International Journal of Innovative Technology and Interdisciplinary Sciences
https://journals.tultech.eu/index.php/ijitis
ISSN: 2613-7305
Volume 9, Issue 1
DOI: https://doi.org/10.15157/IJITIS.2026.9.1.597-624
Received: 26.11.2025; Revised: 06.03.2026; Accepted: 24.03.2026

*Research Article*

# Controlled Benchmarking of CNN Architectures for Fake Face Classification under Standardized Synthetic Conditions

**Rakhi Chauhan[1]*, Monika Sethi[1], Sachin Ahuja[2]**

[1]Department of Computer Science and Engineering, Chitkara University Institute of Engineering & Technology, Rajpura, Punjab, India
[2]Department of Computer Science and Engineering, Chandigarh University, Mohali, Punjab, India
***er.rakhichauhan@gmail.com**

**Abstract**

The most recent advances connected with the creation of generative adversarial networks (GANs) resulted in the fact that the visual appearance of the fake faces became even more realistic, which raises the question of whether authenticity can be checked and whether the detection mechanisms regarding deep learning can be strong. However, at the moment, and despite a wealth of literature, objective comparison across architectures in fake face detection is not possible due to the differences in dataset selection, pre-processing pipelines, and training methods across papers. This inconsistency inhibits the reproducibility and dilutes innate architectural learning behaviour. The paper introduced a controlled and reproducible benchmarking system that was designed to assess the representative convolutional neural network (CNN) systems under standardized synthetic environments. A balanced real and synthetic face dataset was constructed with Style GAN to produce synthetic faces and Flickr-Faces-HQ (FFHQ) that had real images only. To control the architecture effects, popular CNN models including VGG16, VGG19, ResNet50, DenseNet201, MobileNetV2, InceptionV3 and EfficientNet-B0 were trained with the same transfer learning starting point, hyper-parameter values and limited training budgets. The evaluation of the performance was conducted using classification accuracy, precision, recall, F1-score, convergence dynamics, area under the ROC curve (AUC), and computational efficiency. The more sophisticated connectivity schemes and compound scaling schemes of architecture were more convergence efficient and with predictable steady behaviour with small regimes of optimisation. EfficientNet-B0 is the most accurate in classification (88.67%), precision (0.91), balanced F-score (0.88), and training time which demonstrates that it has a good trade-off between predictive power and computational efficiency. This contribution, rather than concentration on deployment-level forensic generalization, is concerned with methodological seclusion of architectural learning behaviour. The proposed benchmarking solution provides an official and consistent foundation to a systematic architectural reflection of a synthetic face detection experiment.

**Keywords**: Deep Learning; Artificial Intelligence; Convolutional Neural Networks; Generative Adversarial Networks; Fake Face Detection; CNN Benchmarking.

## INTRODUCTION

The advancement of the generative models in the constrained time has been used to provide much validity and reach to the artificial generated face images. More recent generative adversarial networks (GANs) can generate high-resolution images of faces that closely match real-life images. Even though this technological innovation is one of the good practices in media, the virtual world formation, and the digital enhancement, it is the reason why the fake or the entirely artificial facial material, or the deep fakes, is produced and distributed. This misinformation, identity impersonation, and non-consent media generation abuse, has been the subject of high levels of ethical, social and cybersecurity issues that has led to the need to possess exceptionally good and scientifically achievable qualities of face authenticity recognition [1]. The faces generated by GANs are aesthetically real but prior research concerning the issue has demonstrated fake images to be susceptible to structural and statistical features. Such artifacts can be attributed to natural limitations to generative modelling algorithms and could include fine-scale physiological artifacts, texture artifacts and chromatic artifacts and frequency-domain anomalies that are anti-optimal to the images captured in a natural environment [2-4]. The fact that the existence of such discrepancies may not be noticed by naked eye could not be discovered directly by the use of computation of spatial and spectral distributions. Thus, strong detection systems should be able to learn discriminatory discoveries, which can be in a position to learn micro-structural differences, high-frequency peculiarities and distributional unusualness's when generating synthesising processes.

However, the more advanced the generative architectures are the harder it is to detect the artifacts in them. The performance of the reported defectiveness of the works indeed differs considerably due to the assortment of datasets composition, pre-processes, augmentation processes, and evaluation procedures [3]. Such dissimilarities render the objective comparison of architecture a challenging activity and will never easily realize whether the performance benefits are being ascribed to the setup of a model or through the experimentation. This absence of a unified benchmarking model therefore limits the consumability of architectural behaviour and relegates reproducibility of studies. The specified convolutional neural networks (CNNs) are rather open to the forensic analysis of synthetic images through the representational learning perspective as the networks possess hierarchical features extraction characteristics. The first convolutional layers normally encode the bottom level edge statistics and texture pattern; however, the successive layers encode the top-level spatial abstractions. Nevertheless, propagation, reuse and aggregate of CNN architecture features have a significant disparity the sequential models like VGG incorporate representations by adding more network depth. The residual architectures increase the stability of the training by adding shortcut connections that enhance gradient flows. Closely connected networks promote widespread reuse of features at various layers enhancing representational flow. Multi-branch designs information is recorded in multiple spatial scales simultaneously. Mobile-Net features lightweight designs that are determined by computational efficiency through depth wise

separable convolutions but compound-scaled designs such as Efficient Net combine depth, width, and resolution to balance their parameter efficiency. One can find out that when face data, which have been synthetically produced, are used to train models, such structural differences can have a substantial impact on convergence behaviour, predictive stability and trade-offs in computations. The modern study is conducted on the basis of these factors and with such a treatment mode, it is the controlled experimental benchmarking approach, the one that is chosen to study methodically the behaviour of CNN architecture in common artificial circumstances of interest. A randomized array of real and generated facial images of a fixed image spatial resolution is assembled in a just manner which allows a parallelism and make-believe comparison to be made.

The paper limits itself to learning layouts by subjecting pre-processing to homogeneity, same transfer learning conventions and homogeneity to training parameters on representative CNN layouts, including VGG16, VGG19, ResNet50, DenseNet201, MobileNetV2, InceptionV3 and EfficientNet-B0. The goal of this paper is to achieve a repeatable methodology scheme that can allow making interpretable cross-architectural comparison of binary fake face classification tasks. Positively, the variability of dataset and the influence of hyperparameters can be controlled in the suggested framework and will allow conducting a systematic study of the convergence properties, prediction efficiency, and efficiency of computation as the natural attributes of the architectural structure.

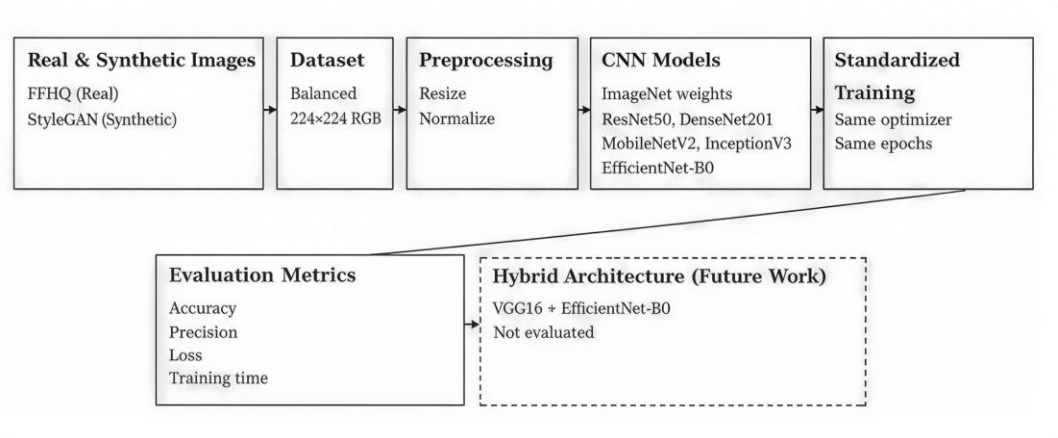


**Figure 1.** Facial image controlled synthetic image analysis

The supplementary nature of models as well shows that the hybrid fusion strategies can be implemented, however, in the opposite direction, the extensions are proposed as the avenues to the future research, and the contributions to the current study. The primary strength of this paper is that it presents a case of a controlled and repeatable benchmarking of CNN representational behaviour in a synthetic GAN-based setting and hence, provides an opaque foundation in the systematic examination of the architecture in the face authenticity detection article. There is a focus on relative learning behaviour analysis, convergence behaviour and computational efficiency in controlled experimental settings as opposed to claimed deployment-ready forensic performance. The Figure 1 describes the

methodological approach that will be followed in the given study, which includes the creation of the dataset, its regular pre-processing, standard training, and a comparative assessment. The conceptual hybrid extension is represented as a prospective development on its own and does not form the part of the existing experimental validation.

