computers

MDPI

*Article*

# Ensemble-Based Biometric Verification: Defending Against Multi-Strategy Deepfake Image Generation

**Hilary Zen [1,*,†], Rohan Wagh [1,†], Miguel Wanderley [2], Gustavo Bicalho [2], Rachel Park [1], Megan Sun [1], Rafael Palacios [3,4], Lucas Carvalho [2], Guilherme Rinaldo [2] and Amar Gupta [1]**

[1] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA; rwagh@mit.edu (R.W.); rpark02@mit.edu (R.P.); megansun@mit.edu (M.S.); agupta@mit.edu (A.G.)
[2] Instituto de Ciência e Tecnologia Itaú -ICTi, Av. Engenheiro Armando de Arruda Pereira 774, Jabaquara - São Paulo, São Paulo 04308-000, Brazil; miguel.wanderley@itau-unibanco.com.br (M.W.); gustavo.bicalho@itau-unibanco.com.br (G.B.); lucas.a.rocha-carvalho@itau-unibanco.com.br (L.C.); guilherme.rinaldo@itau-unibanco.com.br (G.R.)
[3] Cybersecurity at MIT Sloan (CAMS), Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA; palacios@mit.edu
[4] Institute for Research in Technology, Universidad Pontificia Comillas, Alberto Aguilera 23, 28015 Madrid, Spain
* Correspondence: hzen@mit.edu
† These authors contributed equally to this work.

**Abstract:** Deepfake images, synthetic images created using digital software, continue to present a serious threat to online platforms. This is especially relevant for biometric verification systems, as deepfakes that attempt to bypass such measures increase the risk of impersonation, identity theft and scams. Although research on deepfake image detection has provided many high-performing classifiers, many of these commonly used detection models lack generalizability across different methods of deepfake generation. For companies and governments fighting identify fraud, a lack of generalization is challenging, as malicious actors may use a variety of deepfake image-generation methods available through online wrappers. This work explores if combining multiple classifiers into an ensemble model can improve generalization without losing performance across different generation methods. It also considers current methods of deepfake image generation, with a focus on publicly available and easily accessible methods. We compare our framework against its underlying models to show how companies can better respond to emerging deepfake generation methods.

**Keywords:** deepfakes; biometric verification systems; generalization; ensemble learning; deepfake detection model



Academic Editor: Paolo Bellavista

Received: 23 April 2025
Revised: 23 May 2025
Accepted: 2 June 2025
Published: 9 June 2025

**Citation:** Zen, H.; Wagh, R.; Wanderley, M.; Bicalho, G.; Park, R.; Sun, M.; Palacios, R.; Carvalho, L.; Rinaldo, G.; Gupta, A. Ensemble-Based Biometric Verification: Defending Against Multi-Strategy Deepfake Image Generation. *Computers* **2025**, *14*, 225. https://doi.org/10.3390/computers14060225



## 1. Introduction

Recent advancements in Artificial Intelligence (AI) and deep learning have drastically altered our ability to manipulate multimedia content. Modern AI tools, often available for free online, can be easily used to modify the visual appearance of individuals in images or videos, allowing for the generation of images that are near indistinguishable from real world content, coined as Deepfakes. A deepfake is a digitally altered piece of media made by either modifying an existing video by replacing one person's face over another or creating entirely new content involving a target's likeliness. While our ability to create such content highlights developments in Artificial Intelligence and deep learning methods, it also presents a growing set of ethical and security challenges. Malicious uses of generative

AI allow the attacker to create multimedia content including fake images, fake videos, and voice cloning. This multimedia content can be used to impersonate someone, when the content is generated for a target, or to create the profile of a person that does not exist in real life. A survey of methods for creating deepfake content is presented in [1], along with a description of some of the characteristics that help to identify fake content.

As deepfake generation models rapidly evolve, it has become increasingly more difficult for humans to detect the difference between real and fake content. When presented with a deepfake detection task, humans perform no better than randomly guessing to identify deepfakes [2]. As such, modern methods for detecting deepfakes have increasingly relied on machine learning and deep learning models to classify content, such as convolutional neural networks (CNNs), vision transformers, and recurrent neural networks. These detection models, while often achieving high accuracy on their datasets, frequently struggle to maintain high performance on data outside of the scope of their dataset. More specifically, high performance against a specific deepfake generation technique does not translate to effectiveness across a diverse range of generation methods. Detection models easily overfit to artifacts or signals specific to one generation method, creating a tradeoff between achieving high accuracy and ensuring generalization across different techniques.

As generation models continue to diversify, they challenge the effectiveness of overly specialized models, which highlights the need to develop models that are well generalized against a wide variety of deepfake generation methods. We present a new paradigm for efficient deepfake detection: an ensemble of methods that individually perform best on a wide range of deepfakes categorized by their generation method. An ensemble model that takes multiple predictions into account performs better across a mix of deepfakes generated from different sources.

In our experiments across a variety of datasets, the ensemble model shows robust performance across all datasets, whereas the individual models typically perform well on one type of deepfake. We discuss the benefits and limitations of ensemble models, particularly within industry standards and requirements. Although this study has focused on deepfake detection in images, it is important to highlight that recent research [3–5] has explored multimodal analyses, utilizing data such as audio and video. In some cases [5], analyses combining audio and video have demonstrated more relevant results than unimodal systems, positioning themselves as a promising field for fraud detection in the coming years. However, it is crucial to emphasize that the limited availability of multimodal datasets poses a significant challenge to advancing this area of research.

After this introductory section, the rest of this paper is structured as follows: Section 2 discusses the motivation for deepfake detection, highlighting its real-world impact and industry concerns. Sections 3–5 review related work in key generative techniques, the most relevant approaches for deepfake detection, and the general ensemble approach. Section 6 introduces the datasets that we use throughout this work. Section 7 examines the limitations of state-of-the-art models when restricted to single-generation strategies, while Section 8 evaluates their performance across a proposed diverse dataset. Section 9 introduces our ensemble model, designed for improved robustness and generalization, followed by its evaluation in Section 10. Section 11 provides an in-depth discussion of the findings, analyzing the observed behaviors of different approaches, and Section 12 concludes with final remarks and future research directions.

## 2. Motivation

The financial sector increasingly relies on biometric verification to authenticate user identities for banking transactions and other sensitive operations. This reliance stems from the need for secure and user-friendly authentication mechanisms. However, as

biometric verification systems become more widespread, they also face growing threats from adversarial attacks, particularly through the use of deepfake and face-swapping technologies [6,7].

A recent Ipsos poll (2023) [8] conducted for Wells Fargo revealed that nearly one in three Americans (31%) have already been victims of online financial fraud. Credit card fraud is the most common type, affecting 64% of victims, followed by data breaches (32%) and account hacking (31%). Attackers increasingly exploit AI-generated deepfake videos and voice cloning to impersonate victims, sometimes bypassing account access security systems. If a malicious actor is able to gain access to an account, they can issue credit cards, take out loans, order money transfers, etc. leading to financial losses and potential reputational damage for banks and other financial entities.

Face-swapping applications and tools have gained popularity in recent years, offering users the ability to modify facial appearances in real time. These tools are widely available and require minimal technical expertise, making them a growing concern for financial institutions. The accessibility of such technology increases the likelihood of fraudulent activities that bypass traditional biometric verification measures [9].

A major challenge for companies in the financial sector is developing robust biometric verification systems capable of countering the evolving deepfake landscape. These systems must meet several critical requirements: high performance against multiple face generation techniques, low false positive rates, rapid processing speeds, and non-discriminative performance across diverse demographics [10]. Financial institutions must continually refine their verification processes and integrate advanced fraud detection mechanisms to stay ahead of emerging threats.

Deepfakes are increasingly relevant for developing defense mechanisms across various systems, not limited to financial contexts [11]. While promising in areas like entertainment, deepfakes raise concerns about privacy invasion, fake news, and phishing attacks. Advances in generative models like VAEs, GANs, and diffusion models complicate detection, demanding more sophisticated techniques. A key challenge is creating robust models capable of detecting diverse falsification methods. Strengthening detection is critical for securing facial recognition systems and mitigating the risks of this evolving technology. We further include that, in addition to proactive detection of deepfakes, the reactive application of detection methods is equally important to trace fraudulent activities and block fraudsters in their various attempts.

Considering the Top Threats to Cloud Computing 2024 report by the Cloud Security Alliance (CSA) [12], deepfaked images, in our view, have the potential to exacerbate at least 5 of the 11 major cloud security threats: Identity and Access Management (IAM), Insecure Interfaces and APIs, insecure third-party resources, system vulnerabilities, and unauthenticated resource sharing. These threats underscore the critical need to address the risks posed by generative technologies, which can significantly compromise cloud security frameworks.

In our analysis, third-party authentication platforms, especially those utilizing facial biometric systems, are particularly vulnerable to fraudulent activities enabled by advanced deepfake image generation. These platforms represent a significant point of vulnerability for a wide range of applications, with financial systems being among the most affected. Moreover, as noted in the CSA report, the growing reliance on digital platforms for biometric authentication, coupled with increasing leaks of sensitive data—including audio, video, biometric images, and employee Personally Identifiable Information (PII)—amplifies the likelihood of deepfake-driven biometric attacks. However, the CSA report provides limited discussion on the specific impacts of generative technologies, such as

facial deepfakes, emphasizing the urgent need for further exploration of this topic from a cybersecurity perspective.

## 3. Deepfake Generation

Many different state-of-the-art models are available for deepfaked image generation. These models are accessible online through websites or applications that fraudsters can use to produce fraudulent images. These methods follow very different processes for developing images and as a result leave different artifacts that classifiers can take advantage of to identify fake images. While the breadth of generation methods is quite broad, this paper chooses to focus on three distinct generation methods: GANs, VAEs, and diffusion models. By exploring how ensemble models can generalize across these three methods, a framework can be developed to further expand to incorporate a broader class of images.

### *3.1. GANs*

GANs are currently one of the most powerful tools for image generation. In some cases, the images created by GANs are so realistic that even humans find it difficult to distinguish them. They can generate images of various objects, animals, or even body parts, depending on the training dataset used [13].

In general, GANs consist of two types of neural networks: the generator and the discriminator. As the name suggests, the generator is responsible for creating images, while the discriminator evaluates the quality of the generated images to make them appear more authentic. The discriminator assigns a score to each image, where a higher score indicates greater authenticity. Training is thus performed by optimizing this score [14].

Two well-known GAN architectures are the Deep Convolutional GAN (DCGAN) and the Progressive Growing GAN (PGGAN). DCGAN consists of convolutional layers with max pooling, with batch normalization being applied in both the generator and the discriminator networks. On the other hand, PGGAN starts with a smaller number of layers, progressively increasing the number of layers in the generator and discriminator networks, thus producing higher-resolution images throughout training. It is worth noting that other GAN variants are also widely used for image generation, such as BEGAN, Adversarial Network, Cycle GAN, Style GAN, among others [15].

