International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

A Leakage-Aware Framework for Evaluating Fake News and Image Manipulation Detectors

Sri Kushal V
School of Computer Science and Engineering
VIT-AP University
Amaravati, India

Durbha Sriniketh
School of Computer Science and Engineering
VIT-AP University
Amaravati, India

Kunisetty Vamsi Krishna
School of Computer Science and Engineering
VIT-AP University
Amaravati, India

Gunda Rahul
School of Computer Science and Engineering
VIT-AP University
Amaravati, India

Khasim Syed*
School of Computer Science and Engineering
VIT-AP University
Amaravati, India

Abstract*—Published misinformation detectors often report accuracies close to 100%, but these figures usually come from a random split of a single benchmark. We revisit this practice for two forms of misinformation, news text and images, using evaluation protocols that limit data leakage. For text, we fine-tune RoBERTa-base and DeBERTa-v3-base on three news sources derived from ISOT, McIntire and WELFake. Before splitting, we remove exact and MinHash near-duplicates and separate the overlapping datasets, and in doing so we found that WELFake already contains every ISOT and McIntire article. Over 20 training runs, the two transformers reach 0.99–1.00 macro-F1 in-domain and cannot be told apart statistically. Their performance drops to 0.78–0.95 on a news source withheld from training, to about 0.51 on LIAR claims, and below chance on COVID-19 social-media posts (ROC-AUC 0.455–0.469). On these out-of-domain sets the models are strongly overconfident (expected calibration error 0.38–0.56), and their errors change markedly with the random seed. An error analysis and a small controlled style probe indicate that the models respond to writing style rather than to factual truth. For images, we merge CASIA 2.0 tampered photographs with a dataset of AI-generated images into a 45,614-image benchmark and report preliminary single-run baselines. In our setup, every model finds AI-generated images much easier to detect than tampered photographs (ROC-AUC 0.977 versus 0.843 for ResNet50 and 0.975 versus 0.778 for ViT-B/16), although part of this gap may stem from differences between the two datasets and not only from the difficulty of the tasks. On the pooled test set, the ImageNet-pretrained ResNet50 and ViT-B/16 obtain ROC-AUC of 0.954 and 0.946, respectively. In both modalities, pooled or in-domain scores give an inflated view of reliability, and predictions can depend on cues tied to a particular data source. We therefore argue that source-disjoint evaluation, per-source reporting, multiple seeds and explicit shortcut checks should become routine.

***Index Terms*—fake news detection, image forensics, AI-generated images, transformers, shortcut learning, data leakage, domain generalization**

I. Introduction

False or misleading content reaches people both as fabricated news stories and as edited or synthetic images. For news text, the usual approach is to fine-tune a pretrained transformer such as RoBERTa [6] or DeBERTa-v3 [7]. For images, ImageNet-pretrained backbones such as ResNet50 [22] and the Vision Transformer [23] are commonly adapted to forensic tasks. Accuracies close to 100% are frequently reported for both kinds of system.

The difficulty is that most of these numbers come from a random split of one dataset. If the two classes of a dataset were collected from different outlets or image collections, a classifier can separate them simply by recognising where a sample came from, for instance a news agency's dateline or the compression format of a particular image collection, without learning anything about whether the content is genuine. This is a case of shortcut learning [8]: the resulting models look excellent on the benchmark and break down once they are deployed.

In this paper, we evaluate deep detectors for text and image misinformation under protocols designed to bring such failures to light. Our main contributions are as follows:

- We build a leakage-safe text pipeline that combines exact and MinHash near-duplicate removal with a per-source split. Decomposing the benchmarks shows that WELFake [3] contains every ISOT [1] and McIntire article.
- We compare RoBERTa and DeBERTa-v3 under controlled conditions over 20 runs, covering all-source, single-source, leave-one-source-out, giveaway-stripped and multi-seed training, and we support the comparison with bootstrap intervals, paired significance tests and a calibration analysis.

