BINUS UNIVERSITY GRADUATE PROGRAM Master Program

Master of Computer Science

# Detection of Deepfake in Social Engineering Attacks Targeting Construction Work Competency Certificates (SKK-Konstruksi) via the Ministry of Public Works Licensing Portal

## Background

With the rise of social engineering attacks targeting the SKK-Konstruksi certification process through the Ministry of Public Works' online licensing portal, the detection of deepfake content has become increasingly critical. This research proposes a deep learning-based framework to counter such threats by combining face detection YOLO v8 and classification techniques CNN. To enhance transparency and trust in the model's decisions, Explainable AI (XAI) is integrated using Grad-CAM and LIME, providing visual insights into which regions influenced the classification.

## Previous Works

The growing risks associated with deepfakes have prompted extensive research on detection strategies. Prior studies have utilized deep learning for classification tasks, such as Nawaz et al., who employed ResNet-Swish-Dense54 on the FaceForensics++ dataset, achieving 99.88% accuracy. Others, like Abir et al., evaluated multiple CNN architectures (e.g., InceptionV3, ResNet152V2) with interpretable models such as LIME, reporting accuracies above 99%. Furthermore, Ismail et al. introduced Yolo-CNN-Boost, combining YOLO for facial region detection with XGBoost for classification, achieving 93.53% accuracy on the FF++ dataset.

## Result & Discussion

In this research, the selection of the convolutional neural network models are compared to get the best CNN model to detect deepfake, ResNet50V2, InceptionResNetV2, and Xception are employed as base models (without pre-trained layers).

Table 1. Model Performance

| Metric | ResNet50V2 | Inception ResNetV2 | Xception |
|---|---|---|---|
| Accuracy | 91.69% | 90.75% | 91.03% |
| Val. Accuracy | 92.14% | 90.88% | 91.73% |
| Loss | 22.89% | 25.38% | 24.35% |
| Val. Loss | 21.32% | 24.89% | 22.92% |
| Precision | 75.80% | 69.70% | 78.17% |
| Recall | 31.45% | 22.13% | 21.11% |
| F1-Score | 44.45% | 33.59% | 33.24% |

Based on data in Table 1, ResNet50V2 demonstrates the best overall performance for deepfake detection, achieving the highest validation accuracy (92.14%) and lowest validation loss (21.32%).

It also leads in recall (31.45%) and F1-score (44.45%), suggesting better generalization in identifying deepfakes despite the inherent challenge of class imbalance (common in deepfake datasets).

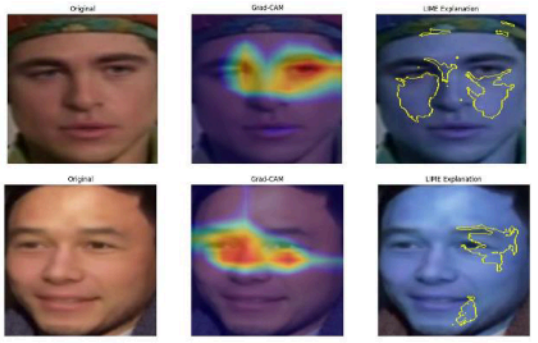



Figure 1. Grad-CAM & LIME Visualization

In this research, two explainable AI techniques, Grad-CAM (Gradient-weighted Class Activation Mapping) and LIME (Local Interpretable Model-Agnostic Explanations) are used to identify the areas in an image to classify whether an image is a deepfake or not as shown in Figure 1.

Based on our comparative analysis, Grad-CAM is the more effective XAI method for deepfake detection tasks. It offers clearer and more focused insights into how the model identifies deepfake artifacts, making it a valuable tool for improving model transparency, supporting forensic investigations, and building trust in automated deepfake detection systems.

## Methodology

The research methodology involves a three-stage process, in the first stage YOLO is employed for data pre-processing and facial detection, followed by second stage is the implementation of three distinct CNN architectures, ResNet50V2, InceptionResNetV2, and Xception to classify deepfake images. The last stage is Explainable AI implementation followed by XAI Models Evaluation to compare which XAI model runs better to interpret deepfake images.

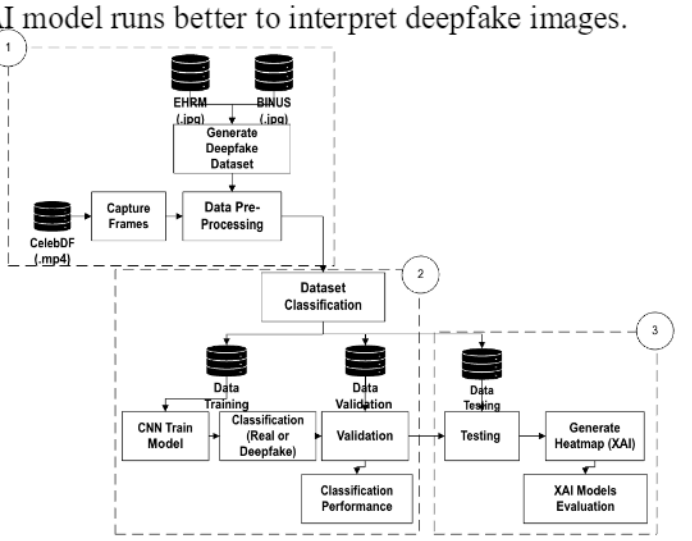
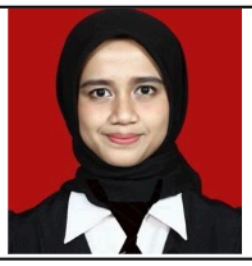



Figure 2. Proposed Method

Figure 2 shows the proposed method carries out the data collection continue to data pre-processing using YOLO. The detected object images then processed with CNN to classify deepfake in the images that creating result of the classification. After combining the primary (EHRM & BINUS) and secondary (CelebDF) datasets, all images are classified into two labels: real or deepfake. Then, before training the model, the dataset is divided into three parts: training data, validation data, and testing data. The training set is used to train CNN models, ResNet50V2, InceptionResNetV2, and Xception. Last step, the data testing is processed and generated in Explainable AI to interpret area that influenced deepfake. This research uses 2 XAI models, Grad-CAM and LIME to be compared in model evaluation stage.


Prof. Benfano Soewito, M.Sc., Ph.D.
bsoewito@binus.edu
Computer Science Department,
BINUS Graduate Program, Master of Computer Science



Gifariani
gifariani@binus.ac.id
Computer Science Department,
BINUS Graduate Program, Master of Computer Science


## Conclusion

Among the CNN evaluated models, ResNet50V2 demonstrated the most promising performance, achieving 91.69% accuracy, 22.89% validation loss, 75.80% precision, and a 44.45% F1-score. While the accuracy is notably high, the relatively low F1-score suggests the model faces challenges in consistently detecting real instances, particularly in avoiding false positives. This could be attributed to class imbalance between real and deepfake samples, variations in training data quality, or the inherent complexity of deepfake patterns that are difficult for the model to capture. In an evaluation comparing 30 deepfake images with LIME, Grad-CAM successfully highlighted relevant manipulated areas in 27 cases, demonstrating its effectiveness in offering human-understandable explanations.