Intechno Journal:
Information Technology Journal
Vol. 8 No. 1 July 2026
Homepage: https://jurnal.amikom.ac.id/index.php/intechno
INTECHNO Journal
INFORMATION TECHNOLOGY JOURNAL
e-ISSN: 2655-1438
p-ISSN: 2655-1632

.

Deepfake Content Analysis Using Error Level Analysis and Metadata

Kurnia Nur Hikmah[1], Ghufron Zaida Muflih[2]
[1,2] Universitas Ma'arif Nahdlatul Ulama, Jalan Kutoarjo Km 5 Jatisaari, Kebumen 54317, Indonesia

Article Info
Article history:

Received: 4 July 2026
Revised: 15 July 2026
Accepted: 20 July 2026
Published: 31 July 2026

Keywords:

Deepfake
Metadata
Error level Analysis
Reality Defender
Digital Forensics
Platform X

ABSTRACT

Purpose: The spread of deepfakes on platform X threatens the integrity of digital information, but previous research has applied Metadata Analysis, Error Level Analysis (ELA), and Reality Defender separately, not yet as an integrated approach. This study applied Metadata Analysis and ELA to identify manipulation in deepfake images, as well as evaluate the effectiveness of Reality Defender in detecting deepfake content in images circulating on platform X as an application case study.
Methods/Study design/approach: Using a descriptive qualitative approach with digital forensic experiment methods, seven purposive selected image samples representing seven content characteristic scenarios (original, face-swap, GAN, full generative AI, anti-forensics, conventionally edited, and platform compressed), analyzed through three layered stages with tiered final classification criteria based on the cross-validated Reality Defender score threshold with ELA.
Result/Findings: Metadata is only informative before uploading, as X deletes EXIF uniformly post-upload. ELA remained effective in both conditions, showing localized intensity anomalies (close to 240-255 from 255) across deepfake samples. Reality Defender correctly classified six of the seven samples (85.7% accuracy, 100% in the deepfake category, 66.7% in the original category), with one false positive (64%) in the original, conventionally edited image.
Novelty/Originality/Value: This study integrated all three methods simultaneously and layered with explicit criteria, showing Metadata lost post-upload diagnostic value while ELA and Reality Defender remained the most reliable for content sourced from platform X.

Corresponding Author:

Kurnia Nur Hikmah
Email: kurnianurhk@gmail.com

1. INTRODUCTION

The rapid development of AI, machine learning, and deep learning makes it easier to manipulate multimedia, which has the potential to be misused to spread false information, trigger political disputes, and extortion and harassment [1]. One of the most alarming impacts is the ease with which deepfake content can manipulate a person's face, voice, or body movements to make them look realistic [2]. The term deepfake comes from a combination of the words "Deep learning" and "fake" [3], generated using the Generative

Adversarial Networks (GAN) method [4]. The term first appeared in 2017 when a Reddit user named "deepfakes" shared a pornographic video of a face swap [5].

Platform X is the main means of spreading deepfake content due to its viral speed as a real-time communication medium [6] [7], with Indonesia ranking third with 25.2 million active users, [8]. This research focuses on images because the format is the most dominant and easier to create and distribute than videos [9]. The lack of technical knowledge of ordinary people makes it difficult for them to distinguish between real images and AI manipulations, while the threat of digital disinformation to public opinion and political stability continues to increase.

Digital forensics approaches are relevant to identifying deepfake content through digital trace checks, such as compression inconsistencies, file structure anomalies, or metadata inconsistencies [10] [11]. Metadata analysis is an important initial stage because it is able to provide information such as file names, creators, creation dates, and source devices [12]. Error level analysis (ELA) detects manipulation through error level differences, calculating the difference between luminance and chrominance values between image areas [13] [3]. Meanwhile, Reality Defender is a multimodal deepfake detection platform for text, photos, videos, and audio [14] [15].

The effectiveness of ELA and Metadata Analysis has been proven by various previous studies Rafique, et al. (2023) the combination of ELA with CNN was able to achieve an accuracy of 89.5% [2]. In the study of Ahmed, et al. (2026) ELA with EfficientNetB0 resulted in an accuracy of 96.37% in the GAN dataset [16]. Bisri & Marzuki (2023) the combination of ELA and Exif Metadata resulted in 94.6% and 98.3% respectively [17]. Syafar, et al. (2025) show that Metadata is effective in identifying changes in information in images, while ELA is effective in detecting manipulation areas based on contrast differences [13]. However, these methods are generally still applied separately, there has been no research that integrates Metadata Analysis, ELA, and Reality Defender simultaneously on image content sourced from the X platform.

Based on the description above, this study aims to apply Metadata Analysis and Error level Analysis (ELA) methods to identify manipulations in deepfake image content, as well as evaluate the effectiveness of Reality Defender Tools in detecting deepfake content in images circulating on platform X as a case study of the application of an integrated forensic approach. This research is expected to support the development of effective digital forensics methods in dealing with the threat of deepfake-based disinformation in the era of generative artificial intelligence.

2. METHOD

2.1 Types of Research

This study uses a descriptive qualitative approach with digital forensic experiment methods, focusing on systematically depicting the characteristics and anomalous patterns of objects [11]. Each analysis instrument is applied sequentially and layered to the same object, so that the results of each stage can be compared and validated with each other. Platform X was chosen as an application case study because of its characteristics as a real-time communication medium with a large number of active users in Indonesia [6].

