From Lacks to Leverage: Bilingual Benchmarking for Thumbnail Misinformation Detection: An Empirical Evidence from Visual-Textual Cues

Anonymous ACL submission

Abstract

Social media misinformation frequently exploits visually persuasive thumbnails in which image and text function as a unified deceptive artifact. Despite growing interest in multimodal fake news detection, empirical understanding of modality contributions under weak visual–textual alignment and in bilingual low-resource settings remains limited. We present 19,890 Bangla and English thumbnails from 36 news channels (9,936 for visual analysis; 9,954 OCR-extracted text samples from the same source pool), and conduct a systematic empirical comparison of multilingual transformer text classifiers, task-specific CNN variants, ImageNet-pretrained architectures, CLIP encoders, and decision-level late fusion. XLM-RoBERTa-base achieves a macro F1 of 97.20 on the mixed-language test set and 98.0 on Bangla alone. Task-specific CNNs beat all the popular pretrained models by 6 to 8 percentage points. That tells us the cues for thumbnail authenticity don't really overlap with those found in natural-image representations. When we ran a controlled alignment analysis—looking at distributional, synthetic same-class, and deliberate cross-class fusion setups—we measured a 16.5 point drop when pairing quality dropped, but also found that distributional alignment lands just 0.77 points away from the best possible results. So, our work gives you a solid, reproducible baseline and some real, practical advice for automating content moderation, even in messy, weakly aligned bilingual situations.

1 Introduction

Social media thumbnails convey deception through a compact, collaboratively constructed multimodal signal, consisting of a styled image overlaid with embedded text. This format is commonly found in Bangla and English news ecosystems, where politically motivated content mimics authentic media branding while employing sensationalist language. Unlike traditional fake news articles, which can be assessed independently, both the image and text together form a deceptive artifact in the case of thumbnails.

Existing multimodal detection systems typically assume that paired image-text samples allow for cross-modal alignment (Giachanou et al., 2020; Jing et al., 2022; Zhang, 2023).However, this assumption does not hold in the thumbnail context. Additionally, most prior research has primarily focused on English, with limited evaluation in multilingual environments. (Islam et al., 2020; Hossain et al., 2022). We aim to address these shortcomings through a controlled empirical study framed as source-authenticity fingerprinting. In this study, models learn the visual style, layout conventions, and linguistic tone distinguishing fake publishers from real ones, rather than verifying individual factual claims.

The main contributions of this work are:

- We have compiled a bilingual thumbnail dataset featuring a total of 9,936 images alongside 9,954 OCR-extracted text samples. This collection sources data from 36 different Bangla and English news channels utilizing source-level veracity labels to capture the unique stylistic fingerprints of both authentic and deceptive publishers.
- A systematic comparison of visual CNN architectures and multilingual transformer models, evaluating each modality independently under identical data splits.
- An analysis of multimodal fusion with weak alignment, including controlled experiments that measure how pairing quality affects detection accuracy.
- Cross-class fusion experiments that strategically pair fake images with real text, and vice versa, to evaluate how multimodal disagreement affects detection performance.

Together, these experiments provide valuable insight into multimodal misinformation signals when the correspondence between images and text is weak, establishing a baseline for detecting bilingual thumbnail misinformation.

2 Related Work

2.1 Text and Multimodal Fake News Detection

In the early stages of fake news detection, systems utilized classical machine learning methods such as Naive Bayes, Support Vector Machines (SVM), and logistic regression, applying these techniques to datasets like LIAR and Kaggle Fake News. As deep learning techniques advanced, models based on CNN text classifiers (Saleh et al., 2021), FNDNet using GloVe embeddings (Kaliyar, 2020), and hybrid CNN-LSTM architectures (Nasir et al., 2020; Umer et al., 2020) demonstrated improved performance on English-language datasets. More recently, transformer-based models, especially BERT (Devlin et al., 2019) and its multilingual variants, have become the dominant approach, as they effectively capture contextual semantic relationships within text. These models serve as the foundation for our text classification experiments.

