scientific reports

Check for updates

OPEN

A quantum resilient deepfake detection framework using enhanced resnext and post quantum cryptography defence

Kollipara Naga Shreeya[1], Brindha Subburaj[1]✉, Kollipara Sai Govinda Saketh[1], T. V. Padmavathy[1], Sherly Alphonse[1] & Girish Subramanian[2]

Deepfakes are digital-media which may contain audio, videos or images synthesised by usage of Generative Adversarial Networks (GANs) with the aid of Artificial Intelligence (AI) technologies. These deepfakes have the capability to replicate and mimic the behaviour of real people. Although deepfakes are beneficial in filmmaking and in education sector, but they can have serious implications within the other areas or fields such as in politics, social network platforms, human security and law. The deepfakes can mislead several people with false information. By creating false evidence that can behave like real people and damage the reputation of one. There are several ways for deepfake detection. They work on the basis of analysing and discriminating between various features such as facial features, movements, blinking, variations in voice, variation in tones and background noises. However, these detection systems may be vulnerable to quantum adversaries and futuristic adversarial attacks. This may result in an ambiguity within trustworthy detection systems, which creates an overriding necessity for effective and trustworthy detection framework. The paper highlights a reliable deepfake image detection framework based on the ResNeXt architecture optimized with use of lattice-based adversarial training that is learning with errors (LWE) mechanism to make it resilient against several adversarial manipulations. In addition to this, when followed by the unification of Kyber and Dilithium with quantum cryptography methods, these ensure the authenticity and encryption of the detection results. The proposed scheme DeepQShield is quantum-resistant because it incorporates the executions of post cryptography algorithms and is trained and tested on the Deepfake Detection Challenge dataset (DFDC). On the DFDC database it achieved significantly higher accuracy of 99.28% and an impressive AUC value of 0.9997. When compared to the existing systems such as EfficientNet-B7 (accuracy: 97.2% on DFDC), Vision Transformers (ViT) (90 to 98% on Celeb-DF and DFDC), Multi-attentional CNN-LSTM networks (98.2% on DFDC), FuzzyDFD (accuracy: 99% FF++ and 93% on (Celeb-DF). DeepQShield outshines the conventional models in terms of security, scalability, accuracy and robustness making it best suitable for various applications in real-world scenarios like face forensics, social media data authentication.

Keywords Deepfake detection, ResNeXt, Lattice-based adversarial training, Post quantum cryptography

The rapid growth of generative models has given rise to the widespread surge of various deepfake generation methods. These generations can convincingly mimic the real individual, which can lead to many security and safety issues. At the same time, video-based deepfakes also draw widespread coverage. Along with increasing sophistication and the growth of the high-resolution based generative adversarial networks (GANs) for generating image-based deepfakes. The need for the presence of an effective detection system has grown significantly. Photo-focused deepfakes pose a major threat in the zone of digital forensics and can often lead to severe complications and implications. These include the tampering of the lawful evidence and spreading of falsehoods, which can result in unfair conclusions and judgments in the judiciary, to be specific. To be in sync with the various advances and updates in deepfake generation methods, advances in detection mechanisms to distinguish them have also become an alarming necessity.

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India. [2]School of Business Administration, Penn State Harrisburg, Middletown, PA, USA. ✉email: brindha.s@vit.ac.in

At present, the current deepfake detection methods cope up with several constraints and limitations. But still, they do not provide robustness against some attack adversaries where minute and well-thought-out undetectable alterations are introduced and injected intentionally. These alterations try to avoid and fool the detection methods[1–3]. The adversarial attacks utilise these vulnerabilities that can take place during feature extraction. Hence, this results in highly confident misclassifications, which can become almost impossible to trace or find. Moreover, these detection models are computationally intensive, as they rely on large-scale deep learning based neural networks, which have complicated and complex architectures. Hence, making them inefficient for deployment in various real-time or in limited-resource environments or scenarios[4–6]. Besides that, no existing detection system offers any mechanism to verify the authenticity or integrity of detection results by data encryption, undermining their credibility for use in legal and forensic settings[7–9]. Most importantly, the key concern is that the current detection methods lack resistance against post-quantum cryptographic threats, as future quantum computing scenarios could expose their vulnerabilities[10,11]. The lack of widespread knowledge about manipulation techniques can further inhibit their use in real-world and dynamic contexts with new and emerging deepfake methods[12–14]. The existing research demands continuous adaptation and more efficient counter methods to ensure trust and reliability.

To overcome the above drawbacks, this research introduces a quantum-resilient, adversarial-resistant robust deepfake image detection system named "DeepQShield", built using ResNeXt convolutional neural network (CNN) architecture[15]. Here, ResNext architecture is chosen due to its modularity, high computational efficiencies, and improved ability to learn subtle manipulations in images, which is crucial for differentiating original and manipulated images. The proposed framework is further reinforced by the incorporation of lattice-based adversarial training through Learning with Errors (LWE) based modifications to improve and enhance the robustness of the system against both traditional and quantum-based adversarial attacks. The integration of the above techniques ensures model efficiency even in the time of adversarial attacks powered by quantum computing.

The major research contributions are:

- Design and development of Quantum-Resilient Detection Framework: A robust deepfake image detection framework DeepQShield is developed, which is capable of withstanding against both traditional adversarial attacks and future post-quantum-based threats by the integration of Learning with Errors (LWE) based changes and cryptographic enhancements.
- ResNeXt based Deepfake detection with Lattice Adversarial Training through LWE: To develop a robust deepfake detection framework leveraging ResNeXt-based Convolutional Neural Network (CNN) as backbone model to proficiently learn and identify fine-grained manipulations in images, enabling precise discrimination between real and GAN-generated deepfake images. To boost the model's resilience, the study integrates lattice-based adversarial training built upon the Learning with Errors (LWE) problem, thereby reinforcing the system against sophisticated adversarial attacks like quantum threats.
- Implementation of Post-Quantum Cryptographic Safeguards: The integration of Kyber for key exchange and Dilithium for quantum-safe digital signatures not only provides data encryption but also ensures authenticity. The verifiable integration of detection outputs supports forensic credibility and maintains reliability in various cases, like legal evidence verification.
- To develop a Fast API-based model framework that is most suited for practical deployment in limited resource scenarios like social media platforms, fact-checking and legal evidence verification systems.
- Verification mechanism: Provides a mechanism for the client to confirm detection results through cryptographic signature verification methods to guarantee the authenticity and integrity of the results.
- Extensive experiments and performance evaluations are carried out to assess the robustness, computational efficiency, security and scalability of the proposed system across various deepfake image datasets such as Deepfake Detection Challenge (DFDC).

This system architecture readily allows its smooth incorporation in the real world and practical domains like fact checking, content moderation in social media platforms and forensic analysis of evidence in legal settings. Unification of optimised and lightweight CNN-based (ResNext-based) detection boosted with post-quantum adversarial robustness and output verification to maintain integrity shows the refinedness and sets it apart as a next-generation deepfake detection system that can be secure and scalable. The further in-depth discussion about the technicality of the DeepQShield, is further explained in the methodology section given below. The remaining paper is arranged in the following way. Section "Related works" explores about the related works in Deepfake detection. Additionally, Sect. "Methodology" gives us the precise explanation of the proposed methodology used. Furthermore, Sect. "Results" reveals the experimentation and results, comparison analytics, ablation studies results and cross-dataset test results. Finally, Sect. "Conclusion" presents the conclusion of the overall model's outcomes.

Related works

The swift advancements in deepfake detection have led to an urgent need to find, assess and reduce social risks. This resulted in the rapid development of deepfake technologies in various fields, primarily focusing on the development of robust and generalized detection mechanisms which can withstand increasingly sophisticated fake media generation techniques. Various methods have been proposed to detect deepfakes that work on diverse modalities and techniques like feature extraction. Numerous existing and conventional models improve detection and reliability across diversified frameworks using different approaches, each explained further in this section.

GAN-based deepfake detection frameworks

Sharma et al.[12] have introduced an ensemble-based framework for the detection of GAN-generated deepfakes among several social media platforms. Their method smartly addresses the challenge raised by the generative adversarial networks (GANs), by refining its performance by setting apart the synthetic GAN-generated images from the original ones by achieving a total reasonably high accuracy of 89.3% on benchmark datasets like Faceswap, Face2Face, Deepfakes and Neutral textures datasets. Another method is the Dual-descriptor method by Jin et al.[1] that works on the reconstruction of a frequency-domain mechanism. This method proves its superiority in the detection of localised manipulations of facial features within the deepfake videos. Furthermore, the framework proposed by Li et al.[16], that is the FDPNet architectural framework, mainly works by using a strong detector that is trained using self-generated multi-scale forged pictures, which ensures improved performance on unseen datasets. D2Fusion, another notable work presented by Qiu et al.[17], this method includes feature superposition to enhance the refinements for the detection of potency, primarily for the multi-model frameworks. Due to these enhancements, they have achieved an impressive AUC score of 0.9942 and 97.77% accuracy on the FaceForensics + + dataset (HQ). An in-depth examination of the imposter bias challenge in deepfake detection was put forward by Casu et al.[2]. This method highlights the use of Generative AI illusions to confuse human as well as automated observers. Hence, calling out for the need for knowledge in forensics and adaptive learning techniques to overcome the current issues. Additionally, a study conducted by Tran et al.[3] provides a deepfake detection system based on meta learning for generalising unknown forgery methodologies and types. Also, it can be inferred that this system has efficiently addressed the cross-dataset performance drops that restrict the traditional detectors. In a recent study by Zhao et al.[18], they explored attention-based architectures in order to capture long-range dependencies in deepfake videos. They proposed ISTVT, an interpretable spatial–temporal video transformer that uses both frame-wise as well as temporal inconsistencies and also provides good explainability through attention visualisation. Their results show improved robustness of deepfake detection on complexly manipulated video datasets but lacks in computational efficiency due to the use of transformers.

Lightweight and real-time deepfake detection models

Lightweight models like MaD-CoRN proposed by Budhiraja et al.[6] are one of the best models for real-time implementation, employed with convolutional reservoir networks enhances the efficiency with reduced computational overhead, achieving an accuracy of 74.7% on manually modified to almost 99.95% GAN-generated data. A comparative study conducted by Kingra et al.[19] evaluates various detection methods on a large-scale Asian deepfake dataset, showing the strengths and weaknesses of the existing models, with top-performing methods reaching above 90% accuracy. Also, emerging technologies are being integrated into many areas, as well as into the judicial system and forensics, to reduce and detect forgery and plagiarism of incident evidence. Among the many emerging technologies, one of these technologies is the plasmonic resonance-enhanced biosensor suggested by Maheswari et al.[4]. This technique detects the expressions and anomalies in facial movements at the microlevel for deepfake detection, but this is still in the experimental phase. Additionally, this technique opens the scope for future advancements in lie detection systems powered by nanotechnology, with potential usage in the judiciary and forensics. One more deepfake detection framework is the DELOCATE detection framework given by Hu et al.[20], which majorly focuses on finding any randomly distributed tampered clues across the video to be detected. This DELOCATE framework ensures precise forensic capability, unlike the simple classification and reaches a moderately good score of 84% on the DFDC dataset trained using FaceForensics++. Apart from these, the study of Schenk et al.[21] displays how deployment of lie detection and deepfake detection algorithms can impact and disrupt the socio-dynamics, such as trustworthiness and accusations. All these point out the vast social impact and implications of the real-time detection modules and frameworks.

Multimodal and audio-based deepfake detection approaches

Several studies have given a technical approach or an ideology for deepfake detection, but Kumar et al.[5] suggested fusion techniques that use both single as well as multimodal setups. This raises the necessity for integration of various complementary features for better detection accuracy. MSFF-Net given by Raveena et al.[22] improves the deepfake detection model by including spatial, frequency and deep semantics feature spaces. Though the use of various feature spaces enhances the detection framework but it lacks adversarial or cryptographic security. Also, the application of deepfake detection for audio-based deepfakes acts as another active area of study, highlighted by Sharma et al.[12] and Xiaoke et al.[23]. But in audio-based deepfake detection use of various accents and languages creates a barrier for getting an optimal solution. The MADD dataset[23], which consists of multilingual support and multi-speaker samples, provides a partial solution to this problem of accents and language differences. The use of this MADD dataset al.so aids in the audio-based detection in the realistic settings. But still, an alarming challenge here is the domain variability in generalising the speaker. Also, the demonstration published by Tahaoglu et al.[24] explains about how ResNeXt-based architecture can be used for audio-based detection. This method uses spectral features such as Mel spectrograms and reached a moderately fine equal error rate (EER) of 1.05% on the ASVspoof 2019 Logistical access (LA) dataset. This EER rate signifies the appreciable performance of convolutional networks in voice forensics. Furthermore, to reduce the problem of domain availability, Samhita et al.[25] gave a self-distillation approach which showcases improved resilience also in scenarios where there are audio variations due to various speakers as well as due to the use of various languages. This technique has achieved a commendable EER of 0.286 using ResNet, 0.337 using ECANet and also 0.371 using SENet on the in-the-wild dataset. Moreover, Gandhi et al.[26] also brought upon a unified multi-modal detection model framework that combines both visual as well as audio data to achieve an appreciable performance enhancement during adverse settings. This framework has got up to 98% on audio and up to 93% on video-based deepfakes.

Post-quantum cryptography and adversarial robustness

The primary focus of our work will be prominently relying on visual deepfakes. To be specific, it addresses about the vulnerabilities caused by adversarial attacks and quantum-level threats. Although this is still proportionately a lesser explored domain, our model will be focusing on building up advancements across various domains. Our proposed model will be working on the ResNext architecture framework as the backbone, enhanced with the help of the lattice-based adversarial defence mechanisms, learn with errors (LWE) to get a quantum-resilient deepfake detection system. This strategy will be addressing the possible potential risks by the use of quantum computing techniques, such as lattice-based cryptographic analysis methods, to take care of the system against futuristic adversarial attacks. Besides this, the incorporation of quantum-competent cryptographic procedures such as Kyber for secure as well as guaranteed key distribution and Dilithium for post-quantum digital signatures acts as a safeguard for the overall system's security against numerous quantum-based threats.

Work put forward by Chawla and Mehra[27] gave a clear roadmap for transitioning from classical to post-quantum cryptography in IoT environments which are 5G-enabled. In this work, they have systematically outlined the security challenges, deployment-related constraints, lattice-based pathways and more. However, their study remains architectural and conceptual. Whereas our study implements the post-quantum-based algorithms for deepfake detection. Further studies suggested by Tian et al.[28] evaluated ADMM-based adversarial false data injection attacks on the multi-label detection systems. This study revealed how optimisation-driven changes can fool the model predictions and signifies the need to address these vulnerabilities. A study by Tian et al.[29] provides a comprehensive analysis of the adversarial attacks and the defence mechanisms to safeguard against those attacks on deep-learning-based unmanned aerial vehicle systems. This work unveiled the vulnerability of vision models due to crafted perturbations. However, this work focuses primarily on classical adversarial settings motivates us to incorporate lattice-based and post-quantum resilient defence mechanisms for securing our systems. Also, as suggested by Maheshwari et al.[10], Lin et al.[11], and Yogarajan et al.[30], the quantum-based enhancements play a major role in the security and safety of the framework. These methods ensure the system's whole integrity by using encryption and authentication of detection results. This works as a barrier between both classical and quantum-based adversarial attacks. To be more specific about the model given by Maheshwari et al.[10], which is the Quantum PlasmoVision imaging, works for enhancing the model's capability for real-time deepfake verification on various deepfake manipulations. This explains the need for emerging improvements by integration of nanotechnology and quantum-based techniques for data authentication.

Moreover, Xu et al.[31] emphasised how public trust and people's perceptions play a key role in gaining acceptance of the various technologies. This highlights the necessity for including both social and ethical implications along with technological constraints. Finally, as said by Chakraborty and Naskar[32], to withstand against sophisticated attacks, the incorporation of human physiology and face biometrics plays a major role in making the system resilient. Collectively, all these advancements point towards the urgent need for an interdisciplinary combined approach consisting of robust machine learning algorithms, cryptographic defences and considerations for humanity to ensure a trustworthy, quantum resilient and morally aligned deepfake detection system. Table 1 depicts the related works section in brief.