2.2 Research Flow

This research was carried out through six main stages in succession. The research stage begins with the collection of image datasets, then the image data is analyzed with metadata using ExifTool, then the Error level Analysis stage using Forensically Beta and verified using Reality Defender. The data from the analysis of the three methods were analyzed based on the final classification criteria that had been set. The final stage of the research is the aggregate evaluation and drawing conclusions about the effectiveness of the integration of the three methods, as illustrated in Figure 1.

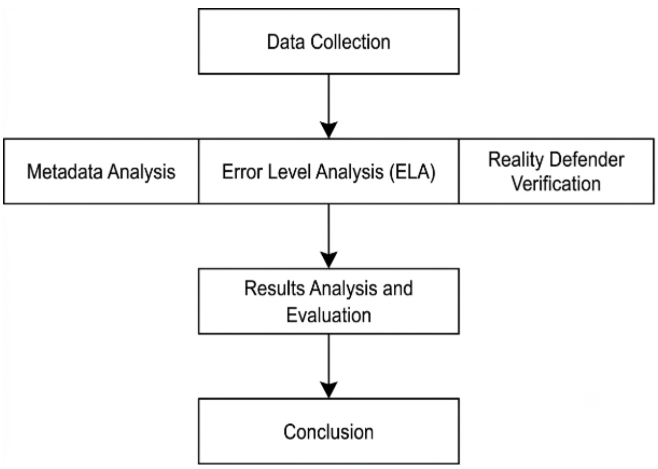

Figure 1. Research Flow Diagram

2.3 Data Collection

The dataset of this study consists of a total of seven image samples, which were purposively selected to represent seven typical scenarios of digital content found on platform X, both original and manipulated using artificial intelligence technology, as summarized in Table 1. Each scenario is represented by a single sample image designed to highlight the specific technique characteristics of the related manipulation category, so this design is an exploratory qualitative multi-case forensic study rather than a large-scale quantitative study. The results obtained were descriptive illustrative of the forensic patterns in each category of manipulation, not estimates of statistical accuracy in large populations.

Table 1. Research scenario

Id.	Scenario	Test Object	Expected Output
1	Original images without manipulation	Original image of platform X	Normal metadata, stable ELA, Reality Defender: authentic
2	Face-swapping deepfake image	Face-swapping deepfake image	Metadata: traces of editing software, ELA: facial area inconsistencies, Reality Defender: deepfake/manipulated
3	Deepfake image of GAN results	Synthetic image of GAN results	Metadata: none or anomaly, ELA: uneven bright areas, Reality Defender: deepfake/manipulated
4	Generative AI-generated fully generated images	Images created entirely by generative AI	Metadata: no data, ELA: unnatural compression pattern, Reality Defender: manipulated
5	Incomplete metadata deepfake images (anti-forensic)	Deepfake images whose EXIF Metadata is edited or deleted	Metadata: inconsistency (inconsistency between creation timestamps, software tags), ELA: not affected by Metadata editing, Reality Defender: manipulated
6	Original images are lightly edited without AI	Original image with cropping/resize/filter	Metadata: traces of editing software, ELA: mild anomalies, Reality Defender: authentic
7	Low-resolution original image (X compression)	Original image automatically compressed by platform X	Metadata: partial changes due to compression, ELA: consistent, Reality Defender: authentic

The seven samples include:

a. S(a) Original images without manipulation. It is an authentic control image or ground truth in the form of a selfie of the Samsung Galaxy A15 camera in lighting conditions.
b. S(b) Deepfake face-swapping images. It is the result of visual manipulation using the face-swapping technique. In this image, the face of the original subject from the first sample is pasted onto the body of the subject in a baseball uniform.
c. S(c) Deepfake image of GAN (Generative Adversarial Network). It is an image that shows a portrait of a curly-haired woman in a red shirt with a white collar on a blue background, characteristic of an airbrushed look.
d. S(d) Images generated by Generative AI. It is a full-body image of an AI prompt, with high rendering but has a tendency to distort text elements, finger details, or small details in the background of the room.
e. S(e) Incomplete Metadata deepfake images (Anti-Forensic). Identical to the GAN result image in the S(c) sample but by removing or reducing the Metadata information structure.
f. S(f) The original image is lightly edited. It is a modification of the original S(a) image edited with Adobe Lightroom (cropping, contrast, brightness, filter).

.

g. S(g) Low-resolution original images. Using the original image object S(a) but uploaded and downloaded multiple times from X, it suffers from quality degradation.

All images used as research samples are the researcher's own and are taken, manipulated using deepfake techniques, and published with the full consent and knowledge of the researcher as the subject. The use and publication of images in this study is solely for academic purposes. The entire sample is presented in Figure 2.

