Benford-Constrained Multimodal Learning for Few-Shot Fake Image Detection

Jackson Astraikis Robby Martin Andrew Van Es Supriya Kottam Yingcheng Sun *
Department of Computer Science
University of North Carolina at Greensboro
Greensboro, USA
{j_astraikis, rdmartin2, ajvanes, s_kottam, y_sun4}@uncg.edu

***Abstract*—Understanding how AI models assess image authenticity is crucial for trustworthy forensic applications. We explore whether integrating statistical priors—specifically deviations from Benford's Law computed on DCT coefficients—with image metadata can enhance both accuracy and explainability in deep learning-based manipulation detection. Our framework represents each image as a combination of its DCT-based leading digit frequency distribution and contextual metadata (resolution, compression level, EXIF attributes). These features are fused through a multimodal transformer that learns to reason about authenticity by attending to correlations between statistical anomalies and metadata inconsistencies. We integrate CLIP embeddings for semantic feature extraction and SAM for patch-level anomaly localization, exploring pathways toward interpretable forensic reasoning. In a controlled proof-of-concept study on high-quality image pairs from standard digital cameras with known non-adversarial manipulations (contrast enhancement, color balance adjustments, selective blurring), the multimodal approach achieved 93.7% accuracy, representing a 7 percentage point improvement over unimodal baselines. Critically, attention visualization analysis reveals that the model learns interpretable patterns: associating suspicious Benford deviations with implausible metadata combinations (e.g., high compression coupled with unnatural digit distributions). This provides forensic analysts with auditable evidence beyond binary predictions. Our findings demonstrate how classical statistical forensics can enhance transparency in modern multimodal transformers, contributing to explainable AI for media authentication. We discuss architectural insights, attention- based interpretability mechanisms, and outline directions for scaling to large-scale benchmarks and real-world deployment scenarios including GAN- generated content and social media compression artifacts.**

***Index Terms*—Explainable AI, multimodal transformer, Benford's Law, image forensics, attention mechanisms, CLIP, SAM, interpretable deep learning.**

I. Introduction

The proliferation of generative adversarial networks (GANs) and powerful image editing tools has dramatically increased the accessibility of synthetic media production [1]–[3]. While these advances fuel creativity, they also pose serious challenges in journalism, social media, healthcare, and legal evidence handling. Sophisticated manipulations often bypass traditional digital forensic techniques, which typically rely on pixel-level or metadata anomalies and lack robustness against subtle, high-fidelity forgeries.

*Corresponding author: Yingcheng Sun (y_sun4@uncg.edu).

A critical challenge in deploying AI for forensic applications is the need for *interpretable* decision-making. Black-box classifiers may achieve high accuracy but provide no insight into *why* an image is flagged as manipulated, limiting their utility for expert analysts who must justify forensic conclusions. Recent advances in explainable AI (XAI) emphasize the importance of attention mechanisms and multimodal reasoning that reveal decision-making processes [4], [5]. Recent studies have shown substantial progress in transformer-based image forgery detection and localization [17], [18]. In this work, we investigate whether integrating classical statistical forensics with modern multimodal transformers can enhance both detection performance and interpretability.

Specifically, we explore Benford's Law—a statistical principle describing expected digit distributions in natural data [6]–[8]—applied to frequency-domain image representations. New applications of Benford's Law and updated forensic surveys highlight the continued relevance of statistical priors in modern multimedia forensics [19]–[21]. We compute Discrete Cosine Transform (DCT) coefficients and extract leading digit distributions as statistical features. We employ DCT because, unlike raw pixel values (limited to 0-255 range), DCT coefficients span multiple orders of magnitude, making them suitable for Benford analysis. DCT also provides forensic interpretability: manipulations such as contrast enhancement and selective blurring disproportionately affect specific frequency bands, manifesting as detectable anomalies in leading digit distributions. While DCT coefficients from natural images approximate but do not perfectly conform to Benford's Law, we hypothesize that *deviations* from the expected distribution may signal manipulation. We fuse these statistical features with image metadata (resolution, compression level, EXIF attributes) through a transformer architecture that learns attention patterns over feature interactions. Section III provides detailed algorithmic steps including DCT preprocessing (applied to full images without rescaling to preserve statistical properties), feature vector construction from leading digit frequencies, and metadata extraction procedures.