Methodology

This section presents the design and implementation of our proposed quantum-resilient deepfake image detection framework. The methodology involves integration of a modular ResNeXt deepfake detection CNN framework as backbone, lattice-based adversarial training enhancements and post-cryptographic safeguarding to support system's robustness against both traditional and future quantum computing threats. The methodology section involves the following subsections, and the system architectural flow is shown by the Fig. 1 as given below:

References	Method used	Dataset(s)	Features	Accuracy	Limitations
Li et al.[16]	Multi-scalar self-forgery image generation using CNN	Celeb-DF, FF++	Multi-scale feature extraction and self-forgery	99.41% (FF++) and 82.46% (Celeb-DF)	Not tested on adversarial attacks
Asha et al.[33]	Spatial-temporal feature extraction	FF++, DFDC, Celeb-DF and custom dataset	CNN + LSTM hybrid, temporal & spatial clues	97% (FF++), 87% (DFDC) and 88% (Celeb-DF)	Limited model complexity and needs GPU optimisation
Brindha and Ragavendra[34]	Structural anomaly learning and decision fusion using spatial-temporal features	FF++ and Celeb-DF(v2)	Fuzzy logic decision fusion	99% (FF++) and 93% (Celeb-DF)	High training time
Yogarajan et al.[30]	Robust deepfake detection using multi-scale feature fusion	DF-TIMIT, FF++ DF, DFD, DFDC, and Celeb-DF,	Multi-layer feature fusion using CNN	97.74% (DF-TIMIT), 97.05% (FF++), 95.22% (DFD), 87.54% (DFDC) and 82.47% (Celeb-DF)	Decreased robustness against real-time forgeries
Budhiraja et al.[6]	MaD-CoRN: Lightweight deepfake detection with CNN and using Reservoir Computing	FFHQ, 100 K-Fake (12 K, 14 K, full dataset)	Efficient convolutional-reservoir fusion (Inception + ResNet + RC)	99.6% (12k), 99.7% (14k), and 99.95% (full)	Needs evaluation on adversarial or low-resolution forgeries
Maheshwari et al.[10]	Quantum PlasmoVision-Based Imaging	DFDC, FF++, and custom dataset	Quantum-enhanced image sensing, plasmonic imaging	98.2% (FF++) and 94.1% (DFDC)	High hardware cost, complex integration for real-time deployment
Ben Jabra et al.[35]	Ensemble learning using InceptionV3, VGG16, Xception	140k Real and Fake Faces	Deep ensemble of feature extractors with CNN ensemble fusion	97.4% (140k Real and Fake Faces)	Slightly heavier runtime, ensemble tuning needed

Table 1. Comparative analysis table of existing deepfake detection works.

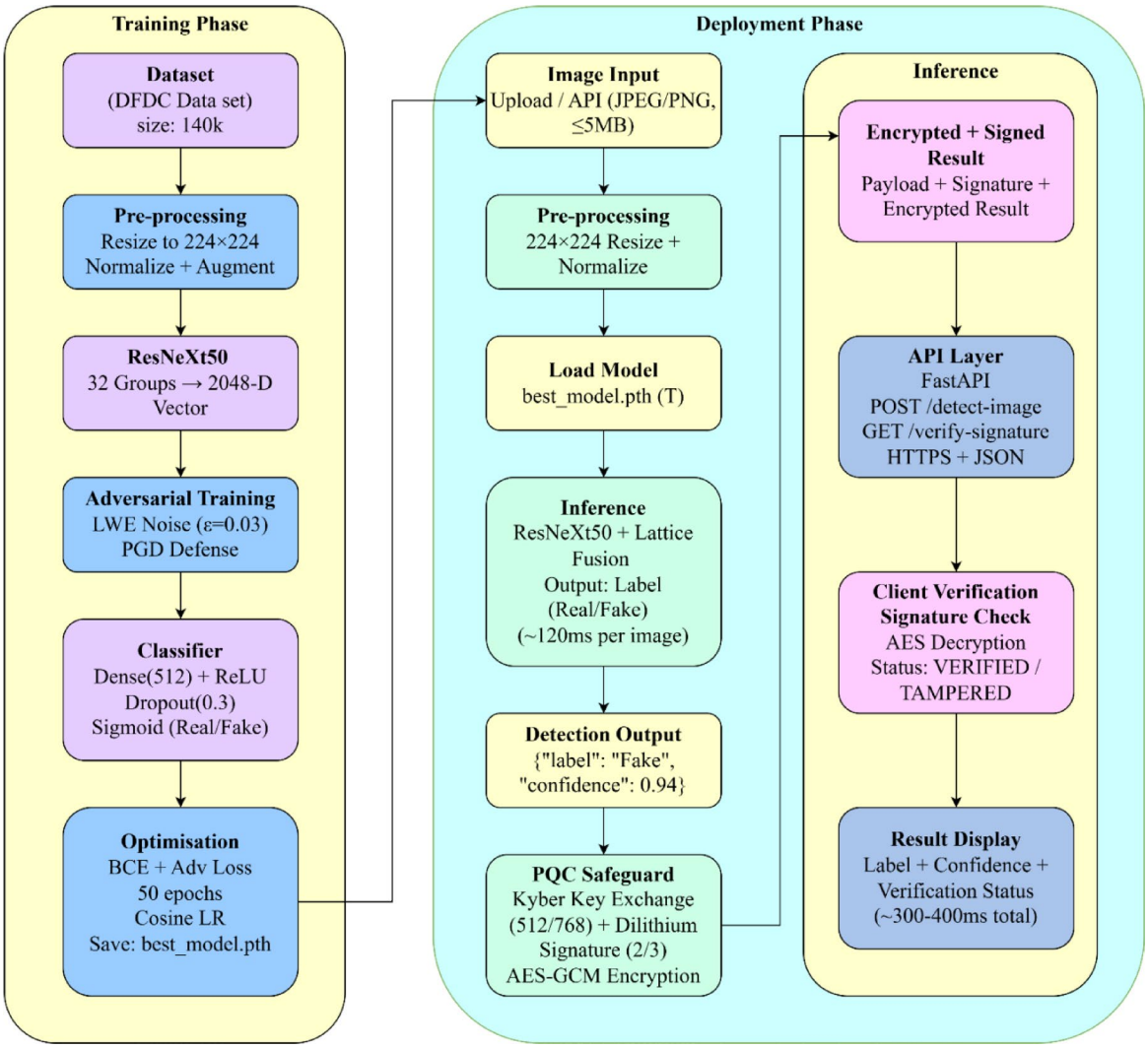

Fig. 1. System architecture of the proposed DeepQShield deepfake detection framework.

Figure 1 illustrates a secure quantum resilient deepfake detection system workflow of DeepQShield. The workflow starts with the use of a large dataset (DFDC, 140k images), a pre-trained image dataset, which is then resized to 224×224, normalised and augmented to make it ready for training. ResNeXt50 is used as the backbone for the detection mechanism, which is enhanced using lattice-based adversarial training. The adversarial training involves the introduction of perturbed noise using the learn with errors mechanism (LWE) with PGD (Projected Gradient Descent) defence. This step is followed by a classifier (Dense + ReLU + Dropout) to predict whether labels are real or fake. Further, the model is optimised using BCE (Binary-Cross entropy) and Advanced loss functions over 50 epochs, also with the help of cosine learning rate scheduling. Then the images are passed as input through upload or API, and the model is set for training, and the trained model (best_model.pth) is loaded for use in inference. Inference combines both the ResNeXt50 and lattice-based enhancements to predict the output along with the confidence scores and throughput (~120 ms/image). The detection results are then encrypted and signed using the post-quantum cryptography (PQC) security algorithms like Kyber key exchange and Dilithium signatures. An API layer (Fast API) is further used to handle requests and responses for the detection and verification of signatures for a safe detection result prediction. Finally, the results are decrypted and verified by the client to maintain integrity. Label confidence is then displayed to the client for safe and secure quantum resilient deepfake detection.

Data preparation and preprocessing

This proposed architecture uses the publicly available Deepfake Detection Challenge (DFDC) image dataset[36]. In total 140k images, the dataset consists of both 70k real images and 70k fake images, which is used for both training and validation of data to keep a balance between representation of authentic and generated visual images. Data processing was also done to normalise the inputs to enable efficient and effective training of the model via the steps shown in Fig. 2. The following procedures have been carried out on each image, which includes,

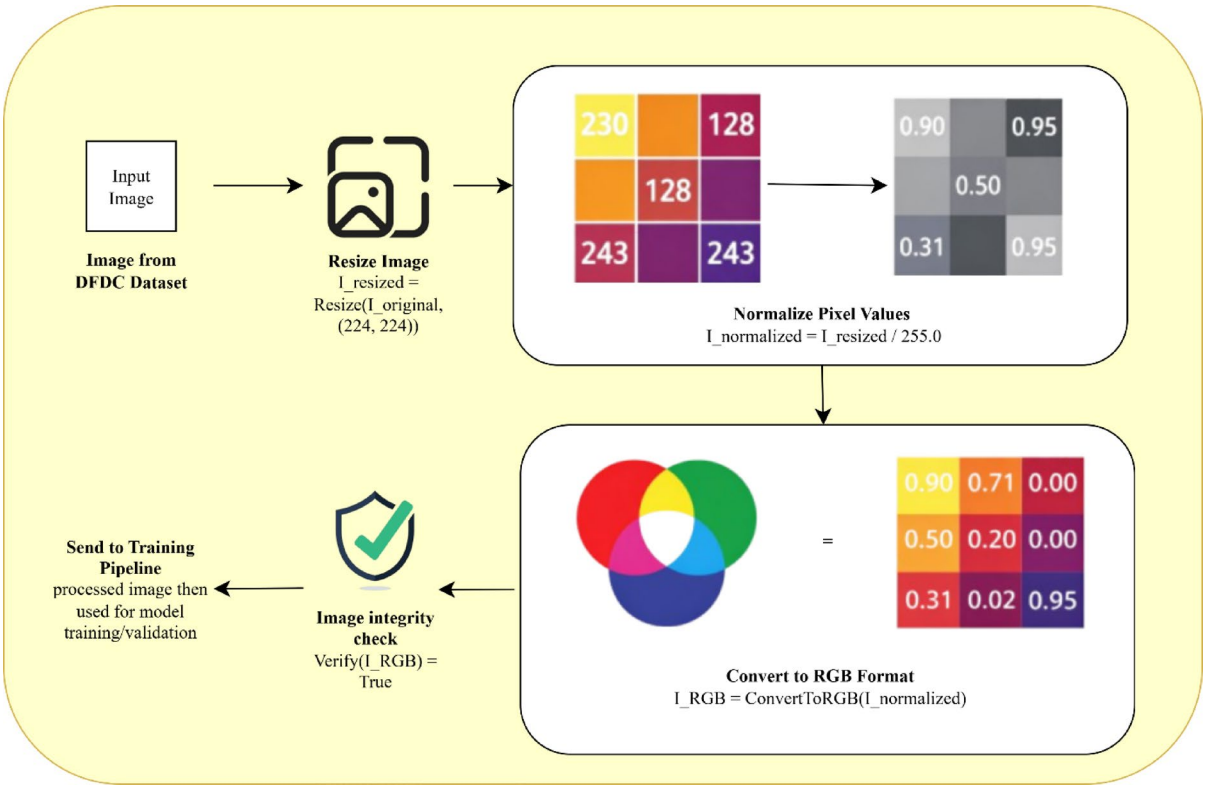

Fig. 2. Dataset preparation and pre-processing.

- Resizing: Each input image is resized into a uniform spatial resolution of about 224 × 224 pixels as shown in Eq. (1), to cope with input size requirements of the convolutional neural network architecture.

$$I_{resized} = Resize\left(I_{original}, (224{,}224)\right) \tag{1}$$

- Normalisation: For each resized image, its pixel intensity levels have been normalised to the [0, 1] range by dividing by 255 as per Eq. (2), which stabilises the learning rate as well as speeds up the convergence while training the dataset.

$$I_{normalized} = \frac{I_{resized}}{255.0} \tag{2}$$

- Colour Channel Conversion: After normalisation of images, each normalised image is then converted into RGB format as per Eq. (3) to ensure consistency in representation of diverse colours and to make it compatible with pretrained convolutional layers, which were initialised on the RGB datasets.

$$I_{RGB} = ConvertToRGB\left(I_{normalized}\right) \tag{3}$$

- Data Integrity Checks: All images then go through a few checks to confirm that they are corruption-free and are in a valid file format before training with the help of the validation in Eq. (4). The checks are designed to prevent any type of corruption or incompatibility of files during training to ensure the model's stability. The checks include file format validation checks, file readability checks, checks if image height and width are greater than zero, confirms the presence of valid colour channels and scans pixel data to test for invalid numeric entries after the normalisation.

$$Verify\left(I_{RGB}\right) = True \tag{4}$$

The uniformity in the preprocessing pipeline guarantees the dataset is clean and consistent and signifies that the dataset is ready to use as an input to the model, aiding towards stabilised learning rates and valid result classifications.

ResNeXt CNN architecture enhanced with lattice cryptographic framework

This section in the methodology for DeepQShield construction combines the ResNeXt-50[15] backbone with LWE-based lattice adversarial training for improved deepfake detection. Followed by introducing a structured noise for cryptographic resilience. Furthermore, the feature fusion approach blends the backbone and lattice

features for a more comprehensive representation. An attention mechanism further optimises the importance of these features. Additionally, by incorporation of highly advanced loss functions, such as focal loss and label smoothing, enhances classification under imbalanced data conditions. Each architectural component of the proposed DeepQShield framework is explained below.

Learning with errors (LWE) foundation and its implementation in lattice layer

This subsection introduces a quantum-resilient defence using LWE-based lattice perturbations during adversarial training to improve the model's robustness, as shown in Fig. 3. This quantum resilient defence mechanism is built on the Learning with Errors (LWE) problem[37], which makes up the mathematical foundation of post-quantum cryptography. The LWE instance is defined by Eq. (5) as:

$$(A,\ b) \in\ Z_q^{n\times\ m} \times\ Z_q^m,\ \ where\, b\ =\ A \cdot s\ +\ e\ (mod\, q) \tag{5}$$

where, A represents a randomly chosen matrix, s a secret vector, e a small error vector sampled from a discrete Gaussian distribution and q represents the modulus parameter used to define the arithmetic space.

In the field of deep learning, the introduction of a small error vector e injects structured perturbations into the model's learning process. The introduction of this controlled noise makes the model noise-tolerant and enhances its ability to generalise. This also reduces sensitivity to adversarial manipulations and preserves its discriminative capacity.

In order to enhance resilience against quantum-based threats, Learning with Errors (LWE)-based lattice perturbations during the adversarial training is introduced. This procedure utilises the computational hardness of lattice-based problems to create various perturbations that create difficulty for the systems to detect the changes easily. In particular, lattice adversarial examples are generated by making subtle and minute variations in the input images using lattice noise design patterns given by the LWE assumptions. The model was then trained concurrently on cleaned and lattice-perturbed data for inducing invariance against such advanced attacks. Intuitively, this training strategy pushes the model to focus more on stable semantic cues and not the brittle noise-sensitive patterns. This helps the model to improve its robustness without giving up on the discriminative performance. Generally, the lattice adversarial training goal is given by the following Eq. (6).

$$\min_{\theta} E_{(x,y)}\left[\max_{\delta\ \in\ S} \mathcal{L}\left(f_\theta\ (x+\delta\ ),\, y\right)\right] \tag{6}$$

where, θ denotes parameters of the model, $\mathcal{L}$ refers to the binary cross-entropy loss function, $f_\theta\ (\cdot)$ represents the ResNeXt deepfake detection model and δ denotes the adversarial lattice perturbation given by the LWE-based lattice space S.

