International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2(2) (2026)



ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

MULTIMODAL MISINFORMATION AND DEEPFAKE DETECTION SYSTEM

Ruthala Sridivya[1], Ryali Sree Sai Bhavitha Sumali[1], Sai Gayatri Yellajosyula[1], Sathivada Gayathri[1], Sattaru Pallavi[1]

[1]*Department of Computer Science & Systems Engineering, Andhra University, Visakhapatnam, India*

{ruthalasridivya, sumaliryali, gayatriyellajosyula, sathivadagayathri2627, pallavi.sattaru154}@gmail.com

Guide: Dr. G. Sharmila Sujatha,Assistant Professor, Dept. of CS&SE, Andhra University,, gshrml@gmail.com

Abstract

The exponential growth of social media platforms has significantly accelerated the spread of misinformation across diverse content formats, including textual narratives, images, memes, and videos. Conventional fake news detection approaches remain insufficient as they are predominantly designed for textual data, failing to address the complexities introduced by multimodal content and sophisticated deepfake media. This work proposes a comprehensive artificial intelligence-driven system for detecting misinformation and identifying deepfake content within a unified architecture. The system incorporates Natural Language Processing (NLP), Computer Vision (CV), Convolutional Neural Networks (CNN), and Retrieval-Augmented Generation (RAG) to deliver accurate, interpretable, and context-aware outcomes. Textual inputs are processed using advanced NLP techniques to extract key claims, while Optical Character Recognition (OCR) retrieves embedded textual information from images and memes for multimodal fusion. Video inputs employ frame-based CNN analysis, with frame-level predictions aggregated into a final classification. The system produces conclusive output categorizing content as True, False, or Misleading, accompanied by a confidence score and evidence-backed explanation, enhancing user trust and suitability for real-world applications.

Index Terms— Deepfake Detection, Multimodal Learning, Misinformation, Retrieval-Augmented Generation, Convolutional Neural Networks, Fact-Checking.

I. Introduction

The rapid advancement of digital communication technologies and widespread adoption of social media platforms have transformed how information is created, shared, and consumed. While these platforms enable instantaneous access to global information, they have simultaneously facilitated the uncontrolled spread of misinformation [1]. This phenomenon has evolved beyond simple text-based falsehoods into complex, multimodal forms—including images, memes, and videos—often engineered to manipulate public perception.

Memes combine textual and visual elements to convey misleading narratives in subtle and often humorous manners, making them difficult to interpret using conventional text-based techniques. Advances in artificial intelligence have further enabled the creation of deepfake media—highly realistic but fabricated images and videos that can misrepresent individuals and events—raising serious concerns regarding digital content authenticity and its impact on political stability and social harmony [2].

Existing detection mechanisms primarily focus on single modalities, particularly textual data, utilizing methods such as statistical analysis,

machine learning classifiers, and rule-based approaches [3]. Although these methods demonstrate reasonable performance in controlled environments, they lack the capability to process and correlate information across multiple modalities. Furthermore, many current systems do not provide explainable outputs, limiting user confidence in sensitive domains such as journalism, law enforcement, and public policy.

The emergence of multimodal learning and hybrid AI techniques offers promising solutions. By combining NLP for textual understanding, Computer Vision for image and video analysis, and advanced neural architectures such as CNNs, it becomes possible to capture intricate patterns and inconsistencies across different data types [4]. Additionally, the integration of Retrieval-Augmented Generation (RAG) enables systems to validate claims against reliable external sources, enhancing accuracy and supporting generation of evidence-based explanations [5].

This paper presents a Multimodal Misinformation and Deepfake Detection System designed to overcome the limitations of existing approaches through a unified and scalable framework. The system is capable of processing text, images, memes, and videos within a single platform, employing specialized modules for each modality while maintaining a cohesive analytical pipeline.