### *3.2. VAEs*

Variational Autoencoders (VAEs) are a type of generative model that builds on traditional autoencoders by incorporating a probabilistic framework to better capture the underlying distribution of data. They consist of two main components: an encoder and a decoder. The encoder maps input data into a latent space, producing the parameters (mean and variance) of a probability distribution, typically Gaussian, instead of deterministic points. This probabilistic approach allows VAEs to model the variability inherent in the data. During training, the decoder reconstructs the input by sampling from these distributions, minimizing the difference between the original and reconstructed data. The training process optimizes two objectives: a reconstruction loss, ensuring accurate reconstruction, and a regularization term, the Kullback-Leibler (KL) divergence, which encourages the latent space to resemble a predefined prior, usually a standard normal distribution. This design enables accurate data reconstruction and also allows the generation of new, realistic data samples, making VAEs highly effective for applications like data synthesis and anomaly detection [16,17].

VAEs have proven to be powerful tools in detecting deepfakes, thanks to their ability to model the distribution of authentic data and flag anomalies. When trained exclusively on real images, a VAE learns to reconstruct genuine data effectively. However, when

presented with a deepfake, the reconstruction error tends to be significantly higher due to the mismatch between the fake input and the learned distribution of real data. This increase in reconstruction error serves as a reliable indicator of deepfakes.

An example of this application is the OC-FakeDect framework [18], which employs a one-class VAE trained exclusively on real face images. This method treats deepfakes as anomalies, achieving 97.5% accuracy on the NeuralTextures dataset and does not use fake images during training.

### 3.3. Diffusion Methods

Traditional generative approaches such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) have limitations in terms of stability, training complexity, and mode collapse, where the generator produces a limited variety of outputs, failing to capture the full diversity of the training data. Diffusion models provide an alternative approach that has demonstrated superior performance in various domains. Diffusion models generate data by progressively introducing noise into a dataset and subsequently learning to reverse this process. The goal of the model is to approximate the reverse of this process, using a neural network trained to predict and remove the added noise at each step. By iteratively refining noisy data, the model reconstructs realistic samples from pure noise [19].

There are two processes to understand how the diffusion works. The forward diffusion process and reverse diffusion process are foundational components of diffusion models, which have shown great promise in generative tasks. The forward process involves gradually adding noise to an initial data sample, such as an image, over a series of time steps, until the data is completely corrupted into pure noise. Mathematically, this is represented by a sequence of conditional distributions that progressively degrade the original data, typically modeled as a Markov process [20]. The reverse diffusion process aims to recover the original data from the noise by learning to reverse the corruptive steps of the forward process. This is carried out through a learned function that predicts the noise added at each time step, allowing for the stepwise denoising of the data [19]. By training the model to predict the noise at each step, the reverse process becomes capable of generating high-quality data, making diffusion models highly effective for generative tasks such as image synthesis [21]. More recently, this technique has allowed researchers to develop text-to-speech models for a variety of applications, such as video generation [22] or 3D image generation [23].

In the context of deepfake generation, diffusion models are particularly advantageous due to their ability to learn complex data distributions, resulting in more realistic facial expressions, seamless blending, and improved temporal consistency in video synthesis [24]. Moreover, the iterative refinement process of diffusion models allows for fine-grained control over attributes such as identity, expression, and lighting, which is crucial for high-quality deepfake manipulation [25]. A novel approach is trying to combine the image quality from diffusion models with the speed of GANs, adversarial training, and score distillation—the Adversarial Diffusion Distillation (ADD) [26]. These properties make diffusion-based deepfake models a promising direction for applications in entertainment, virtual reality, and synthetic media, while also raising concerns regarding misinformation and ethical considerations surrounding their use, in particular, frauds in biometric systems for the financial sector.

## 4. Deepfake Detection

Driven by the need to combat the misuse of generative AI in deepfake creation, several state-of-the-art detection methods have emerged. This section explores these approaches and their key contributions.

### 4.1. CNNs

Convolutional neural networks (CNNs) have historically been a core technology in image processing and is one of the most common frameworks for deepfake detection models. CNNs typically distinguish between real and deepfaked images by observing small features consistent with generated content, typically looking for pixel level differences [27].

A particular type of CNNs, CNN-F, leverages the detectable fingerprints left by CNN-based generators. The model is pre-trained with a dataset, then fine-tuned on these fingerprints. Wang et al. analyzed a detection model trained on one specific CNN-generator. It was shown that with proper data augmentation, the model could generalize across many other datasets generated by different architectures [28]. Various post-processing techniques were tested, and data augmentation proved to increase the generalizability of the model.

An alternative method for deepfake detection arises by focusing on inconsistencies in the texture of generative images. The texture of fake faces in deepfaked images generated using GANs is noticeably different from the texture of faces in real images. GramNet, a CNN-based framework outlined in Liu et al., captures inconsistencies in style and texture to identify generated content. This model architecture introduces a "Gram Block" to the CNN backbone which computes a global texture representation [29]. By changing focus to global style features instead of localized inconsistencies, GramNet has improved robustness against image-editing techniques, such as compression, blur, and noise.

### 4.2. Transformers

Vision transformers have also been applied as a framework for identifying deepfake images. In Ghita et al, a proposed ViT model is proposed that leverages the self-attention mechanisms of ViTs to link relationships between regions, or patches, of the image. This allows the model to identify artifacts across patches that are introduced during the deepfake generation. The experimental results show that the ViT model was able to achieve accuracy at or above traditional CNN-based approaches. The study also showed that ViTs are better at generalization when classifying images generated by novel deepfake methods [30]. Since ViTs have global attention mechanisms, they are better able to consolidate information from across the entire image, which improves the robustness of the classifier. This suggests that ViTs could hold promise in building more resilience in a deepfake detection system, which would help improve against newly developing deepfake generation models.

### 4.3. Patch-Based Classifiers

Another method for deepfake detection are patch-based classifiers. This method processes the images by dividing them into smaller regions called patches and analyzing these localized areas of the initial image. The localized focus allows these classifiers to observe minute artifacts from deepfake generation, such as texture mismatch, spatial inconsistency, and unnatural transitions between elements of the image. This allows patch-based models to become highly effective for isolating fine-grained patterns that can be overlooked in global classifiers and can improve their generalizability to new deepfake generation methods.

Patch-Forensics, a patch-based classifier proposed in Chai et al, segments the input images into non-overlapping patches and processes each patch independently through a

truncated backbone CNN model and projects a per patch score. These scores are aggregated to form an overall prediction on the nature of the image [31].

Extensions of these patch-based models seek to combat images generated using diffusion models. Images generated using these models are known for holding increased photorealism and very few visual artifacts, presenting a difficult challenge for deepfake detection. A framework for detecting and localizing manipulated regions of diffusion-generated images is proposed in Tantaru et al. Within their set of three proposed frameworks in Dolos, the study showed that Patches, a weakly supervised training of a modified Patch-Forensics with Xception as the backbone model, showed promise in not only detecting diffusion-generated images but also localizing the manipulated regions within the images. The experimental results showed the framework was also able to generalize effectively across seen and unseen datasets, and outperformed existing detection models when applied to unseen diffusion-based manipulations [32].

### 4.4. GANs

GANs are mainly used to generate high-quality deepfakes, but the discriminator within a GAN can also be a powerful tool for detection. The discriminator of a Deep Convolutional GAN achieves 100% accuracy on detecting fake images from its corresponding generator within 10 iterations, even as the generator produces highly realistic images [15]. However, its accuracy on identifying real images as well as its performance on other datasets vary. Asan et. al tested three GANs for deepfake detection: Steganography-GAN, MRI-GAN, and MDD-GAN [33]. Each model employed the same architecture, but their techniques differed. Steganography-GAN is a preventative model, embedding watermarks into content. These watermarks are difficult for deepfake generators to decipher, leading to a 100% accuracy in deepfake prevention. MRI-GAN uses perceptual differences in images. It was proven to be viable in detecting deepfakes but performed worse than other models. MDD achieved the highest AUD score and was capable of generalizing effectively to new datasets without any model updates. Jheelan and Pudaruth trained several deep learning models to detect deepfake images generated with GANs, and they even developed a web-based interface to upload images and run their detection algorithm [34].

### 4.5. Strategies

Deepfake detection can also be achieved by analyzing continuity between the face and the context surrounding it (e.g., neck, ears, hair). Nirkin et. al proposes a system with two complementary networks: one for detecting features on the segmented face and another for its context. By identifying discrepancies between the face and its context, their model incorporates multiple Xception networks to differentiate fake and real images. This method significantly outperforms the baseline and successfully identifies many fakes that other state-of-the-art models missed [35]. There are ways to overcome this detection model, like generating deepfakes that reconstruct the entire image or head. However, this would require a much broader integration of identity, which would be challenging to bypass a fully developed biometric verification system.

Another emerging strategy for deepfake detection uses other modalities related to an image, mainly text or audio, to provide extra information or context. SIDA makes use of vision–language models, which learn from images and text, allowing them to perform well on tasks that require both modes of communication [36]. By prompting LISA, a vision–language model with good reasoning capabilities, to identify whether an image is a deepfake and which part has been tampered, if any, Huang et al. found consistent high accuracy across real, partially tampered, and completely generated images.

Although the majority of this paper focuses only on image deepfakes, video and audio deepfakes have rapidly improved over the last several years and pose an even greater challenge for biometric verification systems. These multimodal models typically encode the audio data separately from the corresponding video frames, but can then choose to produce independent predictions for the audio and visual data streams [37], or combine the two into one feature set [38].

### 4.6. Model Generalization

Despite many models achieving high accuracy on benchmark deepfake datasets, this research area suffers from a lack of generalizability and diversity. As deepfake generation becomes more powerful and accessible to a global audience, relying on current models poses major security risks for biometric verification systems.

Many deepfake generation applications are based on a handful of popular GANs such as StyleGAN and CycleGAN. However, verification systems cannot ignore the alternative and emerging GANs that are constantly developing. Many detection models are trained and tested on a dataset generated from a limited number of GANs, leading to sharp performance drops when tested on novel GAN variants and diffusion-based methods [39]. Biometric verification systems cannot limit what methods their adversaries use, and in fact should expect to contend with all types of emerging methods. It is unrealistic for companies to rely on one model without putting in considerable effort to train the model on a wide variety of GANs and diffusion models, and retrain as new deepfake generation methods are developed.

Biometric verification systems, particularly in financial systems, must prepare for adversarial attacks as financial fraud and information leaks can be devastating. Malicious actors will likely make repeated attacks to obtain access to bank accounts. If a detection model depends on specific patches of the face or certain fingerprints left by GANs, the verification system is vulnerable to new deepfakes as attackers learn about the model's weaknesses.

## 5. Ensemble Model

Ensemble learning, or multiple classifier systems (MCS), can be understood as a machine learning technique that improves the model by combining multiple methods to enhance recognition accuracy in pattern recognition systems. Ensembles were introduced as a way to improve the generalization ability of individual neural networks [40]. They were implemented as a "mixture of experts" that vote for the final answer and may specialize in different aspects of the problem or different types of inputs [41,42].

Multiple Classifier Systems (MCSs) have emerged as a crucial paradigm in machine learning and pattern recognition, offering a robust approach to improving predictive accuracy. The core principle of MCS lies in leveraging the diversity of multiple base classifiers, combining their strengths to achieve superior performance compared to individual models. Ensemble learning methods, such as bagging, boosting, and stacking, exemplify this strategy by aggregating decisions from multiple classifiers. Bagging reduces variance through bootstrap aggregation, boosting enhances weak learners sequentially, and stacking employs meta-learning to refine predictions. These approaches have been widely studied and applied in various domains, with modern implementations like Random Forest, AdaBoost, and XGBoost demonstrating significant improvements in practical applications. The success of ensemble methods largely depends on the tradeoff between classifier accuracy and diversity, ensuring that base models complement each other's strengths and mitigate individual weaknesses. Consequently, the study and development of MCS con-

tinue to be a vibrant research area, driving advancements in adaptive learning and complex decision-making systems [43,44].

