ILKOM Jurnal Ilmiah Vol. 18, No. 2, August 2026, pp. 281-291
Accredited 2nd by KEMDIKTISAINTEK No. 10/C/C3/DT.05.00/2025; E-ISSN 2548-7779 | P-ISSN 2087-1716

Forensic Analysis for Detecting Deep-Fake Images Using A Convolutional Neural Network (CNN) and The National Institute of Standards and Technology (NIST) Methods

Muhammad Na'im Al Jum'ah [a,1,*]; Hamid Wijaya [a,2]; Muh. Hajar Akbar [a,3]; Suwito Pomalingo[b,4]
[a] *Universitas Sembilanbelas November Kolaka, Jl. Pemuda, Kolaka, 93517, Indonesia*
[b] *Universitas Muslim Indonesia Makassar, Jl. Urip Sumoharjo, Makassar, 90231, Indonesia*
[1] *Muhnaimusn@gmail.com;* [2] *Hamidwijaya@usn.ac.id;* [3] *Mhajarakbar@usn.ac.id;* [4] *Suwito.pomalingo@umi.ac.id;*
** Corresponding author*

Article history: Received November 23, 2025; Revised December 11, 2025; Accepted April 28, 2026; Available online August 10, 2026

Abstract

The development of Artificial Intelligence (AI) has significantly influenced audio, video, and image manipulation techniques, commonly known as deepfakes. Image forensics faces an urgent challenge in identifying and mitigating the impact of deepfake content to maintain the integrity and credibility of digital information. This research aims to perform forensic analysis in accordance with NIST standards and to implement Convolutional Neural Network (CNN) methods to detect deepfake images. Based on the test results, the Convolutional Neural Network (CNN) method can be effectively applied to deepfake image detection. The CNN architecture used can identify the distinct visual characteristics of deepfake images with high performance. The model demonstrates the ability to learn and minimize prediction errors on training data. Accuracy graphs indicate that the model has successfully learned data patterns, as evidenced by consistent improvements in both training and validation data as the number of epochs increases. Furthermore, the loss graph shows a downward trend, signifying a continuous reduction in model error. The precision graph demonstrates the model's effectiveness in reducing false positives, thereby minimizing errors in detecting the original data. The recall graph also indicates improved detection performance on the training data. The ROC curve suggests that the model possesses superior classification capabilities compared to random guessing. Additionally, the Area Under the Curve (AUC) of 0.6544 serves as a quantitative indicator of performance, indicating that the model has moderate capability for class differentiation. Detection results from the CNN model on a dataset of real and deepfake images show that the Confidence and Raw Score values can distinguish between the two; however, the confidence levels still fluctuate around the classification threshold. Low confidence values in certain images suggest that the extracted features are not yet optimal at distinguishing between real faces and manipulated images. Moreover, the application of the National Institute of Standards and Technology (NIST) standards (Collection, Examination, Analysis, and Reporting) for forensic analysis ensures that the evidence gathered is legally accountable in court. Thus, these standards can serve as a scientific reference to ensure a more structured and standardized investigation process for deepfake images.

Keywords: Image Deep-fake; Image Forensics; Convolutional Neural Network (CNN); Digital Forensics; National Institute of Standards and Technology (NIST).

Introduction

The rapid advancement of artificial intelligence (AI) has significantly influenced manipulation techniques for audio, video, and images, commonly known as deepfakes. A deepfake is a manipulation technique involving complex neural networks trained on vast datasets of images, audio, or video to create manipulations that appear realistic and authentic [1]. Through the use of deepfakes, perpetrators can manipulate information regarding individuals, leading to the spread of false information and identity manipulation [2], [3], [4]. Such actions are highly dangerous when deepfake images are disseminated on social media, causing the rapid spread of misinformation [5]. The world's reliance on electronic technology is also a significant factor in the spread of false information, making it essential for individuals to understand evolving information trends [6]. Furthermore, the emergence of deepfake technology poses a serious threat to the authenticity and integrity of multimedia forensic content [7]. A primary concern is the potential for large-scale information dissemination that can erode public trust, such as the spread of

fake news or the manipulation of public opinion [8]. Additionally, deepfake technology poses a threat to privacy and security by facilitating identity theft and impersonation [9].

Advancements in digital image manipulation have led to the widespread adoption of deepfake technology, presenting significant obstacles in the fields of digital forensics and cybersecurity [10]. Increasingly sophisticated deepfake techniques have made it progressively difficult to detect and link manipulated images; their realistic and convincing nature can deceive both humans and traditional forensic detection methods [11]. Perpetrators can utilize deepfakes to defraud individuals, gain unauthorized access to sensitive information, or manipulate others into believing they are interacting with a different person [12].

Various image manipulation tools can generate deepfakes, such as FaceApp, FaceSwap, DeepFaceLab, DeepNude Faceswap, Faceswap-GAN, FSGAN, and StyleGAN [13], [14]. Given the abundance of technology available for creating deepfakes, there is a pressing need for solutions in deepfake image analysis. One method that can be utilized for this purpose is the Convolutional Neural Network (CNN). As a deep learning model, CNNs can recognize anomalies and patterns within images and identify artifacts or deviations using generative deepfake detection models. The CNN method offers several advantages in deepfake detection, including robustness in pixel-level anomaly detection and high accuracy [15].