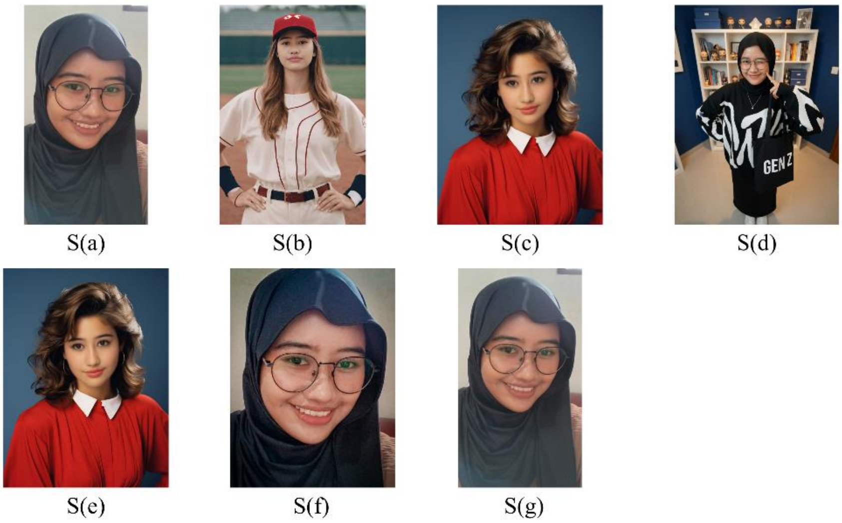

Figure 2. Sample Images

2.4 Stages of Analysis

Each image sample is analyzed through three sequential, layered stages on the same object.

a. Metadata Analysis

Metadata Analysis uses ExifTool to extract technical information stored in EXIF data (format, dimensions, source camera, resolution, software tag) [18]. The analysis was performed on two conditions, namely before the image is uploaded to platform X and after the image is re-downloaded, to illustrate the extent to which platform X affects the metadata structure of the image. The extraction results are categorized into two conditions, namely complete metadata or Incomplete/anomalous metadata as an initial indicator of the authenticity of the image.

b. Error level analysis (ELA)

The application of the Error level Analysis (ELA) method uses Forensically Beta Tools with consistent parameters (JPEG Quality 50, Error Scale 50, Magnifier Enhancement None, and Opacity 0.95) to identify image compression inconsistencies. The results are categorized descriptively into two conditions, namely Consistent error level distribution and Inconsistent/anomalous error level distribution as supporting indicators. Based on the estimated RGB pixel intensity on a scale of 0-255 [19]

c. Reality Defender Verification

Each image sample is verified using Reality Defender, classified as authentic or manipulated with a probability score, then compared to ground truth. Reality Defender was chosen as the main verification instrument by considering three reasons, namely its accessibility as a web-based platform that does not require installation or special computing infrastructure, and its multimodal detection coverage includes text, photos, video, and audio in one integrated system [14], as well as its rapidly growing adoption in the industrial sector as an enterprise-level deepfake detection solution [15]. Compared to similar tools like Sensity AI or Hive Moderation that are generally paid or require custom API integration, Reality Defender offers a more open testing interface for small- to medium-scale academic research needs.

2.5 Final Classification Criteria

To integrate the results of the three methods into one final interpretation per sample, this study establishes a tiered classification criterion that refers to the triangulation principle of multi-layer forensic methods, where the combination of mutually independent methods will increase the validity of conclusions compared to the use of a single method [10] [17]. The criteria used are as follows:

a. If the Reality Defender manipulation probability score is ≥ 80%, the sample is classified as Manipulated with a high confidence level without the need for additional validation.
b. If the Reality Defender score < 80%, the results are cross-validated with the ELA. If the ELA shows a localized intensity anomaly with a difference of more than 100 points between areas (scale 0-255), the

sample is classified as Manipulated. If the ELA shows a consistent and even distribution of error levels with no localized anomalies, the Reality Defender classification is corrected to Authentic.

c. The results of the Metadata Analysis on pre-upload conditions are used as additional supporting evidence (not the primary determinant criterion).

3. RESULTS AND DISCUSSIONS

3.1 Metadata Analysis

Metadata analysis is performed on two conditions, namely before the image is uploaded to platform X and after the image is re-downloaded. This division is done because X applies compression and automatic standardization to uploaded images, which has the potential to change or remove most of the original Metadata. By comparing the two conditions, this study can describe the extent to which platform X affects the structure of image metadata as well as its impact on the process of forensic deepfake identification.

a. Metadata analysis before uploading to platform X

The results of metadata analysis have several important patterns. Sample S(a) as the original image has the most complete metadata with all the technical attributes of the camera being consistently met. In contrast, the samples of S(b), S(c), and S(d) as deepfakes show the absence of almost all of the camera's EXIF data, even S(c) explicitly listing "AI generated" in the Make column. The S(e) sample, as an anti-forensic effort, has metadata that appears to be from the camera but contains significant inconsistencies between attributes. The S(f) sample has the richest Metadata, but it leaves a clear trail of Adobe Lightroom usage. Meanwhile, the minimal metadata on the S(g) sample is more due to platform compression rather than manipulation. Based on the results of the ExifTool analysis, all analyses are summarized in Table 2.

