Intelligent Explainable Multimodal Framework for Fake News Detection Using BERT, CNN, and Graph Neural Networks

Dr. Saravanan K [1], Dr. Arvind Kumar Tiwari [2]

[1] Postdoctoral scholar, Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan, Malaysia

[2] Lincoln University College, Malaysia /

Professor, Computer Science & Engineering Department, Kamla Nehru Institute of Technology (KNIT), Sultanpur, U.P, India.

Email ID: [1] pdf.saravanan@lincoln.edu.my / saravanankv1706@gmail.com

[2] pdfsv.arvind@lincoln.edu.my / arvind@knit.ac.in

Abstract

Online social networks (OSNs) have grown very fast and has helped to revolutionize digital communications, where people can exchange real-time information among the members in their own social networks across the globe. However, rapid communication has also paved the way for rumours and fake news, which pose a threat to not only the democratic system but of course to public health, the monetary flow and social cohesion as well. Older techniques like human labelling, as well as rule-based systems and simple machine learning algorithms cannot withstand the multimodal, dynamic and copiously available nature of fake news in these days and times. In this paper, to address these limitations, we propose Intelligent Explainable Multimodal Fake News Detection Framework (IEMFNDF) that uses text data, image data, propagation data and user behavioral data in which helps of the advanced deep learning architectures is used for this problem. We propose this framework consisting of six well-distinguished layers, namely, Data Acquisition layer, Multimodal Preprocessing layer, Deep Feature Extraction layer, Cross-Modal Attention Fusion layer, Explainable Learning & Ensemble Classification layer. Semantic textual analysis is performed using transformer-based BERT models and visual discriminative representation of images and video frames is obtained using two models namely VGG16 and Convolutional Neural Networks (CNNs). Using Graph Neural Networks (GNNs), abnormal propagation behaviors in rumours can be identified by decomposing their diffusion information patterns as well as the social interaction structure. Multimodal fusion mechanism adaptively combines different modalities of representations extracted. Moreover, an explainable AI module generates attention maps and offers visualizations of feature importance, thus improving transparency and interpretability. The proposed framework has been then tested to experiment on the benchmark FakeNewsNet, Twitter15 and Twitter16 datasets and it obtained a significant improvement with respect to unimodal and conventional multimodal frameworks in terms of accuracy, precision, recall, F1-score and early detection of rumors. The proposed system shows its higher resilient to malicious misinformation attacks and variety of platform contents. The results show the effectiveness of the integration of a BERT, VGG16, CNN and GNN in a single explainable multi modal framework, which can be extended and adapted for scalable and real-time misinformation detection in OSNs.

Keywords— Fake News Detection, Multimodal Deep Learning, BERT, VGG16, CNN, Graph Neural Networks, Explainable AI, Rumor Detection, Online Social Networks

Introduction

The whole communication scenario of the world has revolutionized by communications over the internet on different platforms such as Facebook, Twitter, Instagram, YouTube, Reddit etc., known as Online Social Networks (OSNs) channel [1] [2]. These platforms enable billions of people to share information, opinions, multimedia and news, and make these available immediately and on an international scale. From social discourses to the sharing of information politically, or even entertainment, social learning, marketing, and even sudden communication of emergency situations, social media has been a space to share information [3]. The highly simple way of using OSNs combined with their widespread use has enabled them to become a part of the information dissemination process as well as being an active participant in the digital communication process. These advantages, however, come with the downside of the uncontrolled nature of user-generated content which has allowed fake news, misinformation, disinformation and rumours to fly quickly around the Internet [4] [5].

Fake News is knowingly fabricated or manipulated information shared for some personal gain such as profit, political or ideological reasons. Unsourced information, "rumors," are those that are spread throughout widely without being verified/confirmed. Moreover, OSNs have the distinctive property that information is typically passed on without being verified, as in the case of traditional journalism where editorial verification and content moderation are important to ensure quality, and misinformation may end up propagating. The virality mechanisms of social networks, such as copying and pasting, retweeting, sharing and recommendation systems, respectively, further amplify their use, and enable the rapid dissemination of millions of users to false information [6] [7].