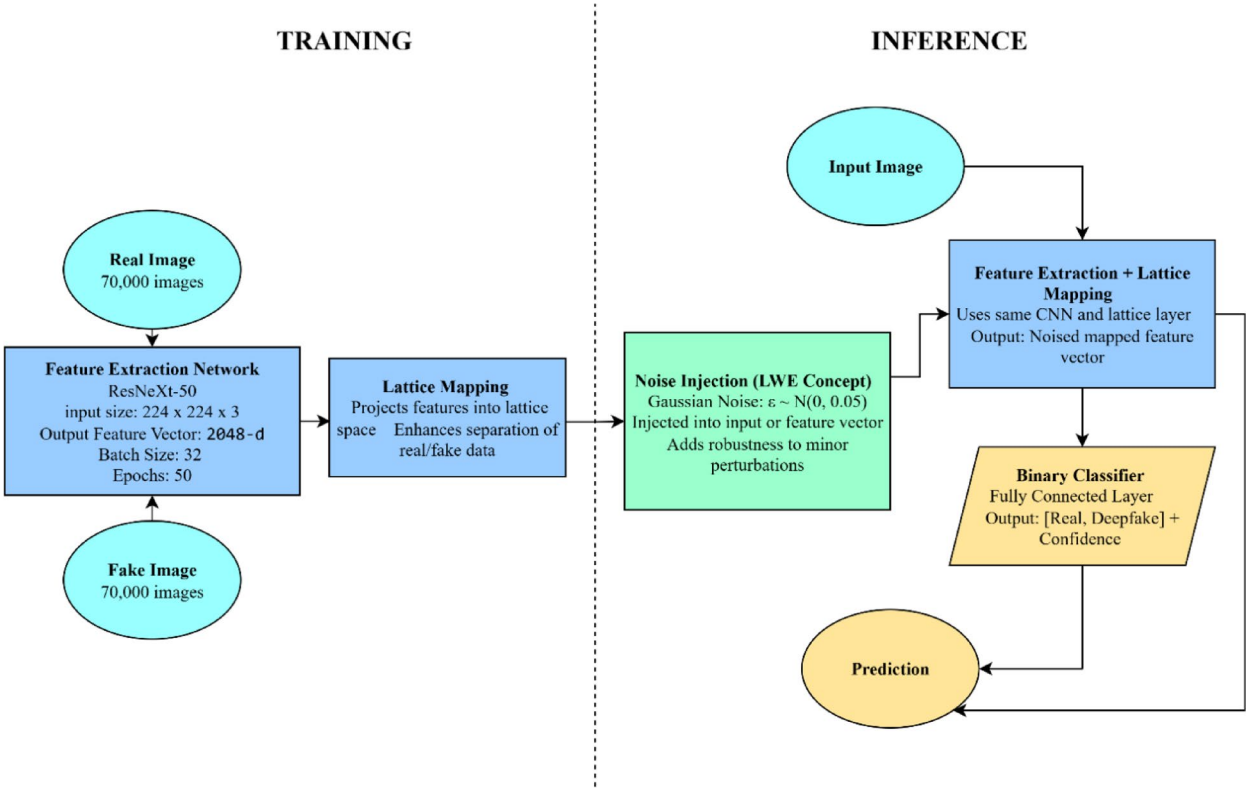

Fig. 3. Learning with errors (LWE) foundation and its implementation in lattice layer.

The inner maximisation gets a perturbation. δ inside the lattice constraint set that tends to maximise the prediction error given by Eq. (7):

$$\delta^{*} = \underset{\delta \in S}{\operatorname{argmax}} \mathcal{L}\left(f_{\theta}\ (x+\delta\ ), y\right) \tag{7}$$

The δ^{*} represents the specific perturbation that achieves the maximum of $argmax$. While the outer minimisation updates θ for minimisation of loss in the worst-case caused due to these perturbations δ^{*}. This ResNeXt detection model cohesively trained with lattice-perturbed samples, promotes robustness and invariance to adversarial manipulations that advancements in quantum computing may pose in future.

To further reinforce the robustness, a lattice-based learning layer is added. The lattice-based learning layer introduces structured noise following the Learning with Errors (LWE) as shown in Fig. 4, and applies a transformation defined as a lattice transformation given by Eq. (8):

$$L\left(x\right) = LayerNorm\left(x \cdot B + \varepsilon\ \right) \tag{8}$$

Where, B is the learnable lattice basis matrix and ε is the structured error term added to simulate LWE-based noise.

Intuitively, the lattice transformation uses controlled randomness inside the feature space. This forces the model to learn the features that are structured manipulations resilient instead of just relying on fragile noise-sensitive features. By including this layer, it moreover stabilises the feature distributions and makes sure of consistent feature learning at the time if training.

The Lattice Layer further transforms the input features using a learned basis matrix then the structured noise is injected to enhance the robustness. Subsequently, it is normalised to stabilise training and improve the system's generalisation capability. This can be understood by the lattice layer forward pass mechanism, as given in Algorithm 1.

Input: Feature vector $x \in R^{d}$

Output: Transformed features $L(x)$

Steps:

1. Initialise lattice basis $B \in R^{d \times k}$
2. Compute the linear transformation of the input feature vector $y = x \cdot B$
3. Generate structured gaussian noise $\varepsilon \sim N(0, \sigma^{2} I)$
4. Add noise during the training phase $y' = y + \varepsilon$
5. Normalisation is applied to each layer for stabilising the feature distribution $L(x) = LayerNorm(y')$
6. Return the transformed feature vector $L(x)$

Algorithm 1. Lattice layer forward pass

By adding various transformation stages to the model, the advanced lattice cryptography module builds on the fundamental lattice layer. Complex and expressive representations are enabled through a series of lattice layers. Each layer is followed by a non-linear activation function rather than a single transformation. The Eq. (9):

$$L_{multi}\left(x\right) = f_{N}\left(L_{N}\left(\ldots\ f_{2}\left(L_{2}\left(f_{1}\left(L_{1}\left(x\right)\right)\right)\right)\right)\right) \tag{9}$$

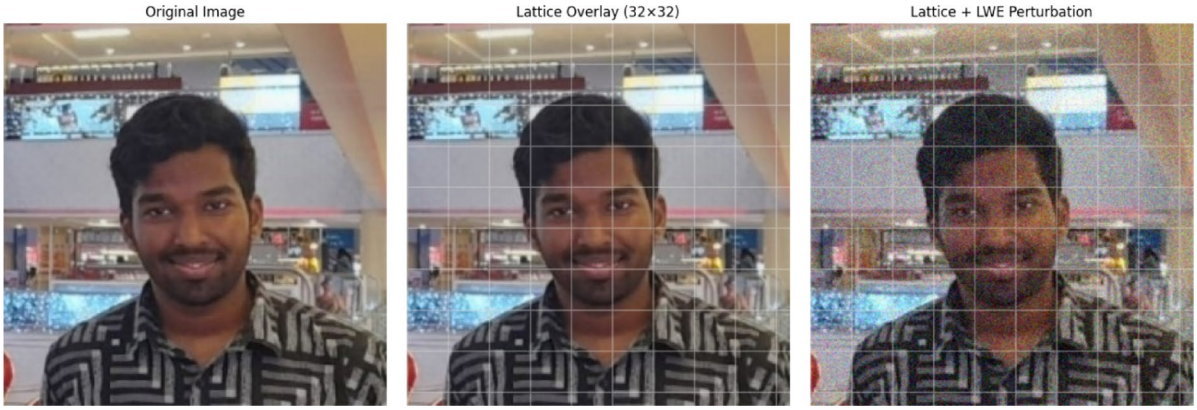

Fig. 4. Introduction of lattice overlay and LWE Perturbations.

Here, the i^{th} lattice layer is represented by L_i, and the matching activation function is indicated by f_i. The model's learning ability and cryptographic strength are both improved by this layered structure.

In the field of deepfake detection, our DeepQShield uniquely operates Learning with Errors (LWE) for attaining feature-level robustness rather than just using pure cryptographic mechanisms. In place of just applying perturbations at the input image space, we have injected structured LWE-based noise into the intermediate feature representations through lattice-based embedding within the ResNeXt backbone. This style of design forces the network to learn invariants of manipulated facial features through the disruption-based shortcut learning. This is associated with generator-specific artifacts that are widely and commonly exploited by the deepfakes. Unlike several conventional adversarial training techniques that solely depend on gradient-based attacks or explicit adversarial examples generation methods, the LWE-based perturbation works like a regularisation mechanism. The LWE mechanism is rooted in lattice hardness assumptions and hence acts for improving resilience against both adversarial manipulated features as well as unseen forgery patterns. This combination bridges the adversarial robustness with post-quantum safeguards inside a single detection system.

Enhanced ResNeXt architecture

ResNeXt-50 architecture works as a backbone for the proposed deepfake detection framework. This architectural backbone is selected due to its cardinality characteristics, as it eases parallel path aggregation inside the grouped convolutions, as shown in Fig. 5. This leads to improvised accuracy and computational efficiency of the design, and enables the model to efficiently discriminate facial features needed for differentiating real and GAN-generated visual data. Then, to effectively extract complex hierarchical structures, group convolutions are used with various parameters. Mathematically grouped convolution operations can be shown in Eq. (10), which is as follows:

$$Output = \bigoplus_{i=1}^{C} Conv_i \left(Split_i \left(input \right) \right) \tag{10}$$

where C denotes the cardinality, $Split_i$ stands for the partitioning of the input feature map, $Conv_i$ represents the convolution applied to each partition, and $\oplus_i$ represents aggregation by summation. This is followed by the addition of bottleneck layers to reduce the computational load without inhibiting representation abilities. Each bottleneck block is based on the following transformation, given below in Eq. (11):

$$y = x + F\left(x\right) = x + W_3\sigma\ \left(BN\left(W_2\sigma\ \left(BN\left(W_1x\right)\right)\right)\right) \tag{11}$$

where x denotes the block input, $W_1,\ W_2,\ W_3$ are the weight tensors corresponding to $1 \times\ 1,\ \ 3 \times\ 3$ grouped, and $1 \times\ 1$ convolutions respectively, BN indicates batch normalisation, and σ Denotes the ReLU activation.

The final dense layer generates a scalar output for specifying the probability of an input image being fake, which is then calculated using Eq. (12) by applying a sigmoid activation function:

$$\overline{y} = \sigma\ \left(w^T h + b\right) \tag{12}$$

where h is the flattened feature vector gained from the final pooling operation and w, b These are the trainable weights and bias of the output layer. The sigmoid function $\sigma\ \left(z\right)$ is used to map the output z to the range (0, 1) and is defined by Eq. (13). This probabilistic formula that allows the model to denote its prediction confidence.

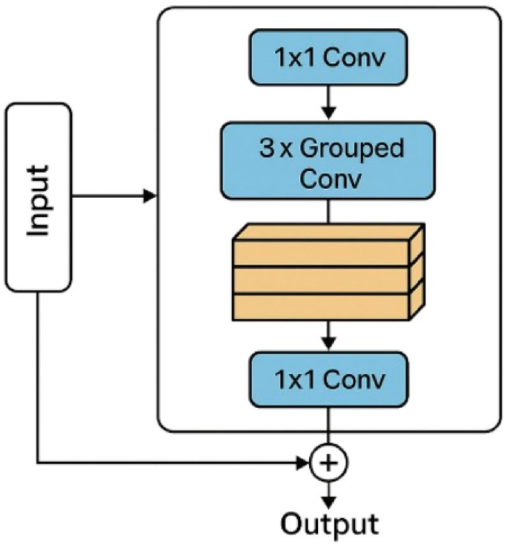

Fig. 5. ResNeXt layers.

$$\sigma\ (z) = \frac{1}{1 + \exp\ (-z)} \tag{13}$$

The network is then trained using binary cross-entropy loss as mentioned in Eq. (14), which measures the divergence between the predicted probability and true label:

$$L = -\left[u \cdot\ \log\left(\overline{u}\right) + (1 - u) \cdot\ \log\left(1 - \overline{u}\right)\right] \tag{14}$$

where $u \in\ \{0,1\}$ denotes the label for ground truth and $\overline{u} \in\ (0,1)$ denotes the predicted probability. Hence, this architectural setup ensures intensive feature extraction without compromising on computational feasibility, and thus making it the best fit for robust deepfake detection applications.

The output of the ResNeXt model is given by Eq. (15):

$$y = x + \sum_{i=1}^{C} T_i\left(x\right) \tag{15}$$

where C represents the cardinality and T_i represents the i^{th} transformation path. By increasing the cardinality instead of depth or width, the ResNeXt results in richer feature representations. Also, it maintains computational efficiency making it suitable for scalable deepfake detection applications.

Feature fusion strategy

The model uses a complex feature fusion mechanism that is equipped with lattice-enhanced representations with backbone features, as explained in Algorithm 2. The feature-fused model is achieved by using Eq. (16):

$$F_{combined} = \ [F_{backbone};\ F_{lattice}] \tag{16}$$

where $[;]$ represents the concatenation operation.

In step 4 of Algorithm 2, the concatenated features are then passed through the multi-layer perceptron (MLP) model that is used for dimensionality reduction, non-linear projections, and for interactive modelling. This step ensures that the fused features preserve the complementary information from the backbone and lattice spaces and also enhances the differentiating power by reducing redundant components. This highlights the informative cues relating to the fake speech or image detection.

Input: Raw image $I \in R^{3\times224\times224}$

Output: Enhanced features $F_{enhanced}$

Steps:

1. Extract backbone features using ResNeXt backbone: $F_{backbone} = ResNeXt(I)$
2. Apply a multi-level lattice-based transformation to the extracted features in the backbone: $F_{lattice} = L_{multi}(F_{backbone})$
3. Concatenate features: $F_{combined} = [F_{backbone}; F_{lattice}]$
4. Apply feature enhancement: $F_{enhanced} = MLP(F_{combined})$
5. Compute attention weights: $\alpha = Attention(F_{enhanced})$
6. Apply attention: $F_{final} = \alpha \odot F_{enhanced}$
7. Return F_{final}

Algorithm 2. Feature enhancement pipeline

Attention mechanism

The model's ability to work on the prominent areas for the input representation is improved by integrating the attention mechanism, which dynamically weights feature importance. The attention weights are computed with the help of Eq. (17):

$$\alpha\ = \sigma\ (W_2 \cdot\ ReLU\ (W_1 \cdot\ F_{enhanced} + \ b_1) + \ b_2) \tag{17}$$

Where W_1, W_2 are the learnable weight matrices, $b_1,\ b_2$ are the bias terms, and σ is the sigmoid function.

Advanced loss functions

It presents focal loss and label smoothing to tackle class imbalance and overconfidence in predictions, combined into a weighted total loss.

Focal Loss Implementation: To address the class imbalance and the hard sample mining, the system implements focal loss[38] given by Eq. (18):

$$FL\left(p_t\right) = \ -\alpha_t(1\ -\ p_t)^{\gamma\ log}\ (p_t) \tag{18}$$

Where α_t denotes the weighting factor, p_t represents the model's estimated probability and γ denotes the focusing parameter.

Label Smoothing Loss: Label smoothing regularisation is then applied to the prediction model to prevent overconfident predictions, given by Eq. (19):

$$L_{smooth} = (1 - \varepsilon) \cdot L_{CE} + \varepsilon \cdot L_{uniform} \tag{19}$$

where ε denotes the smoothing parameter, L_{CE} the cross-entropy loss and $L_{uniform}$ denotes the uniform distribution loss.

Combined Loss Function: The final loss function combines both the focal loss and the label smoothing mechanism, where the combined loss is derived with the help of Eq. (20) given below:

$$L_{total} = \lambda_1 \cdot F_{loss}(p_t) + \lambda_2 \cdot L_{smooth} \tag{20}$$

Where $\lambda_1 = 0.7$, $\lambda_2 = 0.3$ and λ_1, λ_2 are the weighting coefficients. The L_{total} represents the total loss obtained by the combination of focal loss represented by F_{loss} and label smoothening represented by L_{smooth}.

Training strategy and optimization

This section gives us an overall summary of the training process. The training process involves studying various adversarial samples, regularization, and optimization techniques to ensure stability and accuracy.

Optimizer configuration: The training procedure makes use of the AdamW optimizer[39] with restarts using cosine annealing. The AdamW optimizer uses the Adam update rule to update the parameter at each step t given by Eq. (21):

$$\theta_{\{t+1\}} = \theta_t - \eta_t \cdot \left(\frac{\widehat{p}_t}{\left(\sqrt{\widehat{q}_t} + \varepsilon \right)} + \lambda \cdot \theta_t \right) \tag{21}$$

where η_t represents the learning rate of the model, $\widehat{p}_t$ and $\widehat{q}_t$ are the bias-corrected moment estimations, and λ represents the weight decay parameter of the AdamW optimiser. The θ_t is the parameter vector at each step t and ε denotes a small constant to avoid division using zero.

Learning Rate Scheduling: The cosine annealing warm restart scheduler is used to implement the cosine annealing schedule, which is calculated with the help of Eq. (22):

$$\eta_t = \frac{\eta_{min} (\eta_{max} + \eta_{min}) \left(1 + cos\left(\pi \cdot \frac{T_{cur}}{T_{max}}\right)\right)}{2} \tag{22}$$

where, T_{cur} denotes the current epoch and T_{max} represents the maximum epochs in the current restart cycle. The η_{max}, η_{min} are the maximum and minimum learning rates with the cosine annealing and λ denotes the weighting decay coefficient.

The overall training procedure, which has taken place in building DeepQShield, is conveyed with the help of Algorithm 3, which precisely shows the step-by-step method involved in the training of the lattice-enhanced ResNeXt model.

