q) were formed from text features, keys and values are obtained from the image vectors and then, by computing similarity scores between query Q and each key K, this module learns to map descriptive words to relevant image segments. These attention scores indicate relevance between each text token and every image patch, enabling rich multimodal fusion. Softmax transforms these attention scores into attention weights and normalizes them to probability distribution. The Algorithm 3 offers a detailed overview of the process of cross-attention fusion.

Algorithm 3: Cross-Attention Fusion

Input: Text features F_{text} , Image features F_{image}

Output: Cross-fused vector F_{fused}

1: Project text features F_{text} to Queries Q using Equation 9

$$Q = W_Q \cdot F_{text},\ W_Q \in \mathbb{R}^{d_{text}\ X\ d_{attn}} \tag{9}$$

2: Project image features F_{image} to Keys K and Values V using Equation 10

$$K = W_k \cdot F_{image},$$

$$V = W_v \cdot F_{image},\quad W_k, W_v \in \mathbb{R}^{d_{image}\ X\ d_{attn}} \tag{10}$$

3: Compute attention scores with scaled dot-product attention and normalize the values using Softmax (Equation 11):

$$\alpha = \text{softmax}\left(\frac{QK^T}{\sqrt{d_{attn}}}\right) \tag{11}$$

where attention size d_{attn}= 256, d_{text}=768D, d_{image}= (49 X 512)D

4: Fuse representations by computing weighted-sum of attention scores using Equation 12

$$F_{fused} = \alpha \cdot V \tag{12}$$

5: Return fused feature vector F_{fused}

4.5. Multi-Layer Perceptron Classifier

The resulting fused vector F_{fused} was passed through an MLP classifier with two hidden layers of sizes 1024 and 256, respectively. LeakyRelu activation function with leaky slope α = 0.01 was used after each hidden layer to address the vanishing gradient problem and a dropout value of 0.4 is applied between layers as a regularization method to avoid overfitting. Additionally, the model is optimized using Adam optimizer with a learning rate of 0.0001, β1 = 0.9, and β2 = 0.999 to ensure more precise parameter tuning. A softmax classifier using Equation 13 was employed to classify the input into categories: Real, Fake.

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=0}^{1} e^{z_j}} \tag{13}$$

where $P(y_i)$ = probability of class i

$$L = -(y \log(\hat{y}) + (1 - y)\log(1 - \hat{y})) \quad (14)$$

where y = true label, $\hat{y}$ = predicted probability for real, 1- $\hat{y}$ = predicted probability for fake.

Cross-entropy loss computed using Equation 14 is used to calculate classification errors and to improve the accuracy of the model in predictions.

The proposed model combines all four modules into a unified framework. The DistilBERT-based textual encoder captures semantic and contextual embeddings from input text, while the VGG-19-based visual encoder extracts spatially preserved image patch sequences, allowing the attention mechanism to focus on more relevant regions of the image. These heterogeneous features are then fused through a cross-attention fusion module, which models rich interactions between text and image sequences there by enhancing multimodal understanding. Finally, the fused representation is passed through an MLP classifier with a softmax layer to determine whether the information is real or fake. As shown in Figure 3, this pipeline ensures that both textual and visual modalities contribute to the classification, improving the reliability in the detection of fake news.

In the proposed model, textual and visual encoders are designed using DistilBERT and VGG-19. DistilBERT, a compressed version of BERT, has 40% fewer parameters and requires less training time while retaining nearly the same performance, making it more suitable for processing textual content. VGG-19, a robust feature extractor, is particularly effective in preserving spatial patterns in noisy and low-resolution images commonly found in social media posts. Given that the experiments were conducted on Google Colab Pro with a Tesla T4 GPU and the scale of the Fakeddit dataset, the computational efficiency of VGG-19 and DistilBERT—both having lower parameter count—provides an optimal balance between performance and efficiency. Although more recent architectures such as ResNet, Vision Transformers (ViT) and RoBERTa have demonstrated significant performance, they are computationally heavier and require extensive resources for training. Moreover, both models enhance interpretability: VGG-19 allows for the visualization of spatial feature maps, while DistilBERT generates contextual embeddings that are both lightweight and powerful. This combination makes the proposed model not only computationally efficient, but also highly explainable.

