Explainable AI for Multimodal Fake Profile Detection: An Integrated Framework with Trust Scoring

Dr. Aakanshi Aggrawal
CSE Department
Noida, Uttar Pradesh, India
aakanshi.aggrawal@amity.edu

Sumit Kumar Verma
CSE Department
Amity University
Noida, Uttar Pradesh, India
sumitkumarverma09012003@gmail.com

Suhani Sidhu
CSE Department
Noida, Uttar Pradesh, India
suhanisidhu07@gmail.com

***Abstract*—The proliferation of AI-generated media has enabled widespread creation of sophisticated fake profiles across social, professional, and dating platforms, undermining digital trust and user safety. This paper presents *RealityCheck AI*, an explainable multimodal fake profile detection framework that integrates deep learning-based image analysis, natural language processing, and metadata forensics to compute an interpretable Trust Score. The system employs ResNet-18 with YOLOv8 for face authenticity detection, DistilBERT for text classification, and PyExifTool for metadata analysis. Explainability is achieved through Grad-CAM visualizations for image reasoning and SHAP/LIME methods for textual feature interpretation. Experimental evaluation on diverse datasets yielded an overall F1-score of 0.82 with 85% accuracy, demonstrating robust performance across modalities. The integration of Explainable AI (XAI) techniques enhances transparency, accountability, and user trust in automated authenticity verification systems. This work addresses the critical gap between high-performing black-box models and interpretable, deployable solutions for real-world platforms.**

***Index Terms*—fake profile detection, explainable artificial intelligence, deep learning, natural language processing, metadata forensics, multimodal fusion, trust scoring, Grad-CAM, SHAP**

I. Introduction

The rapid advancement of generative artificial intelligence has fundamentally transformed the landscape of online identity verification. Sophisticated models such as StyleGAN [1], Stable Diffusion [2], and large language models like GPT-4 [3] can generate photorealistic images and coherent textual content that effectively deceive both human observers and traditional automated detection systems. This technological capability has been exploited to create fake profiles at scale across social media platforms, professional networking sites, and online dating applications, posing significant threats to digital trust, user safety, and platform integrity.

Current detection approaches predominantly focus on single-modality analysis—either image or text—and typically operate as black-box systems that provide binary classifications without interpretable justifications. This opacity presents substantial barriers to user trust, regulatory compliance, and practical deployment in production environments where stakeholders require transparent decision-making processes.

This research introduces *RealityCheck AI*, a comprehensive explainable multimodal framework that addresses these limitations through three key contributions:

1) A multimodal fusion architecture combining image, text, and metadata analysis with weighted integration to produce a unified Trust Score
2) Integration of state-of-the-art Explainable AI techniques including Grad-CAM [4], SHAP [5], and LIME [6] to provide interpretable predictions
3) Comprehensive evaluation demonstrating both high detection performance (F1-score of 0.82) and effective explainability across modalities

The remainder of this paper is organized as follows: Section II reviews related work in deepfake detection and explainable AI; Section III details the system architecture and individual components; Section IV presents experimental setup and datasets; Section V discusses results and explainability evaluation; and Section VII concludes with future research directions.

II. Related Work

A. Deepfake and Fake Content Detection

Research in synthetic media detection has evolved significantly over the past decade. Ro¨ssler et al. [7] introduced FaceForensics++, a comprehensive benchmark dataset containing over 1.8 million manipulated video frames, and demonstrated effective classification using convolutional neural networks. Verdoliva [8] provided an extensive survey of media forensics techniques, emphasizing the importance of generalization across different generation methods and the challenge of adversarial robustness.

For image authenticity, Wang et al. [9] explored CNN-based approaches and identified characteristic artifacts in GAN-generated images, particularly in high-frequency components. Tolosana et al. [10] surveyed deepfake detection methods and highlighted the arms race between generation and detection technologies.

In the text domain, recent work has focused on detecting AI-generated content from large language models. Gehrmann

et al. [11] developed GLTR to identify automatically generated text through statistical analysis of token probabilities. Mitchell et al. [12] proposed DetectGPT, which exploits the phenomenon that model-generated text tends to occupy negative curvature regions of the model's log probability function.

