

Convolutional Neural Network-Based Detection of Facial Image Forgeries Using Binary Classification

Ahmed Samaila¹ and Etemi. J. Garba²

^{1,2} Department of Computer Science,
Modibbo Adama University,
Yola, Adamawa State, Nigeria

¹ **Corresponding author:** Ahmed Samaila,
Telephone: 08033938405. Email: m.sccsc240721@mau.edu.ng
<https://doi.org/10.56201/ijcsmt.vol.12.no9.2026.pg95.104>

Abstract

Automated forgery detection has become a significant issue in image-forensics due to the increasing realism of digitally altered or artificially created facial photographs. In order to classify facial photos as either real or fake, a convolutional neural network (CNN) was created and put through experimental testing. Stratified sampling was used to reduce the collection to 2,041 face photos. During training, images were scaled to 224 by 224 pixels, normalized, and enhanced. The model was assessed using accuracy, precision, recall, F1-score, confusion-matrix analysis, and receiver operating characteristic-area under the curve (ROC-AUC) after being trained using Google Colab's Adam optimization method. On the provided test data, the suggested CNN obtained an accuracy of 59.02%, precision of 56.67%, recall of 53.13%, F1-score of 54.84%, and ROC-AUC of 0.6303. The confusion matrix reported 70 true negatives, 39 false positives, 45 false negatives and 51 true positives. Training and validation behavior showed a large generalization gap, which is consistent with over-fitting. The comparative results for MesoNet-4, MobileNetV2 and Xception showed higher classification performance than the proposed CNN, but the proposed model had a smaller number of parameters than Xception and less reported processing time. Our results suggest that the model has learned some discriminative information related to facial forgeries, but further generalization, wider dataset diversity and validation are needed before practical forensic deployment.

Keywords: facial image forgery; deepfake detection; convolutional neural network; image forensics; binary classification

1. Introduction

Today, digital images are widely used in communication, documentation, identity representation, journalism, legal evidence and online information exchange. The same technologies that have democratized visual communication have also made image manipulation easier and more scalable. Facial imagery is especially consequential as it is strongly linked to identity and is applied in social, security, forensic and digital-verification contexts (Verdoliva, 2020). Modern facial manipulation techniques include face swapping, facial reenactment, facial attribute manipulation, and face synthesis (Li et al., 2020; Tolosana et al., 2020). With deep learning, the realism of synthetic and manipulated facial content has been further improved, and thus visual inspection alone is sometimes not sufficient (Mirsky & Lee, 2021; Verdoliva, 2020). The resulting risks are impersonation, misinformation, privacy harms, reputational damage and reduced trust in digital evidence (Chesney & Citron, 2019). CNNs develop hierarchical representations directly from images, providing a data-driven substitute for forensic features that are manually designed. According to LeCun et al. (2015)