Input: Training dataset D_{train}, Validation dataset D_{val}

Output: Trained model $M*$

Steps:

1. Initialise the model M by assigning the pre-trained weights of ResNeXt
2. Set the optimiser AdamW with an $learning\ rate = 1e^{-4}$, $weight_{decay} = 1e^{-4}$
3. Configure the AdamW scheduler with $CosineAnnealingWarmRestarts$ ($T_0 = 10$)
4. For $epoch = 1$ to max $epochs$:
5. Training phase:
6. For each $batch$ (X, y) in D_{train}:
7. Carry out the forward pass and get the predictions ŷ and features F: $F = M(X)$
8. Compute the total loss: $L = L_{total(\hat{y},y)}$
9. Using the backward pass, compute the gradients: ∇L
10. Update the model parameters: $M \leftarrow AdamW(M, \nabla L)$
11. Validation phase:
12. Evaluate the model on D_{val}
13. Compute the metrics required for validation
14. Update learning rate: $scheduler.step()$
15. Early stopping check
16. Return the best model $M*$

Algorithm 3. Training procedure

Implementation of post-quantum cryptographic safeguards

In this deployment phase, our model DeepQShield is introduced with Post-Quantum Cryptographic safeguarding algorithms and various essential elements for attaining quantum-resilience. The DeepQShield's architecture is divided into three tiers. The first tier is the Presentation layer (Frontend) with a React.js-based web interface that includes user interaction features, a real-time feedback mechanism, as well as cryptographic evidence visualisation. The next layer is the Application Layer (Backend) based on a Flask-based REST API server. This layer helps in coordinating with the AI model inference as well as taking care of the working of cryptographic operations and data handling. Also, this server makes use of a concurrent processing mechanism with the help of ThreadPoolExecutor for extra optimisation. The third and final tier is the Security Layer (Quantum-Safe Module), which is a cryptographic module that is based on NIST-accepted post-quantum algorithms for validating results and for ensuring security as well as privacy. System Design Principles formulate the fundamental requirements for cryptographic resilience integration without affecting the system performance. The Base-Architecture forms the basis, constituting the ResNeXt50 backbone. Standing on this foundation, the Novel Lattice-Based Enhancement introduces LWE-driven perturbations for quantum-resistant learning. For authenticity and verification, cryptographic algorithms are incorporated, the choice outlines the reasoning for selecting secure NIST-approved post-quantum algorithms. The Cryptographic Workflow explains how Kyber is used for encryption and Dilithium for digital signatures. Lastly, Security Metadata Generation provides traceability and integrity through inclusion of cryptographic proofs within model outputs, summing to a complete post-quantum defence strategy.

Base network architecture

The ResNeXt-50 architectural model is used as the base network for our model, DeepQShield. It is a state-of-the-art convolutional neural network architecture, and it comes in work as a detection core for the model. This is a well-known base architecture due to its excellent performance as well as due to its modular design. It operates on $224 \times 224 \times 3$ RGB images, along with the incorporation of several trainable parameters consisting of over 50 million. Due to the presence of a very large number of trainable parameters enables it to learn numerous intricate visual patterns. It has a depth of 50 convolutional layers and is integrated using residual connections, this ensures the gradient flow is efficient at the time of model training. The key characteristics of ResNeXt-50 includes the cardinality factor. This cardinality factor utilises 32 grouped convolutions within each of the residual blocks, making it suitable for increasing model capability without the use of additional overheads for computation. To upgrade the feature extraction mechanism, the architecture works with the pre-trained ImageNet weights. This makes the model effective for transfer learning as well as to speed up the convergence.

Novel lattice-based enhancement

One of the significant innovations introduced in this work is the incorporation of the lattice-based cryptographic methods into the neural network architecture. This integration aids in enhancing deepfake detection. These enhancements include principles inspired by the Learning with Errors (LWE) problem. This integration of errors into the lattice layers adds some structured randomness into the feature space. This results in both a regularised and secure mechanism. The transformation layer comes up with a formulation on lattice-based features that introduces cryptographic noise during the stage of feature transformation. This stage takes into account in improving generalisation ability for defending against adversarial manipulations and changes. Algorithm 4 explains to us in detail about the working of the Lattice-Based Feature Transformation Layer and also the forward pass mechanism linked to it.

Input:

Feature vector $x \in \mathbb{R}^{input\_dim}$

Lattice basis matrix that is learnable $B \in \mathbb{R}^{input\_dim \times output\_dim}$

Error scale vector which is learnable $\sigma \in \mathbb{R}^{output\_dim}$

Standard deviation of the optimal gaussian noise $noise_{std} \in \mathbb{R}$ (default:0.1)

Output:

Lattice features that are transformed $vector\ y \in \mathbb{R}^{output\_dim}$

Steps:

1. Initialization:
2. Initialize the lattice basis matrix B with the help of Gaussian weights scaled by 0.1.
3. Initialize an error scale vector $\sigma \leftarrow noise_{std}\ 1\ (default\ noise_{std} = 0.1)$.
4. Forward Pass:
5. Compute the linear Projection of the lattice basis matrix B: $z \leftarrow x \cdot B$
6. Sample the Gaussian noise using the learned error distribution: $\varepsilon \sim \mathcal{N}(0, diag(\sigma^2))$
7. Inject noise into the projected features z: $z \leftarrow z + \varepsilon$
8. Return the output y: $y \leftarrow z$ (transformed lattice features)

Algorithm 4. Lattice-based feature transformation layer

The implementation of the lattice layers is a structured noise that is based on the LWE problem. This problem provides:

- Enhanced robustness against several adversarial attacks.
- Cryptographic security assurance.
- Improved generalisation with the help of regularisation.

Cryptographic algorithm selection

For attaining further enhanced security, the system makes use of the Kyber and the Dilithium. These are two lattice-based post-quantum cryptographic algorithms. Kyber is used for safe and secure key exchange. It provides 256-bit resistance to quantum attacks (Kyber-1024)[40] and predominantly works on lattice-based key establishment methodologies. This technique also supports a sub-millisecond key generation mechanism, thus making it best-suited for establishing efficient and secure session setup between servers and clients. In contrast to Kyber, Dilithium is used for the generation of digital signatures. This also provides the same 256-bit quantum resistance (Dilithium-5)[41]. Due to the small signature sizes present in the Dilithium, it guarantees efficient verification of signatures. This method is also very efficient only, with a time for signature generation generally being less than 100 milliseconds. This combination takes in care of performing both the roles. That is to guarantee the authenticity of detection results as well as non-repudiation of the detection results. Hence, these algorithms act as a safeguard to the system by providing it with strong quantum-safe adversarial protection.

Cryptographic workflow

This subsection explains about the cryptographic security workflow for the system. This workflow contains various steps starting with the session setup phase. In this phase, the client raises the need for a quantum-safe and secure session to attain the system's reliability. This phase is followed by a secure key exchange process. In this process, the Kyber-1024 algorithm is used for generating a shared symmetric key according to the requirements in lattice-based cryptography with 256-bit quantum resistance. After securing the session using Kyber, images are processed. The processed images that are obtained are further encrypted and safely passed for further analysis. After the images are processed, the system then produces a digital signature for authentication. The digital signature produced will be in the form of the Dilithium-5 digital signature to obtain the detection results. This offers both authenticity and non-repudiation. A safe and secure package is then built, consisting of the detection result as well as the Dilithium-5 signature, which contains the algorithm's metadata, timestamps, and compliance information. To ensure the overall cryptographic evidence design. Finally, the client-side verification enables clients or third parties to independently authenticate the cryptographic proofs or results using public keys to ensure the system's integrity, authenticity, and acceptability of the results for any legal or forensic purposes. This hybrid approach provides maximum security by combining the best of both worlds - quantum resistance from PQC algorithms and security from well-established classical algorithms. The visualised understanding of the workflow is given in Fig. 6, as shown below. Additionally, the core components of Post-Quantum and Classical Cryptographic Algorithms are provided in Table 2.

The Algorithm 5 given below is used to generate the post-quantum secure keypair with the help of Dilithium-5 and optionally includes the use of classical keys like RSA-4096 (Rivest–Shamir–Adleman, 4096-bit) and ECDH-P384 (Elliptic Curve Diffie–Hellman, P-384) to generate a hybrid key. The hybrid key structure helps in the enhancement of a future-proof and backwards-compatible mechanism.

Input: Security level (default: 5)
Output: Hybrid key pair ($HybridKeyPair$)
Steps:
1. Generate PQC keypair using Dilithium-5
2. If there are classical cryptographic primitives are available, then:
3. Generate $RSA-4096$ keypair
4. Generate $ECDH$ keypair using $P-384$ elliptic curve.
5. Combine the PQC keypair and classical key pairs into $HybridKeyPair$ structure
6. Store and return the $HybridKeyPair$ along with the metadata and expiration time

Algorithm 5. Hybrid key generation

After the hybrid key generation, Algorithm 6 gives us a precise idea about the dual key exchange, that is, the Kyber-5 (PQC) based key for quantum resistance and ECDH for classical compatibility. The resulting encrypted results are then concatenated and processed with help of HKDF-SHA3-256 (HMAC-based Key Derivation Function) with SHA3-256. to generate a well-built shared key.

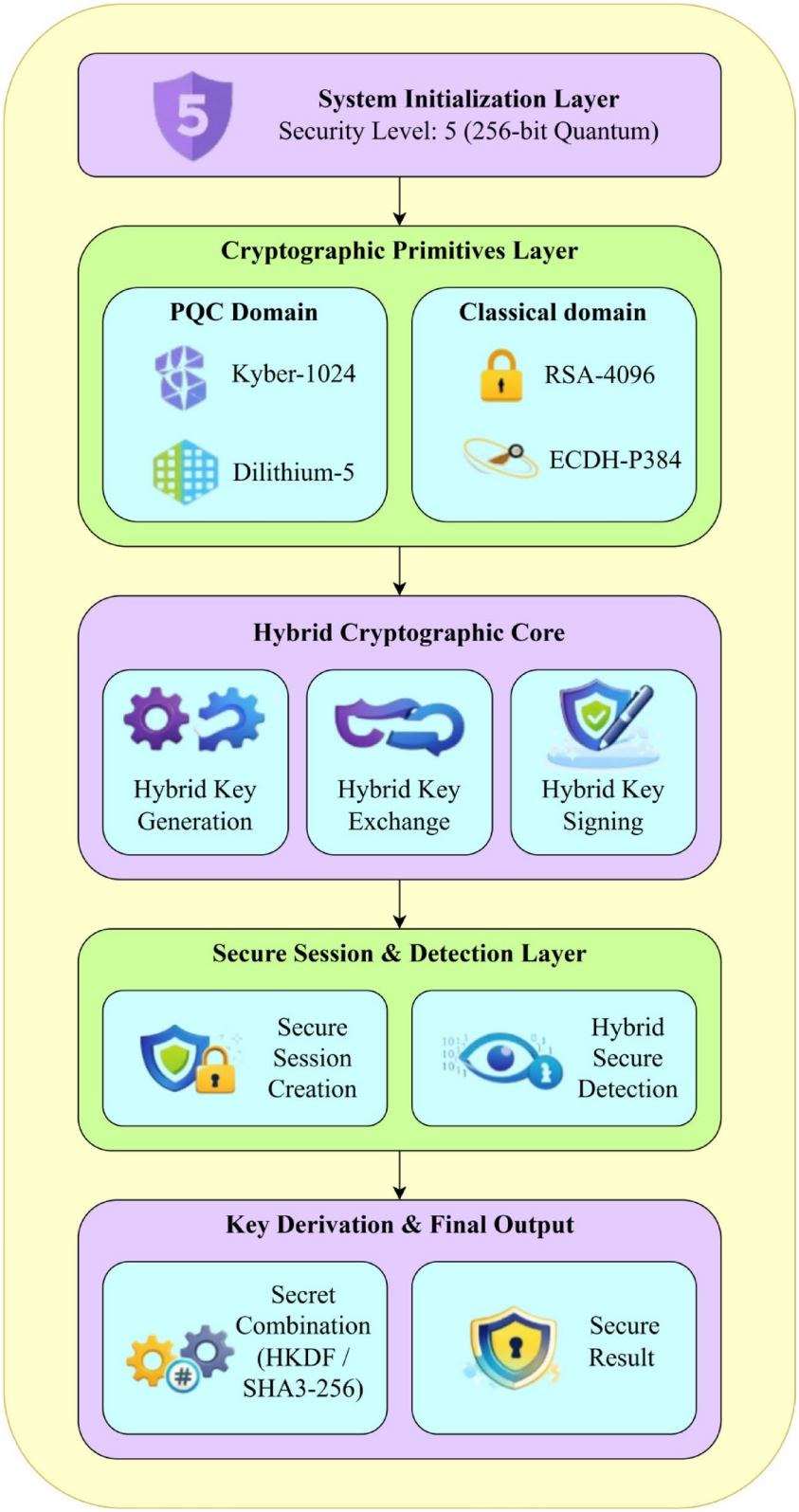

Fig. 6. Post-quantum Cryptographic workflow for secure deepfake detection.

Core components	Algorithm	NIST Security Level	Function
Post-Quantum Algorithms (Primary Security)	Kyber-1024	Level 5	Key Encapsulation Mechanism (KEM)
	Dilithium-5	Level 5	Digital Signatures
Classical Algorithms (Backward compatibility)	RSA-4096	~Level 5 Equivalent	Digital Signatures
	ECDH-P384	~Level 5 Equivalent	Key Exchange

Table 2. Core components: post-quantum and classical cryptographic algorithms.

Input: Peer public keys $\{Kyber_{public}, ECDH_{public}\}$
Output: Combined shared secret $final_{secret}$
Steps:
1. PQC Key Exchange:
2. Perform Kyber encapsulation with the help of the Kyber peer public key and get both the shared and encapsulated Kyber secret:
$shared_{pq}, encaps_{pq} = Kyber.Encapsulate(peer_{kyber\ public})$
3. Classical Key Exchange:
4. Compute the shared classical secret using the elliptic curve:
$shared_{classical} = ECDH.Exchange(peer_{ecdh\ public})$
5. Secret Combination:
Combine the derived secrets along with proper separation labels for each domain:
$combined\_input = shared\_pq\ ||shared\_classical\ ||"pq"\ ||"classical"$
6. Key Derivation:
Derive the final shared secret using the HKDF with the help of SHA3-256:
$final_{secret} = HKDF - SHA3 - 256(combined_{input}, salt = "hybrid_{pqc_{salt}}")$
7. Return $final_{secret}$

Algorithm 6. Hybrid key exchange

Furthermore, after the hybrid key exchange, Hybrid signing takes place as explained in Algorithm 7, where it signs a message with both Dilithium (PQC) and RSA-PSS (classical) algorithms, and the symbol | | denotes the concatenation of byte strings. Due to the presence of dual signatures, they provide resistance against both quantum and classical attacks at the same time, ensuring long-term verifiability.

Input: Message to sign m
Output: Hybrid signature set $\{dilithium_{signature}, rsa_{signature}\}$
Steps:
1. PQC Signature: Generate a quantum-safe signature using Dilithium and the PQC private key
$sign_{pq} = Dilithium.Sign(privatekey_{pq})$
2. Classical Signature: Generate a classical signature using RSA-PSS and the RSA private key
$sign_{classical} = RSA - PSS.Sign(privatekey_{rsa}, m)$
3. Return the hybrid signature set $\{dilithium_{signature}: sign_{pq}, rsa_{signature}: sign_{classical}\}$

Algorithm 7. Hybrid signing

Once the hybrid signing completes, the hybrid verification takes place as given in Algorithm 8 where the system verifies the input message through both Dilithium and RSA signatures. The hybrid signature becomes valid only if both verifications are successful, to maintain a safe and secure multi-layered trust system. The complete Hybrid digital signature process is illustrated in Fig. 7.