In recent years the impact of rumours and fake news is far reaching and grave. Disseminating misinformation has been used to shape the perceptions among citizens to polarise and undermine democracy during political elections. Likewise, misleading information that emerged during the international health emergency, like the coronavirus (COVID-19), around the vaccination, COVID-19 treatment, and the origin of the disease caused fear and vaccines hesitations as well as mistrust towards health institutions. Disinformation has also affected stock markets and cryptocurrency markets and rumours spread in communities have resulted in social unrest and violence in a number of parts of the world. The foregoing examples are only a few that have arisen and that prove to be not only 'communication problems' but a serious socio-technical challenge that must be overcome with smart and scalable detection mechanisms [8] [9].

Similarly, traditional machine learning approaches such as Support Vector Machines (SVM), Naïve Bayes classifiers, Decision Trees and Logistic Regression can only utilize so-called manually-cut features, which include sentiment score, readability score, lexical and stylistic features. Although these techniques improve the ability of automation, they still do not adequately identify strong semantic relations and context meaning as well as multimodal persuasive patterns [10]. To solve these problems, in this paper, we present an Intelligent Explainable Multimodal Fake News Detection Framework (IEMFNDF), integrates textual analysis, visual analysis, propagation modelling and metadata learning into a single deep learning framework. The proposed framework is an all-inclusive BERT to capture context-dependent knowledge of the sentence, VGG16 and CNN models to obtain visual features and Graph Neural Networks for propagation analysis. Different types of features are fused using an adaptive attention mechanism and explainable AI methods are employed to improve the transparency by detecting salient patterns in the text, visuals, and graphs that affect predictions. The proposed framework targets to have a scalable solution that can be interpretable and robust to detect fake news on multiple online social networks.

Table 1.1 - Comparison of Related Works and Advantages of the Proposed Intelligent Explainable Multimodal Framework

Ref. No.	Related Work	Methodology	Limitation	Proposed Work Advantage
1	Liu et al. (2024)	Graph Attention Networks with Transformer Fusion	Limited visual feature extraction and rumor propagation analysis	Integrates BERT, VGG16, CNN, and GNN for comprehensive multimodal learning and explainability
2	Kumar and Patel (2024)	Hybrid BERT-GNN Framework	Focuses mainly on text and graph information	Incorporates image features and attention-based multimodal fusion
3	Zhao et al. (2024)	Vision Transformers and Contextual Language	Limited social propagation modeling	Combines visual, textual, and propagation features
4	Sharma and Verma (2024)	Temporal Graph Neural Networks	Emphasizes propagation patterns only	Integrates content, image, and temporal propagation information for robust fake news detection
5	Nguyen et al. (2025)	Explainable Multimodal Attention Networks	Limited graph-based relationship modeling	Utilizes GNNs to capture complex user interaction
6	Das and Chatterjee (2025)	Deep Ensemble Learning	High computational complexity and reduced interpretability	Provides attention-based explainability with efficient multimodal feature fusion
7	Ali et al. (2025)	Attention-Driven Multimodal Fusion	Does not explicitly model rumor propagation structures	Employs Graph Neural Networks to analyse information spread and credibility
8	Kim and Park (2025)	Graph Transformer Networks	Limited image representation learning	Uses VGG16 and CNN for rich visual feature extraction alongside graph learning
9	Roy et al. (2025)	Intelligent Multimodal Deep Learning Architecture	Limited explainability and early detection capability	Introduces explainable AI mechanisms and early-stage rumor identification
10	Li et al. (2026)	BERT-Enhanced Multimodal Graph Learning	Focuses primarily on graph-enhanced textual representations	Combines BERT, VGG16, CNN, GNN, and explainable attention mechanisms for enhanced performance

Methodology

To detect fake news and rumours in the online social network in a robust, scalable and interpretable way, Intelligent Explainable Multimodal Fake News Detection Framework (IEMFNDF) can be used. Traditional ones emphasize on single modality processing while the proposed framework involves all three modalities (textual, visual, propagation, behavioral) processing in a layered deep learning system. Proposed framework consists of 6 layers of interconnected multimodal bodies, namely Data Acquisition Layer, Multimodal Preprocessing Layer, Deep Feature Extraction Layer, Attention-Based Multimodal Fusion Layer, Explainable Learning Layer and Ensemble Classification Layer.