Table 2. Comparison of Metadata attributes before upload

Metadata Attributes	Original S(a)	S(b) Face-swapping	S(c) GAN	S(d) Generative AI	S(e) Anti-Forensic	S(f) Lightly Edited	S(g) Compression
File format	JPEG	JPEG	JPEG	JPEG	JPEG	JPEG	JPEG
File size	1746 kB	258 kB	256 kB	233 kB	256 k	2.7 kB	83 kB
Dimensions	3712×2088 px	1024×1536 px	928×1232 px	864×1152 px	928×1232 px	2088×2784 px	675×1200 px
Process encoding	Baseline DCT, Huffman	Baseline DCT, Huffman	Progressive DCT, Huffman	Baseline DCT, Huffman	Progressive DCT, Huffman	Baseline DCT, Huffman	Progressive DCT, Huffman
Camera brand (Make)	Samsung	None	AI generated	None	Samsung	Samsung	None
Camera models	Galaxy A15	None	None	None	Samsung	Galaxy A15	None
Software	A155FXXS5BXJ3 (Samsung firmware)	None	None	None	Galaxy a15	Adobe Lightroom	None
Metadata Conditions	Normal/ Complete	Anomalies (without EXIF)	Anomaly (Death: AI generated)	Anomalies (without EXIF)	Anomalies (False/inserted metadata)	Normal, Trace Editing	Minimal (X compression)

b. Metadata analysis after uploading to platform X

The next stage of testing analyzes the image after it has gone through the upload and re-download process on platform X, to determine the resistance of the EXIF Metadata to third-party systems and the effect of the

.

platform's automatic compression on publicly distributed files. The results of the analysis show a uniform and consistent pattern of change, as presented in Table 3.

Platform X automatically deletes almost all of the image's original metadata and renames the file with a random, alphanumeric formatted name. Most EXIF information, such as camera brand, device model, shooting parameters, and image processing software, is no longer available, so the Metadata structure between samples becomes almost uniform.

The change causes post-upload metadata to have limited ability to distinguish between real images and deepfake images. Of the seven samples, all lost the camera EXIF data and software tags without exception, so that in aggregate 0% of the samples were still identifiable through the post-upload attribute. Thus, metadata after an image circulates on platform X is more useful for describing changes in file structure than as a key indicator in deepfake detection.

Table 3. Comparison of Metadata attributes after upload

Metadata Attributes	Original S(a)	S(b) Face-swapping	S(c) GAN	S(d) Generative AI	S(e) Anti-Forensic	S(f) Lightly Edited	S(g) Compression
File size	223kB	219 kB	256 kB	217 kB	256 kB	1150 kB	83 kB
Dimensions	1152×2048 px	1024×1536 px	928×1232 px	864×1152 px	928×1232 px	2088×2784 px	675×1200 px
Process encoding	Progressive DCT	Progressive DCT	Progressive DCT	Progressive DCT	Progressive DCT	Progressive DCT	Progressive DCT, Huffman
Camera brands	Lost	Lost	Lost	Lost	Lost	Lost	None
Software Tag	Lost	Lost	Lost	Lost	Lost	Lost	None
Metadata Conditions	Compressed X Platform	Not significantly changed	Not significantly changed	Not significantly changed	Not significantly changed	Compressed X Platform	Compressed X Platform

3.2 Error Level Analysis (ELA)

Error Level Analysis (ELA) was performed using Forensically with the same parameters across all samples, namely JPEG Quality 50, Error Scale 50, Magnifier Enhancement None, and Opacity 0.95, as shown in Figure 3. The analysis was performed based on the error level distribution and the difference in RGB pixel intensity to identify image compression inconsistencies.

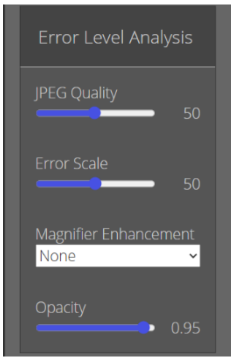

Figure 3. Configure ELA Forensically

The interpretation of the ELA results is based on two main indicators, namely the distribution of brightness in Error Level Map and the level of contrast between areas. On an RGB pixel intensity scale of 0-255, an area with a value close to 0 (dark) indicates a level of Error High compression [19]. The results of ELA are seen in Figure 4.

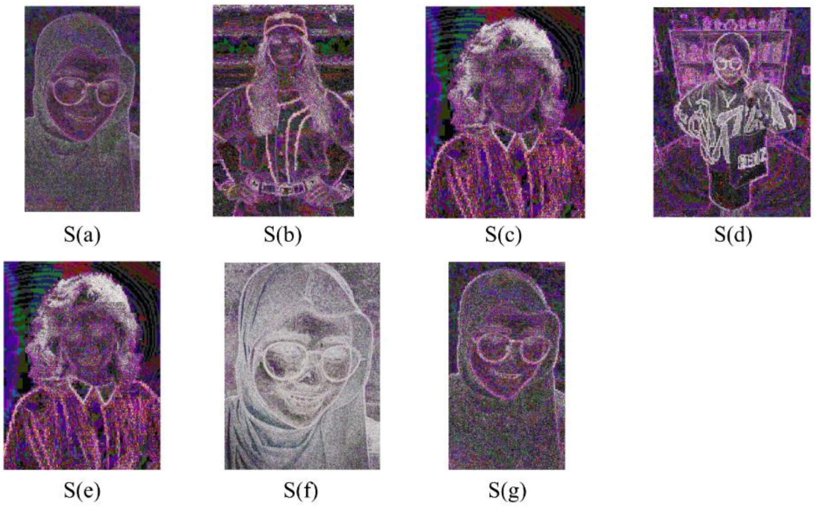

Figure 4. Error Level Analysis (ELA) Results

a. S(a) Original image without manipulation

The ELA results in the S(a) sample showed a relatively consistent and even distribution of error levels. Based on the estimated RGB pixels, this sample has an average error intensity level of about 95 on a scale of 0-255, with a standard deviation of about 40-44 points on each color channel. No blocks of areas were found with a contrast that differed much from the surroundings, either in the face, veil, or background areas. Thus, the S(a) sample is categorized as having a consistent error-level distribution, in accordance with the theoretical expectation that images taken directly by the camera have a uniform single compression history throughout its sections.

