JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 106-116
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1386

CNN EfficientNetB0 Approach Based on Transfer Learning for Deepfake Detection and Disinformation Mitigation

Muhammad Erico Revaldo[1*], Burhanudin Rabbani[2], Wildan Humaidi[3], Syifa Nur Fadhilah[4], Reva Echa Putri[5]
[1,2,3,4,5]Information Systems Study Program, Faculty of Engineering and Informatics, Bina Sarana Informatika University, Indonesia
[1*]erikorevaldo4@gmail.com, [2]brhnrabbani23@gmail.com, [3] wildaannhumaidii@gmail.com
[4] syifanurrrrrr@gmail.com , [5] revaputri578@gmail.com

Abstract
Rapid advances in deep learning technology have led to the emergence of artificial intelligence (AI) media that is very similar to reality, called deepfakes, which have the potential to pose a serious threat to information integrity and public trust. Although detection methods using Convolutional Neural Networks (CNN) have been developed, most still struggle with generalization, particularly in distinguishing modern deepfakes from non-standard original images such as selfies, which often leads to high false positive rates. This study introduces a robust detection model based on the EfficientNetB0 architecture implemented through transfer learning techniques. To improve generalization capabilities and minimize bias, we compiled a large and balanced combined dataset by combining three different public datasets (including classic deepfakes, face swaps, and many authentic selfies). The model was trained using a two-stage strategy: first for feature extraction, then refinement with a very low learning rate. The model's performance was thoroughly evaluated on stratified test data using five key metrics. The results of the experiment showed outstanding performance, achieving 99.81% accuracy and a Macro F1 score of 99.81%. Additionally, the reliability metrics ROC-AUC, Average Precision (AP), and True Positive Rate (TPR) all reached 99.99%, while the False Positive Rate (FPR) remained strictly at 1%. As proof of concept, this optimized model was implemented in a web prototype built using the Django framework, allowing users to upload images and receive classification results in real-time.
Keyword : *Deepfake Detection, Transfer Learning, EfficientNetB0, Convolutional Neural Network, Disinformation Mitigation*

1. INTRODUCING

Artificial intelligence (AI) has undergone rapid development, giving rise to advanced techniques such as Deepfake, which is a combination of deep learning and fake [1]. Deepfake is capable of manipulating human face photos and videos with a very realistic level of authenticity [2]. On one hand, this opens up creative opportunities in various sectors, but on the other hand, deepfake has the potential to pose a serious threat to the authenticity of information through the spread of fake news (disinformation) and illegal content that is difficult to distinguish [3][4][5]. The phenomenon of spreading manipulated photos of public figures is one of the tangible forms of the social impact of deepfakes in

Muhammad Erico Revaldo: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 106-116
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1386

various countries [6][7]. Based on this potential threat, the development of reliable deepfake detection methods has become urgent for the mitigation of disinformation [8].

Various studies in the field of digital image detection show that Convolutional Neural Networks (CNN) have proven to be the most dominant and promising approach [9]. Convolutional Neural Networks are specifically designed for visual data and are effective in automatically learning and extracting subtle features from image data [10][11]. Various studies have applied the Convolutional Neural Network method, both those built from scratch and those using advanced pre-trained architectures through transfer learning such as VGG, ResNet, and EfficientNet [12].

However, a significant research gap still exists. Many current detection models show weaknesses in the generalization of models trained on one type of dataset (e.g., old face swap GANs) whose performance declines dramatically when tested on deepfakes created with new generative AI techniques, or, more importantly, fail to recognize non-standard original photos (such as selfies) and misclassify them as fake [13]. This false positive phenomenon is a major problem in practical applications [14]. In addition, performance evaluations often focus only on accuracy metrics, which can be misleading if the dataset is imbalanced, and ignore reliability metrics that are more crucial for security systems [15].

In this context, choosing the EfficientNetB0 architecture is a strategic move due to its proven efficiency and scalability in various image classification tasks [16]. With the concept of compound scaling, EfficientNetB0 is able to achieve an optimal balance between accuracy and computational efficiency, making it suitable for use in web-based detection systems that require high performance but remain lightweight [17]. In addition, the application of transfer learning in this model speeds up the classification process [18]. by utilizing trained visual representations from large datasets such as ImageNet, thereby accelerating the convergence process and improving the model's generalization capabilities on diverse deepfake data [19].

This research aims to address these gaps. Our novelty lies in our dataset curation methodology and comprehensive evaluation. First, we built a balanced combined dataset from three different data sources, specifically adding tens of thousands of authentic selfies to train the model to recognize variations in authentic photos and reduce false positives. Second, we applied the modern EfficientNetB0 Convolutional Neural Network architecture through transfer learning with a two-stage fine-tuning strategy. Third, we conducted a holistic performance evaluation using five different metrics (Accuracy, F1-Macro, ROC-AUC, Average Precision, and TPR@FPR=1%) to rigorously validate the reliability of the model [20][21].

