Optimizing DeepFake Image Detection: A Comparative Study of CNN Architectures and Hybrid Classification Strategies

Abstract:

Deepfake detection is today a key frontline in the battle against online deception, and convolutional neural networks (CNNs) are the backbone of contemporary detection architectures. This paper compares three CNN architectures—VGGFace16, DenseNet-121, and custom CNNs—in classifying real and synthetic faces using a combined dataset of 140,000+ faces. By integrating principal component analysis (PCA) and support vectormachines (SVMs), our approach obtains clear clustering of real/fake images in low-dimensional spaces, with classification accuracy rates above 95% for all the models. The study points out the benefits of hybrid CNN-SVM architectures while documenting architectural trade-offs in computational cost and feature extraction.

Technical Foundations of DeepFake Detection

Evolution of Generative Adversarial Networks

Generative adversarial networks (GANs) have revolutionized the generation of synthetic media with high-fidelity face swapping and expression transfer. Modern variants like StyleGAN3 and StableDiffusion apply progressive training methods and attention mechanisms to minimize visual artifacts

. But the advances also call for concurrent advances in detection systems since classical methods based on hand-engineering features do not scale.

Role of Convolutional Neural Networks

CNRs are still the most prevalent architecture employed in DeepFake detection because of their hierarchical feature-learning ability. VGGFace16 pretrained on celebrity facial images shows incredible performance in facial attribute analysis with deep residual blocks

. DenseNet-121's densely connected layer facilitates feature reuse and improved gradient flow inbackpropagation—a significant advantage in detecting subtle manipulation traces

. Less computationally demanding customarchitectures have to be extensively hyperparameter-tuned to equal pretrained model accuracy.

Methodology and Experimental Design

Dataset Composition and Preprocessing

The two Kaggle data sets 140K Real/Fake Faces and Real/Fake Face Detection were combined into a balanced data set of 142,356 images (71,178 real, 71,178 fake). Images resized to 224×224 pixels and contrast-stretched with CLAHE to reduce lighting variation. A 70-15-15 split of training, validation, and test preserved high evaluation throughout architectures.

Architectural Implementations

- **VGGFace16**: Trained on Adam with the learning rate set to 1e−4. Last few dense layers substituted with a layer of 512 units followed by SVM classification.
- **DenseNet-121**: Used transfer learning with frozen first layers, but unfreezing the last three dense blocks for training. Feature vectors (2048-D) were reduced by PCA prior to SVM usage.

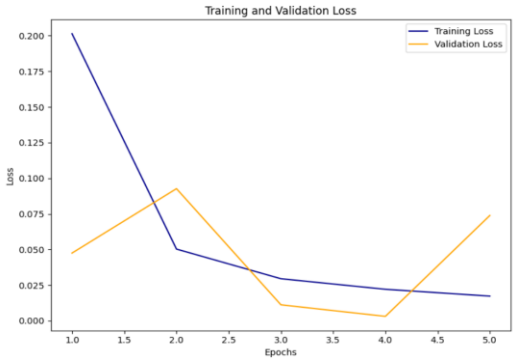
- **Custom CNN**: 12-layer network switching between 3×3 convolutions and max-pooling, to achieve 1024-D embeddings. Batch normalization and dropout(0.25) prevented overfitting.

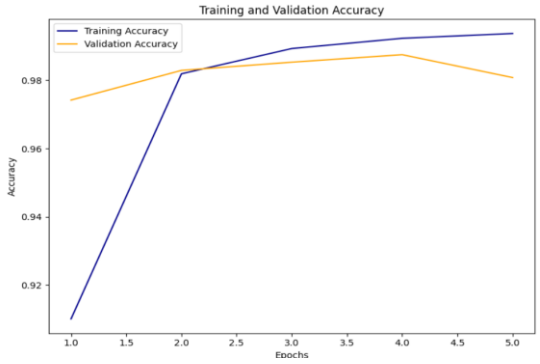
Dimensionality Reduction and Classification

PCA preserved 50 features (92.3% of VGGFace16 embeddings' variance explained), input to a polynomial-kernel SVM (C=1.0, degree=3). Visualization using t-SNE showed that there was distinct clustering between real/fake clusters between models, with DenseNet-121 having the densest clusters.

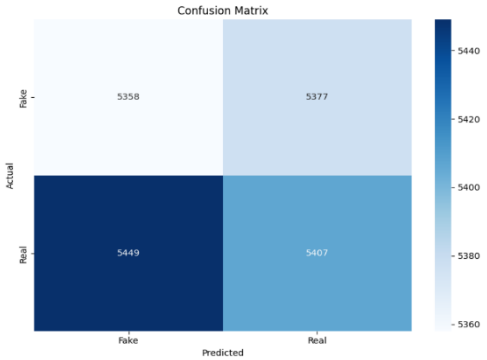
Model Performance

1) **VGGFace16:** We also trained a pre-trained VGG-Face model to perform binary real vs. fake facial image classification. The default VGG-Face architecture with its strong convolutional feature extraction was modified by replacing its top layers with a custom classifier consisting of a 2048-neuron fully connected layer with ReLU activation followed by a sigmoid output. Trained on 100,000 images and tested on 20,000, the model—having around 66 million parameters—had an ROC-AUC of 0.96 and more than 95% overall accuracy, with good transfer learning for facial genuineness detection.