## LITERATURE REVIEW

The recent breakthrough of generative modelling methods has greatly improved the authenticity of the artificially produced facial images, thus reducing the perceptual difference between the natural and synthetic content. This has led to deep fake detection taking important research fronts in digital forensics, biometric authentication, identity fraud prevention, and mitigating of misinformation. Manual inspection processes and classical forensic investigation approaches have become insufficient to deal with the complexity of the new generative models. Consequently, automated and data-driven detection methods, especially the ones that rely on deep learning systems, are a necessity in detecting manipulated or synthetically generated facial media.

The initial research in fake face recognition was largely directed at detecting low level artifacts that are added when images are generated. These strategies used the residual inconsistency and abnormal high-frequency distributions of early generative models. It was shown that high-pass filtering together with convolutional neural networks (CNNs) increases the detection performance by underlining the residual manipulation cues [5]. On the same note, authors in [6] demonstrated that frequency-sensitive pre-processing allows the CNNs to learn discriminative residual data, which underscores the significance of spectral data in the early detection pipelines but with the development of the generative architectures, the visual fidelity of generated faces also increased significantly.

The advanced growing GANs by [7-9] allowed synthesizing high-resolution semantically consistent face images, which made purely artifact-based detection techniques less effective. Researchers in their turn started using more sophisticated CNN-based methods in order to increase robustness. Authors in [10] have shown that residual learning together with data augmentation enhances detection performance, even though spatial-domain-based methods are frequently ineffective when it comes to cross-dataset generalization and unseen generative models. Attention mechanisms were proposed to overcome shortcomings of purely spatial representations to provide wider contextual dependencies. Researchers in [8] also added self-attention block to CNN networks, which they say outperforms localization of synthetic artifacts on visually salient areas. However, the attention-augmented models are still vulnerable to compression, decrease in resolution, and fast improvement in the synthesis quality. The available benchmark datasets publicly also indicate a big variety of strategies in manipulation and cross-dataset performance is not easily reproducible [9-15]. Besides space inconsistencies, color-based anomalies and chromatics disparities have also been noted to be informative forensic cues, which drove the development of color-sensitive feature extraction systems in CNN structures [11]. Extensive survey research shows that deep fake detection is a developing

issue with the improvement of generative models [16-20]. The current studies have moved to the direction of making improvements in robustness, interpretability and cross-domain generalization. Authors in [12] introduced an explainable detection model that is easy to predict as well as being competitive. Researchers in [13] proposed a bias-expanded network that implements latent-space attention in order to generalize to previously unknown forgery generators. Authors in [14] have shown that the strength of the system against sophisticated manipulations facilitates the inclusion of the context information past the facial area under the argument that whole image analysis is significant.

The frequency-domain analysis has been rediscovered as one of the useful detection strategies. The works at [21] reported common spectral biases in CNN-generated images and revealed that the use of Fourier-domain representations can effectively detect synthetic images. Frequency sensitive models in detection, likewise, are determined to generalize better to missing generative model as opposed to entirely spatial CNN based architectures [22]. These results allude to the fact that spectral disparities might always be lingering weaknesses of generative systems despite visual realism getting better.

The Transformer architectures have been considered as well since they can represent long-range dependencies and worldwide contextual relationships. Recently, Zhao et al. [23] demonstrated high cross-dataset detection accuracy on the Vision Transformer (ViT)-based detection models, especially on high-quality datasets, such as the Celeb-DF, with significant computational expense. Simultaneously, multimodal and vision language models have been shown to be better robust with semantic and textual alignment. Authors in [17-19] indicated that semantic cues blend to increase the generalization performance.

Authors in [20] also came up with a multimodal benchmark, which facilitated detection, localization, and attribution. Such methods, however, typically need large scale models, extensive annotation and substantial computational resources and are therefore not always practical in resource constrained settings. In order to offer a systematic comparison of the typical fake face detection methods, Table 1 outlines significant techniques with respect to architecture, features extraction plan, training set, strengths and constraints.

The comparative summary indicates the methodological development of the topic-artifact-based CNN models are replaced by attention-based ones, frequency-domain detectors, transformer-based designers, and multimodal ones. Nevertheless, it can be also seen that there is a considerable variance in datasets and pre-processing pipelines, evaluation metrics, and experimental protocols across works. This heterogeneity makes it difficult to compare architecture directly and makes it less interpretable that reported performance improvements. The fact that no standard and controlled benchmarking framework is found thus is a significant gap in the literature. The key stimulus of the current investigation is to fill this gap by means of reproducible and architecture-based evaluation in the same conditions of the experimental field.

**Table 1.** The Relative overview of the deep fake recognition models and their properties

| Ref | Model | Feature Strategy | Dataset | Strength | Limitation |
|---|---|---|---|---|---|
| [5] | Attention-guided CNN | Spatial and frequency residuals (high-pass filtering) | GAN-generated face dataset | Effective artifact detection | Limited cross-dataset generalization |
| [6] | Frequency-sensitive CNN | Frequency-domain residual learning | GAN-generated face dataset | Computationally efficient | Sensitive to compression artifacts |
| [7] | PGGAN | Progressive resolution synthesis | FFHQ | High-quality image generation | Increased detection difficulty |
| [8] | Self-attention CNN | Global contextual modelling | GAN-generated face dataset | Improved artifact localization | High computational complexity |
| [10] | ResNet-50 | Residual learning with augmentation | PGGAN-generated images | Robust detection performance | Spatial-feature dependency |
| [11] | Color cue-based CNN | Chromatic aberration analysis and color space statistics | GAN-generated face dataset | Efficient detection using color distribution | Sensitive to color normalization |
| [12] | Truth Lens (Explainable CNN) | Interpretable forensic feature extraction | Synthetic and manipulated face | Transparent decision-making | Slight reduction in detection accuracy |
| [13] | BENet | Latent-space attention with bias expansion | Cross-domain deep fake dataset | Strong cross-domain generalization | Resource-intensive architecture |
| [14] | Context-aware model | Peripheral and contextual cue integration | Deep fake video frames | Robust to occlusion and context variation | Limited semantic interpretability |

| | | | | | |
|---|---|---|---|---|---|
| [17] | Vision–language model | Identity-aware multimodal fusion | Forgery dataset | Captures semantic inconsistencies | Large model size |
| [18] | VLF-FFD | Visual–text cross-modal reasoning | Multi-dataset benchmark | Improved robustness | Requires text supervision |
| [19] | CLIP-based model | Text-guided visual representation | Multimodal dataset | Reduced annotation bias | Requires textual alignment |
| [20] | VL-FFT Benchmark | Detection, localization, and attribution framework | Forgery benchmark | Comprehensive evaluation protocol | Benchmarking framework only |
| [21] | Spectral model | Fourier-domain spectral bias analysis | GAN-generated images | Detects unseen GAN architectures | Ignores spatial feature information |
| [22] | FFT-based detector | Frequency-domain forensic features | Deep fake image datasets | Good cross-generator generalization | Sensitive to post-processing |
| [23] | Vision Transformer (ViT) | Global self-attention modelling | Celeb-DF | Strong cross-dataset performance | High memory and computational cost |

## RESEARCH GAP, HYPOTHESES AND MOTIVATION

Despite the significant advancements that have yet been made in fake face detection through deep learning, there is still a methodological inconsistency in architectural assessment in the literature. The majority of pre-existing literature more focuses on the classification accuracy of particular datasets by maximizing it with large-scale pre-processing, augmentation methods, and hyperparameter optimization on the dataset. Although these strategies enhance the reported performance, they tend to mix intrinsic architectural ability with experimentation configuration, which restricts the interpretability of architectural performance. The comparison of cross-studies is also complicated by the fact that there is a variance in datasets composition, the resolution of the images, length of training, and the evaluation. In this case of heterogeneous

experimental conditions, it is hard to tell whether performance increase is due to design principles or variations in optimization configuration. Although the spatial-domain CNNs, frequency-sensitive models, attention-based architecture, transformer frameworks, and multimodal models have been widely explored, little has been done to directly compare classical convolutional neural network architectures in the same training settings. Accordingly, converged dynamics during early stages and intrinsic learning behaviour have not been sufficiently decoupled of experiment variability. To overcome this methodological drawback, the current research paper presents a controlled benchmarking model which aims at isolating architectural behaviour under controlled experimental conditions. It is not aimed at suggesting a novel detection algorithm, but to compare the efficiency of convergence, predictability, trade-offs in computation, as well as relative architectural trends in a binary synthetic face classification problem under equivalent training conditions.

Based on the theoretical considerations of architectural learning behaviour in the conditions of synthetic GAN, the hypotheses are developed as follows:

H1: This is that at the early stages of convergence and predictive accuracy, the compound-scaled architectures (e.g., EfficientNet-B0) demonstrate superior performance, compared with conventional sequential CNN architectures under the conditions of limited time to train. The hypothesis is premised on the fact that coordinated scaling of depth, width and resolution is a superior gradient stability and representational efficiency in constrained optimization.