Furthermore, this research is also expected to contribute to the development of digital media manipulation detection technology as part of efforts to maintain the authenticity and integrity of information in the era of disinformation. With the increasing misuse of deepfakes in social and political contexts, accurate and adaptive detection systems are urgently needed. The Convolutional Neural Network-based approach proposed in this study focuses not only on improving accuracy but also on model generalization to complex data variations. The results of this study are expected to form the basis for the development of multimodal detection systems such as deepfake video detection, as well as the exploration of advanced architectures such as Vision Transformer (ViT) or Swin Transformer in the future.

2. METHODOLOGY

This research methodology is structured in an orderly manner to ensure a planned approach to creating and evaluating deepfake detection models. The overall research process, from data collection to final implementation, is depicted in the following flowchart:

Muhammad Erico Revaldo: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 106-116
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1386

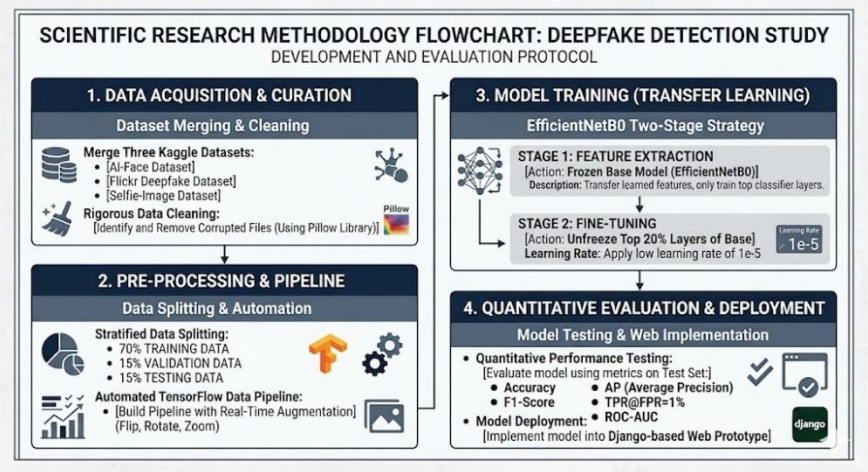

Figure 1. Research Methodology Flowchart

As shown in Figure 1, the research process is divided into four main steps: data collection and management, initial preparation and workflow automation, model training through transfer learning, and finally, quantitative evaluation along with the implementation of a web prototype. All computational processes are implemented using Google Colab with the TensorFlow and Keras frameworks. Below is a detailed explanation of each step:

2.1 Dataset Acquisition and Curation

This stage covers the process of collecting and preparing data to be used in the research. Acquisition is carried out by combining three public datasets from Kaggle that are relevant to deepfake detection, while curation is carried out to ensure data quality and balance by removing corrupted files, standardizing labels, and balancing the amount of data between classes (real and fake). The result of this stage is a structured, combined dataset that is ready for use in the pre-processing stage.

To overcome the problem of generalization and reduce the false positive rate (e.g., selfies being detected as fake), this study does not rely on a single data source. Data acquisition was carried out by combining three large-scale public datasets from Kaggle to create a more representative and varied data set, namely:

1. AI-Face-Detection Dataset, providing basic data of real and AI-generated (fake) face images. All data from this dataset is used.
2. Flickr Deepfake dataset providing a large number of fake face swap images. We randomly sampled 50,000 images from this dataset to enrich the variety of fake data.
3. Selfie-Image-Detection Dataset, providing non-professional real images (selfies) to balance the amount of real data and train the model to recognize real photo variations. We took all images from the Validation_data/Selfie folder (approximately 3,931 images) and a random sample of 44,868 images from the Training_data/Selfie folder.

Next, in the dataset curation stage, all file paths and labels (real or fake) from the three sources were collected into a single Pandas DataFrame. The curation process included removing corrupted files, verifying labels, and balancing the amount of data between classes to avoid class imbalance. The final result of this stage was a clean, structured, combined dataset that was ready for use in the next pre-processing stage.

Muhammad Erico Revaldo: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 106-116
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1386

2.2 Data Pre-processing

The data pre-processing stage is an important process that ensures the data used is of good quality and ready to be processed by the model. The first step is to verify the integrity of the image files using the Pillow library (Image.verify()) to detect and remove damaged or unreadable files. This step is important so that the model training process runs smoothly.

Next, data splitting is performed using the train_test_split function from Scikit-learn with the stratify=True parameter to maintain a balanced proportion of real and fake classes in each data subset. The dataset is divided into three parts: 70% for training, 15% for validation, and 15% for testing. To make the data reading and training process more efficient, a data pipeline is built using tf.data.Dataset from TensorFlow. This pipeline performs a series of automated steps, including:

1. Converting text labels into binary format (0 for real, 1 for fake);
2. Reading and decoding image files into RGB format with a uniform size of 224 × 224 pixels;
3. Normalizing pixel values to match the EfficientNet preprocessing standard;
4. Applying data augmentation (such as RandomFlip, RandomRotation, and RandomZoom) to increase the diversity of training data and avoid overfitting;
5. Applying an error handling mechanism to skip files that fail to be read during runtime;
6. Forming batches of 32 images with a prefetch feature to maximize GPU user efficiency.