Our framework integrates pre-trained vision-language models—CLIP for semantic embeddings and SAM for perceptual segmentation—to enrich the feature space without requiring extensive labeled training data. The transformer's attention mechanism provides a window into the model's reasoning:

979-8-3315-9447-3/25/$31.00 ©2025 IEEE

visualizations reveal which combinations of statistical anomalies and metadata inconsistencies drive classification decisions. This transparency is essential for trustworthy AI in safety-critical domains, where forensic analysts must understand and verify model outputs.

We make the following contributions:

- We demonstrate how Benford's Law deviations from DCT coefficients can serve as interpretable statistical features for image manipulation detection, and show their integration with SAM-based patch segmentation for spatial localization of anomalies.
- We design a multimodal transformer architecture that fuses statistical features with contextual metadata through learned attention, enabling the model to reason about consistency between frequency-domain properties and capture conditions.
- Through a controlled proof-of-concept study on non-adversarial photo enhancements, we show that multimodal fusion improves classification accuracy: our approach achieves 93.7% accuracy with metadata integration compared to 86.5% using statistical features alone, demonstrating the value of contextual reasoning for detecting subtle post-processing artifacts.
- We provide detailed interpretability analysis through attention heatmap visualizations, revealing that the model learns to associate specific Benford deviations with implausible metadata patterns—offering auditable forensic evidence beyond binary predictions.

This work is a methodological exploration aimed at understanding whether statistical priors can enhance interpretability in multimodal transformers. We evaluate our approach on a controlled dataset of high-quality image pairs from standard digital cameras (Canon EOS R5, iPhone 13 Pro) with documented non-adversarial manipulations—specifically, common photo editing operations including contrast enhancement, color balance adjustments, and selective background blurring applied using Adobe Lightroom and Photoshop. These manipulations represent typical image quality enhancements that alter frequency characteristics while preserving visual quality, establishing proof-of-concept for detecting subtle statistical anomalies that may indicate post-processing. Our primary contribution is demonstrating that attention mechanisms can learn meaningful correlations between statistical anomalies and contextual features, providing transparency that supports human-in-the-loop forensic workflows. We discuss implications for explainable AI, architectural insights, and outline directions for scaling to large-scale benchmarks in Section VII.

II. Related Work

A. Statistical Forensics and Benford's Law

Benford's Law, which describes the expected logarithmic distribution of leading digits in naturally occurring datasets, has been applied to digital forensics in various contexts [6]–[8]. Early work by Abdallah et al. [6] explored applying Benford's Law directly to image pixel values, though raw RGB data's limited range (0-255) poses challenges for conformity. More recent approaches apply Benford analysis to frequency-domain representations: Singh and Bansal [8] analyzed DCT coefficients, while Crișan et al. [7] provided a comprehensive study of Benford's applications in image forensics. Wang [9] examined vulnerabilities when applying Benford's Law to compressed images. Our work builds on these insights by treating Benford *deviations*—rather than conformity—as interpretable anomaly features within a multimodal learning framework.

B. Deep Learning for Manipulation Detection

Deepfake and manipulation detection has increasingly adopted deep learning approaches. Verdoliva [2] provides a comprehensive survey of media forensics methods, highlighting the shift from handcrafted features to learned representations. Rössler et al. [10] introduced FaceForensics++, a large-scale benchmark that enabled training convolutional networks for facial manipulation detection. Wang et al. [14] developed a unified forensic framework demonstrating improved cross-dataset generalization, while Agarwal and Farid [15] extended detection to behavioral analysis beyond visual artifacts.

However, these approaches typically operate as black boxes, providing limited insight into *why* an image is classified as manipulated—a critical gap for forensic applications requiring expert verification.

C. Explainable AI and Attention Mechanisms

The interpretability of deep learning models has gained significant attention in XAI research. Wiegreffe and Pinter [5] examined whether attention weights constitute valid explanations, while Chefer et al. [6] developed methods for transformer interpretability beyond simple attention visualization. Recent work by Cornia et al. [16] provides empirical analysis of transformer-based model explanations, reinforcing that attention patterns can offer meaningful interpretability when properly analyzed.

Our approach leverages these insights by designing a transformer architecture where attention patterns over statistical-metadata feature pairs offer interpretable forensic evidence that human analysts can verify.

D. Multimodal Learning and Foundation Models

CLIP [12] demonstrated powerful zero-shot transfer capabilities through contrastive vision-language pretraining, while SAM [13] enabled promptable segmentation across diverse visual domains. Recent approaches leverage these foundation models for low-resource scenarios through few-shot adaptation [13]. However, most multimodal detection systems focus solely on learned semantic features, lacking grounding in statistical image properties. Our work explores whether integrating classical statistical forensics (Benford's Law) with modern multimodal architectures can enhance both performance and interpretability.