Input:
Message m
Signatures set $\{dilithium_{signature}, rsa_{signature}\}$
Public keys $public_{keys}$
Output:
Verification result V
Steps:
1. Verify PQC Signature using PQC public key:
$valid_{pq} = Dilithium.Verify(publickey_{pq}, message, sig_{dilithium})$
2. Verify Classical (RSA-PSS) Signature using the classical public key:
$valid_{classical} = RSA - PSS.Verify(publickey_{rsa}, message, sig_{rsa})$
3. Compute the overall verification:
$hybrid_{valid} = valid_{pq} AND\ valid_{classical}$
4. Return $V = \{dilithium: valid_{pq}, rsa: valid_{classical}, hybrid_{valid}: hybrid_{valid}\}$

Algorithm 8. Hybrid verification

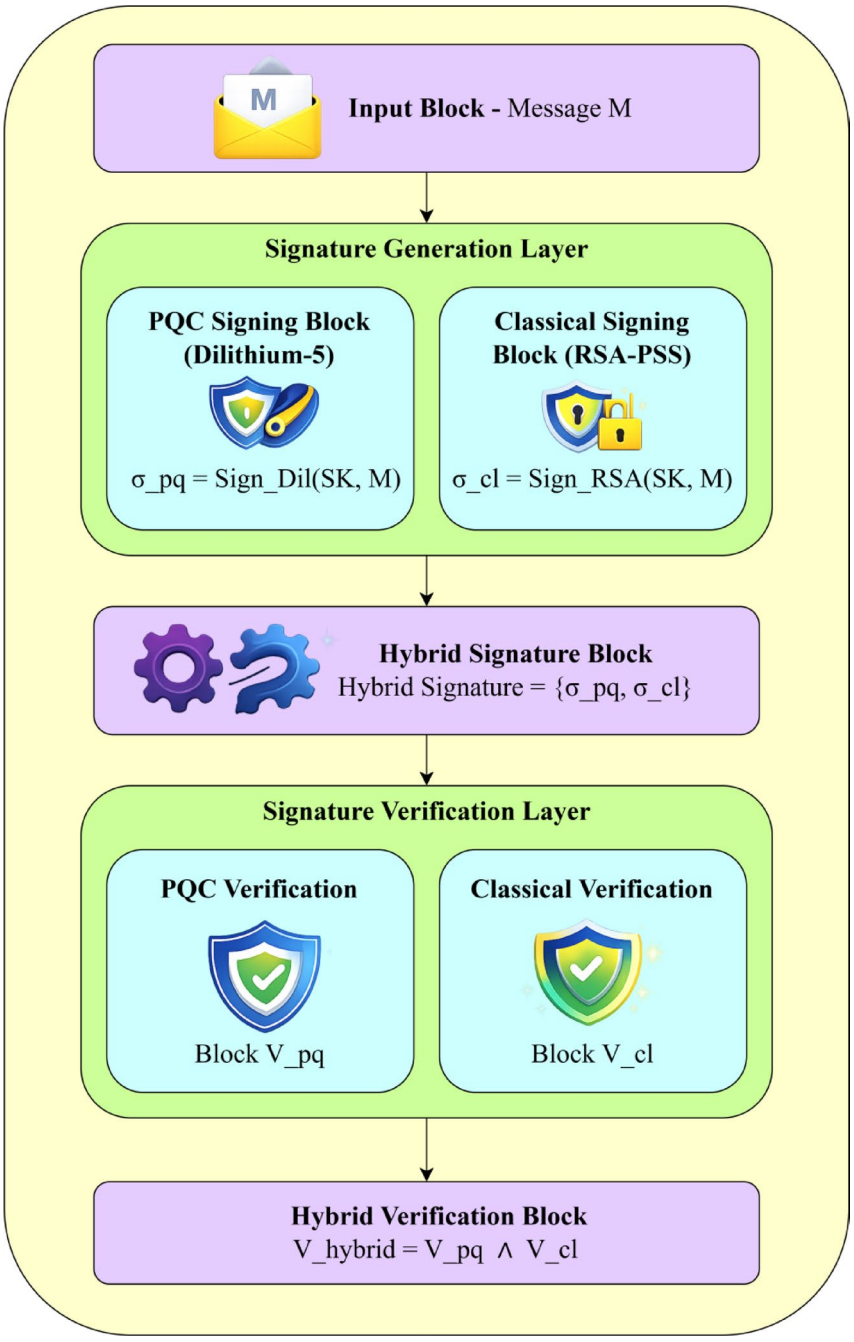

Fig. 7. Hybrid digital signature flow.

Security metadata generation

During each operation for detection, a comprehensive metadata for security is generated. This metadata is used for taking care of the integrity, authenticity, as well as compliance of the detection results. There are two main components in the generated metadata. The first component is the quantum-safe proof, consisting of a digital signature that is generated with the help of Dilithium-5. The digital signature present in the first component is encoded according to Base64, and the first component also contains an ISO 8601-formatted timestamp for marking the exact time when the results are generated. The first component also has the details of the algorithm; it also specifies a 256-bit security level for obtaining post-quantum resilience. This also additionally includes Boolean flags for indicating the NIST PQC compliance as well as forensic-grade applicability in real-world scenarios or cases.

The second part of the metadata that is generated is the Enterprise Security Attributes. These attributes have a flag ensuring the readiness of legal evidence and they also have a unique ID for Chain of Custody. These attributes are present as it is very necessary to support traceability as well as a to cope up with the list of compliance certifications like NIST PQC and FIPS standards. This metadata is then securely packed along with every detection result. This packing is done to ensure quantum-resilience and for enterprise-level regulatory compliance.

The Hybrid Secure Detection of data is done by the use of a hybrid cryptographic technique for protecting the deepfake detection results as shown by Algorithm 9. This algorithm delivers a tamper-proof package of outcomes with metadata. It also gathers forensic evidence by PQC and classical signatures generations in addition to SHA3-256 integrity hashing.

Input: Detection result R, Metadata M
Output: Hybrid secure detection result $RSDR$
Steps:
1. Create and secure session with the help of hybrid key exchange as per Algorithm 6
2. Generate forensic metadata along with the hybrid context: $M \leftarrow Forensic_{Metadata}(R, M)$
3. Create a comprehensive data package: $P = \{R,\ M_{forensic}\}$
4. Then generate a hybrid cryptographic proof by:
5. Generating the PQC proof using Dilithium-5: $\sigma_{pq} = Dilithium.Sign(sk_{pq}, P)$
6. Generating classical signature using RSA-4096: $\sigma_{classical} = RSA - PSS.Sign(sk_{rsa}, P)$
7. Combining them as an $HybridCryptographicProof$: $\Pi_{hybrid} = \{\sigma_{pq}, \sigma_{classical}\}$
5. Compute and calculate the integrity of the hash using $SHA3-256$: $h = SHA3-256(P)$
6. Package the detection results along with the cryptographic proof and the integrity of the hash into HSDR: $HSDR = \{P, \Pi_{hybrid}, H\}$
8. Return $HSDR$

Algorithm 9. Hybrid secure detection

Finally, after securing with the hybrid detection, the system creates a secret combination. This secret combination is created by following the steps as shown in Algorithm 10. Here, the DeepQShield system joins or concatenates several common secrets collected from various algorithms (both PQC and classical). After concatenation, it extracts a single secret key using the HKDF algorithm. These results give a cryptographically secure key data that can be used while performing secure processing in the downstream for getting the deepfake detection system results.

Input: Set of shared secrets from different algorithms $S = \{(A_1, s_1), (A_2, s_2), \ldots, (A_n, s_n)\}$ where A_i and s_i denotes the algorithm identifier and the corresponding shared secret, respectively.
Output: Single combined cryptographic secret $final_{secret} \in \{0,1\}^{256}$
Steps:
1. Initialise an empty byte string: $combined_{input} = \varepsilon$
2. For each $(algorithm\ A_i, secret\ s_i)$ pair:
$combined_{input} += s_i + A_i$
3. Derive the final secret using the HMAC-based Key Derivation Function (HKDF):
$final_{secret} = HKDF_{SHA3-256}(combined_{input}, salt = hybrid_{pqc_{salt}}, info, 32)$
4. Return the derived secret: $final_{secret}$

Algorithm 10. Secret combination

Results

This section thoroughly explains about the experimentation setup that is used along with the implementation details. This is followed by a clear and in-detail explanation as well as analysis of the obtained results of our model DeepQShield. This section also specifies the details about the results, along with discussions and their analysis about it. Additionally, this section also delivers us a comparative study of the DeepQShield with a few pre-existing deepfake detection models to get to know in-depth an overall evaluation of its performance. Also, this section holds details about the cross-dataset verification results as well as ablation studies. Finally, this section also explains how the system provides robustness against adversarial attacks.

Hyperparameter	Values used/experimented
Backbone model	resnext50_32×4d (pretrained)
Image size	224×224
Batch size	32
Total Epochs	50
Learning rate	$1e^{-4}$
Weight decay	$1e^{-4}$
Optimizer	AdamW
Scheduler	CosineAnnealingWarmRestarts ($T_0 = 10,\ \eta_{\min} = 1e^{-6}$)
Dropout rate	0.3
Loss functions	Focal Loss (α = 2, γ = 2), Label Smoothing (ε = 0.1)
Lattice dimension	256
Lattice layers	2
Noise std (LWE)	0.1
Augmentation	Resize + RandomCrop + Flip + Rotate + Brightness/Contrast + Hue/Saturation + Noise/Blur + CLAHE + CoarseDropout

Table 3. Hyperparameter values used for the model training.

Experimental setup

The experiments were performed using the ResNeXt50_32×4d ImageNet-pretrained backbone, accompanied by a lattice-based learning module for improved feature representations. The input images were then resized to 224×224 and further processed with a batch size of 32 and set to train for a total of 50 epochs. After image processing, AdamW optimiser was used with a learning rate $1e^{-4}$, weight decay as $1e^{-4}$ along with CosineAnnealingWarmRestarts scheduler ($T^0 = 10,\ \eta_{\min} 1e^{-6}$). After optimisation, regularisation methods were applied, which involved a dropout rate of 0.3 and lattice noise (std = 0.1) for enhancing the robustness of DeepQShield. Furthermore, two lattice layers of size 256 each were used in the model for perceiving the fine-grained. Additionally, loss functions combined both Focal Loss ($\alpha = 2,\ \gamma = 2$) and Label Smoothing ($\varepsilon = 0.1$) methods to reach stable optimisation. Also, complex data augmentation methods involving resizing, cropping, flipping, rotation, brightness, contrast, hue, saturation, noise, blur, CLAHE and coarse dropout were used to increase generalising ability as well as to mimic variable real-world manipulations and changes.

After training on the DFDC dataset, the top-performing model was saved and then combined with post-quantum cryptographic (PQC) methods. To be particular about the PQC methods, Kyber and Dilithium were used for secure key exchange and for authentication of results to guarantee the integrity as well as confidentiality even during the time of quantum adversarial attacks. The hyperparameters that were used for model training are clearly given in Table 3.

Evaluation metrics

In various experiments, evaluation metrics, like Area under the curve (AUC), Accuracy (Accu), Precision (Pre), Recall (Rec), F1 score (F1) given by Eqs. (23), (24), (25), (26), (27) respectively, are used for performance evaluation:

$$AUC = \sum_{i=1}^{n-1} \frac{TPrate_{(i+1)} + TPrate_i}{2} \left(FPrate_{i+1} - FPrate_i\right) \tag{23}$$

Where $TPrate$ and $FPrate$ represent the true positive and the false positive rates, respectively. The $TPrate = \frac{Tpos}{Tpos+Fneg}$ and $FPrate = \frac{Fpos}{Fpos+Tneg}$. Furthermore, $Tpos, Tneg, Fpos, Fneg$ stands for True positives, True negatives, False positives, and False negatives, respectively.

$$Accu = \frac{Tpos + Tneg}{sumof\left(Tpos, Tneg, Fpos, Fneg\right)} \tag{24}$$

$$Pre = \frac{Tpos}{sumof\left(Tpos, Fpos\right)} \tag{25}$$

$$Rec = \frac{Tpos}{sumof\left(Tpos, Fneg\right)} \tag{26}$$

$$F1 = \frac{2 \times Pre \times Rec}{sumof\left(Pre, Rec\right)} \tag{27}$$

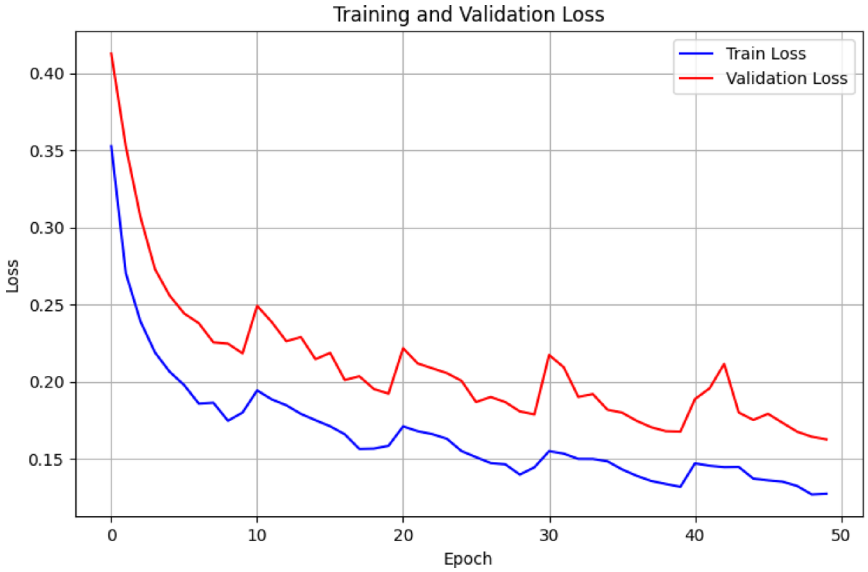

Fig. 8. Graph showing training versus validation loss.

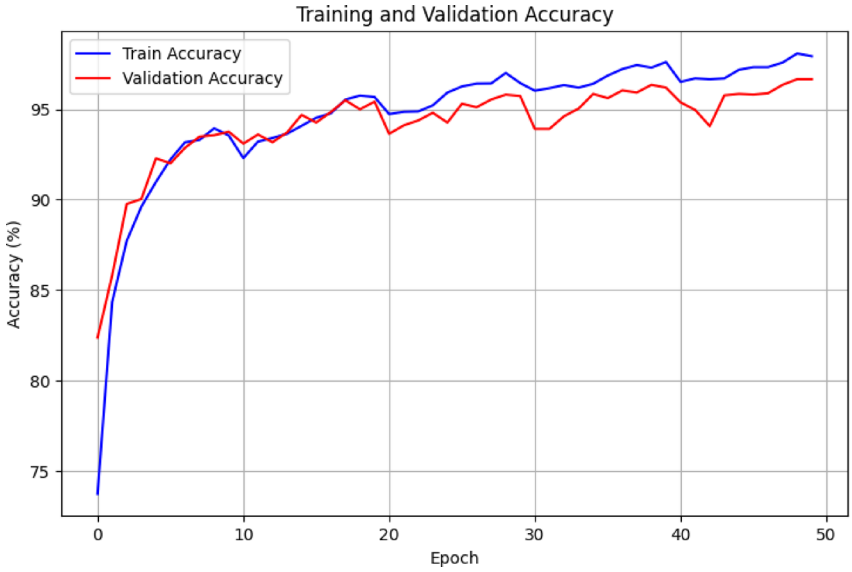

Fig. 9. Graph showing training versus validation loss.

Datasets

The training and testing of the model DeepQShield were mainly done with the help of the Deepfake Detection Challenge (DFDC) dataset[36]. The DFDC dataset was selected as it is a large and well-known dataset for deepfake detection. The dataset has about 140,000 images, which are extracted from video frames. The dataset used is balanced with both real and synthetic images. Also, each frame, which was extracted from pre-processed videos, is given as an image. The obtained images are then cropped and resized to 224×224 pixels to support the ResNeXt-50 backbone.

The DFDC dataset was also used because of its high diversity based on lighting, poses, ethnicity, and backgrounds. This makes this dataset a best-fit for testing and training model's robustness for various real-world scenarios. All frames in the dataset are labelled as real or generated. This ensures the working of supervised learning as well as the right measurement of different classification metrics. To get reproducibility, only frame-level images are used, along with common train-validation splits to test DeepQShield's accuracy, precision, recall, F1-score, and AUC. Further, the dataset of 140,000 images was split into 88,980 images of training, 38,130 images of validation and 12,890 images for testing DeepQSheild. The split is done in such a way that it ensures the balance of both real and fake images.