This preprocessing stage produces clean, structured image data that is ready for use in the model training stage, while ensuring that the training process runs stably and free from interference from invalid data.

2.3 Model Architecture (Transfer Learning)

This stage describes the design of a deepfake detection model using the Transfer Learning approach with the EfficientNetB0 architecture, which is known to have an optimal ratio between accuracy and computational efficiency. This approach allows the utilization of knowledge previously learned from a large dataset (ImageNet), so that the model can recognize visual patterns with less training data and more efficient computing time. In general, the model architecture design consists of several main components as follows:

1. Base Model: The EfficientNetB0 model is loaded without the top layer (include_top=False) so that it can be used as the main feature extractor from facial images.
2. Preprocessing Layer: The Efficientnet.preprocess_input layer is integrated directly into the architecture to normalize pixel values to match the standards used in the pre-trained model.
3. Head Layers (Classification Head Layers): Above the base model, a GlobalAveragePooling2D layer is added to summarize spatial features, followed by Dropout(0.3) for regularization, Dense(128, activation='relu') as a hidden layer, an additional Dropout(0.3) to prevent overfitting, and a Dense(1, activation='sigmoid') output layer for binary classification (real or fake).

Muhammad Erico Revaldo: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 106-116
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1386

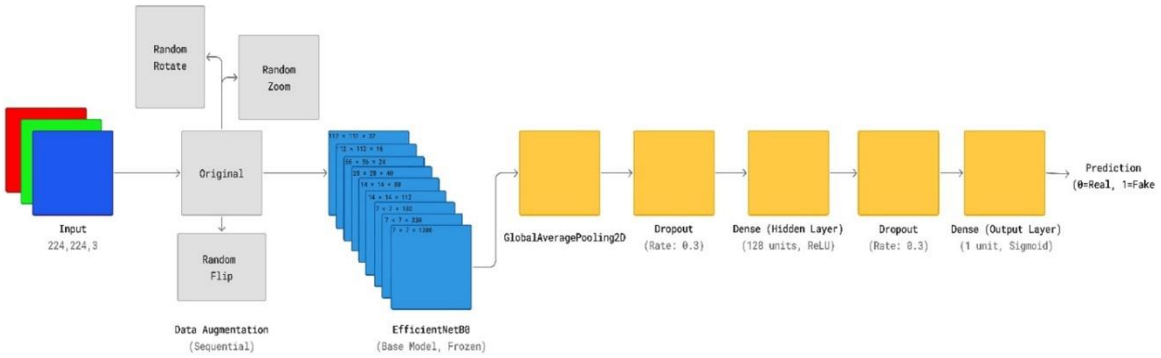

Figure 2. EfficientNetB0 Transfer Learning-based CNN Model Architecture

This design was chosen because EfficientNetB0 incorporates compound scaling techniques, an approach that proportionally balances the depth, width, and resolution of the network. With these characteristics, the model can achieve high accuracy without a significant increase in complexity, making it ideal for implementation in web-based prototypes that require efficient inference time.

2.4 Training Strategy

The model training phase was carried out using a two-stage training strategy to achieve optimal stability and performance. This approach was chosen because Transfer Learning with pre-trained models such as EfficientNetB0 requires a gradual adaptation process so that features learned from the ImageNet dataset can adjust to the characteristics of the new deepfake data without losing general knowledge. In general, the training strategy is carried out as follows:

1. Feature Extraction: In the initial stage, all layers of the EfficientNetB0 base model are frozen so that only the head layer is trained. The goal is for the new layer to adapt to the general features already extracted by the base model. The model is compiled using the Adam optimizer with a standard learning rate (1e-3) and trained for 10–15 epochs until convergence.
2. Fine-Tuning: Once the head layer is stable, the top layer (approximately 20%) of the base model is unfrozen for retraining. The model is then recompiled with a much smaller learning rate (1e-5) so that adjustments are made smoothly without damaging the initial weights. This stage allows the model to increase its sensitivity to specific patterns in deepfake images.
   During the training process, two main callbacks are applied: EarlyStopping to automatically stop training if performance does not improve, and ModelCheckpoint to save the best weights (save_best_only=True) based on the highest validation score.

This two-stage approach has proven effective in maintaining a balance between generalization ability and classification accuracy. With this strategy, the model not only learns to recognize subtle differences between real faces and manipulated results, but also maintains performance stability when tested on new data that it has never seen before.