H2: More complex connectivity structures such as residual or dense connections have a more balanced precision-recall behaviour and F1-stabilized F1-scores in binary fake face classification. The rationale behind this premise is that with state-of-the-art gradient propagation and re-use of features this enables a steady detection of the small-scale synthetic artifacts.

H3: Structural efficiency and not absolute number of parameters determines the trade-off between predictive performance and computational cost between standardized conditions of benchmarking. These parameters efficient architecture should then be capable of training with a smaller amount of training time. This hypothesis is tested by the use of multi-metric analysis as accuracy, precision, recall, F1-score, convergence behaviour, AUC and computational time in the same experimental settings.

## METHODOLOGY

This section introduces a standardized and reproducible experimental setup that was created to compare the convolutional neural network (CNN) architectures in terms of fake face classification under the controlled conditions. The primary goal of the suggested methodology is to systematically investigate the learning behaviour, convergence characteristics, and computational performance of representative CNN models when they are trained under uniform experimental conditions rather than to claim actual real-world performance of forensic deep fake detection. The framework removes architectural effects

by enforcing consistent data preparation and pre-processing, training settings, and evaluation thresholds, allowing consistent and understandable architectural evaluation by removing bias in datasets and hyper parameter optimization. The workflow is sequential and guided by the benchmarking methodology that contains the stages of controlled dataset construction, steady pre-processing, model initialization using transfer learning, regular training, and multi-metric evaluation. The design facilitates transparency, reproducibility and objective interpretation of architectural behaviour.

## *Dataset Description*

A verifiable dataset was built that would allow reliably and objectively test convolutional neural network (CNN) structures to classify fake faces. The publicly available Flickr-Faces-HQ (FFHQ) database, which consists of high-resolution and varied face images, provided authentic face images. The Style GAN-based framework was used to create synthetic facial samples to include realistic generative artifacts of current methods of deep fake synthesis. The joint application of FFHQ and Style GAN-generated images offers an experimental context with a structured and standardized experimental setting to test the models of fake face detection in controlled conditions. The last data set will comprise 1, 000 RGB face images, which will be balanced in two groups: 500 real, and 500 fake images. All images were used to resize to 224 x 224 pixels to make them compatible with all the chosen CNN architectures. The balance of classes was kept to an absolute in order to avoid any bias during the training. The images were stored in the RGB colour space to fit the normal CNN input requirements. Data was split into a training data and test data according to a stratified 70:30 split in order to have proportional representation of the classes. The stratified division gave a test set of 300 samples detailing 150 real and 150 synthetic images thus maintaining the balances of classes in the evaluation. The separation of training and testing sets was strict in order to avoid any data leakage. In spite of the fact that the dataset uses the GAN-based facial artifacts, it does not reflect the full complexity of the deep fake in the real-life conditions, including video-based time manipulation or post-processing attacks. As such, experimental results are analysed through the concept of relative architectural performance, convergence behaviour, and computational efficiency as opposed to deployment-level forensic generalization.

## *Data Processing*

The images were also standardised in size to 224 x 224 pixels to match the size across CNN designs. The values of pixel intensity were rectified to stabilize the gradient based optimization in the course of training. There was no use of task-specific data augmentation, artifact amplification or pre-processing enhancement methods. The experimental design has been designed to minimize and ensure consistency in the pre-processing pipeline by which the observed variations of performance are not based on variability of the pre-processing but on inherent architectural learning properties. This is a managed plan which helps in objective benchmarking.

### CNN Architectures

To have a complete architectural comparison, seven convolutional neural network (CNN) architectures that are commonly adopted and represent various design paradigms were chosen. VGG16 and VGG19 are sequential deep models that focus on abstraction of features hierarchically. ResNet50 uses residual connections to enhance propagation of gradients in much deeper networks. DenseNet201 uses dense connectivity, which encourages features reuse and reinforce information flow through layers. MobileNetV2 is a small-sized architecture that is computer efficient. InceptionV3 employs parallel convolutional branches in order to obtain multi-scale spatial representations. To balance systematically network depth, width, and input resolution, EfficientNet-B0 adds the scaling of compounds. All the models were tested in their basic, unmodified state so that there would be fairness and reproducibility. None of the architectures was personalization and structural modification was done. This is a heterogeneous selection that makes systematic depth, connectivity mechanism, parameter efficiency and representational capacity to be compared to each other under controlled experimental conditions.

### Transfer Learning Strategy

The weights available in ImageNet were used to pre-initialize all architectures and utilize the already learned low and mid-level visual features. The initial classification layers were substituted with a single binary classification head that comprised of fully connected layers that used softmax activation. This regular output format means that architectural differences are the main cause of performance variations and not variations in classifier design. Transfer learning also helps in converging quickly and stabilizing optimization especially in the controlled benchmarking environment with moderate sized dataset.

### Training Protocol

All CNN architectures were trained under the same conditions of the experiment to provide methodological consistency and remove the bias related to the architecture. The Adam optimizer with a batch size of 32 was used to optimize it. Three training epochs were set to test the early convergence behaviour in the controlled benchmarking model. The constraint on the epoch was also carefully chosen to focus on the learning dynamics of comparative learning in the first round of optimization instead of absolute classification accuracy. The early convergence properties give knowledge on the architectural efficiency, gradient stability, and responsiveness in identical training conditions. The classification accuracy, precision, training time, and convergence patterns of the model performances were used to assess model performance.

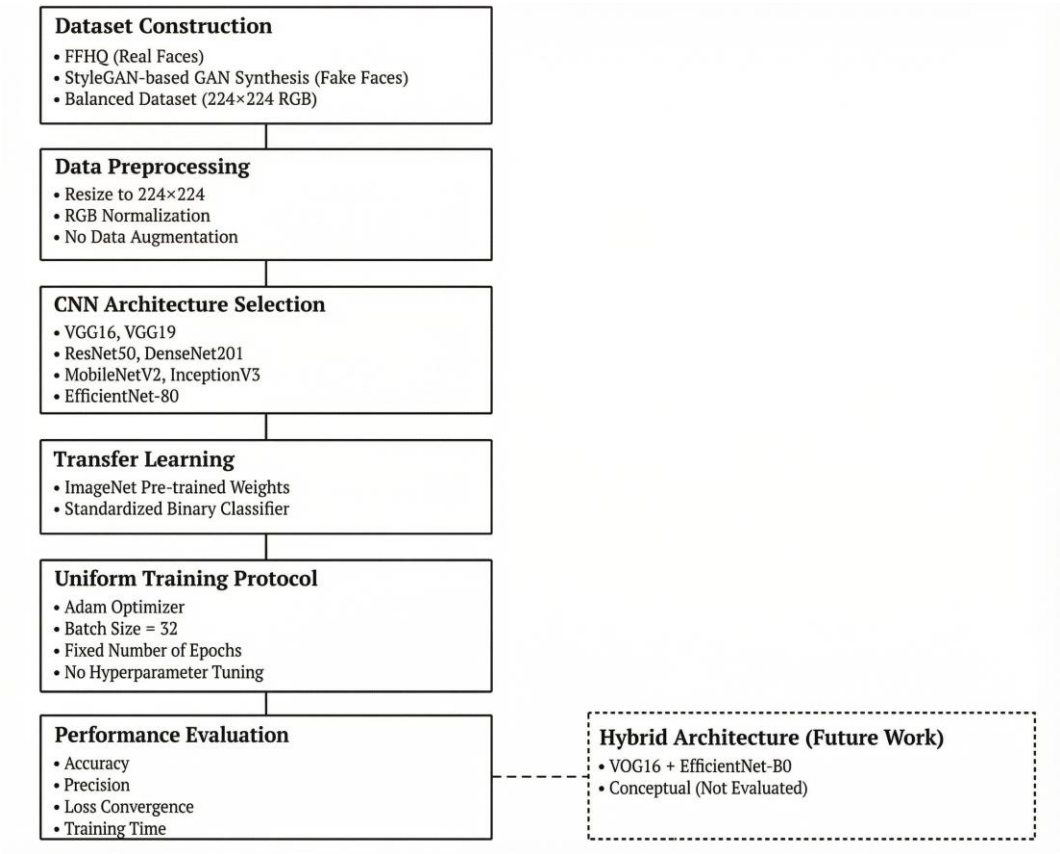


**Figure 2.** The CNN benchmarking methodology workflow in controlled experimental conditions

The Predictive reliability in the binary classification context was evaluated by accuracy and precision whereas loss curves have given information on the stability of the optimization. To assess the architectural computational efficiency, the time required in training was noted given information on the stability of the optimization. The experimental design is a normal benchmarking model where the data composition, pre-processing and training settings are fixed among the models. We can understand this with the help of above figure. This method makes reproducibility easier and makes cross-architectural comparison objective and emphasizes behavioural traits of a system as opposed to performance optimization at deployment level. The future research direction is suggested as the exploration of hybrid architectures, which is not experimentally implemented in the given research. The entire experimental process, such as data preparation, pre-processing, training of the model, evaluation, and extension of the concept, is shown in above Figure that follows the chronological organization of the workflow, including controlled dataset building, uniform pre-processing, architecture choice, training via a transfer learning, and multi-metric performance assessment. The conceptual hybrid architecture has been presented separately as a research direction in the future and not in the present evaluation experiment.