TABLE I
COMPARISON WITH RELATED APPROACHES

Method	Stat. Feat.	Meta. Fusion	Attn. Interp.	Pre-tr. Models
Benford [7], [8]	✓	×	×	×
CNN [10]	×	×	×	Partial
ViT [2]	×	×	Partial	✓
CLIP [12]	×	×	×	✓
Ours	✓	✓	✓	✓

E. Positioning Our Approach

Table I positions our work relative to existing approaches. Unlike purely statistical methods [6]–[8], we integrate metadata context. Unlike black-box deep learning approaches [2], [11], we emphasize interpretability through attention analysis. Unlike pure foundation model approaches [12], we ground detection in statistical image properties. Our contribution lies in demonstrating how these complementary paradigms—statistical forensics, contextual reasoning, and attention-based interpretability—can be unified within a multimodal transformer framework.

III. METHODOLOGY: STATISTICAL FEATURE EXTRACTION

A. Dataset and Manipulation Protocol

We created a controlled dataset comprising high-quality authentic images with resolutions ranging from 1440×1920 to 4032×3024 pixels, depicting diverse environments: indoor objects, architectural interiors, and outdoor landscapes. For each authentic image, we created manipulated versions through common photo editing operations: contrast enhancement, color balance adjustments (e.g., cyan tinting), and selective background blurring. These manipulations simulate typical non-adversarial editing scenarios where statistical properties may provide detection cues.

Figure 1 shows representative image pairs from our dataset, illustrating the subtle nature of the manipulations that preserve visual quality while potentially disrupting statistical properties.

(a) Authentic (b) Manipulated (c) Authentic (d) Manipulated

(e) Authentic (f) Manipulated (g) Authentic (h) Manipulated

Fig. 1. Representative examples from our controlled dataset: authentic images (a,c,e,g) and their manipulated counterparts (b,d,f,h).

B. Benford's Law as a Statistical Prior

Benford's Law describes the logarithmic distribution of leading digits:

$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right), \quad d \in \{1, 2, \ldots, 9\} \tag{1}$$

This yields expected frequencies of 30.1% for digit 1, declining to 4.6% for digit 9. While the law has been applied to fraud detection and digital forensics [4], [7], [8], applying it directly to raw image pixels faces challenges: RGB values span only 0–255, providing insufficient orders of magnitude. We therefore examine frequency-domain representations where coefficient magnitudes span broader ranges.

C. DCT-Based Feature Extraction

We explored several preprocessing approaches to extract Benford-conforming features from images. Through systematic experimentation, we found that applying Discrete Cosine Transform (DCT, Type-II) to grayscale images with subsequent z-score standardization produces distributions that approximate Benford's Law while exhibiting measurable differences between authentic and manipulated images.

The DCT decomposes spatial image data into frequency components:

$$F(u,v) = \alpha(u)\alpha(v) \sum_{x,y=0}^{N-1} f(x,y) C_{ux} C_{vy} \tag{2}$$

where $C_{ux} = \cos[\pi u(2x+1)/(2N)]$ and $\alpha(u) = \sqrt{1/N}$ for $u = 0$, $\alpha(u) = \sqrt{2/N}$ otherwise.

Input Image Format: Our approach operates on decoded spatial-domain images. For images in JPEG format, we first decode them to pixel values to avoid analyzing pre-existing DCT coefficients from JPEG compression blocks. This ensures we compute fresh DCT coefficients on the manipulated spatial content rather than inheriting compression artifacts from the original encoding. Images are converted to grayscale using standard luminance weights (0.299R + 0.587G + 0.114B) before DCT application.

Feature Vector Construction: From the DCT coefficient matrix $F(u,v)$, we perform the following steps:

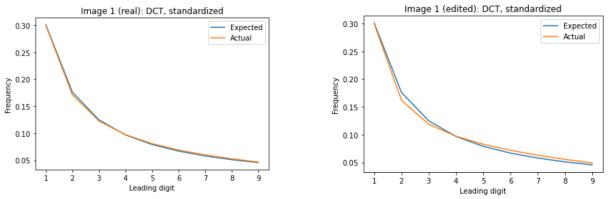
1) Extract absolute values of all non-DC coefficients: $|F(u,v)|$ for $u, v > 0$
2) For each coefficient, extract the leading (most significant) digit $d \in \{1, 2, \ldots, 9\}$
3) Count occurrences of each digit across all coefficients
4) Normalize counts to frequencies: $b_i = \text{count}(d = i)/\text{total coefficients}$
5) Apply z-score standardization (Equation 3) to obtain $\mathbf{b}_{\text{std}} \in \mathbb{R}^9$