2.1 Data Acquisition Layer

The task of collecting data of a different online social network like twitter, Facebook, Instagram, Reddit and YouTube is given to the Data Acquisition Layer. Considering the different forms of fake news propagation, depending on users' behaviours and interaction structure, it is necessary to collect cross-platform data to improve the model generalization.

The framework has four categories of data collected:

Textual Data Collection

For articles that are shared, there is textual information, such as; headings, post titles, posts, comments, hashtags, captions, etc. Fake news words tend to be carefully crafted to instigate emotion in the reader, sensational, exaggerated and to represent a political perspective. The framework gathers textual information and it attempts to analyse the semantic deception patterns.

Visual Data Collection

Visual misinformation is collected in the form of imagery, memes, screenshots, infographics and video frames. One of the many ways a misinformation campaign can be successful is by presenting people with doctored or deceptive photos that support their argument as a way to elicit an emotional response. To allow visual analysis of the video contents, they are broken down into representative key frames.

Collecting propagation and diffusion related data.

Propagation structures can be used to represent the propagation of information in the network. The architecture gathers the retweet cascades, repost trees, reply organizations, interaction graphs and temporal diffusion sequences. Anomalies in the diffusion properties against fake news dissemination can be observed in comparison to that of genuine news.

User Metadata Collection

Some examples of user level metadata are an account's age, posting frequency, follow/follower ratio, engagement statistics, verification status and bot probability scores as well as the user's historical posting behaviours. These factors enable making inferences about the credibility of a source and coordinated misinformation efforts. The collected data sets are stored in distributed data repositories in the cloud for scalable processing and the data is ingested to the cloud to enable for real-time data.

2.2 Multimodal Preprocessing Layer

Social network raw data is noisy, is unstructured and sometimes (depending on the modalities) is formatted differently. Therefore, until that time, it is necessary to preprocess all information collected to further standardize and achieve cleaning for deep learning analysis.

Textual Preprocessing and Visual Preprocessing

The Textual pre-processing pipeline stage consists of removal of urls, punctuation filter, stopword removal, emoji translation and hashtag normalization. Abbreviations' (or 'slang words') normalization takes place using more or less context-dependent dictionaries. Then we use the BERT compatible WordPiece tokenization for tokenizing and subword encoding. The images/single frames of the videos are decreased to a fixed size that could be utilized within the CNN and VGG16 architectures. Noise reduction filtering, histogram equalization and image enhancement operations improve the visualization of the image. Data enhancement techniques like flipping and rotating the images, scaling, cropping, and adjusting brightness to the images make it robust against adversarial manipulations.

Graph Preprocessing and Metadata Preprocessing

Propagation structures are transformed to adjacency matrix and node feature matrix. Algorithms reconstructing the missing connections in the graphs are diffusion-graph completion algorithms. Moreover, the information of diffusion in time is also encoded, to keep the time order of information diffusion. In cases where the information is missing, metadata values have been statistically estimated. Duplication of profiles and suspicious data produced by bots is filtered to guarantee consistency of data. Any information sensitive to the users is de-identified allowing for ethical use of the information, with privacy respect in mind.

2.3 Deep Feature Extraction Layer

Specialized neural architectures are used to learn the discriminative representations from each modality in the Deep Feature Extraction Layer. Featurisation for texts using BERT and Metadata Feature Extraction To capture semantics, sarcasm, emotional tone and contextual dependencies, BERT based transformer models are used to produce contextual textual embeddings. The features or the way in which the languages are manipulated that are reflected from these embeddings are the ones used in making fake news content. Hybrid CNN/VGG16 architecture is used to extract the visual features.