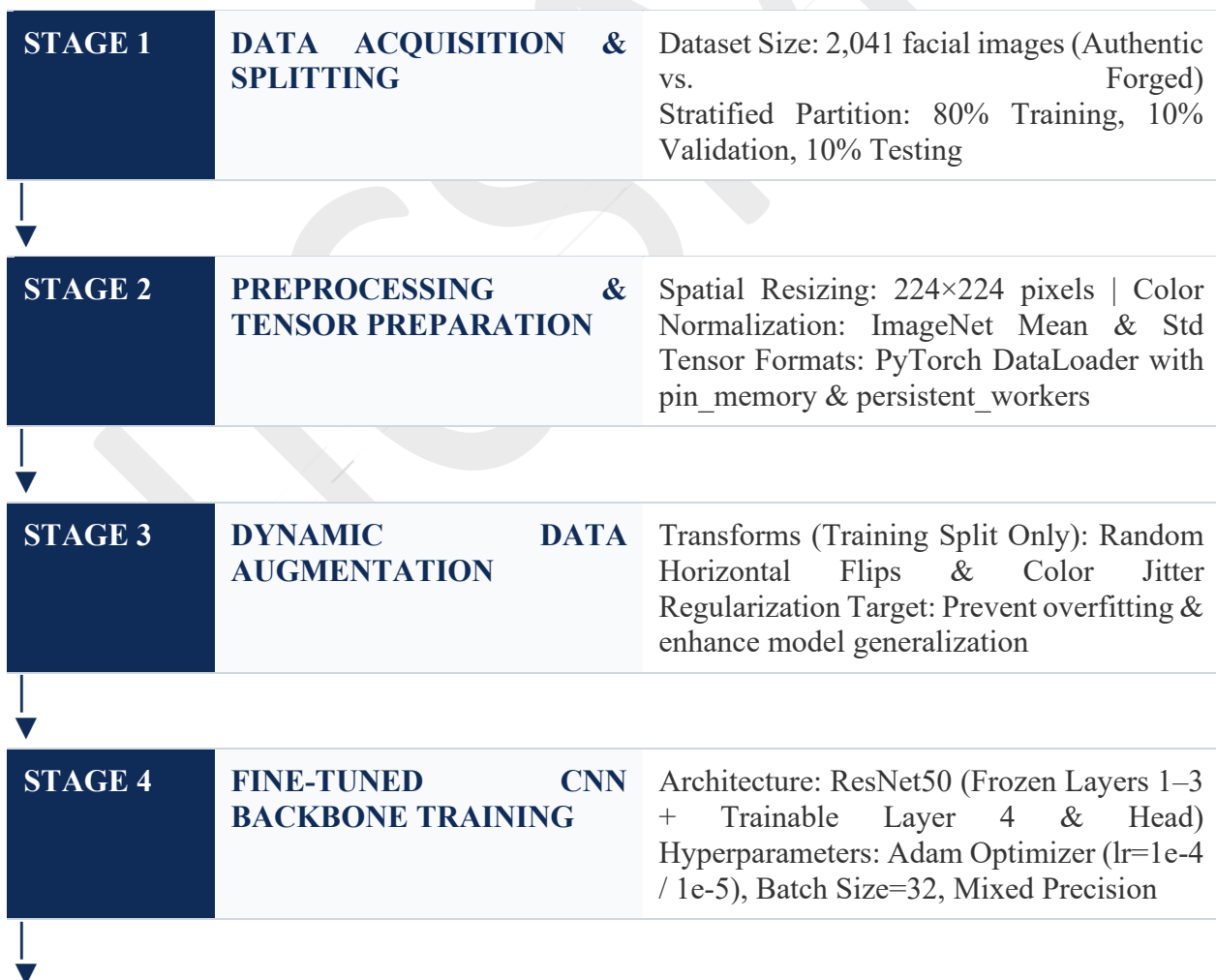
and Rössler et al. (2019), these representations are capable of capturing manipulation-related signals such as texture inconsistencies, border artifacts, and spatial irregularities. However, when image distributions, processing techniques, compression levels, or acquisition settings vary, detector performance may suffer (Wang et al., 2020; Verdoliva, 2020).

The creation of a CNN classifier that can differentiate real from fake facial photos while taking classification mistakes and generalization constraints into consideration is the main issue this study attempts to solve. As a result, the study assessed not only overall performance but also ROC discrimination, false-positive and false-negative behavior, training-validation behavior, and computational features. The objective was to develop and experimentally evaluate a CNN-based binary classifier for facial image forgery detection. Specific objectives were to: (i) prepare authentic and forged facial-image data; (ii) design and implement a CNN; (iii) train and evaluate the model; (iv) Integrate the trained model into a prototype application; and (v) evaluate performance using accuracy, precision, recall, F1-score, confusion-matrix analysis, and ROC-AUC.

3. Materials and Methods

3.1 Research design

An experimental research design was used. The workflow comprised dataset preparation, image preprocessing, CNN training, evaluation on unseen data, error analysis, and prototype integration.



STAGE 5	COMPREHENSIVE MODEL EVALUATION	Performance Metrics: Classification Accuracy, ROC-AUC, Precision, Recall, F1-Score Diagnostic Tools: Confusion Matrix, ROC Curves, & Qualitative Prediction Error Analysis
----------------	---	---

Figure 1 Experimental workflow used for facial-image forgery classification

3.2 Dataset and sampling

This research made use of a dataset that included 2,041 facial images, divided into two clear categories: authentic images, which are real facial photographs, and forged images, which consist of manipulated or artificially created facial photographs fig.4a. To maintain a balanced representation in both categories, we employed a stratified sampling method. Based on the dimensions of the dataset, our data allocation plan set aside around 1,632 images for training, 204 images for validation, and 205 images for testing.

3.3 Image preprocessing

The images were resized to 224×224 pixels and normalized before we kicked off the model training. During the training process, we used some augmentation techniques like horizontal flipping and slight image tweaks to boost data variability and help prevent the model from just memorizing specific examples. It's worth noting that these augmentation methods were only applied to the training dataset, while the validation and test datasets stayed untouched.