The rise of deepfake images represents a significant threat to security and privacy. Several studies have been developed for deepfake analysis, such as combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) for detection [16] and the Secure Deepfake Detection Network (SecDFD Net) for facial deepfake detection [17]. Other research has developed deep learning models that combine CNNs with Recurrent Neural Networks (RNN) to detect deepfakes generated by Generative Adversarial Networks (GANs) [18]. Additionally, the GAN-CNN ensemble model was developed to address catastrophic forgetting in CNN detectors during deepfake detection [19]. NA-VGG has also been developed to detect deepfake face images based on image noise and augmentation [20]. Furthermore, CNN models can be combined with RNNs and hybrid CNN-LSTM models to detect deepfake content [21].

Moreover, forensic techniques for deepfake detection have been established, such as the Digital Forensics for the Socio-Cyber World (DF-SCW) framework, which is used to investigate multimedia deepfakes on social media platforms [22]. Other methods include the use of face image feature vectors extracted via Facenet and Euclidean distance [23], as well as the utilization of digital watermarking to embed markers for real-time deepfake detection [24]. Additional research has developed a new framework for digital forensic investigations on social media platforms specifically for deepfake detection, also referred to as DF-SCW. This model integrates AI-supported deep neural network-based feature identification and recognition with multimedia forensic investigation processes, making digital investigations more precise by providing real-time qualitative assessments as media is uploaded to social platforms [25].

Image forensics faces an urgent task: develop innovative, reliable methods to effectively identify and mitigate the impact of deepfake content, preserving the integrity and credibility of digital information [11]. A key challenge in forensic analysis is ensuring that the processes used meet established standards and are legally admissible. Various models can be used in forensic analysis, including the National Institute of Standards and Technology (NIST) method. NIST provides guidance to industries, government agencies, and other organizations for managing cybersecurity risks, offering a high-level taxonomy of cybersecurity outcomes [26]. The NIST method consists of four stages: Collection, Examination, Analysis, and Reporting. Therefore, this research aims to conduct forensic analysis using the NIST method while implementing Convolutional Neural Network (CNN) methods to detect deepfake images. In this study, the NIST method serves as the reference for the forensic analysis process, while the CNN method is used to detect and analyze deepfake images.

Method

A. Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) are designed with a hierarchical, layered architecture that has proven highly effective for image-based data analysis. Comprising convolutional layers that scan input images using learnable filters, pooling layers for spatial dimension reduction, activation layers to introduce non-linearity, and fully connected layers for global feature aggregation, CNNs excel at capturing intricate patterns and hierarchical

representations within images [27]. Generally, a Convolutional Neural Network (CNN) consists of four primary components: (a) the convolution layer, (b) the pooling layer, (c) the activation function, and (d) the fully connected layer [28]. The functions of these four components are illustrated in the figure below:

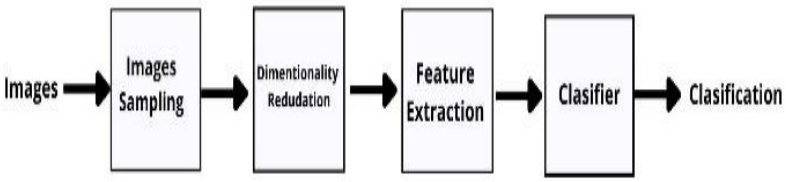

Figure 1. Elementary constituents of CNN[29]

The architecture of a Convolutional Neural Network (CNN) consists of several components, including [30]:

- Input Layer: The input layer receives raw images and passes them to subsequent layers for feature extraction.
- Convolution Layer: Each output is connected to a small neighborhood within the input through a weight matrix (filter or kernel). Multiple kernels can be defined for each convolution layer, with each producing an individual output. Each filter slides across the input to generate a 2D output. The outputs from each filter are then stacked to form an output volume.
- ReLU (Rectified Linear Unit): Following the convolution layer, the next stage is the ReLU layer. The ReLU function sets all negative values in the convolutional layer to zero, reducing the effective time required to train the model.
- Pooling: This layer provides translation invariance through subsampling, thereby reducing the dimensions of the feature maps. Average pooling and max pooling are commonly used for this process.
- Fully Connected Layer: As the final layer of the model, the fully connected layer takes all filtered images and maps them into specific labels and categories.
- Softmax Layer: Located immediately before the output layer, this layer assigns decimal probabilities to each class. These decimal probabilities range between 0 and 1.

B. National Institute of Standards and Technology (NIST)

In digital forensics, compliance with standard procedures is the foundation of every investigation. This is practiced to ensure a comprehensive and scientific analysis. Establishing well-verified standard protocols will yield comprehensive reports when analyzing evidence. Several prominent organizations, such as the National Institute of Standards and Technology (NIST), have played a vital role in developing robust standards in the field of digital forensics. This research will adopt NIST SP 800-86 for forensic analysis [31].

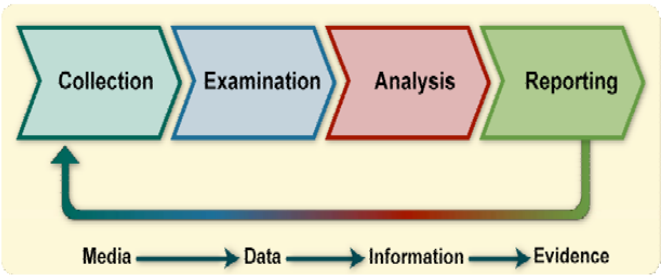

Figure 2. NIST SP 800-86 Framework

1) Data Collection

The first stage of the forensic process begins with identifying and acquiring relevant data sources related to deepfake images. In this process, the acquisition of deepfake image data must adhere to forensic principles;

therefore, practical steps must be taken to ensure that data collection aligns with incident response protocols and that the evidence remains admissible in legal proceedings.

