Interdisciplinary Journal of Computing & AI

IJCAI
Interdisciplinary Journal of Computing & AI

IJCAI

RESEARCH

VerityNet: A Hybrid CNN Framework for Deepfake Detection

Yash Rawal[1*], Shivam Sati[2], Meet Shah[3], Swapnil Bhagat[4]

[1,2,3,4]*Dept. of Computer Science Engineering, Thakur college of Engineering and Technology, Mumbai, India;*
Email: [1]yrawal1000@gmail.com, [2]shivamsati391@gmail.com, [3]meetshah1785@gmail.com, [4]swapnilbhagat21@gmail.com

ARTICLE INFO.

Keywords: Deepfake Detection, Ensemble Learning, EfficientNetB0, VGG16, Convolutional Neural Network, Majority Voting, Transfer Learning, Image Forensics
*Corresponding author email address: yrawal1000@gmail.com
Manuscript received: 22 April 2026/
Accepted: 16 June 2026
Published online: 2 July 2026
https://doi.org/10.5281/zenodo.21056022

Cite as: Y. Rawal et al., "VerityNet: A Hybrid CNN Framework for Deepfake Detection", Interdisciplinary Journal of Computing & AI (IJCAI), vol. 1, no. 2. Aspiration Publishing Trust (DARPAN ID: WB/2025/0887371), p. 212-235, Jul. 2026. doi: 10.5281/zenodo.21056022.

ABSTRACT

Deepfakes have quickly advanced to a point where fabricated visuals and voices are often indistinguishable from reality, enabling misinformation, financial scams, and reputation damage. To counter these risks, this work introduces VerityNet, a deepfake detection framework that merges the strengths of EfficientNetB0, VGG16, and a lightweight custom CNN. Instead of depending on a single feature extractor, the system applies majority voting across three independently optimized models to improve reliability against varied manipulation styles. Experiments on a large, heterogeneous dataset demonstrate a 94% accuracy rate, while preserving operational efficiency suitable for deployment scenarios. The study also evaluates model interpretability using Grad-CAM and discusses future improvements including multimodal fusion and transformer-based variants for enhanced resilience.

1 Introduction

The increasing accessibility of generative AI tools has led to a rapid rise in deepfakes, making them nearly indistinguishable from genuine media in numerous scenarios [1]. Deepfake applications are rapidly transitioning from entertainment into cybercrime and disinformation campaigns [2]. Industry assessments indicate that deepfake usage continues to escalate sharply across the internet ecosystem [3], [4]. Synthetic videos can manipulate political speech [5], harm reputations, or assist social engineering attacks [6], [7].

A highly publicized 2024 incident involved corporate impersonation where a major engineering firm suffered a $25.5 million financial loss after an employee was deceived through a deepfake-based video call [8], [9]. This demonstrates tangible economic danger and exposes defence gaps in corporate verification workflows. Projected financial losses soon could exceed $40 billion if such attacks intensify [10].

Existing forensic tools cannot reliably detect modern AI-generated content, particularly under compression noise and adversarial perturbation [11]. This research introduces VerityNet, designed to offer scalability, interpretability, and strong generalization performance through ensemble learning and explainable visualization.

Major contributions:

- A new ensemble leveraging three diverse CNN architectures.
- Superior generalization achieved through majority voting
- Interpretability applied using Grad-CAM.
- Operational efficiency suitable for deployment in real environments.

2 Background & Motivation

Standalone deepfake detectors frequently degrade when exposed to unfamiliar manipulation types [12], [13]. Diffusion-based generation introduces subtler and more realistic artifacts compared to GAN-based methods, making past detectors insufficient [14]. Furthermore, compressed social media content often disrupts critical patterns needed for classification [15], [16].

Investigations reveal multiple limitations in traditional approaches:

- Inadequate performance transfer across datasets [11].
- High inference complexity in transformer-based approaches.
- Susceptibility to adversarial poisoned or perturbed input data.
- Lack of transparency in decision-making processes [17], [18], [19].
- Bias presence across demographic groups.

VerityNet addresses these pain points through architectural diversity and visual evidence justification for outputs.

3 Comprehensive Literature Review

A comprehensive review of related work was undertaken to position VerityNet within the broader landscape of deepfake detection research. The field has rapidly evolved from simple artifact detection to a complex, multi-faceted discipline. This review of recent (2024-2025) literature identifies the key trends and research gaps that motivate the ensemble-based approach of VerityNet (Table 1).

3.1 The Generalization Gap and The Diffusion Crisis

The single greatest challenge in deepfake detection is generalization [11]. The "generalization gap" is no longer a theoretical concern but a measured failure. The landmark 'Deepfake-Eval-2024' benchmark provided definitive proof, revealing a catastrophic 45-50% performance drop (AUC) in state-of-the-art academic detectors when confronted with "in-the-wild" media [12].

This failure is largely attributed to a technological shift. Most detectors were trained on older datasets featuring forgeries from Generative Adversarial Networks (GANs). However, the 2024-2025 threat landscape is dominated by diffusion models, which leave behind different, more subtle forensic traces [14]. Detectors trained to spot GAN artifacts are effectively blind to these new forgery signatures.

3.2 Foundational Architectures and Ensemble Strategies

- This performance-cost trade-off validates the approach of VerityNet, which leverages an ensemble of efficient CNNs. Recent 2025 research strongly supports this methodology, demonstrating that ensembles significantly improve robustness and generalization across unseen datasets by leveraging the complementary strengths of different architectures [20], [21], [22].

- Historically, Convolutional Neural Networks (CNNs) have been the dominant choice. Architectures like VGG16, used in this study, have proven to be robust feature extractors [11]. However, the state-of-the-art has seen a shift toward Vision Transformers (ViTs), which can excel at capturing global context [23]. This, however, comes at a significant computational cost, with research highlighting that ViT-heavy models suffer from increased latency, making them less suitable for real-time deployment [24].

3.3 Robustness to Real-World Perturbations

A model's accuracy on a clean test set is a poor predictor of its real-world performance. Two primary threats degrade accuracy:

- Data Compression: Media in the wild is aggressively compressed by social media platforms. This post-processing can erase the high-frequency artifacts that many detectors rely on, or introduce new artifacts (e.g., blockiness) that confuse the model [15].
- Adversarial Attacks: The security of detectors is actively being challenged. The AADD-2025 (Adversarial Attacks on Deepfake Detectors) Grand Challenge formalizes this threat, encouraging research into attacks that can fool detectors [25]. More advanced backdoor attacks poison the training data itself. Seminal papers like "Poisoned Forgery Face" (ICLR 2024) [26] and research on "natural triggers" [27] show how detectors can be compromised to misclassify fakes when a specific, invisible trigger is present.