Multimodal misinformation detection extends this idea by integrating both visual and textual evidence. This approach typically assumes that images accompanying an article are matched with corresponding text captions at the instance level. Giachanou et al. (Giachanou et al., 2020) combined BERT and VGG16 using an attention-based fusion method; Jing et al. (Jing et al., 2022) proposed progressive fusion of BERT and Swin Transformer features; Sharma et al. (Sharma et al., 2021) applied EfficientNetB0 and RoBERTa with late fusion and Error Level Analysis; Zhang et al. (Zhang, 2023) modeled cross-modal relationships using ViLBERT. In contrast to these approaches, our work focuses on thumbnail media where image and text signals may only be weakly aligned rather than forming direct semantic pairs.

Vision–language models such as CLIP (Radford et al., 2021) have recently gained attention for learning shared image–text representations from large-scale contrastive training on 400M image–text pairs. These models have been applied to misinformation detection tasks (Papadopoulos et al., 2023), and we include CLIP ViT-B/32 as a unimodal baseline. However, its 77 token text limit and caption-oriented pretraining make it less suitable for OCR-extracted thumbnail headlines, motivating the use of multilingual transformers for our textual branch.

2.2 Bilingual and Low-Resource Detection

Fake news detection research remains strongly English-centric despite the global spread of misinformation. Bangla misinformation detection has received comparatively limited attention. Hossain et al. (Hossain et al., 2022) introduced CB-Fake, a multimodal framework combining capsule neural networks and BERT for Bangla, demonstrating that multimodal signals can improve detection even in low-resource settings. Islam et al. (Islam et al., 2020) further highlight the challenges faced by English-centric models when applied to low-resource languages due to limited annotated data. This issue is particularly evident in Bangla NLP, where historical data scarcity has constrained model development (Hasan et al., 2020). Cross-lingual transformer models such as XLM-RoBERTa (Conneau et al., 2020) have therefore become a common solution, as cross-lingual transfer has shown strong improvements for misinformation detection in South Asian languages (Khalil et al., 2023). While Shibu (Shibu, 2025) explored LLM-based augmentation for low-resource misinformation detection, their work remained limited to textual data. Jing's study further demonstrates that multimodal systems can still operate under weak alignment conditions through progressive fusion, where distributional correspondence substitutes for strict instance-level pairing Segura-Bedmar and Alonso-Bartolome (Segura-Bedmar and Alonso-Bartolome, 2021) similarly emphasize the importance of bilingual datasets with reliable annotations for advancing misinformation detection beyond English. To our knowledge, no prior dataset jointly combines bilingual thumbnail imagery with OCR-extracted text under source-level veracity labels for both Bangla and English.

2.3 CNN Architectures for Digital Media Analysis

In image forensics and manipulation detection, task-specifi's CNN architectures often outperform ImageNet-pretrained models when the target domain differs substantially from natural photography (Rossler et al., 2019). Our experiments extend this observation to thumbnail-based misinformation detection, where authenticity cues such as compression artifacts, overlay typography, and stylized

color grading are not well represented in ImageNet-trained feature hierarchies.

3 Dataset Development

3.1 Collection

Lang.	Real Channels	Fake Channels
EN	5Pillars, AlJazeera, BangladeshPost, IslamChannel, MiddleEastEye, NBCNews, NewAgeBD, OwenJones, SecularTalk, TRTWorld, TheEconomist, TheFrontLines	americannews_en_fake, business_basic_en_fake, cipher_signal_en_fake, crime_watchers_en_fake, danny_haiphong_en_fake, mega_operations_en_fake, safety_zone_en_fake, the_left_lens_en_fake, truth_stream_en_fake, ukraine_today_en_fake
BN	BangladeshProtidin, BusinessInspectionBD, BusinessStandard, DeshTV, EkusheyTV, JamunaTV, MaasrangaTV, SorwarAlam	news_alert_bn_fake, newstoday_bn_fake, public360_bn_fake, public_insight_bn_fake, somoy_bn_fake, time24_bn_fake

Table 1: Source Channels Used for Thumbnail Collection

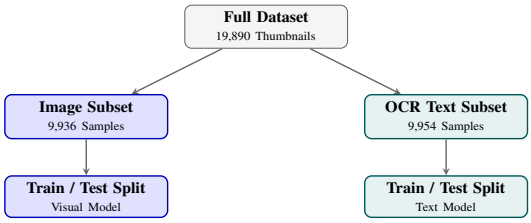

Figure 1: Dataset construction strategy. The full dataset of approximately 19k thumbnails is divided into two modality-specific subsets: one retained as images for visual modeling and the other processed via OCR for textual modeling. Each subset is independently split into training and testing partitions.

