JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

TruthLens: A Lightweight Multimodal Framework for Fake News Video Detection

[1]Anshul Bhatt, [2]Sehar Khan

[1]MSc CS Specialization in Cybersecurity, [2]Assistant Professor

[1,2]Department of Advanced Computing
[1,2]Nagindas Khandwala College,
[1,2]University of Mumbai, Maharashtra, India

Abstract—The rapid growth of online video platforms has increased the spread of fabricated and manipulated news content. Unlike traditional text-based misinformation, fake news videos can contain misleading information across multiple modalities, including visual frames and spoken content. This paper presents TruthLens, a lightweight multimodal framework for detecting fake news videos by combining visual and textual information. The proposed system uses ResNet50 to extract visual features from sampled video frames, Whisper for automatic speech transcription, BERT for contextual textual feature extraction, and XGBoost for final classification. The extracted 2048-dimensional visual representation and 768dimensional textual representation are concatenated to produce a 2816-dimensional multimodal feature vector. The current dataset contains 305 videos, consisting of 180 REAL and 125 FAKE samples. Each video is processed using eight sampled frames, audio transcription, and multimodal feature extraction. On the held-out test set, the system achieved 82.61% accuracy, 82.35% precision, 73.68% recall, 77.78% F1score, and 0.961 ROC-AUC using the reference threshold of 0.50. A threshold of 0.45 was previously selected during validation experiments, while the current production and reporting decision threshold is 0.30 (30%). To improve interpretability, SHAP is used for feature-level explanation and Grad-CAM is used to visualize influential regions within video frames. The proposed framework is intended to evaluate generalization to unseen fakevideo types and sources rather than claim universal detection of all possible fake videos.

Index Terms—Fake news detection, multimodal learning, video classification, ResNet50, Whisper, BERT, XGBoost, SHAP, GradCAM, explainable artificial intelligence.

I. INTRODUCTION

The rapid development of social media and video-sharing platforms has significantly changed the way people consume news and information. News videos can reach large audiences within a short period of time, making video-based misinformation an important research problem. Fabricated or manipulated videos may contain false claims, misleading narration, selective editing, manipulated visual content, or combinations of genuine and misleading material.

Traditional fake-news detection approaches primarily Focus on textual information. However, news videos contain multiple sources of information, including visual frames,spoken language, and audio.

A system that relies on only one modality may therefore Fail to identify important evidence contained in another modality.

Multimodal machine learning provides an opportunity to combine information from different sources. In the proposed TruthLens framework, visual information is extracted using ResNet50, while speech is converted into text using Whisper. BERT is subsequently used to obtain contextual textual representations from the generated transcript. The visual and textual representations are fused and classified using XGBoost.

The primary objective of TruthLens is to develop a lightweight and explainable framework for detecting fake news videos. The system is not intended to claim that every possible fake or manipulated video can be detected. Instead, the research focuses on evaluating classification performance on the constructed dataset and studying the ability of the framework to generalize to unseen fakevideo types and sources. The main contributions of this work are:

A lightweight multimodal pipeline combining visual and speech-derived textual information. • Integration of ResNet50, Whisper, BERT, and XGBoost for fake news video classification.

A 2816-dimensional fused representation combining visual and textual features.

Integration of SHAP and Grad-CAM for explainable predictions.

Evaluation using separate training, validation, and test datasets.
Consideration of generalization to unseen fakevideo types and sources.

II. Literature Review

Fake news detection has traditionally focused on analyzing textual content using natural language processing and machine-learning techniques. However, the increasing use of short-video and social-media platforms has made fake news detection a multimodal problem. News videos may contain misleading speech, manipulated visual content, selective editing, false captions, or combinations of genuine and fabricated information. Consequently, recent research has increasingly investigated methods that combine textual, visual, and contextual information.

A. Fake News Detection on Short-Video Platforms

Wang et al. introduced FakeSV, a multimodal benchmark designed for fake news detection on short-video platforms [1]. The study emphasized that short videos contain multiple forms of information, including visual content, textual information, and social context. FakeSV demonstrated the usefulness of multimodal information for distinguishing fake and genuine news videos.

