Quality Over Quantity: Image Curation for Multimodal Fake News Detection

Amrit Raj
Dept. of CSE (AI/ML)
Manipal University Jaipur
Jaipur, Rajasthan, India
rajamrit52656@gmail.com

Dr. Manish Rai
Dept. of CSE (AI/ML)
Manipal University Jaipur
Jaipur, Rajasthan, India
manish.rai@jaipur.manipal.edu

***Abstract*—Most multimodal fake news detection research assumes that more training images lead to better performance. We test this assumption directly. Using a Light Cross-Modal Attention model that fuses frozen DeBERTa-v3-base and CLIP ViT-B/32 embeddings through a 260K-parameter attention module, we compare detection accuracy across three image collection strategies: generic web scraping via Bing (679 images), curated article-specific images from FakeNewsNet (60 images), and manually collected images from 15 fact-checking organizations (150 images). In a size-controlled comparison (150 vs. 150), curated images outperform generic ones by 14.7 percentage points (93.3% vs. 78.6%, 5-fold CV). The 150 curated images also beat all 679 generic images by 4.5 points. A second experiment on mixed-source data reveals that multimodal fusion degrades by only 1.4% when data sources are combined, while text-only and image-only models drop by 19.6% and 13.5% respectively. These results point to image-text relevance as a more important factor than dataset size in multimodal fake news detection.**

***Index Terms*—fake news detection, multimodal learning, cross-modal attention, image quality, CLIP, DeBERTa**

I. Introduction

Fake news is more likely to be spreading faster with pictures. A false headline and a falsified photograph or a photo out of context are more convincing than written words. Misinformation research is well documented to this effect [8], [15], and it is the reason that detection systems are now attempting to analyze text and images in tandem [1], [2], [6].

The traditional method of constructing multimodal training data is simple: download a set of labeled articles and search them with images by article titles or URLs and download whatever is returned. Such datasets constructed in this manner, such as portions of FakeNewsNet [8], include hundreds or thousands of pairs of images and text. A lot of those pictures though are stock images, website thumbnails, or are simply tenuously related to the claim in the article itself. One question to consider is: are these generic images contributing to the model learning anything about veracity or is it merely noise?

To determine, we conducted a series of experiments. And the simple answer is that image quality is far more important than image quantity and the difference is wide. In particular, this paper contributes in three ways:

1) When compared with matched sample sizes (150 vs. 150) in a controlled comparison, manually curated article-specific images result in 14.7 percentage points more accurate than generic web-scraped images when using the same model and training procedure.
2) The curated images still prevail by 4.5 percentage points even when the generic dataset is 4.5 times bigger (679 vs. 150).
3) Multimodal fusion mechanisms can perform reasonably on a mixed-source dataset with two distinct curation styles (1.4% drop), whereas single-modality models fail (up to 19.6% drop), which confirms the practical usefulness of cross-modal fusion on heterogeneous real-world data.

The rest of this paper is organized as follows. Section II summarizes the related work, which also contains a comparative analysis of the existing multimodal detection techniques. Part III outlines our structure, data and training environment. Section IV shows experimental findings in four sets of experiments. Section V addresses implications, confounds, and limitations in a candid manner. Section VI concludes.

II. Related Work

A. Text-Based Detection

Detecting fake news in text has a long history of research, beginning with manually designed features such as n-grams and linguistic clues [13]. The shift to deep learning brought models like BERT [12] and DeBERTa [9], which achieve over 90% accuracy on standard benchmarks like ISOT [13] and LIAR [14]. Large language models have also been considered more recently to do this task. Jiang et al. [17] examined the use of LLMs in detecting fake news, discovering that although they offer robust language comprehension, they continue to be unable to address statements that need external knowledge validation.

Such text-only models are effective in cases when the article in question carries with it the telltale linguistic patterns of sensationalism, unprovable claims, or style differences between professional journalism and made-up messages. However, they fail to notice situations when the text is in normal form and the lie is made by a distorted or out-of-context image.

B. Multimodal Detection

A number of architectures have been suggested to combine text and image signals. Table I provides a comparative

overview.