2) Examination

Once data collection is complete, the next stage is the examination process. This stage involves extracting relevant data and information to filter and identify files containing critical evidence, while segregating irrelevant data. This process utilizes various relevant tools to ensure the information obtained is actionable. In this research, the dataset will undergo preprocessing, which includes resizing, normalization, and augmentation.

3) Analysis

After the relevant information has been extracted, the analysis process is conducted. In this stage, the goal is to obtain deeper insights by linking people, places, objects, and events to draw evidence-supported conclusions. To achieve this, the process implements specific methods and tools to establish correlations between evidence and data. Specifically, in this study, the analysis involves implementing the Convolutional Neural Network (CNN) method. Following this, model training and validation are performed, and finally, evaluation and forensic visualization are conducted.

4) Reporting

The final stage is the reporting process. In this phase, the forensic analyst compiles and presents the analytical findings in a report, ensuring that all necessary information is comprehensive. The report must be clearly articulated to ensure that the judge can effectively comprehend the information presented. In this study, the detection results will be documented alongside forensic interpretations.

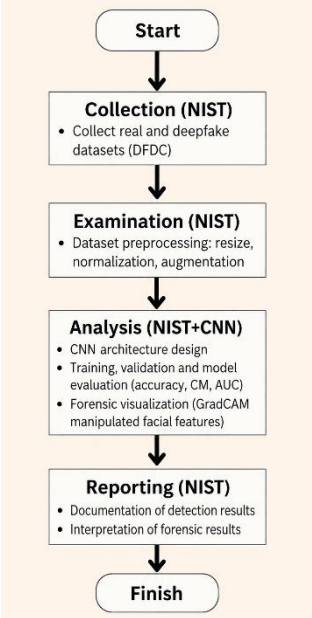

Figure 3. Research Flow

Results and Discussion

A. Data Collection

During data collection, the dataset was sourced from www.kaggle.com and comprised real and manipulated images. It consists of 3,006 JPG images (256 x 256 pixels), partitioned into three subsets training, testing, and validation each containing two classes 'Fake' and 'Real' with 501 images per class. 'Fake' images represent faces generated or modified through various techniques. The dataset was processed as needed to achieve optimal research results and features both authentic and manipulated human faces.

Figure 4. Dataset

Subsequently, a dataset management process was conducted to ensure the data was more organized. This process involved partitioning the dataset into three distinct subsets: training, testing, and validation. Each subset contains the previously collected 'fake' and 'real' images. Furthermore, this structured organization facilitates more efficient model training and evaluation.

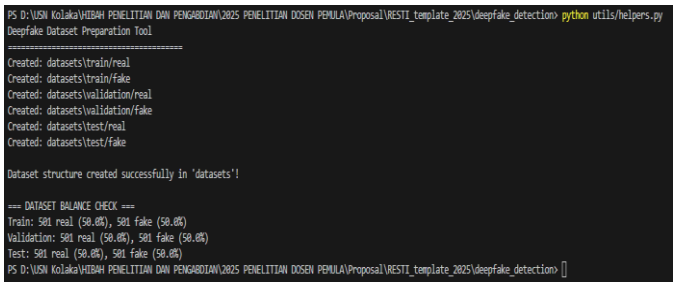

Figure 5. Dataset Management

B. Examination.

In this experimental phase, model training was performed on the dataset. This training was conducted to implement a Convolutional Neural Network (CNN) architecture that distinguishes between real and fake images. The CNN implementation utilizes several convolutional blocks specifically for image feature extraction. Each convolutional block includes a convolutional layer, an activation function, and a pooling layer. The architecture employed in this process comprises four such convolutional blocks, followed by batch normalization, dropout, and a dense layer. Subsequently, data preprocessing and augmentation were carried out. These procedures were implemented to facilitate the training and validation of the collected dataset. During this stage, images were resized to 224 x 224 pixels. Furthermore, pixel values were normalized to [0, 1], and data augmentation was applied to the training set to enhance data diversity and mitigate overfitting.

C. Analysis

The next stage involves the analysis process. In this phase, the initial step is to implement the CNN architectural model. The model employed is a Sequential model, constructed incrementally with a linear structure, starting with the first Conv2D layer and extending to the Dense layer. Based on the training results illustrated in **Figure 6**, the implemented Sequential model has 27,001,377 parameters, comprising 26,997,921 trainable and 3,456 non-trainable. The substantial number of parameters indicates that the model has a high capacity for learning complex features; however, it also carries a potential risk of overfitting if the dataset utilized is relatively small.