5. Results and Discussions

Fakeddit is a multimodal dataset consisting of more than one millon samples. After preprocessing, the 124,530 multimodal samples (text+image) from the dataset were available for experimental analysis in this study. The dataset is then partitioned into train, test and validation sets as shown in Table 3. Tests were conducted on Google Colab Pro, an open-source Python notebook environment featuring a Tesla T4 GPU with 16GB of VRAM and approximately 13GB of system RAM. To ease the T4 GPU's memory constraints, mixed precision training using 'torch.cuda.amp', reducing both training time and VRAM usage, is used. Checkpoints were saved after every epoch to resume the training process if Colab disconnects.

Table 3. Dataset Proportions

Set	Samples	Percentage
Training	99,624	80%
Test	12,453	10%
Validation	12,453	10%

Table 4 illustrates the performance of the MCAF-FND model across various epochs and batch sizes, measured using standard metrics. Batch sizes of 16, 20, and 32 were evaluated, with the best accuracy (93.2%) achieved using a batch size of 16. The model was trained for a maximum of 20 epochs, with early stopping implemented based on validation loss to prevent overfitting. Figure 7 illustrates the average performance metrics of the model across different batch sizes. As shown in the figure, MCAF-DND model achieved highest accuracy when the batch size was 16 and the number of epochs was 10.

Table 4. Model performance across different epochs and batch sizes.

Epochs	Batch Size	Cross-Entropy Loss	Accuracy	Precision	Recall	F1-Score	AUC	Matthews Correlation Coefficient
5	16	0.52	0.90	0.80	0.82	0.81	0.83	0.58
	20	0.50	0.89	0.79	0.80	0.80	0.82	0.57
	32	0.48	0.88	0.78	0.79	0.79	0.81	0.56
10	16	0.38	**0.932**	0.93	0.92	0.93	0.92	0.70
	20	0.40	0.92	0.90	0.91	0.90	0.89	0.64
	32	0.41	0.91	0.89	0.90	0.90	0.88	0.62
20	16	0.36	0.93	0.93	0.93	0.93	0.91	0.68
	20	0.38	0.93	0.91	0.92	0.91	0.90	0.67
	32	0.39	0.92	0.90	0.91	0.91	0.89	0.65

