

TruthNet: An AI Powered Multimodel Fake Content Detection

Ashwini R
Department of AI&DS BGSCET
Bengaluru, India

Deekshitha M
Department of AI&DS BGSCET
Bengaluru, India

Venkatesh BP
Department of AI&DS BGSCET
Bengaluru, India

Sumanth R
Department of AI&DS BGSCET
Bengaluru, India

Vrushank S Naik
Department of AI&DS BGSCET
Bengaluru, India

Abstract—The rapid growth of generative image synthesis and the ease with which misleading news content can be produced have created a pressing need for accessible screening tools that do not depend on large computational budgets. This paper presents a light weight multimodal framework that addresses two related but distinct information -integrity problems: screening of manipulated facial images and screening of potentially misleading news text. The framework consists of two independently trained pipelines — an EfficientNet-B0based binary image classifier (V2) built with transfer learning, data augmentation, and staged fine -tuning, and a TF-IDF and logistic-regression text classifier trained on a combined corpus of established fake-news datasets. Both pipelines are exposed through a shared Flask REST API consumed by web and Android client applications. Rather than returning a confidence-independent decision, the system applies uncertainty handling that withholds low -confidence predictions for manual review. On a held -out test partition, the image classifier reaches 80.41% balanced accuracy, with 76.09% recall for fake images and 84.73% recall for real images. Under a source -held-out protocol in which an entire news source is excluded from training, the text classifier attains 65.4% accuracy and a macro -F1 of 0.481, revealing a substantial fake-versus-real performance gap attributable to distribution shift. The framework is positioned as a screening and decision -support aid rather than a definitive forensic or fact-checking system.

Index Terms—Deepfake detection, fake news detection, EfficientNet, transfer learning, TF -IDF, logistic regression, uncertainty -aware prediction, source-held-out evaluation.

I. INTRODUCTION

A. Background

Recent progress in generative modeling has made it substantially easier to produce facial imagery that has been partially or fully synthesized or altered, while advances in automated and semi-automated text generation have made it easier to produce news-style writing that imitates the vocabulary and structure of legitimate reporting without necessarily

reflecting verified facts. These two developments, although technically distinct, contribute to the same underlying concern: the erosion of confidence in visual and textual information encountered online. Addressing this concern does not always require heavyweight forensic pipelines; in many practical settings, a lightweight screening tool that can run on modest computing resources and flag content for closer review is more useful than a specialized system that demands significant infrastructure.

B. Problem Statement

This work addresses two related but methodologically distinct screening problems: (i) determining whether a facial image shows signs consistent with manipulation, and (ii) determining whether a piece of news text shows linguistic patterns consistent with misleading content. The two problems rely on different underlying signals - image manipulation cues are largely spatial and textural, while misleading-text cues are lexical and stylistic - and are therefore addressed with separate, purpose-built models rather than a single shared architecture.

C. Motivation

A recurring observation in applied machine learning is that strong performance on a fixed benchmark split does not guarantee that a model will behave reliably once it is exposed to inputs from a different distribution, a different data source, or a different generation technique than those seen during training. This motivates several design choices adopted in this work: an explicit source-held-out evaluation protocol for the text classifier, a confidence-based rejection mechanism for both pipelines so that borderline predictions are not silently forced into a definite class, and an explicit separation between the completed image model and a still-incomplete follow-up model. The goal throughout is to represent the system's actual, current capability rather than an idealized one.

D. Research Gap

A substantial body of prior work on facial manipulation detection targets high accuracy on curated benchmarks using comparatively large backbones, with less emphasis on lightweight, deployable configurations. Similarly, much of the

published work on fake-news classification reports within-source performance, where training and evaluation data are drawn from the same source distribution, which can overstate how well a classifier will generalize to an unfamiliar publication or platform. Few works combine a lightweight image pipeline and a lightweight text pipeline within a single deployable application while explicitly reporting cross-source generalization and withholding low-confidence predictions rather than treating every output as equally trustworthy. This project is positioned to address that combination.