2.5 Evaluation Metrics

The evaluation stage is conducted to comprehensively assess model performance, not only in terms of accuracy but also in terms of balance, reliability, and generalization capabilities. A comprehensive evaluation approach is very important because deepfake

Muhammad Erico Revaldo: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 106-116
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1386

detection models should not only focus on classification accuracy but also be able to maintain performance stability on new data that has never been seen before (unseen data).

Model performance is assessed using five main metrics commonly used in image detection and cybersecurity research, namely:

1. Accuracy: Measures the percentage of correct predictions from the entire test data. This metric provides an overview of the model's success rate.
2. F1-Score (Macro): The harmonic mean between precision and recall of both classes (real and fake). The macro average value is used to balance the influence between classes, especially if the dataset is imbalanced.
3. ROC-AUC (Receiver Operating Characteristic – Area Under the Curve): Measures the model's ability to distinguish between positive and negative classes as a whole. An AUC value close to 1 indicates excellent classification performance.
4. Average Precision (AP) / PR-AUC: Measures the performance of the model on positive classes (fake) by considering the relationship between precision and recall. This metric is very important in digital security scenarios that are sensitive to detection errors.
5. TPR@FPR=1%: Measures the True Positive Rate (Recall) at a strict security threshold (when the False Positive Rate is limited to only 1%). This metric is used to assess the reliability of the system in real-world contexts that require a high level of confidence.

Through the use of these five metrics, model performance can be evaluated objectively and comprehensively. The evaluation not only ensures that the model is capable of detecting deepfakes with high accuracy, but also assesses the extent to which the model is reliable in extreme conditions relevant to real-world applications in visual disinformation detection systems.

3. RESULT AND DISCUSSIONS

This section presents the results of experiments based on the methodology described in Chapter 2. The results are presented in the form of quantitative analysis through model performance evaluation metrics and proof of concept of web-based system implementation.

3.1 Quantitative Performance Evaluation Results

After going through two training stages, the best EfficientNetB0 model automatically saved by ModelCheckpoint during the fine-tuning process was evaluated using test data consisting of 14,417 images. This number varies slightly due to the filtering of damaged files during the pre-processing stage. The model's classification results for two classes (real and fake) are summarized in the Confusion Matrix, as shown in Figure 2.

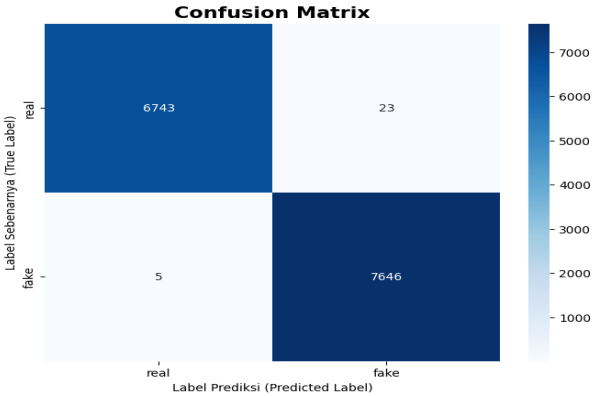

Figure 3. Heatmap Confusion Matrix of evaluation results on test data

Muhammad Erico Revaldo: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 106-116
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1386

From Figure 2, it can be seen that the model produced 6,743 True Negatives (TN), 23 False Positives (FP), 5 False Negatives (FN), and 7,646 True Positives (TP). These values indicate that the model shows exceptional detection performance in distinguishing between authentic and manipulated facial images. In particular, the very low number of False Positives (FP = 23) indicates that the model is able to overcome the issue of classifying non-standard images, such as selfies, as fake images. This was the main objective of this study.

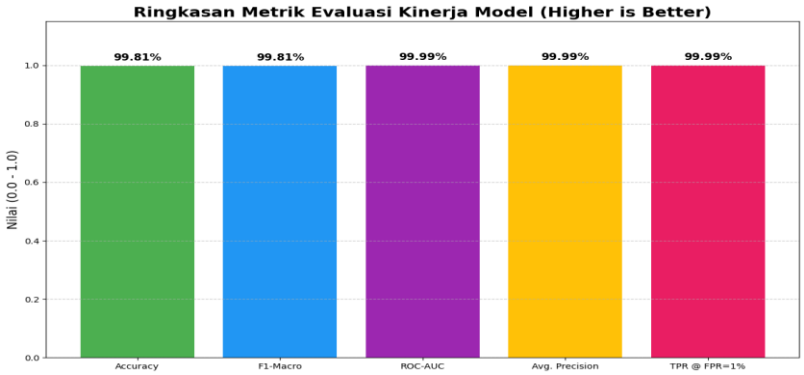

Figure 4. Visualization of model performance metrics summary

Based on the test results, the model achieved an Accuracy of 99.81% and an F1-Score (Macro) of 99.81%, indicating near-perfect performance in two-class classification. The ROC-AUC and Average Precision (AP) values each reached 99.99%, indicating a very high class separation capability, and the False Positive Rate (FPR)=1% was 99.99%, confirming the reliability of the model in strict detection scenarios.

