International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(16s) 2026 pp. 156-171
https://doi.org/10.70917/ijcisim-2026-4565
Article

Unified Multimodal Misinformation Detection Model Based on EfficientNet-B0, BERT and Stack Ensemble

[1]Javeriya Naaz I. Syed, [2]Dr. Ranjit R. Keole

[1] Department of Computer Science, HVPM COET, Amravati (M.S.), India
Email: javeriyanisyed@gmail.com
[2] Department of Information Technology, HVPM COET, Amravati (M.S), India
Email: ranjitkeole@gmail.com

Abstract: Multimodal misinformation has been identified as a key challenge in social media platforms, where the misleading information is presented using both textual and visual modalities. The current state-of-the-art methods for detecting misinformation are primarily unimodal or based on monolithic deep learning models that lack the ability to generalize across different patterns of misinformation. This paper presents a Hybrid Vision-Language Stacked Ensemble Model that combines deep semantic features from EfficientNet-B0 and BERT with traditional ensemble learning techniques to handle multimodal misinformation detection. In particular, high-level image features were extracted using EfficientNet-B0 and text embeddings using a pre-trained BERT model. The learned multimodal feature representations were then used by a stacked ensemble classifier that leverages the complementary capabilities of Random Forest, Gradient Boosting, and a Logistic Regression meta-learner. Experiments conducted on the MMFakeBench dataset for binary and multiclass misinformation classification tasks have shown that the proposed framework achieves better performance compared to individual base learners and conventional fusion approaches. Further detailed evaluation had also validated the efficacy of modality-wise feature fusion and the robustness of the ensemble learning approach against class imbalance and noisy data. The proposed framework provides a scalable, interpretable, and robust solution for real-world multimodal misinformation detection on social media platforms.

Keywords: multimodal misinformation, ensemble learning, EfficientNet-B0, BERT, stacked classifier, social media

1. INTRODUCTION

The proliferation of social media platforms has greatly accelerated the spread of misinformation, which poses a serious challenge to public trust, social stability, and decision-making. In recent years, misinformation has shifted from text-based deception to multimodal deception, where the misleading information is presented through carefully selected or manipulated images and videos. Large-scale survey studies have shown that multimodal misinformation is more persuasive and harder to detect than unimodal information, which has triggered the need for advanced multimodal detection systems [1], [2].

To cope with the challenge, there has been a growing interest in exploring multimodal misinformation detection (MMD) methods that incorporate both text and image information. Current studies have proposed taxonomies, datasets, and deep learning-based methods for cross-platform and multimodal misinformation detection, which have shown the limitations of unimodal models when used on real-world social media data [3]. Comprehensive reviews have also stressed that although significant progress has been made, many existing systems are based on end-to-end deep models, which are computationally intensive and vulnerable to modality imbalance and noisy data [4].

Recent studies on deep learning-based misinformation detection show that transformer language models and convolutional neural networks (CNN) have emerged as the most popular feature extractors for text and image modalities, respectively [5]. However, most existing works are based on single-model or closely-coupled fusion

architectures, which generally lack the ability to generalize well across datasets and types of misinformation. In addition, traditional machine learning (ML) paradigms such as ensemble learning have not been adequately explored in combination with deep multimodal features, although they have been shown to be highly robust and interpretable [6].

To address the above challenges, this paper presents a hybrid vision-language stacked ensemble model for multimodal misinformation detection. This approach is expected to promote robustness, prevent overfitting, and improve generalization performance for binary and multiclass misinformation classification tasks. The proposed approach is evaluated on the MMFakeBench dataset, and the results show that it outperforms unimodal baselines and individual classifiers. The main contributions of this paper are summarized as follows:

- Integrated EfficientNet-B0 for visual feature extraction and BERT for contextual text representation, enabling effective modeling of heterogeneous multimodal content.
- Introduced a stacked ensemble classifier combining random forest and gradient boosting with a logistic regression meta-learner.
- Evaluated multimodal misinformation detection framework under both binary and multiclass misinformation detection configurations.
- Analyzed the contribution of text-only, image-only, and hybrid modalities, as well as the impact of individual classifiers versus the stacked ensemble framework.

The rest of this paper is organized as follows: Section 2 provides an overview of related work on multimodal misinformation detection and ensemble learning; Section 3 introduces the proposed hybrid vision-language stacked ensemble approach; Section 4 presents the experimental results and discussion; and Section 5 concludes the paper with key findings and future work.

2. Related Work

This section critically reviews the existing literature on misinformation detection, with a special emphasis on unimodal and multimodal methods, feature representation, and classification techniques. A multimodal ensemble-based fake news detection system incorporating deep visual and textual feature extraction and multiple classifiers proved that ensemble fusion performs substantially better than unimodal and single-model methods in terms of classification accuracy and robustness [7]. Machine learning (ML) and deep learning (DL) models used for health-related COVID-19 information on Twitter/X revealed that transformer-based text representation methods combined with fact-checking techniques have better precision and accuracy in misinformation classification compared to conventional classifiers [8]. An evidence-grounded multimodal misinformation detection (MMD) system incorporating attention-based graph neural networks (GNN) effectively captured the interactions between text, images, and external evidence, leading to better detection accuracy and improved generalization capability for complex misinformation [9]. A domain-aware test-time adaptation framework for MMD adapted model parameters at test time for unseen target domains, resulting in consistent performance improvements without the need for additional labeled data [10]. A reasoning-based explainable multimodal fake news detection model using large language models (LLM) and transformer models improved detection performance on low-

resource languages with interpretable reasoning paths for better model explainability [11]. A hybrid optimization-driven fake news detection method combining reinforced transformer models and metaheuristic optimization strategies demonstrated better classification performance and faster convergence than baseline transformer models [12]. A multilingual and multimodal disinformation dataset with additional contextual and supportive information facilitated comprehensive benchmarking and demonstrated improved multimodal detection model performance across languages and out-of-context settings [13]. A new dataset and benchmark for grounding multimodal misinformation focused on evidence alignment across modalities, with experimental results demonstrating the effectiveness of grounded models over non-grounded models in terms of detection accuracy and reliability [14]. An early fusion MMD model incorporating linguistic, visual, and social cues demonstrated improved classification accuracy, especially for ambiguous and context-dependent misinformation samples [15]. A comprehensive framework for misinformation analysis and detection on various multimodal media sources demonstrated scalability and flexibility, with experimental results validating its efficacy in processing diverse and large-scale misinformation data [16]. A multi-granularity consistency-driven MMD framework incorporating joint modeling of fine and coarse semantic alignments between the textual and visual modalities demonstrated improved misinformation detection accuracy and robustness over traditional fusion approaches [17].