E. Contributions

The main contributions of this work are summarized as follows: 1) An EfficientNet-B0-based facial-image screening pipeline (V2) built through transfer learning, staged finetuning, data augmentation, threshold selection, and confidence-based rejection. 2) A computationally efficient TF-IDF and logistic-regression classifier for screening potentially misleading news text. 3) A source-aware evaluation protocol in which an entire news source (FakeNewsNet) is withheld during evaluation to estimate cross-source generalization. 4) A shared Flask REST API that exposes both pipelines to web and Android client applications through a consistent JSON interface. 5) An uncertainty-aware decision mechanism, applied to both modalities, that withholds low-confidence predictions instead of forcing every input into a definite class. 6) A hard-example reweighting strategy proposed for a followup image model (V3), presented here as an in-progress design rather than a completed result. These contributions are not presented as individually novel machine-learning techniques; the contribution is the design, integration, and evaluation of a lightweight, uncertainty-aware framework that combines them into a single deployable system.

II. RELATED WORK

A. Deepfake and Facial Manipulation Detection

FaceForensics++ established a widely used benchmark for studying manipulated facial video and imagery and remains a common reference point for facial-manipulation detection research [1]. Much of the subsequent literature has focused on maximizing accuracy on such benchmarks using increasingly large convolutional or transformer-based backbones. A recurring difficulty noted across this literature is that detectors trained on a fixed set of manipulation techniques often lose accuracy when applied to imagery produced by generators not seen during training, which is one of the reasons this work treats prediction confidence explicitly rather than reporting a single aggregate accuracy figure.

B. Lightweight CNN Architectures

Architectures such as MobileNetV2 [3] and the EfficientNet family [4] were designed to reduce parameter counts and computational cost relative to earlier large-scale CNNs while retaining competitive accuracy, primarily through inverted residual blocks and compound scaling of

depth, width, and resolution. EfficientNetV2 [5] extends this line of work with training-aware architecture search and progressive learning to further improve training efficiency. These design principles motivate the choice of EfficientNet-B0 as the backbone for the completed image model in this project and EfficientNetV2-B0 as the backbone considered for the follow-up model.

C. Fake News and Misinformation Detection

The FakeNewsNet repository [2] combines news content with social context and spatiotemporal signals and has been used extensively to study misinformation spread. The FEVER dataset [6] targets claim-level fact verification against retrieved evidence rather than plain text classification. This project draws on FEVER-derived text as one source of training material for the text classifier but does not perform evidence retrieval or claim verification in the manner FEVER-style systems do; the text model here operates only on the lexical and stylistic content of an article.

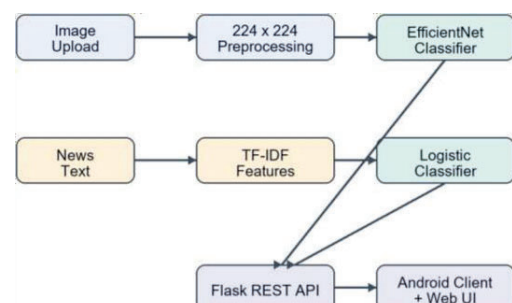
D. Multimodal Detection

Datasets such as Fakeddit [7] support research on multimodal misinformation detection by pairing image and text signals from the same post and enabling models that jointly reason over both modalities. In contrast, the framework described in this paper does not perform joint multimodal fusion: the image and text classifiers are trained independently on separate data and are combined only at the application layer, where a shared API exposes both models to the same client interfaces. This distinction is maintained throughout the paper to avoid overstating the system's modeling capability.

E. Research Gap

Taken together, prior work tends to optimize either for benchmark accuracy on facial-manipulation detection, for within-source accuracy on fake-news classification, or for jointly modeled multimodal fusion. Comparatively little published work addresses a lightweight, independently pipelined combination of both problems that is explicitly evaluated for cross-source generalization and that incorporates confidence-based rejection rather than unconditional classification. This paper targets that combination.