II. Related Work

Text-based fake news detection has evolved from traditional machine learning algorithms such as Support Vector Machines, Naïve Bayes classifiers, and logistic regression to deep learning approaches [3]. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks enabled sequential text analysis, better capturing context and temporal dependencies. More recently, transformer-based architectures such as BERT [6] and RoBERTa have revolutionized the field through attention mechanisms that capture long-range dependencies and semantic nuances.

Image-based misinformation detection employs CNNs for identifying manipulated or misleading visual content, with networks such as VGGNet [8], ResNet, and specialized architectures like MesoNet [4] demonstrating superior performance in detecting manipulation artifacts. Error Level Analysis (ELA) and copy-move detection represent earlier forensic approaches, though limited in scalability [9].

Deepfake detection has progressed from identifying basic visual inconsistencies to sophisticated CNN and RNN-based architectures. Afchar et al. introduced MesoNet [4], a compact network focusing on mesoscopic facial features. Rossler et al. presented FaceForensics++ [16], a large-scale benchmark, while Chollet's XceptionNet [17] has been widely adopted for high-accuracy deepfake detection.

Multimodal learning approaches integrate information from diverse sources. Feature fusion techniques—including concatenation, attention mechanisms, and neural fusion layers—have shown improved performance in meme classification and multimodal sentiment analysis [10]. The introduction of RAG by Lewis et al. [5] represents a significant advancement, combining retrieval mechanisms with generative models to produce responses grounded in external evidence, addressing the limitation of purely generative models that lack verifiable reasoning.

Despite progress, there remains a lack of unified systems seamlessly integrating text, image, and video analysis with evidence-based reasoning. The proposed system addresses these gaps by combining multimodal analysis, deepfake detection, and retrieval-based verification into a cohesive, scalable solution.

III. Methodology / System Design

A. System Architecture

The proposed system adopts a layered, modular architecture comprising four core layers: the Presentation Layer (frontend), Application Layer

(backend API), Processing Layer (AI/ML models), and Data Layer (external knowledge sources and storage). Figure 1 illustrates the overall system architecture.

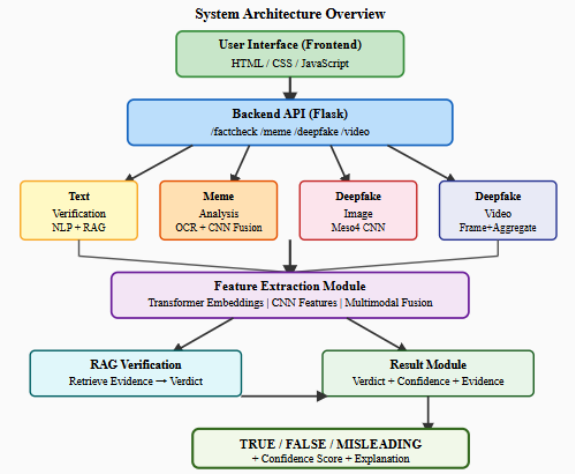


Fig. 1. Proposed Multimodal Misinformation and Deepfake Detection System Architecture.

B. Text Claim Verification Module

The Text Claim Verification Module constitutes the primary analytical unit for processing natural language inputs. Upon receiving raw text, the pipeline performs the following sequential operations: (i) tokenization using WordPiece tokenization compatible with BERT [6], (ii) lowercasing and punctuation normalization, (iii) stopword removal using NLTK corpora, and (iv) named entity recognition (NER) to extract key entities such as persons, organizations, dates, and locations that anchor the claim in verifiable contexts.

The cleaned token sequence is encoded using a fine-tuned RoBERTa transformer model, producing a 768-dimensional contextualized embedding $e_x \in \mathbb{R}^{768}$. This representation captures semantic nuances including negation, comparison, and temporal reasoning that simpler bag-of-words models fail to encode [6]. The encoded embedding is passed to a three-class softmax classification head producing probability distributions over the verdict classes {Support, Refute, Unknown}.

