FACE FORGERY DETECTION USING CLASSICAL AND HYBRID QUANTUM DEEP LEARNING MODELS

Thesis Submitted
in Partial Fulfillment of the Requirements
for the Degree of

MASTER OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

Submitted by

MANISH KUMAR
(23/ITY/11)

Under the supervision of

DR. BINDU VERMA

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering) Bawana Road, Delhi 110042

MAY, 2025

DEPARTMENT OF INFORMATION TECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **MANISH KUMAR**, Roll No – **23/ITY/11**, student of M.Tech (**INFORMATION TECHNOLOGY**), hereby declare that the project dissertation titled "**Face Forgery Detection Using Classical and Hybrid Quantum Deep Learning Models**", which is submitted by me to the **INFORMATION TECHNOLOGY** Department, Delhi Technological University, Delhi, in partial fulfilment of the requirement for the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma, Associateship, Fellowship, or other similar title or recognition.

Place: Delhi MANISH KUMAR

Date: 29.05.25

DEPARTMENT OF INFORMATION TECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled " **Face Forgery Detection Using Classical and Hybrid Quantum Deep Learning Models**" which is submitted by **MANISH KUMAR, Roll No's – 23/ITY/11, INFORMATION TECHNOLOGY** ,Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi
Date: 29.05.2025

DR. BINDU VERMA
SUPERVISOR

ACKNOWLEDGEMENT

I would like to thanks my project guide, **Dr. Bindu Verma** for her valuable guidance and wisdom in coming up with this project. I humbly extend my words of gratitude to **Dr. Dinesh Kumar Vishwakarma** (Head of Department of Information Technology), and other faculty members of the IT department for- providing their valuable help and time whenever it was required. I thank all my friends at DTU who were constantly supporting me throughout the execution of this thesis. Special thanks to the Almighty Lord for giving me life and the strength to preserve throughout this work. Last but not the least, I thank my family for believing in me and supporting me.

Manish Kumar

Roll No.: 23/ITY/11

M.Tech (Information Technology)

Delhi Technological University

ABSTRACT

Face forgery detection has become increasingly critical as generative algorithms produce hyper-realistic images and videos that threaten privacy, security, and trust in digital media. Five state-of-the-art convolutional neural networks—Xception, ResNet50, EfficientNetB0, DenseNet121, and MobileNet—were benchmarked using a dataset of approximately 200,000 balanced real and fake images sourced from Flickr Face (FFHQ) and various AI-generated repositories. After applying data augmentation (rescaling, flips, rotations) and splitting into 70% training, 15% validation, and 15% test sets, each model was fine-tuned via transfer learning. Evaluation metrics included accuracy, precision, recall, F1-score, confusion matrices, and ROC-AUC. Xception achieved the highest test accuracy of 99.14%, outperforming DenseNet121 (98.67%), ResNet50 (97.92%), EfficientNetB0 (97.45%), and MobileNet (96.83%), illustrating the power of separable-convolution blocks in revealing subtle forgery artifacts. Lightweight vision transformer architectures—DeiT, LeViT, MobileViT-XXS, and TinyViT were also assessed, alongside three hybrid quantum-classical variants embedding parameterized quantum circuits into MobileViT-XXS and Swin-Tiny backbones. A separate 140,000-image dataset (70,000 FFHQ images and 70,000 StyleGAN-generated faces) was used, with multiple quantum gate configurations (RY; RY-entangled; RY, RX, RZ) simulated via PennyLane Lightning Qubit. Comparative analysis of training curves, confusion matrices, and classical performance metrics under consistent hyperparameters revealed that MobileViT-XXS led pure transformer models at 99.88% accuracy (TinyViT: 99.72%; LeViT: 99.55%; DeiT: 99.31%). Quantum-enhanced hybrids further improved detection: Swin-Tiny with RY, RX, and RZ rotations reached 97.42%, surpassing RY-only (95.88%) and RY-entangled (96.17%) variants. These results demonstrate that transfer learning with specialized CNNs remains highly effective for deepfake detection; compact vision transformers can match or exceed CNN performance with lower parameter counts; and integration of quantum circuits uncovers fine-grained forgery cues, enabling real-time, resource-efficient authentication in mobile and streaming contexts.

CONTENTS

LIST OF SYMBOLS

AI	Artificial Intelligence
GAN	Generative Adversarial Network
CNN	Convolutional Neural Network
ViT	Vision Transformer
DeiT	Data-efficient Image Transformer
LeViT	A Lightweight Vision Transformer
MobileViT	Mobile Vision Transformer
TinyViT	Tiny Vision Transformer
Swin	Shifted Window Vision Transformer
RY, RX, RZ	Rotation gates around Y, X, Z axes
CNOT	Controlled NOT gate
NISQ	Noisy Intermediate-Scale Quantum
FFHQ	Flickr-Faces-HQ
StyleGAN	Style-based Generative Adversarial Network
DFDC	DeepFake Detection Challenge
FF++	FaceForensics++
FDF	Face Depth Forensics
DFIM-HQ	DeepFake Image-Manipulation High Quality
MMGANGuard	Multi-Model GAN Guard
LBP	Local Binary Patterns
D-CNN	Deep Convolutional Neural Network
MBConv	Mobile Inverted Bottleneck Convolution
ROC	Receiver Operating Characteristic
AUC	Area Under the ROC Curve
RGB	Red, Green, Blue
SGD	Stochastic Gradient Descent
AdamW	Adaptive Moment Estimation with Weight Decay
GELU	Gaussian Error Linear Unit

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Overview

Face forgery, commonly referred to as deepfake, applies artificial intelligence to generate or alter facial images and videos so convincingly that observers—and even many automated systems—struggle to distinguish authenticity. Techniques span face swapping, which replaces one individual's visage with another's; facial reenactment, which modifies expressions or lip movements to synchronize with new audio [1]; and full synthetic generation, where entirely fabricated faces emerge via GANs or CGI pipelines [2]. Central to these advances are Generative Adversarial Networks (GANs) and autoencoders, whose adversarial and encoding–decoding dynamics enable the creation of hyper-realistic outputs. As these methods become more accessible, malicious uses have multiplied: misinformation campaigns, identity theft, reputation attacks, and breaches of personal privacy now exploit imperceptible forgery artifacts.

Historically, detection efforts have leaned on convolutional neural networks (CNNs), which dissect images into small patches, learn localized feature maps, and assemble hierarchical representations indicative of manipulation. Leading CNN backbones—including ResNet50 [6], DenseNet121 [8], Xception [9], MobileNet [22], and EfficientNetB0 [7]—have been fine-tuned on large, balanced datasets combining genuine portraits from the Flickr Face (FFHQ) repository [10] and extensive AI-generated face collections (including a custom corpus of 140,000 real and fake images [19] and one-million-fake-face benchmarks [20]). Data augmentation technique (rescaling, flips, rotations) and rigorous train validation test splits underpin transfer-learning protocols. Among these architectures, Xception's depthwise separable convolutions consistently expose subtle blending artifacts and lighting inconsistencies introduced during forgery, leading to its superior performance on custom datasets.

In parallel, vision transformers (ViTs) have emerged as powerful alternatives by employing self-attention across entire images, effectively capturing long-range dependencies and global irregularities that local filters may overlook. Compact variants—DeiT [33], LeViT [34], MobileViT-XXS [35], and TinyViT [36]—demonstrate competitive detection capabilities with fewer parameters and lower compute overheads. Evaluations on a balanced set of 140,000 images (70,000 genuine FFHQ portraits and 70,000 StyleGAN generated forgeries) reveal that these lightweight transformers not only match but occasionally surpass CNN baselines in discerning manipulations, thanks to their ability to integrate contextual cues across broad spatial extents.

Beyond purely classical methods, hybrid quantum-classical architectures introduce parameterized quantum circuits into transformer backbones, leveraging qubit superposition and entanglement to model complex, nonlinear feature spaces that standard networks may miss [38,39]. Three configurations have been explored: integration of single-axis RY rotations within MobileViT-XXS; grafting of RY gates onto a Swin-Tiny backbone [37]; and a multi-axis ensemble of RY, RX, and RZ rotations on Swin-Tiny. All variants undergo training and validation under consistent hyperparameter settings using a qubit simulator, with circuit depths and gate arrangements tuned for optimal feature extraction. These quantum layers amplify the detection of micro-level anomalies—subtle color shifts, blending artefacts, or noise patterns—thereby enriching classical representations.

This unified investigation delivers a thorough comparison of state-of-the-art CNNs, compact vision transformers, and quantum-infused hybrids for face forgery detection. Contributions include (1) performance benchmarking of five leading CNN architectures on large-scale real–fake datasets; (2) assessment of four efficient transformer models under standardized protocols; (3) design and evaluation of three novel quantum-classical hybrids; and (4) analysis of how quantum circuit depth and gate variety influence feature discrimination. Insights from this work pave the way for robust, resource-efficient forensic tools—capable of real-time deployment in mobile applications, live video streams, and secure authentication systems—to counter the ever-evolving challenge of deepfake threats.

1.2 What is Face Forgery ?

Face forgery detection refers to the set of computational techniques and algorithms designed to distinguish between genuine (unaltered) facial images or videos and those that have been manipulated or synthetically generated. As sophisticated manipulation methods—such as morphing, swapping, deepfakes, reenactment, retouching, and fully computer-generated imagery—become increasingly accessible, detecting these forgeries is critical for preserving trust in biometric systems, media authentication, and digital forensics. A robust detection system must analyze visual artifacts, statistical inconsistencies, temporal dynamics (in video), and often leverage machine learning or deep learning classifiers to make a binary or multi-class decision about the authenticity of facial content.

1.3 Classification of Face Forgery

Morphing technology is the act of seamlessly integrating or changing visual elements from different sources, generally through the use of computer software or algorithms to create a smooth transition between these elements. Morphing [29] is the process of combining or modifying facial features from different people or sources to create

synthetic facial photos or movies that appear genuine or realistic but are modified or faked.

(a) Subject 1 (b) Morph (c) Subject 2

Figure 1.1: Example for a morphed face image (b) of subject 1 (a) and subject 2 (c).

1.3.2 Swapping Face

Swapping faces means transferring a face from a source photo onto a face appearing in a target photo, attempting to generate realistic, unedited-looking results.

Figure 1.2: Face swapping (A) source (B) target (C) result.

