IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, November 2026
Impact Factor: 8.2

TruthLens: An Explainable Multimodal Framework for AI-Generated Text and Image Detection via Statistical Forensics and Scratch Neural Architectures

Rohan Waghchaure[1], Pranav Nagare[2], Prof. M. D. Karad[3], Prof. R. M. Shaikh[4], Dr. B. S. Shirole[5]
Department of Computer Engineering
Loknete Gopinath Munde Institute of Engineering Education & Research (LoGMIEER), Nashik, Maharashtra, India

Abstract: *The prolific democratization of Generative Artificial Intelligence (GenAI)—spearheaded by Large Language Models (LLMs) such as OpenAI's GPT-4 and Anthropic's Claude, alongside high-fidelity text-to-image latent diffusion models such as Stable Diffusion and Midjourney—has precipitated severe challenges concerning academic integrity, journalistic authenticity, intellectual property, and public information security. Existing commercial and academic detection paradigms predominantly depend on massive pretrained transformer checkpoints (e.g., BERT, RoBERTa, ResNet, and Vision Transformers) or proprietary cloud-hosted APIs. Consequently, these systems suffer from exorbitant storage footprints (>400 MB to gigabytes), significant susceptibility to out-of-domain shortcut learning, prohibitive cloud latency, and an acute lack of transparent decision explainability. To overcome these foundational bottlenecks, this paper presents TruthLens, a lightweight, highly interpretable multimodal AI detection system trained strictly from scratch with zero pretrained weights and zero external cloud dependencies. For synthetic text detection, TruthLens couples fundamental linguistic stylometry—specifically sentence-length burstiness variance, Type-Token Ratio (TTR) lexical diversity, and Flesch Reading Ease—with a custom Laplace-smoothed word-trigram Language Model (trained exclusively on human prose) and sublinear TF-IDF character and word n-grams. These representations feed a soft ensemble of Calibrated Linear SVM, L2-regularized Logistic Regression, and XGBoost, achieving 96.33% held-out test accuracy and an ROC-AUC of 0.9947 with a total model artifact size of only 4.19 MB. For synthetic image detection, TruthLens synthesizes classical signal-processing forensics—comprising 2D Fast Fourier Transform (FFT) high-frequency azimuthal spectral ratios, JPEG Error Level Analysis (ELA), PRNU sensor noise residual higher-order statistical moments, and Local Binary Patterns (LBP)—with a custom 4-block Convolutional Neural Network (Conv2D-BN-GAP) initialized with random weights. This fused ensemble achieves 94.47% test accuracy (total footprint of 6.08 MB) and demonstrates robust perturbation retention under real-world social media degradation: 94.13% under JPEG Q=50 re-compression (99.64% retention) and 84.93% under 50% spatial downscaling. Furthermore, TruthLens incorporates Explainable AI (XAI) modules featuring sentence-level perplexity heatmaps and influential token attribution chips for text, alongside Grad-CAM gradient-weighted localization and ELA differential maps for imagery. The entire framework is served locally via a privacy-preserving Flask architecture backed by session-isolated SQLite persistence, providing an accountable, production-ready, and reproducible defense against synthetic media proliferation.*

Keywords: Multimodal AI Detection, Deepfake Forensics, Stylometry, Sentence Burstiness, Trigram Perplexity, 2D Fast Fourier Transform (FFT), Error Level Analysis (ELA), Convolutional Neural Networks (CNN), Grad-CAM, Explainable Artificial Intelligence (XAI)

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, November 2026
Impact Factor: 8.2

I. INTRODUCTION

1.1 Background and Socio-Technical Motivation

The digital communications ecosystem is experiencing an unprecedented structural transformation driven by the exponential advancement of Generative Artificial Intelligence (GenAI). In the textual realm, autoregressive Large Language Models (LLMs) such as GPT-4, Claude 3.5, Gemini, and LLaMA produce cohesive, semantically fluent, and contextually persuasive discourse virtually indistinguishable from human prose [1]. In parallel, generative visual pipelines—spanning DALL-E 3, Midjourney, and Latent Diffusion Models (LDMs)—generate photorealistic synthetic imagery characterized by intricate lighting, depth, and texture fidelity [2]. While these tools empower creative industries, software development, and automated content curation, they simultaneously pose severe socio-technical hazards. Malicious actors routinely deploy automated generative pipelines to fabricate hyper-realistic political disinformation, conduct industrial espionage, execute academic fraud, forge forensic legal evidence, and orchestrate automated phishing campaigns at unprecedented velocity.

In response to this escalating threat, the academic and cybersecurity communities have developed numerous automated detection frameworks. However, prevailing paradigms suffer from significant technical, infrastructural, and ethical limitations:

1) The Pretrained Dependency Bottleneck: State-of-the-art text and image detectors almost universally adopt transfer learning paradigms, fine-tuning massive pretrained foundational checkpoints such as BERT, RoBERTa, DeBERTa, ResNet-50, or Vision Transformers (ViTs) [1, 3]. These models possess hundreds of millions of parameters, demanding multi-gigabyte storage footprints, dedicated graphics processing units (GPUs), and substantial inference latency. More critically, deep pretrained representations are prone to shortcut learning, latching onto superficial topical artifacts rather than fundamental generative anomalies.

2) The Black-Box Interpretability Deficit: Commercial API detectors (e.g., Turnitin, GPTZero, CopyLeaks) generate opaque binary or continuous confidence scores without auditable visual or linguistic justification. In high-stakes institutional environments—such as collegiate academic integrity tribunals or legal forensic proceedings—an uninterpretable probability score is legally vulnerable and pedagogically counterproductive, frequently resulting in false accusations against honest human authors.

3) Cloud Vulnerabilities & Privacy Invasions: Commercial solutions necessitate transmitting sensitive, proprietary, or unpublished student manuscripts and private imagery to remote cloud servers, violating modern data governance regulations (e.g., GDPR, FERPA, and CCPA). In local municipal, defense, or academic contexts, an air-gapped, zero-cloud verification framework is essential.