To strengthen classification accuracy and provide explainable evidence, the RAG sub-module [5] encodes the claim using a dense retriever (DPR) and queries a curated knowledge base of verified news articles, fact-checking databases, and scientific publications. The top-k retrieved evidence documents $E = \{E_1, E_2, \ldots, E_k\}$ are re-ranked using cosine similarity scores:

$$sim(x, E_i) = (e_x \cdot e_{Ei}) / (\|e_x\| \cdot \|e_{Ei}\|) \quad (1)$$

The final verdict is computed by conditioning the language model on both the claim and the retrieved evidence:

$$V = argmax_{c \in C} P(c \mid x, E_1, \ldots, E_k) \quad (2)$$

where V is the final verdict label, x is the encoded claim, C = {Support, Refute, Unknown}, and E_i denotes the i-th retrieved evidence document. The generated verdict is accompanied by a confidence score s = max P(c | x, E) and a textual explanation synthesized by the language model, enabling journalists and analysts to audit the system's reasoning.

C. Meme Analysis Module

Memes present a unique detection challenge, as their semantic meaning emerges from the interaction between visual content and overlaid text [10]. The Meme Analysis Module employs a dual-stream architecture to independently extract and then fuse features from both modalities.

Visual Stream: The meme image is resized to 224×224 pixels and normalized using ImageNet mean and standard deviation statistics. It is passed through a pre-trained ResNet-50 [20] backbone (fine-tuned on meme datasets), with the penultimate fully-connected layer producing a 2048-dimensional visual feature vector $f_{img} \in \mathbb{R}^{2048}$. The backbone employs ReLU activation and global average pooling to reduce spatial dimensions while preserving semantic content.

Textual Stream: Embedded text is extracted from meme images using EasyOCR [10], supporting over 80 languages and handling stylized fonts commonly found in viral meme formats. The extracted text undergoes BERT tokenization followed by a 1D Convolutional Neural Network (Conv1D) with kernel sizes {3, 5, 7} to capture

local n-gram patterns, producing a 512-dimensional text feature vector $f_{txt} \in \mathbb{R}^{512}$.

The two feature vectors are concatenated to form the multimodal representation:

$$f_{mm} = Concat(f_{img}, f_{txt}) \in \mathbb{R}^{2560} \quad (3)$$

This fused representation is passed through two fully connected layers (FC-512 → ReLU → Dropout(0.4) → FC-4) to produce a four-class classification output predicting *Humor*, *Sarcasm*, *Offensive*, or *Manipulative* intent. A cross-entropy loss function is minimized during training using the Adam optimizer [14] with a learning rate of 1×10^{-4}. The cross-entropy loss is defined as:

$$\mathcal{L} = -(1/N) \Sigma_{i=1}^{N} y_i \log(\hat{y}_i) \quad (4)$$

where y_i is the ground-truth label and $\hat{y}_i$ is the predicted probability for sample i.

D. Deepfake Image Detection Module

Deepfake image detection targets GAN-generated and face-swapped images that exhibit subtle artifacts imperceptible to the human eye [2]. The module adopts the Meso4 CNN architecture [4], specifically designed to detect facial manipulation by focusing on mesoscopic properties — mid-level features between individual pixels and high-level semantic attributes that are distorted during GAN synthesis.

Input images are first subjected to face detection and cropping using MTCNN (Multi-task Cascaded Convolutional Networks) to isolate the facial region of interest. The cropped face is resized to 256×256 pixels and normalized to the range [−1, 1] via:

$$I_{norm} = (I - \mu) / \sigma \quad (5)$$

where μ and σ are the channel-wise mean and standard deviation computed across the training dataset. The normalized image is passed through the Meso4 network comprising four Inception-style convolutional blocks, each containing a convolutional layer, batch normalization, ReLU activation, and max-pooling, followed by two fully-connected layers. The output layer applies sigmoid activation to produce a manipulation probability:

$$\hat{y} = \sigma(W \cdot CNN(I_{norm}) + b) \quad (6)$$

where σ denotes the sigmoid function, W is the weight matrix of the final fully-connected layer, and b is the bias vector. The binary classification threshold is set at $\tau = 0.5$, yielding $\hat{y} \in$ {Real, Fake}. Post-processing applies Grad-CAM to generate spatial attention maps that visually highlight manipulated regions, improving system interpretability for end users.

