AALBORG UNIVERSITY
DENMARK

Detecting Synthetic Media:
Generative AI and its Impact on Cybersecurity

In collaboration with

Trifork

January 5, 2026

Aalborg University
http://www.aau.dk

AALBORG UNIVERSITY

STUDENT REPORT

Title:
Detecting Synthetic Media: Generative AI and its impact for Cybersecurity

Theme:
Scientific Theme

Project Period:
Fall 2025

Project Code:
GitHub Repository

Participant(s):
Simão Ferreira

Supervisor(s):
Hamid Bekamiri

Page Numbers: 48

Date of Completion:
January 5, 2026

Abstract:

Generative AI is everywhere now. While it has revolutionized how we work, it has also lowered the barrier for realistic deepfakes. This thesis examines the growing danger of these AI-generated images and puts deep learning to the test to see if it is a reliable defense. The research centers on four main pillars: how effective current architectures are, how they perform against unseen data, how a human and a model detection capabilities compare, and where AI fits into the broader cybersecurity workflow.

For this study, we evaluated three specific architectures—ResNet50, EfficientNet-B0, and Vision Transformer—training them on a balanced mix of real photos from FFHQ and synthetic ones from StyleGAN. ResNet50 turned out to be the best performer of the group, proving the most reliable for actual deployment with 94.85% accuracy and 95.31% recall on the test set. The comparison with human ability was surprising. The "Deepfake Game" showed that in human visual verifications, the user only managed 57.00% accuracy, whereas the best model did 90.32%.

The research also uncovered some cracks in automated detection. When the models encountered image generation methods they had not trained on, performance took a noticeable hit to 62.95% accuracy. On top of that, a demographic audit showed worrying bias, particularly with EfficientNet-B0, which had a 17% accuracy gap between majority and minority ethnic groups. Ultimately, the thesis concludes that while AI is a massive help for security teams, it requires a "Human-in-the-Loop" approach to manage the algorithmic fragility and bias.

Contents

Acknowledgements

This thesis is the final chapter of my Master's in Business Data Science at Aalborg University. This is the result of the research made from September 2025 to January 2026. This represents a period of intense research and reflection on the topic of deepfake detection.

I would like to start by giving a massive thank you to my supervisor, Hamid Bekamiri, whose dedication and mentorship were key to this project.

Having the opportunity to collaborate with a great company such as Trifork was a privilege. I want to specifically thank Philip Lyngø for accepting the challenge and giving me a warm welcome into the corporate world.

To my partner, my family, and friends: thank you. I am aware these were challenging times, but your support, patience, and love were everything I could hope for. I am incredibly grateful for your patience and support, which kept me motivated throughout my studies.

Lastly, I hope this research offers meaningful insights into the role of deepfake detection in the field of cybersecurity.

Simão Ferreira
January 2026

Abbreviations

A list of the abbreviations used in this report, sorted in alphabetical order:

AI	Artificial Intelligence
API	Application Programming Interface
ATLAS	Adversarial Threat Landscape for Artificial-Intelligence Systems
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
CEO	Chief Executive Officer
CNN	Convolutional Neural Network
CSF	Cybersecurity Framework
DF40	DeepFake 40 (Dataset)
DFDC	Deepfake Detection Challenge
DM	Diffusion Model
DORA	Digital Operational Resilience Act
EER	Equal Error Rate
EFS	Entire Face Synthesis
FBI	Federal Bureau of Investigation
FE	Face Editing
FF++	FaceForensics++ (Dataset)
FFHQ	Flickr-Faces-HQ (Dataset)
FPS	Frames Per Second

FR	Face-Reenactment
FS	Face-Swapping
GAN	Generative Adversarial Network
GenAI	Generative AI
Grad-CAM	Gradient-weighted Class Activation Mapping
IC3	Internet Crime Complaint Center
LLM	Large Language Model
MFA	Multi-Factor Authentication
MTCNN	Multi-task Cascaded Convolutional Networks
NIST	National Institute of Standards and Technology
OOD	Out-of-Distribution
RQ	Research Question
SD	Standard Deviation
SGD	Stochastic Gradient Descent
SOC	Security Operations Center
VAE	Variational Auto-Encoder
ViT	Vision Transformer
VRAM	Video Random Access Memory
XAI	Explainable AI
ZTA	Zero Trust Architecture

Chapter 1

Introduction

1.1 Background and Context

For decades, the saying "seeing is believing" served as a foundational pillar of evidentiary standards. Images were regarded as almost unquestionable proof of reality. However, we are currently witnessing a paradigm shift that fundamentally challenges this assumption. We have entered the era of "synthetic media", a general classification that barely conveys how disruptive the technology in question is. In particular, we are dealing with the emergence of deepfakes, a phenomenon in which artificial intelligence is used to create hyper-realistic audio-visual content from scratch, as well as to modify existing media.