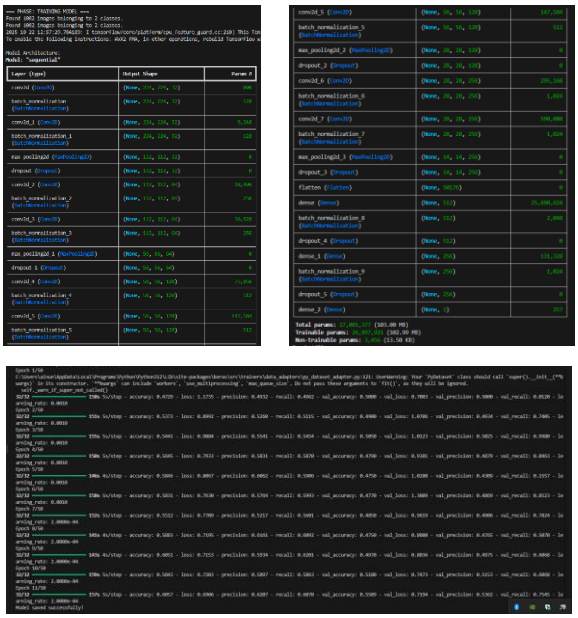

Figure 6. Training dataset

Upon training the model for 50 epochs, the results evidenced a gradual improvement in performance. Training accuracy rose from 47.29% in the initial epoch to approximately 60.57% by the eleventh epoch. Simultaneously, the training loss decreased significantly from 1.1735 to 0.7194, indicating that the model progressively captured the underlying data patterns. The training process is further detailed in **Figure 7**. As illustrated in the accuracy and loss graphs, training accuracy steadily increased from approximately 0.49 to 0.59, while the loss decreased from 1.0 to 0.7. This trend suggests that the model successfully learned to minimize prediction errors on the training set. However, the validation accuracy and loss fluctuated significantly and underperformed relative to the training metrics. This discrepancy points to slight overfitting, in which the model adapts better to the training data than to unseen datasets.

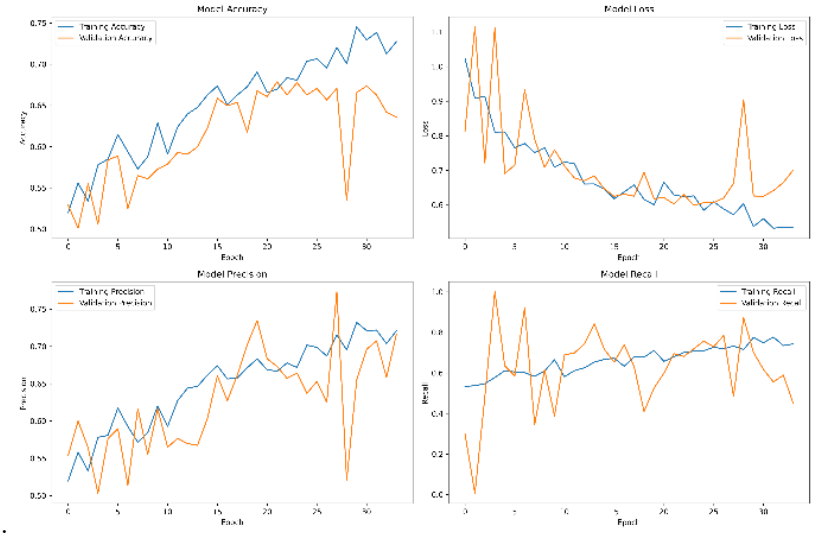

Figure 7. Training dataset

Meanwhile, the precision and recall graphs demonstrate similar trends. Training precision increased to approximately 0.59, while validation precision remained between 0.44 and 0.52. It is worth noting that validation recall varied widely, from 0 to 1.0, indicating instability in the model's ability to accurately identify all classes. Overall, although the CNN model has demonstrated robust learning from the training data, further optimization is required. This could be achieved by increasing the dataset size, implementing regularization techniques such as dropout, and performing hyperparameter tuning to ensure a more balanced performance between training and validation.

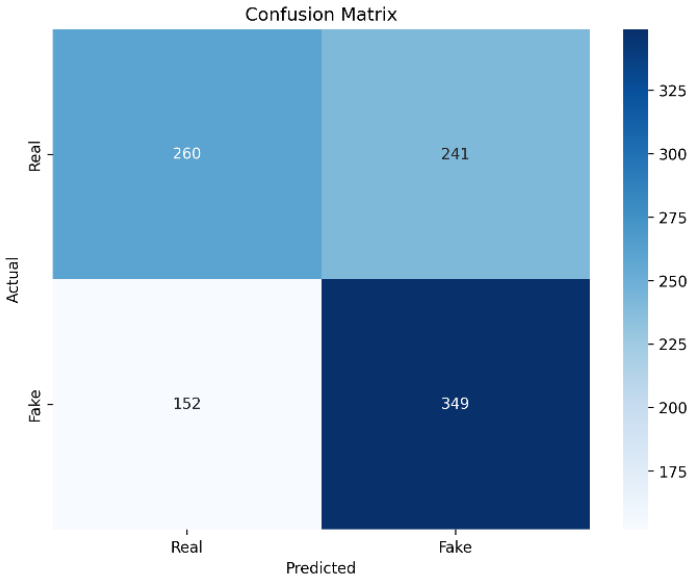

Figure 8. Confusion Matrix

The results of the Confusion Matrix, as illustrated in **Figure 8**, indicate the following four primary values: True Positive (TP) = 260, True Negative (TN) = 349, False Positive (FP) = 241, and False Negative (FN) = 152. These four values demonstrate that the model has a high capacity to identify real images but is less effective at detecting deepfakes. The high number of false negatives indicates that a significant portion of deep-fake images were incorrectly classified and identified as real. This suggests an imbalance in performance across classes, with the model exhibiting a bias toward the 'real' class. This is attributed to unbalanced data distribution between the real and fake classes, the high visual similarity in deep-fake imagery, and a model architecture that lacks sufficient complexity to extract the subtle patterns needed to differentiate the two classes.