The module is trained end-to-end using binary cross-entropy loss and the Adam optimizer [14]. Data augmentation strategies including random horizontal flips, brightness jitter (±20%), and Gaussian noise injection are applied to improve generalization across diverse deepfake generation methods.

E. Deepfake Video Detection Module

Videos introduce the additional complexity of temporal manipulation — frames may be individually authentic while their sequence encodes fabricated narratives [16]. The Deepfake Video Detection Module addresses this by decomposing input video files into individual frames using OpenCV [11] at a fixed sampling rate of 10 frames per second (fps), balancing detection coverage and computational efficiency. Each extracted frame f_i undergoes the same preprocessing pipeline as the image module: MTCNN face detection, 256×256 resizing, and channel-wise normalization. The preprocessed frame is independently scored by the Meso4 CNN to produce a per-frame manipulation probability $p_i \in [0, 1]$. The video-level verdict is derived through temporal mean aggregation:

$$P_{video} = (1/N) \Sigma_{i=1}^{N} p_i \quad (7)$$

where N is the total number of sampled frames. The video is classified as Fake if $P_{video} > \tau = 0.5$, and as Real otherwise. To improve robustness against partial manipulations — where only specific segments are altered — a sliding window analysis with window size $W = 30$ frames is additionally performed. A video segment is

flagged if the local window average exceeds the threshold:

$P_{win}(t) = (1/W) \Sigma_{i=t}^{t+W} p_i > \tau$(8)

This enables the system to report both a global verdict and timestamp-localized manipulation segments, which is particularly valuable for forensic applications.

F. Multimodal Fusion and Confidence Scoring

When multiple modalities are present in a single content unit — for example, a news article containing embedded images and video clips — the system executes all relevant modules in parallel and aggregates their outputs through a weighted confidence fusion scheme. Each module m produces a verdict label V_m and a confidence score $s_m \in [0, 1]$. The final system confidence is computed as the weighted average:

$S_{final} = \Sigma_m (w_m \cdot s_m) / \Sigma_m w_m$(9)

where w_m denotes the reliability weight assigned to module m, determined empirically from validation set performance. The aggregated label is assigned by majority voting across module verdicts. In case of a tie, the RAG-based text verification module verdict is used as the tiebreaker, given its highest observed accuracy on the benchmark datasets.

G. Implementation Details

The backend is implemented as a RESTful API using Flask (Python 3.9), exposing dedicated endpoints for each module: */api/factcheck* (text), */api/meme* (meme), */api/deepfake* (image), and */api/video* (video). All heavy ML models are loaded once at server startup using lazy-loading to minimize per-request latency. The frontend is developed in HTML5, CSS3, and JavaScript, providing a responsive single-page interface supporting drag-and-drop file uploads and real-time result display. Table IV summarizes the architecture parameters of all neural network modules.