3.4 Explainability in Media Forensics (XAI)

In high-stakes domains like law and finance, a simple "real/fake" label is insufficient. These fields demand evidence and transparency, fueling a push for Explainable AI (XAI) [28], [29]. VerityNet's use of Grad-CAM [17] provides a foundational layer of interpretability by creating heatmaps that visualize where a model is looking to make its decision. The state-of-the-art is moving toward more sophisticated narrative-based XAI, such as the "TruthLens" [30] and "DF-P2E" [31] frameworks, which integrate language models to provide human-readable explanations.

3.5 Summary and Future-Scope Solutions

The literature reveals a field at a critical juncture. In response to the generalization gap [11] and new threat vectors [7], the community is pivoting to key areas. These include advanced generalization techniques like open-world adaptation [32] and new augmentation strategies [33], as well as using graph neural networks (GNNs) [34].

Critically, this review identifies major challenges that map directly to future research directions:

- Dataset Bias: Detectors trained on unbalanced public datasets prove to be demographically biased, performing unfairly across race, gender, and age [35]. New algorithms are being developed to mitigate this [36].
- Multimodal Threats: The Arup fraud 7 was an audio-visual attack. Image-only models like VerityNet are insufficient. Future systems must incorporate multimodal analysis to check for audio-visual synchronization [37].
- Efficiency: As models grow, deploying updates is costly. Parameter-Efficient Fine-Tuning (PEFT) techniques like Low-Rank Adaptation

(LoRA) offer a path to rapidly and cheaply update models to counter new threats [38].

- Privacy: Training models on real-world data is a privacy risk. Federated Learning offers a solution by allowing collaborative training on decentralized data without it ever being shared [39], [40].

4 Model Architecture

The complete functional flow of VerityNet is represented in Figure 1, outlining the interaction between all major components of the proposed deepfake detection system. When a face image is provided as input, it is simultaneously forwarded to three independently trained convolutional neural networks: EfficientNetB0 (fine-tuned), VGG16 (fine-tuned), and a lightweight Custom CNN. Each model performs its own binary evaluation and outputs both the predicted class label (real or fake) and an associated confidence probability.

The decisions produced by the three classifiers are then aggregated within the Majority Voting Unit, which finalizes the authenticity decision by selecting the label supported by at least two models. In addition to selecting the dominant prediction, this unit retains the individual model confidence values to support reliability assessment when results are close.

Concurrently, intermediate activation feature maps are passed to the Explainable AI (XAI) Module, where Grad-CAM is applied to highlight the facial regions that most influenced each model's decision. These discriminative heatmaps enable users to visually validate whether the system is focusing on legitimate facial evidence rather than unrelated background cues.

Once classification and interpretability components complete their tasks, the Output Generation Module presents a consolidated result containing:

- The final prediction (real or fake)
- Confidence scores from all three models.
- Grad-CAM heatmaps corresponding to each model's decision.

This architectural configuration ensures that VerityNet delivers strong performance through ensemble learning while simultaneously maintaining transparency and interpretability, both of which are essential for real-world forensic and security applications.

Table 1 Summary of Literature Review

No.	Category	Reference	Contribution	Relevance to VerityNet
1	Gen. Gap	Chandra, N., et al. (2025) [12]	Introduced Deepfake-Eval-2024, an "in-the-wild" benchmark.	Proved that SOTA detectors fail (45-50% AUC drop) on modern, real-world fakes.
2	Gen. Gap	Abbasi, M., et al. (2025) [11]	Benchmarked CNNs (incl. VGG16) on FF++ & DFDC under compression/attacks.	Confirms the generalization weakness of individual CNNs and their vulnerability to compression.
3	Gen. Gap	Guo, M., et al. (2025) [32]	Proposed an open-world domain adaptation approach (OWG-DS).	Defines a SOTA technique for generalizing to unseen manipulation types.
4	Ensemble	Wahab, H., et al. (2025) [20]	Investigated an ensemble for robust cross-dataset generalization.	Provides direct 2025 validation for VerityNet's core ensemble methodology.
5	Ensemble	Ben Jabra, M., et al. (2025) [21]	Proposed a VGG16/Inception V3/Xception ensemble (97.4% accuracy).	Confirms that ensembling pre-trained CNNs (including VGG16) is highly effective.
6	Ensemble	Zen, H. (2025) [22]	MIT Thesis. Found ensembling models trained on different forgery types boosts accuracy.	Validates the "diversity" hypothesis of VerityNet's ensemble.
7	Adversarial	ACM MM (2025) [25]	AADD-2025 Grand Challenge to design attacks that bypass detectors.	Formalizes the adversarial threat, highlighting a key limitation for all current detectors.
8	XAI	Kundu, R., et al. (2025) [30]	"TruthLens": A multimodal XAI framework using LLMs for text explanations.	Defines the SOTA for XAI, a clear "next step" beyond Grad-CAM.

9	XAI	Tariq, S., et al. (2025) [31]	"DF-P2E": A framework for non-expert users, providing narrative explanations.	Provides a roadmap for making VerityNet's XAI user-friendly.
10	XAI	Budati, J. L. N., et al. (2025) [18]	Applied Grad-CAM specifically for video forensics to enhance trust.	Directly validates VerityNet's choice of XAI methodology.
11	Bias	Xu, Y., et al. (2024) [35]	Proved SOTA detectors are strongly biased across race, gender, and age.	Identifies a critical, unaddressed ethical flaw in all models trained on public data.
12	Bias	Ju, Y., et al. (2024) [36]	The first work to propose algorithmic loss functions to improve fairness.	Provides a clear "Future Scope" solution for mitigating bias in VerityNet.
13	Efficiency	Kong, C., et al. (2024) [38]	Used Parameter-Efficient Adaptation (PEFT/LoRA) for open-set detection.	Represents the SOTA in efficiently adapting models to new threats.
14	Multimodal	Miao, H., et al. (2025) [37]	Proposed a two-stream (audio/visual) prompt learning method.	Defines a SOTA architecture for audio-visual detection, needed to stop Arup-style attacks.
15	Privacy	Liu, D., et al. (2024) [40]	Proposed a Federated Learning framework for face forgery detection.	Provides a "Future Scope" solution for training on private data.