- We show that out-of-domain text performance is low, overconfident and sensitive to the random seed, and we use a style probe to show that predicted labels follow writing style rather than truth.
- We assemble a benchmark that combines traditional image tampering with AI-generated images and evaluate four image classifiers on each source separately, which reveals a large performance gap between the two threats.

II. Related Work

Fake news benchmarks. ISOT [1], [2] pairs real articles from Reuters with fake articles from websites flagged by PolitiFact. WELFake [3] merges four earlier datasets into roughly 72,000 articles. LIAR [4] consists of short claims checked by PolitiFact, and the Constraint COVID-19 dataset [5] consists of social-media posts. Several shortcuts in ISOT have already been documented. Texts that begin with "Reuters" are always real [11], and the subject field by itself gives F1 = 1.000, while models trained on ISOT perform close to chance on LIAR [13]. Cross-dataset studies consistently find a sharp fall from in-dataset to cross-dataset performance [9], [10], [12], and rewriting articles in the style of a trusted outlet or a tabloid lowers detector F1 by as much as about 38% [14].

Image forensics. CASIA 2.0 [15] is a widely used tampering benchmark. However, its tampered images share a single JPEG quantisation table [16], and they have been reported to differ from the authentic images in file format [17]. Detectors of generated images can transfer across GAN architectures [18], yet they often fail on images from diffusion models [19], [20], and their accuracy decreases whenever the generators used for training and testing differ [21].

Evaluation validity. Data leakage is common in science that relies on machine learning [25], and McNemar's test is the recommended way to compare two classifiers evaluated on the same test set [26]. To our knowledge, few fake-news studies account for duplicates, overlap between benchmarks and training randomness at the same time, and few image detectors are evaluated separately for each type of threat. This work addresses both gaps.

III. Data

A. News Text

Training uses three sources: ISOT, McIntire (*fake_or_real_news*) and *WELFake-rest*. WELFake-rest consists of the 18,893 WELFake rows left after we remove the rows that exactly match ISOT (44,054 rows) or McIntire (6,335 rows). We checked the label direction of every dataset empirically and use 0 = real and 1 = fake throughout; the WELFake file uses the opposite direction to the one given in its documentation. Text normalisation is deliberately conservative (HTML unescaping, NFKC and whitespace), and articles whose body has fewer than 20 words are discarded. Duplicates are removed both exactly and with MinHash LSH (5-word shingles, 128 permutations, Jaccard ≥ 0.80), and clusters whose members carry conflicting labels are dropped. Each source is then split 80/10/10 with stratification by label,

TABLE I
Text corpus after cleaning

Source	Total	Train	Test	Fake
ISOT	37,935	30,348	3,794	45.0%
McIntire	5,877	4,701	588	49.7%
WELFake-rest	17,121	13,696	1,713	39.1%
LIAR (external)	10,174	–	10,174	55.5%
COVID-19 (external)	10,546	–	10,546	47.0%

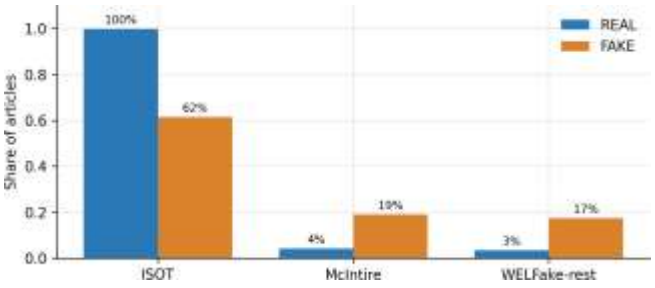


Fig. 1. Proportion of articles that contain a source giveaway, broken down by source and class.

and an automated audit verifies that no exact or near-duplicate text is shared between splits. LIAR and COVID-19 are used only as external test sets. For LIAR, the pants-fire, false and barely-true labels are mapped to fake, mostly-true and true are mapped to real, and half-true is dropped. Table I summarises the resulting corpus.