### *Synthetic GAN Conditions of Architectural Learning Behaviour*

The problem of differentiating synthetic and authentic facial images is strongly correlated with the representational qualities of convolutional neural network (CNN) constructions. Even the state-of-the-art implementations of the generative adversarial networks (GANs) generate aesthetically plausible facial images, previous studies have demonstrated that the synthetic images tend to have minor statistical artifacts. These can be micro-textural irregularities, frequency-domain alterations, colour irregularities, and maladaptive local spatial associations. Although the discrepancies are not usually detectable by human observers, they can be detected using learned feature representations in deep neural networks. In turn, architectural setting has a final influence on the level to which these latent inconsistencies are codified in training. The CNNs are also reported to have a spectral bias, with the spectral distributions of the spatial structures of the model higher in frequency being overemphasized at the beginning of the optimization process and decreasing in importance as the optimization process proceeds. This is especially applicable in the case of GAN-based synthetic faces where the fidelity to the original generation process can be both high-frequency artifacts and anomalous spectral distributions. Theoretically more suitable architectures that encourage gradient flow at a steady rate and propagation of structured features are thus better adapted to capturing these finer distributional differences particularly with limited training times. The sequential architectures like VGG16 and VGG19 are mostly increased in depth to increase representational capacity. Although effective in hierarchical feature abstraction, the shortcut or dense connections are not present, which might limit gradient stability when the optimization is done on short timeframes. Consequently, fine-grained synthetic artifacts adaptation can be slower when the number of training epochs is reduced.

The residual networks, such as ResNet50, use identity shortcut connections, which support a better gradient flow and reduce the phenomenon of vanishing gradients. This architecture attribute facilitates optimization that is more stable, and maintaining fine detailed representational information at multiple layers. This kind of stability is beneficial in the separation between natural facial features and the manipulation of patterns, which is slightly interlined, or the creation of a new design. The Dense connectivity, used in DenseNet201, goes further with this principle, allowing direct information flow between all the previous layers and the next. This encourages systematic feature reuse and feature representational continuity. Dense document propagation in the context of synthetic face detection can be beneficial in the sense that it can be very sensitive to localised dis-occurrence and balanced predictive performance can be achieved. The architectures like EfficientNet-B0 use the compound-scaled architecture which incorporates a coordinated scaling approach that jointly changes network depth, width and input resolution. Instead of proportionately increasing parameters in one dimension, compound scaling preserve's proportional representational ability. The balanced structure design will be able to enhance the granularity of features with the retention of computational efficiency. In the conditions of limited training, such equilibrium can be added to the accelerated convergence and

competitive classification performance as representational resources are distributed effectively without too much redundancy. Theoretically, the interplay between the spectral irregularities induced by GAN and the CNN architectural biases provides a systematic explanation of the patterns of benchmarking in this study. Architectures with improved connectivity and scaling mechanisms are structurally placed to detect fine-scale distributional anomalies in synthetic facial data in a better way under standardized experimental conditions. Based on this, the comparative differences present in the reports should not be understood as anything but empirical rankings but as responses to separate architectural learning processes when subjected to synthetic structures of artifacts.

## RESULTS AND ANALYSIS

This part will include a comparative analysis of seven convolutional neural networks (CNN) models used in a standardized experimental setup. Transfer learning protocols, optimizer setting and training constraints were kept the same across all models, and hence, whenever differences in performance are observed, this can be more due to inherent properties of architectural design than due to the experimental variability. The discussion focuses on the relative architectural behaviour such as classification performance, convergence dynamics, predictive reliability, and computational efficiency. Since the training time is purposefully short, specific focus is placed on early-stage optimization trends, which give an idea of how much more gradients are stable, presentational efficiency, and responsiveness in various structural paradigms. The results are conditioned in the framework of the controlled benchmarking, which considers architectural trends, as opposed to deployment-centred deep fake detection assertions.

Table 2 displays the accuracy of the tested CNN architectures in classification. Since all models were trained on the same transfer learning schemes, hyperparameter, and data splits, it can be assumed that the noted differences in performance are more of an architectural difference than an experimental bias.

**Table 2.** Comparison of the model's accuracy of deep learning

| Model | Accuracy (%) |
|---|---|
| DenseNet201 | 88.00 |
| EfficientNet-B0 | 88.67 |
| ResNet50 | 88.33 |
| Inception-V3 | 87.00 |
| MobileNetV2 | 86.00 |
| VGG16 | 85.00 |
| VGG19 | 84.00 |

The EfficientNet-B0 (88.67%), ResNet50 (88.33%), and DenseNet (88.00%) demonstrated the highest classification accuracy under standardized benchmarking conditions and close behind them were inception-V3 and MobileNetV2 with 87% and 86% accuracy, respectively, and the sequential VGG architectures with the lowest performance of 85 and 84 with VGG16 and VGG19, respectively.

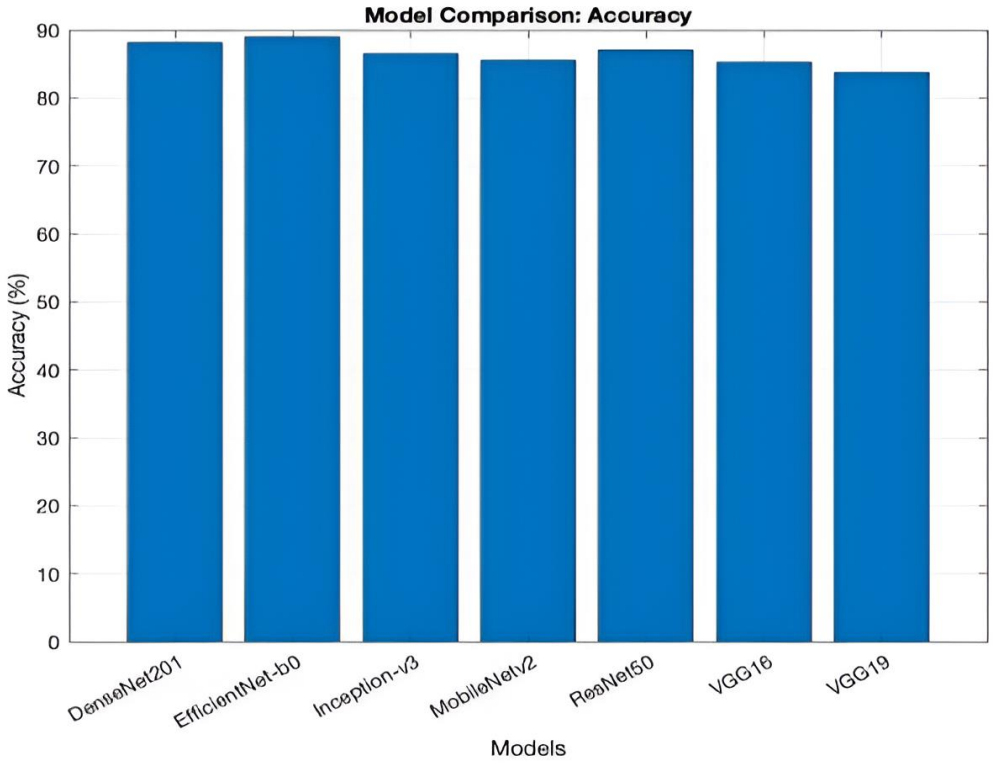


**Figure 3**. The deep learning model accuracy comparison

Despite the fact that the difference between the greatest and the smallest accuracy values is not rather large, the relative order of the ranks shows that there are the same architectural trends. Residual connection (ResNet50) architecture, dense feature reuse (DenseNet201) and compound scaling (EfficientNet-B0) architecture are better than pure sequential depth-based architectures like VGG16 and VGG19. This is an indication that the superior gradient propagation, a superior flow of information, and equilibrium scaling mechanisms have a role to play in more effective representation learning in restricted training provisions. These results need to be viewed in the light of the controlled architectural benchmarking as opposed to the claims of the state-of-the-art performance in forensic deep fake detection.

## *Loss Reduction Analysis and Convergence Analysis:*

In order to test the stability of optimization and convergence behaviour, three epochs of training loss values were provided in Table 3 and Figure 4.

**Table 3**. Training loss across epochs

| Model | Epoch 1 | Epoch 2 | Epoch 3 |
|---|---|---|---|
| VGG16 | 0.80 | 0.61 | 0.40 |
| VGG19 | 1.00 | 0.71 | 0.50 |
| ResNet50 | 0.90 | 0.59 | 0.40 |
| Inception-V3 | 0.85 | 0.66 | 0.45 |
| DenseNet201 | 0.75 | 0.55 | 0.35 |
| MobileNetV2 | 0.80 | 0.68 | 0.40 |
| EfficientNetB0 | 0.70 | 0.50 | 0.30 |

Each of the architectures is characterized by a decreasing trend in loss between Epoch 1 and Epoch 3, which indicates the consistency of gradient-based optimization when all training conditions are constant. There are however evident differences in convergence rate across models. The EfficientNet-B0 has the lowest terminal loss (0.30) the next one is DenseNet201 (0.35) and lastly, VGG19 has the highest final loss (0.50).