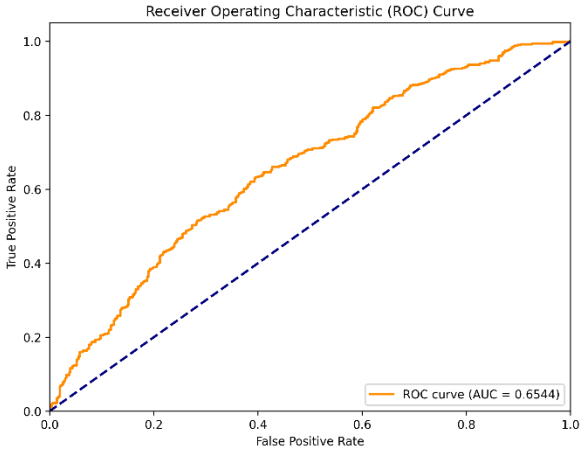

Figure 9. ROC curve

Furthermore, an ROC analysis was conducted, as illustrated in **Figure 9**. The ROC curve plots the relationship between the True Positive Rate (TPR, or sensitivity) and the False Positive Rate (FPR) at various classification thresholds. An ROC curve positioned above the diagonal line indicates that the model performs better than random guessing. Moreover, the Area Under the Curve (AUC) of 0.6544, a quantitative performance indicator, indicates that the model has a moderate ability to distinguish between classes.

Subsequently, a detection process was performed on the dataset using a Convolutional Neural Network (CNN) prediction model on a set of images comprising real and deep-fake categories. As shown in the results in Figure 10, confidence levels and raw scores were used as the basis for predicting real and deep-fake images. Confidence represents the model's degree of certainty (probability) for its predictions, expressed on a scale from 0 to 1. Meanwhile, the raw_score is the raw probability value before it is converted into a final decision.

no	filename	result	confidence	raw_score	status
1	real_0.jpg	FAKE (Deepfake)	0.50219285	0.50219285	FAKE
2	real_1.jpg	REAL	0.52316996	0.47683004	REAL
3	real_10.jpg	REAL	0.80770025	0.19229975	REAL
4	real_100.jpg	FAKE (Deepfake)	0.90109307	0.90109307	FAKE
5	real_101.jpg	REAL	0.85735789	0.14264211	REAL
6	real_102.jpg	FAKE (Deepfake)	0.74996567	0.74996567	FAKE
7	real_103.jpg	FAKE (Deepfake)	0.75902039	0.75902039	FAKE
8	real_104.jpg	FAKE (Deepfake)	0.90014791	0.90014791	FAKE
9	real_105.jpg	REAL	0.89641405	0.10358595	REAL
10	real_106.jpg	FAKE (Deepfake)	0.6968745	0.6968745	FAKE
11	real_107.jpg	FAKE (Deepfake)	0.77624428	0.77624428	FAKE
12	real_108.jpg	FAKE (Deepfake)	0.51443875	0.51443875	FAKE
13	real_109.jpg	FAKE (Deepfake)	0.51939368	0.51939368	FAKE
14	real_11.jpg	FAKE (Deepfake)	0.91429621	0.91429621	FAKE
15	real_110.jpg	REAL	0.58192137	0.41807863	REAL
16	real_111.jpg	FAKE (Deepfake)	0.71408021	0.71408021	FAKE
17	real_112.jpg	REAL	0.68340978	0.31659022	REAL
18	real_113.jpg	FAKE (Deepfake)	0.54681563	0.54681563	FAKE
19	real_114.jpg	FAKE (Deepfake)	0.59809053	0.59809053	FAKE
20	real_115.jpg	REAL	0.51118109	0.48881891	REAL
21	real_116.jpg	FAKE (Deepfake)	0.79152834	0.79152834	FAKE
22	real_117.jpg	FAKE (Deepfake)	0.75927514	0.75927514	FAKE
23	real_118.jpg	FAKE (Deepfake)	0.69174868	0.69174868	FAKE
24	real_119.jpg	FAKE (Deepfake)	0.76249921	0.76249921	FAKE
25	real_12.jpg	FAKE (Deepfake)	0.89916825	0.89916825	FAKE
26	real_120.jpg	FAKE (Deepfake)	0.80176151	0.80176151	FAKE
27	real_121.jpg	REAL	0.87190036	0.12809964	REAL
28	real_122.jpg	FAKE (Deepfake)	0.54340625	0.54340625	FAKE
29	real_123.jpg	FAKE (Deepfake)	0.70215261	0.70215261	FAKE
30	real_124.jpg	FAKE (Deepfake)	0.70293188	0.70293188	FAKE
31	real_125.jpg	REAL	0.76092015	0.23907985	REAL
32	real_126.jpg	FAKE (Deepfake)	0.73005784	0.73005784	FAKE
33	real_127.jpg	FAKE (Deepfake)	0.75741863	0.75741863	FAKE
34	real_128.jpg	FAKE (Deepfake)	0.52051353	0.52051353	FAKE
35	real_129.jpg	REAL	0.9382024	0.0617976	REAL
36	real_13.jpg	FAKE (Deepfake)	0.89637941	0.89637941	FAKE
37	real_130.jpg	REAL	0.82178375	0.17821625	REAL
38	real_131.jpg	REAL	0.61356172	0.38643828	REAL
39	real_132.jpg	FAKE (Deepfake)	0.87590635	0.87590635	FAKE
40	real_133.jpg	FAKE (Deepfake)	0.71090502	0.71090502	FAKE