We also wrote a regular-expression diagnostic that marks source "giveaways", such as agency datelines, image credits and outlet names, and produces a stripped copy of every text. In ISOT, giveaways occur in 100% of the real articles but in only 62% of the fake ones. In the other sources, they appear in fewer than 20% of the articles (Fig. 1).

B. Images

The image benchmark joins CASIA 2.0 (7,491 authentic and 5,123 tampered images) with a public AIGC detection dataset (15,000 real and 18,000 AI-generated images; according to the dataset source, the real images come from COCO [27]). The result is 45,614 images, labelled 0 = authentic and 1 = fake. CASIA is split 70/15/15, with images grouped by a hash of their file content. For AIGC we keep the published split, after checking that no original image appears in more than one split. Because the CASIA grouping relies on byte-identical file hashes, tampered images are not kept together with their donor images, and the CASIA scores may therefore be optimistic. The test set has 6,839 images, of which 1,889 come from CASIA and 4,950 from AIGC.

IV. Methods

Fig. 2 gives an overview of the two modules and of the evaluation framework they share.

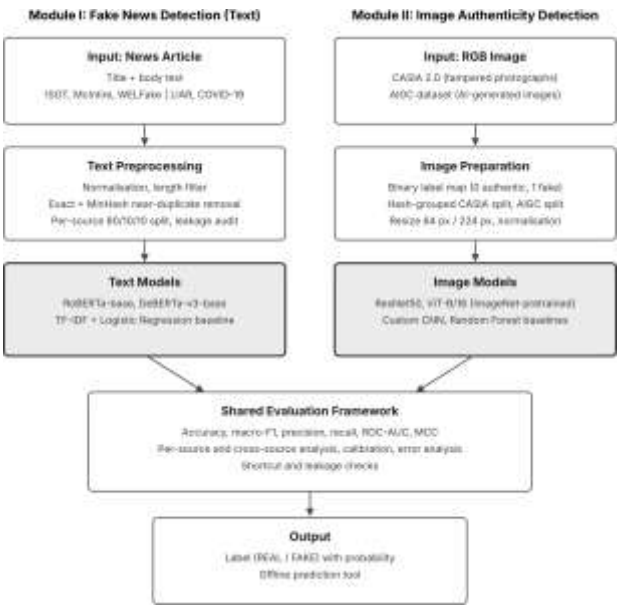


Fig. 2. Overview of the two detection modules and the shared evaluation framework.

A. Text Models and Protocol

RoBERTa-base [6] and DeBERTa-v3-base [7] take the title and the body as a sentence pair, truncated to 256 tokens. Every run uses the same settings: AdamW with a learning rate of 2×10^{-5} and weight decay of 0.01, linear warm-up over the first 6% of steps, a batch size of 32, mixed-precision training and gradient clipping at 1.0. Training lasts at most two epochs, with validation every half epoch, early stopping on validation macro-F1 and a fixed decision threshold of 0.5. Each model is trained on all three sources (seeds 42, 13 and 7), on ISOT alone (seeds 42, 13 and 7) and in three leave-one-source-out variants, and each is trained once on giveaway-stripped text, for a total of 20 runs. As a reference, we also fit a TF-IDF (1–2-grams) and logistic-regression baseline whose hyperparameters are tuned on validation data only.

B. Image Models

We compare two baselines trained from scratch, a Random Forest (200 trees on flattened 64×64 pixels) [24] and a custom three-block CNN (64×64 input), with ResNet50 [22] and ViT-B/16 [23] pretrained on ImageNet [28] and fine-tuned at 224×224. The pretrained models are trained for one epoch with a frozen backbone and then with discriminative learning rates, cosine decay and early stopping on validation AUC. Note that the two groups differ in more than model family: the Random Forest and the custom CNN see 64×64 inputs and have no pretraining, whereas ResNet50 and ViT-B/16 see 224×224 inputs and start from ImageNet weights. The gap between them should therefore not be read as a pure deep-versus-classical comparison. All four models use a fixed threshold of 0.5 and seed 42, so the results compare complete pipelines rather than architectures in isolation.