A common benchmark dataset for MMD on social media was proposed together with baseline multimodal models, which set up a standardized evaluation framework and highlighted substantial performance differences between unimodal and multimodal models [18]. A multimodal misinformation dataset for media authenticity assessment included various manipulation tasks, which enabled comprehensive evaluation and demonstrated that multimodal models significantly outperformed unimodal baselines in identifying forged and misleading information [19]. Large vision-language models were employed for MMD by simultaneously reasoning about images and texts, and the experimental results demonstrated excellent zero-shot and few-shot performance compared to task-specific multimodal models [20]. Annotated multi-event tweet data was created to facilitate misinformation detection for various real-world events, which enabled comprehensive evaluation and demonstrated the improved generalization of detection models trained on event-rich data [21]. An attentive fusion-based MMD system dynamically fused text and image features, which achieved better classification performance than early and late fusion baselines [22]. A computational modeling method was suggested for studying susceptibility to misinformation by learning patterns of

misbeliefs from linguistic and behavioral cues, resulting in efficient prediction of acceptance tendencies of misinformation [23]. A knowledge-augmented large vision-language model (VLM) paradigm leveraged external factual knowledge in multimodal reasoning, greatly improving the accuracy and interpretability of fake news detection compared to knowledge-agnostic models [24]. A novel task definition emphasizing misinformation with legal implications brought forth dedicated datasets and metrics, shedding light on the social implications of misinformation and the inadequacy of current models for misinformation detection in critical tasks [25]. A strong multimodal misinformation benchmark addressing unimodal bias with strict evaluation constraints enforced a bias-aware model that showed better detection performance in terms of reliability and generalizability [26]. A misinformation detection system combining emotion analysis with user behavior modeling in social networks proved that affective and behavioral information can improve detection accuracy compared to content-based approaches [27].

The use of multimodal LLMs to generate synthetic multimodal data was used to augment the training of misinformation detection models, leading to significant performance improvements and better generalization capabilities when tested on real-world multimodal datasets [28]. A framework for evaluating data quality in misinformation detection identified high-quality training data using statistical and feature-driven approaches, demonstrating that performance gains can be achieved by training models on carefully curated and noise-cleaned datasets [29]. A two-branch multimodal model for fake news detection using multimodal bilinear pooling with attention mechanisms effectively modeled cross-modal interactions, achieving better classification accuracy than traditional fusion-based models [30]. A supervised learning-based approach for misinformation classification using user credibility features improved the trustworthiness of misinformation detection, especially in social media settings with diverse user behavior [31]. An explainable misinformation detection system supporting multiple social media platforms integrated transparent ML methods, allowing for accurate detection with model transparency and trustworthiness [32]. A thorough analysis of fake news propagation and detection methods on social networking sites demonstrated that ML methods are superior to rule-based methods, although also pointing out scalability and early detection issues [33]. ML-based prevention model for online misinformation spread employed classification and intervention strategies, with experimental results indicating effective reduction in misinformation dissemination across social media platforms [34].

Although there has been considerable progress in the multimodal misinformation detection task, the existing literature shows several shortcomings. Most of the existing works are based on a single fusion approach, including early fusion, late fusion, or attention-based fusion, which may not be able to effectively leverage the complementary information provided by the different modalities, especially in the presence of noisy or weak modalities. Some of the existing works are based on complex models such as graph neural networks, complex vision-language models, or knowledge-augmented models, which, although efficient, are computationally expensive and lack scalability. Dataset-focused works mainly concentrate on benchmarking and grounding but lack sufficient focus on the generalization of the model on different types of misinformation and modality imbalance. In addition, optimization and explainability models mainly focus on interpretability or low-resource configurations rather than the efficiency of multimodal learning. In contrast, the proposed model addresses these gaps by employing lightweight yet powerful feature extractors coupled with a stack ensemble learning strategy that integrates multiple base learners at the decision level

3. METHODOLOGY

The proposed methodology adopts a hybrid vision–language stacked ensemble framework for robust multimodal misinformation detection, as shown in Figure 1. The system jointly processes visual content (images) and

textual content (captions/posts) using modality-specific deep feature extractors, followed by a stacked ensemble learning strategy to fuse predictions from multiple base classifiers.

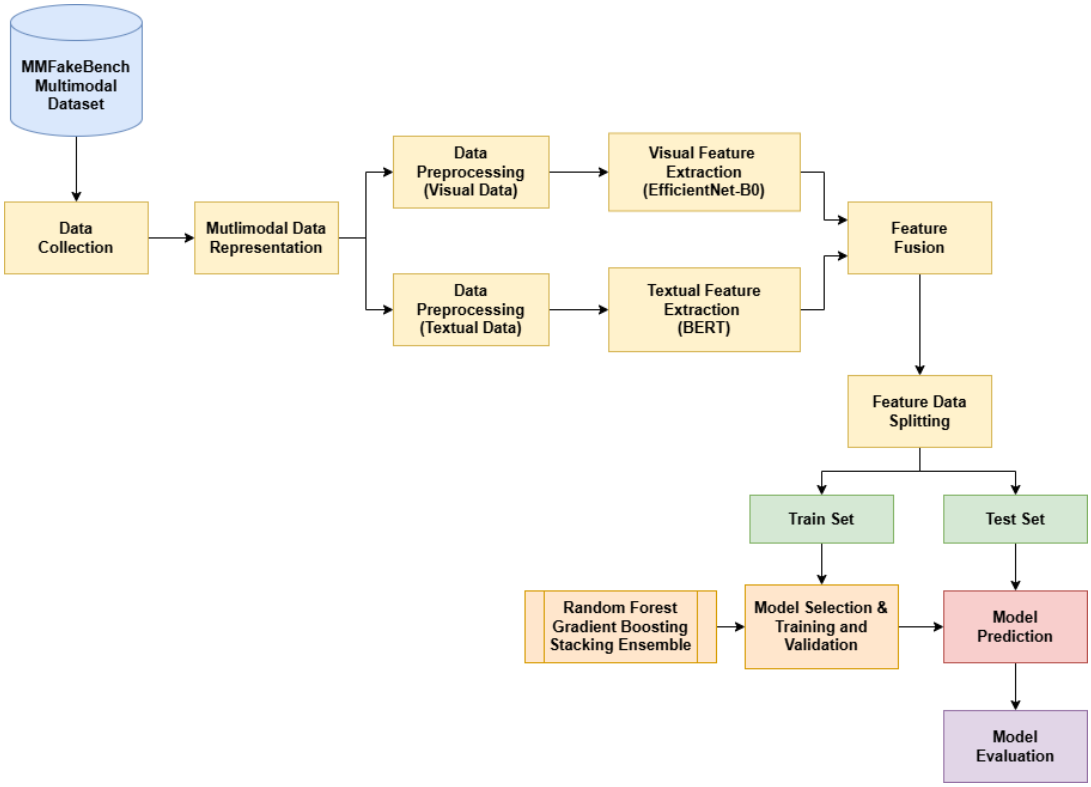

Figure 1: Multimodal Misinformation Detection Framework

3.1 MMFakeBench Dataset and Data Collection

The proposed framework utilizes the MMFakeBench dataset [35], a publicly available benchmark designed for multimodal misinformation detection (MMD), consisting of paired visual content (images) and textual content (captions or claims) annotated with ground-truth labels. Each data instance represents a real-world social media post, enabling realistic evaluation of multimodal misinformation patterns. The dataset covers diverse misinformation topics and visual–textual inconsistencies, enabling rigorous evaluation of multimodal learning models under realistic conditions. MMFakeBench encompasses one real category and three misinformation categories, such as textual veracity distortion, visual veracity distortion, and cross-modal consistency distortion. Its balanced structure, standardized splits, and rich multimodal annotations make MMFakeBench well suited for benchmarking and comparing advanced vision–language models and ensemble-based approaches for robust misinformation detection.