Figure 10. Real and Deep-fake Image Prediction

The prediction results based on Confidence and Raw_Score values indicate that the CNN model can distinguish between real and deep-fake images; however, confidence levels remain near the classification threshold. The low confidence values observed in some images indicate that the extracted features are not yet optimal for distinguishing between authentic and manipulated faces.

D. Reporting

The final stage is the reporting phase. This process entails compiling results that are suitable for presentation as digital evidence. Based on the data testing procedure, classifications were determined utilizing confidence values and raw scores. From the comprehensive image dataset evaluated, it was determined that images are classified as deep-fakes when their confidence values exceed 0.5. Conversely, images with confidence values below 0.5 are considered authentic. These data function as supporting documentation in legal proceedings and serve as digital evidence.

no	filename	result	confidence	raw_score	status	Penjelasan
1	real_0.jpg	FAKE (Deepfake)	0.502192855	0.502192855	FAKE	Nilai confidence di atas 0.5
2	real_1.jpg	REAL	0.523169965	0.476830035	REAL	Nilai confidence di bawah 0.5
3	real_10.jpg	REAL	0.807700247	0.192299753	REAL	Nilai confidence di bawah 0.5
4	real_100.jpg	FAKE (Deepfake)	0.901093066	0.901093066	FAKE	Nilai confidence di atas 0.5
5	real_101.jpg	REAL	0.857357889	0.142642111	REAL	Nilai confidence di bawah 0.5
6	real_102.jpg	FAKE (Deepfake)	0.749965668	0.749965668	FAKE	Nilai confidence di atas 0.5
7	real_103.jpg	FAKE (Deepfake)	0.759020388	0.759020388	FAKE	Nilai confidence di atas 0.5
8	real_104.jpg	FAKE (Deepfake)	0.900147915	0.900147915	FAKE	Nilai confidence di atas 0.5
9	real_105.jpg	REAL	0.896414049	0.103585951	REAL	Nilai confidence di bawah 0.5
10	real_106.jpg	FAKE (Deepfake)	0.696874499	0.696874499	FAKE	Nilai confidence di atas 0.5
11	real_107.jpg	FAKE (Deepfake)	0.776244283	0.776244283	FAKE	Nilai confidence di atas 0.5
12	real_108.jpg	FAKE (Deepfake)	0.514438748	0.514438748	FAKE	Nilai confidence di atas 0.5
13	real_109.jpg	FAKE (Deepfake)	0.519393682	0.519393682	FAKE	Nilai confidence di atas 0.5
14	real_11.jpg	FAKE (Deepfake)	0.91429621	0.91429621	FAKE	Nilai confidence di atas 0.5
15	real_110.jpg	REAL	0.581921369	0.418078631	REAL	Nilai confidence di bawah 0.5
16	real_111.jpg	FAKE (Deepfake)	0.714080215	0.714080215	FAKE	Nilai confidence di atas 0.5
17	real_112.jpg	REAL	0.68340978	0.31659022	REAL	Nilai confidence di bawah 0.5
18	real_113.jpg	FAKE (Deepfake)	0.546815634	0.546815634	FAKE	Nilai confidence di atas 0.5
19	real_114.jpg	FAKE (Deepfake)	0.598090529	0.598090529	FAKE	Nilai confidence di atas 0.5
20	real_115.jpg	REAL	0.511181086	0.488818914	REAL	Nilai confidence di bawah 0.5
21	real_116.jpg	FAKE (Deepfake)	0.791528344	0.791528344	FAKE	Nilai confidence di atas 0.5
22	real_117.jpg	FAKE (Deepfake)	0.759275138	0.759275138	FAKE	Nilai confidence di atas 0.5
23	real_118.jpg	FAKE (Deepfake)	0.691748679	0.691748679	FAKE	Nilai confidence di atas 0.5
24	real_119.jpg	FAKE (Deepfake)	0.762499213	0.762499213	FAKE	Nilai confidence di atas 0.5
25	real_12.jpg	FAKE (Deepfake)	0.899168253	0.899168253	FAKE	Nilai confidence di atas 0.5
26	real_120.jpg	FAKE (Deepfake)	0.801761508	0.801761508	FAKE	Nilai confidence di atas 0.5
27	real_121.jpg	REAL	0.871900365	0.128099635	REAL	Nilai confidence di bawah 0.5
28	real_122.jpg	FAKE (Deepfake)	0.543406248	0.543406248	FAKE	Nilai confidence di atas 0.5
29	real_123.jpg	FAKE (Deepfake)	0.70215261	0.70215261	FAKE	Nilai confidence di atas 0.5
30	real_124.jpg	FAKE (Deepfake)	0.702931881	0.702931881	FAKE	Nilai confidence di atas 0.5
31	real_125.jpg	REAL	0.760920152	0.239079848	REAL	Nilai confidence di bawah 0.5
32	real_126.jpg	FAKE (Deepfake)	0.730057836	0.730057836	FAKE	Nilai confidence di atas 0.5
33	real_127.jpg	FAKE (Deepfake)	0.757418633	0.757418633	FAKE	Nilai confidence di atas 0.5
34	real_128.jpg	FAKE (Deepfake)	0.520513535	0.520513535	FAKE	Nilai confidence di atas 0.5
35	real_129.jpg	REAL	0.9382024	0.0617976	REAL	Nilai confidence di bawah 0.5