Similarly, SV-FEND investigated multimodal fake-news detection on short-video platforms [2]. The framework considered multiple information sources rather than depending exclusively on video text. These studies demonstrate that video-based misinformation detection benefits from combining complementary modalities.

However, many existing approaches depend on large and complex architectures or additional social-context information. TruthLens instead focuses on a lightweight architecture that can be implemented and evaluated using a comparatively smaller dataset while retaining multiple modalities.

B. Visual Deep Learning for Video Analysis

Visual information represents an important component of fakenews videos because manipulation may occur through editing, scene changes, reused footage, or other visual characteristics.

He et al. proposed ResNet, a deep residual learning architecture that introduced residual connections to facilitate the training of deep convolutional neural networks [3]. ResNet architectures have subsequently been widely used for visual feature extraction.

In TruthLens, ResNet50 is used to extract high-level visual representations from eight sampled frames of each video. Instead of training the complete network from scratch, the pretrained network is used as a feature extractor. This reduces computational requirements and makes the framework suitable for a lightweight research implementation.

Although frame-level CNN features can capture useful visual information, they do not explicitly model the temporal relationships between consecutive frames. This represents an important limitation and an opportunity for future work involving temporal video models.

C. Speech Recognition and Textual Information

Spoken information can contain important evidence about whether a news video is misleading or fabricated. Automatic speech recognition therefore provides a method for converting the audio modality into text that can subse quently be analyzed using natural language processing techniques.

Radford et al. introduced Whisper, a large-scale weakly supervised speech-recognition system capable of transcribing speech from diverse audio and video sources [5]. Its ability to generate transcripts makes it suitable for multimodal video processing pipelines.

TruthLens uses the Whisper base model to convert the speech component of each video into an English transcript. The transcript is then passed to BERT for contextual feature extraction.

Devlin et al. introduced BERT (Bidirectional Encoder Representations from Transformers), which uses bidirectional transformer representations to capture contextual relationships within text [4]. BERT has become a widely used architecture for natural language understanding.

In TruthLens, BERT generates a 768-dimensional textual representation from the Whisper transcript. This representation provides the textual component of the multimodal feature vector.

D. Multimodal Feature Fusion

A major challenge in fake-news video detection is determining how information from different modalities should be combined. Visual information may reveal characteristics of the footage, while textual information may provide information about the claims or narration. Existing multimodal fake-news research demonstrates that combining multiple modalities can provide additional information compared with relying exclusively on a single modality [1], [2]. However, multimodal fusion can increase model complexity and computational requirements.

TruthLens uses a straightforward feature-level fusion strategy. The 2048-dimensional ResNet50 representation is concatenated with the 768-dimensional BERT representation, producing a 2816-dimensional multimodal feature vector.

This approach was selected because it is comparatively simple to implement, computationally manageable, and easy to explain during evaluation.

E. Machine-Learning Classification

XGBoost was introduced by Chen and Guestrin as an efficient scalable gradient-boosting framework [6]. It has demonstrated strong performance on structured and highdimensional feature representations.

TruthLens uses XGBoost as the final classifier rather than adding another deep neural-network classification layer. The fused 2816-dimensional feature representation is provided to XGBoost, which produces the probability of the video belonging to the FAKE class. This approach separates feature extraction from classification and reduces the amount of end-to-end neural-network training required.

F. Explainable Artificial Intelligence

Explainability is particularly important for fake-news detection because users may want to understand why a system classified a video as fake or genuine.

Lundberg and Lee proposed SHAP (SHapley Additive exPlanations) as a unified approach for interpreting machinelearning predictions [7]. SHAP estimates the contribution of individual features to a model's prediction.

TruthLens applies SHAP to the XGBoost classifier to identify influential features within the fused representation.

For visual explanation, Selvaraju et al. introduced GradCAM, which uses gradients to identify regions of an input image that contribute to a neural-network prediction [8]. TruthLens applies Grad-CAM to the ResNet50 visual component using the layer4[-1] target layer.

The combination of SHAP and Grad-CAM provides two complementary forms of explanation: SHAP provides featurelevel information, while Grad-CAM provides visual localization.

G. Research Gap

The reviewed studies demonstrate the effectiveness of multimodal approaches for fake-news detection. However, several gaps remain relevant to the development of a lightweight research framework.