B. Explainable Artificial Intelligence

The field of XAI has emerged to address the interpretability crisis in modern AI systems. Lundberg and Lee [5] introduced SHAP (SHapley Additive exPlanations), a unified framework based on cooperative game theory that provides consistent feature attribution. Ribeiro et al. [6] proposed LIME (Local Interpretable Model-agnostic Explanations), which explains individual predictions by learning local linear approximations.

For visual explanations in deep neural networks, Selvaraju et al. [4] developed Grad-CAM (Gradient-weighted Class Activation Mapping), which produces visual explanations by utilizing gradients flowing into the final convolutional layer. Zhou et al. [13] introduced Class Activation Mapping (CAM) as a precursor technique.

C. Research Gap

Despite significant progress in both detection and explainability domains, limited research has integrated multimodal analysis with comprehensive XAI techniques for fake profile detection. Most existing approaches either focus on single modalities or, when multimodal, lack interpretable outputs that justify their predictions. Our work bridges this gap by combining robust multimodal detection with multiple XAI methods tailored to each modality, creating a transparent and trustworthy system suitable for real-world deployment.

III. Methodology

A. System Architecture

RealityCheck AI employs a modular three-stream architecture that processes image, text, and metadata independently before fusing their outputs into a unified Trust Score. The system architecture is illustrated in Fig. **??**.

Each module produces a probability score representing authenticity likelihood:

- I: Image authenticity score $\in [0, 1]$
- T: Text authenticity score $\in [0, 1]$
- M: Metadata authenticity score $\in [0, 1]$

The final Trust Score is computed as a weighted combination:

$$\text{Trust Score} = w_I \cdot I + w_T \cdot T + w_M \cdot M \tag{1}$$

where $w_I = 0.4$, $w_T = 0.3$, and $w_M = 0.3$ represent empirically determined weights based on validation set performance. These weights reflect the relative reliability and discriminative power of each modality, with image features receiving slightly higher weight due to their strong performance in preliminary experiments.

B. Image Authenticity Module

1) Face Detection and Preprocessing: We employ YOLOv8 [14], a state-of-the-art object detection model, for face localization. YOLOv8 provides real-time performance with high accuracy, extracting facial regions with bounding box coordinates. Detected faces are cropped, resized to 128×128 pixels, and normalized using ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$).

Data augmentation techniques include random horizontal flipping, rotation ($\pm 15°$), brightness and contrast adjustment ($\pm 20\%$), and Gaussian noise injection to improve model robustness.

2) Classification Model: The core classification model is ResNet-18 [15], a residual convolutional neural network pretrained on ImageNet. We fine-tune the model on a combined dataset of real faces (CelebA [16]) and synthetic faces (This Person Does Not Exist, Kaggle fake face datasets).

The model architecture consists of:

- Initial convolutional layer: 64 filters, 7×7 kernel
- Four residual blocks with increasing channel dimensions (64, 128, 256, 512)
- Global average pooling layer
- Fully connected layer with binary classification head

Training hyperparameters: batch size 32, Adam optimizer with learning rate 10^{-4}, binary cross-entropy loss, 25 epochs with early stopping.

3) Explainability: Grad-CAM: To provide visual explanations, we implement Grad-CAM [4], which generates class-discriminative localization maps. For a target class c, Grad-CAM computes:

$$L^{c}_{Grad-CAM} = ReLU\left(\sum_{k} \alpha^{c}_{k} A^{k}\right) \tag{2}$$

where $\alpha^{c}_{k} = \frac{1}{Z}\sum_{i}\sum_{j} \frac{\partial y^{c}}{\partial A^{k}_{ij}}$ represents the importance weight for feature map k, and A^{k} is the activation map of the last convolutional layer.

The resulting heatmap highlights regions that most strongly influence the fake/real classification, such as facial inconsistencies, background artifacts, or unnatural texture patterns characteristic of GAN-generated images.

C. Text Authenticity Module

1) Model Architecture: We utilize DistilBERT [17], a distilled version of BERT that retains 97% of its language understanding while being 60% faster and 40% smaller. The model is fine-tuned on a custom dataset of profile biographies.