The precise, well-financed manipulation of audio or video was once limited to state actors and large film studios that had the financial means to do so. This high-end manipulation needed great money, processing power, and technical expertise. That era is over. The birth of generative AI has altered the course of events. With powerful public-based models like Midjourney and a plethora of GANs (Generative Adversarial Networks), the barrier to entry has basically disappeared.

What used to need a server farm can now be accomplished on consumer-grade hardware or accessed via a cloud API for pennies. It is safe to argue the technology has moved from controlled research environments to widespread consumer availability. Therefore, advanced fabrication tools are in the hands of the masses, with developers, and the eventuality of falling into the hands of malicious actors. The sophisticated manipulation of reality is no longer a question of if it can be done, but who is doing it and to what goal.

These generative tool's universal accessibility has caused repercussions that go well beyond the novelty of digital art. We are witnessing a dramatic disintegration of information integrity. The problem here is not only that false information might be manufactured, but that it can be generated at a size and pace that overwhelms existing verification procedures.

1.2 Problem Statement

The central challenge facing modern cybersecurity is not only the existence of synthetic media, but also the rapid velocity at which it is evolving. The quality of generative models, particularly in deepfake imaging, is now surpassing human understanding. Simply put, the naked eye is no longer a reliable defense mechanism. Standard protocols are designed to detect malware signatures or unauthorized access attempts; they are not prepared to detect a video call that looks and sounds exactly like a company's CEO authorizing a wire transfer. This represents a fundamental failure in our current defensive posture. The problem is compounded by scale. The democratization of these tools allows adversaries to automate social engineering on an industrial level, launching synthetic identity fraud and impersonation attacks that are both targeted and voluminous. Without the development of robust detection systems, organizations face more than just financial damage. Their credibility is being sabotaged. We are now combating a 21st-century danger with last-century technologies, creating a fundamental vulnerability that our research tries to remedy.

1.3 Research Purpose and Objectives

The primary purpose of this research is to bridge the widening gap between synthetic media generation and detection. As generative models increasingly disconnect visual evidence from objective reality, reliance on manual verification has become a security liability. Therefore, the goal of this project is to develop and evaluate the efficiency of an image detection system that can distinguish between algorithmic fabrication and organic photography. But the goal goes beyond binary classification. This thesis aims to propose measures on how AI can be used in cybersecurity to increase the efficiency of the field and restore a foundational layer of trust in digital media

1.4 Research Questions

- How effective are current deep learning methods at detecting AI-generated image manipulations?
- How does a detection model react to unseen out of distribution data?
- How does a detection model compare to the ability of humans to identify manipulated media?
- How can AI be used to increase efficiency in Cybersecurity?

Chapter 2

Literature Review

Generative AI has been evolving rapidly, setting the stage for a strategic inflection point in the digital landscape, with these models considered powerful dual-use technologies that have serious implications for cybersecurity. The unprecedented creative and productive capabilities opened by the evolution from early, simple, rule-based AI systems to state-of-the-art generative models, including GANs and diffusion models, have been described in various literature [1]. The very technologies in question have also been equally instrumental in aiding malicious actors in launching advanced phishing campaigns, executing new forms of fraud, and amplifying disinformation campaigns to an extent and sophistication hitherto unimaginable [2]. The lowering of barriers when creating synthetic media, or "deepfakes," has fueled an arms race between innovation and forensic detection. The very accessibility that has accelerated innovation has now diminished efforts at creating synthetic media that can convincingly impersonate an individual, manipulate public opinion, and undermine trust in digital information.

This paper presents an extensive literature review of the topic within both academic and industry domains from 2018 through 2025. The goals of this research include examining the technical basis of malicious uses of generative AI through an analysis of the malicious uses of these advances within the cyber community and digging deeper into an escalating war that intensifies between these malicious uses of generative tools and those designed as countermeasures towards detection. The latter part of the academic journey that led to the present state of the art in generative AI includes an exposition of the underlying principles of neural networks from their inception through transformer and diffusion networks to the latest efforts at updating these underlying principles towards new uses of these advances.