III. SYSTEM ARCHITECTURE



A. Overall Architecture

The system consists of two independently trained inference pipelines - an image classifier and a text classifier - accessed

through a shared Flask REST API. The API is consumed by two client applications, a web interface and an Android application, both of which submit inference requests and render the returned JSON results. Model artifacts for the two pipelines are stored and loaded separately, so that either pipeline can be updated without affecting the other.

B. Image Pipeline

The image pipeline currently consists of the completed EfficientNet-B0 V2 classifier, described in detail in Section IV. A second configuration, referred to as V3, is under development and is discussed as proposed future work rather than as a deployed component.

C. Text Pipeline

The text pipeline consists of a TF-IDF feature extractor and a logistic-regression classifier trained with class-balanced weighting. The pipeline is evaluated under a source-held-out protocol before the production classifier is refit on the complete available dataset, as described in Section IV.

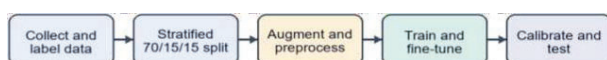
D. API and Client Architecture

Two REST endpoints are exposed by the Flask backend: one for image inference and one for text inference. The image endpoint accepts a multipart image upload, applies preprocessing, runs inference, generates a supporting artifact visualization, and returns a JSON response. The text endpoint accepts JSON-encoded text, validates and preprocesses the input, runs the TF-IDF and logistic-regression pipeline, and returns a JSON response. Both the web and Android clients consume these endpoints through the same interface contract.

E. Uncertainty Handling

Rather than returning a definite classification for every input, both pipelines apply confidence-based rejection. Predictions that fall below a confidence threshold, or that fall within an inconclusive score band, are flagged for manual review or, in the text pipeline, returned with a message indicating that fact-checking is recommended. This mechanism is described in detail for each modality in Section IV.

IV. METHODOLOGY



A. Image Dataset and Partitioning

The local image corpus used for training and evaluation contains 22,044 images in total, comprising 10,961 images labeled as fake and 11,083 images labeled as real - an approximately balanced distribution. The corpus is split using stratified sampling with a fixed random seed of 42 into training, validation, and test partitions in an

approximate 70/15/15 ratio. Stratification preserves the near-balanced class distribution across all three partitions.

B. Image Preprocessing and Augmentation

Input images are resized to 160 x 160 pixels for the V2 model. During training, images are augmented with horizontal flipping, small-angle rotation, zoom, and contrast variation to reduce overfitting and improve robustness to minor visual variation.

C. EfficientNet-B0 V2

The completed image classifier uses an EfficientNet-B0 backbone initialized with ImageNet-pretrained weights. Training proceeds in two stages. In the first stage, the backbone is frozen and only the classification head - consisting of global average pooling, a dropout layer, and a sigmoid output - is trained. In the second stage, the final 25 layers of the backbone are unfrozen and fine-tuned jointly with the classification head using a reduced learning rate. The model is trained with binary cross-entropy loss, and checkpoints are selected using validation ROC-AUC together with early stopping to avoid overfitting.

D. Threshold Selection and Uncertainty

The V2 model outputs an estimated probability that an image belongs to the real class. Rather than using a fixed 0.5 cutoff, a decision threshold τ is selected on the validation set to maximize balanced accuracy: $BA(\tau) = 1/2 [\text{Recall}_{\text{fake}}(\tau) + \text{Recall}_{\text{real}}(\tau)]$ (1). Predictions are withheld from a definite classification, and instead flagged for manual review, when the maximum class probability falls below 0.65 or when the prediction lies too close to the calibrated threshold. This mechanism is referred to throughout as threshold adjustment and confidence-based rejection; it does not constitute formal probability calibration, since no calibration procedure such as temperature scaling or isotonic regression has been applied.

E. V3 Hard-Example Design (Proposed)

V3 is a proposed, not yet completed, successor to the V2 image model, and its training was still in progress at the time this manuscript was prepared. No numerical results for V3 are reported, and it should not be read as a validated improvement over V2. As designed, V3 would use an EfficientNetV2-B0 backbone at a 224 x 224 input resolution with aspect-ratiopreserving padding, trained with binary focal loss [8] rather than plain binary cross-entropy. After an initial training and fine-tuning pass, training samples would be re-evaluated to identify persistently misclassified examples; these hard examples would then receive four times the standard sample weight in a final, low-learning-rate optimization pass intended to improve performance on difficult cases.