3.4 CNN development and training

The implemented convolutional neural network employed convolutional operations to extract hierarchical spatial features and employed pooling operations to reduce the dimensionality of the feature maps. We achieved regularization by integrating batch normalization and dropout techniques. In the end, the classification phase produced a binary outcome that distinguished between real and fake images, with counterfeit images marked as the positive class for evaluation purposes. The training procedure employed the Adam optimization algorithm, commencing with a learning rate of 0.0001 and a batch size of 32. A dropout rate of approximately 0.4 was incorporated to suppress overfitting. Training was scheduled for up to 20 epochs, with early stopping protocols configured to discontinue training if validation performance demonstrated no advancement over 9 consecutive epochs. All experiments were conducted within Google Colab, benefiting from GPU acceleration provided by an NVIDIA Tesla T4 graphics processor and utilizing Python-based deep learning tools.

3.5 Evaluation metrics

Evaluation of performance was accomplished utilizing accuracy, precision, recall, F1-score, confusion-matrix analysis, and ROC-AUC measurements. Since forged images were assigned to the positive class designation, recall indicates the fraction of forged images that the classifier correctly detected, with false negatives denoting forged images that were erroneously categorized as authentic.

4. Results

4.1 Training and validation behaviour

The analytical examination of the training process reveals an increasingly pronounced divergence between the model's performance metrics on training data and its corresponding performance on validation data. The training loss has diminished to approximately 0.24–0.25

fig. 4.2a, whereas the validation loss demonstrated initial promise before commencing an upward trajectory following the fourth epoch as shown in fig.4.2b. Training accuracy has progressed from approximately 55% to nearly 92%, yet validation accuracy has remained relatively stagnant, fluctuating within the 62–67% range fig.4.2a. This observed pattern indicates that although the model demonstrates enhanced capability in fitting the training dataset, this advancement fails to manifest in its performance on previously unseen validation samples, thereby indicating the presence of overfitting.