The next sections explore two of the most prominent threat types: visual deepfakes and audio manipulation. For these two topics, the review outlines the technical processes of creation, examines the adversarial relationship that exists in both creation and detection as a form of cybersecurity threat, and outlines the collective cybersecurity risks that have been identified. These topics then inform a much wider conversation that exists between

generative forms of AI and existing cybersecurity frameworks that require adaptation. Finally, the review will explore existing forms of defense against these threats.

The literature review states that the dual nature of generative AI and its rapid growth necessitate such a response. Preventing the cybersecurity threats that come with it requires such an approach. The findings highlight the need for such a response as important and relevant at this timely moment for the transformative potential of generative AI to be realized without negative consequences for digital security.

2.1 Generative AI: Background and Evolution

2.1.1 From GANs to Diffusion

To gain a full understanding of the cybersecurity threats posed by using generative artificial intelligence, it is important to reflect on the background and evolution of the system as well as the factors that led to its increased adoption. This section briefly introduces the events and models that constitute the landscape of the use of generative artificial intelligence. The technology evolution that follows involves more than an academic consideration due to the rising levels of sophistication of threats.

The evolution of generative technologies marks a significant shift from the concrete phase of AI to one defined by data adaptability in contemporary times. The transition has been from rule-based AI towards something flexible, such as machine learning [1].

Generative Adversarial Networks (GANs), introduced by Ian Goodfellow and his colleagues in June 2014, were a significant milestone in this evolution. It represented a novel design in which two neural networks competed to generate very realistic synthetic data [3]. This breakthrough generated unprecedented interest and research, leading to the construction of increasingly sophisticated media synthesis models. This was a period in which the field rapidly grew with the introduction of transformer architectures and diffusion models. This is a significant transition, as diffusion-based models have recently advanced to the state-of-the-art in the field of entire face synthesis (EFS), with signs that photorealism will considerably outperform many earlier GAN-based techniques [4].

Generative AI's capabilities are based on a common set of foundational models and methods poised for synthesis of media. At the basis of image/video synthesis is the auto-encoder architecture, which typically consists of a shared encoder network learning identity-independent attributes (e.g., facial expressions) and multiple identity-specific decoders that reconstruct the faces of different subjects [3, 5]. A multitude of other model families have evolved from there, either in quality or complexity.

Considering that VAEs constitute a different family of models, it is worth noting that they provide for generation by first learning a compressed latent representation of input data. The generation of compatible new images can then be enforced by sampling from this learned latent vector and feeding it through the decoder network [6].

More recently introduced, diffusion models have become an important technology for

appropriately performing high-fidelity entire-face synthesis. Models like StableDiffusion, by learning to reverse a gradual process of adding noise to an image, produce extremely detailed and coherent images from inputs composed purely of random noise [4]. Whereas GANs were the predecessor in this field, diffusion models are seen to achieve much more realistic outputs, which, in turn, is considered to make them more difficult to differentiate between original and fake.

2.1.2 Architectural Differences: CNNs vs ViTs

Understanding the basic architectural split between Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) in contemporary computer vision is essential to appreciating the advantages and disadvantages of current detection techniques. These indicate radically different perspectives on how to view, rather than being alternative instruments for the same task.

CNNs, for years the undisputed champions of image analysis, operate by learning spatial hierarchies of patterns. In the early layers, the model relies on filters—often called kernels—to pick out the essentials of an image, like edges or shifting gradients. As we go deeper, subsequent layers stack these basic patterns together to form richer textures and, eventually, distinct objects [7]. The whole system runs on "convolutional inductive bias", which is a technical way of saying that pixels sitting next to each other are more related than the ones far apart. Because this assumption is hard-wired into them, CNNs tend to be surprisingly efficient learners, especially when we don't have a massive dataset to work with.

Vision Transformers, on the other hand, arrived from the world of natural language processing and brought with them a completely different approach. A ViT discards the notion of a sliding filter. Instead, it essentially takes the image and divides it into a sequence of patches—all fixed size—sort of like how a sentence is broken down into individual words. From there, it uses self-attention to map out the linkages between every patch and every other patch simultaneously [8]. What makes this approach distinct is that it eliminates the intrinsic spatial bias found in a CNN; it is free to pick up on patterns regardless of how far apart they are in the data, making zero assumptions about local correlations.

There are significant trade-offs. A CNN like ResNet notably benefits from its inductive bias, which provides it an edge in understanding visual structure when training data is limited. But when the amount of data increases, the tables rotate. Vision Transformers regularly outperform CNNs when pre-trained on large, proprietary datasets such as JFT-300M [8]. This implies a deep truth: depending on a hard-coded assumption, no matter how helpful, is ultimately less effective than being able to learn pertinent patterns directly from the environment, given sufficient evidence.