3.2 Discussion of Results

This section discusses the methodological foundations that contributed to the model's superior performance and analyzes its relevance to current challenges in deepfake detection.

Model Performance Analysis Based on False Positive Reduction

The evaluation results presented in the Confusion Matrix (Figure 2) demonstrate a high and consistent level of model performance. The dominance of True Positive (TP) and True Negative (TN) values compared to False Positive (FP = 23) and False Negative (FN = 5) values indicates the model's strong capability in differentiating between authentic and manipulated facial images with a minimal error rate.

A significant reduction in False Positive (FP) cases now only 23 marks a notable improvement over previous studies. In the domain of deepfake detection, False Positive errors, where genuine facial images are misclassified as fake, represent the most severe form of misclassification, as they may undermine system reliability and lead to social or legal consequences for affected individuals. This instability is often attributed to the limitations of commonly used benchmark datasets. The FaceForensics++ dataset, for instance, was built from high-quality YouTube videos that were manually screened and predominantly featured front-facing subjects, while the DFDC dataset used staged videos recorded by paid actors under controlled lighting and camera conditions. These curated datasets fail to capture the unpredictable variations found in real-world, non-standard images such as selfies with diverse angles, lighting, and facial expressions which explains why models trained solely on them often struggle to generalize effectively [22][23].

The improvement achieved in this study stems from the integration of non-standard real images into the training process. By augmenting the dataset with tens of thousands

Muhammad Erico Revaldo: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 106-116
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1386

of real-world selfies exhibiting diverse environmental and facial conditions, the EfficientNetB0 model not only learned to identify synthetic artifacts but also developed a more robust generalization of authentic human facial diversity. This enhancement was instrumental in significantly reducing False Positive occurrences, demonstrating the model's improved adaptability to realistic and unconstrained image conditions that better reflect actual use cases.

Justification for the Success of the Transfer Learning Strategy and EfficientNetB0 Architecture

The application of Transfer Learning in this study is a strategic step, focusing on improving the model's ability to generalize, rather than just overcoming the problem of data shortage. By integrating three extensive and varied public datasets, challenges arise regarding how the model can maintain its performance when encountering significant differences in data (domain shifts). Transfer learning provides a solution by utilizing weights that have been previously trained on ImageNet, a dataset consisting of millions of images. Therefore, the EfficientNetB0 model has a strong foundation for extracting visual features, both simple and complex.

The two-stage fine-tuning approach ensures that the general feature foundation (shallow layers) from ImageNet is preserved, while the deeper layers and classification layers are specifically tailored to distinguish deepfake artifacts. This method significantly speeds up the training process and prevents overfitting on features that may only be visible in a specific dataset. In this way, the trained model becomes more adaptive and resilient to various types of deepfakes, while confirming the principle of the effectiveness of Transfer Learning in improving image classification.

The selection of the EfficientNetB0 architecture is the second crucial element in this success. This architecture design, based on the principle of Compound Scaling, aims to achieve the highest level of accuracy with minimal parameter computation. EfficientNetB0 evenly regulates the depth, width, and resolution of the network, resulting in a resource-efficient model.

In deepfake detection situations that require real-time speed (especially for web-based applications), the efficiency advantages of EfficientNetB0 are crucial. Although larger pre-trained Convolutional Neural Network architectures such as VGG19 or ResNet152 can achieve comparable accuracy, they require longer inference times and significantly greater memory consumption. EfficientNetB0 successfully offers accuracy equivalent to larger models but in a much smaller size. This computational advantage makes the model highly practical and scalable for use in production applications.

Justification and analysis of evaluation metrics

To ensure model quality beyond Accuracy and F1-Score, which depend on thresholds, this study evaluates performance based on three important reliability metrics: ROC-AUC, Average Precision (AP), and TPR@FPR=1%.

The ROC-AUC value of 99.99% indicates nearly flawless class separation across all prediction thresholds. AUC assesses the likelihood that the model will assign a higher value to deepfake samples than to randomly selected real samples. This value, which is very close to 1.0, clearly shows that EfficientNetB0 has understood highly discriminative feature representations, allowing the model to effectively separate the two class distributions without being affected by the standard threshold (0.5).

Furthermore, the Average Precision (AP) reaching 99.99% provides additional crucial confirmation, as this metric is highly sensitive to Recall performance at high Precision levels. AP essentially measures the quality of predictions on the positive class (deepfakes), ensuring that the model is able to maintain high Precision levels while attempting to detect as many deepfakes as possible (increasing Recall). This near-perfect AP value confirms

Muhammad Erico Revaldo: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 106-116
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1386

that the model is not only accurate, but also produces highly calibrated probability outputs, providing high confidence in every case identified as a deepfake.