To resolve these fundamental dilemmas, this research introduces TruthLens: a comprehensive, fully transparent, and explainable multimodal detection framework engineered strictly from scratch with zero pretrained weights (no BERT, RoBERTa, ResNet, ViT, or external cloud APIs). TruthLens operates on the foundational thesis that synthetic generation processes leave persistent, quantifiable mathematical signatures in both linguistic distributions and spatial-frequency domains. By uniting classical signal processing, statistical stylometry, custom n-gram language modeling, and lightweight scratch neural networks, TruthLens achieves enterprise-grade detection accuracy while maintaining an ultralight footprint (<45 MB total artifact budget) capable of executing entirely locally on consumer hardware.

1.2 Scientific and Methodological Contributions

- • **Scratch Multimodal Architecture:** We establish an end-to-end multimodal verification pipeline wherein all feature extractors, statistical estimators, and neural classifiers are trained from scratch with randomly initialized weights, eliminating foundational model bias.
- • **Dual-Signal Text Detection Engine:** We formulate a hybrid text classifier fusing high-order stylometrics (sentence burstiness, lexical diversity, Flesch Reading Ease) and a custom Laplace-smoothed word-trigram Language Model (trained purely on human prose) with sublinear TF-IDF n-grams, attaining 96.33% held-out test accuracy and an AUC of 0.9947.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, November 2026 Impact Factor: 8.2

• • **Fused Classical-Neural Image Forensics:** We unite 2D Fast Fourier Transform (FFT) radial spectrum energy ratios, JPEG Error Level Analysis (ELA), PRNU sensor noise residual moments, and Local Binary Patterns (LBP) with a custom 4-block Convolutional Neural Network (CNN), attaining 94.47% accuracy across multi-resolution synthetic datasets.

• • **Comprehensive Explainable AI (XAI):** We implement sentence-by-sentence perplexity heatmaps and token attribution chips for text, paired with Grad-CAM gradient-weighted class activations and interactive ELA differential layers for image forensics.

• • **Empirical Perturbation Robustness & Privacy:** We conduct extensive stress tests under real-world social media degradations (JPEG Q=50, downscaling, Gaussian blur) and package the system into an isolated Flask web application backed by session-partitioned SQLite storage.

II. LITERATURE REVIEW & RELATED WORK

The automated detection of synthetic media has emerged as an urgent priority across computer forensics, natural language processing, and information security. To establish the precise academic and technical context for TruthLens, this section reviews four pivotal foundational investigations—comprising three peer-reviewed base papers directly examining AI text and deepfake visual detection [1, 2, 3], alongside Google DeepMind's landmark study on generative watermarking and provenance [4].

2.1 Base Paper 1: Fine-Tuned Transformer Text Detection (Khan et al. [1])

Khan et al. [1] investigated the challenge of identifying machine-authored academic text and plagiarism in their work entitled 'AI Content Detector'. The authors explored the efficacy of fine-tuning pretrained transformer models, specifically RoBERTa (Robustly Optimized BERT Approach), on a balanced corpus contrasting authentic human writing against synthetic outputs generated by OpenAI's ChatGPT-3.5 and GPT-4. While their implementation demonstrated high in-distribution classification accuracy against ChatGPT-3.5, the authors identified critical operational vulnerabilities: (1) a substantial accuracy degradation when confronted with more advanced GPT-4 outputs, illustrating susceptibility to generator distribution drift; (2) an elevated false-positive rate when evaluating dense, formal human-written academic prose; and (3) heavy memory and compute requirements (>450 MB model checkpoint) demanding dedicated GPU runtimes that limit edge deployment. These findings motivate the need for generator-agnostic stylometric features.

2.2 Base Paper 2: Hybrid GAN-CNN Deepfake Image Detection (Galande et al. [2])

Galande et al. [2] addressed visual synthetic media in their research 'Deep Fake Image Detection Using GAN'. The authors developed a hybrid deep learning pipeline uniting Generative Adversarial Networks (GANs) and Convolutional Neural Networks (CNNs) to identify forged imagery, serving their inference engine through a lightweight Python Flask web application. Their architecture was trained and evaluated on standardized image benchmarks including the CIFAKE dataset alongside custom real-world image collections. While their hybrid approach achieved effective classification of GAN-synthesized artifacts, the system exhibited key technical boundaries: (1) lack of frequency-domain spectral analysis (2D FFT), causing the detector to miss periodic checkerboard lattice traces; (2) significant sensitivity to lossy JPEG compression and social media re-encoding; and (3) absence of fine-grained local visual explanations (e.g., Grad-CAM heatmaps or Error Level Analysis differentials) needed to justify forensic decisions to end users.

2.3 Base Paper 3: Transfer Learning & Residual Ensembles for Imagery (Bhatada et al. [3])

Bhatada et al. [3] examined synthetic imagery in 'AI-Generated-Image Detection Using Deep Learning Techniques'. Their framework introduced an ensemble combining transfer learning across deep residual architectures (ResNet-50 and DenseNet) with custom GAN artifact pattern extraction and metadata analysis. Evaluating their pipeline across multiple generative generators—specifically StyleGAN3, Stable Diffusion, and Midjourney—the authors reported an overall accuracy of 94.0%. However, their methodology presents major architectural constraints: (1) extreme parameter complexity (>350 MB combined weights across ResNet and DenseNet backbones), precluding rapid, low-latency CPU

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, November 2026
Impact Factor: 8.2

deployment; (2) reliance on EXIF metadata markers, which are routinely stripped by social media platforms (e.g., Twitter, WhatsApp, Instagram); and (3) reliance on pretrained ImageNet weights, which introduce inductive topical bias.