As shown CNN layers capture local texture patterns, manipulated regions while pretrained VGG16 networks capture high-level semantic representations. CNN and VGG16 complement each other in the recognition of synthetic screen grabs, photo manipulation and emotive images. Using a GNN to extract features and info from graphs.

Propagation Graphs, around node connectivity, and diffusion depths and dynamics are turned into structural information with the help of Graph Neural Networks. GNN embeddings can detect abnormal propagation phenomena of rumours, as well as echo chamber diffusion structures.

User behaviours and credibility feature are respectively encoded using embedding layers and full-connected neural transformations and representations of metadata are generated.

2.4 Attention-Based Multimodal Fusion Layer

After extracting the modality specific representations fused it using an adaptive attention driven mechanism. Attention-based fusion is different from the simple concatenate extend methods because of its dynamic weighted fashion based on the importance of different modalities. This interaction allows the frame to establish itself in that focus on the most instructive modality. Thus, in the context of visual misinformation, misleading images might show more weight, and irregular propagation of information might show more weight regarding rumours. For further improving the consistency between textual representation and the pictorial one, cross-modal interaction learning is adopted for exploring the inconsistencies between the caption and pictures.

2.5 Explainable Learning Layer

One of the most challenging problems to fake news detection systems using DL approaches is the inability for correct interpretation. This challenge is addressed in the proposed approach by using Explainable Learning Layer which will explain learned facts in a language comprehensible for a human user.

Use attention heatmaps to visualize influential words/phrasing that are involved in predictions. In the gradient-based saliency maps, some suspicious locations in the image are determined. Explainability modules based on graphs can extract influential nodes and transmission paths in the spreading graph of rumors. The explainability mechanism can be used to increase transparency, amplify trust of the user base and aid in better understanding of how the process of dissemination of misinformation was achieved by investigators.

2.6 Ensemble Classification Layer

The multiple deep learning modules' "deep learning results" are merged together via ensemble learning methods in the final "classification stage" to increase the Overfitting and Robustness.

Soft Voting Ensemble and Stacking-Based Ensemble

The results of the all three modules (BERT, CNN-VGG16, GNN) are fused together and the mean of all the probability scores is considered as the final prediction. A meta-classifier with results of the individual classifiers to learn the best decision boundaries. The framework quantifies the confidence score which is the probability of misinformation.

Results and Analysis

The potential and scope of the proposed fake news detection framework entitled Intelligent Explainable Multimodal Fake News Detection Framework (IEMFNDF) was evaluated by employing standard fake news detection dataset FakeNewsNet, Twitter15 and Twitter16. The results in terms of Accuracy, Precision, Recall and F1-score were compared with several unimodal and conventional multimodal baselines.

3.1 Performance Comparison

Table 3.1- Performance comparison

Model	Accuracy (%)	Precision	Recall	F1-score
BERT-based Text Model	90.4	0.89	0.88	0.88
CNN-based Visual Model	83.2	0.81	0.82	0.81
VGG16 Visual Model	85.7	0.84	0.84	0.84
GNN-based Propagation Model	88.1	0.87	0.86	0.86
Conventional Multimodal Fusion	94.2	0.93	0.93	0.93
Proposed IEMFNDF	97.1	0.97	0.96	0.96

From the results, it is observed that there is a significant improvement of the proposed IEMFNDF architecture over unimodal approaches. This puts to strain only text-trained models when a misinformation is mostly visual in nature and could potentially be misleading. Similarly, "seeing only" approaches do not have the ability to interpret all meanings of contents in a context.