EANN [1] proposed adversarial training in order to develop features that are not overfitted to particular news events. The thought was that a model trained on political news would continue to function on health misinformation. SpotFake [2] followed a more basic path and joined VGG-19 and BERT features, and then input them into a classifier. SpotFake+ [4] built upon this by end-to-end fine-tuning of the text and image encoders, although at the cost of increased training. MVAE [3] attempted to learn a shared latent space between text and image modalities with a variational autoencoder by claiming that fake and real news should be in different parts of this joint space.

SAFE [5] took a different approach to the problem. It did not merge features, but estimated the closeness between text and visual data, based on the idea that fake articles tend to have text-image discrepancies. That is a sensible assumption when it comes to situations where a real-life photograph is paired with a fake narrative, but it does not apply to AI-generated images that are created to align with the text.

More recent efforts have shifted to attention-based fusion. HMCAN [7] applied hierarchical attention in the modalities in a variety of context levels. CAFE [6] represented cross-modal ambiguity as it is a feature and not a problem. Li et al. [18] investigated cross-modal contrastive learning to achieve better modalities feature alignment. Wu et al. [20] explored different fusion techniques and revealed that late fusion using learned weights is likely to achieve better results compared to naive concatenation.

Most of these fusion methods are based on the attention mechanism described by Vaswani et al. [11] and its performance with cross-modal tasks is well-established.

The only similarity in all these papers is that they all are concerned with the fusion architecture but assume that the training data is provided. They do not inquire whether the pictures in the training set are useful or not. Scraping is used to create datasets, and the unspoken rule here is that the more image-text pairs one has, the better. We question that assumption head-on.

C. Foundation Models as Feature Extractors

The internet provided 400 million image-text pairs which were trained on CLIP [10], learning to map images and text to a shared embedding space. Abdelnabi et al. [16] have investigated its usefulness in misinformation-related tasks, and discovered that CLIP embeddings encode semantic connections between manipulated images and their original contexts. DeBERTa-v3 [9] is an advancement of BERT with disentangled attention and a more powerful mask decoder, delivering high-performance results across NLP benchmarks.

When there is a lack of labeled data, it is customary to freeze these large models and only train a small task-specific head on top. It does not involve catastrophic forgetting and yet utilizes representations that it has learned on large pre-training data. Our architecture is this frozen-encoder design, which also has the practical advantage of being fast enough to train to perform large ablation experiments.

D. Image Manipulation Detection

There is a related body of work which is interested in detecting manipulated images themselves. Heller et al. [19] demonstrated that compression artifacts and noise patterns can be used to detect the presence of spliced and doctored pictures by image forensics. Although our work does not detect pixel-level manipulation, the manipulations are often present in the curated images in our dataset, and the embeddings of CLIP can implicitly detect some of these forensic signals.

III. Methodology

A. System Architecture

The system consists of three components: a text encoder, an image encoder, and a fusion classifier. Both encoders stay frozen throughout training. The fusion module is the only one that is optimized. Fig. 1 shows the complete pipeline.

Text encoder. DeBERTa-v3-base (110M parameters) is fed on the concatenated title and body text of each article. The [CLS] token embedding, which is a 768-dimensional vector, is taken. Training 110M parameters on 150 labeled samples would overfit in a few epochs, meaning that the encoder is kept fixed.

Image encoder. CLIP ViT-B/32 (86M parameters) represents each image as a 512-dimensional vector using its standard preprocessing. It remains frozen as well.

Fusion module. The images and text embeddings are projected into a common 128-dimensional space with independent linear layers. Multi-head attention (4 heads) is implemented with text as query and image as key/value. This is a deliberate design decision: the text poses the question "what in this image is relevant to confirming this statement?" The attention output is followed by a residual connection and layer normalization. The text that was attended to and the projected image are concatenated (256 dimensions overall) and sent through a two-layer classifier with ReLU and 20% dropout.

Table II breaks down the trainable parameter count. The entire fusion module comprises approximately 260K parameters, which is 0.13% of DeBERTa alone. This allows training to be very fast and minimizes the chance of overfitting on small datasets.

B. Datasets

Our datasets of images are four sets of images alongside the ISOT text corpus. Table III gives an overview.

ISOT text baseline. 44,158 articles (21,417 real from Reuters, 22,741 fake from unreliable sources) with a 70/15/15 stratified split. Applied to the text-only baseline, as there are no paired images.