We collected 19,890 social media thumbnails from YouTube, Facebook, and Instagram using Python scripts that retrieved video identifiers and automatically downloaded the associated thumbnail images 1. The dataset covers 36 news channels: 14 Bangla sources (6 labeled fake, 8 real) and 22 English sources (10 fake, 12 real). Veracity labels are assigned at the source level using cross-referenced fact-checking reports and established media credibility resources. Channels categorized as fake have documented histories of fabricating or systematically distorting news, while real sources correspond to established outlets with editorial accountability.

After collection, thumbnails were merged according to their veracity category. All thumbnails from channels labeled as fake were combined into a single Fake pool, while those from channels labeled as real were combined into a single Real pool. This step ensured that thumbnails were treated as independent samples belonging only to their veracity class, without retaining any channel-level grouping during the subsequent modeling pipeline.

The dataset is divided into two subsets 1. One subset (9,936 samples) only has the raw images for the visual modeling pipeline, while from the other (9,954 samples) text is extracted using the Google Gemini OCR API to obtain the overlaid headline text. Although both subsets originate from the same source-level distribution pool, they do not correspond to identical thumbnails. Meaning the visual and textual modalities are aligned only at the distribution level rather than at the instance level. Table 2 summarizes the composition.

Lang.	Class	Train	Val	Test	Total	Ch.
Bangla	Fake	2,059	441	441	2,941	6
	Real	1,735	372	371	2,478	8
English	Fake	1,751	376	376	2,503	10
	Real	1,411	301	303	2,015	12
Total	—	**6,956**	**1,490**	**1,491**	**9,936**	36

Table 2: Dataset Composition (Image Modality). Text: 9,954 samples from the same sources.

(a) English real news (b) English fake news

Figure 2: Text-embedded thumbnails from the English subset.

(a) Bangla real news (b) Bangla fake news

Figure 3: Text-embedded thumbnails from the Bangla subset.

3.2 Preprocessing

Text was extracted via Google Gemini OCR, selected after comparative evaluation against Tesseract, EasyOCR, and OpenOCR on a pilot benchmark of Bangla-script thumbnails with stylized fonts and compression artifacts— The Gemini API produces context-aware OCR outputs with significantly lower character-level noise compared

to traditional OCR pipelines; therefore, no additional OCR-specific denoising module was required. Preprocessing included URL removal, emoji/symbol filtering, punctuation normalization, lowercasing, and Unicode-aware script preservation. Post-filtering yields 9,954 samples (5,447 fake / 4,507 real). Hash-based deduplication confirms zero cross-split overlap; within-training duplicates (2.3%) from similar templates are retained as real-world noise. Images were resized to 256×256 and normalized to [0,1] for CNNs; per-channel mean subtraction for CLIP. Stratified 70/15/15 splits applied within each language–class stratum.

Language	Class	Example Text
Bangla	Fake	টাইম BANGLADESH ডঃ ইউনূস নির্দেশ পক্ষপাতব দলে ফিরবে সাকিব, তামিম, মাশরাফি ?
	Real	সেতু নির্মাণে ত্রুটি বেরিয়ে পড়েছে গার্ডারের রড — Maasranga HD
English	Fake	DH Russian jets to Iran leaked!
	Real	N.K.H.A Al Jazeera Afghanistan--Pakistan truce offers hope to stranded truckers

Figure 4: Extracted Text Sample

3.3 Source Bias and Alignment Considerations

Here our data were labeled source-wise meaning models may learn publisher-specific fingerprints rather than generalizable misinformation indicators. We frame this as the intended task objective—in the thumbnail setting, fake content is systematically produced by a small ecosystem of outlets mimicking authentic branding, making source-level detection both valid and practically useful. The unpaired modality design precludes instance-level cross-modal reasoning; we therefore adopt decision-level fusion, explicitly avoiding feature-level or attention-based strategies that presuppose semantic alignment between specific image–text pairs.