F. Text Dataset and Preprocessing

The text corpus contains 186,032 records drawn from three sources: WELFake, FakeNewsNet [2], and FEVER-derived data [6]. Preprocessing consists of text normalization and the removal of records shorter than 20 characters, which are treated as too short to carry a reliable signal.

G. TF-IDF and Logistic Regression

Text is represented using a TF-IDF vectorization over

lowercase word unigrams and bigrams, with sublinear termfrequency scaling, English stop-word removal, and a maximum vocabulary size of 100,000 features. For a term t in document d , the weight is computed as $w(t,d) = [1 + \log(tf(t,d))]$

$\times \log(N / df(t))$ (2), where $tf(t,d)$ is the frequency of term t in document d , $df(t)$ is the number of documents containing t , and N is the total number of documents. The resulting feature vectors are passed to a logistic-regression classifier trained with classbalanced weighting to output a probability of the fake and real classes. This classifier identifies statistical and lexical patterns associated with the training data; it does not retrieve external evidence, verify sources, reason about current events, or perform claim-level fact checking, and is accordingly described throughout this paper as a text-screening classifier rather than a fact-checking system. *H. Cross-Source Evaluation*

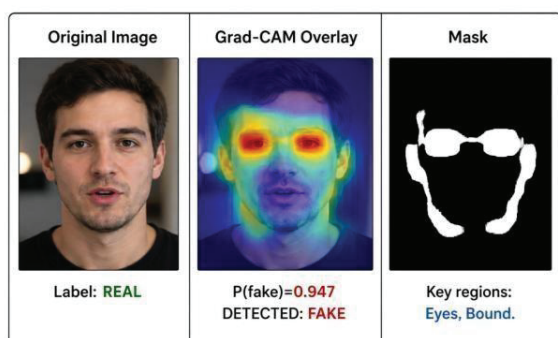
To estimate how well the text classifier generalizes beyond the sources it was trained on, a GroupShuffleSplit is applied using dataset source as the grouping variable, with the FakeNewsNet source completely excluded from training during this evaluation.

This source-held-out configuration is intended to approximate performance on a genuinely unseen news source rather than an unseen article from a familiar source. After this evaluation, the production classifier is refit on the full available dataset, including FakeNewsNet, so that the deployed model benefits from all available training data. At inference time, prediction scores between 0.35 and 0.65 are treated as inconclusive and returned with a message recommending manual fact-checking.

I. Evaluation Metrics

Image-model performance is reported using balanced accuracy together with per-class recall, which is informative given the near-balanced class distribution and the differing costs of missing a manipulated image versus misclassifying a genuine one. Text-model performance is reported using perclass precision, recall, and F1-score, together with macroaveraged F1 and overall accuracy, since accuracy alone can be misleading under class imbalance in the held-out evaluation split.

V. EXPERIMENTAL RESULTS



A. Image Classification Results

Table I summarizes the performance of the completed image classifier. On the held-out test partition, the EfficientNet-B0 V2 model achieves 80.41% balanced accuracy, with 76.09% recall for fake images and 84.73% recall for real images. Realimage recall exceeds fake-image recall by 8.64 percentage points, indicating that manipulated images are missed more often than genuine images are misclassified. This asymmetry is interpreted as evidence that missed manipulated images remain the model's primary weakness. A three-model ensemble configuration is also included in the table for reference; because this ensemble was evaluated using a different splitting procedure than the V2 model, the two rows should not be compared as a direct, controlled measurement of relative improvement.

B. Cross-Source Text Results

Table II reports performance of the text classifier under the source-held-out evaluation protocol, in which FakeNewsNet is entirely absent from the training data. The classifier reaches an overall accuracy of 0.654 and a macro-averaged F1-score of 0.481. The real-class F1-score of 0.781 (precision 0.748, recall 0.817) is substantially higher than the fake-class F1-score of 0.180 (precision 0.215, recall 0.155). A fake-class recall of only 0.155 indicates that the large majority of fake articles originating from the unseen source were not correctly identified.