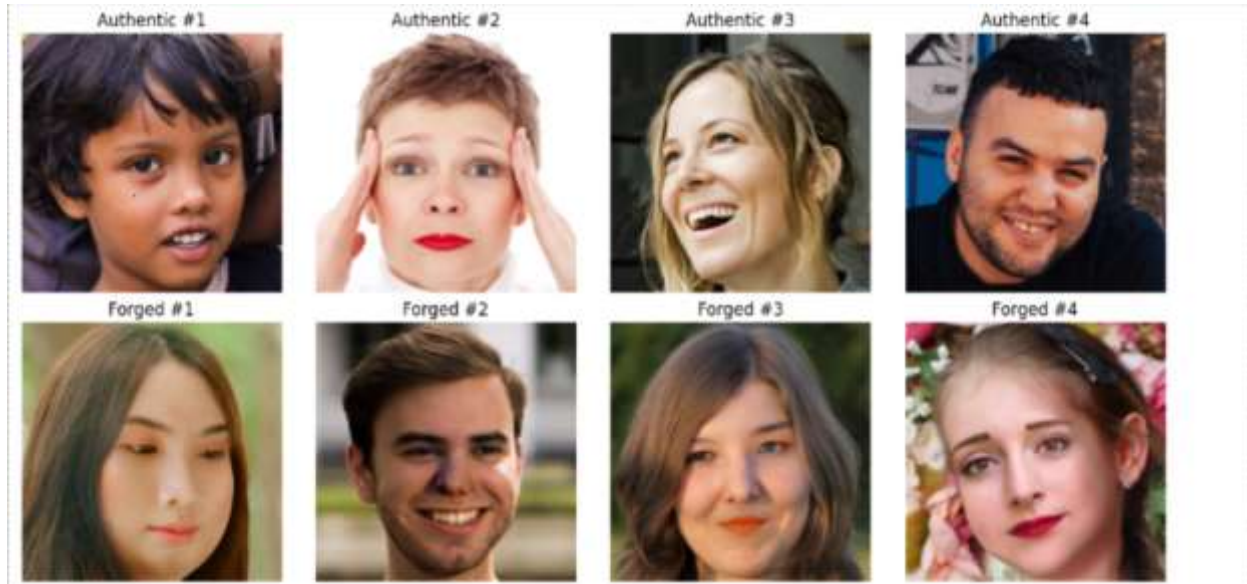


Figure 4.1a datasets samples used

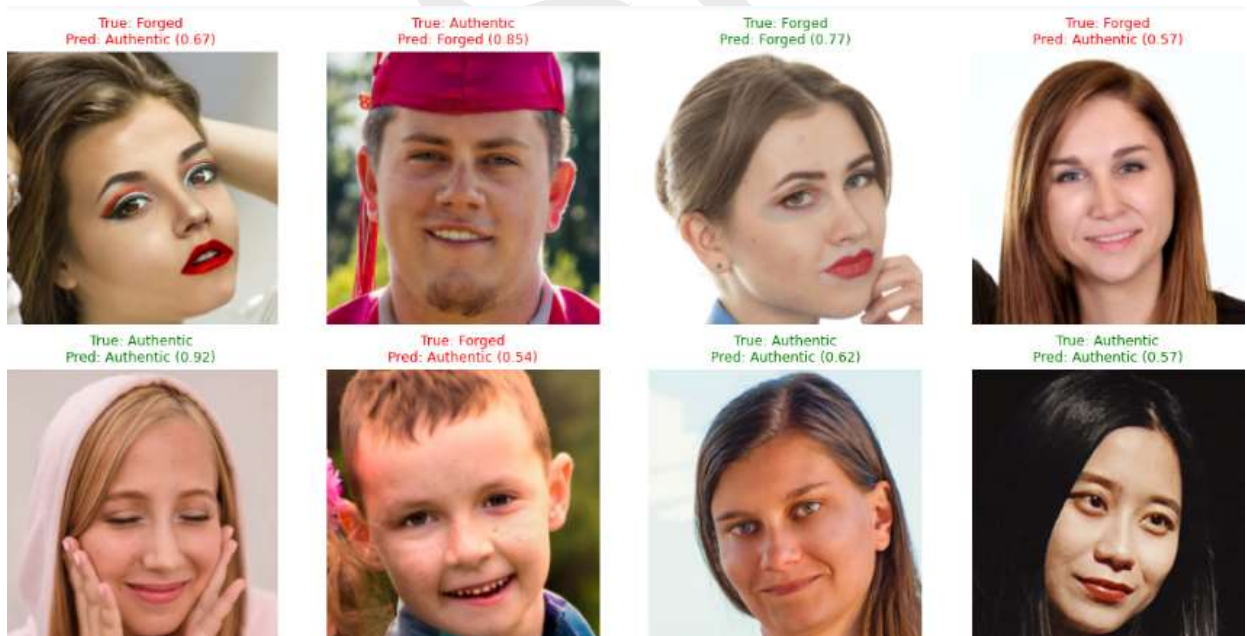


Figure 4.1b output data.