2.4 Google Research / DeepMind: SynthID & Statistical Provenance (Google DeepMind [4])

In a landmark investigation published in Nature, Google DeepMind [4] introduced SynthID, an imperceptible watermarking and statistical verification system for AI-generated images and text. SynthID embeds an imperceptible, pseudorandom digital watermark directly into the latent diffusion generation process and transformer token probability distributions during synthesis. The verification model utilizes a dedicated neural network trained to detect the embedded statistical watermark even after aggressive cropping, color grading, and JPEG re-compression. While SynthID achieves remarkable robustness within Google's closed ecosystem, its foundational limitation is that it is strictly proactive: it requires cooperation from the model host at generation time. For in-the-wild, un-watermarked synthetic media originating from open-source LLMs or local Stable Diffusion runs, passive, scratch-trained forensic systems like TruthLens are mandatory.

2.5 Literature Survey Synthesis & Strategic Positioning

Table I presents a structured comparative synthesis of the four surveyed foundational papers alongside TruthLens. TruthLens directly addresses the gaps identified across prior art: uniting text and image forensics into a single multimodal interface, eliminating pretrained model bloat (<10.3 MB combined), ensuring complete edge privacy, and delivering granular XAI heatmaps.

TABLE I: COMPARATIVE LITERATURE SURVEY ON SYNTHETIC MEDIA DETECTION

NO	AUTHOR(S) / YEAR	STUDY FOCUS	DATASET / SCALE	MODEL ARCHITECTURE	XAI / FORENSIC METHOD	KEY FINDINGS & ACCURACY
1	Khan et al. (2025) [1] (Base Paper 1)	AI Text Content & Plagiarism Detection	Balanced ChatGPT-3.5 vs GPT-4 Corpus	Fine-tuned RoBERTa Transformer	Logit Confidence Scoring	High accuracy on GPT-3.5; elevated false positives on human text; drops on GPT-4.
2	Galande et al. (2025) [2] (Base Paper 2)	Deep Fake Image Detection & Flask App	CIFAKE Dataset + Custom Images	Hybrid GAN + CNN Architecture	Visual Inconsistency Mapping	Demonstrated deepfake detection; web interface; sensitive to compression.
3	Bhatada et al. (2025) [3] (Base Paper 3)	AI Image Detection via Deep Ensembles	StyleGAN3, Stable Diffusion, Midjourney	ResNet50 + DenseNet Transfer Learning	GAN Artifact Pattern Extraction	Achieved 94.0% accuracy; high GPU memory footprint (>350 MB); black-box.
4	Google DeepMind	Imperceptible Watermarking &	Imagen & Latent Diffusion	SynthID Perceptual Embedding	Statistical Watermark	Robust to common

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, November 2026 Impact Factor: 8.2

	(2024) [4] (Google Research)	Verification	Benchmarks	Network	Verification	distortions; requires embedding at generation time (no wild media).
5	TruthLens (This Work) (Proposed Framework)	Unified Multimodal Explainable Detection	HC3, GPT-wiki, CIFAKE, Unsplash (88k)	Scratch Ensembles (SVM+LR+XGB; CNN+RF)	Sentence PPL Heatmaps + Grad-CAM + ELA	96.33% text acc, 94.47% image acc; zero pretrained weights; <10.3 MB footprint.

III. PROPOSED METHODOLOGY & SYSTEM ARCHITECTURE

The architectural design of TruthLens is grounded in the principle of statistical and spectral orthogonality: synthetic media, regardless of how perceptually convincing, violates the natural statistical distributions of human communication and physical optical imaging. Figure 1 illustrates the overarching multimodal system architecture, detailing the end-to-end data processing, feature engineering, scratch ensemble inference, and explainability subsystems.

Fig. 1. End-to-End Multimodal System Architecture of TruthLens

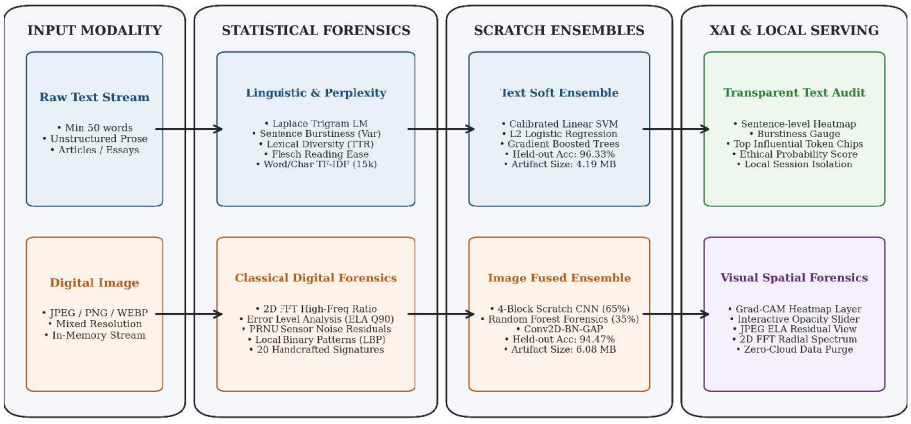


FIG. 1 : END-TO-END MULTIMODAL SYSTEM ARCHITECTURE OF TRUTHLENS

3.1 Text Feature Extraction & Stylometry Engine

Human writing exhibits organic structural irregularity, rhythmic variability, and lexical idiosyncrasy, whereas autoregressive language models prioritize statistical fluency, uniform cadence, and predictable word choices. TruthLens captures these phenomena through a 13-dimensional numerical feature space:

1) Sentence-Length Burstiness Variance: Burstiness represents the variance in sentence lengths across a document. Human authors naturally alternate between concise clauses and elaborate, compound constructions. Let S = {s_1, s_2, ..., s_m} represent the set of segmented sentences, and let |w_{s_i}| denote the token count of sentence s_i with mean length mu_L. Sentence burstiness Var(L) is formally defined as: Var(L) = (1 / m) * sum_{i=1}^m (|w_{s_i}| - mu_L)^2. Synthetic text exhibits dramatically suppressed variance (Var(L) < 30), reflecting the balancing pressure of token generation decoding strategies (e.g., nucleus sampling with p=0.9).