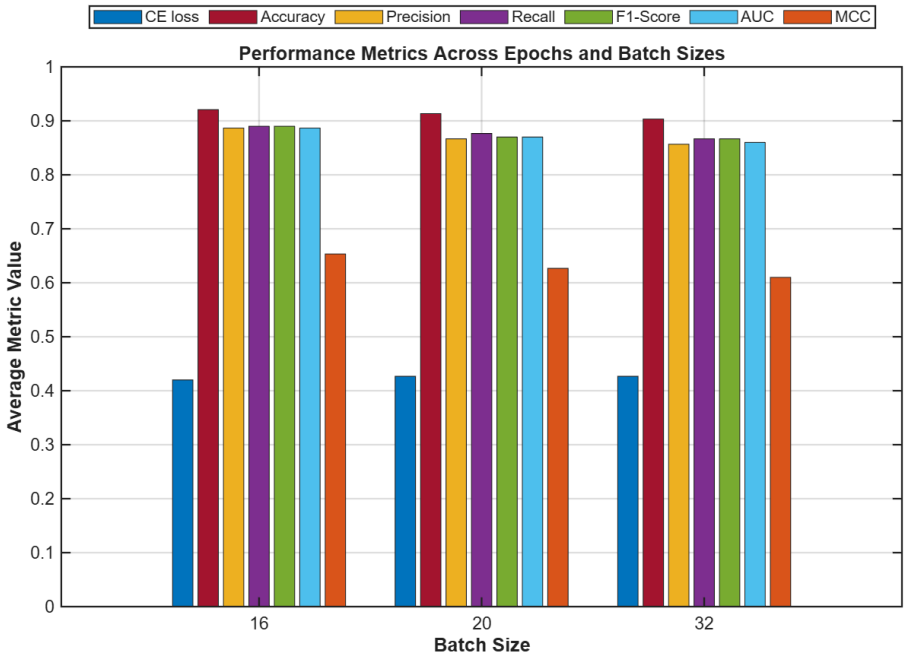


Fig. 7. Average performance across different batch sizes

5.1 Cross-Attention Visualization

- **Attention Heatmap:** To demonstrate the effectiveness of the cross-attention fusion module in capturing multimodal interactions between textual and visual cues, we generated an attention heatmap that visualizes how the model aligns textual tokens with relevant image patches during multimodal fusion. The attention weights obtained from the cross-attention fusion module are then reshaped to a 7 x 7 grid to generate heatmaps. Each intensity value in the heatmap represents the attention between a word in the text sequence and its corresponding spatial patch in the image feature grid. Regions with higher intensity, shown on a Yellow-to-Red color scale, indicate higher attention weights, reflecting a stronger association between image patches and the text query. Regions with lower intensity, shown on Blue-to-Green color scale, indicate weaker attention weights, suggesting that the image patches are less relevant to the text query. For example, when the input text is *"a man floating above a raft"*, the attention heatmap generated, as shown in Figure 8, highlights the upper region of the image containing the man and the raft, and indicates that the model focuses on semantically meaningful regions of the image associated with the input text.

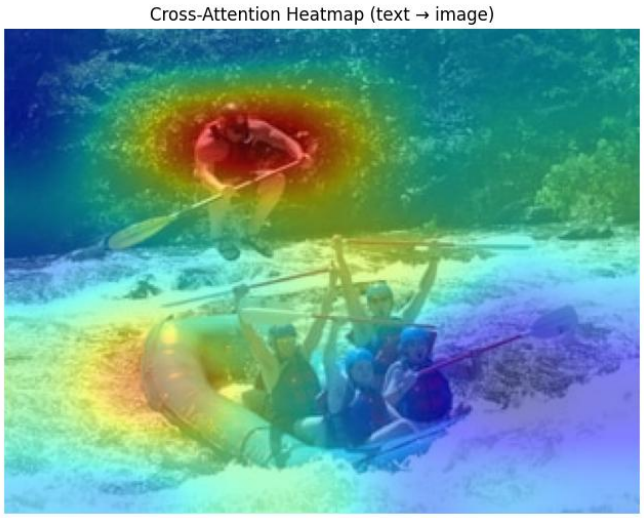


Fig. 8. Cross-Attention Heatmap

- **t-SNE plot**: t-distributed Stochastic Neighbor Embedding (t-SNE) was applied to visualize the fused feature representations learned by the MCAF-FND model on the Fakeedit test set. t-SNE projects the high-dimensional features to two dimensions. Figure 9 shows a clear distinction between fake and real samples

forming two clearly separated clusters indicating that the proposed model efficiently learns differentiative multimodal representations.

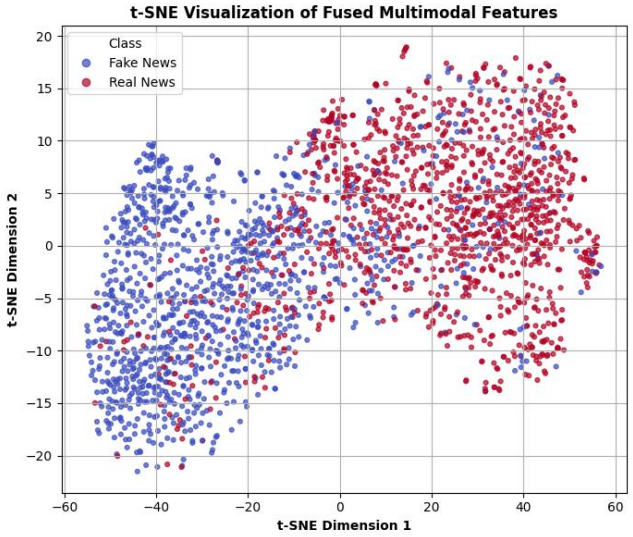


Fig. 9. t-SNE Visualization plot

5.2 Comparison with other baseline models

Baselines: This section compares the results with existing multimodal baseline models for false news detection;