The peak achievement in reliability metrics occurs at TPR@FPR=1% with a value of 99.99%. In the world of cybersecurity, False Positives (FP) are considered the most costly and potentially reputation-damaging type of error. As a result, systems are often required to operate at thresholds where the False Positive Rate (FPR) is reduced to 1% (0.01). This metric indicates that the model has been tuned to prediction thresholds that limit FP errors to only 1%, while maintaining a very high deepfake detection capability (TPR) of 99.99%. The ability to maintain a high Recall rate under strict Precision conditions (with low FPR) demonstrates that the model is very robust and ready for use in operational environments that require high information integrity assurance.

Overall, the results of the three reliability metrics ROC-AUC, AP, and TPR@FPR=1% confirm that this model is valid even outside ideal testing conditions. The combination of perfect performance on metrics that are not threshold-dependent and that focus on security sets this model apart from traditional research, which often only pays attention to accuracy, and clearly proves that the developed EfficientNetB0 has the generalization and reliability capabilities necessary to reduce the risk of real visual disinformation.

Comparison of results, research novelty, and limitations

The results obtained from this study, particularly the 99.81% accuracy and extremely low FPR rate, place the EfficientNetB0 model in a superior position compared to similar deepfake research trends, which often experience performance degradation related to False Positive metrics or domain shifts. This research specifically addresses the issue of False Positives through a dataset curation method that integrates tens of thousands of non-standard authentic selfies. This is a key feature, as the model was trained with the clear objective of understanding diversity in real images, resulting in very few FPs (only 23 cases) and addressing the research gap identified in Chapter 1. The computational efficiency of the EfficientNetB0 design also makes it a practical, scalable solution. Additionally, the novelty of this research is the implementation of a simple website that stakeholders can use to upload human facial images to be classified as real or fake for the mitigation of disinformation.

Our results also outperform traditional deep learning benchmarks. While a comparative analysis by Ashani et al. [24] reported that VGG19 achieved a peak accuracy of 98%, our EfficientNetB0-based approach reached 99.81% with fewer parameters. This demonstrates that EfficientNetB0's compound scaling provides a more efficient and accurate representation for deepfake artifacts compared to deeper but less optimized architectures.

Although this model is very effective in detecting static images, the limitation of this research lies in its focus, which does not cover temporal and audio aspects. Therefore, future research should focus on multimodal analysis that combines visual features from Convolutional Neural Networks with temporal domain analysis (such as RNN/LSTM) and audio manipulation recognition. In addition, investigating the model's resilience in the face of significant compression and malicious attacks should be a concern in order to improve the maturity and application of deepfake detection models in the future.

Overall, these results prove that the EfficientNetB0-based approach with a two-stage fine-tuning strategy and a balanced combined dataset can produce an accurate, stable, and reliable deepfake detection model for mitigating visual disinformation.

3.3 Prototype Implementation

As a final step and proof of concept, the best EfficientNetB0 model that has undergone training and evaluation is then implemented into a prototype web application based on the Django framework. This implementation aims to prove that the model is not only theoretically superior, but can also be practically applied in real systems to detect deepfake-based facial image manipulation.

Muhammad Erico Revaldo: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 106-116
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1386

On the backend, the trained model is loaded using the TensorFlow library when the server is running. The prediction process is performed automatically every time a user uploads an image through the web interface. Uploaded images first go through a preprocessing stage, which includes resizing to 224×224 pixels, normalizing pixel values according to the EfficientNet preprocessing scheme, and transforming them into a tensor format compatible with the model input.

Meanwhile, on the frontend, the application is designed using a combination of HTML, CSS, and JavaScript to create a simple yet interactive interface. Users can upload images of human faces through the main page, and the system will display the classification results in real-time in the form of "REAL" or "FAKE (AI-generated)" labels accompanied by prediction probability values.

The implementation results show that the system successfully detects images with an accuracy level in line with the previous quantitative evaluation results. In direct testing, real face images were detected as REAL with a confidence level of 100%, while deepfake manipulated images were successfully classified as FAKE (AI-generated) with a confidence level of 100%. The web interface display showing the prediction results is presented in Figure 4.

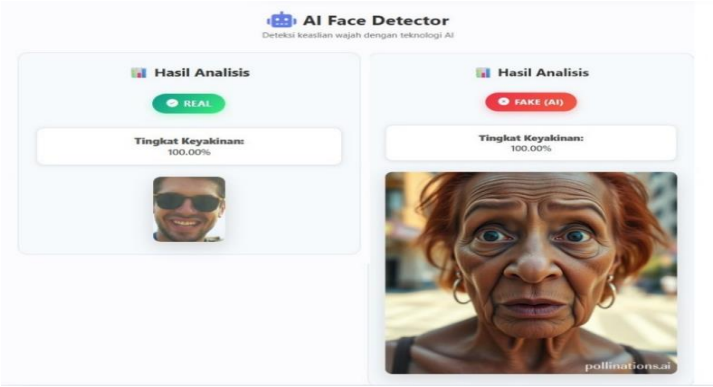

Figure 5. Deepfake Detection Web Prototype Interface Display

The implementation of this prototype successfully proves that the developed deep learning model has high scalability and applicability. The integration between the CNN and Django models can run stably without significant latency, even when processing large images. Prediction results that are consistent with quantitative evaluations reinforce that this model is suitable for use as the foundation of a web-based deepfake detection system.