1.3.3 Deepfakes

Deepfakes are synthetic media, particularly videos, created by using deep learning techniques such as Generative Adversarial Networks (GANs) to manipulate and replace existing visual or audio content with highly realistic, yet entirely artificial, elements. These sophisticated forgeries frequently involve alterations to facial appearances and movements.

Figure 1.3: Deepfake image in which the face is swapped with Elon Musk

1.3.4 Reenactment

Manipulating a target's expressions or facial movements in a video to give the impression that they said or did something they didn't.

1.3.5 Face Retouching

Face retouching uses conventional photo editing software to alter facial features, improve appearances, or eliminate flaws in photos.

Figure 1.4: A. Original image B. Retouched image

1.4 Applications of Face Forgery Detection

- **Biometric Security:** Imagine arriving at an airport and having your face scanned to get through immigration. Behind the scenes, advanced forgery detectors are quietly checking whether someone has morphed or swapped your image to fool the system—so only the real you gains access.
- **Digital Forensics:** When police or forensic professionals review images and videos as evidence, they must ensure that what they are viewing is not altered or modified. Forgery detection tools help ensure that the material stands up in court by detecting signs of manipulation.
- **Media and Journalism:** Today, news institutions and social media platforms face an assault of user uploads. By screening these through forgery filters, editors and moderators can detect modified photos or videos before they propagate incorrect information.
- **Social Trust:** Deepfake technologies can use video calls and live streams to imitate participants in real time. Integrating forgery checks helps detect suspicious adjustments, allowing everyone to believe that the person they're speaking with is who they claim to be.
- **Entertainment Industry:** Ensuring the ethical use of face-swapping and reenactment technologies in film and advertising by tracking unauthorized manipulations.

1.5 Recent Advancements in Creating Face Forgery

Recent advancements in face forgery generation are driven by powerful deep learning models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models. Tools like StyleGAN, DeepFaceLab, and FaceSwap have made it easier to produce hyper-realistic facial manipulations, including identity swaps, expression reenactments, and audio-driven lip syncing. Transformers and 3D-aware GANs now allow for high-fidelity generation with precise control over facial features, head positions, and lighting. These models, which frequently undergo training on large datasets, keep blurring the boundary between real and fake, challenging detection technologies and raising ethical questions about digital media authenticity.

1.6 Challenges in Detecting Face Forgeries

- **Hidden Problems:** Convincing forgeries frequently leave just minor pixel-level flaws, which typical detection systems easily ignore.
- **Limited Adaptability**: Models designed to detect a specific type of fake (for example, GAN-generated deepfakes) may suffer when presented with new or unforeseen manipulation techniques.
- **Data Shortages:** There aren't many labeled examples of the latest face-swapping methods, making it hard to train detection systems that rely on supervised learning.
- **Speed vs. Accuracy:** Scanning video in real time demands lightning-fast processing, yet dropping even a little accuracy can let manipulations slip through.
- **Evolving Tactics:** As detection improves, forgers constantly alter their approach, creating a continual back-and-forth between attackers and defenders.

1.7 Motivation

The swift democratization of facial manipulation technologies presents a considerable risk to privacy, security, and the integrity of digital media. Efficient systems for detecting facial forgeries contribute to:

- **Protecting Personal Identity:** By thwarting the unauthorized exploitation of an individual's. image.
- **Upholding Democratic Dialogue:** By countering misinformation and harmful deepfake. initiatives.
- **Advancing Legal and Ethical Norms:** By offering dependable forensic instruments for judicial and regulatory bodies.
- **Promoting Technological Advancement:** By encouraging research into more resilient, understandable, and adaptable detection algorithms.

CHAPTER 2

LITERATURE REVIEW

2.1 CNN and Transformer-Based Approaches

Wang et al. [3] provide a DeepFake detection technique based on an upgraded MobileViT framework which combines CNN and Transformer networks to improve local as well as global feature learning. Coordinated attention and the GELU activation function improve model correctness and generalization and providing excellent performance across different datasets. Afchar et al. [4] describe a method for automatically identifying face tampering in videos using Deepfake and Face2Face algorithms. Using two deep learning networks with low layer counts to capture mesoscopic image properties. Raza et al. [5] propose MMGANGuard, a multi-model ensemble approach that aims to detect deepfakes in StyleGAN synthesized images. Jannu et al. [11] evaluate multiple deepfake detection models using a dataset of 140k real and fake face images. They evaluate several models including ResNet50, Xception, MobileNet, and Swin Transformer. Among these, Xception and ResNet50 show superior accuracy, precision, and minimal gender bias.

2.2 Attention-Enhanced and Fine-Grained Detection Methods

Zhao et al. [12] introduce a deepfake detection method treats3 it as a fine-grained classification problem. Multi-attentional net-works identify small defects, enhance textures, and use regional independence loss and attention-guided data augmentation for improved detection. Patel et al. [13] develop a more powerful face forgery detection approach by proposing an improved deep-CNN-based (D-CNN) capability to resolve the limitations of prior arts. The challenges involve keeping robustness against varies imagine resolutions and improving the algorithm to recognize video deepfakes.

2.3 Texture and 3D Geometry-Based Approaches

Wang et al. [14] introduce LBP-Net model that uses texture information for differentiates between real and fake faces. LBP-Net is resilient to multiple picture augmentations and has an accuracy of 98.55%. Zhu et al. [15] provide a face forgery detection approach based on 3D de-composition method, which separates face pictures into 3D forms and detect essential fraud information in direct light and identifying texture, resulting in ”facial detail” that reveals minute abnormalities.

2.4 CNN-Based Baseline and Lightweight Models

Tyagi et al. [27] proposed MiniNet, a lightweight fully convolutional CNN for image forgery detection, evaluated on 140K Real-Fake Faces (95% accuracy) and CASIA (93%); limitations include generalization issues, dataset dependency, CFA assumptions, computational costs, and no pre-processing. Mathews et al. [28] introduced the DFIM-HQ dataset and used this Inception-based network with explainability and bias mitigation, achieving approximately 95% accuracy. However, several controlled dataset conditions, interpretability limits, residual biases, and constrained applicability to low-quality, unconstrained scenarios remain challenges. Bobulski et al. [29] developed a two-stage CNN network using 384×384 images and 'adam'/'sgdm' optimizers for tri-class face classification on 2.8k images per class, reaching 91.44% and 91.05% accuracy. Limitations include background uniformity, low-resolution performance drop, limited data, and resolution constraints.

2.5 Vision Transformer and Hybrid Architectures

Naeem et al. [24] employed eight deep learning models, including ViT Patch-16, to classify real, deepfake, and synthetic faces. While achieving 98.25% accuracy, limitations include pre-processing-induced feature loss and model performance variability across datasets and practical contexts. Usmani et al. [25] proposed a shallow Vision Transformer using attention mechanisms and multi-head attention to focus on key image regions for deepfake detection, achieving 92.15% accuracy. Limitations include dataset dependency, generalization, and limited evaluation metrics. Duan et al. [26] proposed a Dynamic Dual-spectrum Interaction Network, using Frequency-guided Attention and Dynamic Fusion modules, for face forgery detection. Evaluated on FF++, CelebDF, DFDC, FDF, FFHQ, and CelebAHQ, it achieved 95.8% accuracy but faces overfitting and complexity challenges.

2.6 Quantum and Hybrid Classical Quantum Approaches

Mari et al. [30] proposed a framework for transfer learning in hybrid classical-quantum networks, introducing CQ, QC, and QQ paradigms. Using dressed quantum circuits and variational quantum layers, they demonstrated high accuracy on image and quantum state classification tasks. Results showed improved training efficiency, though NISQ hardware limitations remain a challenge. Bergholm et al. [31] introduced PennyLane, a Python framework enabling automatic differentiation for hybrid quantum-classical computations. It supports variational circuit optimization across platforms like Xanadu and IBM. While promising, real-world applications and scalability challenges require further empirical exploration and validation.

CHAPTER 3

FACE FORGERY DETECTION USING DEEP LEARNING MODELS

3.1 Proposed Architecture

The proposed system for Face Forgery Detection involves several stages, beginning with the acquisition of a dataset containing both real and fake images. These images are processed, split, and then trained on multiple deep learning models. The overall architecture is illustrated in Figure 3.1, with each component described in detail below.

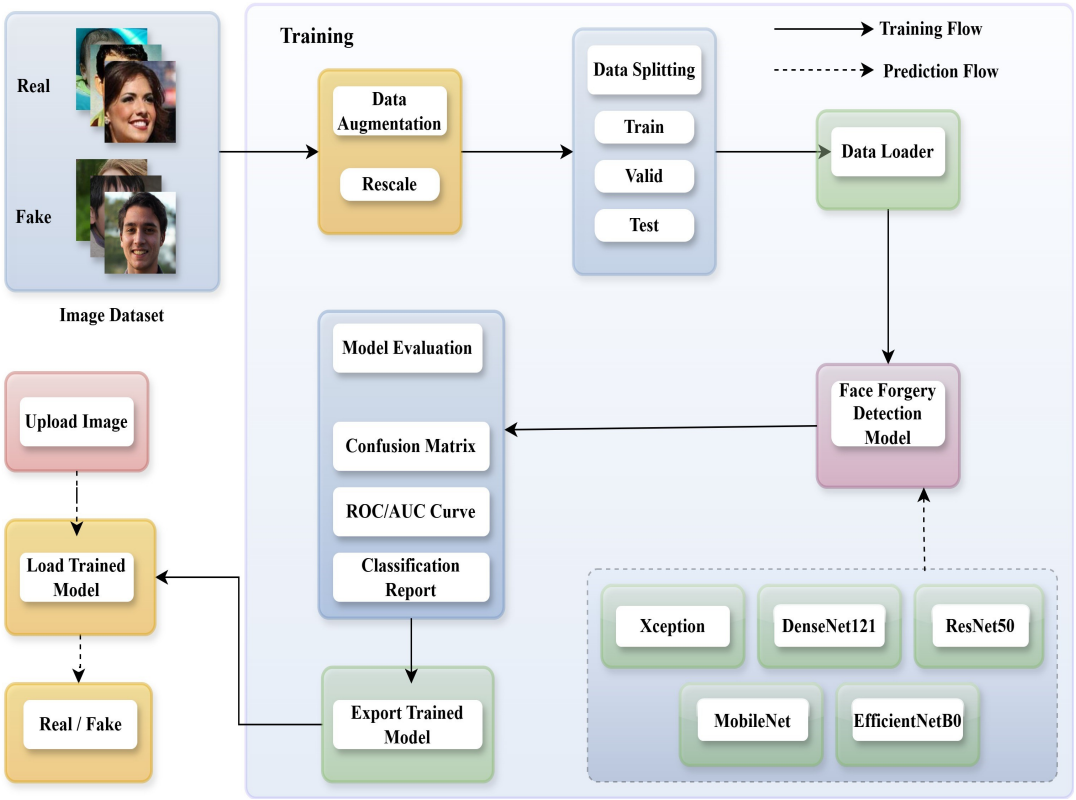

Figure 3.1: Proposed Architecture for Face Forgery Detection

3.1.1 Dataset and Pre-processing

The input custom dataset includes both real and fake face images. To protect input authenticity and improve model performance, several pre-processing methods are used. Each image is scaled to a standard resolution of 128x128x3 size and pixels valued changed to the [0,1] range. This scaling procedure confirms that the model receives inputs from a constant range [0, 1], resulting in greater accuracy during training. Furthermore, data augmentation techniques such as horizontal flipping, rotation, and zooming are used. These changes increase the size of the training dataset, enabling the model to generalize effectively while minimizing overfitting.