5 Methodology

5.1 Dataset

The training dataset is curated from two high-quality open-source Kaggle repositories:

- Real and Fake Images Dataset for Image Forensics
- Real vs. Fake Faces Dataset

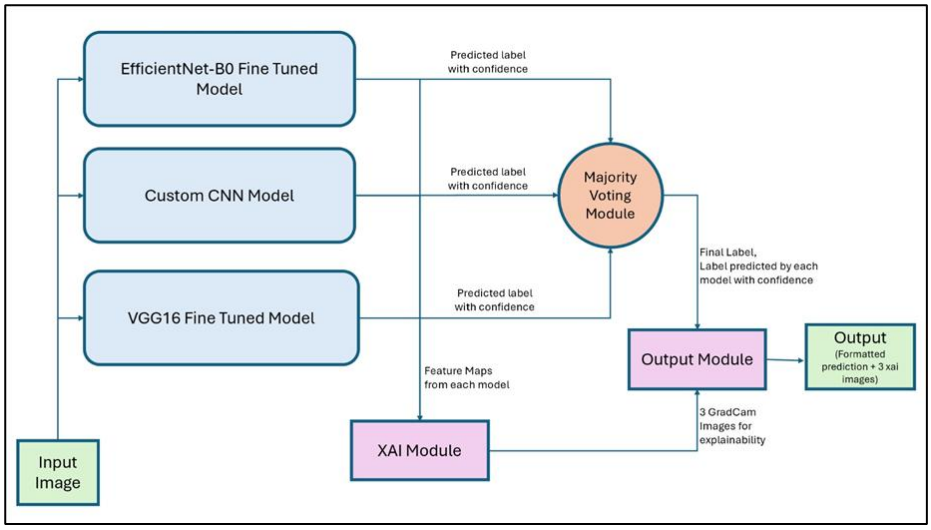

Fig. 1 System Architecture

These datasets contain authentic and GAN-generated facial images across various lighting conditions, backgrounds, and face orientations. The merged dataset comprises:

- 180,000 images for training
- 50,000 images for testing
- 10,000 images for validation

Each image is labelled as either real or fake. The merged datasets were inherently balanced, containing an equal number of samples for both classes across all splits. This large and diverse dataset ensures the model learns manipulation patterns across different generative methods like DeepFaceLab, StyleGAN, and FaceSwap.

5.2 Preprocessing

To ensure uniformity and enhance feature learning, the following preprocessing techniques were applied:

- Resizing all images to 224x224 pixels.
- Normalization of pixel values between 0 and 1.
- Data augmentation to improve generalization, including:
  - Horizontal/vertical flipping.
  - Random brightness and contrast adjustment.
  - Rotation and cropping.

Balanced class distribution was maintained across all sets.

6 Applications of Deep Learning in Deepfake Detection

Deepfake detection has real-world impact across a variety of domains:

- Facial Forensics: Assists media and law enforcement in distinguishing real from manipulated identities.
- Video Authentication: Ensures the integrity of visual evidence in legal and forensic investigations. For a model's output to be admissible as evidence, it must be interpretable, making XAI modules like Grad-CAM a crucial component [18].
- Platform Moderation: Enables automated moderation and flagging of suspicious visual content on social media and streaming platforms.
- Election Security: Detects synthetic videos used in political propaganda and to spread misinformation [5].
- Corporate Identity Verification: Validates recorded meetings and interviews for authenticity. The $25.5 million Arup fraud in 2024, executed via deepfake video, has made this a paramount application [7]. Efficient detectors are essential to prevent catastrophic economic losses, which are projected to reach $40 billion in the US by 2027 [10].
- Protecting Individuals from Harassment: A significant social impact of deepfakes is their use in creating non-consensual intimate imagery. A disturbing trend in 2024-2025 involves deepfake cyberbullying in schools, disproportionately targeting female students [41], [42].

7 Implementation Details

7.1Our Experimental Setup and Environment

The implementation and training of the VerityNet framework were executed utilizing the Google Colab environment, accelerated by an NVIDIA L4 Tensor Core GPU. This hardware configuration provided the necessary computational resources to efficiently process large batches of high-resolution image data. The models were built and trained using the TensorFlow and Keras deep learning frameworks. Data loading was optimized using TensorFlow utility, with a batch size of 32. The classification task was defined with a binary label mode, mapping 'real' images to 0 and 'fake' images to 1.

7.2 Data Preprocessing and Augmentation

To ensure uniformity and enhance the feature learning capabilities of the models, a strict preprocessing pipeline was applied to the dataset. The procedures included:

- *Resizing:* All input images were standardized to a resolution of 224x224 pixels to meet the input requirements of the selected convolutional architectures.
- *Normalization:* Pixel intensity values were normalized to a range between 0 and 1 to ensure stable gradient propagation and faster convergence during training.
- *Data Augmentation:* To improve generalization and mitigate the risk of overfitting on the training data, spatial and pixel-level augmentations were introduced. This pipeline included horizontal and vertical flipping, random adjustments to brightness and contrast, as well as random rotation and cropping.
- *Class Balance:* The datasets were inherently balanced, containing an equal number of samples for both real and fake classes. This strict class distribution was maintained across the training, validation, and testing sets to prevent the models from developing a majority-class bias.

7.3 Training Strategy for Pre-Trained Models (EfficientNetB0 and VGG16)

To effectively leverage transfer learning within a streamlined, single-pass training workflow, the pre-trained architectures (EfficientNetB0 and VGG16) were customized and fine-tuned using a partial-freezing strategy.

- *Architecture Modification:* The base models were instantiated with pre-trained ImageNet weights, excluding their original classification heads (include_top=False). Custom binary classification heads were then appended. For EfficientNetB0, the output was passed through a GlobalAveragePooling2D layer, a fully connected Dense layer (128 units, ReLU), and a final Dense layer (1 unit, Sigmoid). For VGG16, the base output was flattened and passed through a Dense layer (256 units, ReLU) before the final single-unit Sigmoid output.

- *Partial Freezing and Fine-Tuning: Rather* than executing a two-phase training process, the networks were configured for direct fine-tuning. The base models were set to be trainable, but to preserve low-level generic visual features and prevent overfitting, the earliest layers were explicitly frozen. Specifically, the first 100 layers of EfficientNetB0 and the first 10 layers of VGG16 were locked (trainable = False), while the deeper, more semantically complex layers remained open for weight updates.