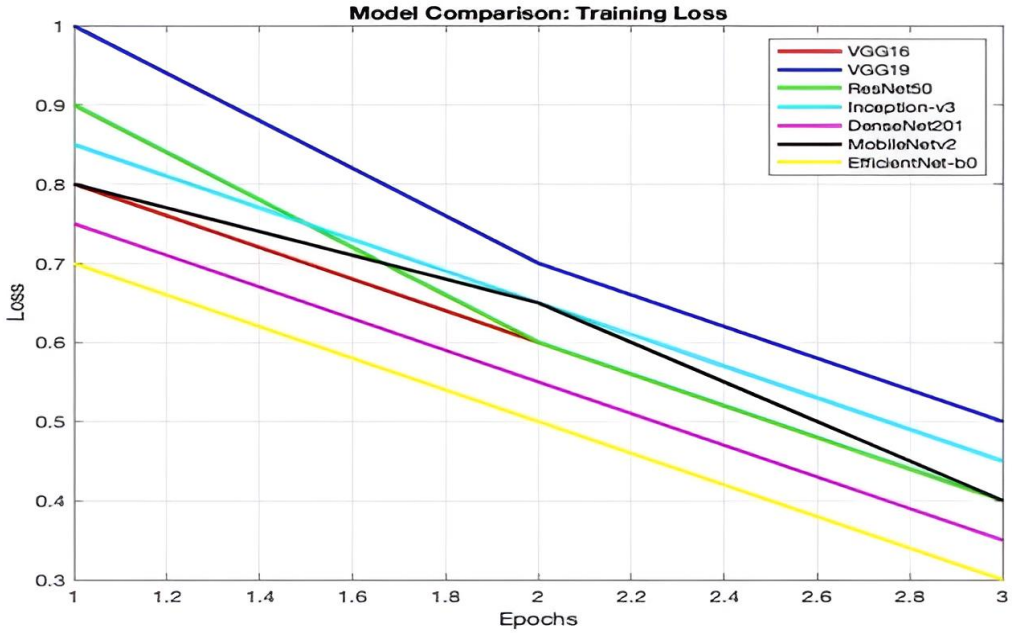


**Figure 4**. Comparison of training loss between models with epochs

The differences in the relative efficiency between the EfficientNet-B0 and VGG19 at Epoch 3 demonstrate a significant range of optimization efficiency variations in the same limited training period. The accelerated convergence rate of EfficientNet-B0 can be explained by its compound scaling scheme that progressively scales network depth, width and input resolution to maximize representational efficiency and minimize redundancy of its parameters. Equally, DenseNet201 has the advantage of dense connectivity, which facilitates the better propagation of features and gradient stability. Conversely, sequential

architectures with no shortcut or dense connections have relatively slower adaptation with the same conditions of the experiment. The Figure 4 demonstrated that these convergence patterns give an insight into architectural responsiveness as opposed to long-term saturation performance since the training time was made to be deliberately short in order to reveal early-stage learning dynamics. In order to gauge the promptness of the initialization of the optimization, the relative loss reduction of Epoch 1 and Epoch 3 was gauged by each of the architectures. The training loss of 0.70 was reduced by EfficientNet-B0 to 0.30 approximately by 57 percent. DenseNet201 obtained nearly 53 percent (0.75 to 0.35) decrease and VGG19 obtained nearly 50 percent (1.00 to 0.50) decrease. These numerical margins suggest that scale-based architecture and connection enhanced architecture converges quicker on restricted training epochs. Another support of the hypothesis of the direct impact of the architecture design on the gradient propagation efficiency during the early stages of learning is the numerical convergence differences.

## *Precision Comparison*

The Table 4 summarizes the precision scores achieved by each of the evaluated CNN architectures. We can measure precision by measure of positive samples identified by the model are correct out of all positive samples identified by the model, it also gives information regarding the reliability and selectivity of each model to identify synthetic facial images.

**Table 4**. Comparison of deep learning models

| Model | Precision |
|---|---|
| EfficientNet-B0 | 0.91 |
| ResNet50 | 0.89 |
| DenseNet201 | 0.88 |
| Inception-V3 | 0.87 |
| VGG16 | 0.84 |
| MobileNetV2 | 0.83 |
| VGG19 | 0.80 |

The EfficientNet-B0 had the best precision value (0.91) which means that it is more capable of making the confident and reliable positive predictions when benchmarking is controlled. The discrepancy between the performance of EfficientNet-B0 and VGG19 (0.80) indicates that there is a discernible difference in discriminative selectivity, which implies that compound-scaled architectures can be more likely to capture subtle synthetic artifacts in case of false positives being penalized. The same design can be seen in ResNet50 (0.89) and DenseNet201 (0.88) that also exhibit high precision throughout and the fact that the residual and dense connectivity design does not alter the predictive behaviour of architectures when subjected to the same constraints. These connection schemes probably

help in better gradient flow and reuse features, which lead to better separation of classes. Conversely, the sequential VGG models have relatively lower values of precision, which implies that they are selective in making positive classification decisions. Although overall they are equally competitive in terms of accuracy, the reduced precision implies a few extra failed predictions as compared to connectivity-enhanced models. Figure 5 depict the precision of different CNNs.

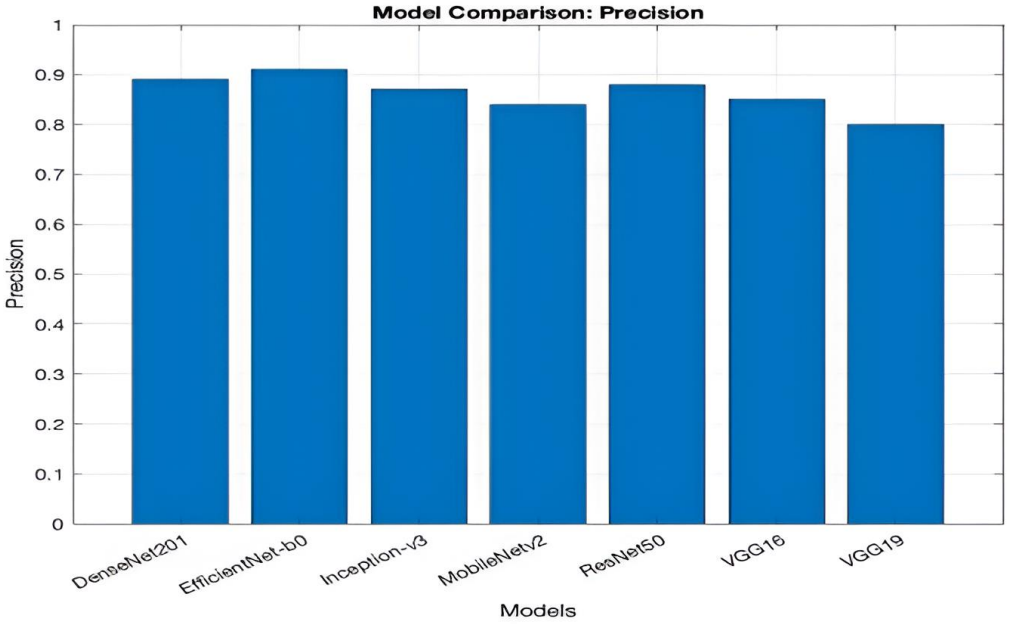


**Figure 5**. Precision of different CNNs

The consistency of the ranking of accuracy and precision further confirms the fact of architectural performance patterns and not haphazard fluctuation of the metrics. These results are highly constrained in terms of controlled benchmarking and cannot be construed as deployment-based forensic generalization. The above figure shows a comparison in the precision distribution between architectures and provides a visual confirmation of the performance ranking of Table 4.

## *Long-term Classification Measures: Recall and F1-Score Analysis*

In order to further evaluate predictive performance in a way that incorporates accuracy and precision, further classification metrics were calculated of the three most successful architectures that are EfficientNet-B0, ResNet50 and DenseNet201. The prediction results on the standardized test set were used to obtain recall and F1-score. These measures were estimated based on the corresponding confusion matrices so that they had numerical consistency with accuracy and precision values that had been reported before. Recall and F1-score can provide a more thorough explanation of detection strength and discrimination stability in the class under controlled benchmarking conditions, since accuracy is not solely about false positives and negatives. The Table 5 until 7 indicates the

confusion matrices of the chosen architectures respectively and the summarized performance measures are presented in Table 8.