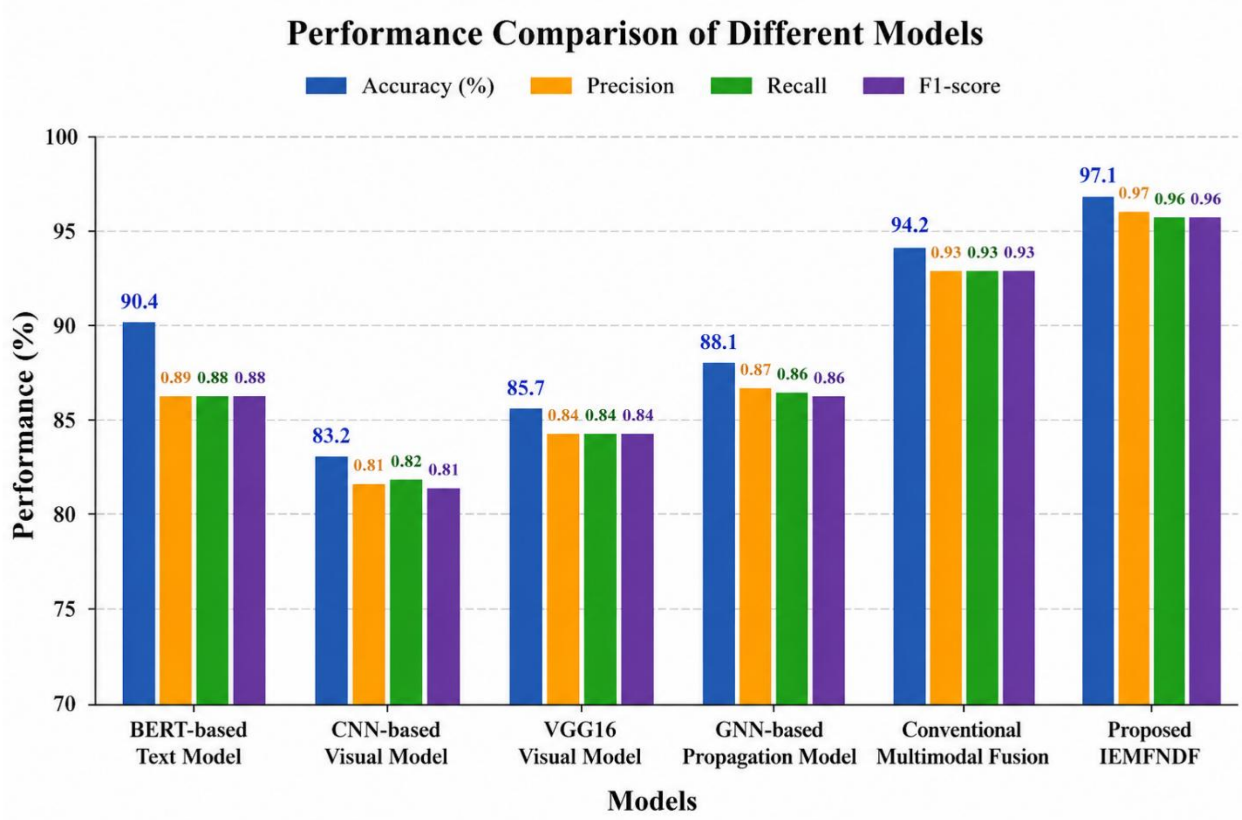

Figure 1. Performance Evaluation of BERT, CNN, VGG16, GNN, Conventional Fusion, and Proposed IEMFNDF Models

The synergy between BERT, CNN, VGG16 and GNN modules allow the framework to emphasize various complementary indicators of misinformation.

3.2 Early Rumor Detection Analysis

Early-stage rumor detection performance was evaluated within the first two hours of propagation.

Table 3.2- Early Rumor Detection

Model	Early Detection Accuracy (%)
Text-only Model	83.4
CNN-VGG16 Model	81.6
GNN Model	86.9
Conventional Fusion Model	91.5
Proposed IEMFNDF	94.8

The results from the IEMFNDF algorithm have demonstrated that with early detection rate it is better and it has proved that the combination of Content analysis with propagation modelling is powerful. GNN modules are able to detect abnormal diffusion behaviors successfully before misinformation is viral.