Each sample is formally represented as:

$Xi = \{Ii, Ti, yi\}$

where Ii denotes the image, Ti denotes the corresponding text, and $yi \in \{0,1\}$ represents the misinformation label.

To preserve modality-specific information while enabling joint learning, each multimodal post is processed as a unified tuple without premature fusion. This structured representation allows independent preprocessing and feature extraction pipelines for visual and textual data before fusion at the feature level.

3.2 Data Preprocessing

The visual modality is preprocessed to standardize image quality and enhance feature extraction consistency. All images are resized to a fixed resolution to ensure uniform input dimensions across the dataset, followed by pixel normalization to stabilize model training. Image quality checks are performed to filter out corrupted or low-resolution samples. Data augmentation techniques such as random flipping, rotation, and scaling are applied during training to

improve generalization and reduce overfitting. Finally, the processed images are passed through pretrained or custom visual encoders to extract discriminative visual features that complement the textual information, facilitating effective multimodal fusion in the misinformation detection framework.

The textual component of the dataset undergoes a systematic preprocessing pipeline to ensure consistency, noise reduction, and effective feature learning. Initially, raw text is cleaned by removing URLs, hashtags, user mentions, special characters, and redundant whitespace, followed by conversion to lowercase to maintain uniformity. Tokenization is then applied to segment text into meaningful units, after which stop words are removed to reduce lexical redundancy. Lemmatization is employed to normalize words to their base forms, thereby preserving semantic meaning while reducing vocabulary sparsity. In addition, punctuation handling and contraction expansion are performed to improve syntactic clarity. The preprocessed text is subsequently transformed into numerical representations using suitable feature extraction or embedding techniques, enabling effective learning by downstream classification models.

Formally:

$Ii' = Pv(Ii), Ti' = Pt(Ti)$

Where, $Pv(\cdot)$ and $Pt(\cdot)$ denote visual and textual preprocessing functions, respectively.

3.3 Feature Extraction

A. Visual Feature Extraction

Visual feature extraction plays a critical role in the proposed multimodal misinformation detection framework, as misleading or manipulated visual content often serves as a strong indicator of misinformation. In this study, visual representations are extracted from images using EfficientNet-B0, which is designed to capture both low-level visual patterns and high-level semantic cues such as objects, scenes, and contextual inconsistencies. By leveraging a deep convolutional structure with optimized scaling, EfficientNet-B0 generates compact yet discriminative feature embeddings that effectively encode visual characteristics relevant to detecting falsified, out-of-context, or misleading imagery commonly observed in misinformation posts.

EfficientNet-B0 follows a compound scaling strategy that uniformly scales network depth, width, and input resolution, enabling efficient learning with fewer parameters compared to traditional convolutional neural networks. Its architecture is composed of a series of mobile inverted bottleneck convolution (MBConv) blocks integrated with squeeze-and-excitation modules, which enhance channel-wise feature recalibration and improve representational capacity. The network begins with a standard convolution layer, followed by multiple stacked MBConv blocks with varying kernel sizes and expansion ratios, and concludes with a global average pooling layer to produce a fixed-length visual feature vector suitable for downstream classification and fusion tasks. The detailed architecture of the EfficientNet-B0 visual feature extraction module is illustrated in Figure 2.

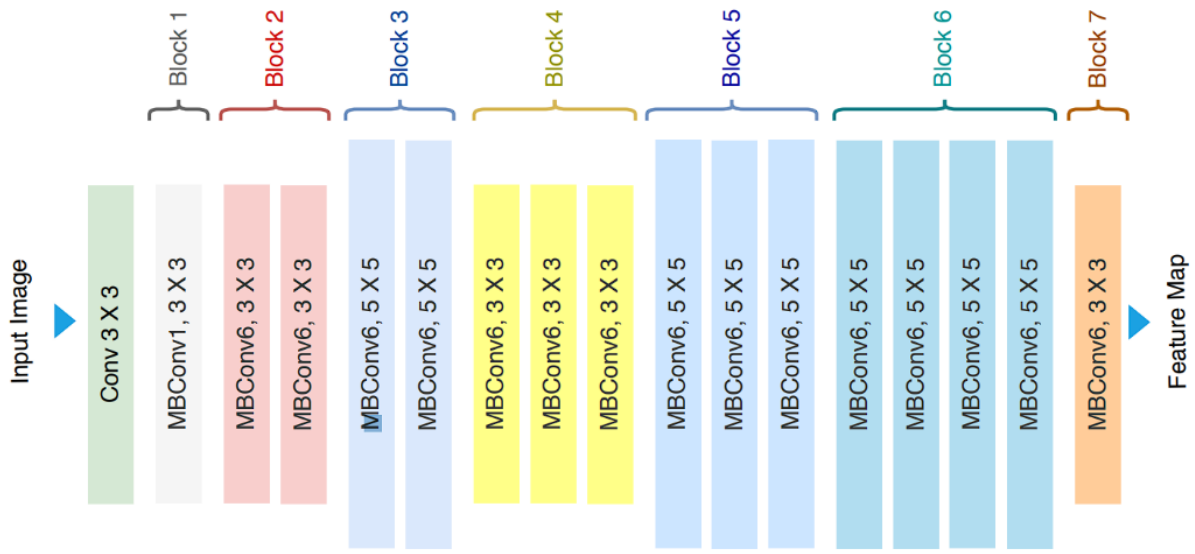

Figure 2: EfficientNet-B0 architecture

Given an image Ii′, the visual feature vector is obtained as:

$$\mathbf{F}_v^{(i)} = f_{\text{EffNet}}(I_i')$$

Where Fv(i) ∈ Rdv captures discriminative visual cues related to manipulated or misleading imagery.

B. Textual Feature Extraction

Textual feature extraction is a key component of the proposed framework, as misinformation is often characterized by deceptive language patterns, exaggerated claims, or misleading contextual cues embedded in textual content. In this study, textual representations are extracted using BERT, which is capable of modeling deep semantic and contextual relationships within text through bidirectional self-attention. By processing the entire text sequence simultaneously, BERT effectively captures word-level dependencies, syntactic structures, and contextual meanings, enabling the model to distinguish between credible information and misleading or manipulative textual narratives present in multimodal posts.

The BERT architecture is based on a multi-layer transformer encoder comprising stacked self-attention and feed-forward layers. Each input text is first tokenized and converted into token embeddings, positional embeddings, and segment embeddings, which are then passed through multiple transformer layers to generate contextualized representations. The special [CLS] token embedding from the final encoder layer is used as the aggregated textual feature vector, as it encapsulates the overall semantic representation of the input text. This feature vector is subsequently utilized for multimodal fusion and classification. The architecture of the BERT-based textual feature extraction module is depicted in Figure 3.