b. S(b) Face-swapping deepfake images

The results of the ELA in the S(b) sample showed a different pattern. The estimated RGB intensity in the hair and jacket area reached 245-255 (solid white), while the facial area was in the lower range, which was around 90-110, the difference was about 140-160 points, with the standard deviation of the overall sample also higher (around 67-70). This extreme contrast pattern in the hair, hat, and jacket area strongly indicates the insertion of elements from different image sources, consistent with the face-swapping technique, where a new face is inserted into the base photo so that the surrounding area has a different level of compression than the swapped face area. Thus, the S(b) sample was categorized as having an inconsistent error-level distribution, with high-contrast anomalies in the hair and jacket areas.

c. S(c) Deepfake image of GAN results

The ELA results in the S(c) sample showed an extreme degree of anomaly. The estimated RGB intensity of the hair area is close to 255 (pure white), while the background shows a concentric line pattern (banding pattern) with an unnatural intensity varying between 70-130. The overall standard deviation of the sample reached about 67, confirming a very high level of error level inconsistency. The concentric banding pattern on the background is a typical artifact of the generative AI process, where the GAN algorithm synthetically generates background textures through a gradual upsampling process, leaving traces of repetitive patterns that do not appear in the camera photos. Thus, the S(c) sample was categorized as having an error-level distribution that was highly inconsistent with the typical generative artifact pattern of GAN, providing strong visual forensic evidence against generative AI manipulation.

d. S(d) Generative AI-generated images

The results of the ELA in the S(d) sample showed a more widespread pattern of anomalies. The estimated RGB intensity at almost all edges of the object is in the range of 240-255, much higher than other areas. The background of the bookshelf shows an intensity of about 100-150 with very high texture detail, each book on the shelf is visualized with individual bright edges characteristic that are not prevalent in the original photo, since objects that are far from the camera's focus should have a lower level of detail and compression. The standard deviation of the overall sample ranged from 64-68. Thus, the S(d) sample is categorized as having an inconsistent error-level distribution, with anomalies evenly distributed across the edges of the object, consistent with the image characteristics that are fully generated by modern generative AI.

e. S(e) Incomplete metadata deepfake image (Anti-Forensic)

The ELA results in the S(e) sample showed a pattern very similar to the findings in the S(c) sample (RGB hair area 250-255, background 65-125), with the overall standard deviation of the sample ranging from 63-66. These findings prove empirically that ELA works based on a direct analysis of image pixel content, so it is not affected by Metadata footprint removal efforts. Thus, the S(e) sample is categorized as having an inconsistent error-level distribution, with a pattern very similar to the GAN deepfake sample, confirming the existence of manipulation even though the Metadata trace has been removed.

f. S(f) Original image lightly edited

The ELA results in the S(f) sample showed a map level error that appeared much lighter and flatter, with a color distribution dominated by light gray shades evenly across the frame. The estimated RGB intensity in this sample shows a relatively high average, which is about 150 on a scale of 0-255 in almost all image areas, but with a lower standard deviation, which is around 42-45. This pattern is in line with the parametric editing process (global adjustment of sharpness, clarity, and contrast), which increases edge contrast evenly throughout the image. Thus, the S(f) sample is categorized as having a relatively consistent error-level distribution, but with increased global brightness due to sharpening, a pattern that is different from deepfakes and can be attributed to conventional editing processes, rather than AI-based manipulation.

g. S(g) Low-resolution original images

The ELA results in the S(g) sample showed a similar pattern to the S(a) sample, which is a relatively dense and evenly distributed noise throughout the frame without any areas of extreme protrusion. The estimated RGB intensity in this sample showed an average of about 100-110 on a scale of 0-255, with a standard deviation of about 64-66. This pattern of increased but still evenly distributed error levels confirms that repeated compression by social media platforms can increase the noise level of the ELA, without producing the localized anomalous patterns that characterize deepfake manipulation. Thus, the S(g) sample is categorized as having a consistent error level distribution even with increased global noise due to the platform's repeated compression, so that it can still be distinguished from the anomalous pattern of deepfakes that are localized.

In aggregate, all four deepfake samples (S(b), S(c), S(d), S(e)) all showed localized anomalies with intensities close to 240-255, while all three original samples (S(a), S(f), S(g)) all showed a consistent distribution with no localized anomalies, thus achieving a very clear pattern separation rate (100% of deepfake samples showed local anomalies vs 0% in the original sample). The comparison of ELA results is shown in table 4.