The success of this implementation also confirms the contribution of research in building a bridge between theory and practice. Namely, the CNN design uses EfficientNetB0, which allows inference (prediction) activities to be carried out in milliseconds. This is crucial for applications that require real-time speed and provides opportunities for the development of detection APIs that can be widely integrated into various digital platforms.

4. CONCLUSION

This study successfully developed and evaluated a deep learning model based on the EfficientNetB0 architecture for classifying real and deepfake facial images as an effort to mitigate digital disinformation. Through a transfer learning approach with a two-stage training strategy (feature extraction and fine-tuning), the proposed model demonstrated a very high and stable level of performance across multiple key evaluation metrics, confirming its effectiveness in detecting AI-generated facial manipulations.

The research process was conducted systematically, starting from the data acquisition and curation phase using three large-scale public datasets to construct a balanced and diverse data corpus. Pre-processing steps were performed to ensure optimal data quality,

Muhammad Erico Revaldo: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 106-116
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1386

followed by model training using an efficient TensorFlow pipeline. The evaluation results showed an accuracy of 99.81%, a Macro F1-Score of 99.81%, and ROC-AUC and Average Precision (AP) values of 99.99%, with a True Positive Rate (TPR) of 99.99% at FPR=1%, confirming that the model provides high reliability for image-based security detection applications.

In addition to its superior quantitative performance, the implementation of the Django-based web prototype proved the model's practical applicability. The system can process facial images in real time and provide classification results of "REAL" or "FAKE (AI-generated)" with a high level of confidence. The integration of the Convolutional Neural Network (CNN) with the web framework demonstrated stable performance without significant latency, proving that the proposed model is not only theoretically robust but also efficient in real-world deployment scenarios.

The success of this research confirms that modern deep learning approaches can be effectively adapted for digital forensics and cybersecurity purposes. The developed model is not only relevant for facial manipulation detection but also holds potential applications in biometric identity verification, digital document authentication, and privacy protection on social media platforms. Hence, this study contributes significantly to the development of reliable and adaptive deepfake detection systems while opening opportunities for future research on multimodal (image and video) detection and advanced architectures such as Vision Transformer (ViT) or Swin Transformer to further enhance model generalization and efficiency.

5. ACKNOWLEDGMENT

The author would like to thank Bina Sarana Informatika University for the facilities, guidance, and opportunities provided during this research process. Thanks are also extended to the supervising lecturer and teammates who contributed to data collection, model training, and the implementation of a web-based prototype, as well as to the open source community that provided datasets and frameworks such as TensorFlow, Keras, and EfficientNet, which enabled this research to be carried out successfully. The author hopes that the results of this research can make a positive contribution to the development of knowledge in the field of deepfake detection and inspire further research into the mitigation of digital disinformation in the future.

6. REFERENCES

[1] J. Rao and T. Uehara, "A Chronological Review of Deepfake Detection : Techniques and Evolutions," vol. 15, pp. 23–49, 2025, doi: 10.5121/csit.2024.150202.

[2] R. Tolosana, R. Vera-rodriguez, J. Fierrez, A. Morales, and J. Ortega-garcia, "DeepFakes and Beyond : A Survey of Face Manipulation and Fake Detection," vol. 64, no. December, pp. 131–148, 2020.

[3] C. Vaccari and A. Chadwick, "Deepfakes and Disinformation : Exploring the Impact of Synthetic Political Video on Deception , Uncertainty , and Trust in News," *Soc. Media + Soc.*, pp. 1–13, 2020, doi: 10.1177/2056305120903408.

[4] K. F. Macdorman, A. Diel, T. Lalgi, I. Carolin, M. Teufel, and B. Alexander, "Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers," *Comput. Hum. Behav. Reports J.*, vol. 16, no. August, pp. 1–13, 2024, doi: 10.1016/j.chbr.2024.100538.

[5] M. R. Shoaib, Z. Wang, M. T. Ahvanooey, and J. Zhao, "Deepfakes , Misinformation , and Disinformation in the Era of Frontier AI , Generative AI , and Large AI Models," pp. 1–8, 2023.

[6] E. P. Boediman, "Exploring the impact of deepfake technology on public trust and media manipulation : A scoping review," *jurnal-komunikasi Explor.*, vol. 19, pp. 312–334, 2025, doi: 10.20885/komunikasi.vol19.iss2.art8.