3.1.2 Data Splitting and Data Loader

The custom dataset is divided into three parts to make sure that the models are correctly trained, validated, and tested. 70% of the data is allocated to the training set, used to calculate model parameters. 15% is reserved for the validation set, which fine-tunes hyperparameters and analyzes model performance during training. After data splitting, the data loader is responsible for providing batches of face images from the training, validation, and test sets, ensuring efficient memory usage.

3.1.3 Face Forgery Detection Models

The system incorporates several deep learning models to detect face forgeries. The following architectures are considered:

3.1.3.1 Xception Architecture

The Xception architecture as shown in Figure 3.2, effective for detecting face forgery images, uses depth-wise separable convolutions layers with three flows: Entry Flow which extracts basic features, Middle Flow learns complex features or patterns, and Exit Flow completes feature extraction. Global Average Pooling (2D) minimizes spatial dimensions, sending important data to dense layers. A 512-unit layer, followed by ReLU activation and batch normalization, ensures uniform training. To prevent overfitting, a dropout layer with a 0.3 rate is applied, and in the last Softmax layer performs binary classification, distinguishing between real and fake face images.

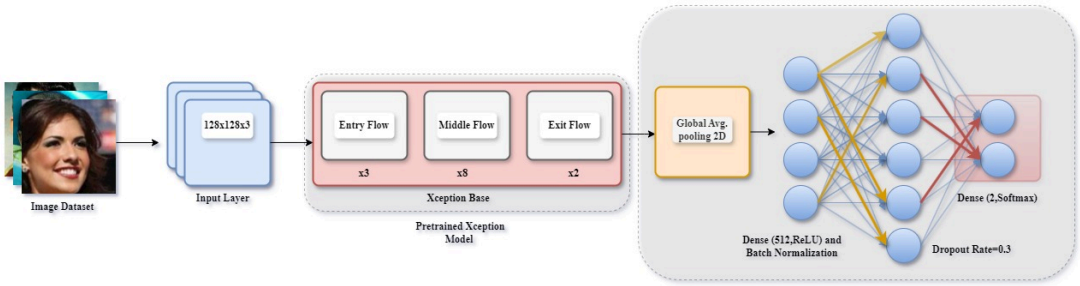

Figure 3.2: Xception model architecture for face forgery detection

3.1.3.2 ResNet50

ResNet50's residual connections enable deep feature learning without needing gradients, making it ideal for face forgery detection. It uses conv, identity, and bottleneck blocks to improve computation and bottleneck blocks reducing dimensions for better performance and feature extraction.

3.1.3.3 EfficientNetB0

EfficientNetB0 model uses the compound scaling method to balance depth, width, and resolution, and this makes it suitable and versatile for face fraud detection. With MBConv layers and squeeze-and-excite optimization, it offers strong performance with fewer parameters and lower computational needs.

3.1.3.4 DenseNet121

DenseNet121's closely linked layers enhance gradient flow and feature reuse, making it effective for detecting face forgeries. Each layer receives input from all previous layers, improving information flow and reducing the vanishing gradients problem.

3.1.3.5 MobileNet

MobileNet designed for face forgery detection, uses depth-wise separable convolutions and a linear bottleneck to reduce parameters and computational workload. Its simplified architecture suits low-power devices like mobile phones, offering high accuracy and efficient operation with limited resources.

3.2 Model Evaluation

Following training, the models are evaluated using key performance indicators such as the confusion matrix and ROC/AUC curve. A classification report also includes additional parameters like precision, recall, F1-score, and overall accuracy. These parameters provide in-depth performance of models.

3.3 Experimental Analysis

The implementations were carried out on Kaggle and Google Colab platforms, utilizing T4 and P100 GPUs to ensure efficient processing and accelerated training of the deep learning models.

3.3.1 Dataset

In Table 4.1 dataset for face forgery detection includes both real and fake images collected from several databases. Real face images are gathered with ”140k Real and Fake Face,” ”Fake vs. Real Faces,” ”Real and Fake Face Detection,” and ”Flickr Face

(FFHQ).” Fake face images contain data from ”140k real and fake faces,” ”fake vs. real faces” (which contains images generated by Style-GAN and ThisPersonDoesNotExist), ”real and fake face detection,” and ”1 Million Fake Faces ” (a large-scale dataset including a million synthetic face images created using multiple generative adversarial networks).

Table 3.1: Custom dataset

Dataset Category	**Dataset Source**	
Real Images	140k Real and fake face	30008
	Fake vs real face	581
	Real and Fake face detection	1081
	Flickr Face (FFHQ)	70000
Fake Images	140k Real and fake face	29996
	Fake vs Real faces	700
	Real and Fake face detection	960
	1Million Fake Face	70000
Total Images		**202326**

3.3.2 Hyperparameter

DeiT-Tiny, LeViT-128S, MobileViT-XXS, and TinyViT use hyperparameters to boost accuracy. Gains are notable. A learning rate of 0.0005 across 40 epochs with a ReduceLROnPlateau scheduler (patience 5, factor 0.5) keeps training stable and recovers from slowdowns. Batch sizes are 16 for DeiT and MobileViT and 32 for LeViT and TinyViT to balance memory and gradients. Adam serves as optimizer, with AdamW for deeper models handling weight decay. Mixed-precision training accelerates computation without losing accuracy. Inputs are resized to 128 or 224, normalized (mean 0.5, std 0.5), and randomly flipped, raising accuracy from 92.7 to 99.88 %, as shown in Table 3.2.

Table 3.2: Hyperparameter settings for different backbone models

Parameter	**DenseNet121**	**MobileNet**	**ResNet50**	**EfficientNet**	**Xception**
Image size	128×128	128×128	128×128	128×128	128×128
Batch size	16	16	16	16	16
Weights	ImageNet	ImageNet	—	—	ImageNet
Activation	Sigmoid	Sigmoid	SoftMax	SoftMax	SoftMax

Loss	Binary Cross-Entropy	Binary CrossEntropy	Categorical CrossEntropy	Categorical CrossEntropy	Categorical CrossEntropy
Learning rate	0.001	0.001	0.001	0.001	0.001
Optimizer	Adam	Adam	Adam	Adam	Adam

3.3.3 Experiment on Custom Dataset

Tables 3.3 and 3.4 present a comprehensive analysis of various models developed for face forgery detection using the custom dataset. These evaluations focus on key performance metrics such as precision, recall, F1-score, True Positive Rate (TPR), True Negative Rate (TNR), and overall accuracy. Among all models, the Xception architecture consistently outperformed the others, achieving a precision of 98.93%, a recall of 99.34%, and an F1-score of 99.14%, effectively distinguishing between authentic and forged images. The MobileNet model followed closely, delivering impressive results with a precision of 99.03%, a recall of 99.16%, and an overall accuracy of 99.10%, slightly lower than Xception. DenseNet121 and EfficientNetB0 also performed well, attaining F1-scores of 98.69% and 98.09%, respectively, though they showed some variability in accurately separating real and fake images. In contrast, ResNet50 underperformed relative to the other models, with an accuracy of 97.62% and a recall of 97.86%, indicating a relatively higher tendency to miss actual forgery cases.

Table 3.3: Performance Metrics for Real and Fake Data across Different Models

Models	Precision (Real) %	Recall (Real)%	F1-(Real)%	Precision (Fake)%	Recall (Fake)%	F1-(Fake)%
DenseNet121	98.64	98.74	98.69	98.73	98.64	98.69
MobileNet	99.03	99.16	99.10	99.16	99.03	99.10
ResNet50	97.62	97.86	97.74	97.85	97.62	97.73
EfficientNet	98.11	98.08	98.09	98.08	98.11	98.09
Xception	98.93	99.34	99.14	99.34	98.93	99.13

Table 3.4: Performance Metrics Summary

Models	TPR (%)	TNR (%)	FPR (%)	FNR (%)	Accuracy (%)
DenseNet121	98.74	98.64	1.36	1.26	98.69
MobileNet	99.16	99.03	0.97	0.84	99.10
ResNet50	97.86	97.62	2.38	2.14	97.74
EfficientNetB0	98.08	98.11	1.89	1.92	98.09
Xception	99.34	98.93	1.07	0.66	99.14

Figure 3.3 illustrates DenseNet121's training process, where training accuracy stabilizes near 1.00 and validation accuracy at 0.98. A spike in validation loss and accuracy around the 10th epoch indicated overfitting, but it resolved later.