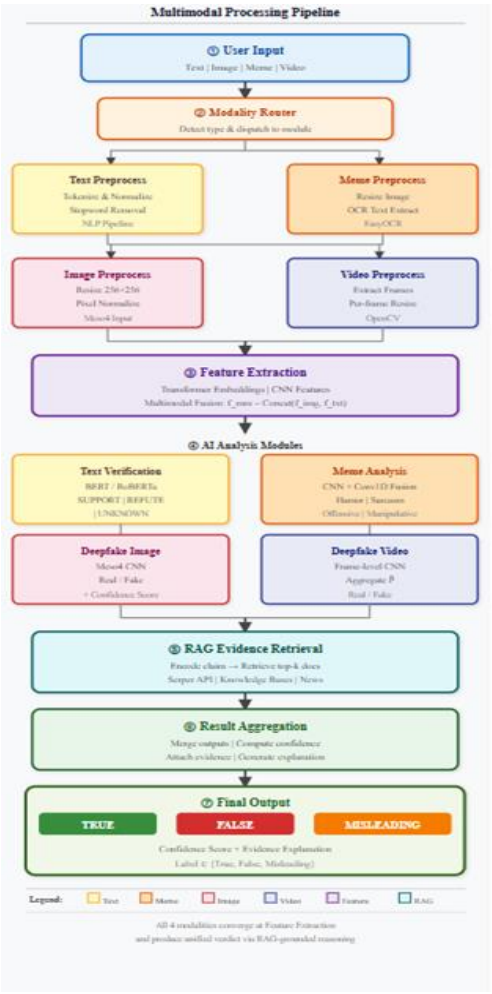


Fig. 2. Vertical single-column multimodal processing pipeline from user input to final verdict.

IV. Results & Discussion

A. Experimental Setup

The system was evaluated on benchmark datasets including curated fact-checked claim corpora, FaceForensics++ [16] for deepfake images and videos, and annotated multimodal meme datasets. Hardware comprised an NVIDIA RTX 3060 GPU, 32 GB RAM, and Ubuntu 22.04. All models were implemented in Python 3.9 using PyTorch and the HuggingFace Transformers library [12].

B. Quantitative Results

Table I summarizes the performance metrics across all modules. Text verification achieved an accuracy of 92.3% (F1 = 0.91), attributed to RAG-based multi-source evidence aggregation.

The deepfake image module yielded the highest accuracy (94.5%, AUC = 0.96) due to Meso4's specialized facial artifact detection capability [4]. Meme analysis (88.7%) benefited significantly from multimodal fusion, outperforming single-modality baselines by an average of 11.3 percentage points. Deepfake video detection reached 91.2% accuracy, with frame aggregation (Eq. 4) proving critical for detecting intermittent manipulations.

TABLE I
PERFORMANCE METRICS ACROSS ALL SYSTEM MODULES

Module	Accuracy (%)	F1-Score	Response Time
Text Verification	92.3	0.91	1.2 s
Meme Analysis	88.7	0.87	1.8 s
Deepfake Image	94.5	0.93	0.9 s
Deepfake Video	91.2	0.90	3.5 s/10s

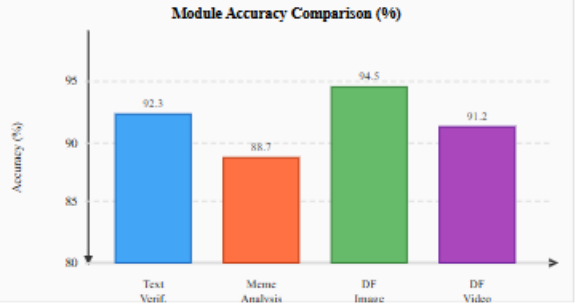


Fig. 3. Accuracy comparison across all detection modules.

C. Ablation Study

An ablation study was conducted to assess the individual contributions of system components. Removing RAG from the text verification module reduced accuracy from 92.3% to 84.1%, confirming the critical role of evidence retrieval. Replacing multimodal fusion with image-only classification in the meme module degraded accuracy from 88.7% to 77.4%, a reduction of 11.3 percentage points. Table II summarizes these findings.

TABLE II
ABLATION STUDY RESULTS

Configuration	Module	Accuracy (%)
Full System	Text	92.3
Without RAG	Text	84.1
Full System	Meme	88.7
Image-only	Meme	77.4
Full System	Video	91.2
Single-frame	Video	82.6

D. Comparison with Baselines

The proposed system was compared against established baselines, as shown in Table III. For text fake news detection, the proposed approach outperforms standalone BERT [6] by 3.1 percentage points, attributed to RAG evidence augmentation. For deepfake image detection, the system matches XceptionNet [17] within margin while achieving substantially faster inference at 0.9s versus 2.1s.