The architecture consists of:

- 6 transformer layers
- 768-dimensional hidden states
- 12 attention heads per layer
- Maximum sequence length: 128 tokens
- Classification head: linear layer with softmax activation

2) Dataset and Training: Training data comprises 10,000 profile biographies:

- 5,000 human-written bios collected from public profiles (with consent)
- 5,000 AI-generated bios created using GPT-3.5 and GPT-4 with various prompts mimicking different writing styles

Training configuration: batch size 16, AdamW optimizer with learning rate 2×10^{-5}, linear warmup for 500 steps, cross-entropy loss, 5 epochs.

3) Explainability: SHAP and LIME: For text explainability, we implement both SHAP [5] and LIME [6]:

SHAP computes Shapley values for each token, representing its contribution to the prediction. For a text input x with tokens $\{t_1, \ldots, t_n\}$, the SHAP value ϕ_i for token t_i is:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (3)$$

LIME generates local explanations by:

1) Perturbing the input text (removing or masking tokens)
2) Obtaining predictions for perturbed samples
3) Fitting a linear model weighted by proximity to the original input
4) Extracting feature importance from the linear model coefficients

These methods identify linguistic patterns indicative of AI generation, such as:

- Excessive use of formal language and sophisticated vocabulary
- Uniform sentence structure and length
- Repetitive adjectives or phrases
- Lack of personal anecdotes or colloquialisms
- Overly coherent and grammatically perfect text

D. Metadata Forensics Module

1) EXIF Data Extraction: We employ PyExifTool [18], a Python wrapper for ExifTool, to extract metadata from profile images. Examined fields include:

- **Camera Information**: Make, model, lens information
- **Capture Settings**: ISO, aperture, shutter speed, focal length
- **Temporal Data**: Creation timestamp, modification timestamp
- **Geolocation**: GPS coordinates, altitude
- **Software**: Editing applications, generation tools
- **Compression**: JPEG quality, compression artifacts

2) Scoring Algorithm: The metadata authenticity score M is computed based on completeness and consistency:

$$M = \alpha \cdot C + \beta \cdot V + \gamma \cdot A \quad (4)$$

where:

- C: Completeness score (percentage of expected fields present)
- V: Validity score (logical consistency of values)
- A: Anomaly score (detection of suspicious patterns)
- $\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$

Synthetic images typically exhibit:

- Missing or minimal EXIF data
- Inconsistent camera parameters
- Software signatures from generation tools (e.g., "Adobe Photoshop", "GIMP", unknown generators)
- Absence of GPS coordinates
- Anomalous timestamp patterns

3) Explainability: Metadata explanations are inherently interpretable, consisting of direct feature reporting:

- List of missing critical fields
- Flagged anomalies with severity scores
- Comparison against authentic image baselines
- Confidence intervals for suspicious attributes

E. Integration and Deployment

The system is implemented as a RESTful API using Flask (Python 3.9), with the following architecture:

Backend Services:

- Model serving: PyTorch (image), Hugging Face Transformers (text)
- Preprocessing: OpenCV, PIL, spaCy
- Metadata extraction: PyExifTool
- Explainability: SHAP, LIME, pytorch-grad-cam libraries

API Response Format:

```
{
  "profile id": "user123",
  "timestamp": "2025-01-15T10:30:00Z",
  "image analysis": {
    "authenticity score": 0.65,
    "prediction": "likely fake",
    "confidence": 0.82,
    "grad cam url": "/viz/gradcam 123.png"
  },
  "text analysis": {
    "authenticity score": 0.72,
    "prediction": "likely fake",
    "confidence": 0.78,
    "top features": [...]
  },
  "metadata analysis": {
    "authenticity score": 0.20,
    "missing fields": [...],
    "anomalies": [...]
  },
  "final trust score": 0.56,
  "interpretation": "Low trust..."
}
```

Frontend: React-based dashboard with Chart.js for Trust Score visualization, interactive Grad-CAM overlays, and SHAP/LIME feature importance displays.