Bing web-scraped (679 images). We used Bing Image Search API with article titles and downloaded the first result of each of the articles. The photos we retrieved back are predominantly thumbnails, stock pictures, and news channel header graphics. Few of them are directly connected with the

TABLE I
COMPARATIVE ANALYSIS OF MULTIMODAL FAKE NEWS DETECTION METHODS

Method	Year	Text Encoder	Image Encoder	Fusion Strategy	Data Quality Focus
EANN [1]	2018	Text-CNN	VGG-19	Adversarial (event-independent)	No
SpotFake [2]	2019	BERT	VGG-19	Concatenation	No
MVAE [3]	2019	Bi-LSTM	VGG-19	Variational autoencoder	No
SpotFake+ [4]	2020	XLNet	ResNet-50	Concatenation + fine-tuning	No
SAFE [5]	2020	Text-CNN	VGG-19	Similarity measurement	No
HMCAN [7]	2021	BERT	ResNet-50	Hierarchical attention	No
CAFE [6]	2022	BERT	ResNet-34	Cross-modal ambiguity	No
CLIP-FND [16]	2023	CLIP text	CLIP ViT	Shared CLIP space	No
FND-LLM [17]	2024	LLM	—	Text-only (LLM-based)	No
Ours	**2026**	**DeBERTa-v3**	**CLIP ViT-B/32**	**Cross-modal attention**	**Yes**

"Data Quality Focus" indicates whether the paper studies the effect of image curation quality on detection performance. All prior methods use automatically scraped or pre-existing dataset images without examining their relevance to article claims.

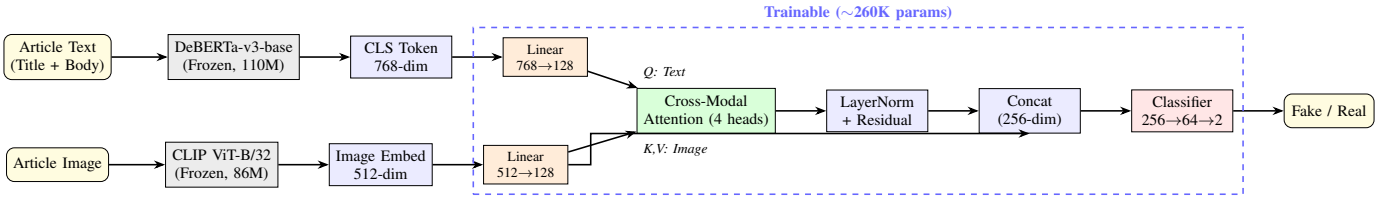

Fig. 1. Architecture of the Light Cross-Modal Attention model. Gray blocks are frozen encoders. The dashed box marks the trainable portion (260K parameters). Text is the Query; image is Key and Value in cross-modal attention.

TABLE II
TRAINABLE PARAMETER BREAKDOWN

Component	Parameters
Image projection (512→128)	65,664
Text projection (768→128)	98,432
Multi-head attention (4 heads)	66,048
Layer normalization	256
Classifier (256→64→2)	16,708
Total trainable	**∼247K**
DeBERTa-v3-base (frozen)	110M
CLIP ViT-B/32 (frozen)	86M

TABLE III
DATASET SUMMARY

Dataset	N	Fake/Real	Image Type
ISOT (text only)	44,158	22,741/21,417	None
Bing scraped	679	351/328	Generic
FakeNewsNet	60	28/32	Curated
Manual	150	75/75	Curated
Combined	210	103/107	Mixed

factual assertions in the article. This is the type of data that most previous research has utilized, hence it serves as our baseline image collection strategy.

FakeNewsNet curated (60 images). From the FakeNewsNet dataset [8], we downloaded images from PolitiFact article pages. These pictures were selected by fact-checkers to accompany their write-ups and they are more relevant than randomly scraped alternatives.

Manual collection (150 images). We sampled 150 samples (75 fake, 75 real) from 15 fact-checking organizations. Table IV lists the breakdown by source. In the case of fake news, the image is usually the image that has been manipulated or misleading, the photo that the fact-checker is refuting — an image made by AI, a photo that has been edited, or a photo used out of context. In the case of verifiable news, it is a confirmed photograph of the actual event. The images that we selected are not generic pictures but are directly related to the factual assertion made in the article.