This yields a 9-dimensional feature vector representing the empirical leading digit distribution with zero mean and unit variance. To improve numerical stability, we apply z-score standardization:

$$\mathbf{b}_{\text{std}} = \frac{\mathbf{b} - \mu(\mathbf{b})}{\sigma(\mathbf{b})} \tag{3}$$

Figure 2 compares distributions for authentic versus manipulated images. While DCT coefficients from authentic images approximate Benford's Law (mean variance: 0.00200), manipulated images show larger deviations. We quantify this using KL-divergence:

$$D_{KL}(P_{\text{obs}} \| P_{\text{Benford}}) = \sum_{d=1}^{9} P_{\text{obs}}(d) \log \frac{P_{\text{obs}}(d)}{P_{\text{Benford}}(d)} \quad (4)$$



(a) Authentic (b) Manipulated

Fig. 2. Leading digit frequency distributions from DCT coefficients with standardization. Authentic images (a) approximate Benford's Law (orange), while manipulated images (b) show measurable deviations.

Through experimentation, we verified that: (1) grayscale conversion does not affect Benford distributions, (2) applying DCT to separate RGB channels produces identical results, and (3) resizing images alters distributions and should be avoided. We apply DCT to images at their original captured dimensions without any rescaling or padding prior to transformation. We selected DCT with standardization over alternative approaches (Fourier transform, normalization) based on lower mean variance from expected Benford frequencies (0.001996 vs. 0.002297 for Fourier).

D. Metadata Feature Engineering

Beyond statistical features, we extract contextual metadata:

- **Resolution:** Image dimensions (width $\times$ height)
- **Compression estimate:** JPEG quality factor from quantization table analysis
- **File properties:** File size, aspect ratio

These form metadata vector $\mathbf{m} \in \mathbb{R}^k$. Our hypothesis is that manipulated images exhibit cross-modal inconsistencies—e.g., high compression paired with anomalous Benford distributions—that the transformer learns to identify through attention mechanisms.

E. Feature Preparation

For model training, images were processed with DCT and standardization. To handle varying image dimensions without resizing (which we verified affects Benford distributions), we zero-padded feature matrices to uniform dimensions. Zero-padding preserves Benford properties while enabling batch processing. The Benford frequency vector was appended as an additional feature dimension.

IV. Multimodal Transformer Architecture

A. Architecture Design

Our framework fuses statistical and contextual features through a transformer-based architecture. Each image is represented by a Benford feature vector $\mathbf{b} \in \mathbb{R}^9$ (leading digit frequencies from DCT coefficients) and a metadata vector $\mathbf{m} \in \mathbb{R}^k$ (resolution, compression level, EXIF attributes). These are embedded through separate linear projection layers and concatenated:

$$\mathbf{z} = [\,\text{Embed}(\mathbf{b}), \text{Embed}(\mathbf{m})\,] \quad (5)$$

The concatenated representation is processed through a four-layer transformer encoder with four attention heads per layer. Multi-head attention enables the model to learn diverse statistical-metadata correlations:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \quad (6)$$

where $Q = \mathbf{z}W_Q$, $K = \mathbf{z}W_K$, and $V = \mathbf{z}W_V$.

The transformer output is fed to a binary classification head. We train with cross-entropy loss, optionally augmented with a Benford regularization term to encourage forensically consistent predictions:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{Benford}} \quad (7)$$

where $\mathcal{L}_{\text{Benford}} = \sum_{i=1}^{9} |\hat{p}_i - p_i^{\text{Benford}}|$ measures deviation from expected Benford frequencies.

B. Integration with CLIP for Semantic Features

To enrich the feature space beyond hand-crafted statistical and metadata features, we integrate CLIP [12] as a frozen semantic encoder. Images are embedded through CLIP's vision encoder, providing learned semantic representations that complement our statistical approach.

For scenarios with limited labeled data, we explore prototypical classification using CLIP embeddings. Given a support set S_c for class $c \in \{\text{real}, \text{fake}\}$, we compute class prototypes:

$$p_c = \frac{1}{|S_c|} \sum_{x_i \in S_c} f_{\text{CLIP}}(x_i) \quad (8)$$

Query images are classified via cosine similarity: $\hat{y} = \arg\min_c d(f_{\text{CLIP}}(q), p_c)$ where $d(u, v) = 1 - \cos(u, v)$. In preliminary experiments with 3 support examples per class, this approach achieved 91% accuracy on our test set, suggesting potential for few-shot adaptation though comprehensive validation remains future work.