Table 4. Comparison of ELA results

Sample Code	Distribution Error level	RGB Intensity Estimation (0-255 scale)	Location of Anomalous Areas	Categories
S(a)	Consistent & evenly distributed throughout the frame	Medium-high range: ±90-140 (average 95), deviation 40-44	There are no areas that stand out significantly from the surroundings	Consistent (Original)
S(b)	Inconsistent, the hair & jacket area is very prominent	Height range: peak close to 255 in the hair & accessories area, average face area 100, deviation 67-70	Hair, jacket and neck accessories (face-swap insert area)	Anomalies (manipulation)
S(c)	Highly inconsistent, extreme contrast in hair	Very high range: hair close to 255 (solid white), concentric striped background 70-130, deviation 67	Hair and facial contours; unnatural ring patterned background	Anomalies (manipulation)
S(d)	Inconsistent, very bright edges of objects evenly	High range: clothing & book edges close to 240-255, bookshelf background 100-150, deviation 64-68	The entire edges of the body, arms, and book objects; Shelf Background Texture	Anomalies (manipulation)
S(e)	Inconsistent, very similar to a GAN sample (S(c))	Very high range: hairs close to 255, deviation 63-66	Hair and facial edges; concentric ring patterned background	Anomalies (manipulation)

S(f)	Relatively consistent yet brighter & even (flat)	Medium range: average 150 across frames (brighter than S(a)), deviation 42-45 (lower/homogeneous)	Slight protrusion on the edges of the veil and glasses due to sharpening	Consistent with light editing trails (not deepfakes)
S(g)	Consistent, dense noise patterns resemble the original image	Medium-high range: 100 average, 64-66 deviation (noise evenly distributed due to recompression)	There are no significant protruding areas; noise is evenly distributed due to repetitive JPEG compression	Consistent (Original, compressively affected)

3.3 Reality Defender Verification

This stage is carried out using Reality Defender as an AI-based deepfake detection system, where each image sample is classified into the Authentic or Manipulated category, accompanied by a probability value as a measure of the system's level of confidence in the detection results.

The test results showed that six of the seven samples were successfully classified according to ground truth. All deepfake images are recognized as Manipulated with 93%, 96%, and 99% confidence rates, while genuine images are classified as Authentic with a low probability of manipulation.

The only discrepancy was found in the sample of the original image undergoing light editing (S(f)). Reality Defender classified the image as Manipulated with a confidence score of 64%, resulting in a false positive. These findings suggest that conventional digital editing can still affect AI-based classifications, so Reality Defender's results need to be confirmed using other forensic methods. The results of the Reality Defender verification are shown in table 5.

Table 5. Reality Defender verification results

Sample Code	Image Type	Ground truth	Reality Defender Results	Probability Score	Status
S(a)	Original Image	Original	Authentic	1%	True
S(b)	Face-swapping	Manipulation	Manipulated	96%	True
S(c)	GAN Deepfake	Manipulation	Manipulated	93%	True
S(d)	Generative AI	Manipulation	Manipulated	99%	True
S(e)	Deepfakes Without Metadata	Manipulation	Manipulated	93%	True
S(f)	Original Images edited	Original	Manipulated	64%	False
S(g)	Compressed Original Image	Original	Authentic	2%	True

In aggregate, Reality Defender managed to classify six of the seven samples precisely, resulting in an overall accuracy of 85.7%. Based on the ground truth category, the accuracy in the deepfake image category reached 100% (4 out of 4 samples were true, with a score of 93-99%), while the accuracy in the original image category reached 66.7% (2 out of 3 samples were true). Details of the accuracy per category are presented in Table 6.

Table 6. Reality Defender Accuracy by Category

Categories	Number of Samples	True	False	Accuracy
Original Image (Ground Truth: Original)	3	2 (S(a), S(g))	1 (S(f))	66,7%

.

Deepfake Images (Ground Truth: Manipulation)	4	4 (S(b), S(c), S(d), S(e))	0	100%
Overall Total	7	6	1	85,7%

The only discrepancy was found in the original image sample that underwent light editing (S(f)), which was classified as Manipulated with a score of 64%, resulting in a false positive. Based on the final classification criteria, since the S(f) score was below the 80% threshold, this result was cross-validated with the ELA. The ELA results on S(f) show a consistent distribution without localized anomalies (the second criterion), so the final classification is corrected to Authentic, according to the actual ground truth.

3.4 Analysis of results and evaluation

Table 7 presents the results of Metadata Analysis after upload to X, Error level Analysis (ELA), and Reality Defender verification side by side for each sample, along with the final category after the application of the classification criteria.