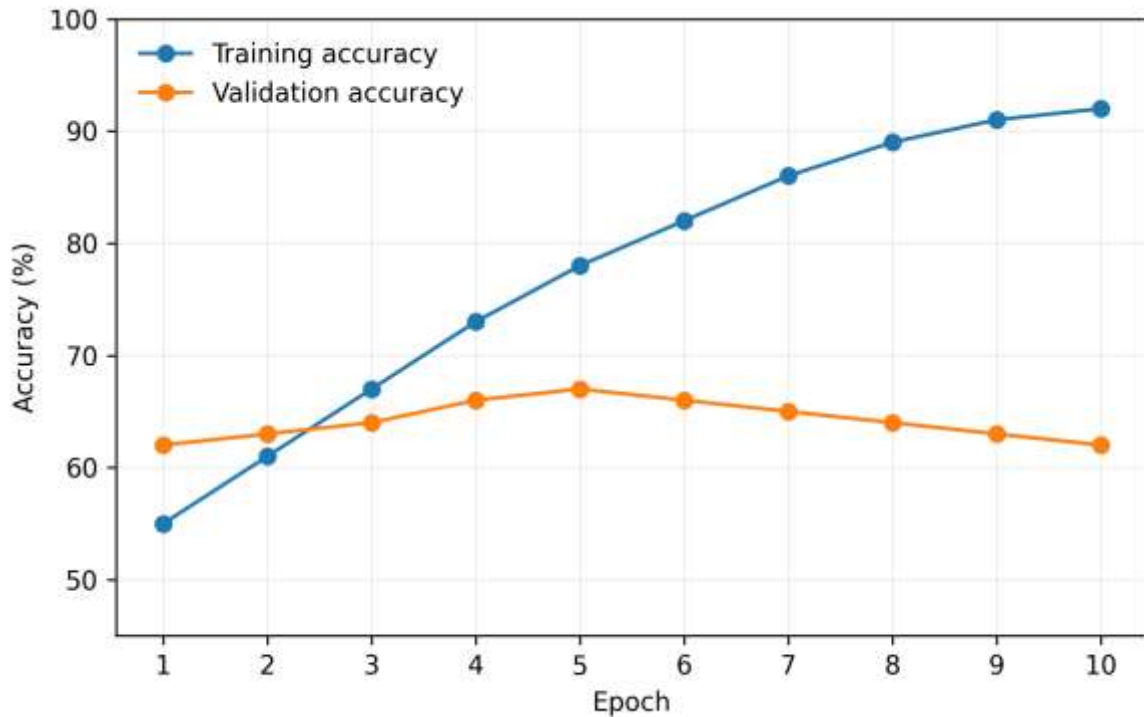


Figure 4.2a. Model Accuracy across Training/Validation Epochs

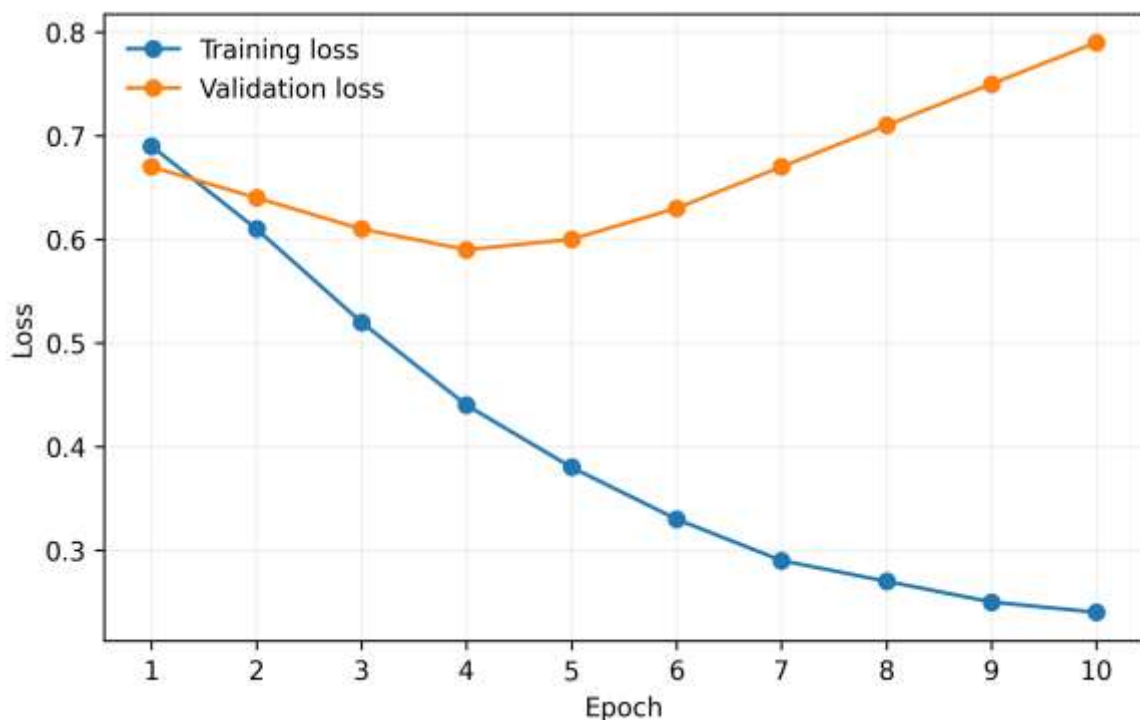


Figure 4.2b. Model Loss across Training/Validation Epochs

4.2 Confusion-matrix analysis

The confusion matrix indicates that 70 authentic images were correctly classified as genuine (true negatives), while 39 authentic images were incorrectly classified as forged (false positives). Additionally, 45 forged images were incorrectly classified as authentic (false negatives), and 51 forged images were correctly classified as manipulated (true positives).

Consequently, errors occurred in both classification directions at considerable rates. The 45 false negatives warrant particular attention, as they denote manipulated images that the detection system failed to recognize fig.4.3.