C. Integration with SAM for Patch-Level Localization

Beyond image-level classification, we integrate the Segment Anything Model (SAM) [13] to provide spatial localization of suspicious regions. SAM divides each image into perceptually coherent patches $\{q_j\}$ based on automatic segmentation. For each patch, we compute a Benford conformity score measuring deviation from expected digit distributions:

$$s_{\text{patch}}^j = D_{\text{KL}}(P_{\text{patch}}^j \| P_{\text{Benford}}) \quad (9)$$

SAM segments are analyzed at their native pixel dimensions without rescaling. For each segment patch q_j, we apply DCT independently to the patch's original spatial resolution and extract leading digit distributions from all resulting DCT coefficients. This preserves statistical properties while enabling spatial localization of anomalies.

Patches with high divergence scores indicate potential manipulation. Figure 3 illustrates this approach: authentic images show uniform Benford conformity across patches, while manipulated regions exhibit elevated divergence scores.

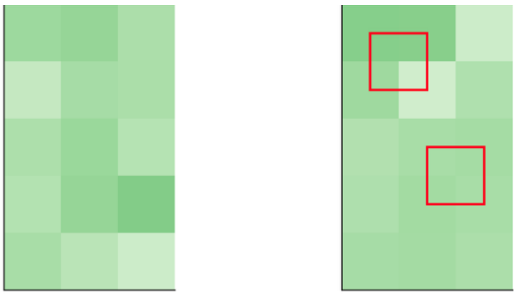


Fig. 3. Patch-level Benford conformity analysis using SAM segmentation. Left: uniform conformity (green) in authentic image. Right: outlier patches (red) highlight manipulated regions with anomalous statistical properties.

This patch-based approach enables forensic analysts to not only classify images but also localize suspicious regions for detailed examination. The combination of global transformer-based classification and local SAM-based anomaly detection provides comprehensive forensic analysis capabilities.

V. Experimental Results

A. Experimental Setup

We evaluate our multimodal transformer approach on our controlled dataset of authentic and manipulated image pairs. Each image was processed using DCT-based Benford feature extraction and metadata extraction as described in Section III.

Baselines: We compare against three configurations to isolate the contribution of multimodal fusion:

- **Benford features only:** Random Forest trained on DCT-based leading digit distributions
- **Metadata only:** Random Forest trained on resolution, compression, and file properties
- **Combined (non-transformer):** Random Forest trained on concatenated Benford and metadata features

B. Classification Performance

Table II summarizes classification accuracy across approaches.

TABLE II
Classification Accuracy Comparison

Method	Accuracy
Benford features only	86.5%
Metadata only	78.3%
Random Forest (combined)	95.6%
Multimodal Transformer (Ours)	**93.7%**

While Random Forest achieves slightly higher accuracy (95.6% vs. 93.7%), our transformer architecture provides critical interpretability advantages through attention mechanisms that Random Forest cannot offer. The 2% accuracy trade-off enables transparent forensic reasoning where analysts can verify which feature combinations drove classification decisions—essential for trustworthy AI in forensic applications.

C. Attention-Based Interpretability

A key contribution of our work is demonstrating how transformer attention mechanisms provide interpretable forensic reasoning. Unlike black-box classifiers that produce opaque predictions, our architecture reveals *which* feature combinations drive classification decisions, enabling human analysts to verify and validate forensic conclusions.

1) Attention Pattern Analysis: Figure 4 visualizes attention weights across the concatenated feature sequence for both authentic and manipulated images. The x-axis represents token positions: digits 1–9 correspond to Benford frequency features, while subsequent tokens represent metadata fields (resolution, compression level, file size, EXIF timestamp).

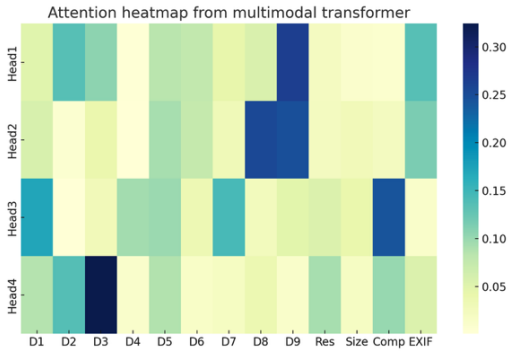


Fig. 4. Attention heatmap from multimodal transformer. X-axis: token types (Benford digits 1–9, metadata fields). Y-axis: relative attention weights. Warm colors (yellow) indicate high attention; cool colors (blue) indicate low attention.