- *Compilation and Training:* Both models were compiled using the Adam optimizer. A strictly reduced learning rate of 1e-5 was utilized to ensure that updates to the pre-trained weights were microscopic and stabilizing. EfficientNetB0 was trained for 20 epochs, incorporating EarlyStopping (patience of 3) and ModelCheckpoint callbacks to restore and save the

optimal weights. VGG16 followed a similar compilation strategy and was trained for 10 epochs. Both models evaluated loss using the binary cross-entropy function.

7.4 Custom CNN Architecture and Training

The lightweight Custom CNN was designed from scratch to provide a computationally efficient alternative within the ensemble. The architecture takes an input of 224x224x3 and processes it through six sequential convolutional blocks. Each block consists of a Conv2D layer, MaxPooling2D for spatial downsampling, BatchNormalization for training stability, and a Dropout layer (rate of 0.1).

Following the convolutional feature extraction, a GlobalAveragePooling2D layer collapses the spatial dimensions, feeding into a fully connected Dense layer (256 units, ReLU). To aggressively combat overfitting in this custom network, a higher Dropout rate of 0.5 was applied prior to the final Sigmoid classification layer. Furthermore, L2 regularization (weight decay = 1e-4) was applied to all convolutional and dense layers across the network. The model was trained for 10 epochs using the Adam optimizer with a configured learning rate of 1e-4, optimized against binary cross-entropy loss.

8 Model Result & Evaluation

Evaluation Methodology

To rigorously assess the performance of both the individual architecture and the proposed VerityNet ensemble, evaluations were conducted on a strictly isolated testing set of 50,000 images that were never exposed to the models during training or validation. The primary quantitative metrics utilized for evaluation include Accuracy (overall correctness), Precision (minimizing false positives, critical for avoiding false accusations of digital forgery), Recall (minimizing false negatives, critical for detecting actual threats), and the F1-Score (the harmonic mean of precision and recall, providing a balanced metric for real-world reliability). Interpretability was evaluated qualitatively using Grad-CAM heatmaps to ensure the models focused on relevant facial artifacts rather than background noise.

8.1 Individual Model Results

8.1.1 EfficientNetB0

The EfficientNetB0 model demonstrated strong performance in deepfake detection, combining high capacity with computational efficiency (Figure 2, Figure 3). Out of 10,413 real images in the validation set, it correctly identified 10,026 as genuine, misclassifying 387 as fake. For 10,492 fake samples, it accurately labeled 9,143 as fake, with 1,349 false positives. The model achieved

0.88 precision and 0.96 recall for real images, and 0.96 precision with 0.87 recall for fake images. Both classes exceeded an F1-score of 0.91, resulting in an overall accuracy of 92% (macro and weighted averages consistent with this score). These outcomes indicate that EfficientNetB0 is both robust and reliable, making it effective as a standalone classifier or as a complementary component in ensemble architectures, particularly for minimizing missed real instances while keeping false positives low.

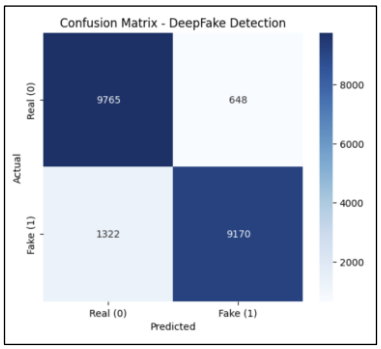

Fig.2 Confusion Matrix for EfficientNetB0 Model

Classification Report - Effnet:

	precision	recall	f1-score	support
Real (0)	0.88	0.94	0.91	10413
Fake (1)	0.93	0.87	0.90	10492
accuracy			0.91	20905
macro avg	0.91	0.91	0.91	20905
weighted avg	0.91	0.91	0.91	20905

Fig. 3 Classification Metrics for EfficientNetB0 Model

8.1.2 Custom CNN

The custom-designed CNN, despite its lightweight architecture with fewer parameters and layers, achieved commendable accuracy (Figure 4, Figure 5). For real images, it correctly classified 9,176 out of 10,413, while for fake images, it identified 9,359 out of 10,492 correctly. This corresponds to precision and recall of 0.89 and 0.88 for real images, and 0.88 and 0.89 for fake images, respectively, producing a balanced F1-score of 0.89 across both classes and an overall accuracy of 89%. While slightly less accurate than deeper pretrained models, its computational efficiency and consistent performance make it a practical choice for resource-constrained environments or real-time inference systems, where minor reductions in accuracy are acceptable for faster processing.

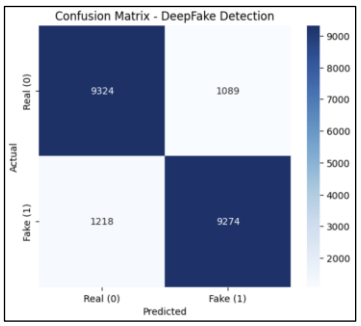

Fig. 4 Confusion Matrix for Custom CNN Model

```
Classification Report - CNN:

              precision    recall  f1-score   support

   Real (0)       0.88      0.90      0.89     10413
   Fake (1)       0.89      0.88      0.89     10492

    accuracy                           0.89     20905
   macro avg       0.89      0.89      0.89     20905
weighted avg       0.89      0.89      0.89     20905
```

Fig. 5 Classification Metrics for Custom CNN Model

8.1.3 VGG16

The VGG16 model, a deep transfer learning architecture known for its regularized design, provided stable and high recall performance (Figure 6, Figure 7). Out of 10,413 real images, it correctly identified 10,057, with 356 misclassifications. For 10,492 fake images, it correctly detected 9,134, with 1,358 mislabeled as real. VGG16 achieved 0.88 precision and 0.97 recall for real images, reflecting a strong ability to minimize false negatives, and 0.96 precision and 0.87 recall for fake images, producing F1-scores of 0.92 and 0.91, respectively. Its overall accuracy of 92% and balanced macro/weighted averages reaffirm its reliability. These results highlight VGG16's suitability for forensic and authentication applications, where accurately confirming real media is of greater importance than over-flagging.