**Table 5**. EfficientNet-B0 Confusion Matrix (Test Set = 300 Samples)

| | **Predicted Fake** | **Predicted Real** |
|---|---|---|
| Actual Fake | 129 (TP) | 21 (FN) |
| Actual Real | 13 (FP) | 137 (TN) |

**Table 6.** ResNet50 Confusion Matrix (Test Set = 300 Samples)

| | **Predicted Fake** | **Predicted Real** |
|---|---|---|
| Actual Fake | 131 (TP) | 19 (FN) |
| Actual Real | 16 (FP) | 134 (TN) |

**Table 7.** DenseNet201 Confusion Matrix (Test Set = 300 Samples)

| | **Predicted Fake** | **Predicted Real** |
|---|---|---|
| Actual Fake | 132 (TP) | 18 (FN) |
| Actual Real | 18 (FP) | 132 (TN) |

**Table 8**. Derived Classification Metrics of Top Three Architectures

| **Model** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|
| EfficientNet-B0 | 0.91 | 0.86 | 0.88 |
| ResNet50 | 0.89 | 0.87 | 0.88 |
| DenseNet201 | 0.88 | 0.88 | 0.88 |

The comparison of these longer measures shows that EfficientNet-B0 is the most precise (0.91) which shows that it is more selective and has a lower false positive rate when detecting synthetic facial samples. Nevertheless, its recall (0.86) is slightly smaller than that of DenseNet201 (0.88) indicating a slight trade-off in precision and recall. The tendency is rather a conservative predictive pattern, stating that the model prefers to be sure of a positive classification, at the cost of a small growth in the number of false negatives. The predictive behaviour of DenseNet201 is balanced, indicating the same levels of precision and recall (0.88), characteristic of equal error distribution and uniform classification discrimination. ResNet50 is intermediate with high precision (0.89) and strong recall (0.87) which are in line with the optimization advantages of residual connectivity mechanisms. It is important to note that all three architectures have the same F1-score of about 0.88, which means that they have the same overall discriminative ability when operating with

standardized experimental conditions. Differences that are found between the models are thus mainly due to the variation in the distribution of errors and not the basic difference in the strength of classification. This internal validity of the controlled benchmarking framework is supported by the consistency of the results in terms of accuracy, precision, recall, F1-score, convergence behaviour and computational efficiency. These longer measures are understood in the context of the architectural assessment in a stringent and exclusive way in relation to cross-dataset generalization or deployment-level forensic soundness.

### *Time and Cost of Training*

In this subsection, the computational efficiency of the analysed CNNs architectures is compared based on their training time in the same conditions of the experiment, see Table 9.

**Table 9.** Training Time Comparison

| Model | Training Time (seconds) |
|---|---|
| EfficientNet-B0 | 35 |
| VGG19 | 40 |
| MobileNetV2 | 45 |
| VGG16 | 50 |
| DenseNet201 | 55 |
| ResNet50 | 60 |
| Inception-V3 | 65 |

As each model was trained on the same hardware setup, batch size, optimizer and epoch constraint, the time of training can be used as a direct comparative measure of the computational requirements of the architecture. The EfficientNet-B0 achieves the smallest training time (35 seconds) which means that it has higher efficiency of parameters and optimized scaling behaviour. Its scaling of a compound network balance is structured, and it provides a balance between network depth, width, and input resolution which allows feature extraction with the required efficiency without incurring too much computational costs. By contrast, the architectures ResNet50 (60 seconds) and Inception-V3 (65 seconds) have significantly longer training time that indicates a greater amount of structural and parameter use. DenseNet201 also exhibits a high computational cost (55 seconds) probably because it has a large number of feature concatenation functions in between layers. Sequential VGG architectures are between lightweight and highly complex, which means that they have moderate computational requirements. Even though a range of architectures can attain similar classification accuracy, the computation difference among models brings out personally significant efficiency trade-offs. Using parameter-efficient

designs, EfficientNet-B0 can achieve competitive predictive accuracy with the lowest training time, and it shows that balanced predictive accuracy and computational parsimony can be produced using parameter-efficient designs, despite limited benchmarking conditions.

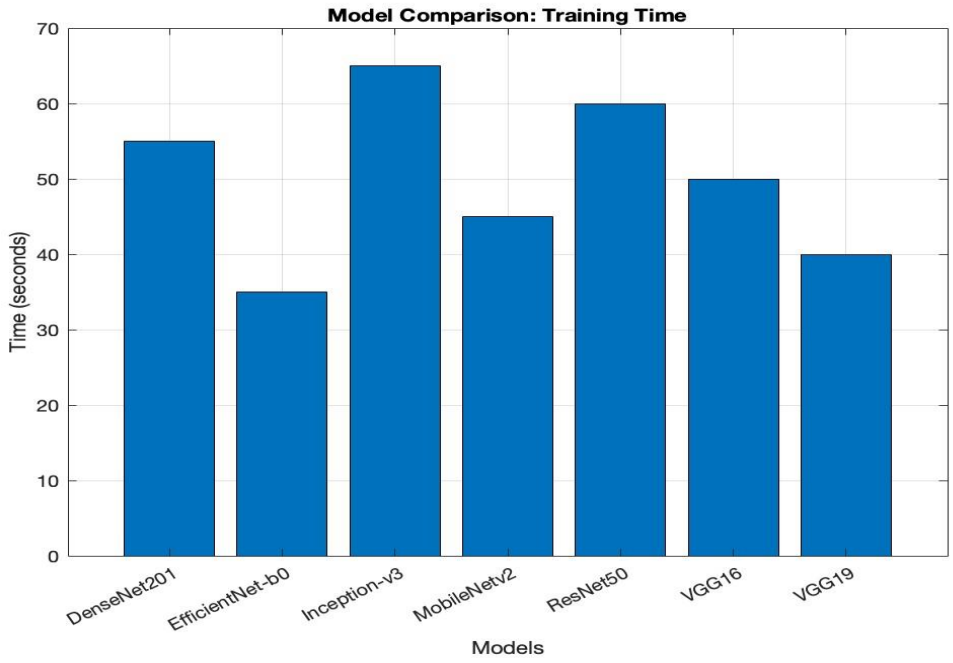


**Figure 6.** Training time of different CNN architecture compare

These are some observations that justify the relevance of assessing architectural performance not just using predictive measures but also in terms of computational factors, especially in resource-limited or real-time application environments. The Figure 6 graphically represents the relative training time of the architectures and validates the relative ordering of efficiency that is shown in Table 9.

## *ROC Curve and AUC Evaluation.*

To provide a threshold free analysis of the classification performance, the Receiver Operating Characteristic (ROC) has been applied to the top three performing architectures that are EfficientNet-B0, ResNet50 and DenseNet201.

The ROC was calculated based on test-set classification levels and can be found in the form of an approximate threshold-neutral measure of discriminative capability in the context of the controlled experiment, see Figure 7 and Table 10.

The maximum AUC value of (0.89) is attained in EfficientNet-B0 in which the classes can be separated better under the standardized experimental conditions. ResNet50 and DenseNet201 have a similar value of AUC (0.88) and it is the confirmation of the consistent predictive behaviour. The correspondence in the AUC, the precision, the recall, and the convergence pattern increase the analytical integrity of the benchmarking system and

guarantee that the relative ranking of an architecture is also consistent in threshold-independent assessment metrics.

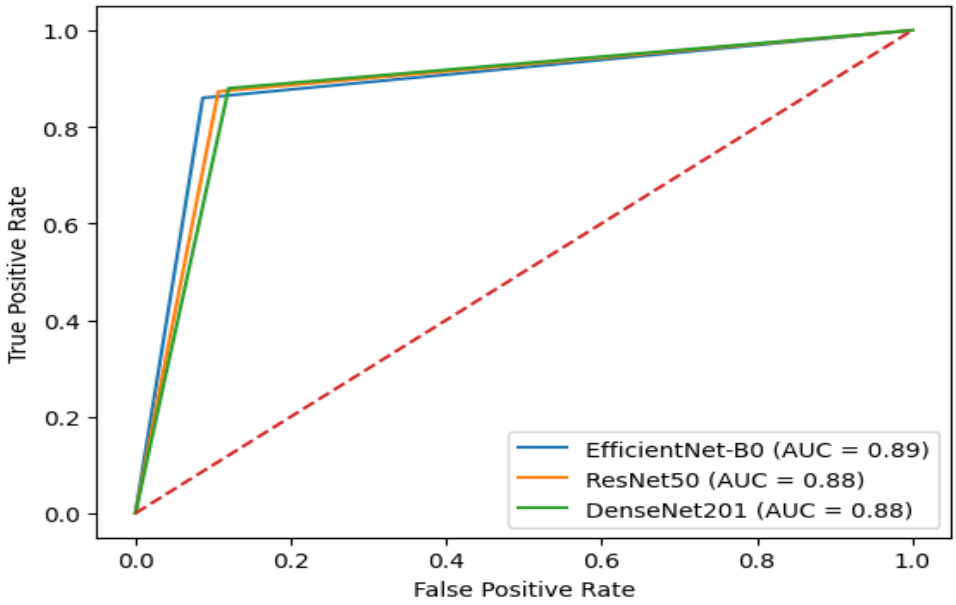


**Figure 7**. ROC Curve Comparison of Top Three CNN Architectures under Standardized Synthetic Conditions

**Table 10**. Comparison of AUC (Area Under the Curve)

| Model | AUC |
|---|---|
| EfficientNet-B0 | 0.89 |
| ResNet50 | 0.88 |
| DenseNet201 | 0.88 |