Voting and combining strategies are fundamental to Multiple Classifier Systems (MCSs), determining how individual classifiers contribute to the final decision. The most common approaches include majority voting, weighted voting, and probability-based fusion. In majority voting, each classifier casts a vote, and the class with the most votes is chosen. Weighted voting assigns different importance levels to classifiers based on their performance, giving more influence to stronger models. Probability-based fusion combines the predicted probability distributions from individual classifiers, selecting the class with the highest aggregated confidence. These strategies allow diverse classifiers—such as decision trees, SVMs, and neural networks—to be integrated effectively, leveraging their complementary strengths to improve robustness and accuracy [45].

In our study, we propose an MCS approach that combines individual classifiers trained on different synthetic data generation strategies, specifically designed for fake image detection in fraud prevention. By leveraging the diversity of these artificially created datasets, our method enhances classification performance, capturing complementary patterns while reducing overfitting to a single data source. This approach is particularly well suited for scenarios where multiple algorithms are used to generate fake images, each introducing unique variations and challenges. By integrating classifiers trained on data from different generation techniques, our ensemble method ensures broader coverage of potential fraud patterns. The structured voting and fusion strategy further enhances robustness and generalization, making it effective in complex and adversarial environments.

Figure 1 presents the generative stage for creating deepfake images, followed by the pre-training phase of individual models. Meanwhile, Figure 2 illustrates the composition stage of a diverse dataset, incorporating samples from each training dataset of the individual models to represent a diverse scenario in deepfake generation techniques. Subsequently, it depicts the combination stage, describing the ensemble learning architecture under a learned voting rule. While system performance might seem degraded by the fact of using several models, it is worth mentioning that these models may be executed in parallel [46] so using today's multi-core processors, the execution time will be equivalent to the running just the slowest model alone.

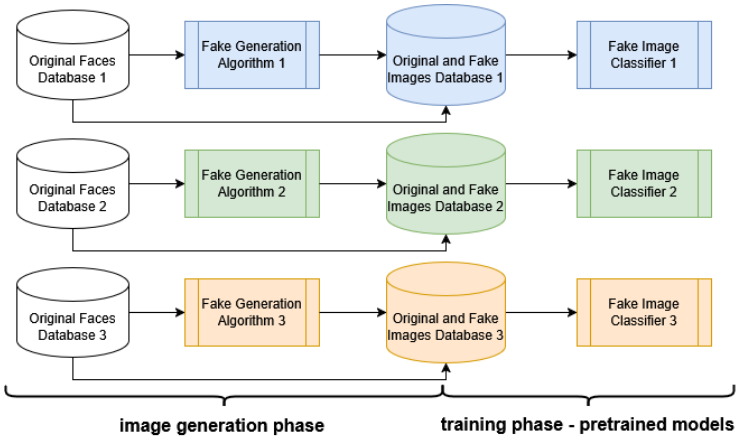



**Figure 1.** A diagram of the pretraining phase of the individual models.

Finally, Figure 3 provides a summarized overview of the proposed ensemble architecture, including the pre-trained models considered in this study for inferring whether an input image is classified as fake or non-fake.

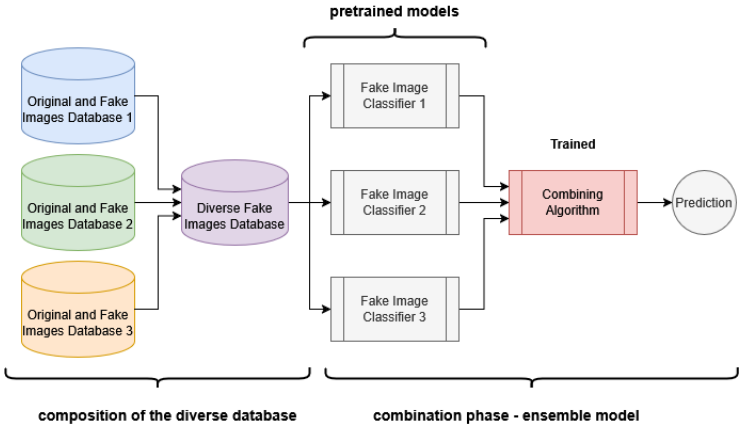



**Figure 2.** An overview of the diverse database composition and ensemble architecture.

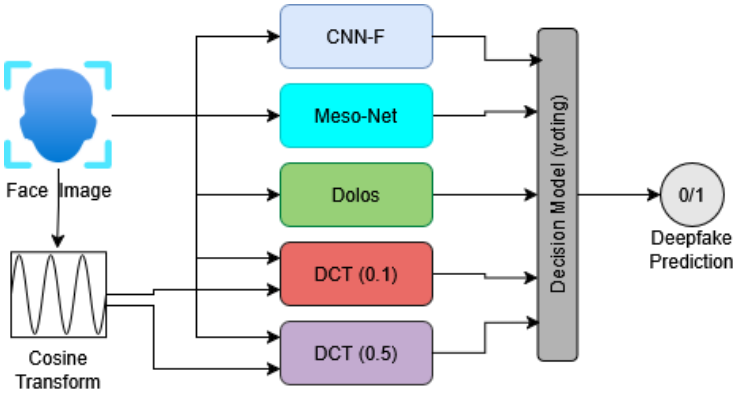



**Figure 3.** A diagram of the proposed ensemble-based approach for robust deepfake image detection.

*Robustness Against Adversarial Attacks*

To mitigate the threat of adversarial attacks, the utilization of ensemble learning and ensemble models has proven to be an effective strategy to increase the robustness of deep learning systems. A prominent technique in this context is ensemble adversarial training, which is particularly beneficial against black-box attacks. The core idea involves training multiple models with distinct architectures or different weight initializations, using adversarial examples generated by diverse attack methodologies.

In the domain of medical image analysis, the advantage of ensemble models has been empirically demonstrated. Studies indicate that custom models integrating multiple CNN architectures exhibit greater resilience against black-box adversarial attacks, outperforming single models in white-box scenarios. Furthermore, implementing an ensemble training approach, where each CNN architecture is trained with multiple initializations, results in a significant reduction in attack accuracy [47,48].

The fundamental principle behind ensemble adversarial training lies in its ability to neutralize the vulnerability of individual models to specific attack types, especially in black-box settings where the attacker has limited knowledge of the target model. By training neural networks with adversarial samples crafted from various attack methods, such as FGSM and PGD, the ensemble benefits from a broader spectrum of training data. This enhanced diversity in the training set improves generalization capability and overall robustness against unseen adversarial examples. In the pursuit of robust medical diagnostic systems, adversarially pre-training multiple models with diverse architectures, followed by averaging their output probabilities, emerges as a promising strategy [49].

While the primary focus is on employing ensembles for defense, it is relevant to note that ensemble-based attacks have shown success in overcoming transferability challenges between

different architectures, such as CNNs and vision transformers. This indirectly underscores the potential of combining multiple models for both offensive and defensive purposes. In summary, evidence suggests that ensemble learning, particularly through the adversarial training of diverse model architectures, offers a promising path towards developing more resilient deep learning systems in critical domains like medical image analysis [47–49].

## 6. Benchmark Datasets

To evaluate the performance of the various state-of-the-art detection models, we selected six benchmark datasets: DeepFakeFace [DFF], FaceForensics++ [FF++], Individualized Deepfake Detection Dataset [IDDD], ProGAN, StarGAN, and WhichFace. Together, these datasets cover three categories of image generation: diffusion models, face-swap models, and GAN models. The details of each dataset, including specific generation methods, year, and size, are presented in Table 1.

**Table 1.** An overview of the datasets used in this study and how many real and deepfake images were included in our combined dataset.

| Dataset | Generation Methods | Year | Real | Fake |
|---|---|---|---|---|
| DeepFakeFace | Stable Diffusion v1.5<br>Stable Diffusion Inpainting<br>InsightFace | 2023 | 1500 | 1500 |
| FaceForensics++ | Deepfakes<br>Face2Face<br>FaceSwap<br>NeuralTextures | 2018 | 2707 | 2698 |
| Individual | Faceswap-GAN | 2021 | 1500 | 1500 |
| ProGAN | Progressive GAN | 2017 | 200 | 200 |
| StarGAN | Star GAN | 2018 | 2000 | 2000 |
| WhichFace | Style GAN | 2019 | 1000 | 1000 |

Each dataset contains images of faces cropped to include minimal background information. All samples are either directly downloaded from open-source repositories or publicly available. These datasets are used to test pre-trained checkpoints of publically available models. This opens up flexibility for dataset construction, as the models are not retrained on the training splits of the data. However, it is important to have each dataset balanced between real and fake classes so as not to bias the evaluation of the model predictions.

### *6.1. DeepFakeFace*

The DeepFakeFace dataset consists of computer-generated images of celebrities using a range of different deepfake generation techniques. Such methods include Stable Diffusion v1.5, Stable Diffusion Inpainting, and InsightFace. The dataset consists of 120,000 different images, made up of 30,000 real images from the IMDB Wiki dataset, and 30,000 fake images generated using the three techniques described above. All images that are part of this dataset have a resolution of 512 by 512. DeepFakeFace was loaded from Hugging Face and re-partitioned in testing sets with 1500 real and 1500 fake images [50].

### *6.2. Individualized Deepfake Detection Dataset*

The Individualized Deepfake Detection Dataset was created using Faceswap-GAN. Unlike existing deepfake detection datasets such as FaceForensics++, IDDD allows us to assess the performance of detection models on images of specific individuals. It consists of images of 45 specific individuals, with a total of 23,000 authentic images and 22,000 deepfake

images. These images are also divided into a training dataset, collected from the CelebDFv2 dataset, and testing dataset, composed of images from the CACD dataset.

### *6.3. Face Forensics++, ProGAN, StarGAN, and WhichFace*

All four of these datasets were downloaded from Hugging Face and were used in the training of the CNN-Fingerprints model [28].

The Face Forensics++ dataset is a partitioned version of the challenge dataset described in Face Forensics [51]. This dataset is composed of images generated from 1000 original video sequences. The videos were generated using four different face manipulation methods: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. The images are all cropped to contain a frontal view of the face without occlusions.

The ProGAN and StarGAN datasets were generated by running the pre-trained ProGAN and StarGAN models on images from CelebA [52] and generate real and fake image pairs. These images are then cropped on the long edge (center crop length is exactly the length of the short edge) and then resized to 256 × 256.

WhichFace includes 1k real face and 1k fake faces images scraped from whichfaceisreal.com. This website holds both real images as well as Style-GAN generated faces. These images are downloaded, compressed in JPEG and resized from 1024 × 1024 to 256 × 256.