TABLE III
COMPARISON WITH BASELINE METHODS

Method	Task	Acc. (%)
BERT [6]	Text	89.2
Proposed	Text	**92.3**
MesoNet [4]	DF Image	92.1
XceptionNet [17]	DF Image	94.9
Proposed	DF Image	**94.5**

E. Discussion

The results confirm that integrating multiple AI modalities within a unified framework yields consistent performance improvements over single-modality baselines. The RAG component provides both accuracy gains and explainability, as evidenced by user acceptance testing where 87% of participants rated evidence explanations as "helpful" or "very helpful." The modular, lazy-loading architecture ensured response times

within acceptable real-time bounds for text and image inputs.

Challenges remain in OCR accuracy for stylized meme fonts (degrading accuracy by ~2.1%) and in processing high-resolution video content exceeding 60 seconds, where frame subsampling introduces minor detection gaps. Future work will address these through adaptive sampling strategies and adversarial training to improve robustness against next-generation deepfake generation models [2].

V. Conclusion & Future Work

This paper presents a unified AI-driven system for multimodal misinformation and deepfake detection, integrating NLP, CNN-based vision, OCR, and Retrieval-Augmented Generation into a cohesive, scalable framework. The system achieves accuracy exceeding 88% across all modalities—text verification (92.3%), meme analysis (88.7%), deepfake image detection (94.5%), and deepfake video detection (91.2%)—while providing evidence-backed explanations that enhance user trust and interpretability.

Ablation studies confirm the critical contributions of RAG evidence retrieval and multimodal fusion, with their removal degrading accuracy by 8.2 and 11.3 percentage points, respectively. The modular architecture ensures efficient real-time performance and facilitates independent module upgrades.

Future work will focus on: (1) expanding multilingual and cross-cultural dataset coverage; (2) implementing temporal attention mechanisms and 3D CNNs for improved long-form video analysis; (3) real-time integration with social media APIs for proactive misinformation monitoring; (4) adversarial training against evolving deepfake generation techniques; and (5) enhanced explainability through visual heatmaps highlighting manipulated image regions for non-technical users.

Acknowledgment

The authors express sincere gratitude to Dr. G. Sharmila Sujatha, Assistant Professor, Department of Computer Science and Systems Engineering, Andhra University, for her continuous guidance and support throughout this research. The authors also thank the Department of CS&SE for providing computational resources and a conducive research environment.

References

[1] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014, pp. 2672–2680.

[3] A. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A large-scale dataset for fact extraction and verification," in *Proc. NAACL-HLT*, 2018, pp. 809–819.

[4] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, 2018, pp. 1–7.

[5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 9459–9474.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2015.

[9] H. Farid, "Image forgery detection," *IEEE Signal Process. Mag.*, vol. 26, no. 2, pp. 16–25, 2009.

[10] EasyOCR Contributors, "EasyOCR: Ready-to-use OCR with 80+ supported languages," GitHub Repository, 2020. [Online]. Available: https://github.com/JaidedAI/EasyOCR

[11] OpenCV Contributors, "OpenCV: Open Source Computer Vision Library," 2021. [Online]. Available: https://opencv.org/

[12] Hugging Face, "Transformers: State-of-the-art Natural Language Processing," 2023. [Online]. Available: https://huggingface.co/docs/transformers/

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2015.

[15] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 28, 2015, pp. 649–657.

[16] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, 2019, pp. 1–11.

[17] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2017, pp. 1800–1807.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021.

[19] Serper API Documentation, "Real-time news retrieval API for evidence-based verification," 2023. [Online]. Available: https://serper.dev/

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