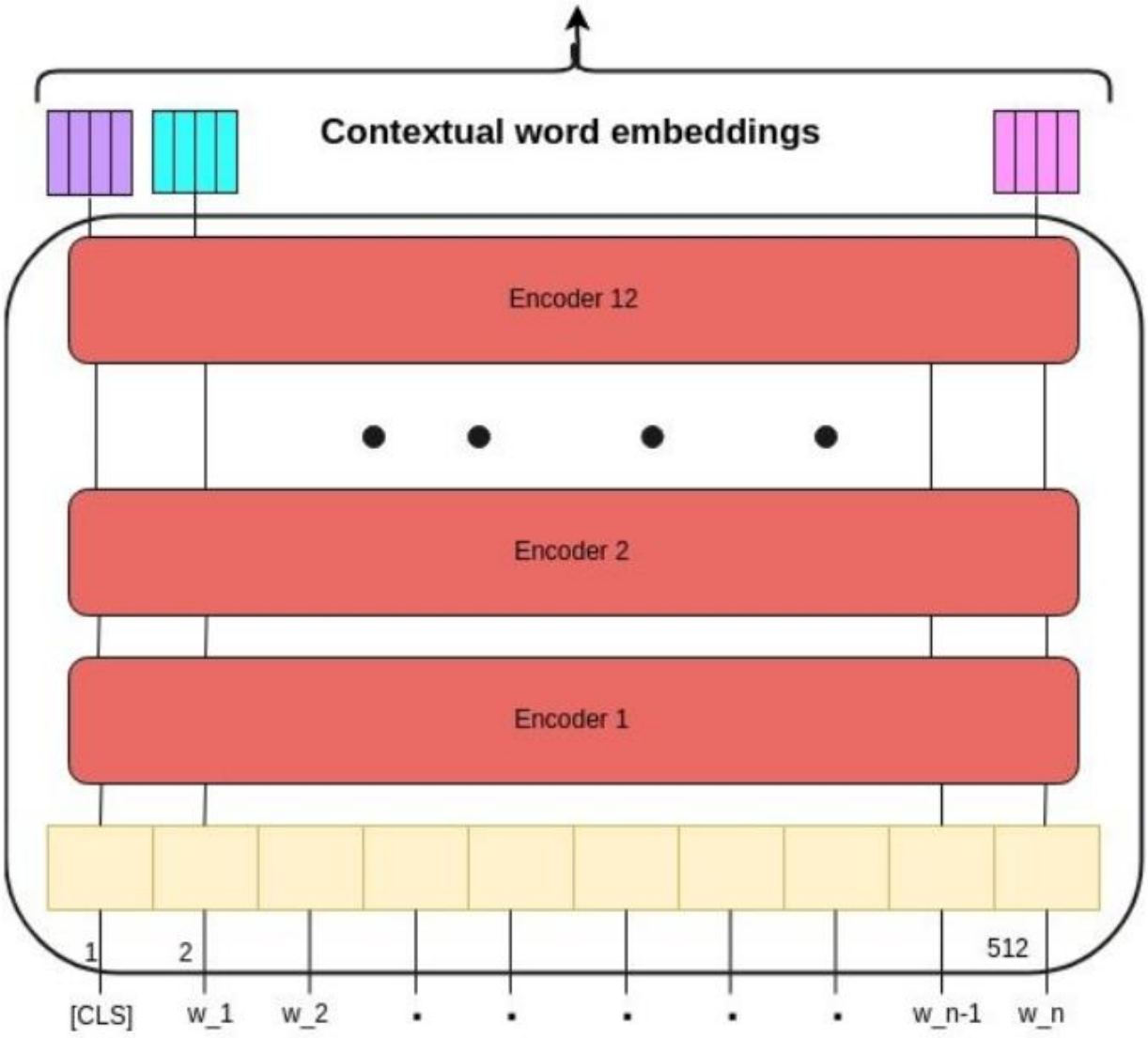

Figure 3: A high level diagram of textual feature extractor (BERT).

The [CLS] token embedding is used as the textual representation:

$$\mathbf{F}_t^{(i)} = \mathrm{BERT}(T_i')_{[\mathrm{CLS}]}$$

Where Ft(i) ∈ Rdt captures high-level semantic and syntactic information.

3.4 Feature Fusion

Feature fusion is employed to effectively integrate complementary information extracted from visual and textual modalities, enabling a unified representation for accurate multimodal misinformation detection. In the proposed framework, feature-level fusion is performed by concatenating the visual feature vector obtained from EfficientNet-B0 with the textual feature vector extracted using BERT, thereby preserving modality-specific discriminative characteristics while facilitating joint learning. This fusion strategy allows the model to capture cross-modal correlations, such as visual–textual inconsistencies and contextual mismatches that are indicative of misinformation, without introducing excessive computational complexity. The resulting fused feature representation serves as a comprehensive input to the ensemble learning models, enhancing robustness and improving detection

performance across diverse and heterogeneous multimodal misinformation scenarios. Feature-level fusion is performed by concatenating visual and textual embeddings to form a unified multimodal feature vector:

$$\mathbf{F}_m^{(i)} = [\mathbf{F}_v^{(i)} \,||\, \mathbf{F}_t^{(i)}]$$

This fused representation preserves complementary modality-specific information while enabling joint learning.

3.5 Feature Data Splitting

Feature data splitting is performed to ensure unbiased training, validation, and evaluation of the proposed multimodal misinformation detection model. After feature fusion, the unified multimodal feature set is divided into training and testing subsets using stratified sampling to preserve the original class distribution across all splits. The training set is used for model learning, the validation set supports hyperparameter tuning and model selection, and the testing set contains unseen samples for final performance assessment. This structured data splitting strategy helps prevent overfitting and ensures reliable evaluation of the model's generalization capability. The fused feature set is divided into training and testing subsets to ensure unbiased performance evaluation:

D = Dtrain ∪ Dtest

Stratified sampling is employed to maintain class balance across all splits.

3.6 Model Learning and Validation

Model training and validation are conducted in a structured manner to ensure robust learning and reliable performance estimation of the proposed multimodal misinformation detection framework. The fused multimodal feature vectors obtained after feature extraction and fusion are used as inputs to multiple ensemble-based classifiers. To mitigate overfitting and improve generalization, k-fold cross-validation is employed on the training set, where the data are partitioned into k equal folds. In each iteration, k−1 folds are used for training and the remaining fold is used for validation, and the process is repeated until every fold has served as the validation set. The average performance across all folds is considered for model selection and hyperparameter tuning.

A. Ensemble Bagging (Random Forest)

The Random Forest classifier is trained using the bagging principle, where multiple decision trees are constructed from different bootstrap samples of the training data. Each tree learns

$$\hat{y}_{RF} = \frac{1}{N}\sum_{j=1}^{N} T_j(\mathbf{F}_m)$$

independently using randomly selected feature subsets, which reduces variance and enhances robustness against noisy and redundant features present in multimodal data. During training, predictions from individual trees are aggregated through majority voting or probability averaging, and cross-validation results are used to determine optimal parameters such as the number of trees, tree depth, and feature sampling size. Each tree learns from a randomly sampled subset of the training data and predicted output is given by:

where Tj denotes the j-th decision tree.