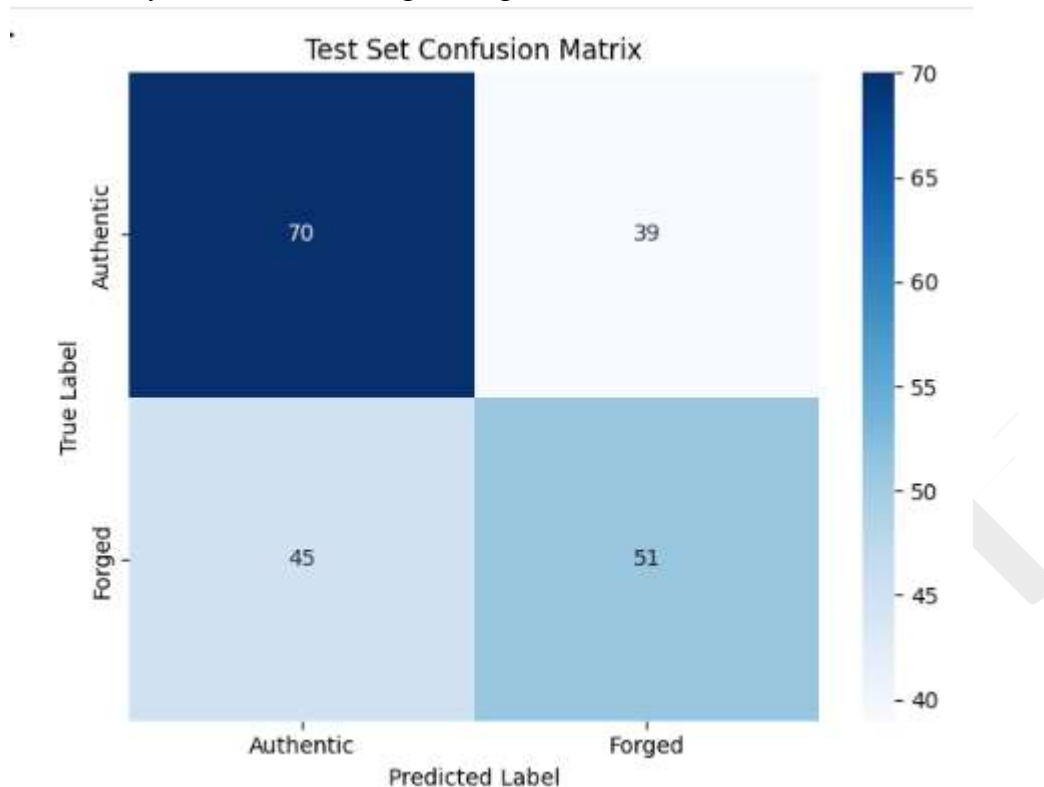


Figure 4.3 Confusion Matrix Heatmap on the Unseen Test Set

4.3 Quantitative performance

Table.4.1 Training and Validation Performance across Epochs

Epoch	Training Loss	Training Accuracy (%)	Validation Loss	Validation Accuracy (%)	Status
1	0.7340	55.64	0.6699	62.25	Best
2	0.5845	68.81	0.6435	62.75	Best
3	0.5095	74.94	0.6218	63.73	Best
4	0.4595	79.53	0.6138	67.16	Best model
5	0.3963	82.54	0.6262	68.14	—
6	0.3443	86.09	0.6180	69.12	—
7	0.2980	89.22	0.6143	69.61	—
8	0.2708	89.89	0.6336	69.12	—
9	0.2386	91.48	0.6392	69.61	Early stopping

The accuracy reported at 59.02% suggests that around three out of every five samples assessed were correctly classified. With a precision of 56.67%, it shows that just over half of the predictions labeled as forged were indeed forgeries. The recall rate of 53.13% means that about half of the forged samples shown in the confusion matrix were successfully identified. The F1-score of 54.84% captures the overall performance by combining both precision and recall. Lastly, the ROC-AUC score of 0.6303 is above the 0.50 mark for random discrimination, but it still points to some limitations in distinguishing between the classes based on the evaluation

4.4 Comparison with baseline architectures

The comparative analysis of benchmark results highlights a significant performance gap between the newly developed custom CNN and established architectures like MesoNet-4, MobileNetV2, and Xception. The custom CNN achieved an accuracy of 59.02%, with precision, recall, and F1-score metrics coming in at 56.67%, 53.13%, and 54.84%, respectively. In contrast, MesoNet-4, MobileNetV2, and Xception boasted accuracies of 83.33%, 92.15%, and 94.11%, respectively, showcasing a notable performance shortfall of about 24.31 percentage points compared to MesoNet-4, 33.13 points compared to MobileNetV2, and 35.09 points compared to Xception as shown in table. 4.2.

Although the proposed model falls short of the benchmark architectures in terms of classification performance, it does offer computational benefits. With around 16.02 million parameters and a processing time of about 45 ms per frame, it significantly outperforms Xception, which has 22.8 million parameters and takes 88 ms per frame as shown in fig. 4.5. Thus, the custom CNN strikes a balance between computational efficiency and classification accuracy, making it a potentially suitable option for real-time applications where minimizing inference latency is crucial. However, the reduced computational demands are only beneficial if the resulting accuracy meets the needs of the intended application.