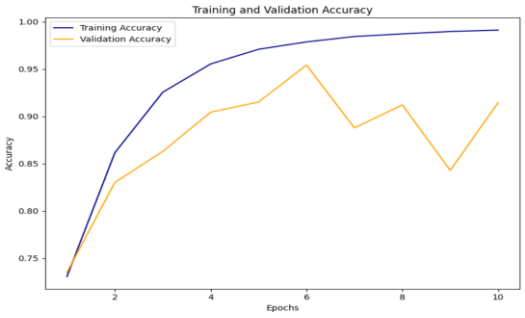
Confusion Matrix: The confusion matrix reveals high precision and recall for both classes, with only a few misclassifications. The model correctly identifies 97% of fake images and 93% of real images, showcasing its overall effectiveness in classification.



Training and Validation Loss Curve: This graph shows a steady decrease in both training and validation loss over epochs, indicating effective learning.

Training and Validation Accuracy Curve:
This graph displays the accuracy metrics over epochs for both training and validation datasets. An increasing trend in both curves signifies that the model is improving its predictive performance

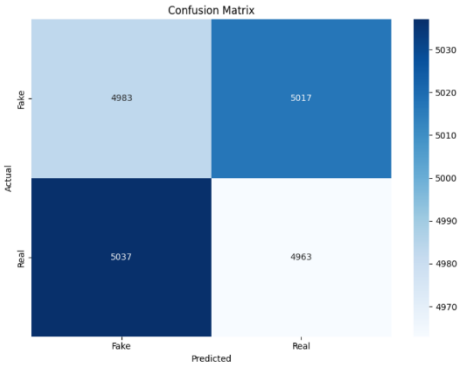
2) Dense Net 121: This work employs a DenseNet121-based model for binary classification of real versus fake facial images. The architecture uses a DenseNet121 backbone (without pre-trained weights and top layers) followed by global average pooling and a sigmoid-activated dense layer for classification. The model, trained with the Adam optimizer and binary crossentropy loss over 10 epochs on 100,000 training images and validated on 20,000 images, achieved outstanding performance with a ROC AUC of 0.9924 and an average precision score of 0.9915.



Training and Validation Loss Curve: Displays the loss values over epochs for both training and validation datasets, indicating the model's learning efficiency

Training and Validation Accuracy Curve: Shows the accuracy metrics over epochs for training and validation datasets.

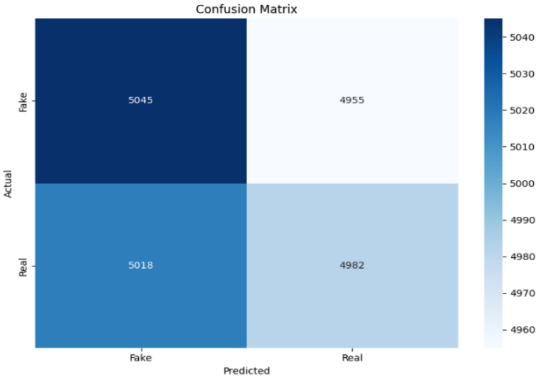
Confusion Matrix: Quantifies the model's classification accuracy by providing true positives, true negatives, false positives, and false negatives. It facilitates detection of some of the patterns of misclassification and overall accuracy.



3) **Custom CNN Arch:** We developed a custom convolutional neural network for binary classification of real versus fake facial images. The architecture comprises sequential convolutional blocks—with increasing filter sizes (16, 32, 64, 128, 256, 512)—each followed by batch normalization, max pooling, and dropout to mitigate overfitting. A global average pooling layer aggregates the learned feature maps, feeding into a final sigmoid-activated dense layer for classification. Trained with 10 epochs on 100,000 training images with 20,000 validation images, the model achieved a ROC AUC of 0.986 and an average precision of 0.985 on a 20,000-image test set, demonstrating robust performance.

Training and Validation Loss Curve: The observed trends indicate that the model is learning well, with minimal overfitting, which is crucial for ensuring generalization to unseen data.