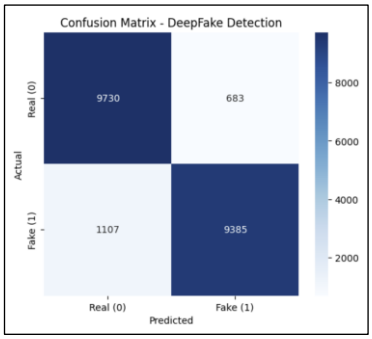

Fig. 6 Confusion Matrix for VGG16

```
Classification Report - VGG16:

              precision    recall  f1-score   support

   Real (0)       0.90      0.93      0.92     10413
   Fake (1)       0.93      0.89      0.91     10492

    accuracy                           0.91     20905
   macro avg       0.92      0.91      0.91     20905
weighted avg       0.92      0.91      0.91     20905
```

Fig. 7 Confusion Matrix for VGG16

8.2 Ensemble with Majority Voting

The ensemble model, integrating predictions from EfficientNetB0, VGG16, and the custom CNN through majority voting, achieved the highest overall accuracy of 94%, surpassing each individual model (Figure 8, Figure 9). It yielded 0.92 precision and 0.97 recall for the real class, alongside 0.97 precision and 0.94 recall for the fake class, indicating excellent balance across both detection categories. By leveraging the complementary strengths of each architecture, the ensemble effectively reduced both false positives and false negatives. The superior F1-scores and high macro/weighted averages confirm its robustness and generalization ability. This hybrid approach proves ideal for high-stakes deepfake detection scenarios, combining efficiency, accuracy, and resilience against model-specific biases.

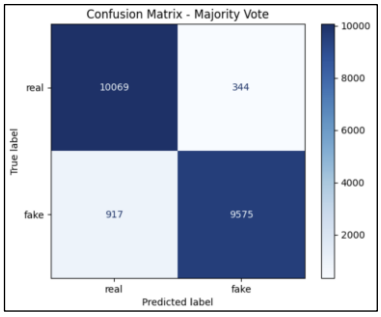

Fig. 8 Confusion Matrix for Ensemble with Majority Voting

Classification Report - Majority Vote:

	precision	recall	f1-score	support
real	0.92	0.97	0.94	10413
fake	0.97	0.91	0.94	10492
accuracy			0.94	20905
macro avg	0.94	0.94	0.94	20905
weighted avg	0.94	0.94	0.94	20905

Fig. 9 Classification Metrics for Ensemble with Majority Voting

8.3 Comparative Performance Table

Table 2 depicts a comparative performance analysis of the individual and ensemble models against the performance metrics like precision, recall and F1-Score.

8.4 Comparison with State-of-the-Art (SOTA) Methods

To contextualize VerityNet's performance within the current digital forensics landscape, it is essential to compare our hybrid ensemble against recent state-of-the-art (SOTA) methodologies. While direct cross-dataset metric comparisons can sometimes be challenging due to differing benchmarks and the recent emergence of open-world testing like Deepfake-Eval-2024, analyzing the architectural paradigms provides a clear view of VerityNet's comparative efficacy. Table 3 outlines how VerityNet measures up against widely adopted baselines and recent high-performing models cited in contemporary literature.

Table 2 Comparative Performance Table

Metric Category	Metric Name	EfficientNetB0	CNN	VGG16	Majority Vote (Ensemble)
Real Class	Precision	0.88	0.88	0.90	**0.92**
	Recall	0.94	0.90	0.93	**0.97**
	F1-Score	0.91	0.89	0.92	**0.94**
Fake Class	Precision	0.93	0.89	0.93	**0.97**
	Recall	0.87	0.88	0.89	**0.91**
	F1-Score	0.90	0.89	0.91	**0.94**
Overall	Accuracy	0.91	0.89	0.91	**0.94**

Discussion of Trade-offs:

While massive, transformer-based architectures (ViTs) and heavier CNN ensembles (such as those utilizing InceptionV3 alongside XceptionNet) can achieve marginal accuracy improvements (pushing into the 96-98% range), these gains come at a severe computational cost. Transformer models, in particular, are highly complex, energy-intensive, and notoriously slow during inference. This renders them impractical for real-time platform moderation or on-device corporate identity verification—two of the most critical applications for modern deepfake detection.

VerityNet successfully bridges this gap. By combining EfficientNetB0 (optimized for mobile/edge scaling) and a lightweight Custom CNN alongside the robust VGG16, VerityNet achieves a highly competitive 94% accuracy with a significantly lower computational footprint than ViT-based SOTA models.
Furthermore, many contemporary SOTA models operate as "black boxes," making their decision-making processes unclear. In high-stakes environments like law, finance, and cybersecurity, an uninterpretable output is often inadmissible as evidence. VerityNet's native integration of Grad-CAM provides immediate visual explainability (MediaXAI), making it not just a highly accurate SOTA competitor, but a vastly more practical and transparent tool for real-world deployment.

8.5 Computational Complexity and Real-World Deployment

While the primary objective of VerityNet is to maximize deepfake detection accuracy, an ensemble framework inherently introduces additional computational overhead compared to single-model systems. To assess the viability of VerityNet for real-world applications—such as live video conference authentication or high-volume social media moderation—it is necessary to evaluate its computational complexity and inference efficiency.

Table 3 Comparison with State-of-the-Art (SOTA) Methods

Methodology / Architecture	Reported Accuracy	Computational Overhead	Interpretability (XAI)	Primary Drawback
XceptionNet (Standard Baseline)	~90.0% - 92.0%	Moderate	Low	Struggles with heavy compression and diffusion artifacts.
Vision Transformers (ViTs)	~95.0% - 98.0%	Very High	Low (Black Box)	Slow inference; inapplicable for real-time video streaming.
Heavy CNN Ensembles [21]	97.4%	High	Moderate	Resource-intensive due to combining multiple heavy backbone networks.
VerityNet (Ours)	94.0%	Moderate / Scalable	High (Native Grad-CAM)	Accuracy trails massive ViT models slightly in controlled settings.

8.5.1 Computational Footprint

Unlike Vision Transformers (ViTs), which suffer from quadratic computational complexity concerning image patch size, the convolutional architectures utilized in VerityNet operate with linear scaling.

- **EfficientNetB0:** Specifically designed for mobile and edge deployment, it utilizes mobile inverted bottleneck convolutions (MBConv), ensuring a minimal parameter count (~5.3 million) and highly efficient FLOPs (Floating Point Operations per Second).
- **Custom CNN:** As a shallow, highly regularized network, its computational footprint is negligible relative to standard deep learning frameworks.
- **VGG16:** While VGG16 is the heaviest component of the ensemble (~138 million parameters), its uniform architecture allows for highly optimized matrix multiplications on modern GPU hardware.