B. Ensemble Boosting (Gradient Boosting)

The Gradient Boosting classifier is trained sequentially, where each weak learner focuses on correcting the errors of its predecessors. By minimizing a differentiable loss function, the model incrementally improves its ability to detect difficult and ambiguous misinformation samples. Cross-validation is employed to tune critical hyperparameters such as the learning rate, number of boosting iterations, and maximum tree depth, ensuring a balance between model complexity and generalization performance. The final prediction is computed as:

$$\hat{y}_{GB} = \sum_{k=1}^{K} \alpha_k h_k(\mathbf{F}_m)$$

where hk represents the weak learners and αk denotes their corresponding weights.

C. Stack Ensemble Learning

To leverage the complementary strengths of bagging and boosting, a stack ensemble learning strategy is adopted. Predictions generated by the trained Random Forest and Gradient Boosting models on the validation folds are used as meta-features to train a Logistic Regression meta-learner. This meta-model learns optimal weights for combining base-level predictions, enabling more accurate and stable decision-making. Validation performance across k-folds is used to select the final stacked model, which is then retrained on the full training dataset before evaluation on the independent test set. This hierarchical training and validation process ensures improved robustness, reduced bias–variance trade-off, and enhanced generalization across diverse multimodal misinformation scenarios. Predictions from Random Forest and Gradient Boosting models are used as meta-features:

$$\mathbf{Z} = [\hat{y}_{RF}, \hat{y}_{GB}]$$

A Logistic Regression meta-learner is trained to produce the final prediction:

$$\hat{y} = \sigma(\mathbf{w}^\top \mathbf{Z} + b)$$

where σ(·) is the sigmoid function.

3.7 Model Testing

Model testing is performed on a held-out test set containing unseen multimodal samples to objectively assess the generalization capability of the proposed framework under both binary and multiclass classification configurations. In the binary scenario, the model evaluates its ability to distinguish genuine content from misinformation, whereas in the multiclass configuration, it is tested on fine-grained misinformation categories defined in the MMFakeBench dataset, including real content, textual veracity distortion, visual veracity distortion, and cross-modal consistency distortion. The trained stacked ensemble model processes fused multimodal features of test samples and produces final predictions without any further parameter updates, ensuring unbiased evaluation. This testing strategy validates the model's effectiveness in handling diverse misinformation types and its robustness in capturing modality-specific as well as cross-modal inconsistencies present in real-world multimodal misinformation.

3.8 Model Evaluation

The performance of the proposed multimodal misinformation detection framework is evaluated using a held-out test set and standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics are derived from the confusion matrix, which summarizes the prediction outcomes in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Such an evaluation strategy enables a reliable assessment of the model's predictive capability, discrimination power, and robustness to class imbalance, which is inherent in misinformation detection tasks.

Accuracy quantifies the overall correctness of the classification model by measuring the ratio of correctly classified samples to the total number of test instances:

$$\text{Accuracy} = \frac{TP + TN}{\sum_{i=1}^{C}\sum_{j=1}^{C} M_{ij}}$$

where MijM_{ij}Mij represents the confusion matrix element corresponding to samples belonging to class iii and predicted as class jjj, and CCC denotes the total number of classes.

Precision evaluates the exactness of the model by measuring the proportion of correctly predicted positive instances among all samples predicted as positive:

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

Recall, also referred to as sensitivity, measures the model's ability to correctly identify actual positive instances:

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

To provide a balanced evaluation between precision and recall, the F1-score is computed as the harmonic mean of these two metrics:

$$\text{F1-score} = \frac{2TP}{2TP + FP + FN}$$

For the multiclass misinformation detection task, including real content and misinformation categories such as textual veracity distortion, visual veracity distortion, and cross-modal consistency distortion, the evaluation metrics are computed using macro-averaging and weighted-averaging strategies. Macro-averaging treats all classes equally by averaging metrics across classes, while weighted-averaging accounts for class imbalance by weighting each class according to its support.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The implementation of the proposed multimodal misinformation detection framework was carried out on a workstation running Windows 11, equipped with 16 GB RAM, an Intel Core i5 processor, and an NVIDIA RTX 3050 GPU, which provided adequate computational resources for deep feature extraction and ensemble-based model training. The experimental environment was developed using Python 3.8, along with widely adopted scientific and machine learning libraries to ensure reproducibility and robustness. NumPy and Pandas were used for efficient data handling and preprocessing, while NLTK and regular expression (re)

libraries supported textual preprocessing prior to BERT-based feature extraction. Deep learning models for visual and textual feature extraction were implemented using standard deep learning frameworks, and scikit-learn was employed for ensemble bagging (Random Forest), ensemble boosting (Gradient Boosting), stack ensemble learning with Logistic Regression as the meta-learner, and comprehensive model evaluation. Matplotlib and Seaborn were utilized for performance visualization and result analysis. The MMFakeBench dataset was downloaded and processed offline to ensure full local execution, experimental reproducibility, and controlled evaluation under both binary and multiclass misinformation detection configurations.

The performance of the proposed framework is evaluated under both binary and multiclass misinformation detection configurations using three classifiers: Random Forest (RF), Gradient Boosting (GB), and Stacked Ensemble (SE). For each classifier, experiments are conducted separately on text-only, image-only, and combined text–image modalities to analyze the contribution of individual and fused feature representations. The evaluation is reported in terms of accuracy, precision, recall, and F-score, as summarized in the tables. Binary classification performance using three classifiers across text, image, and text–image modalities is summarized in Tables 1–3.

Table 1: Result Evaluation of Binary Classification using RF

Modality	Accuracy	Precision	Recall	F-score
Text	0.73	0.78	0.73	0.66
Image	0.70	0.49	0.70	0.58
Text+Image	0.72	0.73	0.72	0.63

Table 2: Result Evaluation of Binary Classification using GB

Modality	Accuracy	Precision	Recall	F-score
Text	0.76	0.75	0.76	0.73
Image	0.70	0.49	0.70	0.58
Text+Image	0.76	0.75	0.76	0.73

Table 3: Result Evaluation of Binary Classification using SE

Modality	Accuracy	Precision	Recall	F-score
Text	0.75	0.75	0.75	0.70
Image	0.70	0.49	0.70	0.58
Text+Image	0.73	0.71	0.73	0.68