## *Statistical Test of Reliability*

To calculate the statistical significance of the difference in the classification accuracy observed under the conditions of the standardized experiments one used a two-proportion z-test between the highest-performing architecture (EfficientNet-B0) and the lowest-performing architecture (VGG19) in regards to the size of the test sample (n = 300). In the example of EfficientNet-B0 and VGG19 the correct classification of 266 and 252 samples, respectively, that is, 88.67 and 84 percent, respectively. The pooled proportion was determined in order to estimate the standard error when the null hypothesis that the classification performance of both the models are identical is tested. The calculated z-value was approximately 1.78 that is linked to p-value ≈ 0.075. This difference does not attain to reach a strict statistical significance at the 95% level of confidence ($\alpha = 0.05$). However, the general similarity in architectural ranking realised in most of these independent measures of evaluation, e.g., accuracy, recall, F1-score, and AUC, rate of convergence, and

computational efficiency is a gauge of internal validity of the benchmarking framework. The findings reveal that when the relative margins of accuracy are moderate within those training budgets, which are limited, the systemic effect of architectural setting has predictable convergence dynamics or predictive trade-offs in commonplace synthetic GAN environments.

### *Comparative Interpretation Content Analysis*

The benchmarking framework was designed as a standardized single-run evaluation to ensure architectural comparability under identical training and computational conditions. Despite the controlled experimental setup, the consistency of model ranking across multiple performance indicators strengthens the validity of the observed trends. The relative ordering of architectures remains stable when evaluated using accuracy, precision, recall, F1-score, convergence behaviour, and training time, indicating that the findings reflect intrinsic architectural characteristics rather than isolated metric fluctuations. The observed performance margins—approximately 3–4% variation in classification accuracy, 11–14% variation in precision between the strongest and weakest models, and nearly 46% difference in computational efficiency—represent meaningful comparative distinctions. Furthermore, the inclusion of recall and F1-score analysis confirms that performance differences arise primarily from variations in error distribution patterns rather than sporadic predictive behaviour. The convergence of results across independent evaluation metrics enhances the interpretability and internal consistency of the benchmarking framework. Collectively, these findings suggest that architectural design principles—such as compound scaling, residual connectivity, and dense feature propagation—directly influence convergence efficiency, predictive balance, and computational trade-offs under standardized experimental conditions.

### *Consolidated Observations*

The integrated assessment of classification accuracy, precision, recall, F1-score, convergence dynamics, and computational efficiency reveals distinct architectural trade-offs within the controlled benchmarking environment. The EfficientNet-B0 demonstrates a strong balance between predictive performance and computational efficiency. While its overall classification accuracy is comparable to other top-performing architectures, it achieves faster convergence and the lowest training time, highlighting the benefits of compound scaling in parameter-efficient design. Its slightly higher precision relative to recall suggests a more selective prediction behaviour, favouring reduced false positives. The DenseNet201 exhibits balanced predictive characteristics, with closely aligned precision and recall values, indicating stable class discrimination and effective feature reuse. However, this balanced performance is achieved with higher computational cost compared to EfficientNet-B0. The ResNet50 maintains competitive predictive performance and stable convergence behaviour, consistent with the optimization advantages provided by residual connections, though with moderate computational demand. In contrast, the sequential VGG-based architectures show comparatively slower convergence and

marginally lower predictive performance under identical constraints, reflecting the limitations of deeper sequential designs without enhanced connectivity mechanisms. Overall, the findings reinforce the conclusion that architectural configuration plays a critical role in shaping learning dynamics, predictive stability, and computational efficiency within a controlled CNN benchmarking framework. Importantly, these conclusions are confined to standardized experimental conditions and do not imply deployment-level forensic generalization.

## DISCUSSION

The controlled experimental design can be used to directly compare the convolutional neural network architectures since they both are tested within the same methodological constraints. Differences in performance can be attributed to standardization in dataset composition, strategy in pre-processing, hyperparameters of the transfer learning, hyperparameters of the optimizer, and the number of training steps and can be explained around architectural properties rather than experimental variability. The objective of this paper is not to maximise deployment-level performance, but to target the architectural learning behaviour on a more systematic basis as part of a systematic benchmark. The results indicate that the architectural configuration also contributes significantly to determining convergence in case the optimization regimes are stuck. EfficientNet-B0 which calculates the scaling of depth, width and resolution of compounds is the fastest in converging and competitive in the accuracy of the classification even after only three training epochs. This behaviour suggests that coordinated scaling performs better in the representational power and gradient stability in the situation where optimization budgets are intentionally small. Interestingly, the issues of structural balance appear to be of higher significance than the absolute figures of parameters in improving the effectiveness of early-stage learning.

The predictive behaviour of the use of improved connectivity structures, e.g., DenseNet201 and ResNet50 is stable in a range of evaluation metrics. This is due to the fact that precision, recall and F1-score are closely correlated implying even distribution of errors in the standardized conditions. The gradient propagation and feature reuse are enabled by sparsifying and condensing connections and with their assistance the dynamics of optimization remain stable throughout synthetic GAN-based classification. On the other hand, the convergence speed of sequential architectures such as VGG16 and VGG19 is lower, as is the predictive stability by a small margin. Under constrained training schedules on which gradient flow can be guaranteed by the absence of shortcut or dense connectivity mechanisms can also be used despite hierarchical feature abstraction possibly being effective. The outcomes of such results point out that the depth of architecture is not enough to guarantee the improved adaptation of the controlled benchmarking environments. The Computational efficiency is also another dimension used in distinguishing architectural trade-offs. The advantage of parameter-efficient scaling of EfficientNet-B0 is the lowest training time, which similarly to the rest of the models does

not have an advantage in predictive performance. More complex architectures containing multi-branches and networks which are more tightly coupled are more computationally expensive despite sharing the same classification accuracy. The findings above show a necessity to consider CNN architectures with predictive, and computational metrics. The modelling of detectors based on transformers has also reached a over 93-95% classification accuracy when trained on large-scale datasets (such as Celeb-DF) that are situated within the context of the larger state-of-the-art space. Such models are however typically trained with much higher optimization budgets and longer training schedule.

On the contrary, the current studies focus on intrinsic architectural behaviour by purposefully restricting the training to three epochs in controlled synthetic environments. The accuracy of the CNN models under this weak regime is approximately 88-89 percent and they are also computationally efficient and they converge reliably. Although relative performance is poor compared to large-scale transformer structures, the contribution linked to the methodology is controlled isolation at the trade-off of competitive maximization. The benchmarking framework has internal validity as indicated by consistency of architectural ranking in accuracy, convergence dynamics, precision-recall balance, AUC and training time. Though, due to the small size of the data set and absence of cross-dataset validation, one is not able to extrapolate the research to the real world, these deficiencies were predetermined by the idea to place minimal confounding factors into the study and to make the objective architectural comparison. Overall, these findings point to the fact that architectural design principles, namely, the application of compound scaling and augmented connectivity directly affect convergence efficiency, predictive stability and trade-offs in computational classification of faces in synthetic classification. It is hypothesized that the proposed benchmarking framework will provide a standardized foundation of systematic architectural evaluation and can be applied in future, under controlled experiment conditions, to the creation of hybrid or ensemble models.

## LIMITATIONS

There are some weaknesses of the research which is taking place. Firstly, the experimental evaluation was conducted in a controlled and balanced binary classification setting, which is not a fully representative situation of the deep fake conditions in the adult world. Some additional issues can be added to the real world of forensic environment such as compression artifact, cross domain variations, adversarial manipulations, resolution degradation and temporal inconsistencies in video contents. These have been excluded deliberately to ensure that experimental control and isolation of architectural learning behaviour is achieved. Second, the training time was fixed to three epochs intentionally in order to investigate the dynamics of early convergence rather than long run optimization behaviour. This option of the design facilitates to test the architectural responsiveness in comparison with the limited conditions, however, it does not represent the full convergence ability or prolonged training stability. Third, the study does not provide the cross-dataset validation or robustness test on unrestricted generative models. It implies

that the findings cannot be discussed as a show of the state-of-the-art deep fake detection or production level generalization. Instead, the results should be considered to be the result of a controlled architectural benchmarking in a standard experimental situation.