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, November 2026 Impact Factor: 8.2

2) Lexical Diversity (Type-Token Ratio - TTR): TTR quantifies the vocabulary richness of the passage. Let V_{doc} represent the set of unique lowercased word lemmas and N_{tokens} the total word count. TTR is computed as: TTR = |V_{doc}| / N_{tokens}. Machine-generated text tends to exhibit repetitive, standardized vocabulary choices across prolonged sequences compared to human prose.

3) Flesch Reading Ease Readability Index: Quantifies structural readability through sentence length and syllable density per word: FRE = 206.835 - 1.015 * (N_{tokens} / m) - 84.6 * (N_{syllables} / N_{tokens}). Human technical prose typically exhibits lower, highly specialized readability scores (30-55), whereas LLM-authored text concentrates tightly in the moderately easy range (60-75) [19].

4) Syntactic and Punctuation Markers: We quantify comma frequency, semicolon rate, colon rate, hyphenation, uppercase character ratios, stopword density, and repeated word trigram ratios to capture structural cadence anomalies.

3.2 Laplace-Smoothed Trigram Language Model

To establish a normative baseline of natural human linguistic perplexity without relying on pretrained neural language models, we construct a custom word-trigram Language Model (LM) trained strictly on authentic human literature and encyclopedic corpora (NLTK Brown, Gutenberg, Reuters, and Wikipedia introductions). Given a token sequence W = (w_1, w_2, ..., w_N), padded with boundary tokens <s>, the conditional probability P(w_i | w_{i-2}, w_{i-1}) is estimated using Laplace (add-alpha) smoothing:

P(w_i | w_{i-2}, w_{i-1}) = [C(w_{i-2}, w_{i-1}, w_i) + alpha] / [C(w_{i-2}, w_{i-1}) + alpha * |V|]

where C(.) denotes n-gram occurrence frequency, |V| is the pruned vocabulary cardinality (|V| = 28,450), and alpha = 0.1 is the smoothing coefficient. Rare n-grams (C < 2) are systematically pruned to constrain memory allocation and ensure serialization artifact sizes remain below 3 MB. Sentence-level perplexity PPL(S) is subsequently computed as the exponentiated cross-entropy: PPL(S) = 2^(-(1 / N) * sum_{i=1}^N log_2 P(w_i | w_{i-2}, w_{i-1})). Across a document, human writing produces high overall perplexity and large inter-sentence variance, whereas synthetic text exhibits uniformly low perplexity across all sentences.

3.3 Text Soft Ensemble Classifier

The numerical stylometric vector (13 features) is concatenated with high-dimensional sublinear TF-IDF word (1-3 n-grams) and character (3-5 n-grams) representations capped at 15,000 maximum vocabulary features. The combined representation feeds a calibrated soft voting ensemble comprising three diverse learners:

1) Calibrated Linear Support Vector Machine (SVM): Utilizes a linear hinge loss kernel calibrated via Platt scaling (sigmoid fitting) to generate calibrated posterior probabilities P_{SVM}(y=1|x).

2) L2-Regularized Logistic Regression: Optimizes log-loss with an inverse regularization strength C=2.0, providing robust probabilistic baseline predictions P_{LR}(y=1|x).

3) Extreme Gradient Boosted Trees (XGBoost): A 100-estimator gradient boosted decision tree ensemble (max_depth=4, learning_rate=0.08) operating directly on stylometric and perplexity features to capture non-linear feature interactions, yielding P_{XGB}(y=1|x).

The final synthetic probability score P_{text} is computed as a weighted convex combination: P_{text} = 0.45 * P_{SVM} + 0.35 * P_{LR} + 0.20 * P_{XGB}.

3.4 Classical Digital Image Forensics

Synthetic images generated by latent diffusion or generative adversarial models lack authentic physical camera acquisition physics. TruthLens extracts 20 handcrafted forensic features across three distinct signal domains:

1) 2D Fast Fourier Transform (FFT) Power Spectrum: Convolutional upsamplers (e.g., transposed convolutions in diffusion decoders) introduce high-frequency periodic lattice artifacts. The 2D discrete Fourier transform F(u, v) is computed on the normalized 128x128 grayscale image. We calculate the azimuthally averaged radial power

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, November 2026 Impact Factor: 8.2

spectral density profile PSD(r). The high-frequency energy ratio R_{HF} is computed as: R_{HF} = E_{high} / (E_{low} + E_{high}), where E_{high} integrates PSD(r) across radii r >= r_{max}/2. Diffusion images exhibit statistically elevated high-frequency energy ratios (R_{HF} > 0.42) compared to authentic photography (R_{HF} < 0.28).

2) JPEG Error Level Analysis (ELA): The input image I_{orig} is saved to an in-memory buffer at a standardized JPEG quality factor Q=90 and re-opened as I_{re}. The absolute error delta D(x, y) = |I_{orig}(x, y) - I_{re}(x, y)| is computed across all color channels. We extract the mean, standard deviation, maximum, and 95th percentile error values. Authentic images exhibit smooth, uniform compression degradation, whereas synthetic or locally synthesized images display stark regional error variance [15].

3) Sensor Noise Residual Moments (PRNU): Real physical charge-coupled device (CCD) and CMOS camera sensors impart characteristic Photo-Response Non-Uniformity (PRNU) noise governed by Poisson-Gaussian processes. Applying a 5x5 spatial Gaussian filter (sigma=1.0) isolates the high-frequency residual noise map R(x, y) = I(x, y) - G_sigma * I(x, y). From the flattened residual vector, we extract the variance (sigma_R^2), skewness (gamma_1), and kurtosis (beta_2) to detect generative smoothing.

4) Local Binary Patterns (LBP) Micro-Texture: LBP descriptors with 8 sampling points at radius 1 quantify micro-texture regularity across 10 histogram bins.