TABLE IV
MANUAL DATASET SOURCES (150 SAMPLES)

Source	N	Source	N
AFP Fact Check	18	Factly	4
Reuters Fact Check	13	The Hindu	3
AP News	11	Full Fact	3
BBC Reality Check	9	Others (8 sources)	14
BOOM Live	4		
Total			**150**

Combined (210 images). FakeNewsNet 60 and Manual 150 were combined to form a dataset with two curation pipelines and source distributions. This tests whether models generalize across data sources.

C. Training Configuration

Multimodal experiments all work with pre-extracted features (CLIP: 512-dim, DeBERTa: 768-dim) stored as NumPy arrays. This enables training to be incredibly fast — a complete 5-fold cross-validation in just under 30 seconds on 150 samples on a Tesla T4 GPU. We train with AdamW (lr =

10^{-4}, weight decay = 0.01), batch size 8, and 20 epochs. A ReduceLROnPlateau scheduler (patience 3, factor 0.5) reduces the learning rate when the validation loss ceases to reduce. All our findings are 5-fold stratified cross-validation mean and standard deviation.

D. Ablation Variants

To see the contribution of each component, we compare five model variants:

1) **Text-Only**: DeBERTa CLS token fed into a 3-layer MLP (768 to 128 to 64 to 2). Checks whether text alone is sufficient.
2) **Image-Only**: CLIP embedding fed into a 3-layer MLP (512 to 128 to 64 to 2). Tests whether images alone carry the signal.
3) **Concat Fusion**: Concatenation of CLIP + DeBERTa to an MLP (1280 to 256 to 64 to 2). No attention, simple fusion.
4) **Cross-Modal Attention**: 256-dim projections, 8 heads, bigger classifier. Approximately 520K trainable parameters.
5) **Light CMA (ours)**: 128-dim projections, 4 heads, compact classifier. About 260K parameters. Half the size of variant 4.

IV. Experiments and Results

A. Text-Only Baseline

The test accuracy of DeBERTa-v3-base at 70/15/15 split is 92.0% on ISOT. We also applied it to 10 manually chosen real-life articles that contained political claims, health misinformation, and manipulated images. It got all 10 right. This verifies a robust text baseline, but can say nothing about image-based detection as ISOT does not have any paired images.

B. Image Quality: Controlled Comparison

This is the main experiment. Three image sources are compared with the same model (Light CMA) and using the same training procedures. Table V has the numbers.

TABLE V
Image Quality vs. Quantity (Light CMA, 5-Fold CV)

Image Source	N	Type	Accuracy
Bing (subsampled)	150	Generic	78.6% ± 1.5%
Manual (curated)	**150**	**Article-specific**	**93.3% ± 4.7%**
Bing (full)	679	Generic	88.8% ± 3.4%

Bing-150: mean over 10 stratified random subsamples. The 1.5% std is across seeds, not CV folds. Individual seeds ranged 70.0%–91.7%.

When matched in terms of sample sizes, the gap is 14.7 percentage points. To ensure that we did not draw an unlucky sample, we ran the Bing-150 subsampling 10 times with varying seeds. Standard deviation among seeds was only 1.5%, hence the result is consistent. No Bing subsample out of 10 achieved the average of the curated dataset of 93.3%.

Even more telling: 150 curated images beat the full 679-image Bing set by 4.5 points (93.3% vs. 88.8%). The addition of more generic images does help (78.6% increases to 88.8% with N increasing from 150 to 679), but it cannot compare to what careful curation can achieve with 4.5 times less data.

One caveat that can be mentioned in advance is that the curated and Bing images belong to different articles. The gap could be partly due to topic-level differences. In Section V, we discuss this. Fig. 2 shows the comparison visually.

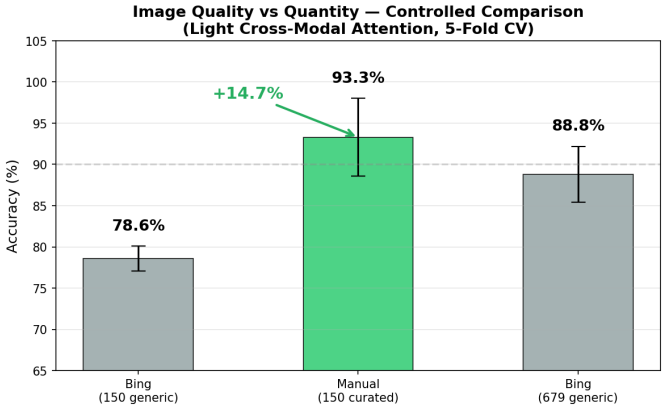

Fig. 2. Controlled comparison of image collection strategies. Curated 150 samples outperform both the size-matched Bing subset (78.6%) and the full Bing set (88.8%).