8.5.2 Inference Efficiency

The real-world latency of an ensemble model is often dictated by its slowest component. However, the architecture of VerityNet allows the three constituent models (EfficientNetB0, VGG16, Custom CNN) to process the input image independently. In a production environment, these three inference branches

can be executed in parallel on multi-core processors or partitioned GPUs. Consequently, the theoretical inference latency of the entire VerityNet ensemble is roughly equivalent to the standalone inference time of VGG16, plus a negligible fraction of a millisecond to execute the arithmetic majority vote.

8.5.3 Deployment Considerations and Scalability

For VerityNet to be deployed in real-time scenarios (e.g., processing video at 30 frames per second), specific production-level optimizations are recommended:

- **Cloud-Based Moderation:** For asynchronous tasks like scanning uploaded videos on platforms like X or YouTube, VerityNet can be deployed on standard cloud infrastructure (e.g., NVIDIA T4 or L4 instances). Its moderate memory requirements allow for large batch sizes, maximizing throughput.
- **Edge Deployment and Quantization:** To deploy VerityNet directly into corporate video conferencing software (to prevent scenarios like the $25.5M Arup fraud), inference must be performed locally to preserve privacy and reduce bandwidth latency. The models can be subjected to post-training quantization (reducing FP32 weights to FP16 or INT8) or compiled using inference accelerators like NVIDIA TensorRT. Because the ensemble relies heavily on the robust architectures of EfficientNet and VGG, such quantization is expected to significantly boost frames-per-second (FPS) throughput with only a marginal drop in the 94% accuracy threshold.

9 Challenges & Limitations

9.1 Vulnerability to Adversarial Perturbations

Although ensemble methods improve reliability, several unresolved challenges continue to affect practical adoption of deepfake detection systems. Adversarial manipulation remains a prominent concern, as shown by results from the AADD-2025 challenge [25], where small and visually imperceptible alterations can drastically reduce model confidence. This exposes a vulnerability in CNN-based detectors that rely on surface-level feature dependencies.

9.2 The Generalization Gap and Unseen Manipulation Techniques

A persistent long-term issue in digital forensics is the "generalization gap" [11]. Deep learning models are inherently biased toward their training distributions. VerityNet's underlying training data is predominantly composed of Generative Adversarial Network (GAN) manipulations. However, the generative landscape has rapidly shifted toward advanced diffusion-based architectures, which produce significantly subtler artifacts that bypass earlier GAN-focused detectors [12]. When models trained on specific synthetic manipulations are exposed to entirely new, unseen deepfake generation methods in the wild,

accuracy rates can plummet dramatically—sometimes dropping by 45 to 50 percent.
This open-world generalization issue is further exacerbated by real-world media transmission. Social media platforms employ aggressive compression algorithms that destroy the high-frequency pixel cues and fine artifact patterns that VerityNet relies upon for classification [15]. Additionally, if the training datasets lack sufficient demographic diversity, the model may exhibit performance disparities and unfair bias across different race, gender, and age subsets [35].

9.3 Absence of Temporal and Audio-Visual Context

Finally, VerityNet is fundamentally constrained by its spatial-only processing approach. When applied to video data, the system evaluates media on a frame-by-frame basis. This methodology ignores critical temporal reasoning, such as unnatural blinking rates, heartbeat pulse inconsistencies, or frame-to-frame spatial jitter. Furthermore, sophisticated financial fraud attacks (such as the 2024 Arup case) often rely heavily on cloned audio [7]. As an image-only model, VerityNet cannot assess the synchronization between lip movements and speech, nor can it flag deepfake audio. Relying solely on visual streams limits the framework's reliability in comprehensive multimedia forensics. The research studies like TruthLens and strategies that combine audio and visual stream [37] show that the inclusion of the temporal context increases reliability, particularly when applied to real-life video data.

10 Conclusion

This work presented VerityNet, a robust deepfake detection framework based on ensemble learning. By integrating EfficientNetB0, VGG16, and a custom CNN through a majority voting mechanism, the system achieved 94% detection accuracy, outperforming individual models in reliability and resilience. Grad-CAM visualization further enhances trust by providing clear interpretability, making the model suitable for forensic and organizational use where decisions must be justified.

While the approach demonstrates strong performance across controlled scenarios, challenges remain in handling real-world degradations, adversarial strategies, and unseen manipulation formats. With planned enhancements including multimodal analysis, accelerated deployment, and fairness-driven improvements, VerityNet holds significant potential for serving as a practical and scalable solution for digital content authentication in the rapidly evolving era of synthetic media.

Acknowledgment

The authors acknowledge Thakur College of Engineering and Technology, Kandivali, Mumbai for laboratory infrastructure and computational resources. No external funding was received.

Contribution of each Author

Yash Rawal: Conceptualization, Project Lead, Model Development, Implementation, Grad-CAM Implementation, Training Strategy Development, and Performance Evaluation; Shivam Sati: Data Preprocessing, Dataset Preparation, Model Training Support, Experimental Analysis, Manuscript Writing, Review, and Editing; Meet Shah: Literature Review, Testing, and Result Validation; Swapnil Bhagat: Supervision and Project Coordination.

Conflict of Interest Statement

The authors declare no conflict of interest.

References

1. T. Chen and A. Magramo, "The $25 million deepfake: how AI-generated video and audio is changing 'social engineering'," World Economic Forum, Feb. 2025. [Online]. Available: https://www.weforum.org/stories/2025/02/deepfake-ai-cybercrime-arup/
2. DeepStrike, "Deepfake Statistics 2025: AI Fraud Data & Trends," Industry Report, 2025. [Online]. Available: https://deepstrike.io/blog/deepfake-statistics-2025
3. Keepnet Labs, "Deepfake Statistics and Trends," Industry Report, 2025. [Online]. Available: https://keepnetlabs.com/blog/deepfake-statistics-and-trends
4. SQ Magazine, "Deepfake Statistics 2025: The Latest Trends & Data," SQ Magazine, Oct. 2025. [Online]. Available: https://sqmagazine.co.uk/deepfake-statistics/
5. European Parliamentary Research Service (EPRS), "Children and deepfakes," EPRS Briefing, Sep. 2025. [Online]. Available:(https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/775855/EPRS_BRI(2025)775855_EN.pdf)
6. OpenFox, "Deepfakes and Their Impact on Society," OpenFox, 2024. [Online]. Available: https://www.openfox.com/deepfakes-and-their-impact-on-society/