Table 7. Synthetic results of Metadata Analysis, ELA, Reality Defender

Sample Code	Metadata Results	ELA Results	Reality Defender	Final Category
S(a)	223 kB, 1152x2048 px, Progressive DCT. EXIF is completely lost after upload. Unable to distinguish samples.	Consistent evenly. Average 95, std 40-44. Homogeneous noise, no localized spikes	Authentic 1%	ORIGINAL (ELA & RD consistent)
S(b)	219 kB, 1024x1536 px, Progressive DCT. EXIF is lost, the file name is renamed system X.	High anomalies. Hair/jacket is close to 255 (solid white). Faces 90-110. The difference is 140-160 points. Std 67-70.	Manipulated 96%	DEEPFAKE (ELA & RD are consistently strong)
S(c)	256 kB, 928x1232 px, Progressive DCT. EXIF is missing.	Anomalies are very high. Hair is close to 255. Concentric ring patterned background 70-130 (typical GAN artifact). Std 67.	Manipulated 93%	DEEPFAKE (ELA & RD reinforce each other)
S(d)	217 kB, 864x1152 px, Progressive DCT. EXIF is missing,	Widespread anomalies. The entire edge of the object is 240-255. Bookshelf background of 100-150 unnatural details. Std 64-68.	Manipulated 99%	DEEPFAKE (supreme confidence)
S(e)	256 kB, 928x1232 px, Progressive DCT. EXIF is completely lost. Fake metadata is also removed, the end result is identical to S(c).	High anomalies. Identical pattern S(c). Hair 250-255, concentric background 65-125. Std 63-66.	Manipulated 93%	DEEPFAKE (anti-forensics is ineffective)
S(f)	1150 kB, 2088x2784 px, Progressive DCT. EXIF and XMP Adobe Lightroom are missing. The file size & dimensions are larger than other samples, indicating a high-quality/high-resolution image source.	Relatively consistent. Average 150 (brighter due to global sharpening). Std 42-45 (homogeneous). There are no localized anomalies.	Manipulated 64% (False-Positive)	ORIGINAL (ELA False-Positive RD correction)

S(g)	83 kB, 675x1200 px, Progressive DCT. The Metadata condition does not change EXIF is missing.	Consistent. Averages 100-110, std 64-66. Noise increases evenly due to repeated compression, no localized spikes.	Authentic 2%	ORIGINAL (ELA & RD consistent)

Based on Table 7, ELA and Reality Defender gave mutually consistent results in six of the seven samples (85.7%), without the need for additional corrections. In one sample (S(f)), the application of the tiered classification criteria successfully corrected the false positive Reality Defender using ELA proof, so that the final accuracy of the integrated classification (Metadata + ELA + Reality Defender with tiered criteria) reached 100% (7 out of 7 samples according to ground truth), higher than the accuracy of a single Reality Defender (85.7%). These findings show the added value of a multi-method integration approach compared to using a single method. As applied to the reference study, this study is in line with the accuracy range reported in the literature.

3.5 Discussions

The results show that the Metadata has limited diagnostic value because the X platform removes most of the EXIF information so that the original image and the manipulated image have an almost uniform post-upload Metadata structure. In contrast, ELA and Reality Defender provided consistent results in six of the seven samples. ELA was able to identify anomalous patterns in deepfake images through error level distribution, while Reality Defender was able to classify manipulated images with high confidence levels (93%, 96%, and 99%). The difference in results was only found in the S(f) sample, which is the original image that was lightly edited. ELA shows a pattern that is still consistent with the conventional editing process, while Reality Defender produces false positives.

The findings suggest that the use of Reality Defender as a supporting tool should be combined with Metadata and ELA analysis through explicit classification criteria to make the deepfake identification process more accurate and reduce the potential for misclassification. Metadata analysis has limited diagnostic value when applied to post-upload content, as platform X systematically removes almost all traces of metadata, both original and anti-forensic engineering. This condition makes Error Level Analysis (ELA) and Reality Defender the two most relevant and reliable primary methods for identifying deepfake content that has been circulating on platform X, as both work on the analysis of the visual content of the image directly and do not depend on the presence of file metadata.

4. CONCLUSION

This study concluded that Metadata Analysis is not effective in identifying deepfake images that have been uploaded to platform X, because the platform deletes almost all EXIF data automatically so that the real and deepfake images become indistinguishable based on Metadata alone. In contrast, Error level Analysis (ELA) remains effective because it works directly on the pixels of the image, including on images whose metadata is deliberately deleted, with a clear pattern separation rate between the original sample and the deepfake across the entire sample being tested. The results of the evaluation showed that Reality Defender was effective in detecting deepfakes with an overall accuracy of 85.7% (100% in the deepfake category, 66.7% in the original category) and was not affected by compression or deletion of Metadata, although it had incorrectly generated one false positive in the original image edited as Manipulated. The application of a tiered classification criterion that integrates ELA as cross-validation successfully corrected these errors, increasing the accuracy of the final classification to 100% in this study dataset. Thus, to identify deepfakes in content already circulating on platform X, ELA and Reality Defender are the main methods that are more relied upon, with Metadata acting as additional supporting evidence when available in pre-upload conditions.

Given the limitations of this study using seven representative samples without validation on the public benchmark dataset, it is recommended that further research use a larger and more diverse sample in each scenario category, validate this approach against public reference datasets such as FaceForensics++ or Celeb-DF, test this method on other social media platforms, and refine the Reality Defender classification threshold based on broader empirical data to avoid errors Detection on the original image can be further minimized.

.

REFERENCES

[1] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022, doi: 10.1109/ACCESS.2022.3154404.

[2] R. Rafique, R. Gantassi, R. Amin, J. Frnda, A. Mustapha, and A. H. Alshehri, "Deep fake detection and classification using error-level analysis and deep learning," *Sci. Rep.*, vol. 13, no. 1, pp. 1–13, 2023, doi: 10.1038/s41598-023-34629-3.