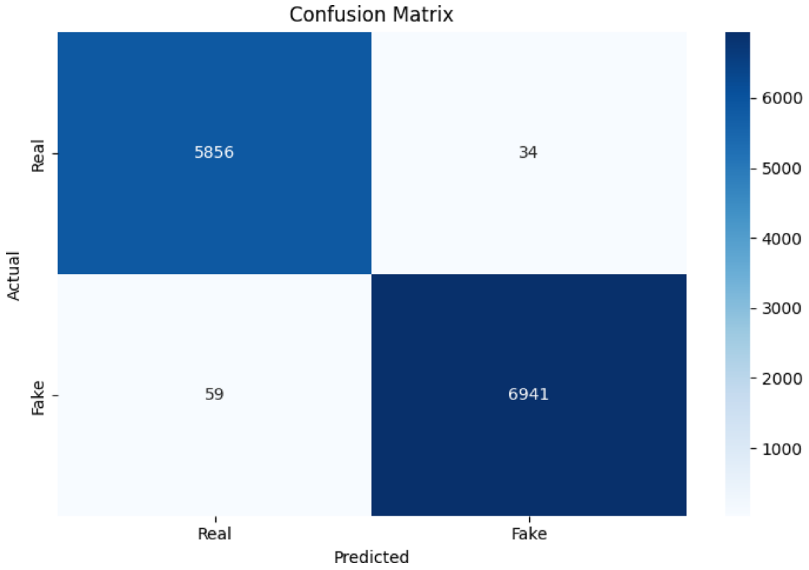

Fig. 10. confusion matrix.

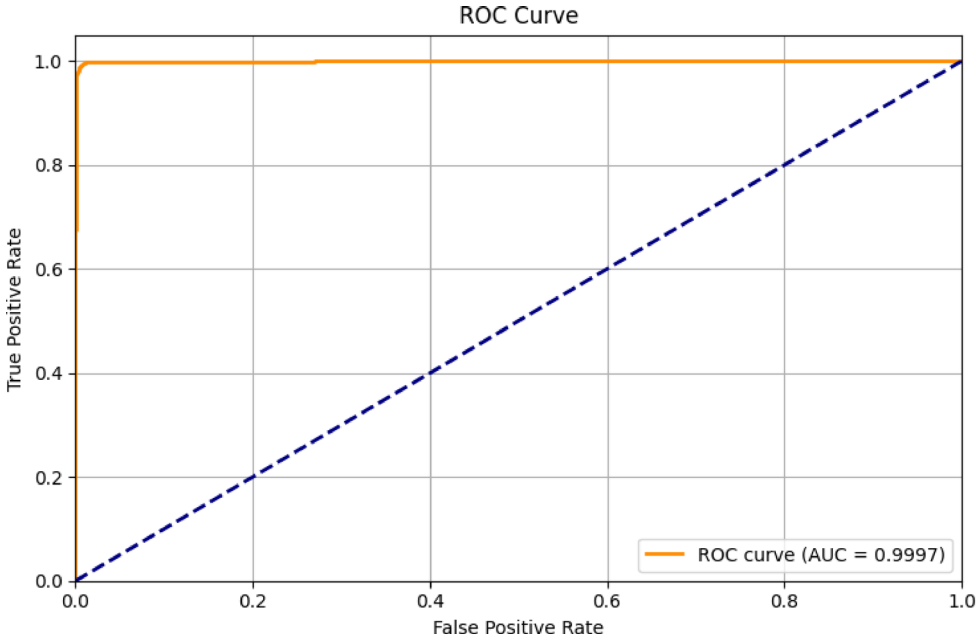

Fig. 11. ROC curve.

Results and analysis

Our proposed framework DeepQShield, combines the ResNeXt50_32 × 4d backbone along with lattice-enhanced learning mechanisms like LWE. Learning with error mechanism-based noise regularisation and the use and integration of post-quantum cryptographic algorithms were evaluated to obtain the results on the DFDC dataset.

DeepQShield has successfully achieved a test accuracy of 99.28% and an AUC score of 0.9997 on the DFDC Dataset. The following Figs. 8, 9, 10 and 11 gives us the visualisations of the 4 different graphs that are the training and validation loss comparison, the training and validation accuracy comparison, the confusion matrix, and the ROC curve.

Figure 8 depicts the training and validation loss curves. It can be seen that these lines, training loss and validation loss, are steadily decreasing and converging, hence suggesting strong generalization.

The accuracy curves in Fig. 9 further prove to us that validation accuracy and training accuracy closely follow each other hand-in-hand and are stabilized consistently above 98%.

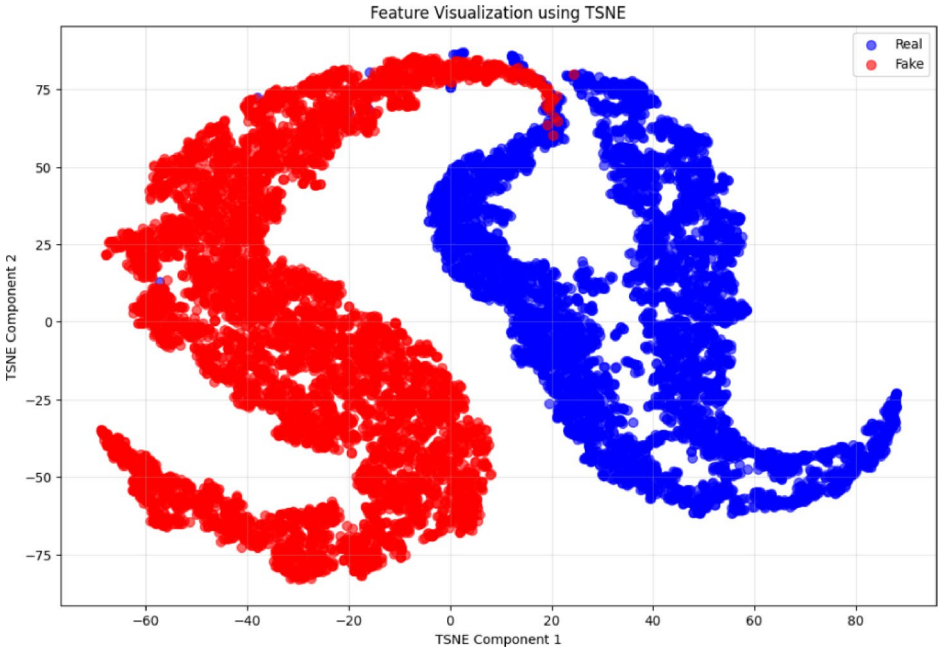

Fig. 12. Graph depicting Feature Visualization using the TSNE.

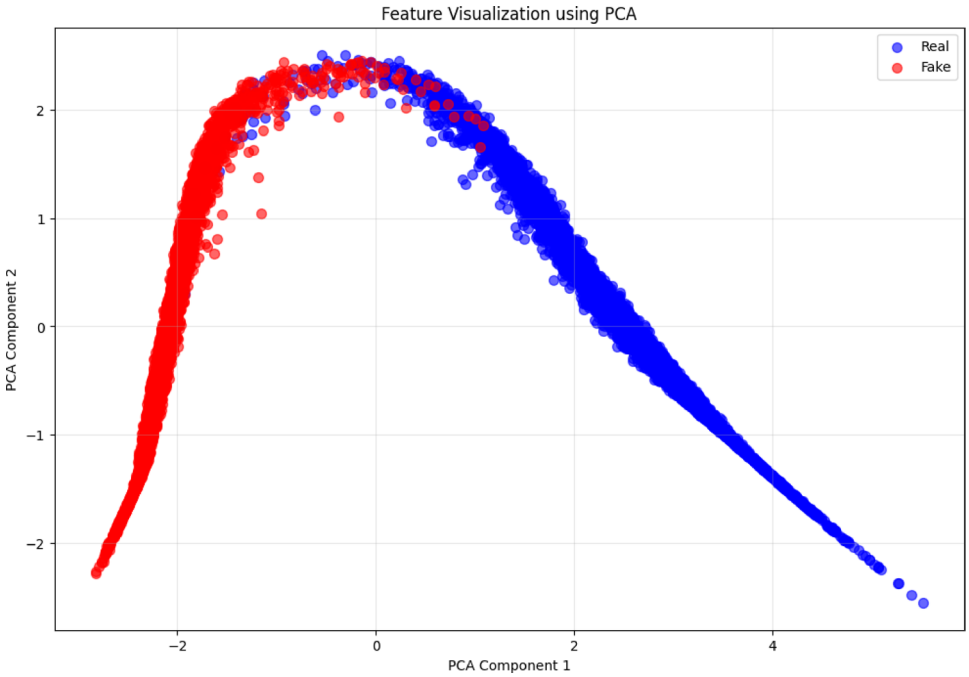

Fig. 13. Graph depicting Feature Visualization using the PCA.

The confusion matrix shown in Fig. 10 highlights the reliability of DeepQShield, with only 34 false positives and 59 false negatives out of 12,890 test samples, proving balanced performance across both classes, real and fake.

The ROC curve in Fig. 11 illustrates the AUC that is about 0.9997, underscoring the model's exceptional discriminative power.

Figure 12 shows us that the t-SNE (t-distributed Stochastic Neighbour Embedding) visualization shows clear clustering of real (blue) and fake (red) samples, confirming that DeepQShield learns highly discriminative embeddings. It also suggests and confirms that there exists an effective separability of features despite of the dataset being very complex and large further gives us the validation that there exists effective feature separability despite the dataset being complex.

References	System	Accuracy	AUC Score
Asha et al.[33]	spatial-temporal feature extraction	87%	0.910
Sudarshana et al.[42]	UAM-Net	98.065%	0.998
Rancy et al.[43]	Vision Transformer	85.41%	0.680
Yogarajan et al.[30]	Multi-layer feature fusion using CNN	87.54%	0.8094
Balafrej et al.[44]	SRNet (YCbCr domain)	81.52%	0.881
Ben Jabra et al.[35]	Ensemble CNN models (InceptionV3, VGG16, Xception)	97.04%	0.997
Nelson et al.[45]	XceptionNet + LSTM	97.57%	Not mentioned
Pintelas et al.[46]	DiffConvNet (diffusion-based CNN)	92.9%	0.962
Zhao et al.[18]	ISTVT (Video Transformer)	92.1%	0.742
Raveena et al.[22]	MSFF-Net (Multi-space feature fusion)	91.5%	0.9671
DeepQShield (Our proposed Model)	ResNeXt50_32 × 4d backbone with lattice-enhanced learning and post-quantum cryptographic security.	99.28%	0.9997

Table 4. Comparison analysis of the results of deepqshield with the existing models for the DFDC dataset.

Figure 13 depicts the PCA (Principal Component Analysis) visualization. This proves that there exists a strong feature separability capability for the DeepQShield. It could be inferred that real and fake image samples form as well-distinguished clusters. It shows the presence of a clear boundary in the reduced feature space and proves the impressive DeepQShield's discriminative power.

Hence, DeepQShield can be titled as a benchmark model for the application in real-life scenarios due to its wonderfully obtained results. Our model DeepQShield has also reached a remarkable average inference speed of almost 15.20 ± 2.17 ms per image (≈ 65.8 images/second). It can also be told that with batch inferencing, the latency has dropped to as low as 3.70 milliseconds per image when the batch size is 32. It is also noticeably clear that GPU usage stats are stable. This marks that the model can obtain a smooth running also on a standard hardware.

Additionally, to obtain superior performance of the DeepQShield, it also incorporates the PQC security mechanisms. These mechanisms include algorithms like the Kyber (for quantum-resistant encapsulated and getting keys) and Dilithium (for generating digital signatures required in deploying the model). These mechanisms are used for securing the best-performing model after training, and they also ensure that the inference outputs are safely sent and attested even during the post-quantum adversarial scenarios. With the combination of the state-of-the-art accuracy and future-proof PQC securities, DeepQShield proves its promising capability for real-world applications in various cases, like multimedia forensics, authentication of content, and for secure media delivery.

Comparative analysis of deepqshield with existing models

There are many existing deepfake detection models with varying strengths and weaknesses. The comparison of the results between the existing models and our model, DeepQShield, on the DFDC dataset is presented in Table 4.

From the study of Asha et al.[33] and Yogarajan et al.[30], they both have used spatial-temporal and multi-layer CNN fusion methods, respectively. Both methods have achieved an accuracy of ~ 87%, but due to the handcrafted feature strategies, they have limited adaptability. Hence, they lack generalisation capacities. Whereas Balafrej et al.[44] relied on SRNet with YCbCr and reached an accuracy of 81.52% and an AUC score of 0.8811. But here, due to the dependence on colour-space reduces the system's reliability. Also, vision transformer-based approaches, such as those by Rancy et al.[43] have gained only 85.41% accuracy and 0.68 AUC score, proving high scalability issues.

A study by Ben Jabra et al.[35] uses high-performing CNN-based approaches using ensemble CNNs reaching accuracy up to 97.04% and 0.997 AUC. This strong performance is maybe due to the combined CNN backbone of InceptionV3, VGG16, and Xception as an ensemble framework. Also, the deepfake detection strategy of Nelson et al.[45] using XceptionNet+LSTM has reached a commendable accuracy of 97.57%, though the AUC score was not mentioned. This commendable accuracy may have been reached due to the temporal dynamics by coupling XceptionNet with LSTM. The framework proposed by Sudarshana et al.[42] was further advanced with UAM-Net and achieved 98.065% accuracy and 0.9982 AUC with help of hybrid attention. But this method lacks robustness during adversarial attacks or during deployment security.

Additionally, the recent works such as the diffusion-based CNN framework (DiffConvNet) explores noise diffusion processes to detect the manipulations. This method achieved an accuracy of 92.9% accuracy and 0.962 AUC, proving its robustness in detecting diverse deepfakes generated using various techniques. The ISTVT, an interpretable spatial-temporal video transformer put forward by Zhao et al.[18], combines spatial and temporal dependencies and has achieved an accuracy of 92.1% and an AUC score of 0.742. This work provides attention-based interpretability due to the use of a transformer-based architecture, but it requires higher computational costs. Also, its resilience is limited during challenging attacks due to the use of transformers. The MSFF-Net, which integrates HOG descriptors along with deep features extracted from ResNet50 and was proposed by Raveena et al.[22]. This framework achieved an accuracy of 91.5% and an AUC score of 0.9671. But due to the dependency on only handcrafted features, the model may have restricted adaptability to the evolving deepfake generation methods.

Reference	Model	DFDC AUC	FF++ AUC	Celeb-DF AUC
Vrizlynn[49]	EfficientNet B7	0.976	0.742	0.971
Hua et al.[50]	Xception + PCC	0.995	0.703	0.802
DeepQShield (Our Model)	ResNeXt50_32 × 4d backbone with lattice-enhanced learning and post-quantum cryptographic security	0.9997	0.853	0.880

Table 5. Comparison analysis of the cross-validation results of deepqshield with the existing models trained using DFDC dataset and tested on FF++ and Celeb-DF.

Model	Accuracy (%)	AUC
Baseline-ResNeXt	82.74	0.9163
DeepQShield (Final)	99.28	0.9997

Table 6. Comparative analysis of ablation study results.

Finally, our proposed model, DeepQShield, overthrows the existing models and frameworks by its accuracy of 99.28% and an AUC score of 0.9997. Our model achieves this high accuracy due to the use of the ResNeXt50_32×4d backbone, enhanced using lattice-based learning and the LWE mechanism. Also, by the integration of PQC security, our model ensures the security and reliability of the detection results. Overall, DeepQShield surpasses the available models in accuracy, scalability, and resilience, making it best suited for real-world application scenarios and deployment.

Cross-dataset evaluation and validation for generalisation

For testing and confirming the generalising capacity of the DeepQshield, many publicly available datasets were used, like FaceForensics++ (FF++) and Celeb-DF(v2) for cross-dataset verification. The FF++ dataset consists of more than 1000 manipulated videos. The videos present in the dataset are synthesised using four distinct types of deepfake creation techniques. In FF++ also the data frames were resized to 224×224 pixels. Furthermore, the images extracted from the FF++ were utilised to test and analyse the model performance under an unseen dataset to get to know about the model's generalising capabilities for achieving high evaluation metrics under highly diverse real-world scenarios. One more dataset, Celeb-DF(v2)[47], was also used for cross-dataset validations. The Celeb-DF dataset includes about 5,639 deepfake or synthetic videos along with 590 original videos, which are sourced from YouTube. The videos mentioned in the Celeb-DF dataset take care of a range of factors like age, ethnicity, gender, etc., to ensure the dataset's diversity. Furthermore, datasets making up extracted images from Celeb-DF and FF++[48] are used for cross-validation and testing of our model, DeepQShield.