First, many existing systems are designed around large datasets or complex multimodal architectures. Such approaches may be difficult to reproduce under limited computational resources.

Second, video-based misinformation can involve different combinations of manipulation, including misleading speech, selective editing, fabricated content, and genuine footage presented with misleading information.

Third, high classification performance on a particular dataset does not necessarily demonstrate generalization to previously unseen fakevideo types or sources.

Fourth, explainability is important when the system is intended for practical use, yet a prediction alone does not explain the visual or feature-level evidence behind the decision.

To address these gaps, TruthLens combines ResNet50, Whisper, BERT, and XGBoost in a lightweight multimodal pipeline. The system additionally incorporates SHAP and Grad-CAM for explainability and considers generalization to unseen fake-video types and sources as an important research objective.

H. Summary of Literature Review

The reviewed literature establishes the importance of multimodal information for fake-news video detection. FakeSV and SV-FEND demonstrate the value of multimodal approaches for short-video misinformation detection [1], [2]. ResNet50 provides an effective mechanism for extracting visual representations [3], while BERT provides contextual textual representations [4]. Whisper enables the conversion of spoken video content into text [5]. XGBoost provides an efficient classifier for the resulting structured feature representation [6]. Finally, SHAP and Grad-CAM provide complementary mechanisms for explaining model predictions [7], [8].

Based on these findings, TruthLens combines these established techniques into a single lightweight framework focused on fake news video detection, explainability, and evaluation of generalization to unseen fake-video types and sources.

III. Proposed Methodology

The proposed TruthLens architecture consists of video preprocessing, frame sampling, visual feature extraction, audio transcription, textual feature extraction, multimodal feature fusion, classification, and explanation generation.

A. Video Preprocessing

The system accepts news videos as local video files. Each video is processed before feature extraction.

Eight representative frames are sampled from every video. This provides a compact representation of the visual content while maintaining manageable computational requirements.

The current dataset contains 305 videos, and all 305 videos successfully contain the required eight frames, audio, and transcripts.

B. Visual Feature Extraction Using ResNet50

ResNet50 is used as the visual feature extractor. Each sampled frame is passed through the pretrained ResNet50 network.

The resulting visual representation contains 2048 features per video.

The extracted representation captures high-level visual information contained within the sampled frames.

ResNet50 is used primarily as a feature extractor rather than being trained from scratch. This reduces computational requirements and makes the framework suitable for a lightweight research implementation.

C. Speech Transcription Using Whisper

The audio component of each video is processed using Whisper.
Whisper converts spoken content into textual transcripts. This allows the system to analyze information that may not be directly observable from video frames.
The current implementation uses the Whisper base model.

D. Text Feature Extraction Using BERT

The transcripts generated by Whisper are processed using BERT (bert-base-uncased).
BERT generates a contextual textual representation with 768 features per video.
The resulting representation captures contextual information from the spoken content.

E. Multimodal Feature Fusion

The ResNet50 and BERT representations are concatenated to create a single multimodal representation. The fused dimensionality is calculated as:

$$2048+768=2816 \text{ features} \quad (1)$$

Therefore, every video is represented using a 2816dimensional multimodal feature vector.

F. XGBoost Classification

XGBoost is used as the final classifier. The class labels are defined as:

- 0= REAL
- 1= FAKE

The classifier produces a probability associated with the FAKE class.
The current production and reporting decision threshold is
0.30 (30%).
Therefore:

- FAKE probability ≥0.30→ FAKE
- FAKE probability < 0.30→ REAL

Changing the threshold changes the final classification decision but does not retrain the XGBoost model or modify its underlying probability output.

IV. Dataset and Experimental Setup

The current TruthLens dataset contains 305 news videos. The dataset distribution is presented in Table I.

Class	Number of Videos
REAL	180
FAKE	125
Total	305

Table 4.1 Dataset Distribution

Duplicate auditing identified zero duplicate video IDs, URLs, or titles. No overlap was identified between the training, validation, and test splits. The split distribution is detailed in.

All 305 videos contain eight sampled frames, audio, and Whisper transcripts.
The generated BERT feature representation contains 305 samples with 768 dimensions, while the ResNet50 representation contains 305 samples with 2048 dimensions.
The final fused representation contains 2816 features per video.