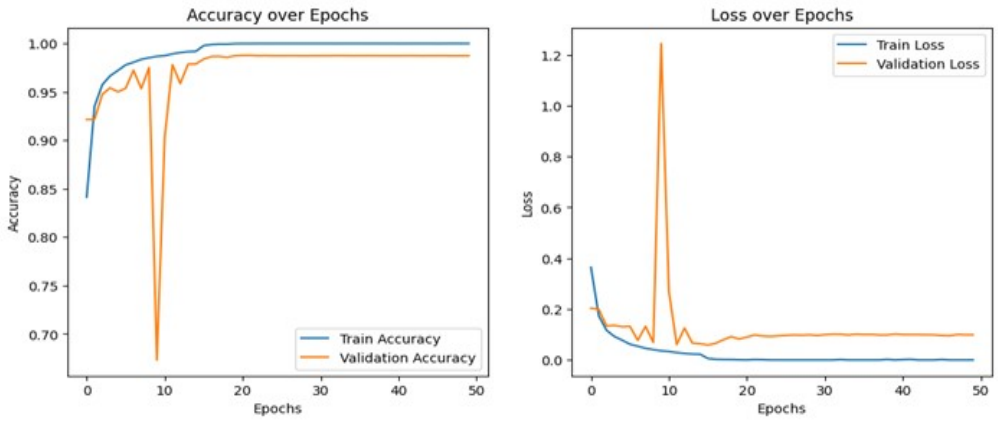

Figure 3.3: DenseNet121 Training and Test accuracy and Loss graph

In Figure 3.4 and 3.5, the ResNet50 and MobileNet training plots reveal early overfitting spikes before the 10th epoch, but both models stabilize later. ResNet50's validation accuracy settles around 0.98, while MobileNet achieves 1.00 training accuracy and 99% validation accuracy, reflecting efficient learning.

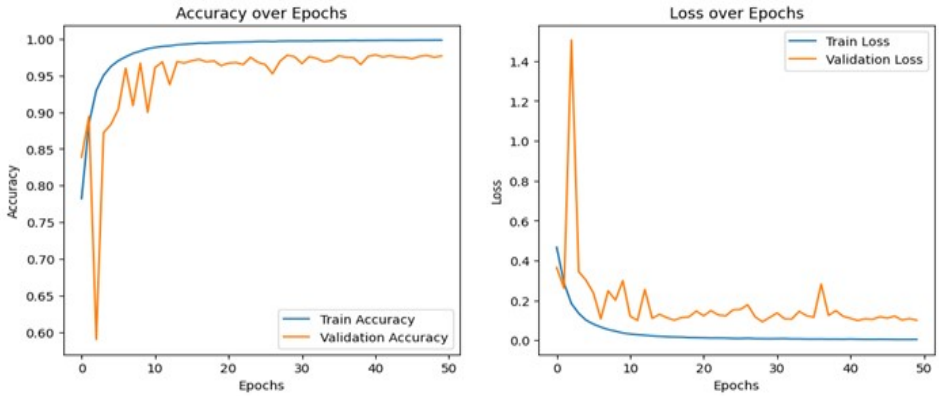

Figure 3.4: ResNet50 Training and Test accuracy and Loss graph

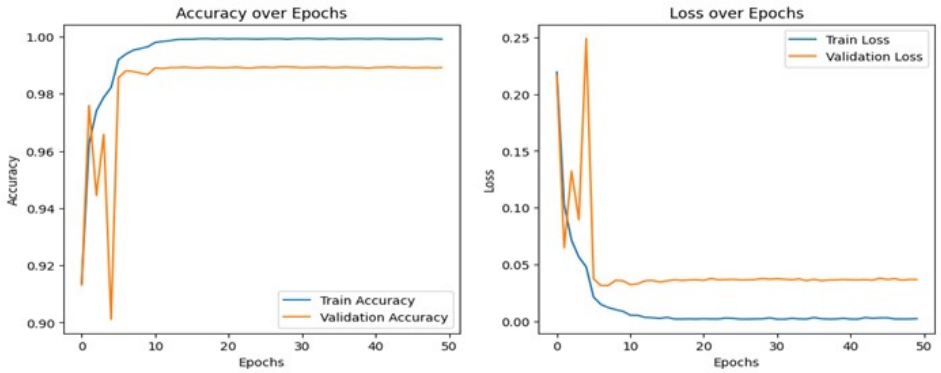

Figure 3.5: MobileNet Training and Test accuracy and Loss graph

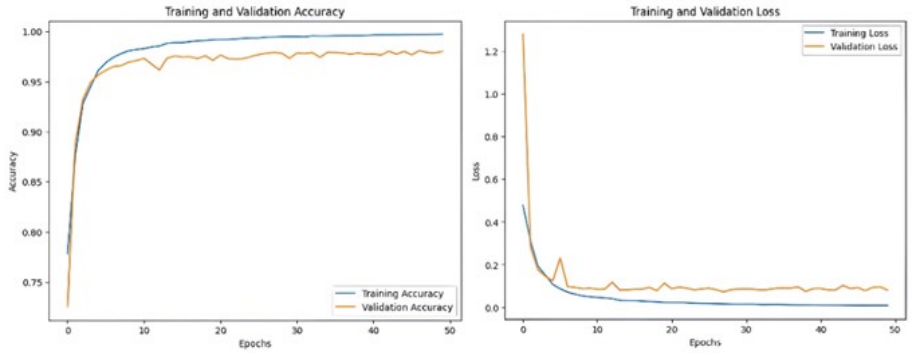

Figure 3.6: EfficientNetB0 Training and Test accuracy and Loss graph

Figure 3.6 and 3.7 show the EfficientNetB0 and Xception training processes graph. Xception shows rapid accuracy improvement, with validation accuracy around 0.99 and training accuracy reaching 1.00. A small spike appears but is less than in ResNet50, DenseNet121, and MobileNet, indicating stable learning. Efficient-NetB0 exhibits a smooth training curve with no significant spikes, reflecting steady learning.

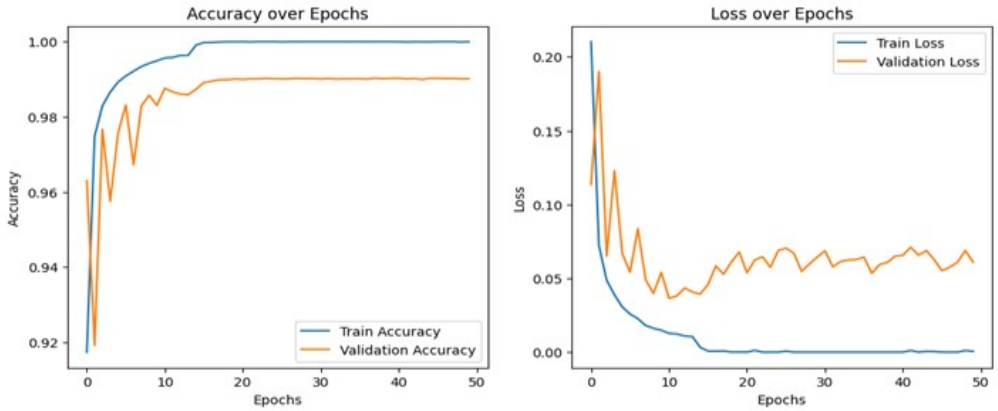

Figure 3.7: Xception Training and Test accuracy and Loss graph

3.3.4 Confusion matrix

The confusion matrices for Xception, ResNet50, EfficientNetB0, DenseNet121 and MobileNet show true positives, true negatives, false positives, and false negatives. Xception model outperforms the other pretrained models with the fewest errors 97

false negatives and 159 false positives.it shows higher accuracy and precision as shown in Figure 3.8.

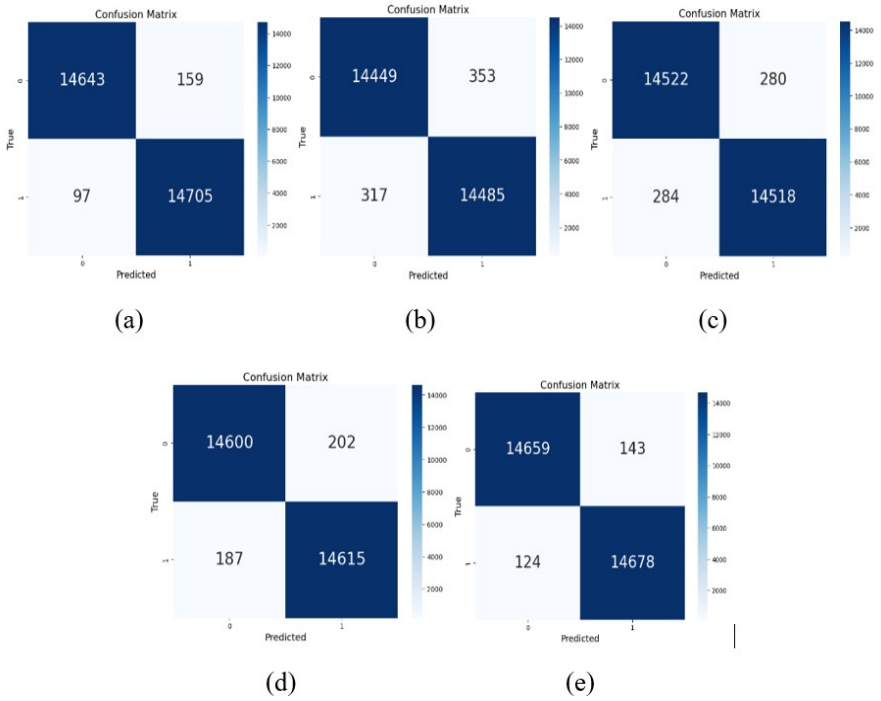

Figure 3.8: Confusion matrices for (a) Xception, (b) ResNet50, (c) EfficientNetB0, (d) DenseNet121, and (e) MobileNet

Table 3.5: Face forgery detection using Xception and comparison with existing models on a similar dataset

References	Objective	Algorithm	Accuracy (%)	Limitations
Manoranjitham et al. (2024)	Detecting fake images generated by GANs	DenseNet-121	98.00	Manual weight assignment and adaptability to new GANs
Raveena et al. (2024)	Compare CNN models and hyperparameters for optimal deepfake detection	ResNet50	88.00	Limited dataset diversity
Jannu et al. (2024)	Comparative analysis of deepfake detection models	MobileNet	75.00	Model bias and limited dataset diversity

Raza et al. (2024)	Evaluate DenseNet121 and InceptionResNetV2 for deepfake detection	MMGAN-GUARD	97.00	Limited dataset diversity and focus on specific architectures
Wang et al. (2021)	Develop and evaluate LBP-Net for robust detection	LBP-Net	98.58	Limited dataset diversity and using a single model
Neha et al. (2023)	Analyze distortions' impact on deepfake detection	DenseNet	94.23	Limited to specific distortions and model
Proposed Model	Comparative analysis of face forgery detection models	**Xception**	**99.14**	Limited data diversity and lack of advanced detection techniques

CHAPTER 4

FACE FORGERY DETECTION USING VARIANTS OF VISION TRANSFORMERS AND HYBRID-CLASSICAL QUANTUM MODELS

4.1 Proposed Architecture

4.1.1 Dataset and Pre-processing

A balanced collection of real and fake face images is split into training, validation, and test sets. Every image is resized to 224×224 pixels, randomly flipped and cropped to introduce variation, then channel-wise normalized to match Vision Transformer inputs. These steps ensure each model sees data that's both consistent and diverse.