Figure 11. Classification of Real vs. Deep-fake Images

Conclusion

Based on the conducted tests, the results demonstrate that the Convolutional Neural Network (CNN) method can be successfully applied to deepfake image detection. The CNN architecture utilized has also proven capable of identifying the distinct visual characteristics of deep-fake imagery with favorable performance. The model effectively learned to minimize prediction errors on the training data, as evidenced by the accuracy graphs, which show the model successfully capturing data patterns. This is further supported by the consistent increase in both training and validation accuracy as the number of epochs progressed. Furthermore, the loss curves exhibited a downward trend, indicating a continuous reduction in model error. The precision metrics suggest that the model is sufficiently effective at reducing false positives, thereby minimizing errors in detecting the original data. Similarly, the recall graphs show an improvement in the training data for detecting false positives. The ROC curve indicates that the model possesses superior classification capabilities compared to random guessing. Additionally, the Area Under the Curve (AUC) value of 0.6544 serves as a quantitative performance indicator, signifying a moderate ability to differentiate between classes.

The detection results from the CNN prediction model on a dataset comprising real and deep-fake images show that the Confidence and Raw Score values can distinguish between the two categories. However, confidence levels still fluctuate around the classification threshold. Low confidence scores for certain images suggest that the extracted features are not yet optimal for distinguishing between authentic faces and manipulated images. Moreover, applying the National Institute of Standards and Technology (NIST) standards (Collection, Examination, Analysis, and Reporting) for forensic analysis ensures that the evidence gathered in this deep-fake study is legally defensible. Consequently, this standard can serve as a benchmark for forensic analysis of deep-fake images, providing a more structured, scientifically standardized investigation framework. For future research, it is recommended to increase the volume and variety of deep-fake datasets to allow the model to learn a broader range of facial manipulation patterns and achieve better generalization.

Acknowledgement

This research was funded through the Higher Education Research Grant program, No. 98/UN56.D.01/PN.03.00/2025, by the Directorate of Research and Community Service (DRPM) of the Ministry of Research, Technology, and Higher Education of Indonesia, in collaboration with Universitas Sembilanbelas November Kolaka.

References

[1] A. Kwok and S. Koh, "Deep-fake: a social construction of technology perspective," *Current Issues in Tourism*, pp. 1–5, Mar. 2020, doi: 10.1080/13683500.2020.1738357.

[2] A. M. Almars, "Deep-fakes Detection Techniques Using Deep Learning: A Survey," *Journal of Computer and Communications*, vol. 09, no. 05, pp. 20–35, 2021, doi: 10.4236/jcc.2021.95003.

[3] H. Wang, Z. Li, Y. Li, B. B. Gupta, and C. Choi, "Visual saliency guided complex image retrieval," *Pattern Recognit. Lett.*, vol. 130, pp. 64–72, Feb. 2020, doi: 10.1016/j.patrec.2018.08.010.

[4] L. Nataraj *et al.*, "Detecting GAN generated Fake Images using Co-occurrence Matrices," Mar. 2019.

[5] L. Verdoliva, "Media Forensics and Deep-fakes: An Overview," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 5, pp. 910–932, 2020, doi: 10.1109/JSTSP.2020.3002101.

[6] Y. Li and Q. Liu, "A comprehensive review study of cyber-attacks and cyber security; Emerging trends and recent developments," *Energy Reports*, vol. 7, pp. 8176–8186, 2021.

[7] H. Farid, "Image forgery detection," *IEEE Signal Process. Mag.*, vol. 26, no. 2, pp. 16–25, 2009, doi: 10.1109/MSP.2008.931079.

[8] B. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *Calif. Law Rev.*, vol. 107, no. 6, pp. 1753–1820, 2019, doi: 10.15779/Z38RV0D15J.

[9] A. Abul Hasanaath, H. Luqman, R. Katib, and S. Anwar, "FSBI: Deep-fake detection with frequency enhanced self-blended images," *Image Vis. Comput.*, vol. 154, Feb. 2025, doi: 10.1016/j.imavis.2025.105418.

[10] B. N. Al-din Abed, J. Karimpour, and F. Mahan, "A deep fake detection approach for cyber security threat based on deep learning and diffusion-osmosis model," *Egyptian Informatics Journal*, vol. 31, p. 100748, 2025, doi: https://doi.org/10.1016/j.eij.2025.100748.

[11] G. Bendiab, H. Haiouni, I. Moulas, and S. Shiaeles, "Deep-fakes in digital media forensics: Generation, AI-based detection and challenges," *Journal of Information Security and Applications*, vol. 88, p. 103935, Feb. 2025, doi: 10.1016/j.jisa.2024.103935.

[12] C. Cross, "Using artificial intelligence (AI) and deep-fakes to deceive victims: the need to rethink current romance fraud prevention messaging," *Crime Prevention and Community Safety*, vol. 24, no. 1, pp. 30–41, 2022, doi: 10.1057/s41300-021-00134-w.