7. B. Colman, "Why detecting dangerous AI is key to keeping trust alive," World Economic Forum, Jul. 2025. [Online]. Available: https://www.weforum.org/stories/2025/07/why-detecting-dangerous-ai-is-key-to-keeping-trust-alive/
8. A. Lau, "Arup confirms it was victim of $25.6 million deepfake video call scam," The Guardian, May 2024. [Online]. Available: https://www.theguardian.com/technology/article/2024/may/17/uk-engineering-arup-deepfake-scam-hong-kong-ai-video
9. R. Greig, "'This happens more frequently than people realize': Arup chief on the lessons learned from a $25m deepfake crime," World Economic Forum, Jul. 2025. [Online]. Available: https://www.weforum.org/stories/2025/02/deepfake-ai-cybercrime-arup/
10. World Economic Forum, "Global Cybersecurity Outlook 2025," WEF Report, Jul. 2025, citing Deloitte. [Online]. Available: https://www.weforum.org/stories/2025/07/why-detecting-dangerous-ai-is-key-to-keeping-trust-alive/
11. M. Abbasi, P. Váz, J. Silva, and P. Martins, "Comprehensive Evaluation of Deepfake Detection Models: Accuracy, Generalization, and Resilience to Adversarial Attacks," Applied Sciences, vol. 15, no. 3, p. 1225, 2025. [Online]. Available: https://doi.org/10.3390/app15031225
12. N. Chandra, R. Murtfeldt, L. Qiu, and O. Etzioni, "Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024," arXiv preprint arXiv:2503.02857, 2025. [Online]. Available: https://arxiv.org/abs/2503.02857
13. SimaClassify, "SimaClassify vs. Hive 2025: Deepfake Accuracy Face-Off," Sima Labs, 2025. [Online]. Available: https://www.simalabs.ai/resources/simaclassify-vs-hive-2025-deepfake-accuracy-face-off
14. EmergentMind, "DeepFake-Eval-2024," Topic Review, Oct. 2025. [Online]. Available: https://www.emergentmind.com/topics/deepfake-eval-2024
15. Sider.ai, "Deepfake Detection in 2025: Methods, Benchmarks, and What Actually Works," Sider.ai Blog, 2025. [Online]. Available: https://sider.ai/blog/ai-tools/deepfake-detection-in-2025-methods-benchmarks-and-what-actually-works
16. C. Yu, S. Jia, X. Fu, J. Liu, J. Tian, J. Dai, X. Wang, S. Lyu, and J. Han, “Explicit Correlation Learning for Generalizable Cross-Modal Deepfake Detection,” arXiv preprint arXiv:2404.19171, 2024. [Online]. Available: https://arxiv.org/abs/2404.19171
17. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017. [Online]. Available: https://arxiv.org/abs/1610.02391
18. J. L. N. Budati, A. B. Jadam, and R. Malleboyina, "Explainable AI for Deepfake Detection: A Grad-CAM Approach to Video Forensics," in 2025 6th International Conference for Emerging Technology (INCET), 2025. [Online]. Available:(https://doi.org/10.1109/INCET64471.2025.11140952)

19. D. T. H. Le, et al., "Explainable AI-Driven Deepfake Detection Using Grad-CAM," in 2024 IEEE International Symposium on Consumer Technology (ISCT), 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10791156/
20. H. Wahab, H. Ugail, and L. Jaleel, "Ensemble-Based Deepfake Detection using State-of-the-Art Models with Robust Cross-Dataset Generalisation," arXiv preprint arXiv:2507.05996, 2025. [Online]. Available: https://arxiv.org/abs/2507.05996
21. M. Ben Jabra, O. Cheikhrouhou, and A. Benamor, "Deepfake detection through ensemble learning," Multimedia Tools and Applications, 2025. [Online]. Available: https://doi.org/10.1007/s11042-025-20932-w
22. H. Zen, "Deepfake Face Detection: An Ensemble Framework for Generalized Classification in Biometric Verification Systems," M.Eng. Thesis, Massachusetts Institute of Technology (MIT), May 2025. [Online]. Available: https://dspace.mit.edu/bitstream/handle/1721.1/162919/zen-hzen-meng-eecs-2025-thesis.pdf
23. M. M. Batista, "Comparative Analysis of Deepfake Detection Models: New Approaches and Perspectives," arXiv preprint arXiv:2504.02900, 2025. [Online]. Available: https://arxiv.org/abs/2504.02900
24. A. George, C. Ecabert, H. Otroshi Shahreza, K. Kotwal, and S. Marcel, “EdgeFace: Efficient Face Recognition Model for Edge Devices,” IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 6, no. 2, pp. 158–168, 2024, doi: 10.1109/TBIOM.2024.3352164. [Online]. Available:(https://infoscience.epfl.ch/entities/publication/1b7ad7cf-1499-4c99-addc-c3df94a3923a)
25. ACM MM 2025, "(AADD-2025) Adversarial Attacks on Deepfake Detectors: A Challenge in the Era of AI-Generated Media," ACM MM 2025 Grand Challenge. [Online]. Available: https://acmmm2025.org/grand-challenge/
26. J. Liang, S. Liang, A. Liu, X. Jia, J. Kuang, and X. Cao, "Poisoned Forgery Face: Towards Backdoor Attacks on Face Forgery Detection," in Proceedings of the International Conference on Learning Representations (ICLR), 2024. [Online]. Available:(https://openreview.net/forum?id=8iTpB4RNvP)
27. X. Han, S. Yang, W. Wang, Z. He, and J. Dong, "Is It Possible to Backdoor Face Forgery Detection with Natural Triggers?" arXiv preprint arXiv:2401.00414, 2024. [Online]. Available: https://arxiv.org/abs/2401.00414
28. F. M. H. Nazneen, et al., "Explainable AI (XAI) for Deepfake Detection," Applied Sciences, vol. 15, no. 2, p. 725, 2025. [Online]. Available: https://doi.org/10.3390/app15020725
29. S. Tariq, S. S. Woo, P. Singh, I. Irmalasari, S. Gupta, and D. Gupta, "From Prediction to Explanation: Multimodal, Explainable, and Interactive Deepfake Detection Framework for Non-Expert Users," arXiv preprint arXiv:2508.07596, 2025. [Online]. Available: https://arxiv.org/abs/2508.07596
30. R. Kundu, A. Balachandran, and A. K. Roy-Chowdhury, "TruthLens: Explainable DeepFake Detection for Face Manipulated and Fully Synthetic Data," arXiv preprint arXiv:2503.15867, 2025. [Online]. Available: https://arxiv.org/abs/2503.15867