4.1.2 Fully Classical ViT-Based Architecture

Four Vision Transformer backbones—ViT-Base, DeiT, MobileViT, and TinyViT—are repurposed by replacing their classification heads with a simple two-node layer that outputs "real" or "fake." Experiments alternate between fine-tuning the entire network and training only the new head, while the rest remains frozen. Training uses the AdamW optimizer with cross-entropy loss, and ReduceLROnPlateau is applied to drop the learning rate when progress stalls. Mixed-precision training helps accelerate computations. Loss and accuracy logs on both training and validation sets guide the optimization process. Final testing reports overall accuracy and confusion matrices, as shown in Figure 4.1.

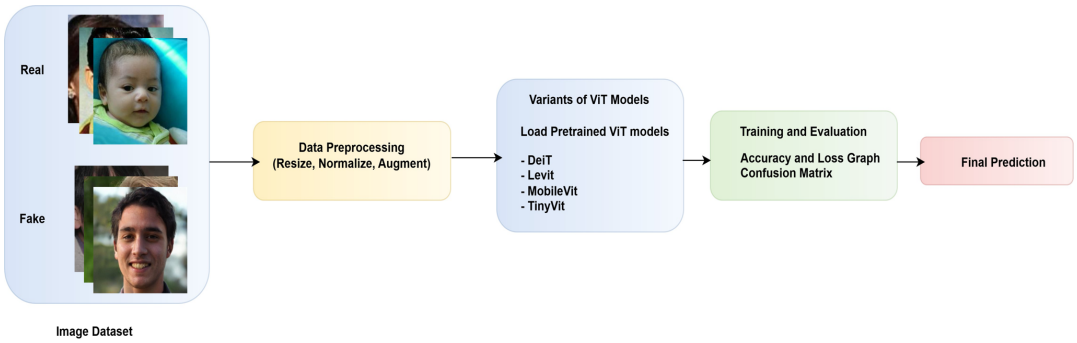

Figure 4.1: Workflow for face forgery detection using Vision Transformers.

4.1.3 ViT base model and variants

The Vision Transformer (ViT) as shown in Figure 3.4, processes images by first dividing them into small patches, like cutting a photo into tiles. Each patch is turned into a numeric vector and given positional info so the model knows their order. These vectors pass through a transformer, which uses self-attention to understand patterns and relationships across the image. Inside the transformer, layers normalize inputs, apply attention, and use a small neural network to refine the output. Finally, a classifier predicts if the image is real or fake. This design mimics how transformers read text, but it works on image pieces instead.

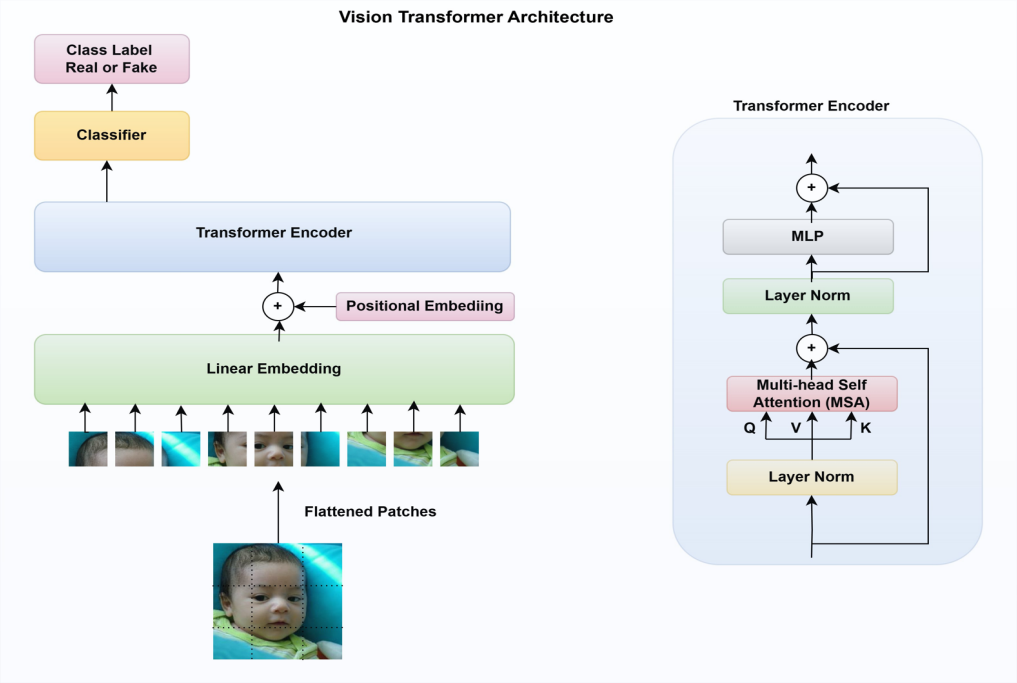

Figure 4.2: ViT base model architecture

4.1.3.1 DeiT

DeiT (Data-efficient Image Transformer) shrinks the hunger for massive datasets by borrowing knowledge from a teacher network. It uses a clever distillation token during training, guiding the transformer to learn rich visual features with far fewer images—making high accuracy more accessible without endless GPU hours.

4.1.3.2 TinyViT

TinyViT trims the fat off a standard Vision Transformer to fit on resource-tight hardware. By crafting compact attention modules and slimming down channels, it delivers surprisingly strong performance with only a few million parameters—ideal for edge devices and real-time applications where every millisecond and megabyte counts.

4.1.3.3 MobileViT

MobileViT blends the best of convolutions and self-attention to adapt transformers for smartphones. It weaves lightweight transformer blocks into a convolutional backbone, capturing both local details and global relationships, yet stays lean enough to run smoothly on mobile CPUs and GPUs—perfect for on-device image tasks.

4.1.3.4 LeViT

LeViT rearranges the transformer playbook by interleaving convolutional stages with attention layers in a pyramidal design. This hybrid structure drastically cuts down inference time and memory use, while still learning expressive representations—an excellent match for scenarios demanding ultra-fast inference on modest hardware.

4.1.3.5 Swin

The Swin Transformer is an efficient, hierarchical vision transformer that computes self attention within shifted local windows, enabling linear computational complexity and cross-window connections. It produces multi-scale feature maps for vision tasks, achieving robust, state-of-the-art performance across classification, detection, and segmentation.

4.1.4 Hybrid Quantum–Classical Model

In this hybrid setup, a frozen Vision Transformer (Swin-Tiny or MobileViT-XXS) serves as a feature extractor. Its output vector is first reduced to an intermediate size (256 or 512), then further down to match two qubits. A tanh activation maps these values into the [–1,1] range, which are then scaled into rotation angles for the quantum layer. The quantum circuit, implemented using Pennylane, applies angle-encoding gates on each qubit and entangles them through a ring of CNOTs. This "rotate → entangle" block is repeated for a fixed quantum depth, as shown in Figure 3.5. During training, only the quantum circuit parameters and the final dense layer are updated. Pauli-Z measurements on each qubit produce classical outputs, which are passed into the dense layer to classify the input as real or fake.

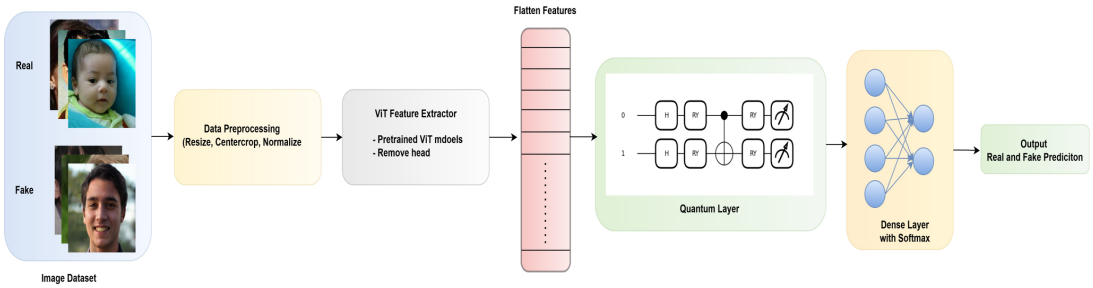

Figure 4.3: Hybrid quantum-classical architecture with Vision Transformer backbone and quantum enhanced classifier.

4.2 Model Evaluation

Both the fully-classical and hybrid quantum–classical ViT models are evaluated by plotting training/validation loss and accuracy curves and computing test-set accuracy. Confusion matrices reveal true versus predicted outcomes, while precision and recall per class quantify correctness and completeness. The quantum variant additionally leverages gates like Hadamard, RY rotations, and CNOT entanglers within its circuit to enhance decision boundaries.

4.3 Experimental Analysis

Both fully-classical and hybrid quantum–classical ViT models were implemented on Kaggle (P100 GPUs and Google Colab), with PennyLane powering the quantum layers. Evaluation includes training/validation loss and accuracy plots, test-set accuracy, and confusion matrices. Precision and recall per class quantify performance, while the quantum model employs Hadamard, RY rotations, and CNOT gates to refine decision boundaries.

4.3.1 Dataset

The 140k Real and Fake Faces dataset contains 70,000 authentic face images from Flickr (Nvidia) and 70,000 fake faces generated by StyleGAN. It is widely used for training and evaluating Face forgery detection models, supporting robust real-vs-fake face classification.

4.3.2 Hyperparameter

DeiT-Tiny, LeViT-128S, MobileViT-XXS, and TinyViT use hyperparameters to boost accuracy. Gains are notable. A learning rate of 0.0005 across 40 epochs with a ReduceLROnPlateau scheduler (patience 5, factor 0.5) keeps training stable and recovers from slowdowns. Batch sizes are 16 for DeiT and MobileViT and 32 for LeViT and TinyViT to balance memory and gradients. Adam serves as optimizer, with AdamW for deeper models handling weight decay. Mixed-precision training accelerates computation without losing accuracy. Inputs are resized to 128 or 224, normalized (mean 0.5, std 0.5), and randomly flipped, raising accuracy from 92.7% to 99.88%, as shown in Table 4.1.

