# Detection of Text Overwriting Manipulation in Indonesian Language News Images Using Deep Learning

Julian Caleb Simandjuntak
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
13522099@std.stei.itb.ac.id

Rinaldi
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
rinaldi@staff.stei.itb.ac.id

***Abstract*—The spread of misinformation through text manipulation in images has become an increasingly common problem on social media platforms like Instagram and Facebook, highlighted by the emerging Indonesian phenomenon known as *timpa teks*, where text within news images is altered to create misleading or satirical content. Conventional image processing methods have shown limited effectiveness in addressing this issue, as they tend to produce incorrect detections when applied to complex or low contrast backgrounds. This study presents a multimodal deep learning detection system that combines EfficientNetV2 for visual feature extraction, IndoBERT for textual feature extraction from text obtained through EasyOCR, and a three-way scaled dot-product attention module to combine both types of information. A dedicated dataset was constructed for this study, comprising 2,400 news images with an equal number of authentic and manipulated samples, collected from Facebook groups and official Indonesian news social media accounts. The dataset was divided into 2,000 training samples, 200 validation samples, and 200 testing samples. Test results show that the proposed multimodal approach achieves an accuracy of 0.8950 and an F1-score of 0.8923, surpassing other approaches, including single-modality (visual or textual) analysis and conventional methods.**

***Keywords—text manipulation in images, image manipulation detection, deep learning, EfficientNetV2, IndoBERT, EasyOCR***

## I. Introduction

The rapid advancement of technology has accelerated the spread of information across social media platforms such as Instagram and Facebook, particularly via visual news containing embedded text. However, irresponsible parties frequently exploit these platforms by altering the text within genuine news images using editing software or AI and redistributing them to spread hoaxes or misinformation. This phenomenon, known in Indonesia as *timpa teks* (text overwriting), poses a significant risk of misleading social media users who cannot easily discern the content's authenticity (as illustrated in Fig. 1).

This issue falls under the category of image tampering, defined as the addition, modification, or removal of significant features within an image without visible traces, in order to alter the information or facts it conveys [1]. In the case of text overwriting manipulation, the specifically altered component is the text embedded within the image itself, typically a news caption or explanation, modified using image editing tools such that the conveyed meaning changes substantially.


Fig. 1. Original news image by Kompas.com (left) and its manipulated counterpart (right)

Prior efforts to detect such manipulations using conventional digital image processing techniques (e.g., thresholding, edge detection, and contour analysis) have proven inadequate. These methods frequently yield high false-positive rates on complex backgrounds and fail to detect low-contrast text regions [2]. Such limitations have motivated the exploration of more adaptive deep learning and multimodal detection frameworks, as discussed further in Section II.

Although a range of fake news and image manipulation detection approaches have been developed, no existing study specifically addresses text manipulation embedded directly within Indonesian-language news images, as observed in the text overwriting phenomenon on social media. This study aims to address this gap through the following contributions:

1) Developing a news image dataset containing text, consisting of authentic images and images manipulated through text overwriting.

2) Developing and training three deep learning models, comprising a visual-based model, a textual-based model, and a multimodal model, to detect the authenticity of text-embedded news images using the constructed dataset.
3) Evaluating and comparing the performance of the three developed deep learning models on the test data.

This research is subject to the following limitations: (1) the dataset contains manipulations applied to textual content within the images; (2) the dataset is restricted to news images containing Indonesian-language text; and (3) the study focuses on authenticity classification without performing manipulation localization.

## II. Related Works

To date, no existing study has applied deep learning to detect text overwriting manipulation, although a traditional image processing approach has been proposed for this specific problem. Waan [2] developed a GUI-based application that detects text overwriting through a sequence of five image processing stages, which include grayscale conversion, Otsu thresholding, Canny edge detection, morphological dilation, and contour analysis. Within this framework, a manipulation decision was made when at least 15 text-like contours with a combined area of at least 8,000 pixels were found. While this pipeline required no training data and could run with minimal computational resources, it produced false positives on genuine news images with complex backgrounds. Furthermore, it remained highly sensitive to lighting variation, low-contrast text, small text size, and transparent text, without yielding a reliable probabilistic output suitable for integration into more advanced detection systems.

To address the limitations of threshold-dependent methods, deep learning approaches to image forgery detection have been explored. Sudiatmika et al. [3] combined Error Level Analysis (ELA) with a VGG-16-based CNN on the CASIA v2.0 dataset (7,491 authentic and 5,123 forged images), reaching 92.2% training accuracy and 88.46% validation accuracy after 100 epochs, demonstrating that JPEG compression inconsistencies serve as a strong forensic signal without relying on high-level semantic features. However, ELA is effective only on JPEG-format images and depends heavily on compression history, degrading after repeated re-saving, and it does not consider any embedded textual modality, so it cannot capture the semantic inconsistency between visual content and text that characterizes text overwriting. Korsipati et al. [4] later proposed a transfer-learning framework based on EfficientNetV2-B0 enhanced with Squeeze-and-Excitation (SE) blocks, evaluated across four forgery benchmarks (CASIA v1, Columbia, MICC-F2000, and Defacto Splicing) with accuracies of 83.76%, 94.24%, 96.92%, and 99.97% respectively, showing that EfficientNetV2 generalizes robustly across manipulation types without depending on compression artifacts. This study, however, addresses general-purpose image forgery detection only, without considering the textual modality embedded in news images. Joshi et al. [5] further investigated this direction by evaluating several state-of-the-art image classification architectures on ELA-processed images from the CASIA ITDE v.2 dataset, testing VGG-19, Inception-V3, ResNet-152-V2, XceptionNet, and EfficientNetV2L through transfer learning from ImageNet weights. Their results showed no single architecture dominating across all metrics. VGG-19 achieved the highest precision (98.25%) but the lowest recall (79.34%), whereas EfficientNetV2L exhibited the opposite pattern, with the lowest precision (85.20%) but the highest recall (94.60%) among all models, while also generalizing better than ResNet-152-V2, which showed a substantially larger gap between training and validation accuracy. This comparative study offered a clearer picture of the precision-recall trade-offs across popular ELA-based forensic architectures. However, it remained restricted to JPEG-compressed images through ELA, did not consider the textual modality, and did not test lighter EfficientNetV2 variants more suitable for computationally efficient, small-scale systems.

In parallel, textual approaches to fake news detection have relied on transformer-based language models. Ramzan et al. [6] compared Logistic Regression, Random Forest, LSTM, and BERT on 44,897 English-language news titles, finding that fine-tuned BERT (80% accuracy) substantially outperformed classical methods (40%) and LSTM (60%) due to its bidirectional, context-aware representation. This study, however, relied solely on English text and did not address Indonesian-language content or OCR-extracted text. Rijal et al. [7] extended this direction by fine-tuning IndoBERT on the Indonesian Fact and Hoax Political News Dataset (18,960 authentic and 10,381 hoax articles), achieving 92% accuracy and F1-score, confirming that IndoBERT captures Indonesian linguistic nuances more accurately than general-purpose models. Nevertheless, the model was trained on complete, well-formed news articles, and its learned representation is not guaranteed to transfer effectively to the short, incomplete, and recognition-error-prone text produced by OCR from news images, as encountered in text overwriting cases.