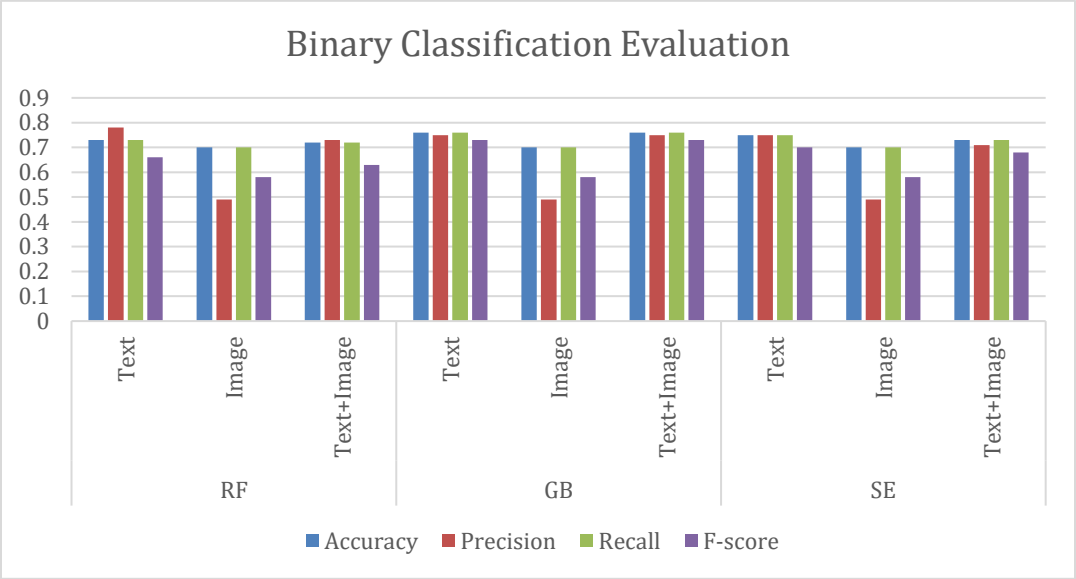

Figure 4: Binary classification performance across classifiers and modalities

In the binary classification configuration as shown in Figure 4, the Random Forest results in Table 1 show that text-based features achieve an accuracy of 0.73, outperforming image-only features, which record an accuracy of 0.70 and a comparatively low precision of 0.49. This indicates that textual cues are more discriminative than visual cues for binary misinformation detection. The multimodal text–image combination achieves balanced performance with an accuracy of 0.72, demonstrating that feature fusion helps stabilize predictions, although the improvement over text-only features remains marginal for RF. Similar trends are observed for the Gradient Boosting classifier, where text-based and multimodal features both achieve the highest accuracy of 0.76 and an F-score of 0.73. This highlights the effectiveness of GB in capturing nonlinear relationships and leveraging complementary multimodal information. The Stacked Ensemble results further reinforce this observation, with text-only features achieving the highest accuracy of 0.75, while the text–image combination provides competitive performance with improved recall and a balanced F-score of 0.68. In the multiclass classification configuration, the performance trends differ notably, as shown in Tables 4–6.

Table 4: Result Evaluation of Multiclass Classification using RF

Modality	Accuracy	Precision	Recall	F-score
Text	0.67	0.68	0.67	0.65
Image	0.28	0.08	0.28	0.13
Text+Image	0.64	0.64	0.64	0.62

Table 5: Result Evaluation of Multiclass Classification using GB

Modality	Accuracy	Precision	Recall	F-score
Text	0.67	0.68	0.67	0.66
Image	0.28	0.08	0.28	0.13
Text+Image	0.67	0.67	0.67	0.67

Table 6: Result Evaluation of Multiclass Classification using SE

Modality	Accuracy	Precision	Recall	F-score
Text	0.67	0.67	0.67	0.66
Image	0.28	0.08	0.28	0.13
Text+Image	0.65	0.65	0.65	0.64

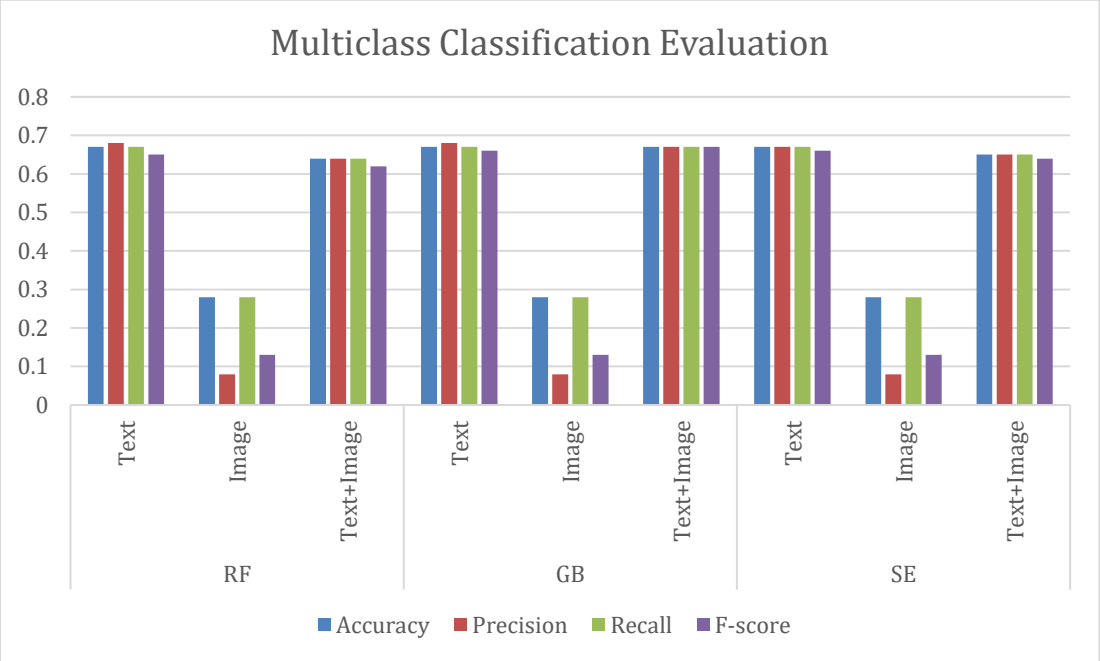

Figure 5: Multiclass classification performance across classifiers and modalities

Figure 5 shows that for the Random Forest classifier, text-based features achieve an accuracy of 0.67, whereas image-only features perform poorly, with an accuracy of 0.28 and an F-score of 0.13, indicating the limited discriminative capability of visual features when used independently for fine-grained misinformation categorization. The multimodal approach improves robustness, achieving an accuracy of 0.64 and an F-score of 0.62. Gradient boosting demonstrates superior and more consistent performance in the multiclass configuration, with the multimodal configuration achieving the highest accuracy and F-score of 0.67, closely matching the text-only performance while benefiting from improved class-wise balance. The Stacked Ensemble results follow a similar pattern, where text-only features achieve an accuracy of 0.67, and multimodal fusion enhances stability with an accuracy of 0.65 and an F-score of 0.64.

Overall, the experimental results clearly indicate that textual features play a dominant role in both binary and multiclass misinformation detection, while image-only features are insufficient when used independently, particularly in the multiclass scenario. However, the integration of text and image features consistently improves model robustness and generalization across classifiers. Among the evaluated models, gradient boosting with multimodal feature fusion yields the most consistent and competitive performance across both classification tasks, validating the effectiveness of the proposed multimodal framework for fine-grained misinformation detection. The comparative state-of-the-art analysis between existing large vision–language models and the proposed hybrid multimodal framework under both binary and multiclass misinformation classification configurations is shown in Table 7 below.