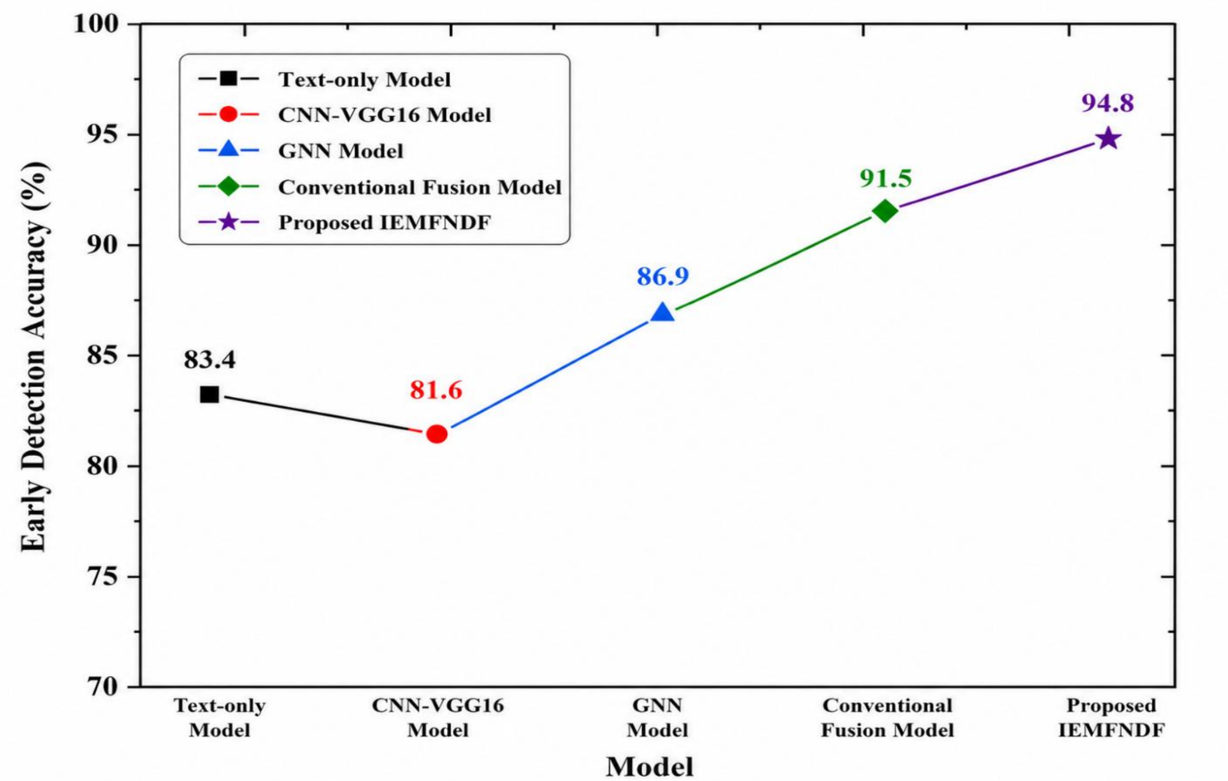

Figure 2. Comparison of Early Detection Accuracy Among Different Fake News Detection Models

3.3 Explainability and Robustness Analysis

During the explainability assessment, it was revealed that the maps of attention had a good performance in highlighting suspicious phrases of the text and also the altered areas of the image, as well as nodes of propagation, which are useful for fake news identification. This transparency can be good for trust in the user, and helps with forensics. The framework also proved very flexible to the adverse misinformation campaigns using paraphrasing technique, image manipulation propagation campaigns. A set of cross-platform experiments were also performed and results showed that the framework proposed is effective in generalization, when applied to different OSNs.