## 7. Individual Models

This paper seeks to explore how methods of constructing combined or ensemble methods can impact the generalizability of deep fake detection classifiers. In order to understand what models constructively fit in an ensemble method, individual pre-trained models were first individually evaluated on deepfake detection tasks prior to constructing and testing the multi-classifier models. This combined model relies on a collection of individual classifiers that serve as the 'voters' in the ensemble. This paper explored five models identified from deepfake detection papers.

As described in Section 8, we do not retrain any of the following models on our chosen benchmark datasets. Instead, we use pre-trained checkpoints from the models' authors to evaluate their performance. The deepfake generation methods that each model checkpoint was trained on are listed in Table 2 below.

**Table 2.** Overview of models used in this study.

| Model Name | Year | Deepfake Generation | Methods |
|---|---|---|---|
| CNN-F [28] | 2020 | GAN | ProGAN |
| MesoNet [53] | 2018 | VAE | Deepfake<br>Face2Face |
| DCT [39] | 2022 | GAN | ProGAN, StyleGAN, ProjectedGAN<br>Diff-StyleGAN2, Diff-ProjectedGAN |
| | | Diffusion | DDPM, IDDPM, ADM, PNDM, LDM |
| Dolos [32] | 2024 | Diffusion | Repaint–P2, Repaint–LDM<br>LaMa, Pluralistic GAN |

### *7.1. CNN-Fingerprints*

CNN-fingerprints describes a specialized CNN designed to identify common artifacts present across GAN generated synthetic images [28]. The architecture of this model seeks to extract fine-grained spatial features by defining carefully chosen kernel sizes, strides, and paddings in the convolutional layers. This allows CNN-F to specifically capture subtle statistical irregularities and artifacts inherent in synthesized images, which are preserved during the pooling layers.

CNN-F is a pivotal part of the ensemble model, as it focuses on the more minute artifacts rather than overall textural elements or context-based detection. CNN-F also shows decent generalization itself, as it focuses on smaller artifacts present across GAN-based deepfake generation methods.

### 7.2. MesoNet

The second pre-trained model explored for the ensemble model is MesoNet, as introduced in "MesoNet: a Compact Facial Video Forgery Detection Network" [53]. MesoNet is a compact deep neural network that was specifically designed to detect facial video forgeries. This model analyzes each frame of a video individually, so the model architecture is still applicable for photo-based deepfake detection. MesoNet focuses on intermediate properties present in facial images by using more largely selected kernel and pooling sizes. The model was trained with the goal of capturing facial specific digital manipulation, which enhances the model's application in identify verification and images forgery detection.

Within the ensemble model, MesoNet plays a critical role as it can efficiently process and extract relevant features from the facial regions in the input image, ultimately enhancing the overall robustness and accuracy of the ensemble model.

### 7.3. Discrete Cosine Transforms

The DCT-based approach [39] uses frequency-domain analysis to identify artifacts created by diffusion-based image generation. The method begins by applying a Discrete Cosine Transform (DCT) to the input image. This transformation allows the model to pickup on artifacts in the frequency domain that are ingrained during the denoising process of diffusion models. By identifying these frequency discrepancies, the DCT + classification approach can successfully identify diffusion-based image generation. Within our ensemble framework for deepfake detection, the DCT-based module enhances our ensemble's robustness by adding a frequency-focused perspective to the detection process, thereby improving overall detection accuracy against diffusion model deepfakes.

### 7.4. Dolos

We evaluate the "Patches" architecture category described by the authors of Dolos [32], which is based on the Patch-Forensics model [31]. This localization model truncates the Resnet and Xception models to obtain binary deepfake predictions on small regions of an image. All the binary patch classifications are averaged to produce a final classification for the image as a whole. The authors of Dolos retrain the Patch-Forensics model with diffusion-generated deepfake images and experiment with three setups that provide the model with varying amounts of information. We choose Setup B, where images are partially manipulated but only classified with real or fake, with no further localization details. This setup is closest to our context of biometric verification systems: most deepfake attacks manipulate the face without changing the image background, and any detection system should classify images as fake if any part has been manipulated.

Including Dolos in our ensemble model introduces a new deepfake detection strategy that takes advantage of differences between deepfaked image portions and the background, which is often not manipulated.

## 8. Individual Model Evaluation

For each model we evaluated, we used the publicly available code and pretrained weights that the original researchers provided, with special care to finding checkpoints loaded from fine-tuning on facial datasets. Since this paper focuses on improving classification for unseen data, we chose to load pre-trained models that any security verification developer would be able to access without re-training or fine-tuning.

Each model was thresholded at 0.5, with 1 indicating a deepfake image and 0 indicating a real image. As these models are being evaluated for their contribution to a voter ensemble model, the quality of the positively tagged images is of higher concern (as a voter with a high false-positive rate would add more noise to the ensemble system). To address this, the ROC curves for each model and precision score were further examined to understand how the models behave. The ROC curves also provide insight into potentially stronger thresholds for the individual models. As the ensemble model will provide a weight to each voter output, it does have the capability to learn a stronger threshold for each sub-classifier. By observing the ROC curve, the maximum precision of the model can be observed based on the true-positive vs false-positive rate.

The six datasets represent a coverage of different concerns with image fraud. DeepFakeFace contains images generated using Stable Diffusion models; FaceForensics and individual contain images generated using Faceswap models; and ProGAN, StarGAN, and WhichFace contain images generated using GANs. Results are shown in Table 3.

**Table 3.** Individual model results for various datasets with a 0.5 threshold. Metrics include accuracy (Acc) split into total accuracy, deepfake accuracy, and real accuracy; precision (Prec); recall (Rec); and area under the ROC curve (AUC). The DCT models' values refer to the quality factor of the pre-processing compression. Bold indicates the highest score on each dataset for some metrics. * Indicates high performance as the benchmark dataset was in the model's training data.

| Model | Dataset | Accuracy | | | Metrics | | |
|---|---|---|---|---|---|---|---|
| | | Total | Fake | Real | Prec | Rec | AUC |
| CNN-F | DeepFakeFace | 52.9 | 70.8 | 35.1 | 52.2 | 70.8 | 53.1 |
| | FaceForensics | 56.7 | 60.9 | 52.6 | 56.2 | 60.9 | 58.9 |
| | Individual | 51.8 | 69.7 | 34.0 | 51.2 | 69.7 | 51.7 |
| | ProGAN | 55.3 | 52.0 | 58.5 | 55.6 | 52.0 | 53.7 |
| | StarGAN | 50.0 | 100.0 | 0.0 | 50.0 | 100.0 | 43.0 |
| | WhichFace | 50.4 | 2.9 | 97.9 | 58.0 | 2.9 | 45.9 |
| | Average | 51.5 | 51.7 | 51.2 | 51.4 | 51.7 | 51.5 |
| MesoNet | DeepFakeFace | 53.2 | 27.8 | 78.5 | 56.4 | 27.8 | 54.8 |
| | FaceForensics | 68.7 | 74.8 | 62.5 | 66.6 | 74.8 | 74.6 |
| | Individual | 61.1 | 64.7 | 57.4 | 60.2 | 64.7 | 64.7 |
| | ProGAN | 53.5 | 21.5 | 85.5 | 59.7 | 21.5 | 49.7 |
| | StarGAN | 51.8 | 67.0 | 36.7 | 51.4 | 67.0 | 50.6 |
| | WhichFace | 51.4 | 68.3 | 34.5 | 51.0 | 68.3 | 50.7 |
| | Average | 57.3 | 39.8 | 74.7 | 61.1 | 39.8 | 57.3 |
| DCT (0.1) | DeepFakeFace | 54.3 | 10.9 | 97.7 | 82.8 | 10.9 | 77.8 |
| | FaceForensics | 67.5 | 42.4 | 92.5 | 88.9 | 44.4 | 85.2 |
| | Individual | 65.6 | 43.4 | 81.7 | 70.2 | 43.4 | 69.3 |
| | ProGAN * | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| | StarGAN | 95.2 | 93.2 | 97.2 | 97.1 | 93.2 | 98.8 |
| | WhichFace | 98.8 | 98.5 | 99.0 | 99.0 | 98.5 | 99.9 |
| | Average | 72.2 | 58.3 | 86.1 | 80.7 | 58.3 | 77.8 |
| DCT (0.5) | DeepFakeFace | 51.0 | 3.1 | 99.5 | 83.0 | 2.6 | 77.1 |
| | FaceForensics | 54.6 | 14.8 | 94.4 | 72.3 | 14.8 | 64.4 |
| | Individual | 56.9 | 23.3 | 90.3 | 70.6 | 23.3 | 65.8 |
| | ProGAN* | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| | StarGAN | 91.7 | 89.4 | 94.0 | 93.7 | 89.4 | 97.1 |
| | WhichFace | 99.3 | 99.2 | 99.4 | 99.4 | 99.2 | 100.0 |
| | Average | 68.6 | 42.8 | 94.3 | 88.2 | 42.8 | 76.9 |
| Dolos | DeepFakeFace | 68.1 | 86.7 | 49.4 | 63.2 | 86.7 | 71.3 |
| | FaceForensics | 50.1 | 0.0 | 99.9 | 0.5 | 0.0 | 19.8 |
| | Individual | 50.8 | 35.1 | 66.4 | 50.9 | 35.1 | 50.5 |
| | ProGAN | 50.0 | 0.0 | 100.0 | 50.0 | 100.0 | 27.4 |
| | StarGAN | 50.0 | 0.0 | 100.0 | 50.0 | 100.0 | 0.9 |
| | WhichFace | 50.0 | 100.0 | 0.0 | 0.0 | 0.0 | 34.1 |
| | Average | 52.9 | 14.9 | 90.7 | 61.6 | 14.9 | 23.9 |

As expected, we see the performance of the models varies across the types of generation methods. For example, we note that the DCT models performed especially strongly on the GAN-generated images, while other models did not display confidence in classifying similar images. This section further explores the results of the individual models across three categories of images: diffusion, VAEs, and GANs.

In addition to comparing model performance for each collection of datasets, this section also identifies combinations of models that would provide insight into developing heuristics for ensemble model formation. The data was partitioned based on the method of generation used by the datasets. Each model's ROC curve was examined to understand the quality of the 'votes' that the models provide. As mentioned earlier, the weighting process of the ensemble model provides an opportunity for the individual model thresholds to be rebalanced. This means that an individual model's performance in the ensemble is likely higher than against a 0.5 threshold.

### *8.1. Diffusion*

The DeepFakeFaces dataset serves as our collection of images for diffusion models, using Stable Diffusion v1.5 and Inpainting to generate samples. Figure 4 displays the ROC curves on this dataset for each of the five individual models from Table 3. For this class of images, Dolos was the only model which showed total accuracy significantly better than a random classifier. That being said, it was noted that Dolos tends to over-classify false positives, with a lower precision metric of 63.2%. In the context of biometric verification, having a higher-than-normal false-positive rate is not inherently a negative, as the samples that are flagged as fraudulent can be passed on for secondary or manual verification. Adding the Dolos model with its 'over eager voting' could help to identity how accurate models with lower precision impact ensemble results.