To overcome the limitations of single-modality approaches, multimodal frameworks combining visual and textual features have also been proposed. Tuan and Minh [8] introduced an architecture that fuses BERTweet-based textual features and VGG-19-based visual features through a three-way scaled dot-product attention mechanism, consisting of text as query with image as key-value, image as query with text as key-value, and image self-attention, evaluated on the MediaEval 2016 dataset of 17,000 Twitter posts, achieving 81.2% accuracy with F1-scores of 84.3% and 76.7% for the fake and real classes. This confirmed that explicit cross-modal attention captures correlations that simple feature concatenation cannot. However, this approach does not incorporate forensic visual signals, was designed for English-language social media captions rather than Indonesian text, and does not target text that is physically embedded within the image itself, as is the case in text overwriting manipulation.

Overall, these studies show that conventional image processing methods lack robustness against complex

backgrounds and varying text properties, while single-modality visual approaches, including ELA-based methods and EfficientNetV2, are designed for manipulations that differ from manipulated embedded text, as they are evaluated on general-purpose forgery datasets such as CASIA v1 that do not reflect the visual characteristics of Indonesian news images, which typically combine a photograph, a fixed typographic layout, and an embedded news caption. Likewise, single-modality NLP approaches, including IndoBERT, are optimized for complete, well-formed articles rather than the short, incomplete, and recognition-error-prone text produced by OCR from news images. Meanwhile, existing multimodal frameworks with cross-modal attention, such as [8], treat text and image as separate, loosely coupled modalities, such as a tweet and its attached photo, rather than text that is physically embedded within the image itself, so the inconsistency to be detected in text overwriting is intra-image rather than cross-source. An additional challenge specific to this domain is the high visual variability across Indonesian news publishers in aspect ratio, typography, and layout, which further limits the applicability of rule-based or hand-crafted feature detection.

## III. Proposed Method

These gaps motivate the multimodal detection framework proposed in this study, which integrates EfficientNetV2 for visual feature extraction, IndoBERT for OCR-derived textual feature extraction, and a three-way scaled dot-product attention module tailored specifically to the text overwriting phenomenon in Indonesian-language news images. Building on this motivation, three classification approaches are developed and compared. The first is a unimodal visual model, motivated by the potential of text overwriting to introduce visible artifacts such as color or typography inconsistencies. The second is a unimodal textual model, motivated by the possibility that the overwritten caption itself contains linguistic cues such as informal language or typographical errors. The third is a multimodal model that fuses both modalities, motivated by the finding of Tuan and Minh [8] that inconsistency between an image and its accompanying text is a strong indicator of manipulation. Fig. 2 to Fig. 4 show the workflow diagrams for the unimodal visual, unimodal textual, and multimodal approaches, respectively. The dataset, system architecture, graphical interface, and baseline comparison are detailed in Subsections A-D.

### *A. Dataset Construction*

A dedicated dataset of 2,400 Indonesian-language news images was constructed, evenly split between 1,200 authentic and 1,200 manipulated samples. Authentic images were collected unaltered from official Indonesian news social media accounts (e.g., Tempo, Kompas). Manipulated images were obtained from two sources. The first source consisted of 600 real-world manipulated images collected from Facebook groups where text overwriting content circulates organically, providing unknown and highly varied manipulation styles that reflect real misinformation conditions. The second source consisted of 600 manually manipulated images created by applying five manipulation types namely copy-move, splicing, insertion, coverage, and inpainting, using Figma for the first four types and Grok for inpainting, applied while preserving each publisher's original visual style for realism. The source images used to generate this second manipulated subset were excluded from the authentic set to prevent data leakage.

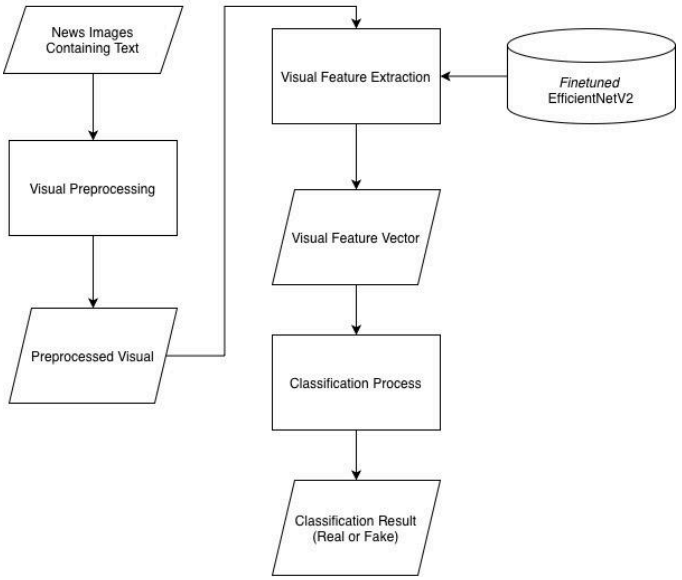


Fig. 2. Workflow diagram of the proposed system using the unimodal visual approach.

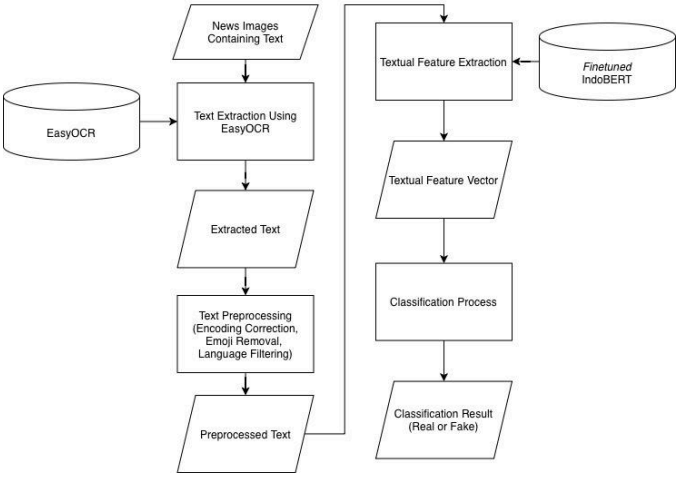


Fig. 3. Workflow diagram of the proposed system using the unimodal textual approach.

The dataset was partitioned using a stratified split into 2,000 training (1,000/1,000), 200 validation (100/100), and 200 testing (100/100) images. All images were standardized to JPEG, converted to RGB, and padded to a square aspect ratio. Training images were further augmented with two randomly selected transformations (e.g., rotation, brightness/contrast adjustment) at a fixed intensity level, followed by tensor normalization using ImageNet statistics. Validation and test images, in contrast, were only resized and normalized without augmentation. As an additional visual preprocessing step, Error Level Analysis (ELA) at a re-compression quality of 90

was explored under three configurations, namely no ELA, ELA-only, and a 70:30 blend of the original image and its ELA map, to evaluate which input best exposes compression-related manipulation residue without excessively distorting visual content.