- *FNDB* [17]: Uses XLNet and VGG-19 to retrieve textual and visual attributes, respectively. They used a cross-modal attention mechanism to combine both linguistic and visual features. Its key ability is to load different pre-trained BERT models based on the language of the dataset.
- *MPFN* [18]: The model uses a BERT-based feature extractor, a transformer-based visual feature extractor, and a progressive feature fusion process. An MLP mixer acts as a fusion module, blending the features between modalities.
- *CAMI* [22]: Using convolutional neural networks, CNN can detect if users are spreading incorrect information by segmenting tweets into sequences of predetermined length. Its key addition is the ability to learn feature representations, which allows it to perform early detection works.
- *att-RNN* [23]: The att-RNN recurrent neural network uses the attention mechanism to combine different features related to different modalities. For feature extraction, LSTM models the text and pre-trained VGG-19 models the pictures.
- *MVAE* [24]: It is a literature-based multitask model that integrates a multimodal variational auto encoder and a fake news detector. It uses a bidirectional LSTM for textual features and a pre-trained VGG-19 for image feature extraction. They constructed an intermediate expression using the spliced features of both models to rebuild input features and classify false news.

Table 5 provides a comparison of various multimodal fake news detection models based on their performance in different multimodal datasets. While Weibo [23], Twitter [25] and Gossipcop [26] datasets consisting of 9.5k, 13k, 13k samples, the Fakeddit dataset consists of 628k samples. The proposed MCAF-FND model achieved its highest accuracy of 0.932 when evaluated on the Fakeddit multimodal dataset.

Table 5. Performance comparison of various models

Methods	Data	Accuracy	Precision	Recall	F1-Score
FNDB [17]	Gossipcop	0.87	0.79	0.44	0.57
	Weibo	0.89	0.89	0.97	0.93
MPFN [18]	Twitter	0.83	0.85	0.92	0.88
	Weibo	0.84	0.86	0.89	0.89
CAMI [22]	Twitter	0.78	0.74	0.85	0.79
	Weibo	0.74	0.74	0.76	0.74
Att-RNN [23]	Twitter	0.66	0.75	0.62	0.68
	Weibo	0.77	0.85	0.66	0.74
MVAE [24]	Twitter	0.91	0.93	0.92	0.92
	Weibo	0.82	0.85	0.77	0.81
MCAF-FND	Fakeddit	0.93	0.93	0.92	0.93

Figure 10 illustrates a comparative bar plot illustrating the best performance of various multimodal fake news detection models, as shown in Table 5. The models are evaluated against the Fakeddit, Twitter, and Weibo datasets. It is evident from the visualization that the proposed MCAF-FND model, evaluated on the Fakeddit dataset, outperforms all the other models. In particular, MCAF-FND achieves the highest accuracy, surpassing the MVAE model by 2%, with an accuracy of 0.932 compared to the best performance of 0.91 by MVAE. These visual trends emphasize that the cross-attention fusion mechanism in MCAF-FND effectively balances all metrics, ensuring robust multimodal fake news detection.