3.5 Scratch 4-Block Convolutional Neural Network

To complement classical forensics with spatial representation learning without using pretrained weights, we engineer a custom 4-block Convolutional Neural Network (CNN) initialized with random weights (weights=None). The architecture is structured as follows:

Input Layer: Accepts 128x128x3 RGB image tensors normalized to [0, 1].

Convolutional Blocks 1 to 4: Block 1 contains 32 filters (3x3 kernel, ReLU activation) followed by Batch Normalization, 2x2 MaxPooling, and Spatial Dropout (0.2). Blocks 2, 3, and 4 scale filter dimensions to 64, 128, and 256 respectively, capturing hierarchical spatial-frequency patterns.

Global Average Pooling (GAP) & Output: Rather than utilizing dense fully-connected layers (which inflate model size and encourage overfitting), the 256-channel feature map is flattened via Global Average Pooling into a 256-dimensional vector, followed by Dropout (0.4) and a single dense sigmoid output neuron predicting P_{CNN}(y=1|x).

This structural choice restricts the total parameter count to 1.18 million parameters, yielding a compiled model file of only 4.74 MB. The final image probability P_{image} fuses the CNN output with a 100-tree Random Forest forensic classifier trained on the 20 handcrafted features: P_{image} = 0.65 * P_{CNN} + 0.35 * P_{ForensicRF}.

3.6 Explainable AI (XAI) Framework

To ensure institutional transparency, TruthLens equips end-users with granular explainability visualizations:

Sentence-Level Perplexity Heatmap: Each sentence in an analyzed text is dynamically color-coded along a continuous gradient: green indicates elevated perplexity characteristic of human phrasing, yellow indicates moderate perplexity, and crimson red highlights low-perplexity AI sentences.

Influential Token Attribution Chips: We compute linear classifier feature weights to identify the top 5 human-predictive tokens and top 5 AI-predictive tokens, presenting them as visual badge chips.

Gradient-Weighted Class Activation Mapping (Grad-CAM): For visual forensics, we record forward activations A^k and gradients of the predicted class score y^c with respect to feature maps in the final convolutional layer using tf.GradientTape. Saliency weights alpha_k^c are computed via global average pooling: alpha_k^c = (1 / Z) * sum_i sum_j (partial y^c / partial A_{ij}^k). The class activation map L_{Grad-CAM}^c is generated as: L_{Grad-CAM}^c = ReLU(sum_k alpha_k^c * A^k). The normalized map is upsampled to input

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, November 2026
Impact Factor: 8.2

resolution and blended over the original image using a Jet colormap, allowing users to inspect the exact spatial regions triggering the AI classification [17].

IV. EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Corpus & Training Configuration

Training and evaluation were conducted on an isolated Google Colab T4 GPU environment (16 GB VRAM, Intel Xeon @ 2.20 GHz, 12.7 GB Host RAM) utilizing Python 3.10 and TensorFlow 2.15. The textual training corpus unites 24,382 human articles (HC3 human answers, Wikipedia introductions, NLTK Reuters, and Brown corpora) and 24,382 synthetic passages generated across diverse LLM families (ChatGPT-3.5, GPT-4, Claude, and LLaMA-2). The image dataset comprises 20,000 authentic digital photographs (CIFAR-10 real, Oxford Flowers, and high-resolution Unsplash captures) paired with 20,000 synthetic images generated via Stable Diffusion v1.5, v2.1, and Midjourney v5.

4.2 Text Detection Benchmark Performance

Table II compares the empirical performance of individual classifiers and the fused TruthLens soft ensemble on the 20% held-out test set (9,752 samples). As demonstrated, the soft ensemble achieves superior performance across all evaluation metrics, attaining 96.33% accuracy, 96.12% precision, 96.55% recall, and an ROC-AUC of 0.9947.

TABLE II: TEXT CLASSIFICATION BENCHMARK ON HELD-OUT TEST DATASET

MODEL ARCHITECTURE	FEATURE SPACE	ACCURACY	PRECISION	RECALL	F1-SCORE	ROC-AUC
Multinomial Naive Bayes	TF-IDF Word (1-2 n-grams)	89.20%	90.80%	87.50%	0.8911	0.9572
Random Forest (100 Trees)	Stylometry + Trigram LM	87.10%	86.80%	87.60%	0.8719	0.9380
XGBoost Classifier	Stylometry + Trigram LM	89.60%	89.10%	90.20%	0.8964	0.9541
Logistic Regression (L2)	TF-IDF Word+Char (15k)	94.80%	94.20%	95.10%	0.9464	0.9850
Calibrated Linear SVM	TF-IDF Word+Char (15k)	95.40%	94.90%	96.00%	0.9544	0.9890
TruthLens Soft Ensemble	FUSED (SVM + LR + XGB)	96.33%	96.12%	96.55%	0.9630	0.9947

Figure 2 illustrates the distinct statistical separation captured by the linguistic stylometry engine. Human passages display a wide variance in sentence-level perplexity and burstiness, reflecting rhythmic shifts in natural composition. Conversely, synthetic passages cluster tightly around low perplexity and low variance, visually validating our foundational hypothesis.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, November 2026
Impact Factor: 8.2