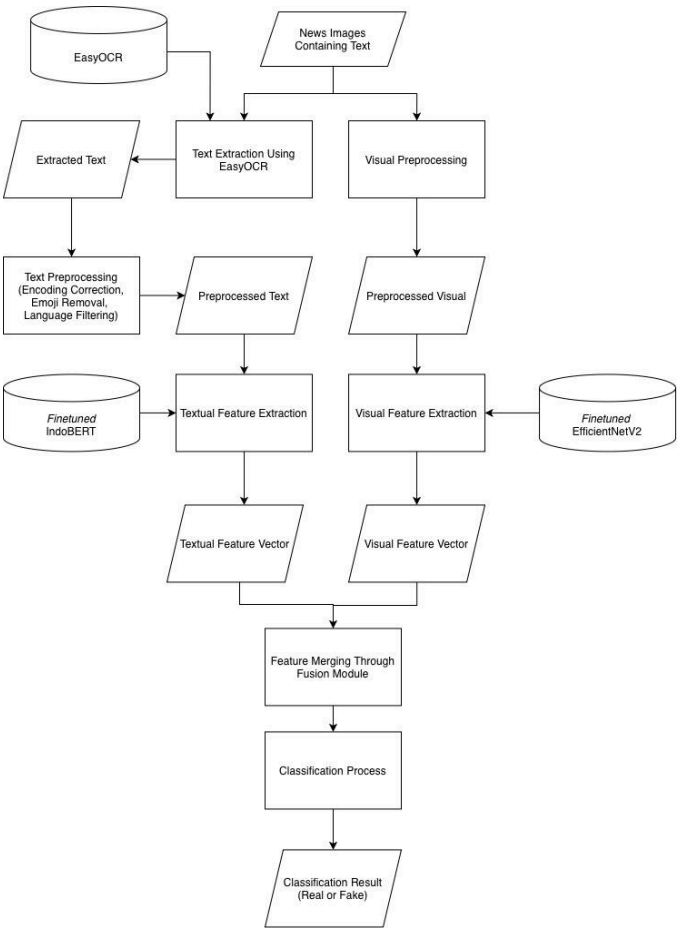


Fig. 4. Workflow diagram of the proposed system using the multimodal approach.

For the textual modality, EasyOCR was applied with a confidence threshold of 0.4 using both Indonesian and English language models, with results cached to a CSV file to avoid redundant OCR computation. Extracted text was cleaned using the ftfy library for encoding repair, emoji removal, and punctuation normalization, then tokenized using the uncased IndoBERT tokenizer with padding and truncation. Class labels were encoded as Fake = 1 and Real = 0.

### *B. System Architecture*

The visual feature extraction pipeline, shown in Fig. 5, passes a preprocessed image tensor of shape $(B, 3, H, W)$, where $B$ is the batch size, $H$ is the height of the input image, and $W$ is the width of the input image, through an ImageNet-pretrained EfficientNetV2 backbone (via the timm library) with its classification head removed, producing a spatial feature map of shape $(B, C, H/32, W/32)$, with $C$ denotes the number of output channels, which depends on the EfficientNetV2 variant used.

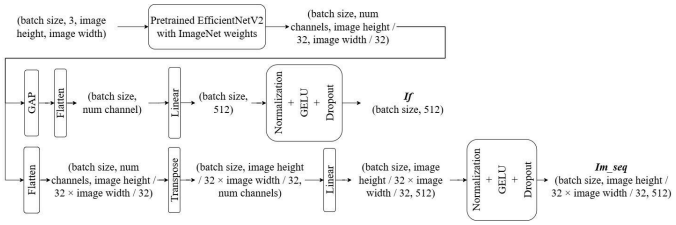


Fig. 5. Detailed visual feature extraction architecture, showing the two processing branches producing $If$ and $Im\_seq$.

This feature map is processed along two parallel paths. The first applies global average pooling and flattening, followed by a linear projection to dimension 512, layer normalization, GELU activation, and dropout, producing a pooled visual representation $If$ with a shape of $(B, 512)$. The second path flattens and transposes the feature map into a sequence of shape $(B, H/32 \times W/32, C)$, which then linearly projected to 512, normalized, activated, and regularized with dropout, producing a spatial sequence representation $Im_seq$with a shape of $(B, H/32 \times W/32, 512)$ that preserves spatial information for the fusion stage.

The textual feature extraction pipeline, shown in Fig. 6, tokenizes preprocessed OCR text using the IndoBERT-Base tokenizer and passes it through a fine-tuned IndoBERT backbone (via HuggingFace transformers), producing token-level embeddings of shape $(B, L, 768)$, where L is the token sequence length. Layer-wise Learning Rate Decay (LLRD) is applied during fine-tuning, assigning smaller learning rates to lower layers, which retain general pretrained linguistic representations, and larger learning rates to upper layers, which must adapt more specifically to the classification task.

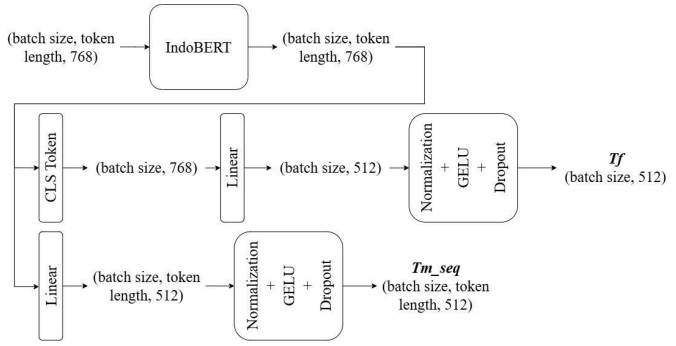


Fig. 6. Detailed textual feature extraction architecture, showing the two processing branches producing $Tf$ and $Tm\_seq$.

Analogous to the visual branch, two processing paths are applied. The first extracts the [CLS] token representation, projects it to dimension 512, and applies normalization, GELU, and dropout, producing a pooled textual representation $Tf$ with a shape of $(B, 512)$. The second projects the full last-layer token sequence to dimension 512 with the same

normalization–activation–dropout sequence, producing $Tm\_seq$ with a shape of $(B, L, 512)$. In the unimodal configurations, $If$ (visual) or $Tf$ (textual) is passed directly to the classification head. In the multimodal configuration, $If$ and $Tf$ serve as attention queries, while $Im\_seq$ and $Tm\_seq$ serve as the corresponding key/value pairs in the fusion module described next.

The fusion module, detailed in Fig. 7 and implemented entirely in torch, receives four inputs: $If$, $Tf$, $Im\_seq$, and $Tm\_seq$. Before attention, $If$ and $Tf$ are unsqueezed to shape $(B, 1, 512)$ and normalized.

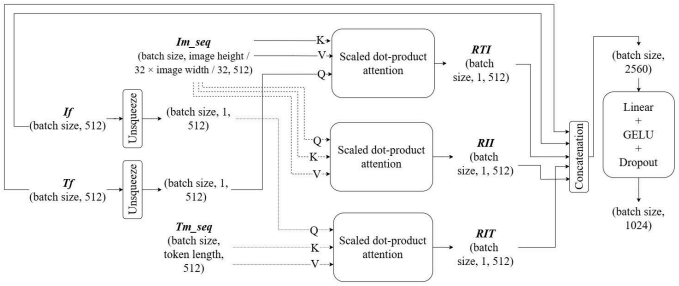


Fig. 7. Detailed architecture of the three-way multi-head attention fusion module, following Tuan and Minh [8].