TABLE II
TEXT MODELS TRAINED ON ALL SOURCES (SEED 42): MACRO-F1 [95% CI] AND ROC-AUC

Model	Test set	Macro-F1 [95% CI]	AUC
RoBERTa	ISOT	1.000 [0.999, 1.000]	1.000
	McIntire	0.993 [0.986, 0.998]	1.000
	WELFake-rest	0.994 [0.990, 0.998]	1.000
	LIAR	0.512 [0.502, 0.521]	0.556
	COVID-19	0.317 [0.313, 0.322]	0.469
DeBERTa	ISOT	1.000 [1.000, 1.000]	1.000
	McIntire	0.988 [0.979, 0.997]	0.999
	WELFake-rest	0.996 [0.993, 0.999]	1.000
	LIAR	0.513 [0.503, 0.522]	0.523
	COVID-19	0.319 [0.312, 0.326]	0.455
TF-IDF+LR	ISOT	0.986 [0.982, 0.989]	0.999
	McIntire	0.918 [0.896, 0.939]	0.973
	WELFake-rest	0.952 [0.941, 0.962]	0.992
	LIAR	0.436 [0.428, 0.445]	0.538
	COVID-19	0.322 [0.317, 0.327]	0.383

C. Evaluation

We treat fake as the positive class and report macro-F1, accuracy, ROC-AUC, PR-AUC, the Matthews correlation coefficient and the expected calibration error (ECE, 15 bins). For the text models, we additionally compute 95% bootstrap intervals (1,000 resamples), McNemar's exact test and paired-bootstrap differences in macro-F1, with Holm correction for multiple comparisons. For the image models, metrics are reported both overall and for each source.

V. RESULTS

A. Text: In-Domain and External Performance

Table II lists the results of the models trained on all sources. In-domain, both transformers reach 0.99–1.00 macro-F1 with ECE $\leq$ 0.009, and both are significantly better than TF-IDF after Holm correction, even though TF-IDF already achieves 0.92–0.99. We find no difference in macro-F1 between RoBERTa and DeBERTa on any test set, since every paired-bootstrap interval contains zero. On LIAR, all models perform close to chance. On COVID-19, ROC-AUC is below 0.5 and its confidence intervals exclude 0.5. Although only 47% of the COVID-19 posts are fake, the transformers label 88–98% of them as fake.

B. Text: Unseen Sources and Single-Source Training

The leave-one-source-out results are given in Table III. Every model performs worse on the source it did not see during training, losing 0.04–0.21 macro-F1 in the case of the transformers and 0.20–0.26 in the case of TF-IDF. Models trained on ISOT alone all score 1.00 on ISOT, but only 0.31–0.84 on the remaining news sources (Fig. 3). ISOT's test set shows how strong the giveaway cue is: 99.4% of the articles