4 Methodology

4.1 Text Classifiers

We fine-tune four multilingual transformers for binary classification: XLM-RoBERTa-base (Conneau et al., 2020) (125M params), mBERT-uncased and mBERT-cased (Devlin et al., 2019) (110M each), and DistilBERT-multilingual (Sanh et al., 2019) (66M). CLIP's text encoder (ViT-B/32; frozen + linear head) is included as a negative control. Training uses AdamW (lr $= 2\times10^{-5}$, weight_decay$=$ 0.01), batch size 16, up to 5 epochs, inputs truncated to 256 tokens. Each model is evaluated in BN-only, EN-only, and mixed configurations.

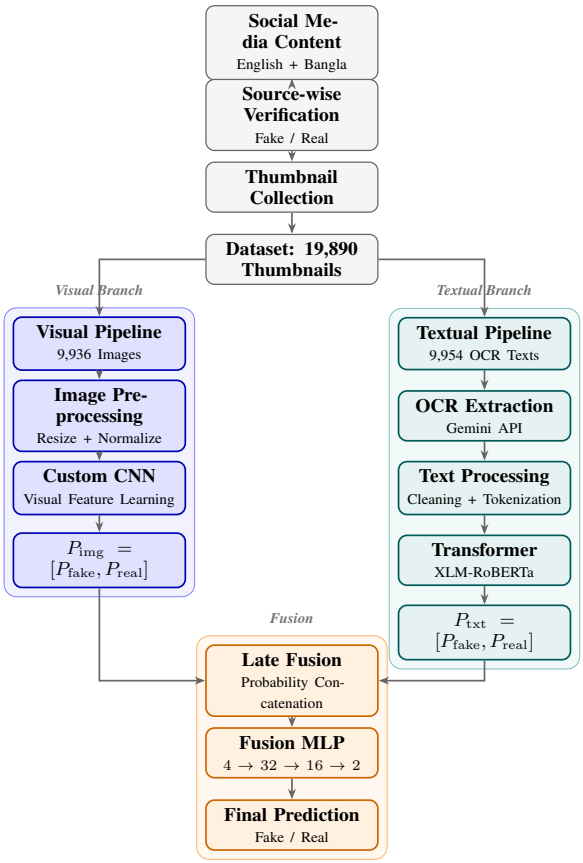

Figure 5: End-to-end architecture of the proposed multimodal fake news detection framework. Visual (blue) and textual (teal) modalities are processed in parallel and combined via probability-level late fusion (orange).

4.2 Image Classifiers

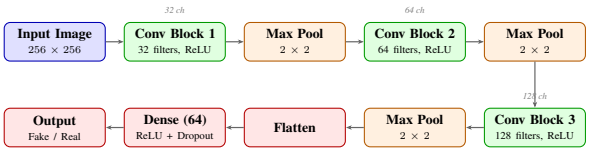

Figure 6: Base CNN architecture for visual fake news detection. Three convolutional blocks with progressive filter growth (32→64→128) Extract typography inconsistencies, compression artifacts, and layout deviations; MaxPooling preserves local feature extrema critical for thumbnail authentication.

We ablate five lightweight CNNs trained from scratch (Adam lr $= 1\times10^{-3}$, batch size 32, EarlyStopping patience=5, Dropout 0.5): (1) Base CNN—MaxPool, 32→64→128 filters, Flatten+Dense (64); (2) Shallow CNN—two

blocks, 32→64; (3) Constant Filter CNN—three blocks, 32→32→32; (4) GAP CNN—Base with Global Average Pooling; (5) AvgPool CNN—Base with AveragePooling. These are compared against four frozen ImageNet-pretrained models (MobileNetV2, InceptionV3, Xception, and ResNet50, all with GAP + linear head) and frozen CLIP ViT-B/32 with a linear head.