Following the three-way attention design of Tuan and Minh [8], three attention blocks are computed: $Tf$ as query attending to $Im\_seq$ as key/value produces $RTI$; $If$ as query attending to $Tm\_seq$ as key/value produces $RIT$; and $Im\_seq$ self-attention produces $RII$. Each block uses multi-head attention with 8 heads, enabling the model to capture cross-modal relationships across multiple representation subspaces in parallel. The three attention outputs are concatenated with the original pooled representations $If$ and $Tf$, forming a combined representation of dimension 2,560, which is projected through a linear layer, GELU activation, and dropout to obtain the final fused representation used for classification.

A shared classification head, shown in Fig. 8, is used across all three configurations, differing only in input dimensionality. The head consists of a normalization layer, a linear layer for dimensionality reduction, GELU activation, dropout, and a final linear layer followed by softmax to produce class probabilities (authentic/manipulated).

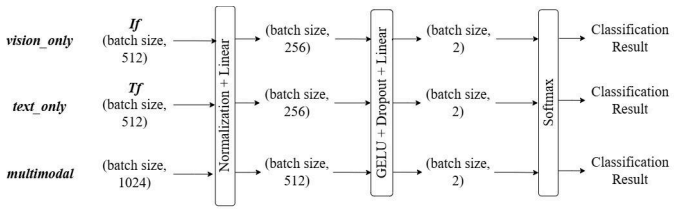


Fig. 8. Shared classification head architecture across the unimodal visual, unimodal textual, and multimodal configurations.

## C. Graphical User Interface

A graphical user interface (GUI) is designed to allow users to upload a news image, select one of the three model configurations (unimodal visual, unimodal textual, or multimodal), and obtain a classification result along with the predicted confidence for each class, as shown in Fig. 9. The interface is implemented as a locally-hosted web application using Gradio, supported by torch, transformers, easyocr, and pillow for inference. The implementation is organized into separate modules for configuration constants, model architecture definitions, and preprocessing utilities, allowing each component to be modified independently.

The interface supports two main workflows. Users first select a model via a dropdown and click "MUAT MODEL" to load it, after which a confirmation message indicates the model is ready. Users then upload a news image and click "KLASIFIKASI", triggering image preprocessing, OCR-based text extraction and cleaning, and joint inference, concluding with the display of the predicted class and its associated confidence percentages.

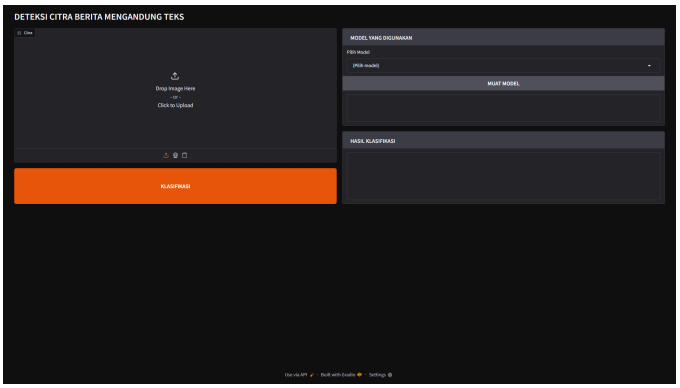


Fig. 9. Design of the graphical user interface for the proposed classification system.

## D. Baseline

To quantify the contribution of the proposed approach, the thresholding-and-edge-detection-based system of Waan [2] is re-implemented as a non-learned baseline using OpenCV. Unlike the three deep-learning configurations, this baseline is fully deterministic and requires no training: each test image undergoes grayscale conversion, Otsu thresholding, Canny edge detection, morphological dilation (3×3 kernel, 1 iteration), and contour analysis filtered by bounding-box area (100-3,000 px) and aspect ratio (0.2-5.0). An image is classified as manipulated if at least 15 qualifying contours are found with a combined area of at least 8,000 pixels, otherwise, it is classified as authentic. The baseline is evaluated on the same 200-image test set and the same metrics (accuracy, precision, recall, F1-score) as the proposed models, enabling direct comparison.

## IV. Experiments and Results

### A. Experimental Setup

The experimental framework was implemented using a hybrid computational environment partitioned by task requirements. Dataset acquisition, system architecture integration, and the graphical user interface (GUI) development were executed locally on a host system equipped with an Intel Core i5-12450H processor and 16 GB of RAM. For computationally intensive phases, such as dataset preprocessing, model training, and quantitative evaluation, operations were offloaded to a cloud-based Kaggle Notebook environment leveraging dual NVIDIA Tesla T4 GPUs and 31 GB of system memory. All deep learning architectures and text-processing workflows were developed in Python utilizing the torch, transformers, timm, easyocr, and ftfy libraries, while the interactive user interface was deployed via the Gradio framework.

Testing was conducted in five stages. The first two stages evaluated the unimodal visual and unimodal textual architectures independently on the validation set to determine the optimal configuration for each backbone. The best-performing configurations were then used in the third stage to examine the effect of training data volume on the multimodal model, evaluated on the validation set. In the fourth stage, the complete system, comprising the unimodal visual, unimodal textual, multimodal, and baseline approaches, was evaluated on the held-out test set to assess generalization performance. The final stage tested the graphical interface to verify that all three inference modes produced correct predictions. Model performance was measured using loss, accuracy, precision, recall, and F1-score, with confusion matrices and a gradient sensitivity analysis additionally used to examine the fusion module in detail. Performance analysis specific to each individual manipulation technique used in the manually created subset was not conducted, since the sample size per technique was too small to yield representative conclusions.

### B. Dataset Analysis

Fig. 10 shows the class distribution of the constructed dataset, confirming a total of 2,400 images evenly split between 1,200 authentic and 1,200 manipulated samples, with the manipulated samples further divided into 600 manually manipulated images and 600 images obtained from social media.

Visual inspection of the dataset (Fig. 11) revealed that authentic images generally consist of a background photograph overlaid with a colored caption box containing explanatory text, along with the publisher's official logo. Manipulated images sourced from social media exhibited visible anomalies such as inconsistent background color behind portions of the text, mismatched fonts, irregular text sizing, and in some cases embedded watermarks indicating manipulation. Manually manipulated images, in contrast, exhibited subtler visual anomalies, though inconsistencies such as misaligned center-justified text were still occasionally observable.

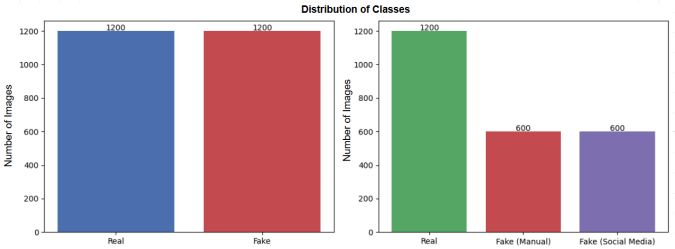


Fig. 10. Class distribution of the constructed dataset, showing the split between authentic and manipulated samples, and between manually manipulated and social-media-sourced manipulated samples.