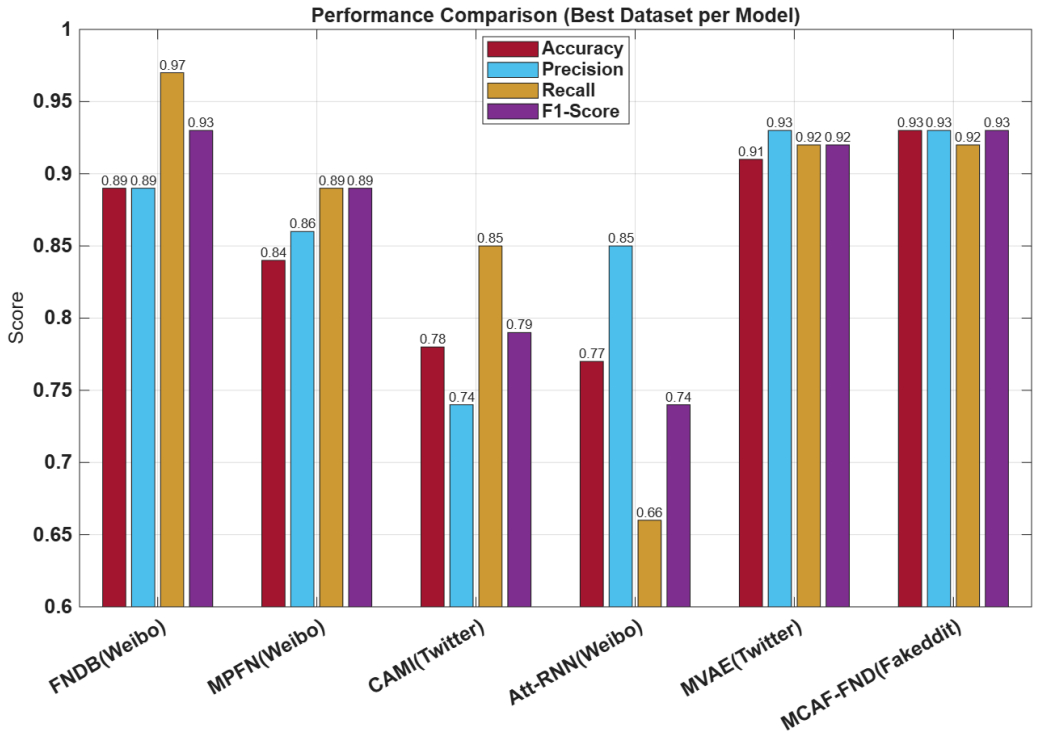


Fig. 10. Performance comparison of different fake news detection models on their respective best-performing datasets.

6. Conclusion

This study proposed a Multimodal Cross-Attention Fusion based Fake News Detection (MCAF-FND) model to address the challenges of fake news detection. The model employs DistillBERT to extract rich semantic and contextual embeddings from textual content and VGG-19 to capture spatially aware visual features from images. These heterogeneous features are then fused using a cross-attention based feature fusion, to capture semantic correspondence between both the modalities, thereby enhancing multimodal understanding. The fused features are then classified using a multi-layer perceptron (MLP) with softmax, enabling robust discrimination between real and fake news. The experimental results on the Fakeddit dataset show that the proposed MCAF-FND model attains a classification accuracy of 93.2%, outperforming the traditional mid-fusion, late-fusion and CNN-based methods. In future research, the framework can be extended to consider other multimodal content, including audio, video, and other metadata along with quantum classifiers, which may further enhance accuracy and computational efficiency through quantum parallelism.

Author Contributions Statement

Vineela Krishna. Suri – Conceptualization, Methodology, Supervision, Model Training, Validation, Performance Evaluation, Formal Analysis, Visualization, and Writing: Proposed research ideas, constructed the overall framework, supervised project execution, led the model training process, validated results using standard metrics, performed performance evaluation, conducted in-depth analysis, prepared visualizations, and drafted the initial manuscript.

Prasad. GVSNRV – Validation and Writing – Review and Editing: Contributed to the verification of experimental outcomes, provided critical feedback to improve the quality and clarity of the manuscript, and assisted in reviewing and refining the final version of the paper.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

This research did not receive external funding.