C. Ablation Study on 150 Manual Samples

Table VI shows all five variants on the 150-sample manual dataset.

TABLE VI
Ablation on 150 Manual Dataset (5-Fold CV)

Variant	Acc.	Std	Params
Text-Only (DeBERTa)	96.7%	5.2%	∼115K
Image-Only (CLIP)	97.3%	5.3%	∼83K
Concat Fusion	93.3%	5.6%	∼345K
Cross-Modal Attention	93.3%	5.6%	∼520K
Light CMA (ours)	95.3%	4.5%	∼260K

One good question: why do single-modality models beat fusion here? Both CLIP and DeBERTa can learn the distribution on their own with 150 samples in a single pipeline of curation. CLIP already possesses much knowledge about visual content because it has been pre-trained on 400M image-text pairs. The language priors provided by DeBERTa are also powerful. The addition of a fusion mechanism itself is simply the addition of additional parameters that have no payoff when a single modality already explains the valid patterns.

But see the standard deviations. Light CMA has the least at 4.5% — the most consistent across folds. And the next experiment shows the point of fusion.

D. Fusion Robustness on Mixed-Source Data

We combined FakeNewsNet 60 and Manual 150 into a 210-sample dataset of two curation styles and retrained all variations. Table VII tells the story.

TABLE VII
150 MANUAL VS. 210 COMBINED (5-FOLD CV)

Variant	150	210	Δ
Text-Only	96.7% ± 5.2	77.1% ± 3.2	–19.6%
Image-Only	97.3% ± 5.3	83.8% ± 6.8	–13.5%
Concat Fusion	93.3% ± 5.6	91.9% ± 2.9	–1.4%
Cross-Modal Attn	93.3% ± 5.6	91.9% ± 5.8	–1.4%
Light CMA	95.3% ± 4.5	90.5% ± 5.0	–4.8%

Text-only collapsed by 19.6 points. Image-only dropped 13.5. In the meantime, Concat Fusion and Cross-Modal Attention went down almost unscathed, with only 1.4 points apiece. Light CMA dropped 4.8 points — more so than the other variants in the fusion, but by far less than either of the unimodal models.

This is why fusion is worth it. On pure, homogeneous data, fusion is unnecessary overhead. Fusion performs consistently whilst single-modality models collapse on messy, mixed-source data, which is more likely to be encountered by a real-world system. When a change in the text distribution occurs (different styles of writing in different sources), the image channel covers. Text compensates when there are changes in the image characteristics. There is no such safety net in unimodal models.

Fig. 3 and Fig. 4 show these results graphically.

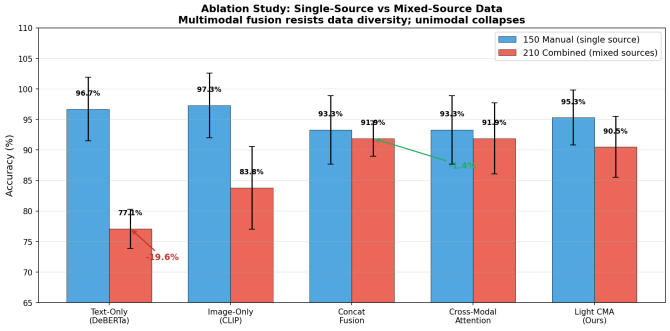

Fig. 3. Ablation: single-source (150) vs. mixed-source (210). Unimodal models degrade sharply; fusion methods hold.

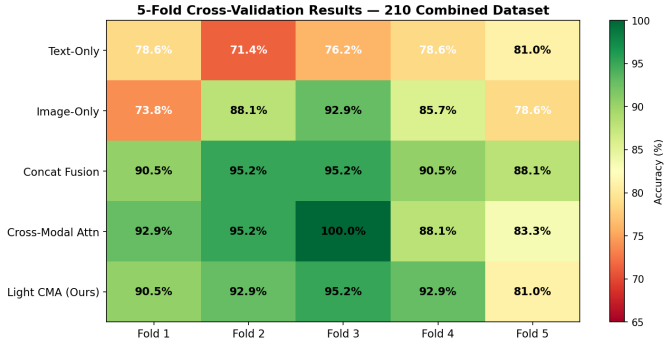

Fig. 4. Fold-level accuracy on 210 combined dataset. Green = high accuracy. Fusion rows stay consistently green; text-only and image-only show red/yellow variation.