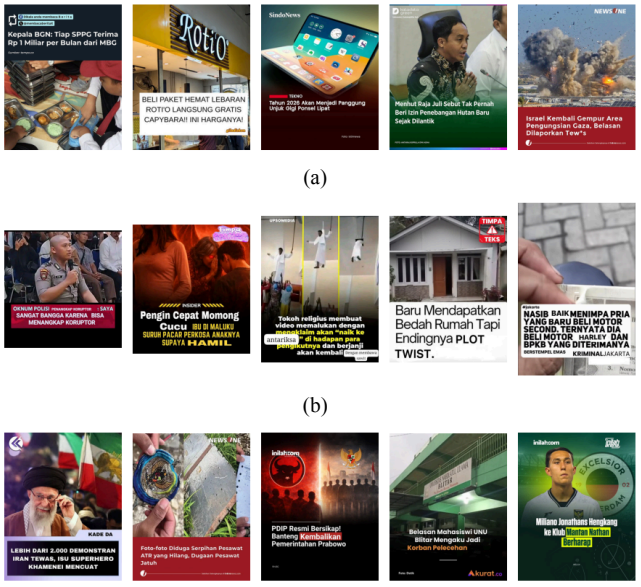


Fig. 11. Representative examples of (a) authentic images, (b) manipulated images obtained from social media, and (c) manually manipulated images.

Applying ELA (Fig. 12) revealed that authentic images produced no sharply concentrated residue blocks under either the ELA-only or blended configuration. Manipulated images obtained from social media showed only faint residue concentration around edited text regions, whereas manually manipulated images exhibited clearer and more concentrated residue blocks with cleaner edges around the edited regions, suggesting that manual manipulation introduces more precise and consistent text insertion compared to manipulations of unknown origin from social media.

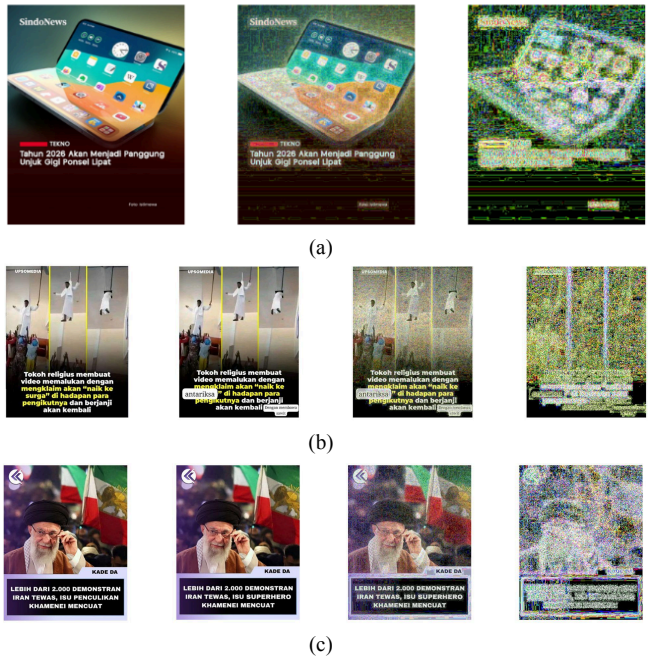


Fig. 12. Comparison of ELA outputs across (a) authentic images, (b) social-media-sourced manipulated images, and (c) manually manipulated images, under the image-only, blended, and ELA-only configurations.

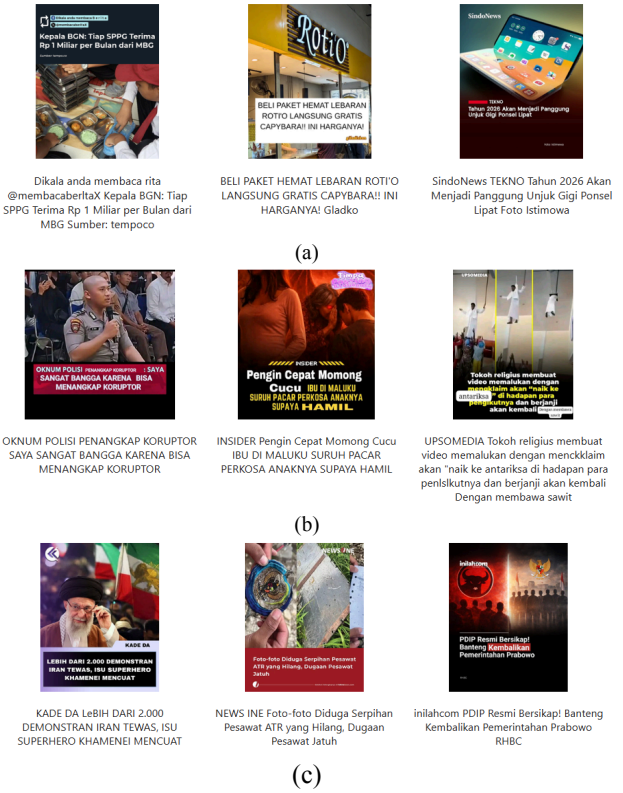


Fig. 13. Examples of text extracted by EasyOCR from (a) authentic images, (b) social-media-sourced manipulated images, and (c) manually manipulated images

EasyOCR extraction (Fig. 13) was found to capture all visible text elements in an image, including publisher account names and source attributions, introducing potential noise into the extracted text. For manipulated images, occasional character recognition failures were observed around edited regions, particularly involving punctuation.

Table I illustrates the effect of text preprocessing, showing that it successfully reduced redundant punctuation and excess whitespace, although residual noise such as irrelevant account names and typographical errors remained.

TABLE I. COMPARISON OF OCR TEXT BEFORE AND AFTER PREPROCESSING

| Text Before Preprocessing | Text After Preprocessing |
|---|---|
| YUK MARI KITA BINGUNG DENGAN DERETAN PRODUK JADUL PILIHAN INI , SOBAT >>>>> giladiskon | YUK MARI KITA BINGUNG DENGAN DERETAN PRODUK JADUL PILIHAN INI , SOBAT > giladiskon |
| BERSAMA BERITASATU Gudang penyimpanan doksili Terbang dan Timpa penimpa Warga di Jember | BERSAMA BERITASATU Gudang penyimpanan doksili Terbang dan Timpa penimpa Warga di Jember |
| INDOMARET JUGAADA ICE CREAM FAIR, HARGA MULAI 6 RIBUAN!!! | INDOMARET JUGAADA ICE CREAM FAIR, HARGA MULAI 6 RIBUAN! |
| HANYA HARI INI! ICE CREAM DAY DI INDOMARET! ADA BUY 2 GET 1!!! | HANYA HARI INI! ICE CREAM DAY DI INDOMARET! ADA BUY 2 GET 1! |

The token length distribution (Fig. 14) was right-skewed, peaking between 20 and 25 tokens, with the majority of samples falling below 40 tokens, confirming that OCR-extracted text from news images is generally short.

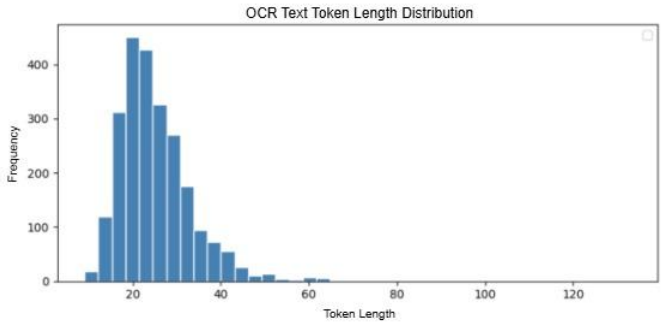


Fig. 14. Distribution of token lengths for tokenized OCR text.