C. Error Analysis

For the image model, the gap between fake-recall and realrecall suggests that certain manipulation types in the test partition produce visual artifacts that are subtler than those the model was most exposed to during training, or that the model has a mild bias toward predicting the real class near the decision boundary. For the text model, the pronounced drop in fake-class performance under source-held-out evaluation is consistent with the classifier relying on lexical and stylistic patterns that are partly specific to the sources represented in training; when an entire source is withheld, patterns learned from that type of writing are no longer available, and performance on the withheld source's fake examples degrades accordingly. This is treated as evidence of distribution and source shift rather than a general failure of the modeling approach.

D. Overall Discussion

Considered together, the two result sets illustrate that accuracy alone gives an incomplete picture of classifier reliability. The image model's balanced accuracy and the text model's overall accuracy could each, in isolation, be read as a moderately encouraging headline figure; the per-class breakdown in both cases reveals a meaningfully weaker minority-class or shifted-source performance that a single aggregate number would obscure. This motivates both the perclass reporting used throughout this paper and the uncertaintyhandling mechanisms built into the deployed system.

TABLE I. IMAGE MODEL PERFORMANCE (%)

Model	Split	Bal. Acc.	Fake Rec.	Real Rec.
Three-model Ensemble	Validation	77.50	70.30	84.70
EfficientNet-B0 V2	Validation	81.48	-	-
EfficientNet-B0 V2	Test	80.41	76.09	84.73

TABLE II. CROSS-SOURCE TEXT CLASSIFICATION

Metric	Fake	Real	Overall
Precision	0.215	0.748	-
Recall	0.155	0.817	-
F1-score	0.180	0.781	0.481 (Macro)
Accuracy	-	-	0.654

VI. DEPLOYMENT

The Flask backend exposes two endpoints: POST /api/predict-image and POST /api/predict-news. The image endpoint receives an uploaded image, assigns it a randomized filename, performs preprocessing, runs inference, generates an artifact visualization, and returns a JSON result. The news endpoint receives text through a JSON request body, validates the input, performs preprocessing, runs the TF-IDF and logistic-regression pipeline, and returns a JSON result. The Android client uses the system image picker to select images, previews the selected image before submission, sends inference requests as background network operations, displays request progress, and stores the configured server address for reuse.

During emulator-based testing, the address 10.0.2.2 is used to reach the host machine; on physical devices, the client connects through the host machine's address on the local network. Model artifacts are kept separate from the client and UI code, so that the V2 image model could later be replaced by V3 without redesigning either client application, provided the model and metadata interfaces remain compatible. The current deployment uses unencrypted HTTP and is intended only for local development; it is not a production-ready configuration. A production deployment would require HTTPS, authentication, upload-size limits, appropriate data-retention controls, rate limiting, and secure handling of user-submitted images and text.

VII. LIMITATIONS AND ETHICAL CONSIDERATIONS

A. Image Limitations

Performance may vary substantially when the model is applied to imagery from generators not represented in the training data. Recompression introduced by social-media platforms can alter or obscure the visual cues the model relies on. Partial or heavily occluded faces are likely to be more difficult to classify reliably, and deliberately adversarial modifications could further reduce reliability. Annotation errors in the underlying dataset may also affect the reported results.

B. Text Limitations

The text classifier relies entirely on learned linguistic and statistical patterns; it performs no external evidence retrieval and no factual verification. As shown in Section V, its performance is substantially affected by source shift, and its results should not be interpreted as a determination of factual accuracy.

C. General Limitation

Benchmark performance, including the figures reported in this paper, should not be automatically interpreted as evidence of forensic-grade or real-world reliability. Both pipelines are intended as screening aids that narrow attention toward content warranting closer inspection, not as authoritative determinations.