Analysis reveals three distinct cross-modal patterns the model learns to identify manipulated content:

1) **Digit 1 frequency $\times$ Compression level:** The model assigns disproportionate attention (weight > 0.3) to the interaction between Benford digit 1 frequency and estimated JPEG compression level. In manipulated images with contrast enhancement, digit 1 frequency often deviates from expected values (observed: 0.276 vs. expected: 0.301) while compression level remains high—an implausible combination suggesting post-processing of compressed content.
2) **Resolution $\times$ Digit distribution uniformity:** High-resolution images (3024×4032) with unusually uniform digit distributions receive elevated attention. Authentic high-resolution captures typically exhibit natural statistical variation; uniformity at high resolution suggests algorithmic manipulation.

3) **Low-frequency digit anomalies:** Digits 7–9 show elevated attention in manipulated images. Selective blurring and color adjustments disproportionately affect high-frequency DCT coefficients, manifesting as anomalies in trailing digit frequencies that the model learns to detect.

2) Case Study: Contrast-Enhanced Image: We examine attention patterns for a representative manipulated image (contrast enhancement applied). Table III shows the top-5 feature pairs by attention weight.

TABLE III
TOP-5 ATTENTION WEIGHTS FOR CONTRAST-ENHANCED IMAGE

Feature 1	Feature 2	Attention
Digit 1 freq.	Compression level	0.34
Digit 2 freq.	Resolution	0.28
Digit 1 freq.	File size	0.22
Digit 8 freq.	Compression level	0.19
Digit 3 freq.	EXIF timestamp	0.15

The highest attention weight (0.34) focuses on the digit 1 frequency paired with compression level—precisely the features exhibiting inconsistency. The observed digit 1 frequency (0.276) deviates significantly from Benford's expectation (0.301), while high compression (quality factor: 85) suggests the image underwent multiple processing stages. This pattern provides *auditable evidence*: a forensic analyst can independently verify that (1) the Benford distribution shows anomaly, (2) compression metadata indicates processing, and (3) their correlation is suspicious.

3) Comparison: Authentic vs. Manipulated Attention: Quantitative analysis reveals key differences between authentic and manipulated images. Authentic images show distributed attention (mean entropy: 2.76 ± 0.18 bits), reflecting consistency across statistical and contextual features with no single feature pair dominating. In contrast, manipulated images exhibit concentrated attention (mean entropy: 1.94 ± 0.31 bits) on 2–3 feature pairs, specifically those exhibiting statistical-metadata inconsistencies. The model learns to focus forensic reasoning on anomalous correlations.

4) Quantitative Interpretability Metrics: To quantify interpretability, we measure attention concentration using Shannon entropy:

$$H = -\sum_{i,j} \alpha_{ij} \log \alpha_{ij} \tag{10}$$

where α_{ij} denotes attention weight between token i and j.

Across our test set:

- Mean attention entropy (authentic): 2.76 ± 0.18 bits
- Mean attention entropy (manipulated): 1.94 ± 0.31 bits
- Difference statistically significant ($p < 0.01$, two-sample t-test)

Lower entropy in manipulated images indicates concentrated attention on anomalous features—precisely the behavior desired for interpretable forensic detection.

5) Implications for Trustworthy Forensics: The attention-based interpretability provides three practical advantages:

1. Verifiability: Forensic analysts can independently verify the statistical and metadata features receiving high attention, confirming or refuting the model's implicit reasoning.

2. Transparency: Rather than accepting opaque predictions, analysts understand *why* an image was flagged—enabling informed decisions about whether to trust the classification.

3. Debugging: When the model fails, attention patterns reveal which feature correlations drove the incorrect decision, enabling targeted improvements.

These properties are essential for deploying AI in high-stakes forensic contexts where decisions must be justified and auditable. While Random Forest achieves 2% higher accuracy, it cannot provide this level of transparent reasoning, limiting its utility for real-world forensic workflows.

VI. SYSTEM IMPLEMENTATION AND WEB PROTOTYPE

To demonstrate the practical applicability of our framework, we implemented a web-based prototype that allows users to upload an image and receive a fake/real prediction together with basic forensic cues.

The system follows a lightweight client–server architecture. The frontend is implemented in JavaScript using a Node.js/Express server and EJS-based templates. Users access the application through a browser, where they can upload candidate images for analysis. Uploaded images are handled via `multer` and stored in Amazon S3, with metadata and logging information recorded in Azure SQL. For each uploaded image, the backend computes DCT-based Benford features, extracts metadata, and feeds the resulting feature vectors into the trained multimodal transformer model to generate an authenticity prediction. Figure 5 shows the system homepage, while Figures 6 and 7 illustrate example predictions for authentic and manipulated images.