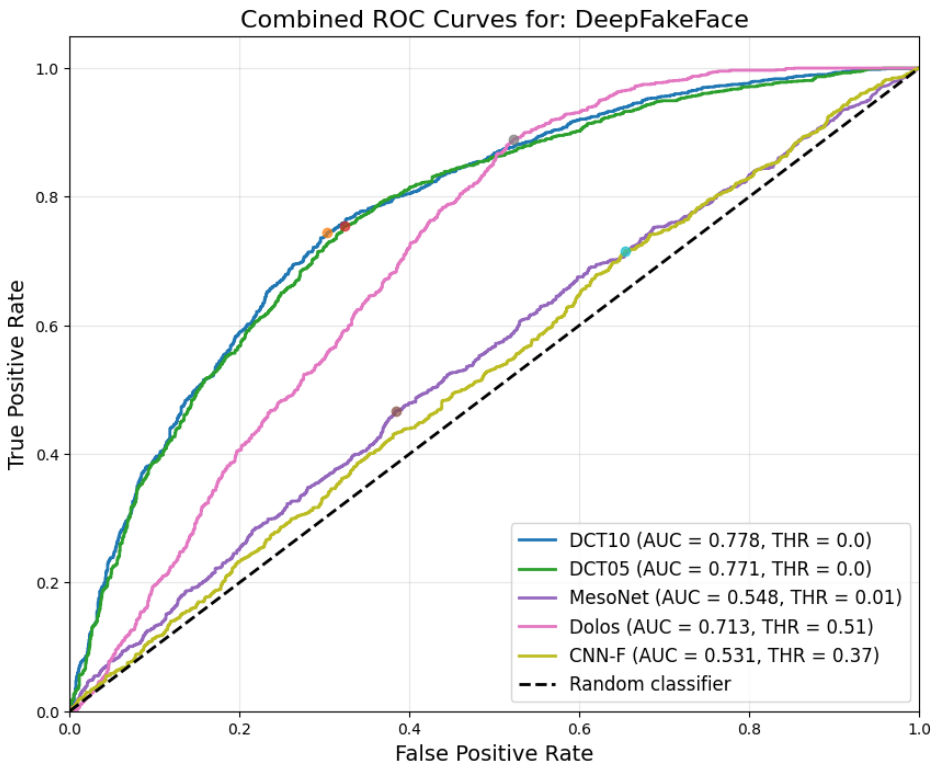
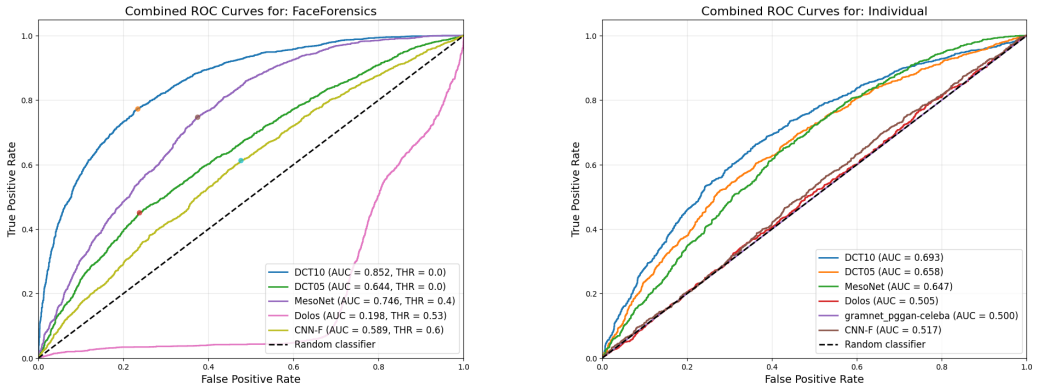



**Figure 4.** ROC curves of each model over diffusion generated images in DeepFakeFace. The point on each curve is placed at the spot corresponding to the threshold with the maximum J-Score. The color of these points has no special meaning.

Looking at the ROC curves of each model, one can see that the Dolos model is not the only model that shows a positive performance. Although DCT models only slightly improve accuracy compared to a random classifier, they show a starkly different ROC curve that suggests successful classification. When coupled with a high precision for DCT (0.1)—83.0—it is clear that while the DCT model is likely to miss a good portion of the diffusion-generated images, the images that it flags as fraudulent are not likely to be false positives. Intuitively, high precision should suggest that a model is a good candidate for the ensemble strategy, as it is unlikely to provide a 'vote' for an image that is not deepfaked. This potential heuristic is further explored in Section 10.

### *8.2. FaceSwap*

The results based on face-swap datasets—FaceForensics and Individual—present a similar tradeoff between accuracy and precision. MesoNet and DCT both showed high total accuracy—between 61% and 69%—though DCT had a higher precision—88.9% and 70.6% on FaceForensics and Individual, respectively. Figure 5 shows the ROC curves of the models for the dataset based on face-swap algorithms.



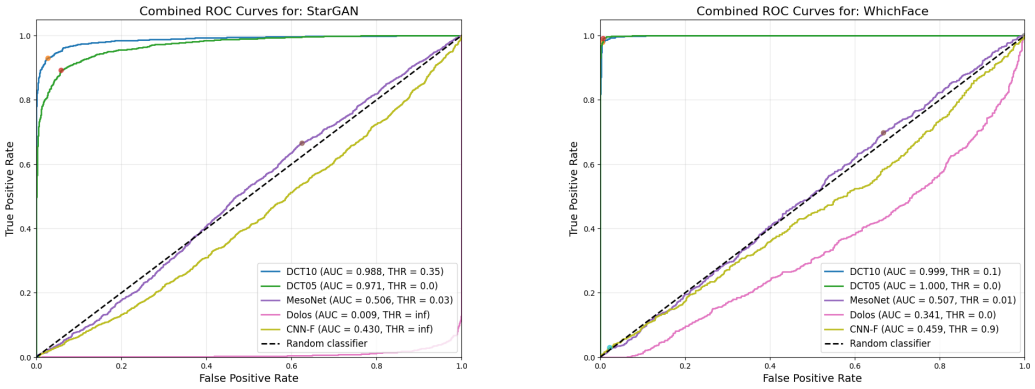

(**a**) ROC curve on FaceForensics (**b**) ROC curve on Individual

**Figure 5.** ROC curves of each model over face-swapped images. The point on each curve is placed at the spot corresponding to the threshold with the maximum J-Score. The color of these points has no special meaning.

While MesoNet had higher total accuracy, the AUC for DCT (0.1) was the largest for both FaceForensics and Individual, at 85.2% and 69.3%, respectively. This high AUC and high precision again suggests that DCT could prove to be a more successful member of the ensemble, which is explored further in Section 10 as well.

### *8.3. GANs*

The DCT models were by far the highest performers in the GAN-generated datasets as can be seen in the total accuracy, precision, and ROC curves. This suggests that our ensemble models should be constructed with DCT models as the primary members and the additional models to add robustness to VAE- and diffusion-generated input samples. Figure 6 shows the ROC curves of the models for the datasets based on GAN generated images.




(**a**) ROC curve on StarGAN (**b**) ROC curve on WhichFace

**Figure 6.** ROC curves of each model over GAN-generated images. The point on each curve is placed at the spot corresponding to the threshold with the maximum J-Score. The color of these points has no special meaning.

## 9. Ensemble Model Architecture

We apply the ensemble methodology to deepfake image detection to gain robustness against adversarial attacks and a larger variety of generative methods. As explored in Section 8, individual detection models can achieve high performance on a specific dataset or generative method, but do not carry this level of performance to unseen datasets. We aim to combine the strengths of individual models that perform well on one generative category: Dolos for diffusion-generated deepfakes, Mesonet for deepfakes created through face-swap methods, and DCT (0.1) and DCT (0.5) for GAN-generated deepfakes.

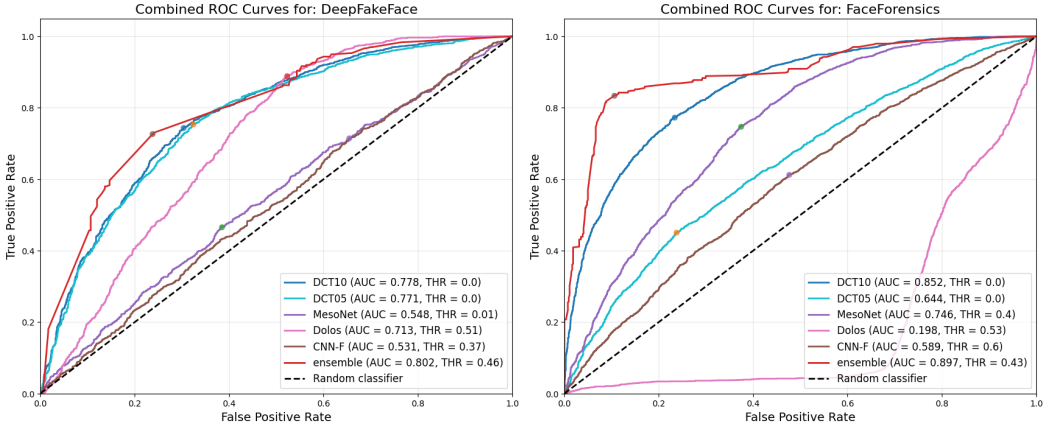
Each combination of these four models was combined into an ensemble model using a random forest classifier. The classifier is composed of a number of decision trees based on Gini impurity, each trained on a different subset of the data. The maximum depth, minimum samples to split and the number of estimators were chosen to be three, five, and nine by parameter search with successive halving. Each decision tree returns a probability between 0 and 1 for the image being a deepfake and the image being real. These results are averaged, and the classifier returns real or fake based on which class has the higher average.

We provided only the individual model predictions, not the images themselves, as input to our ensemble model. Each ensemble model was trained on predictions from 80% of each of our six datasets and tested on the remaining 20%, with 5-fold cross validation. In the following sections, we report the mean and standard deviation for the metrics across the five folds.

## 10. Ensemble Model Evaluation

In this section, we present the results of different ensemble approaches. Table 4 shows the average accuracy, precision, recall, and area under ROC curve (AUC) of exhaustive ensemble combinations between four models: DCT (0.1), DCT (0.5), Dolos, and Mesonet. We find that the combination of all four models produces the best accuracy and recall, without significant drawbacks in precision or standard deviation in any metric. Combining only DCT (0.5) and Dolos produces the highest precision of 87.7%, but with lower accuracy and recall.

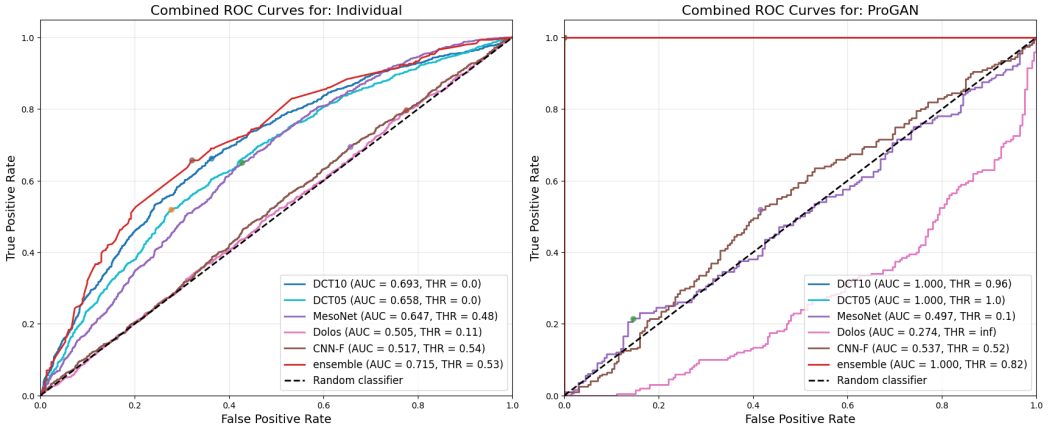
In Figure 7, we have added the ROC curves of the DCT (0.1), DCT (0.5), Dolos, and Mesonet ensemble for each dataset. These are shown in red along with the five individual models we evaluated in Section 8. For all datasets and deepfake generation methods, the four-model ensemble matches the best single model.