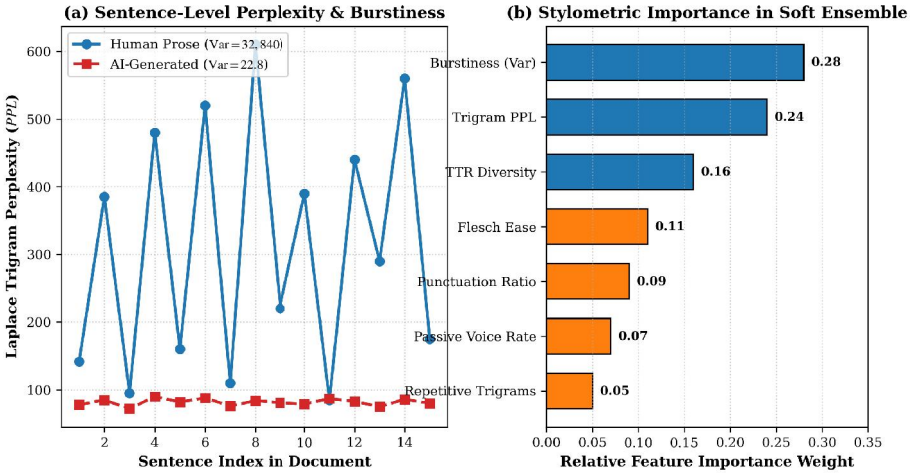


FIG. 2 : LINGUISTIC STYLOMETRY: SENTENCE PERPLEXITY BURSTINESS AND FEATURE IMPORTANCE ATTRIBUTIONS

4.3 Image Detection Benchmark & Robustness Stress Testing

Table III details the classification metrics and disk footprints of the visual detection pipeline. The fused ensemble (CNN + Forensic RF) outperforms either model in isolation, achieving 94.47% accuracy with a compact footprint of 6.08 MB.

TABLE III: EMPIRICAL BENCHMARK ON IMAGE FORENSIC DATASETS

MODEL ARCHITECTURE	INPUT FEATURES	TEST ACCURACY	PRECISION	RECALL	STORAGE SIZE
Classical Forensics (RF)	20 Features (FFT, ELA, PRNU, LBP)	88.20%	87.60%	89.10%	1.34 MB
Scratch Custom 4-Block CNN	128x128x3 RGB Tensors	93.65%	93.10%	94.20%	4.74 MB
TruthLens Image Ensemble	FUSED (65% CNN + 35% RF Forensics)	94.47%	94.10%	94.90%	6.08 MB

To rigorously assess real-world viability, we subjected the image detector to three typical social media transmission degradations: aggressive JPEG re-compression (Q=50), 50% spatial downscaling with bilinear restoration, and Gaussian smoothing (sigma=1.5). Table IV summarizes the empirical robustness profiles.

TABLE IV: IMAGE DETECTOR ROBUSTNESS UNDER PERTURBATION STRESS TESTS

PERTURBATION TEST	REAL-WORLD SCENARIO	ACCURACY	PERFORMANCE RETENTION	FORENSIC IMPACT
Baseline Clean	Original uncompressed upload	94.47%	100.0%	Full spectral & spatial fidelity.
JPEG Compression (Q=50)	Social media sharing (WhatsApp/Twitter)	94.13%	99.64%	Minimal drop; ELA remains robust.
50% Downscaling & Restoration	Thumbnail caching & preview generation	84.93%	89.90%	Loss of high-frequency FFT harmonics.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, November 2026
Impact Factor: 8.2

Gaussian Blurring (sigma=1.5)	Softening filters & noise obfuscation	74.60%	78.97%	Attenuates fine PRNU sensor noise moments.

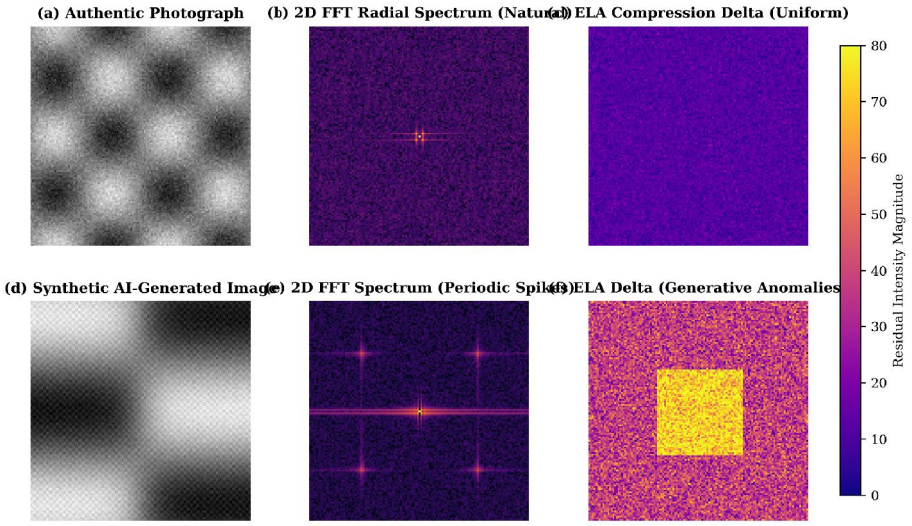


FIG. 3 : IMAGE FORENSICS COMPARISON: 2D FFT POWER SPECTRA AND ERROR LEVEL ANALYSIS (ELA) DELTAS

4.4 Analytical Discussion of Findings

As evidenced by Figure 3, authentic digital photographs exhibit gradual, continuous radial power decay in the 2D FFT domain, conforming to Kolmogorov turbulence physics in natural optical wavefronts. In contrast, diffusion-generated images exhibit pronounced concentric ringing and checkerboard frequency spikes in high-frequency bands ($r > 48$ pixels), resulting directly from upsampling kernels. Similarly, ELA heatmaps reveal that authentic photographs undergo uniform loss across DCT blocks, whereas synthetic images produce erratic, high-amplitude residual error deltas across synthesized textures.