## C. Visual Architecture Experiments

The unimodal visual configuration was finetuned on training set and tested on the validation set across four factors: EfficientNetV2 variant, input resolution, learning rate, and ELA configuration, using the fixed training setup in Table II (AdamW optimizer, cross-entropy loss, batch size 32, max 40 epochs, early stopping patience 10).

TABLE II. VALIDATION PERFORMANCE ACROSS VISUAL HYPERPARAMETERS

| Configuration | Loss | Accuracy | F1-Score | Precision | Recall |
|---|---|---|---|---|---|
| Architecture Variation | | | | | |
| S | 0.5026 | 0.7600 | 0.7391 | 0.8095 | 0.6800 |
| M | 0.4722 | 0.7800 | 0.7556 | 0.8500 | 0.6800 |
| **L** | **0.5869** | **0.8300** | **0.8440** | **0.7797** | **0.9200** |
| XL | 1.3618 | 0.7250 | 0.7120 | 0.7473 | 0.6800 |
| Image Size | | | | | |
| 128 | 0.6021 | 0.6850 | 0.6480 | 0.7342 | 0.5800 |
| 224 | 0.5104 | 0.7400 | 0.6905 | 0.8529 | 0.5800 |
| **400** | **0.5869** | **0.8300** | **0.8440** | **0.7797** | **0.9200** |
| Learning Rate for Backbone | | | | | |
| **2e-5** | **0.5869** | **0.8300** | **0.8440** | **0.7797** | **0.9200** |
| 1e-5 | 0.5173 | 0.7800 | 0.7442 | 0.8889 | 0.6400 |
| 5e-6 | 0.4377 | 0.8100 | 0.8021 | 0.8370 | 0.7700 |
| Learning Rate for Classification Head | | | | | |
| 2e-4 | 0.5110 | 0.7450 | 0.7052 | 0.8356 | 0.6100 |
| **1e-4** | **0.5869** | **0.8300** | **0.8440** | **0.7797** | **0.9200** |
| 5e-5 | 0.5879 | 0.7900 | 0.8205 | 0.7164 | 0.9600 |
| ELA Application | | | | | |
| Without ELA | 0.5869 | 0.8300 | 0.8440 | 0.7797 | 0.9200 |
| ELA only | 0.4170 | 0.8700 | 0.8713 | 0.8627 | 0.8800 |
| **Blended** | **0.3637** | **0.8850** | **0.8844** | **0.8889** | **0.8800** |

The best configuration, EfficientNetV2-L at 400×400 resolution with backbone LR 2e-5, head LR 1e-4, and blended ELA, achieved a validation accuracy of 0.8850 and F1-score of 0.8844, and was carried forward to subsequent experiments.

### *D. Textual Architecture Experiments*

The unimodal textual configuration was tuned on maximum token length and learning rates using the fixed setup in Table III (AdamW, LLRD factor 0.9, batch size 32, max 40 epochs).

The best configuration, IndoBERT with 256 max tokens, backbone LR 2e-5, and head LR 1e-4, achieved a validation accuracy of 0.6650 and F1-score of 0.6912, noticeably lower than the visual branch, indicating that OCR-extracted text alone is a weaker discriminative signal for this task.

TABLE III. VALIDATION PERFORMANCE ACROSS TEXTUAL HYPERPARAMETERS

| Configuration | Loss | Accuracy | F1-Score | Precision | Recall |
|---|---|---|---|---|---|
| Token Length Maximum | | | | | |
| 128 | 0.6096 | 0.6550 | 0.6667 | 0.6900 | 0.6449 |
| **256** | **1.0308** | **0.6650** | **0.6912** | **0.6410** | **0.7500** |
| 512 | 0.6663 | 0.6450 | 0.6380 | 0.6510 | 0.6255 |
| Learning Rate for Backbone | | | | | |
| 5e-6 | 0.6236 | 0.6450 | 0.6728 | 0.7300 | 0.6239 |
| 1e-5 | 0.6154 | 0.6550 | 0.6462 | 0.6300 | 0.6632 |
| **2e-5** | **1.0308** | **0.6650** | **0.6912** | **0.6410** | **0.7500** |
| Learning Rate for Classification Head | | | | | |
| 5e-5 | 0.8880 | 0.6100 | 0.6803 | 0.5764 | 0.8300 |
| **1e-4** | **1.0308** | **0.6650** | **0.6912** | **0.6410** | **0.7500** |
| 2e-4 | 0.6218 | 0.6100 | 0.6355 | 0.6800 | 0.5965 |

### *E. Effect of Training Data Volume*

Using the selected visual and textual configurations, the multimodal model was trained on 50%, 75%, and 100% of the training data (Table IV).

TABLE IV. EFFECT OF TRAINING DATA VOLUME ON VALIDATION PERFORMANCE

| Amount of Training Data | Accuracy | F1-Score |
|---|---|---|
| 100% | 0.8800 | 0.8812 |
| 75% | 0.8250 | 0.8128 |
| 50% | 0.8150 | 0.7910 |

Accuracy and recall increased consistently with more data, while precision was highest at 50%, suggesting that the model trained on less data adopts a more conservative prediction strategy. This trend confirms that further data collection would likely yield additional performance gains.

### *F. Overall System Evaluation*

Table V compares the unimodal visual, unimodal textual, multimodal, and baseline approaches across the training, validation, and test sets.

TABLE V. PERFORMANCE COMPARISON ACROSS ALL DATA SPLITS

| Configuration | Loss | Accuracy | F1-Score | Precision | Recall |
|---|---|---|---|---|---|
| Training Data | | | | | |
| **Unimodal Visual** | **0.1693** | **0.9780** | **0.9779** | **0.9838** | **0.9720** |
| Unimodal Tekstual | 0.2460 | 0.9375 | 0.9371 | 0.9433 | 0.9310 |
| Multimodal | 0.1949 | 0.9645 | 0.9647 | 0.9594 | 0.9700 |
| Baseline | - | 0.4810 | 0.6255 | 0.4893 | 0.8670 |
| Validation Data | | | | | |
| **Unimodal Visual** | **0.5216** | **0.8850** | **0.8856** | **0.8812** | **0.8900** |
| Unimodal Tekstual | 0.9902 | 0.7550 | 0.6797 | 0.9811 | 0.5200 |
| Multimodal | 0.7503 | 0.8800 | 0.8812 | 0.8725 | 0.8900 |
| Baseline | - | 0.4700 | 0.4833 | 0.8700 | 0.6214 |
| Test Data | | | | | |
| Unimodal Visual | 0.5625 | 0.8400 | 0.8261 | 0.9048 | 0.7600 |
| Unimodal Tekstual | 1.1879 | 0.7050 | 0.6093 | 0.9020 | 0.4600 |
| **Multimodal** | **0.6982** | **0.8950** | **0.8923** | **0.9158** | **0.8700** |
| Baseline | - | 0.4850 | 0.6255 | 0.4914 | 0.8600 |

On the training and validation sets, the unimodal visual approach performed best, but on the held-out test set, the multimodal approach achieved the highest accuracy and F1-score, outperforming the unimodal visual approach on every metric, including recall (0.870 versus 0.760), without sacrificing precision. This reversal indicates that the unimodal visual model is more prone to overfitting to training-specific characteristics, whereas the multimodal model generalizes more reliably by leveraging complementary textual information. The unimodal textual approach performed weakest among the learning-based methods, confirming that OCR-extracted text alone is insufficient as a standalone signal, while the baseline underperformed all learning-based methods throughout.