IV. Experimental Setup

A. Datasets

Image Data:

- **Real Faces**: CelebA dataset [16] (30,000 images)
- **Fake Faces**: This Person Does Not Exist (15,000 StyleGAN2), Kaggle 70K fake faces (15,000), custom generations from DALL-E and Midjourney (5,000)
- Total: 65,000 images (50% real, 50% fake)
- Split: 70% training, 15% validation, 15% testing

Text Data:

- **Human-written**: Scraped from LinkedIn, Twitter bios, dating profiles (with consent, anonymized)
- **AI-generated**: GPT-3.5, GPT-4, Claude with diverse prompts
- Total: 10,000 bios (50% human, 50% AI)
- Split: 80% training, 10% validation, 10% testing

Metadata:

- Analyzed EXIF data from both real and synthetic image datasets
- Established baseline statistics for authentic image metadata

B. Evaluation Metrics

We evaluate system performance using:

- **Accuracy**: Overall correctness
- **Precision**: $P = \frac{TP}{TP+FP}$
- **Recall**: $R = \frac{TP}{TP+FN}$
- **F1-Score**: $F1 = 2 \cdot \frac{P \cdot R}{P+R}$
- **ROC-AUC**: Area under receiver operating characteristic curve

For explainability evaluation, we employ:

- Faithfulness metrics: correlation between feature importance and prediction change upon feature removal
- User studies: interpretability surveys with domain experts
- Computational overhead: inference time comparison

C. Baseline Comparisons

We compare against:

- Single-modality detectors (image-only, text-only)
- Non-explainable multimodal fusion
- Existing fake profile detection systems (where available)

V. Results and Discussion

A. Quantitative Performance

Table I presents comprehensive performance metrics across all modules and overall system performance.

Key Findings:

1) **Image Module**: Achieved 84% accuracy with strong detection of GAN artifacts, particularly in facial regions (eyes, skin texture) and background inconsistencies. The model effectively identified characteristic fingerprints of StyleGAN2 and Stable Diffusion generations.

TABLE I
Performance Metrics Across Modalities

Metric	Image	Text	Metadata	Overall
Accuracy (%)	84.0	88.0	79.0	85.0
Precision	0.83	0.86	0.78	0.82
Recall	0.82	0.87	0.76	0.81
F1-Score	0.82	0.86	0.77	0.82
ROC-AUC	0.89	0.92	0.83	0.90

2) **Text Module**: Highest individual performance (88% accuracy, 0.86 F1-score), successfully distinguishing AI-generated text through linguistic pattern analysis. DistilBERT captured subtle differences in vocabulary distribution, sentence complexity, and stylistic coherence.
3) **Metadata Module**: Modest but valuable contribution (79% accuracy). While providing lower discriminative power individually, metadata analysis offered complementary information, particularly for detecting screenshots or re-uploaded images lacking original EXIF data.
4) **Multimodal Fusion**: The weighted combination yielded 85% overall accuracy and 0.82 F1-score, demonstrating effective integration. Fusion significantly reduced false positives compared to single-modality approaches, as conflicting signals from different modalities triggered additional scrutiny.

B. Explainability Evaluation

1) Visual Explanations (Grad-CAM): Grad-CAM heatmaps revealed interpretable attention patterns:

- **Fake Images**: High activation in facial asymmetries, unrealistic eye reflections, inconsistent lighting across face regions, and unnatural background textures characteristic of GAN hallucinations
- **Real Images**: Attention distributed across naturally occurring features, skin imperfections, and realistic environmental context

Qualitative analysis of 100 randomly selected predictions showed that Grad-CAM highlighted relevant features in 87% of cases, with particularly strong performance on StyleGAN2-generated faces where attention focused on characteristic spectral artifacts.

2) Textual Explanations (SHAP/LIME): Feature importance analysis identified consistent patterns in AI-generated text:

High AI-Indicator Features:

- Formal vocabulary: ”passionate,” ”dedicated,” ”innovative,” ”enthusiastic”
- Perfect grammar with no colloquialisms or typos
- Uniform sentence structure: compound sentences with balanced clauses
- Generic descriptions lacking specific personal details
- Overuse of adjectives and adverbs

High Human-Indicator Features:

- Informal language, contractions, slang
- Minor grammatical imperfections

- Personal anecdotes and specific references
- Variable sentence length and structure
- Emotional expressions and humor

SHAP values provided consistent attributions across similar inputs, while LIME offered intuitive local explanations through token highlighting. Cross-validation between SHAP and LIME showed 82% agreement on top-5 most important features.