(**a**) DeepFakeFace



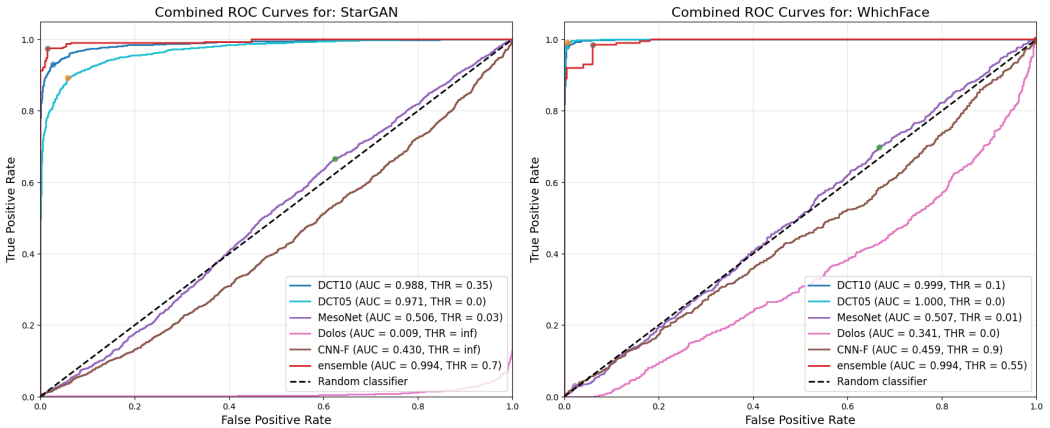

(**b**) FaceForensics




(**c**) Individual




(**d**) ProGAN




(**e**) StyleGAN




(**f**) WhichFace

**Figure 7.** ROC curves (in red) for an ensemble that combines all four submodels, compared against the ROC curves of individual submodels for each dataset. Each of the six subfigures contain results for a specific dataset, listed ini the subfigure caption.

**Table 4.** Performance comparison of different ensemble models. The mean and standard deviation across five runs are reported.

| Ensemble Model Combination | Accuracy | Precision | Recall | AUC |
|---|---|---|---|---|
| DCT (0.1) & DCT (0.5) | 72.1 ± 0.8 | 83.1 ± 2.5 | 61.2 ± 3.5 | 80.1 ± 1.1 |
| DCT (0.1) & Dolos | 75.5 ± 0.8 | 79.5 ± 4.9 | 71.5 ± 3.5 | 85.3 ± 0.9 |
| DCT (0.1) & Mesonet | 76.3 ± 0.2 | 84.7 ± 0.8 | 68.0 ± 1.0 | 82.9 ± 1.1 |
| DCT (0.5) & Dolos | 74.0 ± 0.4 | **87.7 ± 2.4** | 60.1 ± 2.4 | 82.8 ± 1.3 |
| DCT (0.5) & Mesonet | 72.7 ± 0.8 | 78.2 ± 3.3 | 67.2 ± 2.6 | 80.0 ± 0.7 |
| Mesonet & Dolos | 72.6 ± 0.8 | 77.0 ± 3.4 | 68.2 ± 4.7 | 80.4 ± 1.2 |
| DCT (0.1) & DCT (0.5) & Dolos | 74.4 ± 1.8 | 84.2 ± 1.8 | 64.6 ± 4.4 | 84.2 ± 0.7 |
| DCT (0.1) & DCT (0.5) & Mesonet | 75.4 ± 0.8 | 85.5 ± 0.7 | 65.3 ± 2.2 | 82.3 ± 0.9 |
| DCT (0.1) & Dolos & Mesonet | 78.0 ± 1.9 | 86.9 ± 3.0 | 69.2 ± 3.2 | **87.8 ± 1.3** |
| DCT (0.5) & Dolos & Mesonet | 77.1 ± 1.7 | 84.0 ± 3.9 | 70.1 ± 5.6 | 85.2 ± 1.3 |
| DCT (0.1) & DCT (0.5) & Dolos & Mesonet | **79.2 ± 1.5** | 83.4 ± 2.1 | **73.0 ± 2.0** | 87.4 ± 0.9 |

### *10.1. Combining Similar Models*

We explore the combination of the two versions of the DCT model, DCT (0.1) and DCT (0.5), as an example of whether two models that detect similar types of deepfakes can be combined into an improved ensemble. Table 5 presents the accuracy, precision, and recall of each individual model and their combination in all six datasets.

**Table 5.** Individual and ensemble model results split across various datasets. Metrics include accuracy (Acc); precision (Prec); and recall (Rec). The DCT models' values refer to the quality factor of the pre-processing compression. A metric is bolded if it is the highest across DCT(0.1), DCT(0.5), and the ensemble DCT (0.1) & DCT (0.5).

| Model | Dataset | Metrics | | |
|---|---|---|---|---|
| | | Acc | Prec | Rec |
| DCT (0.1) | DeepFakeFace | 54.3 | 82.8 | 10.9 |
| | FaceForensics | 67.5 | **88.9** | 44.4 |
| | Individual | **65.6** | 70.2 | 43.4 |
| | ProGAN | **100.0** | **100.0** | **100.0** |
| | StarGAN | **95.2** | **97.1** | 93.2 |
| | WhichFace | 98.8 | 99.0 | 99.0 |
| | Average | **72.2** | 80.7 | 58.3 |
| DCT (0.5) | DeepFakeFace | 51.0 | **83.0** | 2.6 |
| | FaceForensics | 54.6 | 72.3 | 14.8 |
| | Individual | 56.9 | **70.6** | 23.3 |
| | ProGAN | **100.0** | **100.0** | **100.0** |
| | StarGAN | 91.7 | 93.7 | 89.4 |
| | WhichFace | **99.3** | **99.4** | 99.2 |
| | Average | 68.6 | **88.2** | 42.8 |
| DCT (0.1) & DCT (0.5) | DeepFakeFace | **54.9 ± 0.2** | 81.7 ± 4.6 | **12.7 ± 1.1** |
| | FaceForensics | **69.3 ± 3.6** | 81.6 ± 2.9 | **49.7 ± 8.5** |
| | Individual | 63.2 ± 2.1 | 68.3 ± 2.0 | **48.8 ± 5.3** |
| | ProGAN | 99.3 ± 1.1 | 98.6 ± 2.1 | **100.0 ± 0.0** |
| | StarGAN | 83.4 ± 3.0 | 75.6 ± 3.2 | **98.9 ± 0.4** |
| | WhichFace | 91.1 ± 1.6 | 84.9 ± 2.2 | **100.0 ± 0.0** |
| | Average | 72.1 ± 0.8 | 83.1 ± 2.5 | **61.2 ± 3.5** |

The ensemble model produces slightly higher accuracy on two datasets that DCT (0.1) and DCT (0.5) did not perform well on: DeepFakeFace and FaceForensics. It also shows reduced performance on two GAN datasets that the individual DCT models performed

very well on: StarGAN and WhichFace. This leads to the ensemble of the two DCT models being slightly worse than DCT (0.1).

However, the combination of these two models increases recall across all datasets. Although DCT (0.5) performs significantly worse than DCT (0.1) on most of the six datasets, with ProGAN and WhichFace as the exceptions, the ensemble results suggest that DCT (0.5)'s deepfake predictions help catch deepfake images that DCT (0.1) misses. Recall is particularly important for biometric verification systems, as fewer false negatives mean fewer malicious actors that are able to gain access to systems using deepfakes. The performance of the DCT (0.1) and DCT (0.5) ensemble is particularly encouraging as it shows a strategy to improve security without developing new model architectures. More analysis is needed to determine what features DCT (0.5) may have learned that DCT (0.1) did not, as well as how the two individual models' predictions differ. Exploring other instances of ensembles constructed from similar models and generalizing this into a strategy for creating an effective ensemble given a specific model type are valuable directions for future work.

### *10.2. Combining Different Models*

We next explore an ensemble of two models that achieve their best performance on different types of datasets. DCT (0.1) performs very well on GAN-generated deepfakes, moderately well on the face-swap datasets, and at random on diffusion-generated deepfakes. Dolos performs moderately well on the diffusion-generated deepfakes but at random on all other datasets, suggesting that the model cannot adapt to new generative methods for deepfakes and is limited to detecting one type of deepfake. We show accuracy, precision, and recall for these two individual models as well as their ensemble in Table 6.

**Table 6.** Individual and ensemble model results split across various datasets. Metrics include accuracy (Acc); precision (Prec); and recall (Rec). The DCT models' values refer to the quality factor of the pre-processing compression. A metric is bolded if it is the highest across DCT(0.1), Dolos, and the ensemble DCT (0.1) & Dolos.

| **Model** | **Dataset** | **Metrics** | | |
|---|---|---|---|---|
| | | **Acc** | **Prec** | **Rec** |
| DCT (0.1) | DeepFakeFace | 54.3 | **82.8** | 10.9 |
| | FaceForensics | 67.5 | 88.9 | 44.4 |
| | Individual | **65.6** | **70.2** | 43.4 |
| | ProGAN | **100.0** | **100.0** | 100.0 |
| | StarGAN | **95.2** | **97.1** | 93.2 |
| | WhichFace | **98.8** | **99.0** | 99.0 |
| | Average | **72.2** | 80.7 | 58.3 |
| Dolos | DeepFakeFace | **68.1** | 63.2 | **86.7** |
| | FaceForensics | 50.1 | 50.0 | 4.0 |
| | Individual | 50.8 | 50.9 | 35.1 |
| | ProGAN | 50.0 | 50.0 | 100.0 |
| | StarGAN | 50.0 | 50.0 | 100.0 |
| | WhichFace | 50.0 | 0.0 | 0.0 |
| | Overall | 52.9 | 61.6 | 14.9 |
| DCT (0.1) & Dolos | DeepFakeFace | **71.0 ± 4.3** | 78.2 ± 9.0 | 62.3 ± 17.4 |
| | FaceForensics | **74.2 ± 2.2** | **90.7 ± 3.8** | **54.0 ± 5.5** |
| | Individual | 64.2 ± 3.0 | 68.4 ± 4.8 | **52.9 ± 3.1** |
| | ProGAN | 98.3 ± 1.9 | 96.7 ± 3.4 | 100.0 ± 0.0 |
| | StarGAN | 81.2 ± 1.9 | 72.9 ± 2.0 | 99.3 ± 0.7 |
| | WhichFace | 87.0 ± 3.8 | 79.8 ± 4.5 | 99.5 ± 1.1 |
| | Overall | **75.5 ± 0.8** | 79.5 ± 4.9 | **71.5 ± 3.5** |

The DCT (0.1) and Dolos ensemble improves performance on the DeepFakeFace and FaceForensics datasets compared to the DCT (0.1) model. We see a drop in GAN-generated deepfake detection, most notably in the StarGAN and WhichFace datasets, but the ensemble model still shows significant improvement in overall accuracy and recall. As in the previous DCT (0.1) and DCT (0.5) ensemble, this DCT (0.1) and Dolos ensemble produces a much higher recall percentage, with recall improving across all datasets compared to the individual DCT (0.1) model despite accuracy decreases in the GAN-generated datasets. As recall is the more critical metric for blocking fraud in verification systems, this ensemble example highlights that we do not have to sacrifice performance in DCT (0.1)'s strengths to improve its accuracy in diffusion-generated deepfakes. We can combine two models with different strengths to produce a more robust defense system, even if the individual models generalize poorly.