Fig. 15 presents the test-set confusion matrices for all four approaches. The multimodal approach shows the most balanced error distribution, consistent with its high precision and recall. The unimodal visual approach shows fewer false negatives, the unimodal textual approach shows a bias toward negative predictions, and the baseline shows a bias toward positive predictions.

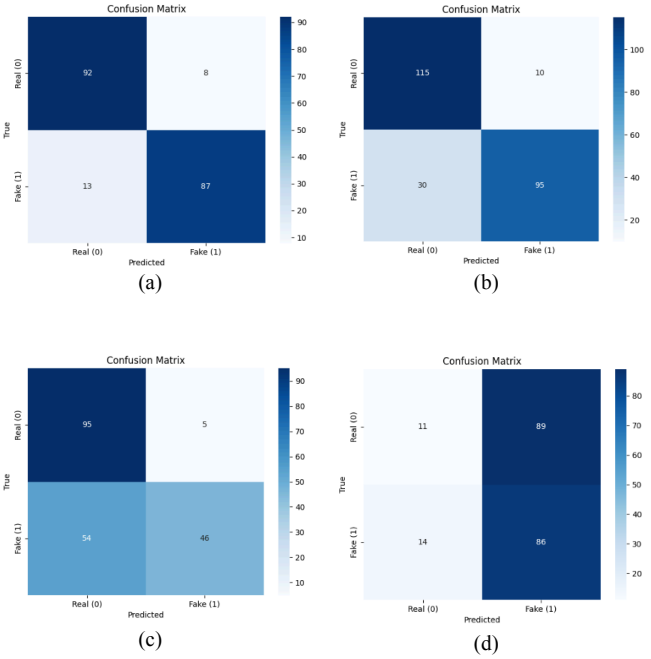


Fig. 15. Confusion matrices on the test set for (a) multimodal, (b) unimodal visual, (c) unimodal textual, and (d) baseline approaches.

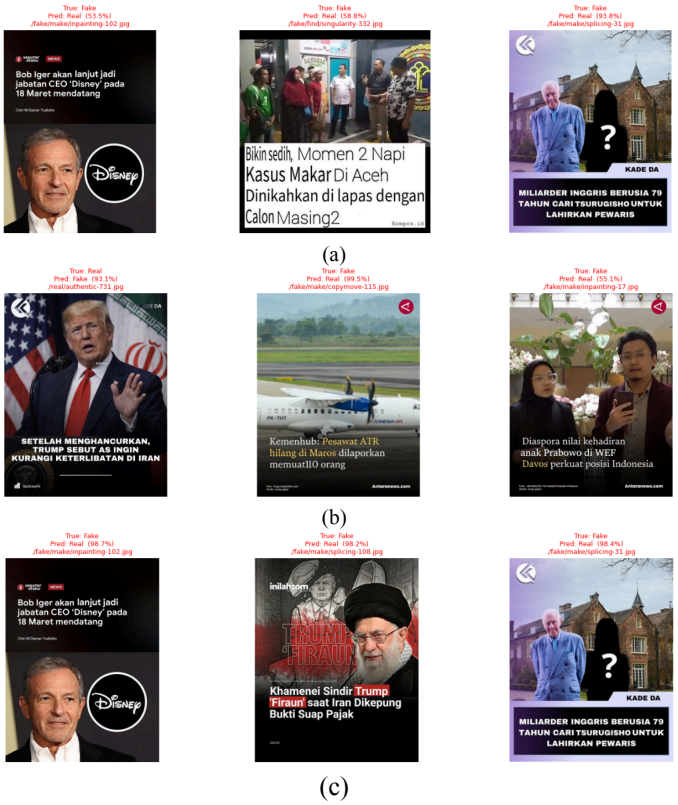


Fig. 16. Examples of misclassified samples under the (a) unimodal visual approach, (b) unimodal textual approach, and (c) multimodal approach.

Qualitative inspection of misclassified samples (Fig. 16) showed that multimodal errors typically occurred on samples

ambiguous in both modalities, unimodal visual errors on images with minimal visual difference between classes, and unimodal textual errors on images with poor OCR quality due to low resolution, occlusion, or unusual layout.

A gradient sensitivity analysis (Fig. 17) further showed that the pooled visual representation $If$ (0.281) and pooled textual representation $Tf$ (0.251) contributed more strongly to the final decision than the attention-derived representations $RTI$ (0.143), $RIT$ (0.154), and $RII$ (0.171), indicating that the fusion module adaptively weights the pooled modality representations as the dominant contributors.

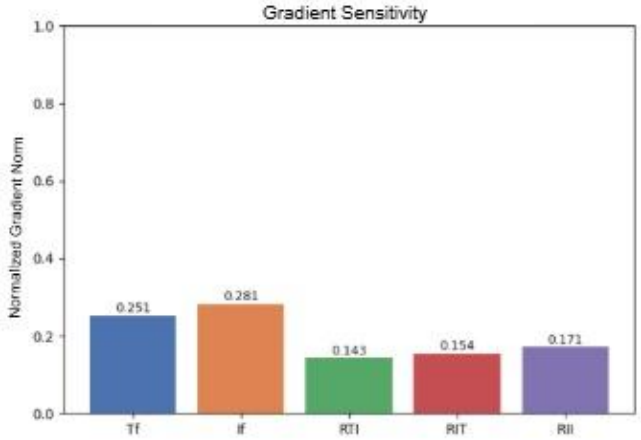


Fig. 17. Normalized gradient sensitivity of each representation within the fusion module.

Overall, these results confirm that the multimodal approach achieves the best generalization among all evaluated methods, supporting the hypothesis that combining visual and textual information yields a richer representation for detecting text overwriting manipulation in Indonesian news images.

### *G. Graphical Interface Testing*

The graphical interface was tested using all three inference modes on the same test sample to compare prediction characteristics directly, as shown in Fig. 20. In the unimodal textual mode (Fig. 20a), the model predicted the Fake class with 99% confidence based solely on OCR-extracted text. In the unimodal visual mode (Fig. 20b), the model predicted the Fake class with 93% confidence based solely on visual features. In the multimodal mode (Fig. 20c), combining both modalities through the fusion module produced the highest confidence among the three modes, at 100%.

All three modes agreed in predicting the Fake class, confirming that the interface correctly and consistently displays inference results across all configurations. The high confidence scores across all modes further suggest that the tested sample exhibited clear manipulation indicators in both modalities. Despite this consistent behavior, the current interface is limited to local execution, requiring installation and lacking online accessibility, and requires users to wait for model weights to load from local storage each time the inference mode is switched.