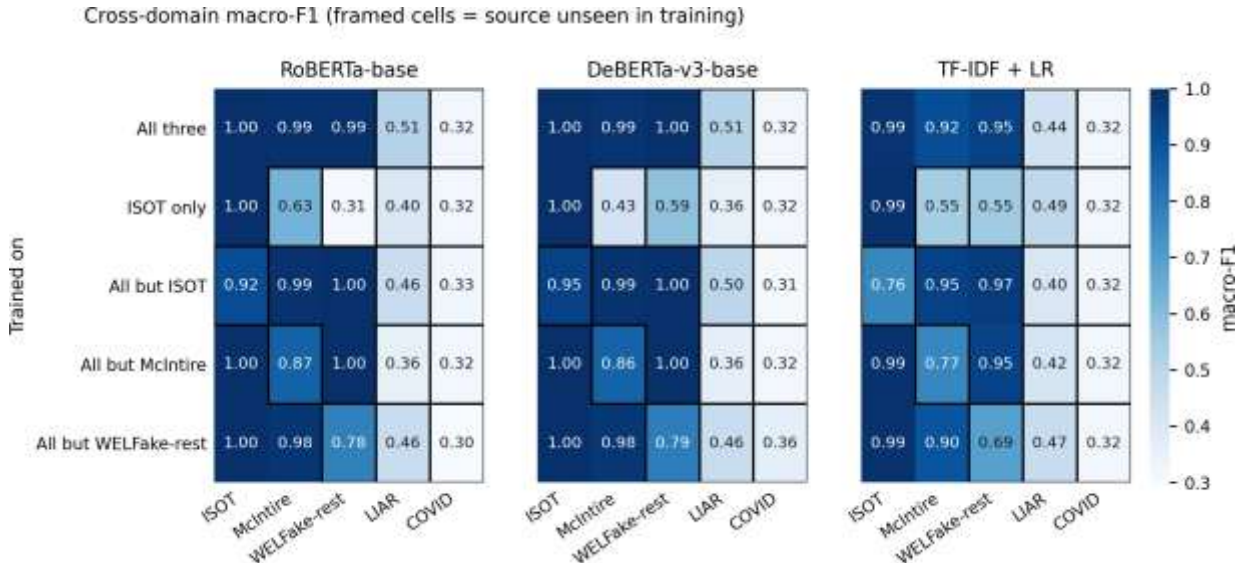


Fig. 3. Cross-domain macro-F1 (seed 42). Each row is a training configuration and each column a test set; framed cells mark sources that were not seen during training.

TABLE III
LEAVE-ONE-SOURCE-OUT MACRO-F1 (SEED 42)

Held out	Model	Seen	Held out	Drop
ISOT	RoBERTa	0.994	0.921	0.073
	DeBERTa	0.994	0.951	0.043
	TF-IDF+LR	0.960	0.764	0.196
McIntire	RoBERTa	0.999	0.865	0.134
	DeBERTa	0.998	0.862	0.137
	TF-IDF+LR	0.968	0.766	0.202
WELFake-rest	RoBERTa	0.991	0.780	0.210
	DeBERTa	0.990	0.789	0.201
	TF-IDF+LR	0.946	0.689	0.257

TABLE IV
SEED VARIATION: MACRO-F1 MEAN ± SD OVER SEEDS 42, 13, 7

Training	Model	LIAR	McIntire	WELFake-rest
All sources	RoBERTa	0.451±0.079	0.990±0.003	0.996±0.001
All sources	DeBERTa	0.510±0.009	0.988±0.002	0.996±0.001
ISOT only	RoBERTa	0.388±0.025	0.569±0.123	0.461±0.206
ISOT only	DeBERTa	0.413±0.049	0.639±0.186	0.733±0.129

TABLE V
STYLE PROBE: P (FAKE) FOR ROBERTA / DEBERTA

Claim	Agency style	Tabloid style
True (water on the Moon)	0.0006 / 0.0030	1.0000 / 0.9999
False (second moon)	0.0002 / 0.0099	1.0000 / 1.0000

without a giveaway are fake (645 articles), compared with 33.8% of the articles that contain one.

C. Text: Giveaway Ablation and Seed Variation

Stripping the listed giveaways has almost no effect on in-domain scores, and stripped ISOT still reaches 1.000. This suggests that the models draw on stylistic cues that go beyond the explicit patterns we removed. Across seeds, in-domain scores vary by no more than 0.003, whereas out-of-domain scores vary considerably (Table IV, Fig. 5). Two conclusions that single runs had suggested turned out to be wrong once more seeds were added: that the two ISOT-only models failed in opposite directions, and that the ISOT-only DeBERTa model depended almost entirely on giveaways.