V. IMPLEMENTATION

The TruthLens implementation consists of preprocessing, feature extraction, multimodal fusion, classification, and explainability components.

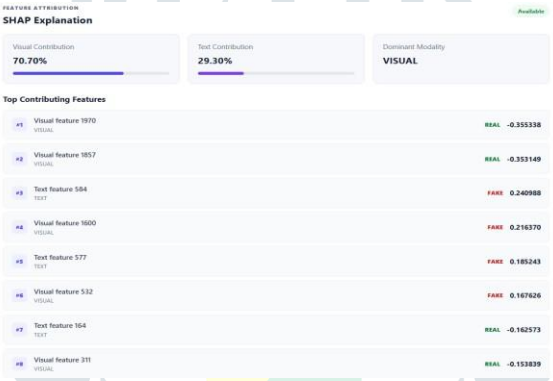
A. Preprocessing Pipeline

The preprocessing pipeline performs the following operations:

1) Loads the input video.

Fg.5.1 uploaded video from system

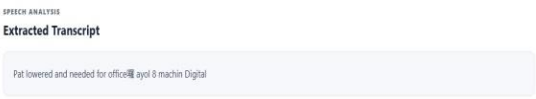
2) Samples eight representative frames.



Fg.5.2 Extracted frames

3) Extracts the audio stream.

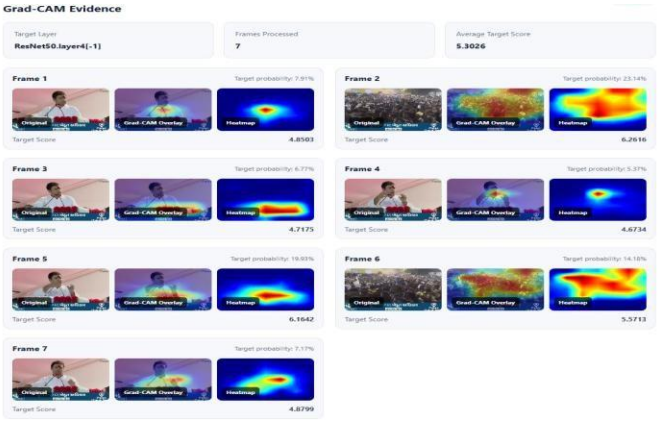
4) Generates an English transcript using Whisper.



Fg5.3 Generated text

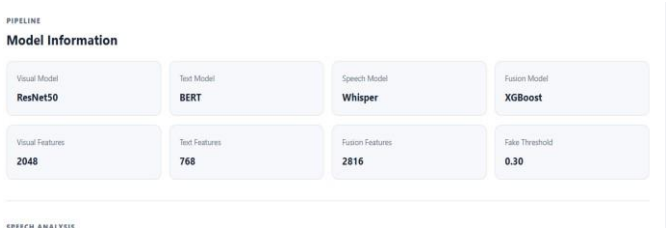
5) Stores processed metadata.

6) Generates ResNet50 visual features.



Fg5.4 Extracted visual features

7) Generates BERT textual features.
8) Combines the features into the final multimodal representation.



Fg.5.5 Mutli-modal information

B. Feature Representation

The visual and textual feature representations are stored independently before fusion.
The BERT representation contains:

305×768 features

The ResNet50 representation contains:

305×2048 features

The final fused representation contains:

305×2816 features

This structure allows the two modalities to be independently analyzed before classification.

C. Classification

The fused 2816-dimensional representation is passed to the XGBoost classifier.
The classifier generates a probability for the FAKE class.
The current TruthLens application uses the 0.30 threshold for its operational decision.

Fg.5.6 Data Split representation

VI. RESULTS

The primary reference evaluation was performed on the held-out test set using a conventional threshold of 0.50. The performance metrics are shown in Table III.

Metric	Result
Accuracy	82.61%
Precision	82.35%
Recall	73.68%
F1-Score	77.78%
ROC-AUC	0.961

Table 6.1 TRUTHLENS TEST PERFORMANCE

The corresponding confusion matrix is shown in

	Predicted REAL	Predicted FAKE
Actual REAL	24	3
Actual FAKE	5	14

Table 6.2 CONFUSION MATRIX AT 0.50 REFERENCE THRESHOLD

Therefore, 24 REAL videos were correctly classified, 3 REAL videos were incorrectly classified as FAKE, 5 FAKE videos were incorrectly classified as REAL, and 14 FAKE videos were correctly classified.