4.3 Late Fusion and Alignment Analysis

Each unimodal model outputs $[P(\text{fake}), P(\text{real})]$. For the primary fusion, the two probability vectors are concatenated into a 4-dim representation fed to an MLP (4→32→16→2, ReLU, Dropout 0.3/0.2), with frozen backbones (Adam lr $= 10^{-3}$, 10 epochs, $N = 6{,}340$ sorted-split-aligned samples). We test Base CNN + XLM-R and Base CNN + mBERT-cased.

To quantify the cost of alignment quality, we run two additional CLIP-backbone experiments. (1) *Alignment upper bound (UB)*: CLIP image and text features (512-dim each) concatenated into 1024-dim and trained with a deeper MLP (1024→256→64→2) under synthetic same-class pairing (fake images matched only with fake text, real with real). (2) *Misalignment lower bound (LB)*: probability-level weighted fusion at α=0.6 image weight, deliberately pairing fake images with real text and vice versa. Both CLIP experiments use frozen ViT-B/32 for both modalities to isolate the alignment strategy from the classifier choice.

These results validate the proposed methodology and highlight the effectiveness of late fusion for multilingual, low-resource, and weakly aligned multimodal fake news detection. Our codes are publicly available in this repository [1].

5 Results and Discussion

5.1 Image Modality

Table 3 summarizes the ablation results for all image models. The Base CNN achieves the highest performance with 96.78% accuracy and a macro F1 score of 0.967. Notably, the Shallow CNN trails by only 0.27 percentage points, suggesting that thumbnail authenticity signals can be captured using relatively shallow feature hierarchies. More importantly, all five task-specific CNNs outperform all pretrained architectures (Xception, InceptionV3, etc.) despite the latter having far more parameters.

[1] https://github.com/seekersg96-png/Anonymous_Repo

Table 3: Image Model Ablation. All custom CNNs trained from scratch. † backbone frozen. Best results in **bold**.

Model	Acc	P	R	F1	Notes
Base CNN	**96.78**	**.967**	**.968**	**.967**	Best
Shallow CNN	96.51	.963	.967	.965	−0.27
Const. Filter CNN	94.67	.950	.943	.946	No growth
GAP CNN	94.47	.943	.949	.944	Spatial↓
AvgPool CNN	92.37	.925	.920	.922	Worst
CLIP ViT-B/32†	92.15	0.92	0.92	.914	−4.6
MobileNetV2†	90.39	.904	.904	.903	−6.4
Xception†	89.47	.895	.895	.894	−7.3
InceptionV3†	88.16	.883	.883	.881	−8.6
ResNet50†	82.96	.833	.830	.827	−13.8

Architecture details: Base CNN = convolution filters (32→128) with MaxPooling; Shallow CNN = filters (32→64) with MaxPooling; Const. Filter CNN = constant 32 filters across layers; GAP CNN = Global Average Pooling instead of flattening; AvgPool CNN = average pooling used instead of max pooling. † Pretrained backbone frozen during training. The differences listed in the Notes column are measured in percentage points relative to the Base CNN.

The best pretrained backbone (CLIP ViT-B/32) achieves 92.15% accuracy, while ResNet50 performs significantly worse at 82.96%. We attribute this to a domain mismatch: ImageNet features target natural object recognition, while thumbnail authentication requires sensitivity to digital artifacts (compression patterns, font discontinuities, and overlay anomalies) typically absent from natural photography.

Pooling strategy also plays a critical role. Replacing MaxPooling with AveragePooling reduces F1 performance by 4.5 percentage points, indicating that discriminative thumbnail cues often manifest as sharp local feature extrema. Similarly, replacing the Flatten layer with Global Average Pooling decreases performance by 2.3 percentage points, confirming that spatial layout information remains important for authenticity classification.

5.2 Text Modality

Table 4: Text Model Comparison (Mixed BN+EN). † backbone frozen.