Table 7: State-of-the-Art Analysis

Ref Models	Overall Accuracy (%)	
	Binary Classification	Multiclass Classification
LLaVA-1.6-34B [35]	67.2	49.9
GPT-4V [35]	74.0	61.6
Proposed Hybrid Model	76.1	67.1

The LLaVA-1.6-34B model achieves an overall accuracy of 67.2% for binary classification and 49.9% for multiclass classification, indicating limitations in handling fine-grained

misinformation categories despite its strong general-purpose vision–language capabilities. GPT-4V demonstrates improved performance with 74.0% accuracy in the binary configuration and 61.6% in the multiclass scenario, reflecting its superior cross-modal reasoning ability. However, the proposed hybrid model outperforms both state-of-the-art baselines, achieving 76.1% accuracy for binary classification and 67.1% for multiclass classification. This performance gain highlights the effectiveness of task-specific multimodal feature fusion and tailored classification strategies over large general-purpose models. Notably, the improvement is more pronounced in the multiclass configuration, emphasizing the proposed model's stronger capability to distinguish between different misinformation types and capture nuanced cross-modal inconsistencies.

Overall, the experimental results across binary and multiclass evaluation settings demonstrate that textual features consistently provide the strongest discriminative capability for misinformation detection, while image-only representations show limited effectiveness, particularly in fine-grained multiclass classification. Multimodal fusion of text and image features improves robustness and stability in several cases; however, it does not uniformly outperform text-only models due to weak cross-modal alignment and the limited semantic contribution of visual content in certain samples. Among the evaluated classifiers, gradient boosting exhibits the most consistent performance across modalities and tasks, highlighting its effectiveness in handling heterogeneous feature spaces. These findings indicate that, while multimodal learning holds promise for misinformation detection, its effectiveness is highly dependent on data quality and fusion strategy. The results further reveal a key limitation of the current framework: naive or feature-level multimodal integration may introduce noise rather than complementary information when visual cues are weak or inconsistent. This underscores the need for more advanced cross-modal alignment and adaptive fusion mechanisms to fully exploit multimodal data in real-world misinformation detection scenarios.

5. CONCLUSION

This paper proposed a multimodal misinformation detection framework to address both binary and multiclass classification tasks. By integrating textual and visual representations and analyzing cross-modal consistency, the framework aims to capture complementary information across modalities. Experimental analysis indicates that textual features provide the strongest and most consistent discriminative signals across both evaluation configurations, while

multimodal fusion improves robustness in certain scenarios but remains sensitive to cross-modal noise and alignment quality. The inclusion of fine-grained misinformation categories, textual veracity distortion, visual veracity distortion, and cross-modal consistency distortion, facilitates a deeper understanding of misinformation characteristics and highlights the varying contribution of each modality. Overall, the results confirm that effective misinformation detection in multimodal configurations requires not only feature integration but also careful modeling of modality relevance and interaction.

Future work will focus on improving multimodal effectiveness by enhancing cross-modal alignment, refining visual feature relevance, and adopting advanced vision–language models with adaptive fusion mechanisms, as well as extending the framework to multilingual and cross-domain misinformation datasets.

Acknowledgement

The authors would like to express their sincere gratitude to all individuals and institutions that indirectly supported this research and contributed to the successful completion of this study.

Funding

The authors declare that this study was carried out without any financial assistance or funding from external sources.

Conflict of Interest

The authors declare that there are no conflicts of interest associated with this research.

Declaration of Artificial Intelligence (AI) Assistance

The authors confirm that generative AI tools were used only for language refinement and grammatical correction. No AI-generated content, analysis, or interpretation contributed to the work. All research design, data processing, and manuscript preparation were carried out independently by the authors without AI involvement.

Author Contributions

All authors contributed equally to this work and participated in every stage of the research. All authors have read and approved the final version of the manuscript.

Ethics approval

This study utilized publicly available multimodal misinformation datasets that are fully anonymized and released for research purposes. No human subjects were directly involved, and no personally identifiable information was accessed. Therefore, additional ethical approval was not required for this research.

Data availability

The datasets used in this study are publicly available from their respective repositories, including the benchmark multimodal misinformation dataset, MMFakeBench.

Abbreviation

EffNet-B0: EfficientNet-B0, VLM: Vision–Language Model, BERT: Bidirectional Encoder Representations from Transformers, GNN: Graph Neural Network, LLM: Large Language Model, CNN: Convolutional Neural Network, Ensemble: Stack Ensemble Learning, MIS: Misinformation Detection System.

Reference

1. J. Lv, Y. Gao, L. Li, et al., “Multi-modal fake news detection: A comprehensive survey on deep learning technology, advances, and challenges,” J. King Saud Univ. Comput. Inf. Sci., vol. 37, art. no. 306, 2025, doi: 10.1007/s44443-025-00317-7.
2. S. Abdali, S. Shaham, and B. Krishnamachari, “Multi-modal misinformation detection: Approaches, challenges and opportunities,” ACM Comput. Surv., vol. 57, no. 3, art. no. 76, pp. 1–29, Mar. 2025, doi: 10.1145/3697349.
3. K. Singh, R. G. Prasad, and S. K. Dwivedi, “Cross-platform multimodal misinformation: Taxonomy, characteristics and detection for textual posts and videos,” Multimedia Syst., vol. 30, no. 2, pp. 495–510, Apr. 2024, doi: 10.1007/s00530-023-00999-w.
4. S. A. Chowdhury, M. Alsmadi, and H. A. Jalab, “Fake news, disinformation and misinformation in social media: A review,” Soc. Netw. Anal. Min., vol. 13, no. 1, art. no. 20, Jan. 2023, doi: 10.1007/s13278-023-00965-1.
5. M. Himelein-Wachowiak, et al., “Bots and misinformation spread on social media: A mixed scoping review with implications for COVID-19,” J. Med. Internet Res., vol. 23, no. 5, Jan. 2021, doi: 10.2196/26933.
6. M. R. Islam, S. Liu, and X. Wang, et al., “Deep learning for misinformation detection on online social networks: A survey and new perspectives,” Soc. Netw. Anal. Min., vol. 10, art. no. 82, 2020, doi: 10.1007/s13278-020-00696-x.
7. M. Abdullah, et al., “A multimodal ensemble-based framework for detecting fake news using visual and textual features,” Mathematics, vol. 14, no. 2, art. no. 360, 2026, doi: 10.3390/math14020360.
8. S. Sharifpoor, M. Okhovati, B. Ghatee, et al., “Classifying and fact-checking health-related information about COVID-19 on Twitter/X using machine learning and deep learning models,” BMC Med. Inform. Decis. Mak., vol. 25, art. no. 73, Feb. 2025, doi: 10.1186/s12911-025-02895-y.
9. S. Duwal, et al., “Evidence-grounded multimodal misinformation detection with attention-based GNNs,” arXiv, art. no. arXiv:2505.18221, 2025. [Online]. Available: https://arxiv.org/abs/2505.18221
10. K. Xu, et al., “DATTAMM: Domain-aware test-time adaptation for multimodal misinformation detection,” Appl. Sci., vol. 15, no. 21, art. no. 11832, 2025, doi: 10.3390/app152111832.