3) Metadata Explanations: Metadata forensics provided directly interpretable outputs:

Common Indicators of Fake Profiles:

- Complete absence of EXIF data (73% of fake images)
- Software signatures: ”Unknown,” ”GIMP,” or generic editors (18%)
- Inconsistent timestamps or future dates (5%)
- Missing GPS data in contexts where location tagging is common (62%)

C. Performance vs. Explainability Trade-off

Adding XAI components introduced minimal performance degradation:

- Accuracy decrease: ¡ 2%
- Inference time increase: 15-25% (primarily from SHAP computation)
- Memory overhead: +30 MB for explanation storage

This trade-off is acceptable for deployment scenarios where interpretability is critical for user trust and regulatory compliance.

D. User Study Results

A pilot user study with 25 participants (cybersecurity professionals, social media analysts, general users) evaluated explanation quality:

- 84% found visual explanations (Grad-CAM) helpful for understanding image predictions
- 78% found textual explanations (SHAP/LIME) increased confidence in text predictions
- 92% preferred the explainable system over black-box alternatives
- 76% reported that explanations would influence their decision to trust or investigate a profile further

Participants particularly valued the ability to verify model reasoning against their own intuition, noting that explanations helped identify both correct predictions and potential model errors.

E. Error Analysis

False Positives (Authentic Flagged as Fake):

- Heavily edited but authentic images (filters, retouching)
- Professional photography with studio lighting (atypical metadata)
- Formal writing styles in human-written bios

False Negatives (Fake Flagged as Authentic):

- High-quality recent generations (DALL-E 3, Midjourney v6)
- AI-generated text deliberately mimicking informal styles
- Manipulated images with preserved original metadata

These failure modes highlight areas for future improvement and the importance of continual model updating to address evolving generation techniques.

F. Comparison with Baselines

Table II compares our approach against baseline methods.

TABLE II
COMPARISON WITH BASELINE APPROACHES

Method	F1-Score	Explainable
Image-only CNN	0.79	No
Text-only BERT	0.82	No
Multimodal (no XAI)	0.84	No
RealityCheck AI	**0.82**	**Yes**

While the non-explainable multimodal baseline achieved slightly higher F1-score (0.84 vs. 0.82), our XAI-integrated system provides substantial advantages in interpretability and trustworthiness with minimal performance sacrifice. The 2% performance difference is statistically insignificant and acceptable given the dramatic improvement in user confidence and regulatory compliance capability.

VI. ADVANTAGES AND LIMITATIONS

A. Advantages

1) **Multimodal Robustness**: Integration of image, text, and metadata analysis provides defense-in-depth against sophisticated attacks targeting single modalities. Adversaries must simultaneously deceive all three detection mechanisms.
2) **Enhanced Transparency**: XAI techniques (Grad-CAM, SHAP, LIME) transform opaque predictions into interpretable explanations, increasing user trust and enabling human oversight.
3) **Modular Architecture**: Independent modules can be updated, retrained, or replaced without system-wide modifications, facilitating adaptation to evolving threats.
4) **Deployable API**: Flask-based RESTful interface enables straightforward integration with existing platforms, supporting real-time analysis at scale.
5) **Regulatory Compliance**: Explainability features align with emerging AI regulations (EU AI Act, GDPR Article 22) requiring transparency in automated decision-making.

B. Limitations

1) **Dataset Bias**: Training data may not represent all demographic groups, image generation techniques, or writing styles, potentially causing disparate performance across populations. Mitigation requires diverse, continuously updated datasets.