[13] J. Jam, C. Kendrick, K. Walker, V. Drouard, J. G. S. Hsu, and M. H. Yap, "A comprehensive review of past and present image inpainting methods," *Computer Vision and Image Understanding*, vol. 203, Feb. 2021, doi: 10.1016/j.cviu.2020.103147.

[14] T. T. Nguyen *et al.*, "Deep learning for deep-fakes creation and detection: A survey," *Computer Vision and Image Understanding*, vol. 223, Oct. 2022, doi: 10.1016/j.cviu.2022.103525.

[15] R. Sunil, P. Mer, A. Diwan, R. Mahadeva, and A. Sharma, "Exploring autonomous methods for deep-fake detection: A detailed survey on techniques and evaluation," Feb. 15, 2025, *Elsevier Ltd*. doi: 10.1016/j.heliyon.2025.e42273.

[16] O. A. H. H. Al-Dulaimi and S. Kurnaz, "A Hybrid CNN-LSTM Approach for Precision Deep-fake Image Detection Based on Transfer Learning," *Electronics (Switzerland)*, vol. 13, no. 9, May 2024, doi: 10.3390/electronics13091662.

[17] B. Chen, X. Liu, Z. Xia, and G. Zhao, "Privacy-preserving Deep-fake face image detection," *Digital Signal Processing: A Review Journal*, vol. 143, Nov. 2023, doi: 10.1016/j.dsp.2023.104233.

[18] Y. Li, M.-C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1–7. doi: 10.1109/WIFS.2018.8630787.

[19] P. Sharma, M. Kumar, and H. K. Sharma, "GAN-CNN Ensemble: A Robust Deep-fake Detection Model of Social Media Images Using Minimized Catastrophic Forgetting and Generative Replay Technique," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 948–960. doi: 10.1016/j.procs.2024.04.090.

[20] X. Chang, J. Wu, T. Yang, and G. Feng, "Deep-fake Face Image Detection based on Improved VGG Convolutional Neural Network," in *2020 39th Chinese Control Conference (CCC)*, 2020, pp. 7252–7256. doi: 10.23919/CCC50068.2020.9189596.

[21] F. T. Winata, N. J. Tanuwijaya, R. Setiawan, and R. Y. Rumagit, "Comparison of deep-fake detection using CNN and hybrid models," *Procedia Comput. Sci.*, vol. 269, pp. 1556–1564, 2025, doi: https://doi.org/10.23919/CCC50068.2020.9189596.

[22] A. Ayub Khan *et al.*, "Digital forensics for the socio-cyber world (DF-SCW): A novel framework for deep-fake multimedia investigation on social media platforms," *Egyptian Informatics Journal*, vol. 27, Sep. 2024, doi: 10.1016/j.eij.2024.100502.

[23] J. Wu, K. Feng, X. Chang, and T. Yang, "A Forensic Method for Deep-fake Image based on Face Recognition," in *Proceedings of the 2020 4th High Performance Computing and Cluster Technologies Conference & 2020 3rd International Conference on Big Data and Artificial Intelligence*, in HPCCT & BDAI '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 104–108. doi: 10.1145/3409501.3409544.

[24] Z. Lai, S. Arif, C. Feng, G. Liao, and C. Wang, "Enhancing Deep-fake Detection: Proactive Forensics Techniques Using Digital Watermarking," 2025, *Tech Science Press*. doi: 10.32604/cmc.2024.059370.

[25] A. Ayub Khan *et al.*, "Digital forensics for the socio-cyber world (DF-SCW): A novel framework for deep-fake multimedia investigation on social media platforms," *Egyptian Informatics Journal*, vol. 27, Sep. 2024, doi: 10.1016/j.eij.2024.100502.

[26] The NIST Cybersecurity Framework (CSF) 2.0, "The NIST Cybersecurity Framework (CSF) 2.0," Feb. 2024. doi: 10.6028/NIST.CSWP.29.

[27] Y. Xu, W. Hong, M. Noori, W. A. Altabey, A. Silik, and N. S. D. Farhan, "Big Model Strategy for Bridge Structural Health Monitoring Based on Data-Driven, Adaptive Method and Convolutional Neural Network (CNN) Group," *SDHM Structural Durability and Health Monitoring*, vol. 18, no. 6, pp. 763–783, 2024, doi: https://doi.org/10.32604/sdhm.2024.053763.

[28] S. Indolia, A. K. Goswami, S. P. Mishra, and P. Asopa, "Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach," in *Procedia Computer Science*, Elsevier B.V., 2018, pp. 679–688. doi: 10.1016/j.procs.2018.05.069.

[29] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: a convolutional neural-network approach," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, 1997, doi: 10.1109/72.554195.

[30] K. Singh, D. Singh, and N. Mishra, "Review: Convolutional neural networks and its architecture," *Int. J. Health Sci. (Qassim).*, May 2022, doi: 10.53730/ijhs.v6nS1.7074.

[31] A. FFaizal and A. Luthfi, "Comparison Study of NIST SP 800-86 and ISO/IEC 27037 Standards as A Framework for Digital Forensic Evidence Analysis," *Journal of Information Systems and Informatics*, vol. 6, no. 2, pp. 701–718, Jun. 2024, doi: 10.51519/journalisi.v6i2.717.