Our model DeepQShield is trained using the DFDC dataset and tested using the FF++ reached an AUC score of 0.853, proving its good generalising capability also on the unseen data. When compared to other models, that is the EfficientNet B7 given by Vrizlynn[49] got only 0.742 AUC score and Xception + PCC given by Hua et al.[50] got only a 0.703 AUC score when trained using DFDC and cross-validated on FF++ datasets. Whereas when our model DeepQSheild when trained on DFDC and tested on Celeb-DF, it reached an AUC score of 0.880. But the EfficientNet B7 given by Vrizlynn[49] reported a superior AUC score of 0.971 and Xception + PCC given by Hua et al.[50] has got only 0.801 AUC score when trained using DFDC and cross-validated on Celeb-DF datasets. These results set up the fact that DeepQShield has a strong cross-dataset robustness as well as good detection performance across various benchmark datasets. Also, it can be inferred that the slight decrease in AUC scores when tested on FF++ and Celeb-DF is maybe due to the difference in the distribution of data and also may be due to the difference in image characteristics between two datasets that is affecting the cross-dataset generalisation. Overall, our model proves itself to be good in classifying unseen images and proves its good generalisation ability. The cross-validation results and generalizing capacity of the existing works and ours is shown in Table 5 and is given below.

Ablation studies

To understand the impact of each enhancement, the proposed framework, DeepQShield is compared with a baseline model that uses only the ResNeXt50_32×4d backbone. This initial model is called Baseline-ResNeXt. This comparison highlights how each added mechanism contributes to the final performance and why the full version of DeepQShield achieves superior results.

After training and testing the Baseline-ResNeXt model, it achieved an accuracy of 82.74% and an AUC score of 0.9163 on the DFDC dataset. This depicts the ability of the ResNeXt50_32×4d to various useful features for classification. It could be seen that our Baseline-ResNeXt model falls short when handling the subtle changes or manipulations in the image data. Also, Baseline-ResNeXt has limited capability for finding and detecting many types of forgeries.

So, various novel enhancements are incorporated to get an evolved version of the Baseline-ResNeXt. After adding various upgrades and enhancements, a new and evolved deepfake detection system, DeepQShield is presented. The various upgrades and enhancements involve many techniques like data processing, augmentation, addition of lattice-enhanced learning (LWE) modules, etc. These changes have increased the strength of the DeepQShield's feature extraction capability. The strengthening of the feature extraction ability makes the model

more efficient and precisely detects and differentiates the real and fake images. The introduction of LWE-based noise regularization helped as a technique to prevent overfitting and improve reliability. Finally, the inclusion of PQC algorithms to safeguard the entire detection system ensured security, reliability, and robustness, also during adversarial attacks.

All enhancements combined resulted in a much stronger and quantum-resilient framework. This hybrid framework, DeepQShield, achieved an impressive accuracy of 99.28% and an AUC score of 0.9997on the DFDC dataset. This shows how all enhancements aid in overthrowing the results of the initial Baseline-ResNeXt model. Hence, proving how all the enhancements work hand in hand to make the model reach its best results. The summarized comparison of ablation study results is mentioned in Table 6, given below.

Adversarial robustness evaluation

In this section, we will evaluate adversarial robustness through established white-box attack families, namely the Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), and Projected Gradient Descent (PGD). These attacks establish well-known baselines for the evaluation of robustness, ranging in optimisation strength, with PGD commonly considered among the strongest practical first-order adversaries.

All adversarial examples were generated in a white-box configuration, granting visibility into model gradients and ensuring that robustness estimates remain pessimistic and realistic. In this setting, it was assumed that the model architecture, parameters, loss function, and training procedure were known to the attacker, who could perform an exact gradient-based optimisation of the perturbation.

Attack model and setup

For an input $x \in \mathbb{R}^d$ with label y, an adversarial example is constructed as shown in Eq. 28

$$x^* = x + \delta \tag{28}$$

where the perturbation δ is constrained under the L_∞ norm. Also, where $\| \ \delta \ \|_\infty \leq \ \varepsilon$ and $\varepsilon > 0$ denotes the perturbation budget. This constraint ensures that each pixel (or feature) is modified by at most ε, preserving perceptual similarity while inducing misclassification. To study robustness at multiple distortion levels, several values of ε were systematically explored.

Let $L(\cdot\ )$ denote the loss function and $\nabla_x L\ (x, y)$ its gradient with respect to the input. The attacks differ in how they optimise δ:

FGSM (single-step): Produces a one-shot perturbation in the direction of maximal loss increase for the input and is given by Eq. 29

$$x^* = x + \varepsilon \cdot \ \text{sign}\ (\nabla_x L(x, y)) \tag{29}$$

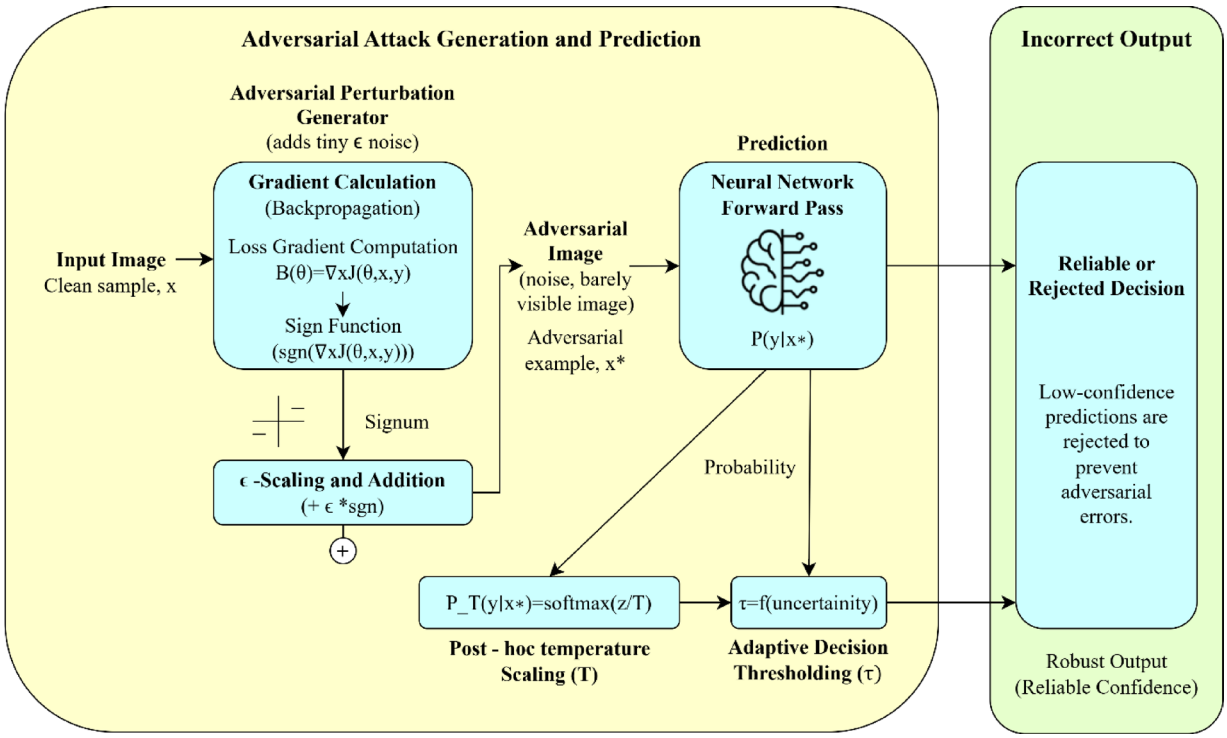

Fig. 14. ML model robustness checking pipeline.

where $x$is the original input to the model, x^* the adversarial example, y denotes the ground-truth label, and ε controls the perturbation magnitude under the L_∞ constraint.

BIM (iterative FGSM): This applies FGSM iteratively with a smaller step size and refines the perturbations over multiple iterations, as shown in Eq. 30

$$x_{t+1} = \Pi_{B_\infty\ (x,\varepsilon)}(x_t + \alpha\ \cdot\ \text{sign}\ \ (\nabla_x L(x_t, y))) \tag{30}$$

where x_t denotes the adversarial sample at iteration t, α is the step size $\Pi_{B_\infty\ (x,\varepsilon)}\ (\cdot)$ projects the perturbed sample back into the L_∞ ball of radius ε center x.

PGD (iterative with random restarts): Initialises from a random point inside the constraint set and repeats the BIM iterations, increasing the probability of reaching a stronger local optimum of the adversarial objective, as given by Eq. 31

$$x_0 = x + \mathcal{U}(-\varepsilon, \varepsilon), x_{t+1} = \Pi_{B_\infty\ (x,\varepsilon)}(x_t + \alpha\ \cdot\ \text{sign}\ \ (\nabla_x L(x_t, y))) \tag{31}$$

where $\mathcal{U}\ (-\varepsilon, \varepsilon)$ denotes uniform random initialisation within the L_∞ ball.

Together, these formulations span a progression from single-step to stronger multi-step optimisation, enabling a rigorous assessment of robustness across attack intensities. We can see the clear adversarial attack generation and prediction pipeline in Fig. 14. This visualises the process of generating adversarial examples and using gradient-based perturbations, which are followed by the prediction of the neural network. Then by post-hoc temperature scaling and adaptive confidence-based decision thresholding for ensuring robust and reliable outputs under various adversarial settings.

Figure 14 depicts how the small, gradient-driven perturbations that are added affect the clean input to create an adversarial sample. Then the obtained predictions are calibrated and filtered out with the help of the uncertainty-aware decision-making thresholds to reduce and mitigate adversarial errors.

Calibration-Aware robustness

Robustness was further explored from a calibration-aware perspective, corresponding to deployment contexts where the system abstains on low-confidence cases. Expected Calibration Error (ECE) was utilised to quantify the discrepancy between predictive confidence and empirical accuracy. The ECE is calculated using Eq. 32 given below:

$$\text{ECE} = \sum_{m=1}^{M} \frac{\mid B_m \mid}{N} \mid\ \text{acc}\,(B_m) - \text{conf}\,(B_m)\ \mid \tag{32}$$

where B_m marks the set of samples in the m-th confidence bin, $\mid\ B_m\ \mid$ denotes the number of samples in the bin, N denotes the total number of samples, $\text{acc}\,(B_m)$ represents the empirical accuracy, and $\text{conf}\,(B_m)$ denotes the average predicted confidence within the bin.

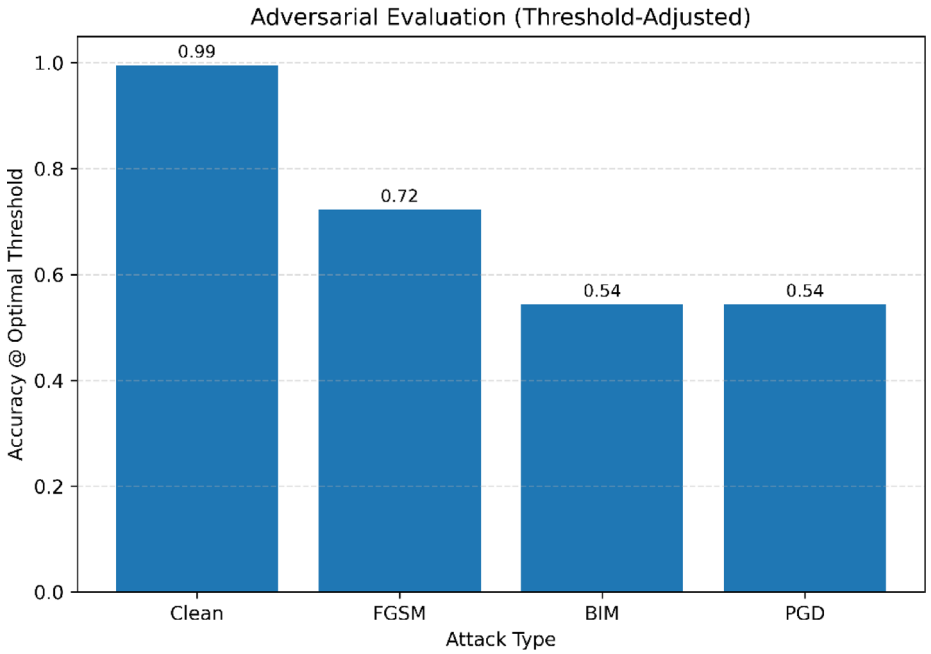

Fig. 15. Adversarial accuracy after threshold optimisation.

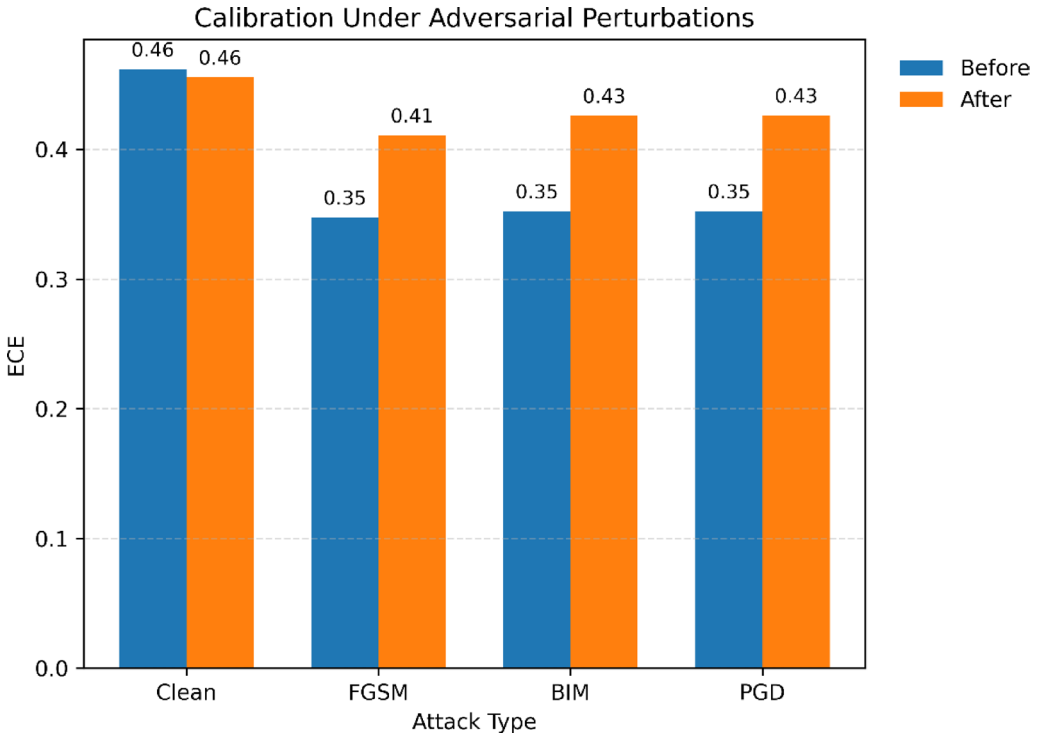

Fig. 16. ECE across attack scenarios.

Attack	ECE (Before)	ECE (After)	Optimal τ	Accuracy @ τ
Clean	0.4616	0.4555	0.39	0.9949
FGSM	0.3473	0.4107	0.12	0.7223
BIM	0.3522	0.4258	0.05	0.5431
PGD	0.3521	0.4260	0.05	0.5431

Table 7. Calibration-aware adversarial evaluation.

Furthermore, temperature scaling was then used to optimise a single parameter T that minimises ECE on validation data. This process does not change class predictions but adjusts confidence estimates. Moreover, dynamic cutoff thresholds were used to suppress high-risk outputs. Put together, temperature scaling and thresholding form a constitute a pragmatic approach for uncertainty management and robustness enhancement without model retraining.

Results: calibration and decision stability

From Fig. 15, we can visualise the adversarial accuracy of various attacks after the optimisation. We can see how threshold adjustment improves performance to roughly 0.54 even in the case of strong PGD and BIM attacks. Compared to substantially lower accuracy under default settings, this behaviour points out that adversarial degradation is mainly due to shifts in confidence rather than a complete loss of discriminative capability.

Figure 16 depicts how the calibration under adversarial perturbations changes before and after the attacks. Here, temperature scaling on clean data drops ECE from 0.4616 to 0.4555, which aligns with effective calibration behaviour. Under adversarial conditions, ECE increases, such as FGSM → 0.4107 and PGD → 0.4260, reflecting the intentional disruption of confidence calibration due to adversarial perturbations. These are summarised in Table 7.

Without adversarial noise, calibration does not degrade, and no loss in accuracy. Under FGSM, optimized decision thresholds maintain performance at approximately 0.72, indicating resilience against single-step perturbations. Stronger multi-step attacks like BIM and PGD result in greater misalignment of confidence ECE $\approx$ 0.42. However, adaptive thresholding restores accuracy to about 0.54, which is a significant enhancement compared to the default decision settings. The findings point to that adversarial attacks affect the distribution over confidence far more than they do the latent discriminative signal, which remains recoverable by and large.