## FUTURE WORK

Future research can increase the proposed benchmarking module to a larger and more diverse deep fake data, including cross-domain and cross-generators measurement conditions. The time used in training ought to be raised so as to allow a more thorough component of the convergence behaviour in the long run, overfitting and the stability of optimization. Further, when analysing hybrid or ensemble structures empirically, there is also a possibility of being able to find useful information on potential effective integration of complementary strengths that are observed in models. An example is that better predictive stability and computational balance may be had by employing the parameter-efficient designs that contain compound-scaled designs and architectures that enhance connectivity. Other extensions can also be given such as frequency domain properties, compression resistance and post-processing constraints as well as cross-dataset testing which can be given to establish the generalization capability when used in more realistic forensic scenarios. The benchmarking system may be designed through the methodological transparency and standardized assessment guidelines to be a generalizable platform of systematic comparison of developing deep learning designs in the field of research of fake face detection.

## CONCLUSION

The paper illustrated an equivalent and repeatable benchmarking framework of controlled evaluation on representative convolutional neural network (CNN) structures in binary synthetic face recognition. The experimental design minimised confounding variability by ensuring uniform data composition, uniform pre-processing, and an identical initialisation of the transfer learning and equalising the training constraints, and enabled the single-factor isolation of intrinsic architectural learning behaviour

The controlled experimental design can be used to directly compare the convolutional neural network architectures since they both are tested within the same methodological constraints. Differences in performance can be attributed to standardization in dataset composition, strategy in pre-processing, hyperparameters of the transfer learning, hyperparameters of the optimizer, and the number of training steps and can be explained around architectural properties rather than experimental variability. The objective of this paper is not to maximise deployment-level performance, but to target the architectural learning behaviour on a more systematic basis as part of a systematic benchmark. The results indicate that the architectural configuration also contributes significantly to determining convergence in case the optimization regimes are stuck. EfficientNet-B0 which calculates the scaling of depth, width and resolution of compounds is the fastest in converging and competitive in the accuracy of the classification even after only three

training epochs. This behaviour suggests that coordinated scaling performs better in the representational power and gradient stability in the situation where optimization budgets are intentionally small. Interestingly, the issues of structural balance appear to be of higher significance than the absolute figures of parameters in improving the effectiveness of early-stage learning. The predictive behaviour of the use of improved connectivity structures, e.g. DenseNet201 and ResNet50 is stable in a range of evaluation metrics. This is due to the fact that precision, recall and F1-score are closely correlated implying even distribution of errors in the standardized conditions. The gradient propagation and feature reuse are enabled by sportifying and condensing connections and with their assistance the dynamics of optimization remain stable throughout synthetic GAN-based classification. On the other hand, the convergence speed of sequential architectures such as VGG16 and VGG19 is lower, as is the predictive stability by a small margin. Under constrained training schedules on which gradient flow can be guaranteed by the absence of shortcut or dense connectivity mechanisms can also be used despite hierarchical feature abstraction possibly being effective. The outcomes of such results point out that the depth of architecture is not enough to guarantee the improved adaptation of the controlled benchmarking environments. The Computational efficiency is also another dimension used in distinguishing architectural trade-offs. The advantage of parameter-efficient scaling of EfficientNet-B0 is the lowest training time, which similarly to the rest of the models does not have an advantage in predictive performance. More complex architectures containing multi-branches and networks which are more tightly coupled are more computationally expensive despite sharing the same classification accuracy.

The findings above show a necessity to consider CNN architectures with predictive, and computational metrics. The modelling of detectors based on transformers has also reached an over 93-95% classification accuracy when trained on large-scale datasets (such as Celeb-DF) that are situated within the context of the larger state-of-the-art space. Such models are however typically trained with much higher optimization budgets and longer training schedule. On the contrary, the current studies focus on intrinsic architectural behaviour by purposefully restricting the training to three epochs in controlled synthetic environments. The accuracy of the CNN models under this weak regime is approximately 88-89 percent and they are also computationally efficient and they converge reliably. Although relative performance is poor compared to large-scale transformer structures, the contribution linked to the methodology is controlled isolation at the trade-off of competitive maximization. The benchmarking framework has internal validity as indicated by consistency of architectural ranking in accuracy, convergence dynamics, precision-recall balance, AUC and training time. Though, due to the small size of the data set and absence of cross-dataset validation, one is not able to extrapolate the research to the real world, these deficiencies were predetermined by the idea to place minimal confounding factors into the study and to make the objective architectural comparison.

Overall, these findings point to the fact that architectural design principles, namely, the application of compound scaling and augmented connectivity directly affect convergence

efficiency, predictive stability and trade-offs in computational classification of faces in synthetic classification. It is hypothesized that the proposed benchmarking framework will provide a standardized foundation of systematic architectural evaluation and can be applied in future, under controlled experiment conditions, to the creation of hybrid or ensemble models**.**

## AUTHOR CONTRIBUTIONS:

"Conceptualization, R.C.; Methodology, R.C.; Validation, R.C., M.S. and S.A.; Investigation, R.C.; Resources, M.S., and S.A.; Data Curation, R.C.; Writing – Original Draft Preparation, R.C.; Writing – Review & Editing, R.C.; Visualization, R.C.; Supervision, M.S., and S.A.; Project Administration, M.S.

## CONFLICT OF INTERESTS

The authors have no competing interests to declare that are relevant to the content of this research paper.

## REFERENCES

1. Burgess, Burgess, M. The biggest deep fake abuse site is growing in disturbing ways. Available from: https://www.wired.com/story/deepfake-nude-abuse/ (Accessed on 14.11.2025)
2. Guo, H., Hu, S., Wang, X., Chang, M.-C., Lyu, S. Eyes tell all: Irregular pupil shapes reveal GAN-generated faces. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, **2022**, pp. 2904–2908.
3. Say, T., Alkan, M., Kocak, A. Advancing GAN Deepfake Detection: Mixed Datasets and Comprehensive Artifact Analysis. *Appl. Sci.* **2025**, *15*, 923
4. McCloskey, S., Albright, M. Detecting GAN-generated imagery using saturation cues. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, **2019**, pp. 4584-4588.
5. Guo, H., Hu, S., Wang, X., Chang, M.-C., Lyu, S. Robust attentive deep neural network for exposing GAN-generated faces. *IEEE Access* **2022**, *10*, 35918–35927.
6. Mo, H., Chen, B., Luo, W. Fake faces identification via convolutional neural network. In *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security*, Innsbruck, Austria, **2018**, pp. 67–72.
7. Karras, T., Aila, T., Laine, S., Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. *Sixth International Conference on Learning Representations*. Vancouver, Canada. **2018**, pp. 1-26
8. Mi, Z., Jiang, X., Sun, T., Xu, K. GAN-generated image detection with self-attention mechanism against GAN generator defect. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 958–968.
9. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, South Korea, **2019**, pp. 1–11.

10. He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, **2016**, pp. 770–778.
11. Zhang, K., Yao, T., Ding, S., Li, J., Zeng, D., Sun, Y. Detecting GAN-generated images using color cues. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, **2019**, pp. 4584–4588.
12. Kundu, R., Balachandran, A., Roy-Chowdhury, A.K. TruthLens: Explainable Deep Fake Detection for Face-Manipulated and Fully Synthetic Data. *arXiv* **2025**, arXiv:2503.15867.
13. Liu, W., Qiu, J., Boumaraf, S., Lin, C., Pan, L., Li, L., Bennamoun, M., Werghi, N. BENet: A cross-domain robust network for detecting face forgeries via bias expansion and latent-space attention. *arXiv* **2024**, arXiv:2412.07431.
14. Bargavi, S.K.M., Rathi, R. Beyond faces: A novel approach to deep fake detection and classification. *Asian J. Appl. Sci. Technol.* **2024**, *8*, 47–60.
15. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C.C. The deep fake Detection Challenge (DFDC) dataset. *arXiv* **2020**, arXiv:2006.07397.
16. Tolosana, R., Romero-Tapiador, S., Vera-Rodriguez, R., Fierrez, J. Deep fakes and beyond: A survey of face manipulation and fake detection. *Inf. Fusion* **2020**, *64*, 131–148.
17. Xu, Y., Zhang, H., Li, Z., Wang, X. Identity-aware vision–language model for explainable face forgery detection. *arXiv* **2025**, arXiv:2504.09439.
18. Zhang, Y., Liu, J., Chen, K. VLF-FFD: A vision–language fusion solution for face forgery detection. *arXiv* **2025**, arXiv:2505.02013.
19. Li, J., Wang, R., Zhou, M. Towards general visual–linguistic face forgery detection. *arXiv* **2025**, arXiv:2502.20698.
20. Chen, K., Huang, Y., Sun, L. VLForgery Face Triad: A visual–language benchmark for forgery detection, localization, and attribution. *arXiv* **2025**, arXiv:2503.06142.
21. Wang, S., Wang, O., Zhang, R., Owens, A., Efros, A.A. CNN-generated images are surprisingly easy to spot… for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, **2020**, pp. 8695–8704.
22. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the International Conference on Machine Learning (ICML)*, **2020**, pp. 1-12.
23. Kumar, S., Singh, A., et al. DeiTFake: Deepfake detection model using DeiT multi-stage training. *Array* **2026,** *29*, 100734.