Model	Acc	P	R	F1	Params
XLM-R	**97.09**	**97.26**	**97.22**	**97.20**	125M
mBERT-u	96.95	96.84	97.08	96.93	110M
mBERT-c	96.82	96.87	96.82	96.80	110M
Distil-mBERT	96.54	96.59	96.54	96.52	66M
CLIP-Text†	82.13	83.00	81.20	78.90	63M

Model abbreviations: XLM-R = XLM-RoBERTa-base; mBERT-u = Multilingual BERT (uncased); mBERT-c = Multilingual BERT (cased); Distil-mBERT = DistilBERT-multilingual; CLIP-Text = CLIP Text Encoder.

Table 4 compares the performance of multilingual transformer models on the combined Bangla–English dataset. XLM-RoBERTa achieves the highest overall performance with an accuracy of 97.09% and a macro F1 score of 97.20.

DistilBERT follows closely, with an F1 score just 0.68 points lower, while utilizing approximately 59 million fewer parameters. This makes DistilBERT a strong option for deployment in resource-constrained environments.

In contrast, the CLIP text encoder significantly lags behind, achieving an F1 score of only 78.9. This notable performance gap can be attributed to CLIP's limit of 77 tokens and its training objective, which focuses on captions that are not well-suited for the dense text extracted from OCR thumbnails.

Interestingly, multilingual BERT models reach near-zero training loss within eight epochs while maintaining a stable test performance of 96.82%. This suggests that they exhibit tight distributional fitting rather than robust generalization.

5.3 Per-Language Evaluation

Table 5: Per-Language Results. Δ vs. mixed-training baseline.

Model	Tr.	Acc	$F1_f$	$F1_r$	Mac.	Δ
XLM-R	BN	**98.00**	.980	.980	**.980**	+0.91
XLM-R	EN	95.70	.960	.950	.957	−1.39
XLM-R	Mix	97.09	.971	.971	.971	base
CNN	BN	96.61	.968	.965	.966	−0.17
CNN	EN	96.49	.965	.965	.965	−0.29
CNN	Mix	96.78	.967	.968	.967	base

Abbreviations: Tr. = Training data; BN = Bengali; EN = English; Mix = Mixed BN+EN training; Mac. = Macro F1; Δ = change vs. mixed-training baseline (percentage points).

Table 5 presents performance across language-specific training configurations. The Bangla-only XLM-R model achieves the highest performance, with a macro F1 score of 0.980. This represents a 0.91 percentage point improvement over mixed-language training. In comparison, training conducted exclusively in English performs 1.39 points worse than the mixed training baseline.

This asymmetry likely indicates more consistent stylistic conventions in Bangla fake-news thumbnails, whereas English thumbnails show greater stylistic diversity.

For the visual modality, Performance is nearly identical across Bangla-only, English-only, and mixed datasets, with differences under 0.3 percentage points. This finding confirms that visual authenticity cues, including layout, color schemes, and branding patterns, function independently of language.

5.4 Fusion Results and Alignment Analysis

Table 6: Fusion Results Across Alignment Conditions.

Config.	Pair	Acc	F1	Δ	Role
Text: XLM-R	—	97.09	97.20	—	Base
Image: CNN	—	96.78	96.70	—	Base
Dec.: CNN+XLM-R	Dist.	**97.15**	**97.14**	+0.06	Primary
Dec.: CNN+mBERT	Dist.	97.15	97.13	+0.33	Primary
Feat. (UB)	Same	97.92	97.68	—	Align. UB
Prob. (LB)	Cross	81.38	81.39	—	Align. LB

Abbreviations: Config. = Configuration; Pair = Pairing type; Dist. = Distributed pairing; Feat. = Feature-level fusion; Prob. = Probability-level fusion; Base = Baseline; UB = Upper Bound; LB = Lower Bound; Align. = Alignment condition.

Primary result. Both decision-level fusion setups reach an accuracy of 97.15%, which is a small improvement of 0.06–0.33 percentage points over the best text-only model (Table 6). The Multi-Layer Perceptron (MLP) converges quickly, finishing in about one epoch as the loss drops from 0.279 to 0.065. Some oscillation appears afterward, which is expected since both unimodal models already achieve 96–97% accuracy. Their errors overlap on only about 3–4% of samples. Although the gain is small, it shows that visual and textual signals provide complementary information. The image branch captures cues that may not appear in OCR text, especially when images are manipulated or text extraction is incomplete.