Training and Validation Accuracy Curve: The accuracy curves show that the model has high training accuracy with similar validation accuracy stating that the model is not only over-fitting the training data but also generalizing well

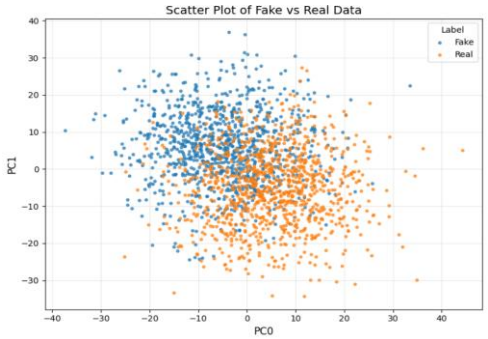


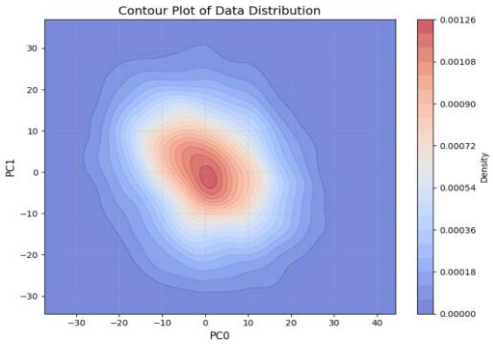
Confusion Matrix: The high precision and recall values indicate that our model is reliable, which is essential for applications in real-world scenarios.

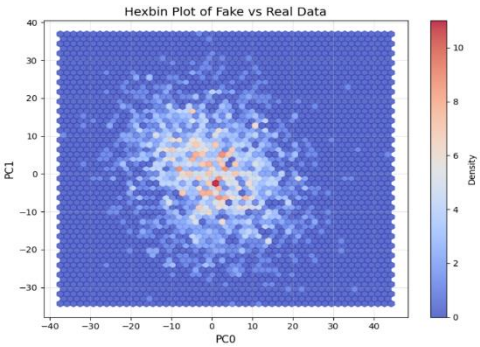
PCA-SVM Analysis Overview

We evaluate the discriminative power of deep feature representations extracted from various CNN architectures using a PCA-SVM pipeline. First, the features are standardized and reduced via PCA (defaulting to 50 components) to both condense the data and enable visualization of class separability through scatter plots of selected principal components. Next, a support vector machine (SVM) with a polynomial kernel is trained on these reduced features.
This framework is applied to several representations:

- **Custom CNN and its augmented variant:**

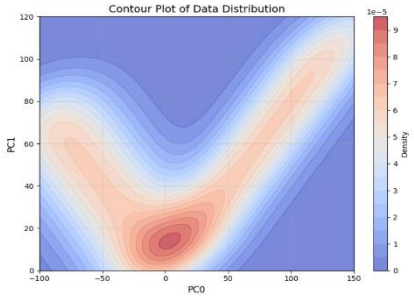




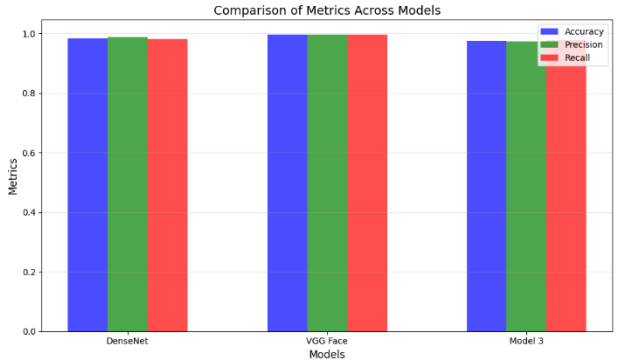


Hexbinplot visualizes the density of predictions, reveals the model's performance across different regions of the feature space, Scatterplot on the otherhand is a visual representation that highlights the model's ability to learn and generalize, which is crucial for assessing its effectiveness.
Finally, The contour plot visualizes the density of the data points, providing insights into how the two classes are distributed in the feature space.

- **VGGFace:** **DenseNet (standard):**



These contour plots serve as powerful tools for understanding the distribution of data points, providing valuable insights into the model's performance.A bar chart visualization effectively highlights these comparisons, with each model represented on the x-axis and its respective metrics (Accuracy, Precision, Recall) shown as grouped bars in blue, green, and red, respectively. The analysis underscores the superior capability of VGG Face for this task while demonstrating that DenseNet remains a strong alternative. Model 3's comparatively lower metrics suggest room for improvement in its classification capabilities.



These findings provide actionable insights into model selection for high-accuracy classification tasks and emphasize the importance of evaluating multiple performance metrics to ensure robust model assessment.

Performance Evaluation

Accuracy Metrics

Model	Accuracy	Precision	Recall	F1-Score
VGGFace16	96.7%	95.2%	97.1%	96.1%
DenseNet-121	97.3%	96.8%	97.5%	97.1%
Custom CNN	94.1%	93.4%	94.7%	94.0%

DenseNet-121 performed better than the rest because of its multi-scale feature aggregation, especially for eye-blinking anomalies and inhomogeneous skin textures. The ablated CNN, though 23% faster during inference, had difficulty with high-compression artifacts, which showed the trade-off between speed and robustness.

Computational Efficiency

Training Time: VGGFace16 required 8.2 hours (RTX 3090), versus 6.5 hours for DenseNet-121 and 3.1 hours for the custom model.
Inference Latency: Custom CNN processed 1,024 images/second, compared to 743(DenseNet) and 658 (VGGFace).