[7] A. Domenteanu, T. George-cristian, L. Cr, A. Mol, L.-A. Cotfas, and C. Delcea, "Living

Muhammad Erico Revaldo: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 106-116
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.v20i2.1386

in the Age of Deepfakes : A Bibliometric Exploration of Trends , Challenges , and Detection Approaches," *information*, pp. 1–31, 2024.

[8] S. M. Qureshi, A. Saeed, S. H. Almotiri, F. Ahmad, and M. A. Al Ghamdi, "Deepfake forensics : a survey of digital forensic methods for multimodal deepfake identification on social media," *PeerJ Comput. Sci.*, pp. 1–40, 2024, doi: 10.7717/peerj-cs.2037.

[9] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, *A review of convolutional neural networks in computer vision*, vol. 57, no. 4. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10721-6.

[10] J. Mu, M. Adrezo, and A. N. Haikal, "Identifikasi Wajah Asli dan Buatan Deepfake Menggunakan Metode Convolutional Neural Network," vol. 13, no. 1, pp. 45–50, 2024, doi: 10.34148/teknika.v13i1.705.

[11] F. A. Jiwani1 and B. S. W. Poetro, "SISTEM DETEKSI GAMBAR DEEPFAKE MENGGUNAKAN CNN DENSENET-121 DENGAN WATERMARKING LEAST SIGNIFICANT BIT (LSB)," *J. Rekayasa Sist. Inf. dan Teknol.*, vol. 2, no. 3, pp. 1157–1170, 2025.

[12] M. S. Momin, A. Sufian, D. Barman, M. Leo, C. Distante, and N. Damer, "Explainable deepfake detection across different modalities : An overview of methods and challenges," *Image Vis. Comput.*, vol. 163, no. August, p. 105738, 2025, doi: 10.1016/j.imavis.2025.105738.

[13] S. Singh and A. Dhumane, "Unmasking digital deceptions: An integrative review of deepfake detection, multimedia forensics, and cybersecurity challenges," *MethodsX*, vol. 15, p. 103632, 2025, doi: 10.1016/j.mex.2025.103632.

[14] R. Apriliansyah, A. Handayanto, and N. D. Saputro, "Application of EfficientNetB0 on Convolutional Neural Network Model for Early Detection of Cataract," *J. Comput. Eng. Syst. Sci.*, vol. 10, no. 2, pp. 435–446, 2025.

[15] M. Abbasi, P. Váz, and J. Silva, "Comprehensive Evaluation of Deepfake Detection Models : Accuracy , Generalization , and Resilience to Adversarial Attacks," *Appl. Sci.*, pp. 1–16, 2025.

[16] H. Almirza, S. Sanjaya, L. Handayani, and F. Syafria, "Klasifikasi Daging Sapi dan Daging Babi Menggunakan Convolutional Neural Network EfficientNet-B0 dengan Augmentasi Citra," vol. 3, no. 6, pp. 1013–1021, 2023, doi: 10.30865/klik.v3i6.910.

[17] M. Tan and Q. V Le, "EfficientNet : Rethinking Model Scaling for Convolutional Neural Networks," 2020.

[18] A. E. Putra, M. F. Naufal, and V. R. Prasetyo, "Klasifikasi Jenis Rempah Menggunakan Convolutional Neural Network dan Transfer Learning," vol. 9, no. 1, 2023.

[19] O. Akinsanmi, A. Sobowale, B. A. Omodunbi, and A. Soladoye, "Detection of Image-Based Deepfake using Deep Transfer Learning Algorithms," *FUOYE J. Eng. Technol.*, vol. 10, no. 2, pp. 0–5, 2025.

[20] E. Altuncu, V. N. L. Franqueira, and S. Li, "Deepfake : Definitions , Performance Metrics and Standards , Datasets and Benchmarks , and a Meta-Review," pp. 1–31, 2022.

[21] P. Rai and P. Kanungo, "A Robust CNN-Siamese Framework for Iris Deepfake Spoof Detection with Superior Accuracy and AUC," *J. Inf. Syst. Eng. Manag.*, vol. 10, pp. 816–833, 2025.

[22] R. Andreas, D. Cozzolino, L. Verdoliva, J. Thies, M. Nießner, and C. Riess, "FaceForensics ++ : Learning to Detect Manipulated Facial Images," pp. 1–11.

[23] B. Dolhansky *et al.*, "The DeepFake Detection Challenge (DFDC) Dataset," pp. 1–13, 2020.

[24] Z. N. Ashani, I. S. C. Ilias, K. Y. Ng, M. R. K. Ariffin, A. D. Jarno, and N. Z. Zamri, "Comparative Analysis of Deepfake Image Detection Method Using VGG16, VGG19 and ResNet50," *J. Adv. Res. Appl. Sci. Eng. Technol.*, vol. 47, no. 1, pp. 16–28, 2025, doi: 10.37934/araset.47.1.1628.

Muhammad Erico Revaldo: *Corresponding Author