Data Availability Statement

This study analyzed publicly available datasets. The data that support the findings of this study can be downloaded from https://github.com/entitize/Fakeddit

Ethical Declarations

This study does not involve any experiments on human subjects or animals conducted by the authors. The research is based on publicly available datasets, and all data used in this study were handled in accordance with relevant ethical guidelines and regulations. No personal or sensitive information was collected or used.

Therefore, ethical approval was not required for this study.

Acknowledgments

We sincerely thank the reviewers and experts for their professional evaluation and valuable recommendations, which have significantly contributed to improving the quality of the study and the reliability of its results.

Declaration of Generative AI in Scholarly Writing

During the preparation of this work, the authors used generative AI tools to assist in language refinement, grammar correction, and improving the clarity of the manuscript. The use of these tools was limited to the writing process only and did not involve data analysis or interpretation of results. The authors take full responsibility for the content of the publication and confirm that all interpretations, results, and conclusions are their own.

Abbreviations

The following abbreviations are used in this manuscript:

CNN – Convolutional Neural Network
DL – Deep Learning
MLP – Multilayer Perceptron
VGG-19 – Visual Geometry Group 19
BERT – Bidirectional Encoder Representations from Transformers
DistilBERT – Distilled Bidirectional Encoder Representations from Transformers
ResNet – Residual Neural Network
ViT – Vision Transformer
RoBERTa – Robustly Optimized BERT Approach
FC – Fully Connected (Layers)
t-SNE – t-distributed Stochastic Neighbor Embedding
MCAF-FND – Multimodal Cross-Attention Fusion for Fake News Detection

Appendix

No supplementary material is provided.

References

[1] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *SIGKDD Explorations Newsletter, 19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

[2] Haque, M. M., Yousuf, M., Alam, A. S., Saha, P., Ahmed, S. I., & Hassan, N. (2020). Combating misinformation in Bangladesh: Roles and responsibilities as perceived by journalists, fact-checkers, and users. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW2), Article 130, 1–32. https://doi.org/10.1145/3415201

[3] Balogun, T. E., Abosede, S. A., Joda, S. C., Balogun, O., Olotu, P. K., & Faluyi, S. G. (2024). Machine learning approaches for detecting and mitigating the impact of fake news in online information ecosystems. In *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)* (pp. 1–6). https://doi.org/10.1109/SEB4SDG60871.2024.10630248

[4] Segura-Bedmar, I., & Alonso-Bartolome, S. (2022). Multimodal fake news detection. *Information, 13*(6), 284. https://doi.org/10.3390/info13060284

[5] Raval, M. S., Roy, M., & Kuribayashi, M. (2022). Survey on vision-based fake news detection and its impact analysis. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 1837–1841). https://doi.org/10.23919/APSIPAASC55919.2022.9980089

[6] Alnabhan, M. Q., & Branco, P. (2024). Fake news detection using deep learning: A systematic literature review. *IEEE Access, 12*, 114435–114459. https://doi.org/10.1109/ACCESS.2024.3435497

[7] Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 422–426). https://doi.org/10.48550/arXiv.1705.00648

[8] Põldvere, N., Uddin, Z., & Thomas, A. (2023). The PolitiFact-Oslo corpus: A new dataset for fake news analysis and detection. *Information, 14*(12), 627. https://doi.org/10.3390/info14120627

[9] Mattern, J., Qiao, Y., Kerz, E., Wiechmann, D., & Strohmaier, M. (2021). FANG-COVID: A new large-scale benchmark dataset for fake news detection in German. https://doi.org/10.18653/v1/2021.fever-1.9

[10] Choi, H., & Ko, Y. (2022). Effective fake news video detection using domain knowledge and multimodal data fusion on YouTube. *Pattern Recognition Letters, 154*, 44–52. https://doi.org/10.1016/j.patrec.2022.01.007

[11] Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., & Yu, P. S. (2018). TI-CNN: Convolutional neural networks for fake news detection. https://doi.org/10.48550/arXiv.1806.00749