E. Phase Progression

Table VIII traces how our experiments built on each other. The turning point was Phase 2: generic web-scraped images actually yielded worse accuracy (76.5%) compared to text alone (92.0%). It is that adverse outcome that led to the hypothesis that image quality, and not image quantity, is important in multimodal detection. The remainder of the experiments tested that hypothesis.

TABLE VIII
EXPERIMENTAL PHASE PROGRESSION

Phase	Experiment	N	Best Acc.
1	Text-only baseline (ISOT)	44,158	92.0%
2	+ Bing-scraped images	679	76.5%
3	+ FakeNewsNet curated	60	88.9%
4	+ Manual curated	150	93.3%
5	Combined (FNN + Manual)	210	91.9%

F. Deployment

We created a web application with Gradio [21]. Users fill in an article title, body text, and add a picture. The system uses frozen DeBERTa and CLIP to extract features, which are fed into the Light CMA classifier, and a prediction with confidence is displayed. Fig. 5 shows the interface classifying a fake news article at 100% confidence.

The system is based on Google Colab with a T4 GPU and creates a public URL. In testing, it made the right classification of articles in the training distribution but was not able to correctly classify entirely new topics. With 210 samples, this problem of generalization is expected and is addressed below.

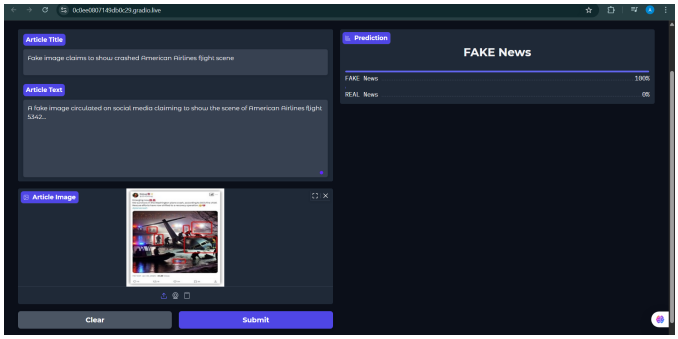

Fig. 5. Gradio deployment interface. The system correctly identifies a fake news article about an AI-generated crash image with 100% confidence.

V. DISCUSSION

A. Why Curated Images Are Better

Images scraped off websites are typically banners on the web site, stock photos, or logos of a news channel. Whether the article is real or fictional, a stock picture of the White House will appear the same. This is coded by CLIP in similar vectors regardless of the label and the classifier has nothing to work with.

Curated pictures are otherwise. A disproved fake image frequently has obvious artifacts: artificial distortions created by AI such as curved hands, shadows that are not consistent, or disjointed components with different resolutions. An actual news image of Reuters or AP is likely to exhibit the characteristics of professional photojournalism: correct exposure, coherent composition, and visual context which corresponds to the caption. These differences are picked up by CLIP and the cross-modal attention layer can be trained to balance them with what is being said in the text.

This is in line with the research on image manipulation detection [19], which has demonstrated that forensic clues in manipulated images have detectable evidence. Although these manipulations may not be explicitly coded by vision encoders such as CLIP, in the event that the training images actually contain manipulations as opposed to irrelevant stock photos, our work indicates that vision encoders can implicitly encode some of these cues.

B. Confounds We Cannot Ignore

There is a confound to the controlled comparison, which we need to be upfront about. The 150 curated pictures and 150 Bing pictures belong to different sets of articles. The fact-checking sites used to curate the articles are about the latest misinformation that includes AI-generated images, political deepfakes, and manipulated photos in 2025–2026. Bing articles were sourced from ISOT, which is U.S. political news from 2016–2017. The difference in topic between these sets might be partly responsible in explaining the difference in accuracy. The model could be capturing topic signals rather than (or in addition to) image quality signals.

An uncontaminated test would associate the identical articles with curated and generic images. This we could not do as the ISOT articles lack ground-truth article-specific images, and our fact-check articles lack Bing-scraped counterparts. The datasets that should be created in future work should enable such kind of a paired comparison.