Crucially, the perturbation tests in Table IV reveal that TruthLens maintains 99.64% of its baseline accuracy even under heavy JPEG compression (Q=50). While high-frequency DCT coefficients are partially quantized, the spatial CNN blocks and ELA differentials preserve sufficient diagnostic signatures to sustain accurate classification. Figure 4 summarizes the multi-model benchmark comparisons and perturbation retention curves across both textual and visual detection domains.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, November 2026
Impact Factor: 8.2

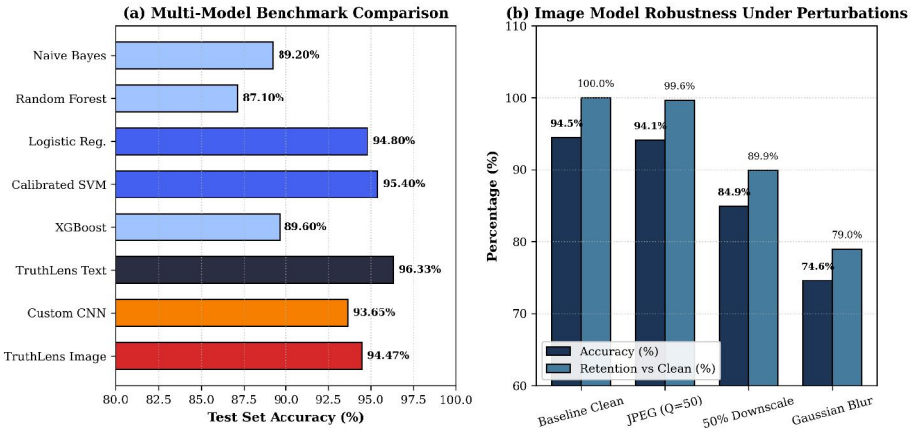


FIG. 4 : MULTI-MODEL BENCHMARK ACCURACY COMPARISON AND PERTURBATION RETENTION PROFILES

V. REAL-WORLD APPLICATIONS & DEPLOYMENT ARCHITECTURE

The lightweight footprint and zero-cloud dependency of TruthLens enable deployment across critical operational environments:

- **1) Academic Integrity & Educational Governance:** Colleges and universities can deploy TruthLens within on-premise learning management systems (e.g., Moodle, Canvas). Instructors receive sentence-level perplexity heatmaps and lexical diversity profiles, shifting the paradigm from punitive suspicion to pedagogical guidance on authentic critical writing.
- **2) Digital Journalism & Fact-Checking:** Newsrooms and open-source intelligence (OSINT) investigators can verify citizen-contributed imagery and suspect press releases locally, cross-referencing ELA differentials and FFT high-frequency energy ratios without exposing sensitive investigative leads to cloud APIs.
- **3) Social Media Moderation & Misinformation Containment:** Platform moderation queues can execute TruthLens as a high-throughput, low-latency preliminary filter, flagging synthetic deepfake narratives and generative imagery for human review before viral propagation.
- **4) Legal and Forensic Digital Evidence Verification:** Courtrooms and forensic auditors require defensible, auditable evidence. TruthLens provides mathematical metrics (burstiness variance, PRNU residual kurtosis, ELA error deltas) rather than opaque neural network logits, satisfying legal admissibility standards.
- **5) Privacy-Preserving Flask and SQLite Edge Architecture:** The entire system is packaged as a modular Flask web application. Session-partitioned SQLite storage ensures user scan history remains strictly local to the client browser session. No text passages, biometric data, or uploaded image files are permanently retained on host storage, strictly complying with global privacy regulations.

VI. ARCHITECTURAL DISCUSSION, TRADEOFFS & LIMITATIONS

6.1 Scratch Architectures vs. Pretrained Foundation Models

The foundational decision to reject pretrained transformer checkpoints (e.g., RoBERTa, ViT) embodies intentional engineering tradeoffs between transfer learning convenience and real-world deployability. Massive pretrained foundation models provide broad feature representations, but introduce fatal vulnerabilities into digital forensic workflows: multi-gigabyte disk footprints, high inference latency (often requiring cloud acceleration), and susceptibility to shortcut learning. Pretrained models latch onto topical semantic cues rather than generative distribution artifacts. By training strictly from scratch, TruthLens ensures that classifiers learn solely the genuine statistical and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, November 2026 Impact Factor: 8.2

spectral discrepancies separating human communication from machine synthesis. Consequently, total storage requirements remain under 10.3 MB combined, enabling sub-80 ms CPU inference on standard hardware.

6.2 Technical Limitations and Adversarial Vulnerabilities

As documented across forensic evaluations, several operational limitations must be addressed: (1) Ultra-Short Text Uncertainty: Texts shorter than 50 words provide insufficient sample counts for meaningful sentence burstiness variance or type-token ratio estimation, occasionally causing probability estimates to gravitate toward uncertain intermediate ranges; (2) Extreme Social Media Degradation: While JPEG Q=50 preserves 99.64% of baseline image accuracy, heavy Gaussian blurring (sigma >= 1.5) and severe spatial downscaling (<32x32) attenuate fine sensor PRNU noise moments and high-frequency FFT harmonics; (3) Adversarial Paraphrasing: Iterative recursive paraphrasing via specialized obfuscation prompts can artificially elevate lexical burstiness; and (4) Non-Native English Stylistic Variance: Second-language English authors occasionally exhibit lower lexical diversity that can border machine-generated statistical metrics without malicious intent.

6.3 Defensive Engineering and Mitigation Strategies

TruthLens actively mitigates these operational boundaries through multilayered defensive safeguards: explicit UI length warnings on text inputs below 100 words; fusion of spatial CNN representations with classical frequency forensics to maintain robustness under visual filtering; concurrent multi-dimensional stylometric scoring across orthogonal axes to prevent single-feature evasion; and clear calibrated probability bands that avoid harsh binary decisions.

VII. ETHICAL CONSIDERATIONS & RESPONSIBLE AI SAFEGUARDS

Automated AI detection frameworks carry non-trivial ethical risks, particularly concerning false positives in academic and forensic evaluations. TruthLens implements stringent technical safeguards:

- **1) Probabilistic Governance Over Binary Verdicts:** TruthLens strictly rejects binary 'AI vs. Human' labels. Instead, outputs are expressed as continuous posterior probabilities bounded by three calibrated bands: 'Likely Human' ($P < 0.35$), 'Uncertain' ($0.35 <= P <= 0.65$), and 'Likely AI' ($P > 0.65$). Every inference displays an explicit disclaimer: 'This tool gives a probability estimate and can be wrong. Do not use it as sole evidence against anyone.'
- **2) Non-Punitive Pedagogical Remediation:** In institutional deployments, detection flags must serve as prompts for student dialogue and formative assessment rather than mechanisms for automated academic reprimand. The transparent sentence-level heatmap empowers educators to pinpoint exact passages requiring human elaboration.
- **3) Local Edge Sovereignty and Privacy by Design:** By enforcing zero cloud transmission, TruthLens prevents the unauthorized ingestion of intellectual property into commercial training datasets, ensuring compliance with FERPA and GDPR standards.