The ROC-AUC value of 0.961 indicates strong discrimination capability on the current held-out test set. However, because the dataset is relatively small, these results should be interpreted cautiously.

VII. Decision Threshold Analysis

Decision threshold selection is important because the classifier's probability must be converted into a final classification.
During development, a threshold of 0.45 was selected during validation experiments. The validation experiment produced:

- Accuracy = 89.13%
- F1-score = 87.18%
- ROC-AUC = 0.9727

However, the current production/reporting threshold is 0.30.

Threshold	Purpose
0.50	Reference test evaluation
0.45	Earlier validation-selected experiment
0.30	Current production/reporting threshold

Dataset Split	Total	REAL	FAKE
Training	213	126	87
Validation	46	27	19
Testing	46	27	19
Total	305	180	125

Table 7.0 Decision Thresholds Used During Development

The 0.30 threshold should be used consistently by the backend and frontend of the current TruthLens application. For example, if the model generates: FAKE probability =0.3934

then: 0.3934≥0.30 and the operational prediction is: FAKE
The threshold does not change the raw probability generated by XGBoost. It only determines how that probability is converted into the final class label.

VIII. Explainability

Explainability is an important component of TruthLens because a fake-news detection system should provide information about why a prediction was generated. Two explanation mechanisms are implemented:

- SHAP
- Grad-CAM

A. SHAP

SHAP is applied to the XGBoost classifier to estimate the contribution of individual features toward the prediction.
Some influential ResNet50 latent features identified in the current explanation outputs are summarized in

Feature	SHAP Importance
ResNet50 Feature 1202	0.375260
ResNet50 Feature 1089	0.306487
ResNet50 Feature 832	0.245840

Table 8.1 Selected SHAP Feature Importance

These feature indices represent latent neural-network representations. They should not be interpreted as direct human readable concepts without additional analysis.

B. Grad-CAM

Grad-CAM is used to generate visual explanations for the ResNet50 component.
The current implementation targets: ResNet50.layer4[-1]

Grad-CAM heatmaps are generated for the sampled video frames.
The purpose is to identify image regions that contributed strongly to the visual representation used by the classifier.

IX. SYSTEM ARCHITECTURE

The overall architecture of the TruthLens framework is illustrated in Figure 9.1 The pipeline executes a dual stream multimodal feature extraction and classification sequence: This architecture allows the system to combine visual and speech-derived textual information within a single classification pipeline.

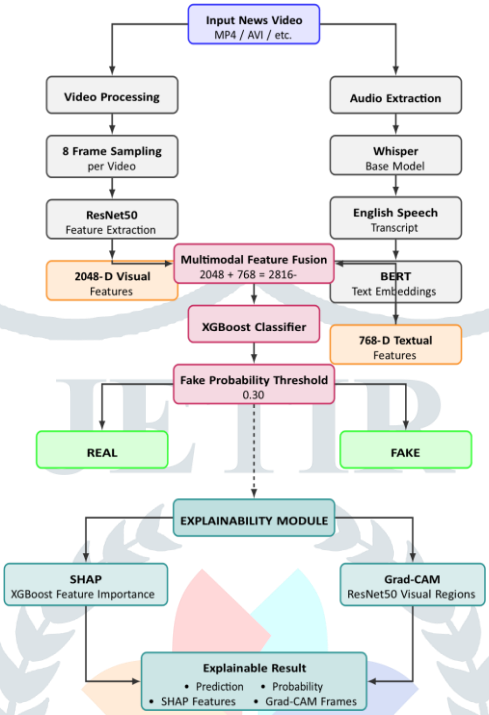


Fig. 9.1 TruthLens System Architecture Diagram

X. DISCUSSION

The experimental results demonstrate that combining visual and textual representations provides useful information for fake news video detection.
The ROC-AUC value of 0.961 indicates strong discrimination capability on the current held-out test set. However, the recall of 73.68% indicates that some FAKE videos remain difficult to identify.
This is expected because fake videos can differ significantly in their manipulation methods. Some videos may contain genuine visual footage combined with misleading narration, while others may involve selective editing or fabricated content.
The multimodal architecture of TruthLens allows evidence from both visual and speech-derived textual modalities to contribute to the final classification.
The explainability mechanisms also improve the usefulness of the system by providing information about influential features and visual regions.
However, the results should not be interpreted as evidence that TruthLens can detect every possible fake video.