Table 4.1: Hyperparameter configurations for the Vision Transformer variants

Parameter	DeiT-Tiny	LeViT-128S	MobileViT-XXS	TinyViT
Batch Size	16	32	16	32
Image Size	128	224	224	224
Num Epochs	40	40	40	40

Learning Rate	0.0005	0.0005	0.0005	0.0005
Optimizer	Adam	AdamW	Adam	AdamW
LR Patience	5	5	5	5
LR Factor	0.5	0.5	0.5	0.5

Quantum circuits use 2 qubits (n _qubits = 2) and a single layer (q depth = 1) to limit noise while enabling CNOT-based entanglement. Gate rotations start at q delta = 0.01 to avoid unstable updates. A learning rate of 0.0004 with batch size 32 ensures steady gradients over 30–40 epochs, balancing overfitting and underfitting. Adaptive schedulers (such as ReduceLROnPlateau or StepLR) adjust the learning rate when validation performance stalls. Pretrained backbones like MobileViT-XXS and Swin-Tiny remain frozen; only the quantum head is trained using the Adam or AdamW optimizer. Mixed precision accelerates training while maintaining numerical stability. Test accuracies range from 90.8% to 97.6%, as shown in Table 4.2.

Table 4.2: Hyperparameter configurations for the hybrid quantum-classical models

Parameter	SWIN-Tiny (RY+RX+RZ)	SWIN-Tiny (RY)	MobileViT-XXS (RY)
Model Backbone	SWIN-Tiny	SWIN-Tiny	MobileViT-XXS
Feature Extractor Size	768	768	384
Intermediate Size	256	256	512
Quantum Layer Structure	Embedding: RZ, RY, RX; Entangling: CNOT (ring); repeated by *q depth*	Embedding: RZ, RY, RX; Entangling: CNOT (ring); repeated by *q depth*	Embedding: RZ, RY, RX; Entangling: CNOT (ring); repeated by *q depth*
Quantum Gates Used	RZ, RY, RX, CNOT	H, RY, CNOT	RY, CNOT
Number of Qubits	2	2	2
Quantum Depth	1	1	1
Qubit Angle Shift	0.01	0.01	0.01
Learning Rate (LR)	0.0004–0.0005	0.0004–0.0005	0.0004

Gate Combination	RY + RX + RZ	RY	RY

4.3.3 Quantum Circuit Structure

Figure 4–6 illustrate hybrid quantum circuits for image classification. Swin-Tiny with RY (Fig. 4) uses Hadamard and RY gates with linear entanglement. MobileViT-XXS with RY (Fig. 5) adds RX and RZ for richer embeddings. Swin-Tiny with RY+RX+RZ (Fig. 6) stacks rotations and double CNOT entanglement for deeper quantum feature encoding.

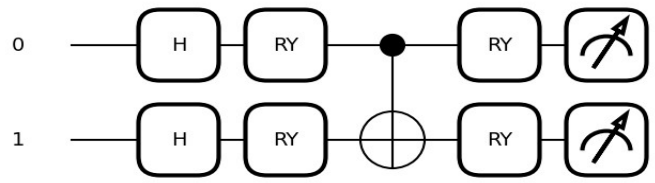

Figure 4.4: Swin-Tiny RY circuit using Hadamard, RY gates, and CNOT-based entanglement

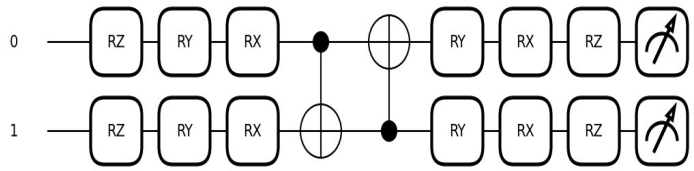

Figure 4.5: Swin-Tiny RY+RX+RZ circuit with full rotation gates and layered entanglement

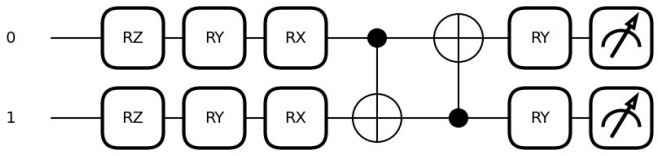

Figure 4.6: MobileViT-XXS RY circuit with RZ, RY, RX embeddings and ring-style CNOTs.