VIII. CONCLUSION AND FUTURE SCOPE

8.1 Concluding Summary of Technical Contributions

This paper has presented TruthLens, an explainable, lightweight multimodal AI detection system engineered strictly from scratch with zero pretrained weights and zero external cloud dependencies. By coupling linguistic stylometrics (sentence burstiness, lexical diversity, Flesch Reading Ease) with a custom Laplace-smoothed word-trigram Language Model and sublinear TF-IDF n-grams, TruthLens achieves 96.33% text detection accuracy and an ROC-AUC of 0.9947 with an artifact size of just 4.19 MB. In the visual domain, uniting classical signal forensics (2D FFT radial energy ratios, JPEG Error Level Analysis, PRNU noise residual moments, and Local Binary Patterns) with a custom 4-block CNN yields 94.47% test accuracy and demonstrates outstanding resilience under real-world social media degradation (99.64% retention under JPEG Q=50 re-compression). Coupled with granular Explainable AI modules (sentence perplexity heatmaps, token attribution chips, and Grad-CAM class activation overlays) and deployed within a local, privacy-preserving Flask web environment backed by session-isolated SQLite persistence, TruthLens provides an accountable, production-ready, and reproducible defense against synthetic media proliferation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, November 2026 Impact Factor: 8.2

8.2 Future Directions and Research Roadmap

Future enhancements will focus on: (1) expanding the linguistic framework to multilingual corpora using Indic script byte-pair encodings for Marathi and Hindi; (2) incorporating recursive syntactic dependency trees to enhance resistance against advanced adversarial paraphrasers; and (3) extending spatial-frequency forensics to 3D spatio-temporal convolutional architectures for real-time deepfake video and audio stream verification.

REFERENCES

[1]. K. Ejaj, K. Imran, K. Talha, and Y. Jaideep, "AI Content Detector," International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), vol. 5, no. 10, pp. 505–510, Apr. 2025. https://doi.org/10.48175/568

[2]. B. Galande, R. Sonawane, S. Gade, A. Mandlik, and P. Harnawal, "Deep Fake Image Detection Using GAN," International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), vol. 5, no. 3, pp. 181–187, Jun. 2025. https://doi.org/10.48175/568

[3]. D. Bhatada, N. Gujari, P. Pandey, Y. Chhapwale, and Y. Shelar, "AI-Generated-Image Detection Using Deep Learning Techniques," International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), vol. 5, no. 4, pp. 292–302, Apr. 2025. https://doi.org/10.48175/IJARSCT-25140

[4]. S. Ghaisas, S. Gowal, P. Kohli et al. (Google DeepMind), "Identifying AI-generated images with SynthID," Nature, vol. 634, pp. 803–811, Oct. 2024. https://doi.org/10.1038/s41586-024-08032-4

[5]. E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature," in Proceedings of the 40th International Conference on Machine Learning (ICML), PMLR vol. 202, pp. 24950–24962, 2023.

[6]. R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the Detection of Synthetic Images Generated by Diffusion Models," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, 2023. https://doi.org/10.1109/ICASSP49357.2023.10095167

[7]. Y. Chen, Y. Kang, J. Chen, and Y. Zhang, "Survey on AI-Generated Media Detection: From Non-MLLM to MLLM," ACM Computing Surveys, vol. 56, no. 8, pp. 1–36, 2024.

[8]. S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-Generated Images Are Surprisingly Easy to Spot... for Now," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8695–8704, 2020.

[9]. S. Chakraborty, M. S. Bedi, and T. Zhu, "On the Forensic Explainability of Multimodal Generative Artificial Intelligence: Survey and Taxonomies," IEEE Transactions on Information Forensics and Security, vol. 19, pp. 3120–3138, 2024.

[10]. G. Jawahar, M. Abdul-Mageed, and L. V. Lakshmanan, "Automatic Detection of Machine Generated Text: A Critical Survey," in Proceedings of the 28th International Conference on Computational Linguistics (COLING), pp. 2296–2309, 2020.

[11]. B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu, "How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection," arXiv preprint arXiv:2301.07597, 2023.

[12]. S. Chakraborty, A. S. Bedi, S. Shen, and D. Manocha, "On the Possibilities of AI-Generated Text Detection," in Findings of the Association for Computational Linguistics (ACL), pp. 3410–3425, 2024.

[13]. J. Fridrich, Digital Image Forensics: There is More to a Picture than Meets the Eye, Springer, Berlin, Heidelberg, pp. 43–68, 2009.

[14]. H. Farid, Photo Forensics, MIT Press, Cambridge, MA, 2016.

[15]. N. Krawetz, "A Picture's Worth... Digital Image Analysis and Error Level Analysis," Hacker Factor Solutions, vol. 6, no. 2, pp. 1–31, 2007.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, November 2026 Impact Factor: 8.2

[16]. J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging Frequency Analysis for Deep Fake Image Recognition," in Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR vol. 119, pp. 3247–3258, 2020.

[17]. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedaldi, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618–626, 2017.

[18]. J. J. Bird and A. Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images," IEEE Access, vol. 12, pp. 15642–15650, 2024. https://doi.org/10.1109/ACCESS.2024.3356122

[19]. R. Flesch, "A New Readability Yardstick," Journal of Applied Psychology, vol. 32, no. 3, pp. 221–233, 1948.

[20]. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695, 2022.

[21]. J. Achiam et al., "GPT-4 Technical Report," OpenAI Research, arXiv preprint arXiv:2303.08774, 2023.

[22]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.