C. Small Dataset Realism

Having 150 training samples, it is not as difficult to achieve a 93.3% accuracy as with 15,000 samples. We are not comparing with systems that are trained on larger data. This assertion is not of absolute performance but of relative performance: identical model, identical code, just changed the pictures, and improved the accuracy by 14.7 points. The number itself, 93.3%, is not the finding but that gap.

The variance is also brought on by the small size. On the 210-sample dataset, on the fold level, the individual folds are between 71.4% and 100.0%. That is a broad spread. A bigger curated set of data would narrow these ranges and render the conclusions more valid.

D. Generalization and Practical Limits

Most of the articles that we tested the deployed system on outside its training distribution were misclassified. The model has 210 samples and frozen encoders, and it learns its training distribution specifically instead of developing a general ability to detect fake news. This is the largest practical limitation.

To reach a production-ready system, either (a) a significantly larger curated dataset, probably thousands of samples with a wide range of topics and types of manipulation, or (b) methods such as few-shot adaptation, continual learning, or retrieval-augmented approaches would be needed. We consider the present work as a proof of concept: curated data is better than scraped data and the next thing is to scale up the process of curation.

VI. Conclusion

We contrasted three image collection methods to detect fake news in multimodal settings: large-scale web scraping (679 images using Bing), small curated collections (60 from FakeNewsNet, 150 manually collected), and a mixed collection (210 images). The selected 150-image curated set was rated higher than the size-matched generic images by 14.7 points and the entire 679-image generic set by 4.5 points. Multimodal fusion only decreased accuracy by 1.4% on mixed-source data, whereas unimodal models decreased by as much as 19.6%.

These findings suggest that researchers creating multimodal fake news datasets should dedicate time to image curation instead of executing larger web scrapers. We achieved good results with a smaller set of images that we carefully paired, compared to a large set of images with loosely related visuals. It is yet to be determined whether this is true at larger scales.

References

[1] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. ACM SIGKDD*, 2018, pp. 849–857.

[2] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "SpotFake: A multi-modal framework for fake news detection," in *Proc. IEEE ICMR*, 2019, pp. 451–455.

[3] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. WWW*, 2019, pp. 2915–2921.

[4] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty, and P. Kumaraguru, "SpotFake+: A multimodal framework for fake news detection via transfer learning," in *Proc. AAAI*, 2020.

[5] X. Zhou, J. Wu, and R. Zafarani, "SAFE: Similarity-aware fake news detection," in *Proc. AAAI*, 2020, pp. 13–20.

[6] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal ambiguity learning for multimodal fake news detection," in *Proc. WWW*, 2022, pp. 2897–2908.

[7] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in *Proc. ACM SIGIR*, 2021, pp. 153–162.

[8] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big Data*, vol. 8, no. 3, pp. 171–188, 2020.

[9] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing," *arXiv preprint arXiv:2111.09543*, 2021.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021, pp. 8748–8763.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[13] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in *Proc. ISDDC*, 2017, pp. 127–138.
[14] W. Y. Wang, "Liar, liar pants on fire: A new benchmark dataset for fake news detection," in *Proc. ACL*, 2017, pp. 422–426.
[15] R. Sharma and M. Gupta, "A survey on multimodal fake news detection," *Multimedia Tools and Applications*, 2025.
[16] S. Abdelnabi, R. Hasan, and M. Fritz, "Open-domain content-based multi-modal fact-checking of out-of-context images via online resources," in *Proc. IEEE/CVF CVPR*, 2022, pp. 14940–14949.
[17] Y. Jiang, X. Zhang, and R. Mao, "Fake news detection via large language models," *PLOS ONE*, vol. 19, no. 12, 2024.
[18] H. Li, J. Wang, and X. Chen, "Cross-modal contrastive learning for multimodal fake news detection," *Complex & Intelligent Systems*, 2025.
[19] M. Heller, J. Marti, and B. Schiele, "Ps-battles: A large-scale image manipulation detection benchmark," in *Proc. ACM Multimedia*, 2016.
[20] S. Wu, M. Li, and Y. He, "A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks," *Information Processing & Management*, vol. 58, no. 1, 2021.
[21] A. Abid, A. Abdalla, A. Abid, D. Khan, A. Alfozan, and J. Zou, "Gradio: Hassle-free sharing and testing of ML models in the wild," in *Proc. ICML*, 2019.