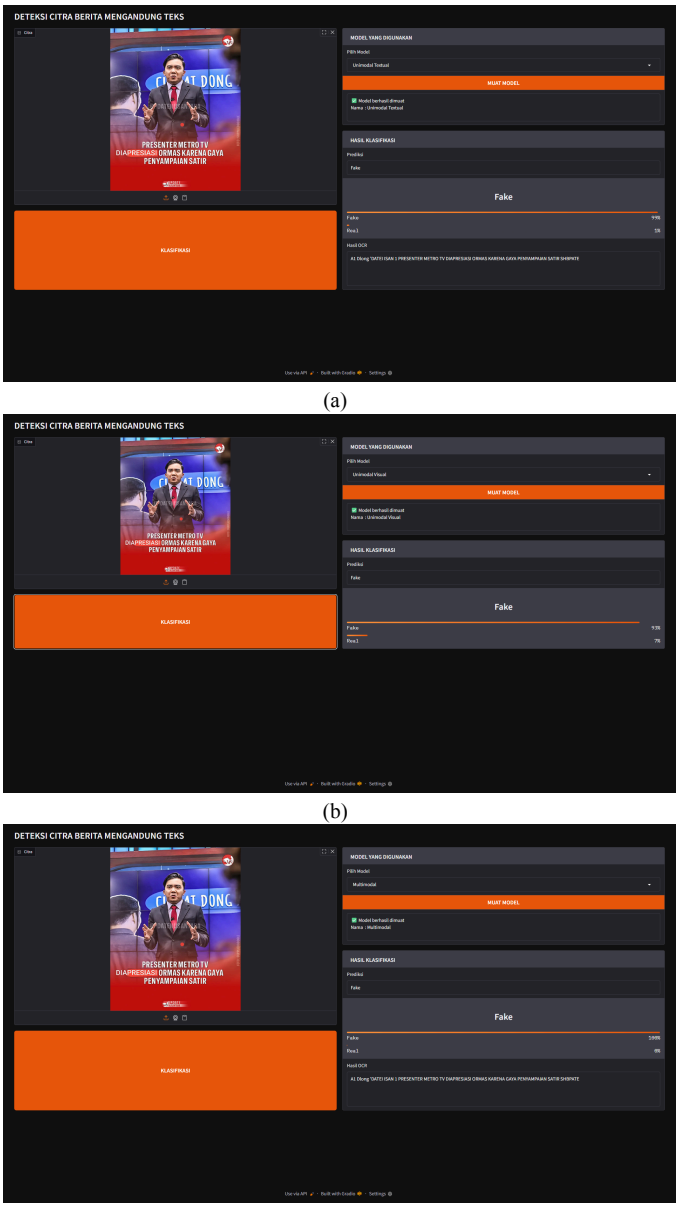


Fig. 18. Graphical interface testing results on the same test sample using (a) the unimodal textual mode, (b) the unimodal visual mode, and (c) the multimodal mode.

## V. CONCLUSIONS

This study presented a multimodal deep learning system for detecting text overwriting manipulation in Indonesian-language news images. A dedicated dataset of 2,400 images was constructed, comprising 1,200 authentic images collected from official Indonesian news social media accounts and 1,200 manipulated images, consisting of 600 real-world samples collected from social media and 600 manually manipulated samples created using Figma. Three classification approaches were developed and evaluated, namely a unimodal visual model based on EfficientNetV2, a

unimodal textual model based on IndoBERT applied to EasyOCR-extracted text, and a multimodal model that integrates both modalities through a three-way scaled dot-product attention fusion module.

Experimental results showed that the unimodal visual and textual components achieved validation accuracies of 0.8850 and 0.6650, respectively, while the multimodal approach achieved the best test-set performance overall, with an accuracy of 0.8950, an F1-score of 0.8923, a precision of 0.9158, and a recall of 0.8700, surpassing both unimodal approaches and the traditional image processing baseline. Although the unimodal visual model performed slightly better on the training and validation sets, the multimodal model generalized more reliably to unseen test data, confirming that integrating visual and textual features produces a richer and more robust representation for detecting text overwriting manipulation. A functional graphical interface was further developed to support all three inference modes, demonstrating the practical applicability of the proposed system.

Despite these results, several limitations remain. The dataset size, while sufficient for the intended scope, does not yet capture the full range of visual styles used by other Indonesian news publishers, and residual OCR noise, such as irrelevant account names, continues to affect the quality of the extracted text. The unimodal visual model exhibited signs of overfitting, and the unimodal textual model achieved comparatively low performance, indicating room for improvement in both modalities individually. The system also currently operates only through a locally-hosted interface without integration into social media platforms, limiting its direct applicability for large-scale misinformation prevention. Future work may address these limitations through larger and more diverse datasets, comparison with alternative detection strategies such as noise inconsistency analysis or vision-language models, exploration of other pretrained backbones and fusion mechanisms, extension toward manipulation localization rather than binary classification, further hyperparameter optimization, and integration of the system into real-world social media environments for direct deployment.

## REFERENCES


[1] Abdulqader, Fakhrulddin., M., Dawod, A. Y., & Zeki Ablahd, A. 2023. "Detection of tamper forgery image in security digital mage". Measurement: Sensors, 27, 100746. https://doi.org/10.1016/j.measen.2023.100746

[2] waan. (2026). Deteksi timpa teks pada gambar menggunakan metode sederhana pengolahan citra. Medium. https://medium.com/@ikhwaanp20/deteksi-timpa-teks-pada-gambar-menggunakan-metode-sederhana-pengolahan-citra-a5faa696dc87

[3] Sudiatmika, I. B. K., Rahman, F., Trisno, T., & Suyoto, S. (2019). Image forgery detection using error level analysis and deep learning. TELKOMNIKA (Telecommunication Computing Electronics and Control), 17(2), 653. https://doi.org/10.12928/telkomnika.v17i2.8976

[4] Korsipati, J. R., Yanamala, R. M. R., Pallakonda, A., Raj, R. D. A., & Prakasha, K. K. (2025). Multi-resolution transfer learning for tampered image classification using SE-enhanced fused-MBConv and optimized CNN heads. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-17799-0

[5] Joshi, R., Gupta, A., Kanvinde, N., & Ghonge, P. (2022). Forged Image Detection using SOTA Image Classification Deep Learning Methods for Image Forensics with Error Level Analysis. In 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1–6). IEEE. 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT). https://doi.org/10.1109/icccnt54827.2022.9984489

[6] Ramzan, A., Ali, R. H., Ali, N., & Khan, A. (2024). Enhancing Fake News Detection Using BERT: A Comparative Analysis of Logistic Regression, RFC, LSTM and BERT. In 2024 International Conference on IT and Industrial Technologies (ICIT) (pp. 1–6). IEEE. 2024 International Conference on IT and Industrial Technologies (ICIT). https://doi.org/10.1109/icit63607.2024.10859673

[7] Rijal, M., Musa, H., Seli, F. Y., Asnawi, N. I., & Firman, A. M. (2025). Fake news detection in Indonesian language using a deep learning approach with Indo-BERT. Proceeding of the 5th International Conference on Social Sciences and Education (ICSSE 2025). https://proceeding.uns.ac.id/icsse/article/view/1025

[8] Tuan, N. M. D., & Minh, P. Q. N. (2021). Multimodal Fusion with BERT and Attention Mechanism for Fake News Detection (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2104.11476