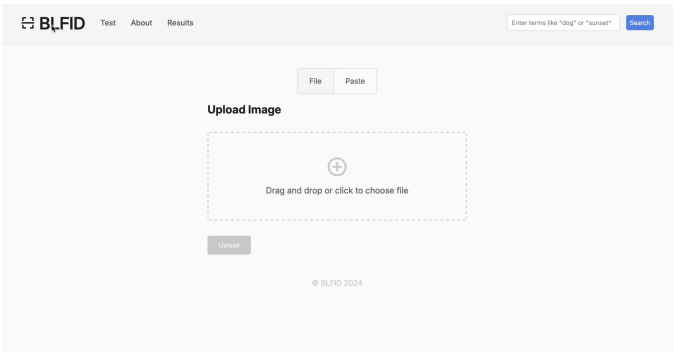


Fig. 5. The BLFID (Benford's Law Fraudulent Image Detector) System.

VII. CONCLUSION

This paper presents a proof-of-concept framework integrating classical statistical forensics (Benford's Law applied to DCT coefficients) with modern multimodal transformers for interpretable image manipulation detection. Our key contribution is demonstrating that attention mechanisms can learn meaningful correlations between statistical anomalies

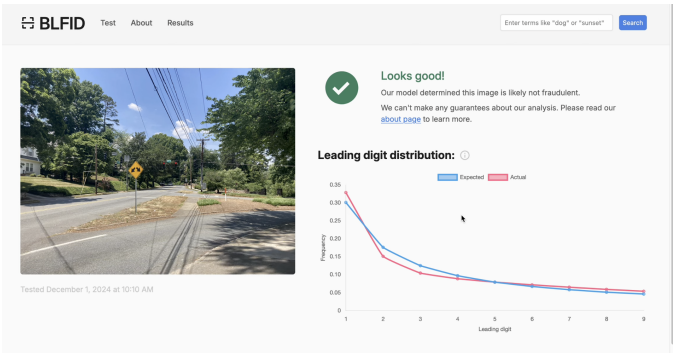


Fig. 6. Example result: image detected as not fraudulent.

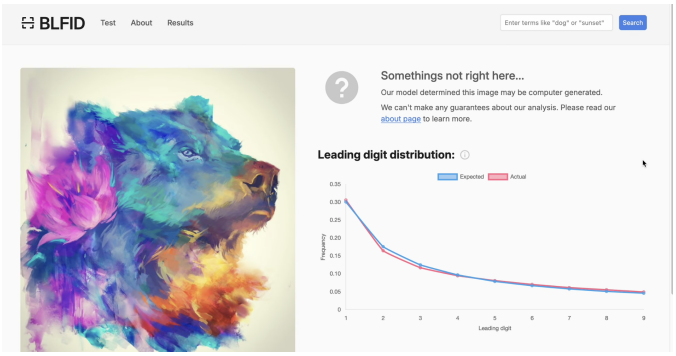


Fig. 7. Example result: image detected as computer generated.

and metadata inconsistencies, providing transparent forensic reasoning rather than opaque binary predictions.

Our design decisions reveal several key lessons for multimodal forensic systems. First, separate embedding layers for statistical and metadata features (Equation 5) proved essential—direct concatenation without learned projections degraded performance by $\sim$5%. Second, four transformer layers with four attention heads provided optimal balance between expressiveness and overfitting on our limited dataset; deeper architectures showed no improvement. Third, the Benford regularization term ($\lambda \mathcal{L}_{\text{Benford}}$) improved interpretability by encouraging forensically consistent attention patterns, though its impact on raw accuracy was modest ($\sim$1%). Finally, integrating frozen CLIP/SAM encoders enabled feature enrichment without increasing training complexity, suggesting that hybrid classical-learned feature spaces are promising for low-resource forensic scenarios.

Experimental results on our controlled dataset show that multimodal fusion improves classification accuracy from 86.5% (Benford features alone) to 93.7% (transformer with metadata integration). While Random Forest achieves slightly higher accuracy (95.6%), our transformer architecture provides critical interpretability advantages: attention visualization reveals which feature combinations drive decisions, enabling forensic analysts to verify and validate model reasoning. Quantitative analysis shows manipulated images exhibit concentrated attention (entropy: 1.94 bits) on anomalous feature pairs, compared to distributed attention (entropy: 2.76 bits) in authentic images.