D. Text: Calibration, Errors and Style Probe

In-domain, the transformers are well calibrated (ECE 0.002). Out of domain they are overconfident, with ECE of 0.38–0.42 on LIAR and 0.54–0.56 on COVID-19 (Fig. 4). On COVID-19, RoBERTa labels 5,570 of the 5,592 real posts as fake. An inspection of the errors shows a consistent pattern. The real posts that receive a fake probability of 1.00 are messages from official accounts, such as retweets from India's Ministry of Health, while the fake posts that receive a probability of 0.00 are fact-check summaries written in formal, agency-like prose. A controlled probe with four hand-written inputs (Table V) reproduces this behaviour, with the predicted label following the style of the text and not its truth.

E. Images: Overall and Per-Source Performance

Table VI reports the image models on the 6,839-image test set. ResNet50 and ViT-B/16 achieve similar pooled ROC-AUC (0.954 and 0.946). Because this small difference comes from a single seed and was not tested for significance, we do not

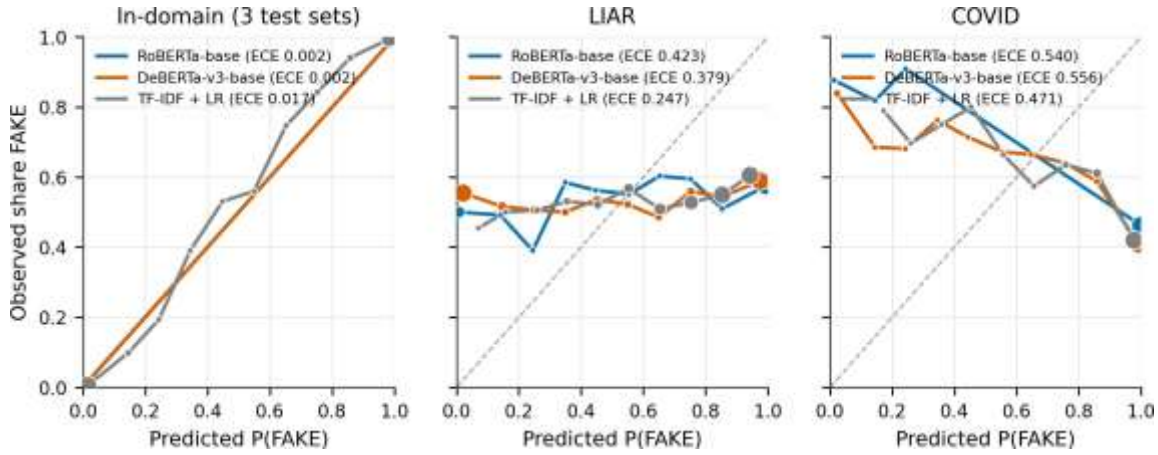


Fig. 4. Reliability diagrams of the all-source text models on the pooled in-domain test sets, LIAR and COVID-19.

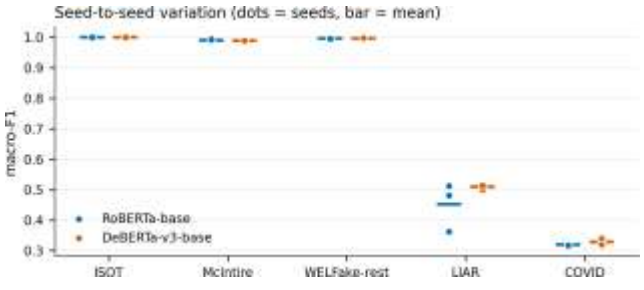


Fig. 5. Macro-F1 of the all-source text models: individual seeds (dots) and their mean (bar).