11. H. R. LekshmiAmmal and A. K. Madasamy, “A reasoning-based explainable multimodal fake news detection for low-resource language using large language models and transformers,” J. Big Data, vol. 12, art. no. 46, 2025, doi: 10.1186/s40537-025-01093-x.
12. M. G. Karthik, et al., “Hybrid optimization driven fake news detection using reinforced transformer models,” Sci. Rep., vol. 15, no. 1, art. no. 14782, Apr. 2025, doi: 10.1038/s41598-025-99936-3.
13. S. Cui, H. Wang, C.-C. Chang, H. H. Nguyen, and I. Echizen, “A multilingual, multimodal dataset for disinformation and out-of-context analysis with rich supportive information,” in Proc. 27th Int. Conf. Multimodal Interaction (ICMI), New York, NY, USA: ACM, 2025, pp. 643–651, doi: 10.1145/3716553.3750813.
14. B. Yang, et al., “A new dataset and benchmark for grounding multimodal misinformation,” arXiv, 2025. doi: 10.1145/3746027.3758191.
15. G. K. Shahi, “Multimodal misinformation detection using early fusion of linguistic, visual, and social features,” in Companion Proc. 17th ACM Web Science Conf. (WebSci Companion), New York, NY, USA: ACM, 2025, pp. 11–18, doi: 10.1145/3720554.3733844.
16. Q. Xu, H. Du, S. Łukasik, T. Zhu, S. Wang, and X. Yu, “MDAM 3: A misinformation detection and analysis framework for multitype multimodal media,” in Proc. ACM Web Conf. (WWW), Sydney, NSW, Australia, 2025, New York, NY, USA: ACM, doi: 10.1145/3696410.3714498.
17. S. Zhang, et al., “Multimodal misinformation detection based on the multi-granularity consistency,” Neurocomputing, vol. 644, art. no. 130393, 2025, doi: 10.1016/j.neucom.2025.130393.
18. H. Li, et al., “Towards unified multimodal misinformation detection in social media: A benchmark dataset and baseline,” arXiv, 2025. [Online]. Available: https://arxiv.org/abs/2509.25991
19. Q. Xu, et al., “M3A: A multimodal misinformation dataset for media authenticity analysis,” Comput. Vis. Image Understand., vol. 249, art. no. 104205, 2024, doi: 10.1016/j.cviu.2024.104205.
20. S. Tahmasebi, et al., “Multimodal misinformation detection using large vision-language models,” arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2407.14321
21. C. Toraman, O. Özçelik, F. Şahinuç, and F. Can, “MiDe22: An annotated multi-event tweet dataset for misinformation detection,” in Proc. Joint Int. Conf. Comput. Linguistics and Language Resources and Evaluation (LREC–COLING), Torino, Italy, May 2024, pp. 11283–11295.
22. S. Mehreen, A. Bhuiyan, and E. Hossain, “An attentive fusion framework for multimodal misinformation detection,” in Proc. 27th Int. Conf. Comput. Inf. Technol. (ICCIT), Cox’s Bazar, Bangladesh, 2024, pp. 1903–1908, doi: 10.1109/ICCIT64611.2024.11022377.
23. Y. Liu, M. D. Ma, W. Qin, A. Zhou, J. Chen, W. Shi, W. Wang, and D. Yang, “Decoding susceptibility: Modeling misbelief to misinformation through a computational approach,” in Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP), Miami, FL, USA, Nov. 2024, pp. 15178–15194, doi: 10.18653/v1/2024.emnlp-main.846.
24. X. Liu, et al., “FKA-Owl: Advancing multimodal fake news detection through knowledge-augmented LVLMs,” arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2403.01988
25. C. F. Luo, R. Shayanfar, R. V. Bhambhoria, S. Dahan, and X. Zhu, “Misinformation with legal consequences (MisLC): A new task towards harnessing societal harm of misinformation,” in Findings Assoc. Comput. Linguistics: EMNLP, Miami, FL, USA, Nov. 2024, pp. 15749–15768, doi: 10.18653/v1/2024.findings-emnlp.924.
26. S. I. Papadopoulos, C. Koutlis, S. Papadopoulos, et al., “VERITE: A robust benchmark for multimodal misinformation detection accounting for unimodal bias,” Int. J. Multimedia Inf. Retr., vol. 13, no. 4, 2024, doi: 10.1007/s13735-023-00312-6.
27. Indu and S. M. Thampi, “Misinformation detection in social networks using emotion analysis and user behaviour analysis,” Pattern Recognit. Lett., vol. 182, pp. 60–66, 2024, doi: 10.1016/j.patrec.2024.04.007.
28. F. Zeng, W. Li, W. Gao, and Y. Pang, “Multimodal misinformation detection by learning from synthetic data with multimodal LLMs,” in Findings Assoc. Comput. Linguistics: EMNLP 2024, Miami, FL, USA, Nov. 2024, pp. 10467–10484, doi: 10.18653/v1/2024.findings-emnlp.613.
29. J. Haber, K. Kawintiranon, L. Singh, A. Chen, A. Pizzo, A. Pogrebivsky, and J. Yang, “Identifying high-quality training data for misinformation detection,” in Proc. 12th Int. Conf. Data Sci., Technol. Appl. (DATA), Rome, Italy, Jul. 2023, pp. 64–76, doi: 10.5220/0012089000003541.
30. Y. Guo, H. Ge, and J. Li, “A two-branch multimodal fake news detection model based on multimodal bilinear pooling and attention mechanism,” Front. Comput. Sci., vol. 5, art. no. 1159063, 2023, doi: 10.3389/fcomp.2023.1159063.
31. M. Asfand e Yar, Q. Hashir, S. H. Tanvir, and W. Khalil, “Classifying misinformation of user credibility in social media using supervised learning,” Comput. Mater. Continua, vol. 75, no. 2, pp. 2921–2938, 2023, doi: 10.32604/cmc.2023.034741.
32. G. Joshi, A. Srivastava, B. Yagnik, M. Hasan, Z. Saiyed, L. Gabralla, A. Abraham, R. Walambe, and K. Kotecha, “Explainable misinformation detection across multiple social media platforms,” arXiv, vol. abs/2203.11724, 2022. [Online]. Available: https://arxiv.org/abs/2203.11724

33. Hassan, M. N. L. Azmi, and A. M. Abdullahi, “Evaluating the spread of fake news and its detection techniques on social networking sites,” Romanian J. Commun. Public Relat., vol. 22, no. 1, pp. 111–125, Apr. 2020, doi: 10.21018/rjcpr.2020.1.289.
34. S. Tyagi, A. Pai, J. Pegado, and A. Kamath, “A proposed model for preventing the spread of misinformation on online social media using machine learning,” in Proc. Amity Int. Conf. Artif. Intell. (AICAI), Feb. 2019, doi: 10.1109/AICAI.2019.8701408.
35. X. Liu et al., “MMFakeBench: A mixed-source multimodal misinformation detection benchmark for LVLMs,” arXiv preprint arXiv:2406.08772, 2024. [Online]. Available: https://arxiv.org/abs/2406.08772