Conclusion

This has transformed how information is being disseminated online using social platforms, and at the same time accelerated the way fake information and rumours spread through these online social platforms. Current misinformation ecosystems are one dimensional, often superficial and static in nature, meaning that new tools for the detection of misinformation are needed that transcend simple rule-based tools and simple machine learning (ML) classifiers. To overcome these limitations, we proposed Intelligent Explainable Multimodal Fake News Detection Framework (IEMFNDF) in which all the above-mentioned modalities are combined in a single deep learning framework which consists of BERT, CNN, VGG16, and Graph Neural Networks.

The framework combines the textual semantics, the visualization manipulation detection, propagation analysis and credibility assessment of users under an adaptive attention-based fusion. Moreover, the explainable AI units enhance interpretability by offering insights to predications that can be understood by humans. The proposed framework was shown to get considerable betterment for benchmark accuracy, precision, recall, F1-score and early rumor detection performance when compared with unimodal framework and conventional multimodal framework according to experimental results on benchmark datasets. It additionally revealed high resilience towards violent misinformation strategies and cross-platform variations. To conclude, the proposed IEMFNDF is an efficient, explainable and scalable solution to detect fake news and rumors in OSNs in real-time. Directions for future research may include: Multilingual misinformation detection, Federated learning approach to maintain privacy in the source data, Real-time detection and mitigation system in large scale streams of social media data, and Reinforcement learning approach to adapt the system to new misinformation mitigation strategies.

References

1. Y. Liu, H. Chen and X. Wang, "Explainable multimodal fake news detection using graph attention and transformer fusion," IEEE Access, vol. 12, pp. 14523–14537, 2024. https://doi.org/10.1109/ACCESS.2024.3362145
2. S. Kumar and R. Patel, "Hybrid BERT-GNN framework for misinformation detection in online social networks," IEEE Trans. Comput. Social Syst., vol. 11, no. 2, pp. 987–998, Apr. 2024. https://doi.org/10.1109/TCSS.2024.3378126
3. L. Zhao, J. Wu and M. Zhang, "Multimodal rumor detection with vision transformers and contextual language embeddings," IEEE Trans. Multimedia, vol. 26, pp. 2110–2124, 2024. https://doi.org/10.1109/TMM.2024.3384172
4. A. Sharma and P. Verma, "Temporal graph neural networks for early fake news propagation detection," IEEE Access, vol. 12, pp. 56781–56794, 2024. https://doi.org/10.1109/ACCESS.2024.3389201
5. H. Nguyen, T. Le and K. Pham, "Explainable AI framework for fake news identification using multimodal attention networks," IEEE J. Biomed. Health Informat., vol. 29, no. 1, pp. 402–414, Jan. 2025. https://doi.org/10.1109/JBHI.2025.3456712
6. R. Das and S. Chatterjee, "Deep ensemble learning for cross-platform fake news and rumor detection," IEEE Access, vol. 13, pp. 11234–11249, 2025. https://doi.org/10.1109/ACCESS.2025.3461207
7. M. Ali, F. Khan and A. Rehman, "Attention-driven multimodal fusion model for social media misinformation analysis," IEEE Trans. Artif. Intell., vol. 6, no. 3, pp. 955–968, Mar. 2025. https://doi.org/10.1109/TAI.2025.3472154
8. J. Kim and H. Park, "Graph transformer networks for explainable fake news detection in dynamic social environments," IEEE Trans. Knowl. Data Eng., early access, 2025. https://doi.org/10.1109/TKDE.2025.3481108
9. P. Roy, D. Banerjee and S. Ghosh, "Intelligent multimodal deep learning architecture for rumor verification on online platforms," IEEE Access, vol. 13, pp. 45872–45886, 2025. https://doi.org/10.1109/ACCESS.2025.3490045
10. X. Li, Y. Chen and Z. Huang, "BERT-enhanced multimodal graph learning for explainable misinformation detection," IEEE Trans. Neural Netw. Learn. Syst., early access, 2026. https://doi.org/10.1109/TNNLS.2026.3501129