4.3.4 Experiment on Dataset

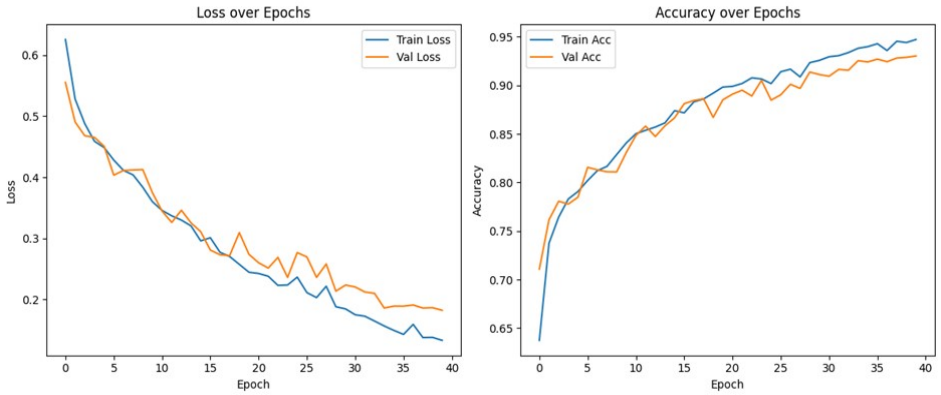

Figure 4.7: Accuracy over epochs for DeiT-Tiny

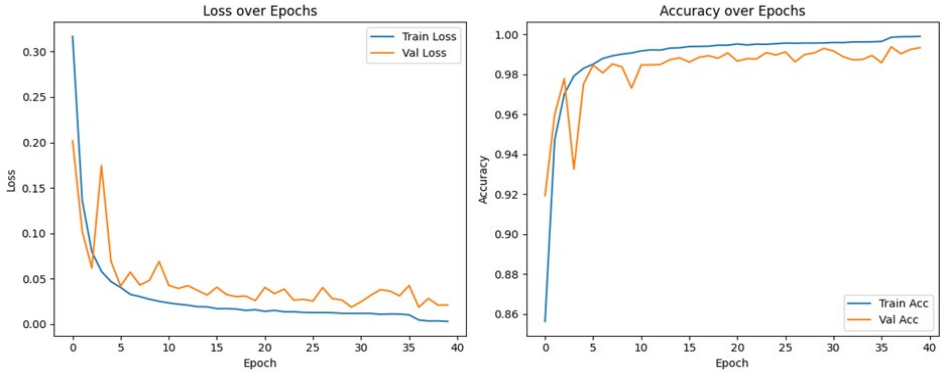

Figure 4.8: Accuracy over epochs for LeViT

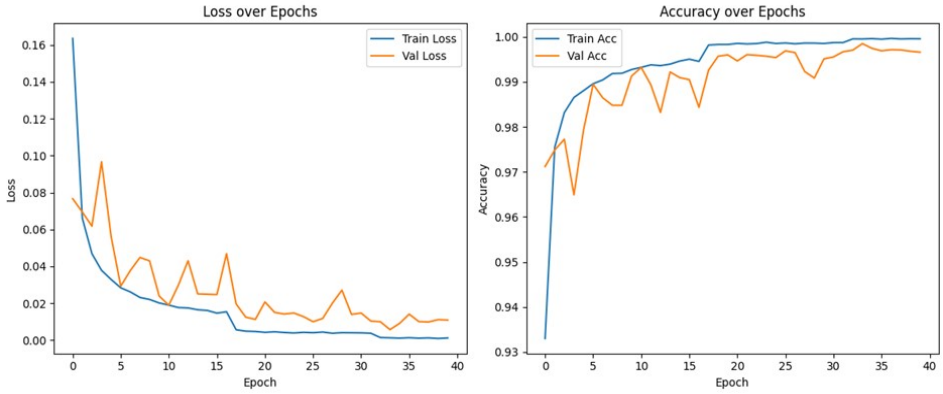

Figure 4.9: Accuracy over epochs for Tiny-ViT

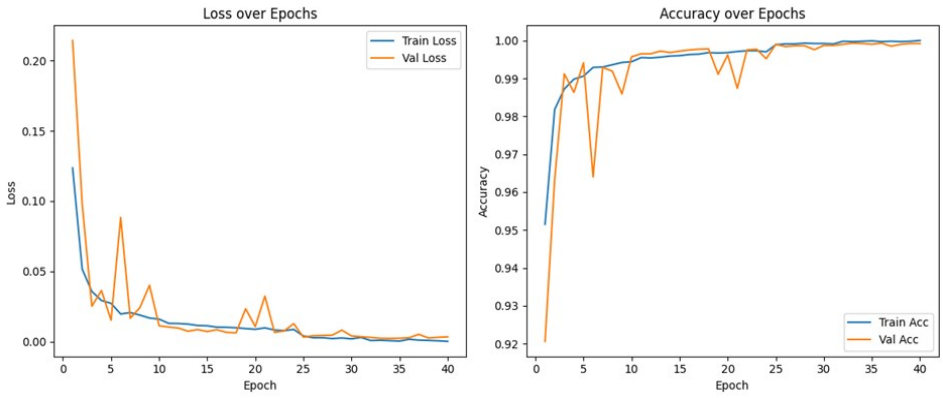

Figure 4.10: Accuracy over epochs for MobileViT

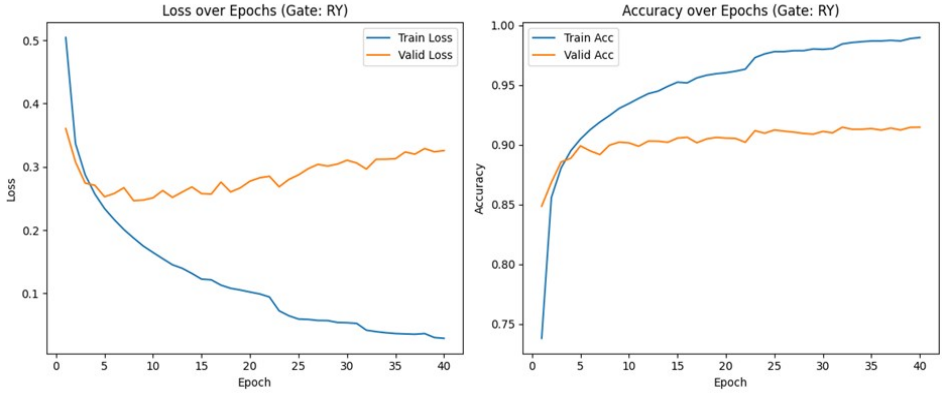

Figure 4.11: Accuracy over epochs for MobileViT with RY

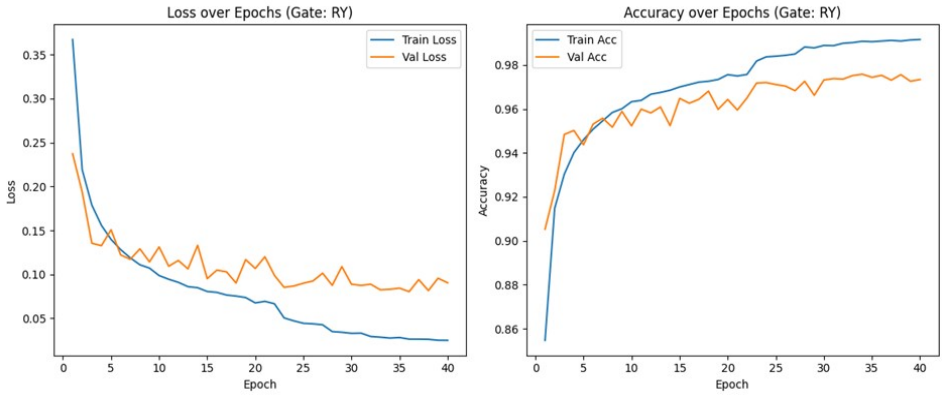

Figure 4.12: Accuracy over epochs for Swin with RY

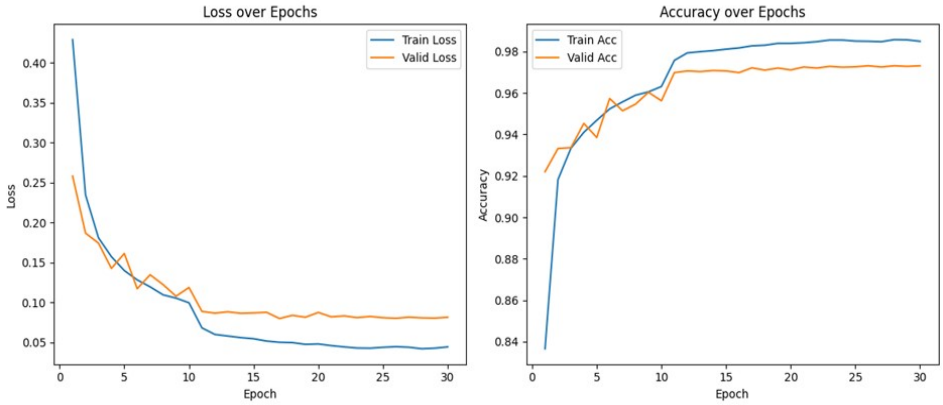

Figure 4.13: Accuracy over epochs for Swin with RY+RX+RZ

4.3.5 Confusion Matrix

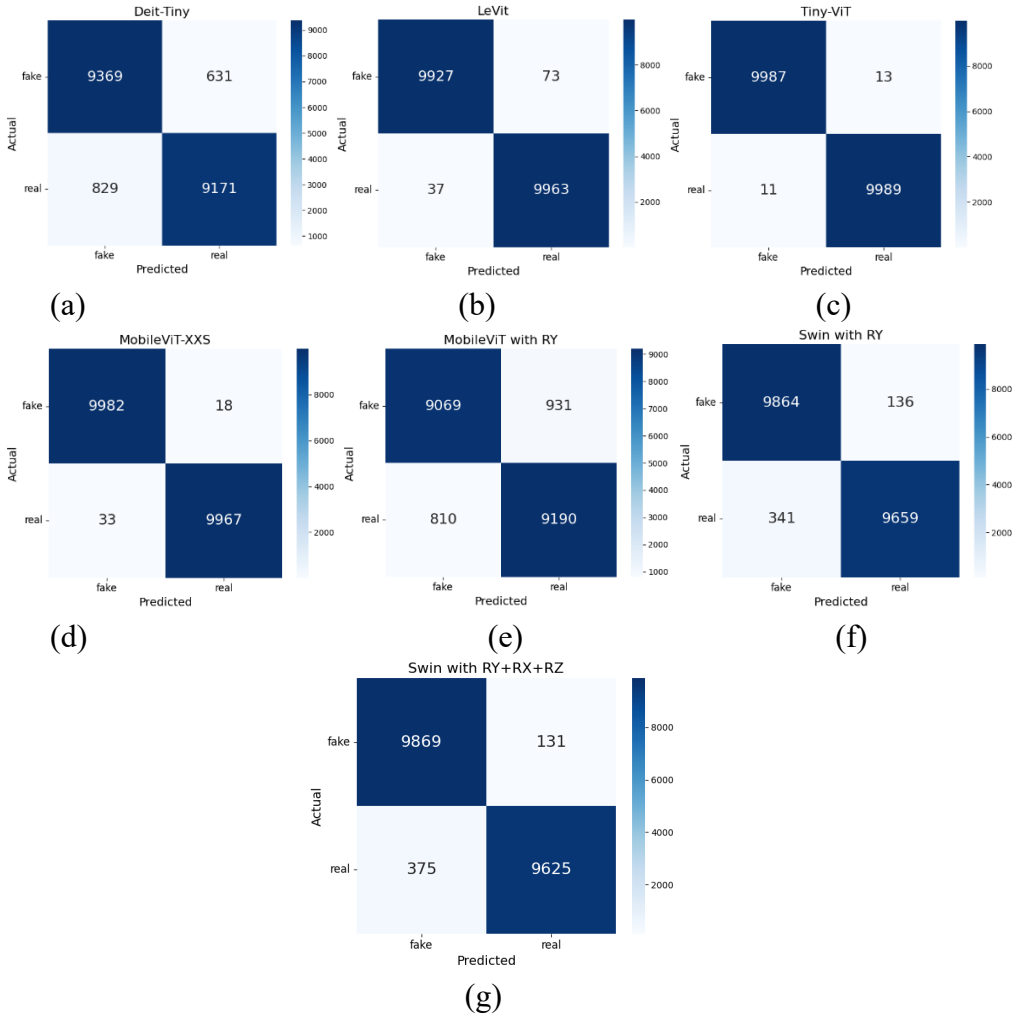

Figure 4.14: Confusion matrices for face forgery detection using different transformer-based and hybrid models.

(a) DeiT-Tiny, (b) LeViT, (c) Tiny-ViT, (d) MobileViT-XXS, (e) MobileViT with RY, (f) Swin with RY, and (g) Swin with RY + RX + RZ.

Table 4.3: Performance metrics of ViT variants and quantum-enhanced models for face forgery detection.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
DeiT	0.9270	0.9356	0.9171	0.9263
LeViT	0.9945	0.9927	0.9963	0.9945
MobileViT-XXS	0.9988	0.9987	0.9989	0.9988
TinyViT	0.9974	0.9982	0.9967	0.9974
MobileViT with RY	0.9130	0.9130	0.9130	0.9129
Swin with RY	0.9761	0.9764	0.9762	0.9762
Swin with RY+RX+RZ	0.9747	0.9750	0.9747	0.9747

Table 4.4: Comparative analysis of state-of-the-art and proposed models for face forgery detection.

Author Name	**Objective of Paper**	**Algorithm Used**	**Accuracy**	**Limitations**
Rao et al.	Detect deepfakes using a CNN for enhanced media security	TruceNet (CNNbased image classifier)	93.01%	May produce false positives and false negatives.
Naeem et al.	Analyze trends in real, deepfake, and synthetic facial images	Eight DL models; ViT Patch16 performed best	98.25%	Limited dataset diversity reduces generalizability.
Usmani et al.	Develop a lightweight model for deepfake detection	Shallow Vision Transformer	88.52%	Limited dataset size hinders generalization.
Shobhit et al.	Propose a lightweight CNN for forged image detection	MiniNet (fully convolutional neural network)	95%	Minimal architecture may limit performance on new datasets.

Sherin et al.	Propose explainable deepfake detection with bias considerations	Inception-based network + explainability framework	99.87%	Potential bias in diverse scenarios and limited scope.
Kerenalli et al.	Detect DL generated fake faces by blending CNN and ViT	EfficientNet + Shifted Window Transformer (Swin)	98.04%	High model complexity may hinder real-time deployment.
Proposed Model (ViT variant)	Detect real vs fake faces using lightweight ViT architecture	MobileViT-XXS	**99.88%**	Limited dataset diversity may affect generalization to unseen domains.
Proposed Model (Quantum)	Quantum enhanced deepfake detection using Swin and hybrid quantum layer	Swin with RY (Quantum enhanced Swin Transformer)	**97.61%**	Quantum layers increase computation time and are sensitive to noise and dataset diversity.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

A diverse set of modern and emerging architectures underwent thorough testing for detecting face forgeries, covering classic convolutional neural networks (CNNs), compact Vision Transformers (ViTs), and hybrid quantum-classical models. A balanced collection of roughly 200,000 images—half genuine faces from the Flickr Face Dataset and half synthetic forgeries crafted with StyleGAN—provided the evaluation framework. Results highlight how transfer learning and architectural innovation can raise detection accuracy to new heights. Within the realm of CNNs, Xception claimed the highest score at 99.14% accuracy on the test set. Other high-performing networks—ResNet50, EfficientNetB0, DenseNet121, and MobileNetV2—each surpassed 97% accuracy. These outcomes underline the advantage of leveraging pre-trained weights tuned on massive datasets, rather than training from the ground up. Such an approach cuts down both development time and resource demands while attaining near-perfect precision.