[3] Y. A. Fernandes and Y. Fatma, "Metode Deep Learning Dalam Teknologi Deepfake : Systematic Literature Review," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 2, pp. 3403–3410, 2025, doi: 10.36040/jati.v9i2.12987.

[4] D. Waringgi and S. Setiasih, "Business Competition Management in the Deepfake and Synthetic Media Era: Corporate Identity Verification as a Strategy to Mitigate Market Disinformation," *J. Ekon. Balanc.*, vol. 21, no. 2, pp. 296–307, Dec. 2025, doi: 10.26618/FT7SDD96.

[5] Fortinet, "What Is Deepfake_ AI Endangering Your Cybersecurity_ Fortinet," 2025, Accessed: Apr. 19, 2026. [Online]. Available: https://www.fortinet.com/resources/cyberglossary/deepfake

[6] N. Halimah and F. Fachri, "Analisis Bukti Digital Twitter Web Menggunakan Live Forensic Pada Kasus Cyberbullying," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 1, pp. 1005–1010, 2024, doi: 10.36040/jati.v9i1.12554.

[7] L. D. V. Anggreni, I. G. A. A. M. Aristamy, K. Laksmi, Maswari, M. L. Radhitya, and i P. A. E. D. Udayana, "Sentiment Analysis of Public Opinion on the Impact of Deepfake Technology Using the Naive Bayes Method on Platform X," vol. 2, no. August, pp. 174–183, 2022, [Online]. Available: http://eprosiding.idbbali.ac.id/index.php/imade

[8] S. Singh, "Twitter (X) Users Statistics 2026 - Latest Global Data," Demandsage. Accessed: Apr. 06, 2026. [Online]. Available: https://www.demandsage.com/twitter-statistics/

[9] F. W. Giffary, "Penggunaan Metode Multilayer Perceptron untuk Identifikasi Sexual harrassment Pada Media Sosial X (Twitter)," 2025, Accessed: Apr. 29, 2026. [Online]. Available: https://dspace.uii.ac.id/handle/123456789/59494

[10] A. A. Khan *et al.*, "Cybersecurity, Digital Forensics, and the IoT for Deepfake Investigation on Social Media Platforms: A Review," *Human-centric Comput. Inf. Sci.*, vol. 15, pp. 1–21, 2025, doi: 10.22967/HCIS.2025.15.038.

[11] R. N. Dasmen, M. R. Pratama, H. Yasir, and A. Budiman, "Analisis Forensik Digital Pada Kasus Cyberbullying Dengan Metode National Institute of Standard and Technology Sp 800-86," *J. Ilm. Inform.*, vol. 12, no. 01, pp. 68–73, 2024, doi: 10.33884/jif.v12i01.8344.

[12] M. A. Kustian, "Analisis Forensik Penggunaan Fungsi Hash Dalam Menentukan Keaslian Video, Metadata Image Dan Magic Number File," *J. Sains, Nalar, dan Apl. Teknol. Inf.*, vol. 2, no. 2, pp. 70–76, 2023, doi: 10.20885/snati.v2i2.21.

[13] Z. J. Syafar, Y. Salim, and A. R. Manga', "Analisis Penerapan Metode Exif Metadata Dan Metode Error Level Analysis Untuk Pengolahan Forensic Digital," *LINIER Lit. Inform. dan Komput.*, vol. 2, no. 1, pp. 129–135, 2025, doi: 10.33096/linier.v2i1.2795.

[14] T. Oorloff *et al.*, "AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 27092–27102, Jun. 2024, doi: 10.1109/CVPR52733.2024.02559.

[15] L. H. Liang, "Reality Defender brings web-based deepfake protection to enterprise | Biometric Update," Biometric Update.com. Accessed: May 12, 2026. [Online]. Available: https://www.biometricupdate.com/202511/reality-defender-brings-web-based-deepfake-protection-to-enterprise#respond

[16] K. R. Ahmed *et al.*, "Enhanced Detection of GAN-Generated Image Forgeries Using Error Level Analysis and Deep Learning Models," *IET Image Process.*, vol. 20, no. 1, p. e70345, Jan. 2026, doi: 10.1049/ipr2.70345.

[17] H. Bisri and M. I. Marzuki, "Forensik Citra Digital Menggunakan Metode Error Level Analysis, Clone Detection dan Exif Untuk Deteksi Keaslian Gambar," *G-Tech J. Teknol. Terap.*, vol. 7, no. 2, pp. 586–595, Mar. 2023, doi: 10.33379/gtech.v7i2.2363.

[18] F. Ferdiansyah, M. R. A. Deazwara, R. R. Billanivo, M. Ardiansyah, and I. Ilham, "Forensic Analysis of AI-Generated Image Alterations Using Metadata Evaluation, ELA, and Noise Pattern Analysis," *J. Inf. Syst. Informatics*, vol. 7, no. 4, pp. 4014–4035, Dec. 2025, doi: 10.63158/journalisi.v7i4.1362.

[19] I. K. Khairullah, A. Yusa, A. D. Hartanto, Hartatik, and Kusnawi, "Deteksi Citra Digital Menggunakan Algoritma Cnn Dengan Model Normalisasi Rgb," *Intechno J. (Information Technol. Journal)*, vol. 2, no. 2, pp. 56–61, 2020, doi: 10.24076/intechnojournal.2020v2i2.1545.