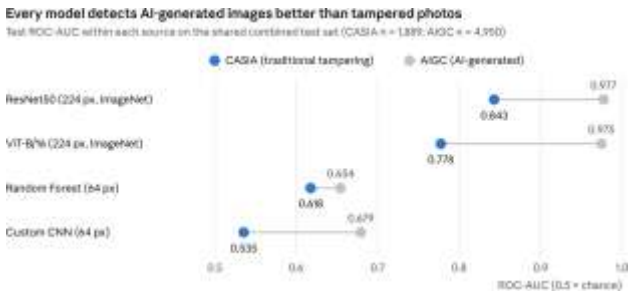


Fig. 6. Test ROC-AUC of each image model computed within each source (CASIA n = 1,889; AIGC n = 4,950).

rank the two models. The two 64-pixel baselines stay close to 0.65. When the results are broken down by source (Fig. 6), every model in our setup detects AI-generated images more reliably than tampered photographs. ResNet50 reaches 0.977 on AIGC but 0.843 on CASIA, and ViT-B/16 reaches 0.975 and 0.778, respectively. The custom CNN recovers only 0.224 of the tampered images in CASIA. This gap may be partly due to differences between the datasets, such as file format, compression and resolution, rather than to task difficulty alone; we did not test this possibility.

TABLE VI
IMAGE MODELS ON THE COMBINED TEST SET (N = 6,839; POOLED RESULTS, SINGLE RUN)

Model	Accuracy	Recall	F1	ROC-AUC
ResNet50	0.888	0.907	0.891	**0.954**
ViT-B/16	0.870	0.909	0.876	0.946
Custom CNN	0.613	0.485	0.560	0.652
Random Forest	0.615	0.628	0.624	0.645

VI. DISCUSSION

The text and image experiments point to the same four lessons. First, pooled or in-domain scores overstate how reliable a detector is. The text models reach 0.99–1.00 in-domain and still fail on new genres, and ResNet50 and ViT-B/16 reach 0.95 pooled ROC-AUC but only 0.78–0.84 on tampered photographs. Second, cues specific to a source can drive the predictions. In ISOT, the writing style of the outlet almost reveals the label, and for images, dataset-level differences in format, compression and content may provide similar shortcuts, although we did not measure this. Third, the choice of training data matters more than the choice between two strong encoders. Adding a second training source raised the held-out text scores from 0.31–0.84 to 0.78–0.95, while RoBERTa and DeBERTa remained statistically indistinguishable. Fourth, a single run is not enough to compare models out of domain, because changing only the seed shifted text macro-F1 by up to 0.4.

Based on these findings, we recommend evaluating misinformation detectors with (i) dataset decomposition and near-duplicate removal before any split is made, (ii) held-out sources together with per-source reporting, (iii) at least three seeds for any out-of-domain claim, (iv) cross-genre tests accompanied by calibration analysis, and (v) shortcut baselines such as TF-IDF or metadata-only models.

VII. Limitations

The labels in ISOT and WELFake reflect the reputation of the outlet, not a fact-check of each individual article. LIAR and COVID-19 differ from the training data in both genre and source, so the two effects cannot be separated. Giveaway removal relies on fixed patterns and will miss cues that the patterns do not cover. Finally, only the all-source and ISOT-only text configurations were repeated with three seeds.

The image results should be regarded as preliminary single-run baselines, for several reasons. Each model was trained once, and no confidence intervals were computed. The CASIA split groups only byte-identical files and does not yet keep tampered images together with their donor images, so the CASIA scores may be optimistic. We did not measure the possible CASIA format shortcut [16], [17], nor did we test differences in format, compression and resolution between the real and AI-generated images in the AIGC data; the gap between AIGC and CASIA may therefore partly reflect dataset differences. Resolution and pretraining are confounded across models, since the baselines use 64×64 inputs without pretraining and the other two models use 224×224 inputs with ImageNet weights. That all four models were evaluated on the same test images was confirmed by counts only. All generators in the test set also appear in training, and the provenance of the AIGC images is taken from the dataset source without independent verification.