Attention then shifted to lightweight transformer designs. MobileViT-XXS and TinyViT achieved over 99.8% accuracy, proving that even small transformer blocks can rival larger models. Their ability to detect long-range dependencies and subtle manipulation artifacts makes them particularly well suited for real-time applications on mobile devices or embedded hardware. Lower memory footprints and reduced inference times further bolster their appeal for on-device deployment. A hybrid quantum-classical experiment integrated minimal quantum circuits into a Swin Transformer backbone. The circuits consisted of two qubits, each undergoing RY, RX, and RZ rotations, followed by a ring of CNOT entangling gates. Despite limitations in current quantum hardware, this quantum-enhanced Swin model reached 97.61% accuracy. Findings suggest that quantum superposition and entanglement may help expose non-linear patterns that sometimes escape classical detection methods. Although the quantum lift was modest, it opens the door to cloud-based quantum resources enhancing future forensic tools.

Many directions offer to move this field forward. Increasing dataset diversity using real-world alterations, video sequences, and audio-visual deepfakes enhances model robustness in diverse settings. Expanding quantum components, including more qubits, deeper circuits, and better hybrid systems, can increase quantum-driven advantages. Adding confidence-scoring systems improves integration into social

media moderation, biometric security, and border-control checkpoints, automatically identifying suspect data for human evaluation. Pre-trained CNNs offer rapid development and high accuracy, making them ideal when computational resources are plentiful. Compact ViTs strike a balanced compromise between speed, resource usage, and performance, fitting edge-device needs perfectly. Hybrid quantum-classical models, which are still evolving, aim to create next-generation detectors that integrate with conventional techniques increase robustness and quantum strengths. In the race against digital deception, careful selection of modelsand techniques, continued dataset expansionand diversity, and deeper quantum integration can result in face forgery detection systems with remarkable accuracy, speed, and durability.

REFREENCES

[1] Nirkin, Y., Keller, Y., Hassner, T. (2019). FSGAN: Subject agnostic face swapping and reenactment. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7184–7193. IEEE.

[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672–2680. https://doi.org/10.3156/JSOFT.29.5_177_2

[3] Wang, T., Lu, X. (2023). Face forgery detection algorithm based on improved MobileViT network. Presented at the IEEE International Conference on Signal Processing. https://doi.org/10.1109/icsp58490.2023.10248802

[4] Afchar, D., Nozick, V., Yamagishi, J., Echizen, I. (2018). MesoNet: A compact facial video forgery detection network. Presented at the European Signal Processing Conference.

[5] Raza, S.A., Habib, U., Usman, M., Cheema, A.A., Khan, M.S. (2024). MMGANGuard: A robust approach for detecting fake images generated by GANs using multi-model techniques. *IEEE Access*. https://doi.org/10.1109/access.2024.3393842

[6] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, pp. 770–778.

[7] Tan, M., Le, Q.V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 6105–6114. PMLR.

[8] Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269.

[9] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807.

[10] NVlabs. (2019). NVlabs/ffhq-dataset: Flickr-Faces-HQ Dataset (FFHQ). Internet Archive. https://archive.org/details/ffhq-dataset

[11] Jannu, O., Sekar, V., Padhy, T., Padalkar, P. (2024). Comparative analysis of deepfake detection models. *IEEE 9th International Conference for Convergence in Technology (I2CT)*, pp. 1–8. https://doi.org/10.1109/i2ct61223.2024.10543823

[12] Zhao, H., Wei, T., Zhou, W., Zhang, W., Chen, D., Yu, N. (2021). Multi-attentional deepfake detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2185–2194.

[13] Patel, Y., Tanwar, S., Bhattacharya, P., Gupta, R., Alsuwian, T., Davidson, I.E., Mazibuko, T.F. (2023). An improved dense CNN architecture for deepfake image detection. *IEEE Access*, vol. 11, pp. 22081–22095.

[14] Wang, Y., Zarghami, V., Cui, S. (2021). Fake face detection using local binary pattern and ensemble modeling. *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 3917–3921. https://doi.org/10.1109/icip42928.2021.9506460

[15] Zhu, X., Wang, H., Fei, H., Lei, Z., Li, S.Z. (2021). Face forgery detection by 3D decomposition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2929–2939.

[16] Manoranjitham, R., Swaroop, S.S. (2024). A comparative study of DenseNet121 and InceptionResNetV2 model for deepfake image detection. *3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pp. 432–438. IEEE.

[17] Raveena, Chhikara, R., Punyani, P. (2024). Exploring deepfake detection: A comparative study of CNN models. *International Conference on Intelligent Systems for Cybersecurity (ISCS)*, pp. 1–6. IEEE. https://doi.org/10.1109/iscs61804.2024.10581012

[18] Neha, Arora, B. (2023). Deep learning-based model for deepfake image detection: An analytical approach. *3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pp. 1019–1027. IEEE. https://doi.org/10.1109/icimia60377.2023.10426561

[19] Xhlulu. (2020). 140k real and fake faces. Retrieved from https://www.kaggle.com/datasets/xhlulu/140k-real-and-fakefaces

[20] Reben, A. (2019). 1 million fake faces. Internet Archive. https://archive.org/details/1mFakeFaces

[21] Dosovitskiy, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *Proceedings of the International Conference on Learning Representations (ICLR)*.

[22] Howard, A.G., et al. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint*, arXiv:1704.04861.

[23] Rao, D., Patil, A., Sarda, E. (2023). TruceNet: A CNN-based model for accurate classification of deepfake images. *IEEE International Conference on Data Science and Network Security (ICDSNS)*, pp. 1–6. https://doi.org/10.1109/ICDSNS58469.2023.10245314

[24] Naeem, S., et al. (2024). Real, fake and synthetic faces - does the coin have three sides? *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–8. IEEE, Istanbul. https://doi.org/10.1109/FG60137.2024.00000

[25] Usmani, S., Kumar, S., Sadhya, D. (2023). Efficient deepfake detection using shallow vision transformer. *Multimedia Tools and Applications*, vol. 83, pp. 12339–12362. https://doi.org/10.1007/s11042-023-15910-z

[26] Duan, J., et al. (2025). Test-time forgery detection with spatial-frequency prompt learning. *International Journal of Computer Vision*, vol. 133, pp. 672–687. https://doi.org/10.1007/s11263-024-02208-2

[27] Tyagi, S., Yadav, D. (2023). MiniNet: A concise CNN for image forgery detection. *Evolving Systems*, vol. 14, pp. 545–556. https://doi.org/10.1007/s12530-022-09446-0

[28] Mathews, S., et al. (2023). An explainable deepfake detection framework on a novel unconstrained dataset. *Complex & Intelligent Systems*, vol. 9, pp. 4425–4437. https://doi.org/10.1007/s40747-022-00956-7

[29] Bobulski, J., Kubanek, M. (2024). Detection of fake facial images and changes in real facial images. In: Nguyen, N.T. et al. (Eds.), *Computational Collective Intelligence. ICCCI 2024*, Lecture Notes in Computer Science, vol. 14811, pp. 110–122. Springer, Cham. https://doi.org/10.1007/978-3-031-70819-0_9

[30] Mari, A., et al. (2020). Transfer learning in hybrid classical-quantum neural networks. *Quantum*, vol. 4, p. 340. https://doi.org/10.22331/q-2020-10-09-340

[31] Bergholm, V., et al. (2018). PennyLane: Automatic differentiation of hybrid quantum-classical computations. *arXiv preprint*, arXiv:1811.04968.

[32] Kerenalli, S., Yendapalli, V., Chinnaiah, M. (2023). Fake face image classification by blending the scalable convolution network and hierarchical vision transformer. *Proceedings of the Fourth International Conference on Computer and Communication Technologies*, Lecture Notes in Computer Science, vol. 13671, pp. 117–126. Springer, Singapore. https://doi.org/10.1007/978-981-19-8563-8_12

[33] Touvron, H., et al. (2021). Training data-efficient image transformers & distillation through attention. *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 10347–10357. PMLR.

[34] Graham, B., et al. (2021). LeViT: A vision transformer in ConvNet's clothing for faster inference. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12259–12269.

[35] Mehta, S., Rastegari, M. (2021). MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint*, arXiv:2110.02178.

[36] Wu, H., et al. (2023). TinyViT: Fast pretraining distillation for small vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12118–12129.

[37] Liu, Z., et al. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022.

[38] Schuld, M., Bocharov, A., Svore, K.M., Wiebe, N. (2020). Circuit-centric quantum classifiers. *Physical Review A*, vol. 101(3), p. 032308. https://doi.org/10.1103/PhysRevA.101.032308

[39] Benedetti, M., Lloyd, E., Sack, S., Fiorentini, M. (2019). Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, vol. 4(4), p. 043001. https://doi.org/10.1088/2058-9565/ab4eb5

LIST OF PUBLICATIONS

1. Manish Kumar, Bindu Verma, “Performance Evaluation of Face Forgery Detection Models: A Comparative Study”, presented at the 6th International Conference on Communication and Intelligent Systems (ICCIS 2024), organized by Maulana Azad National Institute of Technology (MANIT), Bhopal, India.

2. Manish Kumar, Bindu Verma, “Evaluating Vision Transformer Variants and Hybrid Quantum-Classical Models for Face Forgery Detection ”, presented at the 6th International Conference on Data Science and Applications (ICDSA 2025), organized by Malaviya National Institute of Technology (MNIT) Jaipur, Jaipur, India.

6th International Conference on Communication and Intelligent Systems (ICCIS 2024)

Organized by

Maulana Azad National Institute of Technology (MANIT), Bhopal, India

Technically Sponsored by

Soft Computing Research Society

November 08-09, 2024

Certificate of Presentation

This is to certify that **Manish Kumar** has presented the paper titled **Performance Evaluation of Face Forgery Detection Models: A Comparative Study** authored by **Manish Kumar, Bindu Verma** in the 6th International Conference on Communication and Intelligent Systems (ICCIS 2024) held during November 08-09, 2024.

SCRS\ICCIS2024\PC\321

Prof. Sanjay Sharma
(General Chair)

Dr. Harish Sharma
(General Chair)

https://scrs.in/conference/iccis2024

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis : **FACE FORGERY DETECTION USING CLASSICAL AND HYBRID QUANTUM DEEP LEARNING MODELS**

Total Pages : 28

Name of the Scholar : **Manish Kumar**

Supervisor (s) : **Dr. Bindu Verma**

Department : **Information Technology**

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: **Turnitin** Similarity Index: 9% , Total Word Count: 6022

Date: 28/05/2025

Candidate's Signature **Signature of Supervisor(s)**