Alignment upper bound. To estimate the benefit of perfect cross-modal pairing, we test feature-level CLIP fusion using synthetic same-class alignment. A 1024-dimensional MLP trained for 15 epochs reaches 97.92% accuracy with an F1 score of 97.68. The validation F1 increases from 0.961 to 0.975. The 0.77 percentage point gap compared to the primary result likely comes from two factors: deeper fusion (1024-dimensional features vs. 4-dimensional scores) and better alignment, since this setup uses oracle pairing while the main approach relies on distribution-level signals. A truly paired dataset would be needed to fully separate these effects.

Misalignment lower bound. To study modality conflict, we create cross-class pairs where fake images are matched with real text and vice versa. At $\alpha = 0.6$, this misaligned setup reaches 81.38% accuracy (F1 = 0.814), which is 10.8 points lower than the CLIP image baseline of 92.15%. Varying α shows modality asymmetry: accuracy rises

from 64.4% when text dominates ($\alpha = 0.3$) to 90.5% when the image modality has more weight ($\alpha = 0.7$).

Alignment cost. The gap between the lower bound (81.4%) and upper bound (97.9%) is about 16.5 percentage points, giving an estimate of how much pairing quality affects multimodal thumbnail fusion. Importantly, the main result (97.15%) is only 0.77 points below the synthetic upper bound even without instance-level pairing. This finding supports the main idea of this work that source-authenticity fingerprinting can roughly replicate the advantages of explicit multimodal alignment when paired image–text data is not available.

5.5 Error Analysis

Error analysis reveals several consistent patterns across class, modality, and language dimensions. The fusion model achieves 97.15% accuracy, with higher precision for the Fake class (0.986) and higher recall for the Real class (0.984), indicating a conservative prediction behavior that avoids incorrectly labeling legitimate content as misinformation. Examination of the confusion matrix shows that 31 fake samples were misclassified as real, while only 11 real samples were incorrectly predicted as fake, suggesting a low false-positive tendency. At the modality level, the two branches exhibit complementary strengths: the XLM-R text model reaches 97.09% accuracy, effectively capturing linguistic manipulation patterns, while the CNN image model achieves 96.78%, detecting visual cues such as exaggerated typography and characteristic thumbnail styling. Language-specific evaluation further shows that text performance is higher for Bengali (98.00%) than English (95.70%), likely reflecting greater lexical variability in English headlines, whereas the CNN maintains nearly identical performance across languages (96.61% BN vs. 96.49% EN), suggesting that visual misinformation patterns are largely language-independent. Finally, architectural ablations highlight the importance of spatial cues: replacing flattening with global average pooling reduces F1 from 0.967 to 0.944, indicating that layout information such as text placement contributes to fake thumbnail detection.

6 Conclusion

We conducted a study on a bilingual thumbnail misinformation dataset, focusing on weak visual-textual alignment, and derived three important insights. First, we found that task-specific CNNs trained on thumbnail data significantly outperform traditional ImageNet-pretrained models like Xception, InceptionV3, MobileNetV2, ResNet50, and CLIP ViT-B/32, with improvement ranging from 4 to 14 percentage points. This suggests that the authenticity cues in thumbnails differ substantially from those in natural images. Moreover, the choice of pooling methods made a notable difference—MaxPool outperformed AvgPool by 4.5 percentage points, while Global Average Pooling (GAP) had a 2.3 percentage point advantage over Flatten, underscoring their critical role as visual discriminants. Secondly, our analysis emphasized the dominance of linguistic signals, with the XLM-RoBERTa model achieving an impressive F1 score of 97.20 in mixed contexts and 98.0 in Bangla-only contexts, far surpassing the CLIP text encoder's F1 score of 78.9. This highlights the necessity of robust multilingual modeling for text derived from OCR. Lastly, we discovered that decision-level fusion provides a slight yet consistent enhancement, with improvements ranging from 0.06 to 0.33 percentage points over the best unimodal baseline. This finding illustrates that visual and linguistic signals can offer complementary insights, even when the modalities are not perfectly aligned.