VIII. Conclusion

When evaluated with attention to leakage, deep detectors for both fake news and manipulated images turn out to be far less reliable than their headline scores suggest. Transformers that are almost perfect on familiar news sources lose accuracy on unseen sources, fail on short claims and social-media posts, and respond to writing style rather than truth. In our preliminary single-run image baselines, detectors that perform strongly on AI-generated images are clearly weaker on traditional tampering (ROC-AUC 0.977 versus 0.843 for ResNet50 and 0.975 versus 0.778 for ViT-B/16), although this gap may partly reflect differences between the datasets. In future work, we plan to apply donor-aware splitting to CASIA with repeated seeds, run leave-one-generator-out tests, explore style-agnostic training [14] and add evidence retrieval to the text detector.

References

[1] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in *Proc. ISDDC*, LNCS vol. 10618, Springer, 2017, pp. 127–138.
[2] H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spams and fake news using text classification," *Security and Privacy*, vol. 1, no. 1, 2018.
[3] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "WELFake: Word embedding over linguistic features for fake news detection," *IEEE Trans. Comput. Social Syst.*, 2021, doi: 10.1109/TCSS.2021.3068519.
[4] W. Y. Wang, "'Liar, liar pants on fire': A new benchmark dataset for fake news detection," in *Proc. ACL*, 2017.
[5] P. Patwa *et al.*, "Fighting an infodemic: COVID-19 fake news dataset," in *CONSTRAINT@AAAI*, Springer, 2021.
[6] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019.
[7] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing," in *Proc. ICLR*, 2023.
[8] R. Geirhos *et al.*, "Shortcut learning in deep neural networks," *Nature Mach. Intell.*, vol. 2, 2020.
[9] L. Bozarth and C. Budak, "Toward a better performance evaluation framework for fake news classification," in *Proc. ICWSM*, vol. 14, 2020, pp. 60–71.
[10] C. Blackledge and A. Atapour-Abarghouei, "Transforming fake news: Robust generalisable news classification using transformers," arXiv:2109.09796, 2021.
[11] J. Wu and X. Ye, "FakeSwarm: Improving fake news detection with swarming characteristics," arXiv:2305.19194, 2023.
[12] Y. Kishi, Y. Arima, and H. Iyatomi, "Fake news detection strategies under dataset bias: Using large-scale coarse-grained labels," in *Proc. EACL Student Research Workshop*, 2026, pp. 612–621.
[13] Y. Verma, "What does 99% accuracy measure? A reproducible audit of shortcut learning in a widely used fake news corpus," arXiv:2609.25006, 2026.
[14] J. Wu, J. Guo, and B. Hooi, "Fake news in sheep's clothing: Robust fake news detection against LLM-empowered style attacks," in *Proc. KDD*, 2024.
[15] J. Dong, W. Wang, and T. Tan, "CASIA image tampering detection evaluation database," in *Proc. IEEE ChinaSIP*, 2013, pp. 422–426.
[16] M.-J. Kwon, S.-H. Nam, I.-J. Yu, H.-K. Lee, and C. Kim, "Learning JPEG compression artifacts for image manipulation detection and localization," *Int. J. Comput. Vis.*, 2022.
[17] S. K. Mandala, "Format-controlled multi-scale JPEG compression response analysis for image-level forgery screening," arXiv:2607.06615, 2026.
[18] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *Proc. CVPR*, 2020, pp. 8695–8704.
[19] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proc. CVPR*, 2023.
[20] R. Corvi *et al.*, "On the detection of synthetic images generated by diffusion models," in *Proc. ICASSP*, 2023.
[21] M. Zhu *et al.*, "GenImage: A million-scale benchmark for detecting AI-generated image," in *Proc. NeurIPS Datasets and Benchmarks*, 2023.
[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.
[23] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.
[24] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[25] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, 2023.
[26] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1924, 1998.
[27] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.
[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.