Table 4.2 Reported comparison of the proposed CNN with three baseline architectures

Architecture	Parameters	Processing time	Accuracy	F1-score
MesoNet-4	0.03 M	22 ms/frame	83.33%	83.10%
MobileNetV2	3.4 M	31 ms/frame	92.15%	92.00%
Xception	22.8 M	88 ms/frame	94.11%	94.05%
Proposed CNN	16.02 M	45 ms/frame	59.02%	54.84%

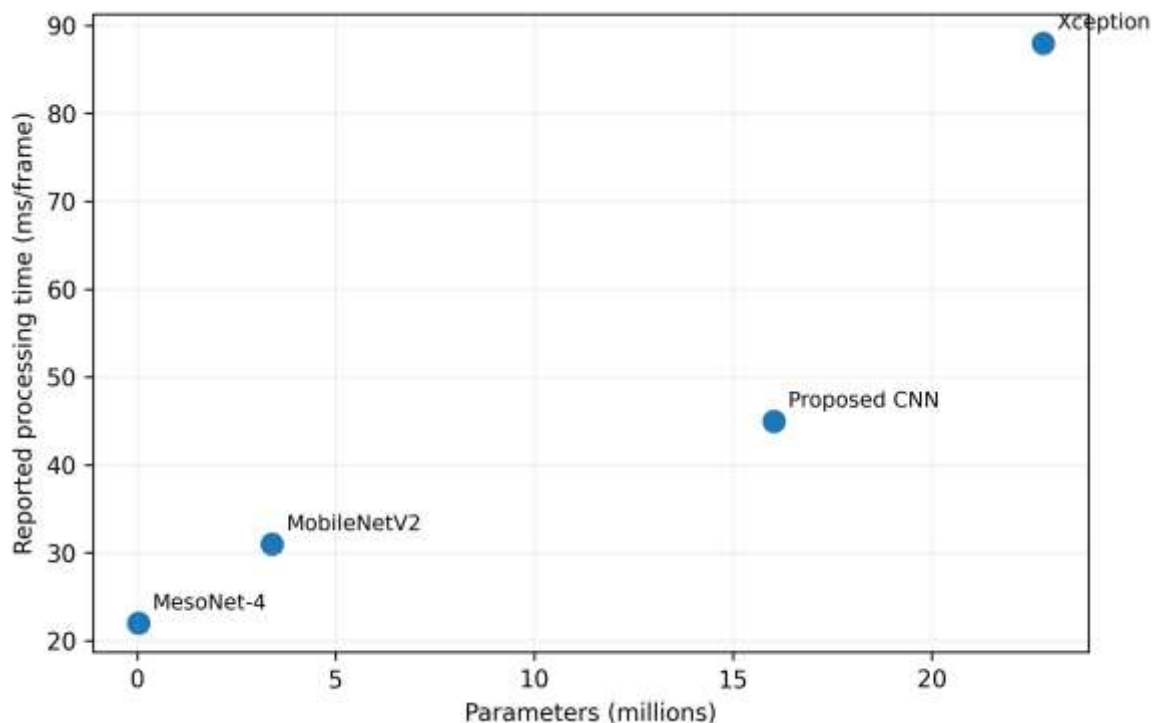


Figure 4.4 Reported parameter count versus processing time. Parameter values are in millions

4.5 Prototype implementation

The pre-trained classifier has been incorporated into an experimental application designated FaceGuard. This application enables users to submit facial imagery and receive an automated assessment determining authenticity or forgery status. Given the documented test results, this application should be regarded as a preliminary screening instrument or investigative platform, rather than an independent forensic adjudication system.