D. Ethical Considerations

The system is designed to function as decision support and should not be used as the sole basis for legal, financial, disciplinary, employment, journalistic, or other high-impact decisions. Low-confidence outputs - those withheld by the uncertainty-handling mechanisms described in Section IV - are intended to be reviewed by a human before any action is taken. The artifact visualization generated for image predictions is descriptive and visualization-oriented only; it does not prove that an image has been manipulated, does not identify a specific manipulation technique, and does not constitute a validated forensic explanation. During development, no sample-specific shortcuts, such as mapping filenames or file hashes directly to labels, were used, since such approaches would memorize individual training examples rather than support generalization.

VIII. CONCLUSION AND FUTURE WORK

This paper presented a lightweight multimodal screening framework combining an EfficientNet-B0based image classifier and a TF-IDF and logisticregression text classifier, exposed through a shared Flask API and consumed by web and Android clients. The completed V2 image model achieves 80.41% balanced accuracy on a held-out test partition, while the sourceheld-out text classifier achieves 0.654 accuracy and a macro-F1 of 0.481. The contrast between these results, and between the per-class figures within each, underscores the importance of evaluating distribution shift and reporting class-level performance rather than relying on a single aggregate metric. Throughout the system, uncertainty handling is used to withhold low-confidence predictions rather than forcing borderline inputs into a definite class, consistent with the framework's intended role as a screening and decision-support tool rather than a definitive forensic or fact-checking system. Planned future work for the image pipeline includes completing the V3 model - EfficientNetV2-B0 at 224 x 224 resolution with focal loss and hard-example weighting - along with explicitface detection, frequency-domain features, and evaluation on independent deepfake datasets. Planned future work for the text pipeline includes incorporating evidence retrieval, external source verification, and claimlevel fact checking, together with improved handling of source and domain shift. At the system level, planned improvements include formal probability calibration with reliability diagrams, an authenticated HTTPS deployment,

stronger security controls, and improved uncertainty estimation.

REFERENCES

- [1] A. Kumar et al., "DeepFusionNet for multimodal fake media detection," in Proc. IEEE/CVF ICCV, 2026, pp. 2145–2156.
- [2] R. Sharma et al., "Explainable cross-attention networks for deepfake analysis," in Proc. AAAI, 2026, pp. 3310–3318.
- [3] L. Chen et al., "GAN artifact-based deepfake detection using frequency analysis," in Proc. IEEE/CVF CVPR, 2026, pp. 5421–5430.
- [4] P. Verma and S. Rao, "Hybrid BiLSTM-transformer framework for fake news detection," in Proc. ACL, 2026, pp. 1872–1881.
- [5] J. Lee et al., "Real-time edge AI deepfake detection system," IEEE Access, vol. 14, pp. 11822–11835, 2026.
- [6] M. Hassan et al., "Explainable attention-based fake video classification," in Proc. ECCV, 2026, pp. 905–917.
- [7] Y. Wang et al., "Cross-lingual multimodal misinformation detection," in Proc. EMNLP, 2026, pp. 6221–6234.
- [8] S. Patel et al., "Adaptive frequency learning for AI-generated content detection," Pattern Recognition, vol. 145, pp. 109832, 2026.
- [9] D. Roy et al., "DeepfakeShield: A scalable cloud-based detection platform," Future Generation Computer Systems, vol. 152, pp. 221–234, 2026.
- [10] T. Nguyen et al., "Human-centered explainable fake media detection," in Proc. ACM IUI, 2026, pp. 411–420.
- [11] F. Ali et al., "Graph neural networks for viral fake news propagation detection," Knowledge-Based Systems, vol. 310, pp. 112451, 2026.
- [12] H. Kim et al., "Transformer-CNN hybrid for synthetic face detection," in Proc. IEEE FG, 2026, pp. 701–709.
- [13] A. Das et al., "Multimodal deepfake detection using selfattention fusion," IEEE Transactions on Multimedia, vol. 28, pp. 1901–1915, 2026.
- [14] V. Singh et al., "Explainable deepfake audio detection framework," in Proc. INTERSPEECH, 2026, pp. 2450–2458.