## 11. Discussion

Several metrics can be used to evaluate the quality of a classifier, in this case the ability to determine if a picture is real or was generated by AI tools. As a general metric, the accuracy provides an overall level of quality, indicating the percentage of correct classification. However, for the specific application of second-factor authentication, it is more acceptable to obtain false positives than false negatives, meaning that it is more desirable to reject access to a real client (and ask for another validation) than to grant access to an impostor. Keeping this in mind, the recall metric penalizes the false positives more and becomes a priority in our consideration of deepfake detection systems. All ensemble model combinations improve on overall accuracy compared to the individual models evaluated in Table 3 and make significant gains in recall.

However, adding more models to an ensemble can risk losing performance for specific types of deepfakes. Table 6 shows that although the DCT (0.1) and Dolos ensemble achieves higher overall accuracy compared to the DCT (0.1) model, its performance on the StarGAN and WhichFace datasets is significantly lower. Selecting another model that is trained on diffusion-generated deepfakes and is more capable of discriminating between GAN-generated and real images may avoid the large drops in accuracy and precision. This tradeoff may also depend on the makeup of deepfake attacks for a specific system, and which types of deepfakes are most common. For instance, GAN-generated images may constitute a very large majority of deepfakes received by a verification system due to the accessibility of GANs. If diffusion-generated deepfakes are rarely received, we would likely prefer to use just DCT (0.1) instead of DCT (0.1) and Dolos. There are many similar cases where lower overall performance is accepted, in exchange for peak performance on a specific deepfake type that will protect more users. One important characteristic of the ensemble strategy it that it can easily be adapted to be more effective against any particular technique being actively used by the attackers.

One additional consideration when using an ensemble approach is how to avoid overfitting on training data. The random forest classifier reduces this phenomena by training each of its decision trees on different subsets of the ensemble training data. We can also tune parameters like the number of trees and the minimum number of data samples in a leaf to prevent individual decision trees from overfitting to outliers. However, the risk of overfitting and the best parameters to use may vary depending on the system and the quality of deepfake attacks it receives.

### 11.1. Guidelines for Ensemble Construction

In Section 10.2, we explored how combining two models that perform well on different deepfake generation methods produces a better ensemble, increasing accuracy by more than 3% compared to combining two similar individual models. This leads us to one general guideline when selecting individual models for an ensemble: choose models with different strengths and avoid overlap in model architecture or training data. It is important that all deepfake generation methods, especially popular ones that are likely to be accessed by malicious actors, are represented by at least one individual model that can contribute strong predictions to the ensemble.

In addition to a model's strong performance on at least one type of deepfake, we must also consider its performance on other deepfake types. From Table 3, we see that both the Dolos and MesoNet models perform well on one deepfake type and perform at chance on the other two types. However, Dolos only predicts one class for four out of six datasets. MesoNet is better able to discriminate between deepfake and real for all datasets, even though its overall accuracy on diffusion- and GAN-generated deepfakes is still 50%. MesoNet's stronger performance across all deepfake types makes it a better candidate for including in ensemble models. We see from Table 4 that the DCT (0.1) and MesoNet ensemble slightly outperforms the DCT (0.1) and Dolos ensemble.

### 11.2. Implications for Biometric Security

Our findings reinforce that single deepfake classifiers that can handle a variety of threats are not readily available. Each of the detection models we evaluated had a distinct generation technique that they performed significantly better on face-swapped deepfakes for MesoNet, GAN-generated deepfakes for DCT, and diffusion-generated deepfakes for Dolos. Classifiers that excel when presented with one method of synthetic data lose performance when confronted with diverse generation techniques, and predict at random on novel datasets. As such, the use of deepfake image classifiers in identify verification applications often struggles against adversarial threats using different methods of image or video generation, and lacks the generalizability required for deployment across secure contexts.

The results of combining submodel architectures through a random forest classifier shows promise in improving the generalizability of a deepfake detection classifier. All ensemble model combinations outperform the individual submodels. In particular, combining two models that work well on different deepfake generation techniques, like diffusion and GANs in Section 8.2, creates a well-rounded model that performs well on both techniques without increasing recall.

From a practical point of view, real-world biometric systems must account for both security and usability. Overly permissive models risk admitting malicious deepfakes, while excessively conservative models degrade user experience by frequently misclassifying legitimate images. The balance achieved by the ensemble approaches in this study is particularly relevant for financial applications, where a wide variety of tools to generate deepfakes are available for adversarial attacks. As adversaries often tailor their attacks to circumvent known detection weaknesses, an ensemble model framework can better accommodate evolving threats, offering enhanced resilience in dynamic operational environments.

One tradeoff that is crucial to consider for real-world systems is the increased latency and overhead when running an ensemble model. The latency of an ensemble predictor can be minimized by running the submodels in parallel, resulting in no additional runtime beyond the longest submodel. This leaves the final classifier, which combines submodel predictions to produce one decision. As this final classifier is often simpler and processes much less input data compared to the submodels, it does not add a large amount of latency. Running our ensemble model on one standard NVIDIA Tesla T4 GPU with 40 cores and

16 GB of memory, we have observed a maximum average latency of 2.83 ms ± 124 µs, which is very efficient for a real implementation. In comparison, we found that for each inference, the individual model MesoNet had an average latency of 78 ms ± 18.4 ms.

### *11.3. Inverting Model Predictions*

As can be seen from the Individual Model Evaluation section, the Dolos model had an interesting performance for the GAN-based datasets (this includes the Faceswap-GAN-generated partition of the data in the FaceForensics dataset). We noted that the curves for these datasets had inverted ROC curves. Upon further examination, the model seemed to produce flipped results. With Dolos predictions on GAN images showing a very low TPR/FPR ratio at certain thresholds. However, the same frozen checkpoint for Dolos still predicts accurately for the Diffusion models. This suggests some promise in dynamically flipping or reweighting the predictions from models like Dolos based on properties of the image. This 'informed' classifier would involve retraining the last layer to also process the input image and learn to additionally evaluate which model would be more successful for the type of image provided. This would allow the performance of 'flipped' classifiers like Dolos to provide a bump in performance across more the image classes. In addition, this would allow for models with spiky performance to be reweighted as specialist and only assigned contribution to the ensemble model when the image contains certain artifacts that it performs well on. This is explained further in the Future Work section below.

## 12. Conclusions and Future Work

This study presents a framework for robust deepfake detection in biometric verification systems, with a particular focus on the financial sector. By combining six benchmark deepfake datasets and evaluating multiple state-of-the-art models across our entire dataset, we have shown that single-model performance often stagnates around chance or moderate accuracy, particularly when faced with new or more complex generation methods. Ensemble approaches offer noticeable improvements, achieving up to 80% accuracy across varied deepfake techniques without requiring expensive model retraining. Practical deployment in biometric systems benefits from these higher detection and lower recall rates, reducing vulnerabilities in identity verification processes. Our results underscore the pressing need for model generalization and ensemble-based strategies to keep pace with the rapidly evolving deepfake landscape.

*Future Work*

Despite the promising performance gains observed in our experiments, several constraints merit further attention. First, the computational overhead of running multiple detection models in an ensemble can impede real-time or large-scale deployment. While we propose that an ensemble framework is much more adaptable to emerging deepfake generation techniques, more research must be carried out to understand the cost of generating an ensemble prediction compared to existing deepfake detection pipelines that process user requests. In addition, testing and selecting a good ensemble model from many available options can be a difficult and time-consuming task. We have proposed some starting guidelines and considerations when choosing a limited set of individual models, but more research can be performed with a broader set of models to develop a formalized set of principals.

In this paper, we do not explore explainability in our ensemble model, even though this is an important concern in fraud detection. Ensemble models may be beneficial in this aspect, as we can analyze how the "votes" of individual models are weighed. However, this

framework adds another classifier layer on top of multiple detection models, and it is hard to understand how different model combinations interact and affect overall performance.

While our combined dataset offers a useful test bed, they may not encompass all emerging deepfake generation methods or demographic factors, limiting the broader applicability of the results. Testing our individual and ensemble models on additional larger datasets, particularly in-the-wild datasets that include real-world malicious uses of deepfakes, would help with evaluating additional generalization. Finally, although ensemble approaches generally improve robustness, their performance may degrade if all base models share the same underlying biases or are trained on similarly skewed datasets. Future research should thus focus on both algorithmic optimizations and more diverse, representative data collection to address these limitations. In addition, using other ensemble methods like gradient boosting may improve the performance of our model and reduce bias.

As mentioned in the discussion, there is also potential for work in developing an informed ensemble method. The specific architecture for informing the classifier may vary. One method for developing the informed ensemble could involve training an 'adjudicator' that outputs a confidence score for each of the individual models. Such an adjudicator may be trained on the same datasets used in individual model benchmarking, with tags as a vector representing which models successfully classified the image in benchmarking. This model's output would then be appending to the classifier outputs and passed into the final decision model. Another method could be to strip the last layers of each of the submodels. The input to the decision model would then instead be embeddings generated from each classifier. This allows the final decision model to consider 'votes' based on properties of the image extracted internally in each classifier.

Finally, we have developed and evaluated our ensemble method for single images only, but it can easily be expanded to deepfake videos. Many biometric verification systems have increased security by requiring video rather than just one image, or prompting the user to do specific actions like turning to one side. As video is a collection of highly correlated image frames, our ensemble model could be run out of the box by performing inference on each video frame. However, a better ensemble strategy would likely use different submodels that can detect temporal inconsistencies across images frames and have been trained on large deepfake video datasets.

**Author Contributions:** Conceptualization, H.Z., R.W., M.W., G.B., R.P. (Rachel Park), R.P. (Rafael Palacios), L.C. and G.R.; methodology, H.Z., R.W., M.W., G.B., R.P. (Rachel Park), R.P. (Rafael Palacios), L.C. and G.R.; software, H.Z., R.W. and M.S.; validation, H.Z., R.W. and M.S.; formal analysis, H.Z. and R.W.; investigation, H.Z., R.W. and M.S.; resources, H.Z. and R.W.; data curation, H.Z. and R.W.; writing—original draft preparation, H.Z., R.W., M.W., G.B., M.S., L.C. and G.R.; writing—review and editing, H.Z., R.W., M.W., G.B., R.P. (Rachel Park), R.P. (Rafael Palacios), L.C., G.R. and A.G.; visualization, H.Z. and R.W.; supervision, A.G.; project administration, A.G.; funding acquisition, A.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Instituto de Ciência e Tecnologia Itaú—ICTi.