Overall, these findings establish a reproducible empirical baseline for detecting bilingual thumbnail misinformation. They also validate source-authenticity fingerprinting as an effective design strategy for multimodal misinformation detection, particularly when explicit image-text pairing is unavailable.

7 Limitations

Four limitations affect the generalizability of the current study. First, because labels are assigned at the source level, the models may partially learn channel-specific stylistic fingerprints rather than broadly transferable indicators of misinformation; evaluation with held-out sources would help quantify this effect. Second, the binary *fake/real* formulation collapses several distinct phenomena—including satire, fabrication, and misleading framing—so future work with finer-grained taxonomies could enable more targeted modality analysis.

- **Source-level labeling:** Models may capture channel-specific visual or linguistic patterns

rather than general misinformation signals; held-out source evaluation would better assess cross-source robustness.

- **Binary labels:** The fake/real dichotomy merges multiple forms of misleading content (e.g., satire, fabrication, framing), limiting more nuanced analysis.

- **OCR variance:** Although Gemini outperformed other OCR systems in benchmarking, degraded thumbnails still introduce recognition errors; the Bangla–English text performance gap (98.0% vs. 95.7%) may therefore partly reflect OCR limitations rather than purely linguistic differences.

- **Unpaired modalities:** Because the visual and textual datasets are aligned only at the distribution level, the fusion training set is limited to 6,340 samples and more advanced feature-level or cross-attention fusion strategies cannot be evaluated.

References

Alexis Conneau and 1 others. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Anastasia Giachanou, Guanyang Zhang, and Paolo Rosso. 2020. Multimodal multi-image fake news detection. In *Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW)*.

T. Hasan and 1 others. 2020. Not low-resource anymore: Alleviating data scarcity in bangla language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

M. Z. Hossain and 1 others. 2022. Cb-fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and bert. *Multimedia Tools and Applications*, 81(5):5587–5620.

M. R. Islam, S. Liu, X. Wang, and G. Xu. 2020. Deep learning for misinformation detection on online social networks: A survey. *Social Network Analysis and Mining*, 10(1).

J. Jing and 1 others. 2022. Multimodal fake news detection via progressive fusion networks. *Information Processing and Management*, 59(1).

R. K. Kaliyar. 2020. Fndnet: A deep convolutional neural network for fake news detection. *Computer Communications*, 153:36–42.

M. Khalil and 1 others. 2023. Cross-lingual transfer learning for misinformation detection in south asian languages. *IEEE Access*, 11:45210–45225.

J. A. Nasir and 1 others. 2020. Fake news detection: A hybrid cnn-rnn based deep learning approach. *International Journal of Information Management Data Insights*, 1(1).

S. I. Papadopoulos and 1 others. 2023. Synthetic misinformation detection using clip-based image-text alignment. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Alec Radford and 1 others. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Andreas Rossler and 1 others. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

H. Saleh and 1 others. 2021. Opcnn-fake: Optimized convolutional neural network for fake news detection. *IEEE Access*, 9:129471–129489.

Victor Sanh and 1 others. 2019. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. In *Proceedings of the NeurIPS EMC^2 Workshop*.

Isabel Segura-Bedmar and Sara Alonso-Bartolome. 2021. Multimodal fake news detection. *arXiv preprint arXiv:2112.04831*.

B. Sharma and 1 others. 2021. Predicting image credibility in fake news over social media using multimodal approach. *Neural Computing and Applications*.

H. M. Shibu. 2025. From scarcity to capability: Empowering fake news detection in low-resource languages with llms. *arXiv preprint arXiv:2501.09604*.

M. Umer and 1 others. 2020. Fake news stance detection using deep learning architecture (cnn-lstm). In *Proceedings of the IEEE International Conference*.

X. Zhang. 2023. Multimodal fake news analysis based on image–text similarity. ResearchGate Preprint.