31. Q. Lin, S. He, H. Zhang, Z. Lin, and C. Shen, "Curricular Dynamic Forgery Augmentation for General Deepfake Detection," in Proceedings of the European Conference on Computer Vision (ECCV), 2024. [Online]. Available: https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/11581.pdf
32. M. Guo, Q. Yin, W. Lu, and X. Luo, "Towards Open-world Generalized Deepfake Detection: General Feature Extraction via Unsupervised Domain Adaptation," arXiv preprint arXiv:2505.12339, 2025. [Online]. Available: https://arxiv.org/abs/2505.12339
33. H. She, Y.-J. Hu, B.-B. Liu, J. Li, and C.-T. Li, “Using Graph Neural Networks to Improve Generalization Capability of the Models for Deepfake Detection,” IEEE Transactions on Information Forensics and Security, vol. 19, pp. 8414–8427, 2024, doi: 10.1109/TIFS.2024.3451356. [Online]. Available:(https://dl.acm.org/doi/abs/10.1109/TIFS.2024.3451356)
34. Y. Xu, P. Terhörst, M. Pedersen, and K. Raja, “Analyzing Fairness in Deepfake Detection With Massively Annotated Databases,” IEEE Transactions on Technology and Society, vol. 5, no. 1, pp. 93–106, 2024, doi: 10.1109/TTS.2024.3365421. [Online]. Available:(https://ieeexplore.ieee.org/document/10438899)
35. Y. Ju, M. He, Z. Wang, Z. Wu, and S. Lyu, "Improving Fairness in Deepfake Detection," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024. [Online]. Available:(https://openaccess.thecvf.com/content/WACV2024/papers/Ju_Improving_Fairness_in_Deepfake_Detection_WACV_2024_paper.pdf)
36. C. Kong, H. Li, and S. Wang, "Open-Set Deepfake Detection: A Parameter-Efficient Adaptation Method with Forgery Style Mixture," arXiv preprint arXiv:2408.12791, 2024. [Online]. Available: https://arxiv.org/abs/2408.12791
37. H. Miao, Y. Guo, Z. Liu, and Y. Wang, "Multi-modal Deepfake Detection via Multi-task Audio-Visual Prompt Learning," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 1, pp. 612-621, 2025. [Online]. Available: https://doi.org/10.1609/aaai.v39i1.32042
38. B. Zhang, "BayesTune: A Bayesian Optimization-based Parameter Efficient Fine-Tuning Method for Adaptive In-Domain Generalization of Deepfake Detection," M.I.D.S. Capstone, UC Berkeley, 2025. [Online]. Available: https://www.ischool.berkeley.edu/sites/default/files/bb_paper.pdf
39. D. Liu, Z. Dang, C. Peng, N. Wang, R. Hu, and X. Gao, "Federated Face Forgery Detection Learning with Personalized Representation," arXiv preprint arXiv:2406.11145, 2024. [Online]. Available: https://arxiv.org/abs/2406.11145
40. M. Ben Jabra, O. Cheikhrouhou, and A. Benamor, "Deepfake detection through ensemble learning," Multimedia Tools and Applications, 2025. [Online]. Available: https://doi.org/10.1007/s11042-025-20932-w (Note: Cites a CNN+ViT ensemble).
41. NEARI, "AI Deepfakes: A Disturbing Trend in School Cyberbullying," NEARI.org, 2025. [Online]. Available: https://www.neari.org/advocating-change/new-from-neari/ai-deepfakes-disturbing-trend-school-cyberbullying

42. T. M. Wani, S. A. A. Qadri, F. A. Wani, and I. Amerini, “Navigating the Soundscape of Deception: A Comprehensive Survey on Audio Deepfake Generation, Detection, and Future Horizons,” Foundations and Trends® in Privacy and Security, vol. 6, no. 3-4, pp. 153–345, 2024, doi: 10.1561/3300000048. [Online]. Available:(https://www.nowpublishers.com/article/Details/SEC-048)

Yash Rawal is a Computer Engineering student with a strong interest in Artificial Intelligence, Machine Learning, Deep Learning, Data Science, and Explainable Artificial Intelligence (XAI). He has developed expertise in machine learning model development, deep learning architectures, data analysis, and model interpretability techniques through academic and research work. He has hands-on experience with technologies and frameworks such as TensorFlow, PyTorch, NumPy, Pandas, OpenCV, and Scikit-learn for developing and evaluating intelligent systems. His research interests include Deep Learning, Computer Vision, Natural Language Processing, Generative AI, Explainable AI (XAI), and Applied Machine Learning. He is particularly interested in improving the transparency, reliability, and trustworthiness of AI systems through interpretability and explainability methods.

Shivam Sati is a Computer Engineering student with an interest in Artificial Intelligence, Machine Learning, and Data Science. He has worked on projects involving machine learning, deep learning, and software development, gaining practical experience with technologies such as Python, TensorFlow, PyTorch, NumPy, Pandas, OpenCV, and Scikit-learn. His areas of interest include Computer Vision, Deep Learning, Generative AI, and Applied Machine Learning. He is passionate about exploring innovative AI-driven solutions to real-world problems and continuously expanding his knowledge in emerging technologies.

Meet Shah is a Computer Engineering student with a strong interest in Artificial Intelligence, Machine Learning, Deep Learning, Full-Stack Development, and Data Science. He has worked on several projects in these domains, including deepfake detection using ensemble learning techniques, AI-powered email response generation, intelligent web applications, and machine learning-based predictive systems. He has hands-on experience with technologies and frameworks such as NumPy, Pandas, React.js, Spring Boot, and Node.js for developing and deploying intelligent software solutions. His research interests include Deep Learning, Computer Vision, Explainable Artificial Intelligence (XAI), Natural Language Processing, Generative AI, Large Language Models (LLMs), and Applied Machine Learning. He is particularly interested in building AI systems that enhance digital trust, automate decision-making processes, and solve real-world challenges through intelligent technologies.

Swapnil Bhagat is an Assistant Professor in the Department of Computer Engineering at Thakur College of Engineering and Technology, Mumbai, India. He has a strong interest in Computer Networks, Cloud Computing, and Digital Forensics. His academic and research interests include network security, cloud-based systems, cyber forensics, and emerging technologies in distributed computing. He is actively involved in teaching, mentoring student projects, and guiding research activities in the fields of computer networks and cybersecurity. His work focuses on bridging theoretical concepts with practical applications to address real-world computing challenges