XI. GENERALIZATION TO UNSEEN FAKE-VIDEO TYPES

A major objective of TruthLens is to investigate generalization. A model may achieve high accuracy on a dataset while performing poorly when exposed to videos from a different source or manipulation technique.

Therefore, TruthLens should be evaluated using unseen fake- video types and sources where sufficient data are available.

Potential evaluation categories include:

- Unseen fake-video sources.
- Unseen manipulation techniques.
- Different news topics.
- Different video qualities.
- Different editing styles.
- Genuine footage combined with manipulated narration.
- Selectively edited genuine news videos.

The system does not claim to detect every possible fake video. Instead, the intended research objective is to evaluate whether the learned multimodal represent ations can generalize beyond the specific examples used during training.

XII. Limitations

Several limitations should be considered. First, the current dataset contains 305 videos, which is relatively small for a multimodal video-classification problem.

Second, the current dataset focuses primarily on English language news content. Therefore, performance on other languages cannot be assumed.

Third, ResNet50 primarily captures frame-level visual information and does not explicitly model long-term temporal relationships between video frames.

Fourth, Whisper transcription may contain errors when the audio quality is poor, speech is unclear, or background noise is present.

Fifth, BERT analyzes the generated transcript but does not independently verify whether the spoken claims are factually correct.

Sixth, the model may learn source-specific characteristics if the dataset does not contain sufficient diversity.

Finally, the reported results should not be interpreted as evidence that TruthLens can detect every form of misinformation, deepfake, manipulation, or fabricated media.

XIII. Future Work

Future work can improve TruthLens in several directions. First, the dataset can be expanded with additional REAL and FAKE videos from diverse sources.

Second, source-held-out and manipulation-type-held out evaluation can be introduced to more rigorously measure generalization.

Third, temporal models such as LSTM, GRU, temporal transformers, or video transformers could be incorporated to capture relationships between consecutive frames.

Fourth, multilingual speech recognition and language models could be introduced to support news videos in languages other than English.

Fifth, claim verification could be integrated with external knowledge sources so that the system can analyze factual consistency rather than relying exclusively on learned visual and textual patterns.

Sixth, more advanced multimodal fusion techniques could be investigated instead of simple feature concatenation.

Finally, explainability could be improved by connecting latent feature importance to more human readable visual and textual evidence.

XIV. Conclusion

This paper presented TruthLens, a lightweight multimodal framework for fake news video detection using ResNet50, Whisper, BERT, and XGBoost.

The framework extracts visual information from eight sampled video frames using ResNet50 and converts spoken information into contextual textual representations using Whisper and BERT. The resulting 2048-dimensional visual representation and 768dimensional textual representation are fused into a 2816dimensional feature vector and classified using XGBoost.

The current dataset contains 305 videos, including 180 REAL and 125 FAKE samples. The held-out reference evaluation achieved 82.61% accuracy, 82.35% precision, 73.68% recall, 77.78% F1-score, and 0.961 ROC-AUC at the 0.50 reference threshold.

SHAP and Grad-CAM provide complementary explanation mechanisms that allow the model's deci- sions to be investigated from feature-level and visual perspectives.

The primary contribution of TruthLens is therefore not a claim of universal fake-video detection, but a practical, lightweight, and explainable multimodal framework for detecting fake news videos and evaluating generalization to unseen fake-video types and sources.

REFERENCES

[1] S. X. Wang et al., "FakeSV: A multimodal benchmark with rich social context for fake news detection on short video platforms," 2022.

[2] M. Qi et al., "SV-FEND: A multimodal framework for fake news detection on short-video platforms," 2022.

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

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

[5] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. Machine Learning (ICML)*, 2023.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017.

[8] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626.

[9] C. Comito, L. Caroprese, and E. Zumpano, "Multimodal fake news detection on social media: A survey of deep learning techniques," *Social Network Analysis and Mining*, vol. 13, Art.no.101, 2023

[10] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).