[12] Patel, S., & Surati, S. (2024). MTL-rtFND: Multimodal transfer learning for real-time fake news detection on social media. https://doi.org/10.1007/978-3-031-53731-8_19

[13] Yadav, A., & Gupta, A. (2024). An emotion-driven, transformer-based network for multimodal fake news detection. *International Journal of Multimedia Information Retrieval, 13*. https://doi.org/10.1007/s13735-023-00315-3

[14] Zhang, L., Zhang, X., Zhou, Z., Huang, F., & Li, C. (2024). Reinforced adaptive knowledge learning for multimodal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(15), 16777–16785. https://doi.org/10.1609/aaai.v38i15.29618

[15] Yadav, A., Gaba, S., Khan, H., Budhiraja, I., Singh, A., & Singh, K. K. (2024). ETMA: Efficient transformer-based multilevel attention framework for multimodal fake news detection. *IEEE Transactions on Computational Social Systems, 11*(4), 5015–5027. https://doi.org/10.1109/TCSS.2023.3255242

[16] Wang, J., Zheng, J., Yao, S., Wang, R., & Du, H. (2023). TLFND: A multimodal fusion model based on three-level feature matching distance for fake news detection. *Entropy, 25*(11), 1533. https://doi.org/10.3390/e25111533

[17] Liang, Z. (2023). Fake news detection based on multimodal inputs. *Computers, Materials & Continua, 75*, 4519–4534. https://doi.org/10.32604/cmc.2023.037035

[18] Jing, J., Wu, H., Sun, J., Fang, X., & Zhang, H. (2023). Multimodal fake news detection via progressive fusion network. *Information Processing & Management, 60*(1), 103120. https://doi.org/10.1016/j.ipm.2022.103120

[19] Fakeddit website. Available at: https://fakeddit.netlify.app/

[20] Fakeddit GitHub repository. Available at: https://github.com/entitize/Fakeddit

[21] Nakamura, K., Levy, S., & Wang, W. Y. (2020). Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 6149–6157).

[22] Yu, F., Liu, Q., Wu, S., Wang, L., Tan, T., et al. (2017). A convolutional approach for misinformation identification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 3901–3907). https://doi.org/10.24963/ijcai.2017/545

[23] Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2017). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia* (pp. 795–816). https://doi.org/10.1145/3123266.312345

[24] Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). MVAE: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference (WWW '19)*. https://doi.org/10.1145/3308558.3313552

[25] Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)* (pp. 675–684). https://doi.org/10.1145/1963405.1963500

[26] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data, 8*(3), 171–188. https://doi.org/10.1089/big.2020.0062

Authors' Profiles



Vineela Krishna. Suri is a Research Scholar in the Department of Computer Science and Engineering at Jawaharlal Nehru Technological University, Kakinada. She holds an M.Tech in Computer Science and Engineering from JNTUK, Kakinada. Her areas of interest include Machine Learning, Deep Learning, and Web Programming. She has contributed to various research publications in reputed journals and conferences. Her research focuses on developing innovative solutions using artificial intelligence and data-driven technologies. Additionally, she is passionate about exploring emerging trends in web development and AI-driven applications.



Dr. GVSNRV. Prasad, M.S, M.Tech, Ph.D., is a Professor of Computer Science and Engineering and Director (PGDCA) at Seshadri Rao Gudlavalleru Engineering College. He earned his Ph.D. in Computer Science and Engineering from JNTUK, Kakinada in the year 2012. With 35 years of teaching experience, he published 39 Journal articles in various reputed National and International journals, 19 research articles in various National and International conferences, published 2 book chapters and guided five research scholars. His areas of interest include Machine Learning, Deep Learning, and Data Mining.

How to cite this paper: Vineela Krishna. Suri, Prasad. GVSNRV, "A Hybrid CNN-Transformer Model for Multimodal Fake News Detection Using Feature Fusion", International Journal of Modern Education and Computer Science(IJMECS), Vol.18, No.2, pp. 132-146, 2026. DOI:10.5815/ijmecs.2026.02.08