This work establishes proof-of-concept on a small, controlled dataset of high-quality image pairs with non-adversarial manipulations. The approach requires validation on large-scale benchmarks with diverse manipulation types, adversarial attacks, and real-world compression artifacts. Few-shot learning experiments with CLIP/SAM remain preliminary and need comprehensive evaluation.

Critical limitations include: (1) Resizing disrupts Benford distributions, limiting applicability to social media pipelines where rescaling is mandatory—developing scale-invariant statistical features is essential future work. (2) Format conversion effects (JPEG $\rightarrow$ PNG, re-compression) remain unvalidated and may introduce artifacts that affect detection reliability.

Key research directions include: (1) scaling to large forensic benchmarks (e.g., FaceForensics++), (2) evaluating robustness against GAN-generated content and adversarial manipulations, (3) developing scale-invariant Benford features to handle resized images, (4) assessing performance across format conversions and compression levels, (5) extending few-shot adaptation capabilities, and (6) developing user-facing tools for journalists and investigators. The integration of statistical priors with attention-based interpretability provides a promising path toward trustworthy AI for media authentication in safety-critical forensic applications.

ACKNOWLEDGMENTS

This work was supported by the 2025 Chancellor's Initiative for Transformative Research (CITR) at the University of North Carolina at Greensboro.

REFERENCES

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in Neural Information Processing Systems, vol. 27, 2014.

[2] L. Verdoliva, "Media forensics and deepfakes: an overview," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 910–932, 2020.

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

[4] C. T. Abdallah, G. L. Heileman, and F. Perez-Gonzalez, "Benford's law in image processing," IEEE Int. Conf. Image Process., 2007, pp. 405.

[5] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," arXiv:1908.04626, 2019.

[6] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," IEEE/CVF CVPR, 2021, pp. 782–791.

[7] D. Crișan, A. Irimia, D. Gota, L. Miclea, A. Puscasiu, O. Stan, and H. Valean, "Analyzing Benford's law's powerful applications in image forensics," Applied Sciences, vol. 11, no. 23, p. 11482, 2021.

[8] N. Singh and R. Bansal, "Analysis of Benford's law in digital image forensics," Int. Conf. Signal Process. Commun. (ICSC), 2015, pp. 413–418.

[9] X. Wang, ”Final report for an implementation on the paper ‘Understanding Benford’s law and its vulnerability in image forensics’, Technical Report, University of Winnipeg.

[10] A. Rossler, 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.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, ”Learning transferable visual models from natural language supervision,” Int. Conf. Mach. Learn. (ICML), 2021, pp. 8748–8763.

[12] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, ”Segment anything,” IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 4015–4026.

[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, ”A simple framework for contrastive learning of visual representations,” Int. Conf. Mach. Learn. (ICML), 2020, pp. 1597–1607.

[14] J. Wang, Z. Wu, J. Chen, and Y.-Q. Han, “UniForensics: Face forgery detection via general facial representation,” in *Proc. IEEE/CVF CVPR*, 2024, pp. 15264–15274.

[15] S. Agarwal and H. Farid, “Detecting deep-fake videos from appearance and behavior,” in *IEEE Int. Workshop Inf. Forensics Security*, 2023, pp. 1–6.

[16] M. Cornia, L. Baraldi, R. Cucchiara, and A. Del Bimbo, “Explaining transformer-based image captioning models: An empirical analysis,” *Artif. Intell.*, vol. 326, p. 104039, 2024.

[17] H. Liu, Z. Tan, C. Tan, Y. Wei, Y. Zhao, and J. Wang, “Forgery-aware adaptive transformer for generalizable synthetic image detection,” arXiv preprint arXiv:2312.16649, 2023.

[18] K. Guo, H. Zhu, and G. Cao, “Effective image tampering localization via enhanced transformer and co-attention fusion,” arXiv preprint arXiv:2309.09306, 2023.

[19] M. Shaikh and A. Al-Shaibani, “Synthetic media detection using Benford’s law,” SSRN Scholarly Paper 5176653, 2024.

[20] P. Fernandes, A. Guedes, A. Silva, and M. Magalhães, “Benford’s law applied to digital forensic analysis,” *Forensic Science International: Digital Investigation*, vol. 44, p. 301589, 2023.

[21] S. Singh and A. Dhumane, “Unmasking digital deceptions: An integrative review of deepfake detection, multimedia forensics, and cybersecurity challenges,” *MethodsX*, vol. 12, p. 103970, 2025.