**Data Availability Statement:** The datasets used in the study are publicly available at the following links: DeepFakeFace https://huggingface.co/datasets/OpenRL/DeepFakeFace (accessed on 1 June 2025), FaceForensics https://www.kaggle.com/datasets/greatgamedota/faceforensics (accessed on 1 June 2025), Individualized Deepfake Detection Dataset https://ieee-dataport.org/documents/individualized-deepfake-detection-dataset (accessed on 1 June 2025), ProGAN https://github.com/PeterWang512/CNNDetection (accessed on 1 June 2025), StyleGAN https://github.com/PeterWang512/CNNDetection, WhichFace https://www.whichfaceisreal.com/ (accessed on 1 June 2025). Original code and data are not readily available due to our efforts to prevent

misuse related to current verification systems. Requests to access data should be directed to the corresponding author.

**Acknowledgments:** We thank Lucas Orosco Pellicer, Sheila Dada and Eduardo Bovo for organizing this collaborative research endeavor and providing their support throughout the project.

**Conflicts of Interest:** The authors declare no conflicts of interest. All data and models used in this paper are open source and publicly available. The full manuscript was drafted and finalized before being presented to the funders, who allowed the manuscript to be published without changes.

## References

1. Naitali, A.; Ridouani, M.; Salahdine, F.; Kaabouch, N. Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions. *Computers* **2023**, *12*, 216. [CrossRef]
2. Diel, A.; Lalgi, T.; Schröter, I.C.; MacDorman, K.F.; Teufel, M.; Bäuerle, A. Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Comput. Hum. Behav. Rep.* **2024**, *16*, 100538. [CrossRef]
3. Salvi, D.; Liu, H.; Mandelli, S.; Bestagini, P.; Zhou, W.; Zhang, W.; Tubaro, S. A Robust Approach to Multimodal Deepfake Detection. *J. Imaging* **2023**, *9*, 122. [CrossRef] [PubMed]
4. Khalid, H.; Tariq, S.; Kim, M.; Woo, S.S. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *arXiv* **2021**. [CrossRef]
5. Khalid, H.; Kim, M.; Tariq, S.; Woo, S.S. Evaluation of an Audio-Video Multimodal Deepfake Dataset using Unimodal and Multimodal Detectors. In Proceedings of the 1st Workshop on Synthetic Multimedia—Audiovisual Deepfake Generation and Detection, ADGD'21, New York, NY, USA, 24 October 2021; pp. 7–15. [CrossRef]
6. Mirsky, Y.; Lee, W. The creation and detection of deepfakes: A survey. *ACM Comput. Surv.* **2021**, *54*, 1–41. [CrossRef]
7. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; Ortega-Garcia, J. Deepfakes and beyond: A survey of face manipulation and fake detection. *Inf. Fusion* **2020**, *64*, 131–148. [CrossRef]
8. Newall, M.; Deeney, C. Nearly 1 in 3 Americans Report Being a Victim of Online Financial Fraud or Cybercrime. 2023. Available online: https://www.ipsos.com/en-us/nearly-1-3-americans-report-being-victim-online-financial-fraud-or-cybercrime (accessed on 1 June 2025).
9. Korshunov, P.; Marcel, S. Vulnerability assessment and detection of deepfake videos. In Proceedings of the 2019 International Conference on Biometrics (ICB), Crete, Greece, 4–7 June 2019; pp. 1–6. [CrossRef]
10. Jain, A.; Ross, A.; Nandakumar, K. *Introduction to Biometrics*; Springer: New York, NY, USA, 2011. [CrossRef]
11. Pei, G.; Zhang, J.; Hu, M.; Zhang, Z.; Wang, C.; Wu, Y.; Zhai, G.; Yang, J.; Shen, C.; Tao, D. Deepfake generation and detection: A benchmark and survey. *arXiv* **2024**. [CrossRef]
12. CSA Top Threats Working Group. *Top Threats to Cloud Computing 2024*; Technical Report; Cloud Security Alliance: Bellingham, WA, USA, 2024.
13. Yadav, D.; Salmani, S. Deepfake: A Survey on Facial Forgery Technique Using Generative Adversarial Network. In Proceedings of the 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 15–17 May 2019; pp. 852–857. [CrossRef]
14. Remya Revi, K.; Vidya, K.R.; Wilscy, M. Detection of Deepfake Images Created Using Generative Adversarial Networks: A Review. In Proceedings of the Second International Conference on Networks and Advances in Computational Technologies, Thiruvananthapuram, India, 23–25 July 2019; Palesi, M., Trajkovic, L., Jayakumari, J., Jose, J., Eds.; Springer International Publishing: Chan, Swltzerland, 2021; pp. 25–35. [CrossRef]
15. Preeti; Kumar, M.; Sharma, H.K. A GAN-Based Model of Deepfake Detection in Social Media. *Procedia Comput. Sci.* **2023**, *218*, 2153–2162. [CrossRef]
16. Kingma, D.P.; Welling, M. An introduction to variational autoencoders. *Found. Trends® Mach. Learn.* **2019**, *12*, 307–392. [CrossRef]
17. Dehghani, A.; Saberi, H. Generating and Detecting Various Types of Fake Image and Audio Content: A Review of Modern Deep Learning Technologies and Tools. *arXiv* **2025**. [CrossRef]
18. Khalid, H.; Woo, S.S. Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 656–657. [CrossRef]
19. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851. [CrossRef]
20. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the International conference on machine learning. *arXiv* **2015**. [CrossRef]
21. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv* **2020**. [CrossRef]
22. Wang, Y.; Chen, X.; Ma, X.; Zhou, S.; Huang, Z.; Wang, Y.; Yang, C.; He, Y.; Yu, J.; Yang, P.; et al. Lavie: High-quality video generation with cascaded latent diffusion models. *Int. J. Comput. Vis.* **2024**, *133*, 3059–3078. [CrossRef]

23. Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; Zhu, J. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 8406–8441.
24. Song, J.; Meng, C.; Ermon, S. Denoising diffusion implicit models. *arXiv* **2020**. [CrossRef]
25. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695. [CrossRef]
26. Sauer, A.; Lorenz, D.; Blattmann, A.; Rombach, R. Adversarial diffusion distillation. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer: Cham, Switzerland, 2024; pp. 87–103. [CrossRef]
27. Li, B.; Sun, J.; Poskitt, C.M. How generalizable are deepfake detectors? An empirical study. *arXiv* **2023**. [CrossRef]
28. Wang, S.Y.; Wang, O.; Zhang, R.; Owens, A.; Efros, A.A. CNN-generated images are surprisingly easy to spot... for now. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8695–8704. [CrossRef]
29. Liu, Z.; Qi, X.; Torr, P.H. Global texture enhancement for fake face detection in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8060–8069. [CrossRef]
30. Ghita, B.; Kuzminykh, I.; Usama, A.; Bakhshi, T.; Marchang, J. Deepfake Image Detection Using Vision Transformer Models. In Proceedings of the 2024 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), Tbilisi, Georgia, 24–27 June 2024; pp. 332–335. [CrossRef]
31. Chai, L.; Bau, D.; Lim, S.N.; Isola, P. What makes fake images detectable? Understanding properties that generalize. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XXVI 16; Springer: Cham, Switzerland, 2020; pp. 103–120. [CrossRef]
32. Țânțaru, D.C.; Oneață, E.; Oneață, D. Weakly-supervised deepfake localization in diffusion-generated images. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 6258–6268. [CrossRef]
33. Asan, J.; Ekaputri, I.; Natalie, C.; Purwandari, K. Exploring Generative Adversarial Networks (GANs) for Deepfake Detection: A Systematic Literature Review. In Proceedings of the 2023 International Workshop on Artificial Intelligence and Image Processing (IWAIIP), Yogyakarta, Indonesia, 1–2 December 2023; pp. 189–194. [CrossRef]
34. Jheelan, J.; Pudaruth, S. Using Deep Learning to Identify Deepfakes Created Using Generative Adversarial Networks. *Computers* **2025**, *14*, 60. [CrossRef]
35. Nirkin, Y.; Wolf, L.; Keller, Y.; Hassner, T. Deepfake detection based on the discrepancy between the face and its context. *arXiv* **2020**. [CrossRef]
36. Huang, Z.; Hu, J.; Li, X.; He, Y.; Zhao, X.; Peng, B.; Wu, B.; Huang, X.; Cheng, G. SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model. *arXiv* **2024**. [CrossRef]
37. Koutlis, C.; Papadopoulos, S. DiMoDif: Discourse Modality-information Differentiation for Audio-visual Deepfake Detection and Localization. *arXiv* **2024**. [CrossRef]
38. Zhang, Y.; Miao, C.; Luo, M.; Li, J.; Deng, W.; Yao, W.; Li, Z.; Hu, B.; Feng, W.; Gong, T.; et al. MFMS: Learning Modality-Fused and Modality-Specific Features for Deepfake Detection and Localization Tasks. In Proceedings of the 32nd ACM International Conference on Multimedia, Melburn, Australia, 28 October–1 November 2024; pp. 11365–11369. [CrossRef]
39. Ricker, J.; Damm, S.; Holz, T.; Fischer, A. Towards the detection of diffusion model deepfakes. *arXiv* **2022**. [CrossRef]
40. Hansen, L.; Salamon, P. Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.* **1990**, *12*, 993–1001. [CrossRef]
41. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive Mixtures of Local Experts. *Neural Comput.* **1991**, *3*, 79–87. [CrossRef]
42. Palacios, R.; Gupta, A.; Wang, P.S. Feedback-based architecture for reading courtesy amounts on checks. *J. Electron. Imaging* **2003**, *12*, 194–202. [CrossRef]
43. Kuncheva, L.I. *Combining Pattern Classifiers: Methods and Algorithms*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
44. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
45. Oriola, O. A stacked generalization ensemble approach for improved intrusion detection. *Int. J. Comput. Sci. Inf. Secur.* **2020**, *18*, 62–67.
46. Latorre, J.M.; Cerisola, S.; Ramos, A.; Palacios, R. Analysis of stochastic problem decomposition algorithms in computational grids. *Ann. Oper. Res.* **2009**, *166*, 355–373. [CrossRef]
47. Dong, J.; Chen, J.; Xie, X.; Lai, J.; Chen, H. Survey on Adversarial Attack and Defense for Medical Image Analysis: Methods and Challenges. *ACM Comput. Surv.* **2024**, *57*, 1–38. [CrossRef]
48. Apostolidis, K.D.; Papakostas, G.A. A survey on adversarial deep learning robustness in medical image analysis. *Electronics* **2021**, *10*, 2132. [CrossRef]
49. Costa, J.C.; Roxo, T.; Proença, H.; Inácio, P.R.M. How Deep Learning Sees the World: A Survey on Adversarial Attacks & Defenses. *IEEE Access* **2024**, *12*, 61113–61136. [CrossRef]

50. Song, H.; Huang, S.; Dong, Y.; Tu, W.W. Robustness and Generalizability of Deepfake Detection: A Study with Diffusion Models. *arXiv* **2023**. [CrossRef]
51. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, South Korea, 27 October–2 November 2019. [CrossRef]
52. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018. [CrossRef]
53. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. Mesonet: A compact facial video forgery detection network. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018; 2018; pp. 1–7. [CrossRef]