5. Discussion

The findings demonstrate that the proposed convolutional neural network successfully acquired patterns from the training dataset; however, this acquired knowledge failed to manifest in comparable performance metrics during the validation phase. The widening disparity between training and validation outcomes points to overfitting, suggesting that certain learned features may be excessively dependent on the training data rather than functioning as dependable indicators of manipulation. The test outcomes corroborate these apprehensions regarding model generalization. An accuracy rate of 59.02% was attained, yet the confusion matrix disclosed substantial misclassifications in both categories. The presence of 39 false positives indicates that authentic images were erroneously classified as forgeries, whereas 45 false negatives demonstrate that numerous forged images were incorrectly accepted as authentic. Since forged images constituted the positive classification, the recall metric of 53.13% proves particularly significant for evaluating the detector's proficiency in detecting manipulated content. A ROC-AUC score of 0.6303 indicates a modest capacity for class differentiation, yet simultaneously reveals that the distinction between classes remains relatively weak. Upon examination in conjunction with the confusion matrix, it becomes evident that the observed errors cannot be solely attributed to an inappropriately calibrated single classification threshold. Multiple factors may contribute to this outcome, including dataset magnitude and compositional diversity, identity variations, pose alterations, lighting conditions, image quality degradation, compression artifacts, manipulation procedures, and insufficient regularization mechanisms. Although the source material documents the implementation of techniques such as dropout, batch normalization, data augmentation, and early stopping, a substantial generalization deficit persists. From a computational perspective, the baseline comparison yields significant insights. The evaluation demonstrates that MesoNet-4, MobileNetV2, and Xception achieved superior performance relative to the proposed CNN regarding accuracy and F1-score metrics. Nevertheless, the proposed model demonstrated fewer parameters than Xception and exhibited faster data processing capabilities. This underscores that computational efficiency and classification performance represent distinct dimensions requiring concurrent evaluation. When interpreting these findings, the experimental limitations warrant careful consideration.

6. Conclusion

Treat the current implementation as an experimental prototype and perform independent verification prior to deployment in any mission-critical forensic applications. On developing and evaluating a convolutional neural network-based binary classifier designed to identify facial-image forgeries. The evaluation results demonstrated that the model attained an accuracy of 59.02%, with precision measured at 56.67%, recall at 53.13%, an F1-score of 54.84%, and a ROC-AUC of 0.6303. The confusion matrix indicated 70 true negatives, 39 false positives, 45 false negatives, and 51 true positives.

These outcomes indicate that although the model has acquired certain discriminative features for distinguishing authentic from forged facial images, it demonstrates limitations in generalization capability. In comparison to alternative models including MesoNet-4, MobileNetV2, and Xception, the proposed CNN failed to match their classification

performance benchmarks. Nevertheless, its computational characteristics establish a foundation for investigating enhanced architectures that could strengthen detection performance while maintaining computational efficiency.

Regarding prospective forensic applications, subsequent research endeavors should prioritize employing substantially larger and more heterogeneous datasets, implementing rigorous regularization techniques to mitigate overfitting, utilizing transfer learning methodologies, enhancing recall performance for forged image detection, executing cross-dataset validation studies, and guaranteeing that test datasets are sufficiently comprehensive and thoroughly documented, supplemented by independent verification.

7. Future Work

To enhance the comprehensiveness of a dataset, it is essential to include a wide range of identities, manipulation techniques, lighting conditions, image resolutions, compression ratios, and demographic characteristics. This diversity will contribute to a more robust model. Additionally, exploring reliable regularization methods such as weight decay, adaptive learning-rate scheduling, advanced augmentation techniques, and validation-driven checkpoint preservation can significantly improve model robustness. The evaluation of transfer learning capabilities using established pretrained models like MobileNetV2 and Xception is crucial for optimizing performance. To achieve better detection accuracy for manipulated images, it is important to focus on the formulation of loss functions, implement class-balancing strategies, and conduct systematic threshold evaluations. Rigorous cross-dataset validation, particularly with manipulation techniques not present in the original training dataset, is a vital step in ensuring model reliability. Utilizing extensive, well-annotated evaluation datasets while maintaining alignment between dataset size, confusion matrix totals, and reported performance metrics will enhance the credibility of results. Furthermore, investigating architectural innovations that can improve classification effectiveness while maintaining the computational efficiency of existing convolutional neural networks is worth pursuing. Finally, treating the current implementation as a preliminary prototype and conducting independent verification before deployment in critical forensic applications is a necessary best practice.

References

- Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A compact facial video forgery detection network. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–7. <https://doi.org/10.1109/WIFS.2018.8630761>
- Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753–1820.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444. <https://doi.org/10.1038/nature14539>
- Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A new dataset for deepfake forensics. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3207–3216.
- Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1), Article 7. <https://doi.org/10.1145/3425780>
- Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1–11.
- Tolosana, R., Romero-Tapiador, S., Vera-Rodriguez, R., Fierrez, J., & Ortega-Garcia, J. (2020). DeepFakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64, 131–148. <https://doi.org/10.1016/j.inffus.2020.06.014>
- Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 910–932. <https://doi.org/10.1109/JSTSP.2020.3002101>
- Wang, S.-Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). CNN-generated images are surprisingly easy to spot... for now. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8695–8704.