Practical deployment considerations

White-box PGD assumes full gradient visibility, unrestricted querying, and iterative optimization-all rarely occurring together in a deployed system. In practice, conservative rejection thresholds and abstention rules

add an extra layer of safety, rejecting suspicious inputs while maintaining reliability on benign ones. In this case, robustness follows from calibrated uncertainty management rather than reliance on brittle overconfident predictions.

Overall, the results have shown that the model is stable under weak perturbations and has controlled confidence distortion under stronger adversarial pressure. Crucially, calibration-aware decision strategies restore significant reliability without extra overhead to retraining or architectural modification. These observations support the claim that the proposed approach offers practical, risk-aware adversarial resistance, even when evaluated under rigorous white-box conditions.

Computational cost, real-time deployability, and privacy considerations

By analyzing the evaluation of metrics, we can see that the DeepQShield outperforms the existing approaches in detection, accuracy, robustness, and deployment feasibility rather than just a single goal. When compared to feature-driven and frequency domain-based approaches that solely rely on the handcrafted or reconstructed cues[1,16], our proposed lattice-enhanced ResNeXt backbone learns more discriminative and stabilised feature representations. Due to this, we have achieved high performance across large-scale benchmark datasets for deepfake detection. Also, Transformer-based architectures such as ISTVT[18] and multi-space feature fusion models like MSFF-Net[22] depict strong results, but these produce high computational complexity, memory overhead and also lack robustness against adversarial attacks. Due these limitations, these models become less suitable for latency-sensitive deployment scenarios.

From the perspective of optimised computation, lightweight models like the MaD-CoRN[6] are optimised to achieve minimal inference latency and low resource usage. This makes it possible for it to reach faster execution via architectural simplification. However, these light-weight models can experience performance degradation over time under the complex manipulations and cross-dataset scenarios[19,44]. In contrast to the above, DeepQShield incorporates moderate additional computational overhead for lattice-based regularisation, but meanwhile it also maintains an inference time of just 15–25 ms per image on a GPU. Due to this inference time, it enables our model to work in real-time operation. This latency also remains quite lower when compared to transformer-heavy approaches like[18] and aligns with the real-world Deployability conditions as listed in empirical and judicial deepfake assessment studies[7,8]. Also, our model takes an average inference latency of 60 ms (~ 16 FPS) per image on CPU, which still takes less time than heavy-transformer-based solutions. This amount of time indicates it can be easily scalable across varying settings where GPUs may not be available.

Primarily, the use of post-quantum cryptographic mechanisms introduces only a negligible overhead on runtime. Also, experimental evaluation indicates that Kyber-1024 key generation, encapsulation require 1ms each. Additionally, Dilithium-5 signature generation and verification takes only 2 ms and 1 ms, respectively. Even hybrid cryptographic verification takes under 5 ms. This not only reassures the fastness but also ensures that security enhancements do not compromise with the real-time responsiveness[27,40,41]. In real-time detection settings, these characteristics help DeepQShield to sustain and maintain a stable throughput, making it suitable for various applications such as online content moderation, forensics, and judicial evidence verification and so on. Overall, by quantitatively managing the accuracy, inference speeds, cryptographic overheads and also preservation of the system's privacy is the main goal. In this way, our deepfake detection framework is a practically deployable and future resilient framework rather than being purely an accuracy or speed-optimised solution.

Conclusion

This research work introduces DeepQShield, a novel deepfake image detection framework that combines the ResNeXt50_32×4d backbone along with the lattice-based enhanced learning techniques and PQC algorithms. This framework has achieved high security, reliability and resilience against various adversarial attacks like quantum attacks due to the integration of post-quantum cryptographic (PQC) safeguards as well as due to the use of learning with errors (LWE) mechanism during model training. By the introduction of LWE-based noise regularisation, the model has reached improved generalisability and resistance to adversarial manipulations of data. The usage of Kyber and Dilithium PQC schemes ensures the system resilience against quantum attacks via encryption and authentication of detection results. The various experimental findings on the DFDC dataset demonstrate the impressive performance of DeepQShield. Our model DeepQSheild has reached a remarkable accuracy of 99.28% and an AUC score of 0.9997, surpassing various existing deepfake detection systems like EfficientNet-B7, UAM-Net, Vision transformers, etc. Additionally, cross-dataset validation on FaceForensics++and Celeb-DF(v2) proves its significant generalisation capabilities even under diverse real-world settings. Apart from the accuracy of DeepQShield's framework, it is also practically deployable and scalable due to its inference speed of ~ 65 images per second. Overall, it is concluded that our model DeepQShield has set a new benchmark in the field of deepfake detection due to its hybrid and novel approach. DeepQShield shines out from the rest due to its exceptional accuracy, scalability, reliability, security safeguards, its robustness as well as due to its quantum resilience. This also certifies itself to be best-suitable for application scope in several environments and settings, such as for use in the digital forensics, for authenticating and verification of data in social media, etc. Further in future, we plan to optimise the DeepQShield's real-time deployment ability. Additionally, we also plan to make changes in future to the model to cope up with the newer, upcoming and evolving deepfake generation techniques.

Data availability

The code developed during the current study can be accessed from: https://github.com/sakethksg/DeepQShield. The data that support the findings of this study are available from: DFDC: https://www.kaggle.com/datasets/xhl

ulu/140k-real-and-fake-faces. Celeb-DF(V2): https://www.kaggle.com/datasets/reubensuju/celeb-df-v2. FF+: https://www.kaggle.com/datasets/greatgamedota/faceforensics.

Received: 28 October 2025; Accepted: 1 February 2026
Published online: 20 February 2026

References

1. Jin, X. et al. A dual descriptor combined with frequency domain reconstruction learning for face forgery detection in deepfake videos. *Forensic Sci. Int. Digit. Invest.* **49**, 301747. https://doi.org/10.1016/j.fsidi.2024.301747 (2024).
2. Casu, M. et al. GenAI mirage: The impostor bias and the deepfake detection challenge in the era of artificial illusions. *Forensic Sci. Int. Digit. Invest.* **50**, 301795. https://doi.org/10.1016/j.fsidi.2024.301795 (2024).
3. Tran, V. N. et al. Generalization of forgery detection with meta deepfake detection model. *IEEE Access.* **11**, 535–546. https://doi.org/10.1109/ACCESS.2022.3232290 (2023).
4. Maheshwari, R. U. et al. Advanced plasmonic Resonance-enhanced biosensor for comprehensive real-time detection and analysis of deepfake content. *Plasmonics* **20**, 1859–1876. https://doi.org/10.1007/s11468-024-02407-0 (2025).
5. Kumar, A. et al. Advances in deepfake detection algorithms: Exploring fusion techniques in single and multi-modal approach. *Inform. Fusion*. **118**, 102993. https://doi.org/10.1016/j.inffus.2024.102993 (2025).
6. Budhiraja, R. et al. MaD–CoRN: An efficient and lightweight deepfake detection approach using convolutional reservoir network. *Multimedia Tools Appl.* https://doi.org/10.1007/s11042-024-20223-w (2024).
7. Hydara, E., et al. Empirical assessment of deepfake detection: Advancing judicial evidence verification through artificial intelligence. *IEEE Access.* **12**, 151188–151203. https://doi.org/10.1109/ACCESS.2024.3480320 (2024).
8. Sandoval, M. P. et al. Threat of deepfakes to the criminal justice system: A systematic review. *Crime. Sci.* **13**, Article 41. https://doi.org/10.1186/s40163-024-00239-1 (2024).
9. Brown, S. D. Virtual unreality: Potential implications of deepfake technology for the course of justice. *ERA Forum*. **24**, 501–518. https://doi.org/10.1007/s12027-024-00780-1 (2023).
10. Maheshwari, R. U. et al. Innovative quantum PlasmoVision-based imaging for real-time deepfake detection. *Plasmonics* https://doi.org/10.1007/s11468-025-02846-3 (2025).
11. Lin, C. A. et al. Quantum–trained convolutional neural network for deepfake audio detection [Preprint]. https://doi.org/10.48550/arXiv.2410.09250 (2024).
12. Sharma, P. et al. A robust ensemble model for deepfake detection of GAN-generated images on social media. *Discov Comput.* **28**, 41. https://doi.org/10.1007/s10791-025-09538-w (2025).
13. Sunil, R. et al. Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation. *Heliyon* **11** (2), e42273. https://doi.org/10.1016/j.heliyon.2024.e42273 (2025).
14. Abbas, F. & Taeihagh, A. Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Syst. Appl.* **252**, 124260. https://doi.org/10.1016/j.eswa.2024.124260 (2024).
15. Tanwar, S. & Singh, J. ResNext50 based convolution neural network-long short term memory model for plant disease classification. *Multimedia Tools Appl.* **82**, 1–19. https://doi.org/10.1007/s11042-023-14851-x (2023).
16. Li, B. et al. FDPNet Deep forgery detection by leveraging multi-scale self-forgery images generating. *J. Supercomput.*. **81**, 778. https://doi.org/10.1007/s11227-025-07229-3 (2025).
17. Qiu, X. et al. Fusion: Dual-domain fusion with feature superposition for deepfake detection. *Inform. Fus.*. **119**, 103087. https://doi.org/10.1016/j.inffus.2025.103087 (2025).
18. Zhao, C. et al. ISTVT: Interpretable spatial-temporal video transformer for deepfake detection. *IEEE Trans. Inf. Forensics Secur.* **18**, 1335–1348. https://doi.org/10.1109/TIFS.2023.3239223 (2023).
19. Kingra, S., et al. Assessing deepfake detection methods: A comparative evaluation on novel large-scale Asian deepfake dataset. *Int. J. Data Sci. Anal.* https://doi.org/10.1007/s41060-025-00741-y (2025).
20. Hu, J. et al. Delocate: Detection and localization for deepfake videos with randomly-located tampered traces. https://arxiv.org/abs/2401.13516 (2024).
21. von Schenk, A. et al. Lie detection algorithms disrupt the social dynamics of accusation behavior. *Article* **27**, 7110201. https://doi.org/10.1016/j.isci.2024.107239 (2024).
22. Raveena, et al. MSFF-Net: A deepfake detection network based on multi-space feature fusion technique. *J. Inf. Syst. Eng. Manag.* **10**(55s), 131. https://www.jisem-journal.com/index.php/journal/article/view/11433 (2025).
23. Qi, X. et al. MADD: A multi–lingual multi–speaker audio deepfake detection dataset. In *Proceedings of the 2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. https://doi.org/10.1109/ISCSLP63861.2024.10800535 (2024).
24. Tahaoglu, G. et al. Deepfake audio detection with spectral features and ResNeXt-based architecture. *Knowl.-Based Syst.* **293**, 113726. https://doi.org/10.1016/j.knosys.2025.113726 (2025).
25. Samhita, V. et al. Self-distillation framework for improving fake speech detection in the domain variability scenario. *Neural Comput. Appl.* **37** (5), 3111–3127. https://doi.org/10.1007/s00521-024-10760-8 (2024).
26. Gandhi, K. et al. A multimodal framework for deepfake detection [Preprint]. Journal of Electrical Systems, 2024. Retrieved from https://arxiv.org/abs/2410.03487 (2024).
27. Chawla, D. & Mehra, P. S. A roadmap from classical cryptography to post-quantum resistant cryptography for 5G-enabled iot: Challenges, opportunities and solutions. *Internet of Things* **24**, 100950 (2023). https://doi.org/10.1016/j.iot.2023.100950
28. Tian, J. et al. ADMM-based adversarial false data injection attacks against multi-label locational detection. *IEEE Trans. Dependable Secur. Comput.*. https://doi.org/10.1109/TDSC.2025.3605689 (2025).
29. Tian, J. et al. Adversarial attacks and defenses for deep-learning-based unmanned aerial vehicles. *IEEE Internet Things J.* **9**(22), 22399–22409. https://doi.org/10.1109/JIOT.2021.3111024 (2022).
30. Yogarajan, G. et al. Robust deepfake detection using multi-scale feature fusion. *Multimed Tools Appl.* https://doi.org/10.1007/s11042-025-20768-4 (2025).
31. Xu, Z. et al. Public perception towards deepfakes through topic modelling and sentiment analysis of social media data. *Soc. Netw. Anal. Min.* **15**, Article 16. https://doi.org/10.1007/s13278-025-01445-8 (2025).
32. Chakraborty, R. & Naskar, R. Role of human physiology and facial biomechanics towards Building robust deepfake detectors: A comprehensive survey and analysis. *Comput. Sci. Rev.* **54**, 100677. https://doi.org/10.1016/j.cosrev.2024.100677 (2024).
33. Asha, S. et al. A defensive framework for deepfake detection under adversarial settings using temporal and spatial features. *Int. J. Inf. Secur.* **22**, 1371–1382. https://doi.org/10.1007/s10207-023-00695-x (2023).
34. Subburaj, B. & Ragavendra, R. Deepfake detection using spatio-temporal-structural anomaly learning and fuzzy system-based decision fusion. *IEEE Access*, **13**, 82747–82758 (2025). https://doi.org/10.1109/ACCESS.2025.3567523
35. Ben Jabra, M. et al. Deepfake detection through ensemble learning. *Multimedia Tools Appl.* https://doi.org/10.1007/s11042-025-20932-w (2025).
36. xhlulu. 140k Real and Fake Faces [Data set]. Kaggle. (2019). https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces

37. Regev, O. On lattices, learning with errors, random linear codes, and cryptography. *J. ACM*. **56** (6). https://doi.org/10.1145/1568318.1568324 (2009). Article 34.
38. Lin, T. Y. et al. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)* 2999–3007. IEEE (2017). https://doi.org/10.1109/ICCV.2017.324
39. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)* (2017). https://arxiv.org/abs/1711.05101
40. Bos, J. et al. CRYSTALS-Kyber: A CCA-secure module-lattice-based KEM. In K. Paterson & J. Möller (Eds.), *Advances in Cryptology – EUROCRYPT 2018* 353–383. Springer (2021). https://doi.org/10.1007/978-3-319-78381-9_13
41. Ducas, L. et al. CRYSTALS-Dilithium: A lattice-based digital signature scheme. In J.-S. Coron & J. B. Nielsen (Eds.), *Advances in Cryptology – EUROCRYPT 2018* (pp. 238–268). Springer. (2018). https://doi.org/10.1007/978-3-319-78375-8_9
42. Sudarshana, K. & Vamsidhar, Y. UAM-Net: robust deepfake detection through hybrid attention into scalable convolutional network. *Expert Syst.* **42** (3), e70009. https://doi.org/10.1111/exsy.70009 (2025).
43. Chepchirchir, R. & Mboli, J. Exploring deepfake detection: A comparative analysis. *TechRxiv*. (2025). https://doi.org/10.36227/techrxiv.174918233.32766378/v1
44. Balafrej, I. & Dahmane, M. Enhancing practicality and efficiency of deepfake detection. *Sci. Rep.* **14**, 31084. https://doi.org/10.1038/s41598-024-82223-y (2024).
45. Nelson, L. et al. Deepfake detection in manipulated images/audio/videos: A three-stage multi-modal deep learning framework. *Inteligencia Artif.* **28** (76), 20–39. https://doi.org/10.4114/intartif.vol28iss76pp20-39 (2025).
46. Pintelas, E. & Livieris, I. E. Convolutional neural network framework for deepfake detection: A diffusion-based approach. *Comput. Vis. Image Underst.* **257**, 104375. https://doi.org/10.1016/j.cviu.2025.104375 (2025).
47. Reubensuju. (n.d.). Celeb-DF (v2) [Data set]. Kaggle. Retrieved September 25. from (2025). https://www.kaggle.com/datasets/reubensuju/celeb-df-v2
48. GreatGameDota. (n.d.). *FaceForensics* [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/greatgamedota/faceforensics
49. Thing, V. L. L. Deepfake detection with deep learning: Convolutional neural networks versus transformers. https://arxiv.org/abs/2304.03698(2023).
50. Hua, Y. et al. Learning patch-channel correspondence for interpretable face forgery detection. *IEEE Trans. Image Process.* **32**, 1668–1680. https://doi.org/10.1109/TIP.2023.3246793 (2023).

Author contributions

Conceptualization: Brindha Subburaj and Kollipara Naga Shreeya; methodology: Brindha Subburaj, Kollipara Naga Shreeya and Kollipara Sai Govinda Saketh; writing— original draft preparation: Brindha Subburaj, Kollipara Naga Shreeya and Kollipara Sai Govinda Saketh, Padmavathy T V, Sherly Alphonse, Girish Subramanian. All authors have reviewed the manuscript.

Funding

Open access funding provided by Vellore Institute of Technology.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.