2) **Computational Overhead**: XAI methods, particularly SHAP for text analysis, increase inference time by 15-25%. While acceptable for profile verification, this may constrain real-time high-throughput applications.
3) **Adversarial Vulnerability**: Determined adversaries aware of detection mechanisms could craft adversarial examples targeting known model weaknesses. Ongoing research in adversarial robustness is necessary.
4) **Generative Model Evolution**: Rapid advancement in generative AI (e.g., Sora, Gemini, DALL-E 3) requires continuous model retraining. Static detectors quickly become obsolete as generation quality improves.
5) **Metadata Availability**: Metadata analysis depends on EXIF data preservation. Social media platforms often strip metadata during upload, limiting this modality's effectiveness in certain contexts.
6) **Explainability Fidelity**: While XAI methods provide interpretable explanations, they are approximations of model reasoning and may not perfectly reflect internal decision processes, particularly for complex neural networks.

VII. Conclusion and Future Work

This paper presented *RealityCheck AI*, an explainable multimodal framework for fake profile detection that integrates image analysis (ResNet-18 + YOLOv8), text classification (DistilBERT), and metadata forensics (PyExifTool) with comprehensive Explainable AI techniques (Grad-CAM, SHAP, LIME). Experimental evaluation demonstrated strong performance (F1-score of 0.82, 85% accuracy) while providing interpretable explanations that significantly enhance user trust and system transparency.

The integration of XAI methods transforms black-box detection into a transparent, accountable system suitable for real-world deployment where stakeholder trust and regulatory compliance are paramount. Our modular architecture facilitates adaptation to evolving threats and seamless integration with existing platforms through a RESTful API.

A. Future Research Directions

1) **Continual Learning**: Implement online learning frameworks enabling automatic model updates as new generation techniques emerge, maintaining detection effectiveness without manual retraining.
2) **Multilingual Support**: Extend text analysis to non-English languages using multilingual transformers (mBERT, XLM-RoBERTa) and culturally-adapted datasets to support global platforms.
3) **Behavioral Analytics**: Integrate temporal interaction patterns, network analysis, and activity fingerprinting to complement content-based detection. Analyzing posting frequency, interaction networks, and engagement patterns can reveal bot-like behavior.
4) **Adversarial Robustness**: Develop defenses against adversarial attacks specifically designed to evade multi-modal detectors. Techniques include adversarial training, certified robustness methods, and ensemble approaches.
5) **Cross-Platform Generalization**: Evaluate transfer learning across different platforms (LinkedIn, Instagram, Tinder) to assess model generalization and identify platform-specific characteristics requiring specialized handling.
6) **Real-Time Optimization**: Optimize inference pipelines through model quantization, pruning, and hardware acceleration (GPU/TPU) to enable high-throughput processing for large-scale platform deployment.
7) **Enhanced Explainability**: Explore counterfactual explanations ("This profile would be authentic if...") and concept-based interpretability to provide more actionable insights for users and moderators.
8) **Privacy-Preserving Detection**: Investigate federated learning and differential privacy techniques to enable detection while protecting user data, addressing growing privacy concerns and regulations.
9) **Video and Audio Modalities**: Extend the framework to analyze profile videos and voice content, addressing emerging deepfake threats in multimedia profiles.

As generative AI continues to advance, the development of robust, transparent, and trustworthy detection systems becomes increasingly critical for maintaining digital trust and platform integrity. This work provides a foundation for next-generation fake profile detection that balances performance with interpretability.

Acknowledgment

The author acknowledges the support of the Department of Computer Science and Engineering at Amity University and thanks the anonymous reviewers for their valuable feedback.

References

[1] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4401–4410.
[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 10684–10695.
[3] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.
[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 4768–4777.
[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2016, pp. 1135–1144.
[7] A. Ro¨ssler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1–11.
[8] L. Verdoliva, "Media forensics and DeepFakes: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 910–932, Jun. 2020.

[9] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 8695–8704.

[10] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, Dec. 2020.

[11] S. Gehrmann, H. Strobelt, and A. M. Rush, "GLTR: Statistical detection and visualization of generated text," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics: Syst. Demonstrations (ACL)*, 2019, pp. 111–116.

[12] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-shot machine-generated text detection using probability curvature," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 24950–24962.

[13] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2921–2929.

[14] Ultralytics, "YOLOv8 Documentation," 2024. [Online]. Available: https://docs.ultralytics.com

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

[16] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 3730–3738.

[17] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[18] P. Harvey, "ExifTool Documentation," 2024. [Online]. Available: https://exiftool.org