2.1.3 Democratization of Access

The democratization of access to powerful creation tools is one of the defining characteristics of the current generative AI landscape. This aspect has intensified the reach for this technology through open-source platforms and user-friendly software. Such pre-packaged implementations of sophisticated deepfake algorithms are, for instance, available through DeepFaceLab and Faceswap, which are freely available on GitHub [3, 5].

The implications of such technical simplifications are quite dramatic. In one sense, it encourages artists, scientists, and developers to use technologies for their intended purposes in much of the creative and scientific worlds. On the other hand, it creates additional avenues for the otherwise less technically proficient to reproduce malicious content with very little effort. This kind of complication further aggravates the adversarial arms race because malicious potential knows no bounds, and neither do state nor scientific organizations produce the sole misuse.

2.2 Threat Landscape

2.2.1 Deepfakes and Synthetic Media

Deepfake technology represents a sophisticated escalation in the cybersecurity threat landscape. We aren't just talking about stealing data or financial scams anymore, this attacks the credibility of digital information itself. A deepfake is synthetic media—images, audio, or video—created by deep learning models. The alarming part is that these fakes are often so realistic that a human being literally cannot tell the difference between what is real and what is manipulated [2, 3]. Consequently, this creates significant challenges for aspects such as national security, political stability, and personal privacy, which is why addressing these issues is critical.

When people say "deepfake," they are referring to a variety of different manipulation tricks. According to Yan et al. [4], we can generally sort these methods into four main categories:

Table 2.1: Categories of Deepfake Manipulation Methods [4]

Category	Description
Face-Swapping (FS)	This is what most people associate with deepfakes. It involves taking a person's face from a source and pasting it onto a target video or photo. It keeps the original expressions and movements, just with a different face attached.
Face-Reenactment (FR)	Sometimes called expression swapping, this is more like Digital reenactment. It tweaks the facial expressions, mouth movements, or head position of the target to match whatever a "source driver" is doing.
Entire Face Synthesis (EFS)	This is generating a completely new face from scratch. It looks photo-realistic, but the person doesn't exist—it's widely used to create fake identities.
Face Editing (FE)	This technique is a bit more subtle. It involves changing specific traits—like making someone look older, swapping gender, or changing hair color without erasing the person's core identity.

2.2.2 Cybersecurity Implications

The fact that deepfake tech is getting so realistic and easy to grab is creating a whole mess of potent cybersecurity threats.

Political Disinformation and Social Manipulation
First off, we have Political Disinformation and Social Manipulation; we are talking about fabricated videos where political leaders appear to say controversial and inflammatory speeches just to stir up unrest or sway an entire election [2, 1]. Then there is the threat of Corporate Espionage and Reputational Attacks. Even if the literature doesn't explicitly document every instance yet, the logic is straightforward: if we can damage a public figure's reputation with a deepfake [2], we can also do it to a CEO to manipulate stock prices or launch social engineering attacks against high-value targets.

Identity Theft, Fraud, and Sextortion
On a more personal level, there is Identity Theft and Fraud, where synthetic media is used to bypass facial recognition login systems, scam insurance companies, or just build fake identities for criminal ops [2]. Then there is Sextortion, easily the most severe of implications. It involves using fake explicit media to bully and intimidate victims. The numbers are just staggering: the FBI's IC3 (Internet Crime Complain Center) received

over 18,000 complaints about this in 2021 alone, with financial losses hitting north of $13.6 million. And while visual deepfakes seem to get all the headlines, the parallel rise of audio deepfakes is becoming an equally urgent and technically distinct problem.

Social Engineering and Phishing
Generative AI isn't just adding a new tool to the toolkit; it is fundamentally exposing the attack surface by amplifying social engineering. Phishing has been the predominant path of threats for years, the 2024 FBI Internet Crime Report flagged it as the top cybercrime type [9], but the field has changed completely. We used to deal with static, unsophisticated templates full of typos that were easy to spot. That era is over.

Now, attackers are quickly generating personalized, context-heavy emails and texts at a massive scale. Look at the scam where fraudsters impersonate a CEO, using panic-inducing messages and fake bank screenshots to fool resellers. It is that level of specific detail that makes these attacks so much harder to catch compared to older attacks.

Then we have the rise of high-fidelity voice cloning, driven by deep learning tech like WaveNet and WaveRNN [10]. This has developed a new type of "vishing." Malicious actors can now clone the voice of someone we trust implicitly—like a boss or a family member—and it is getting nearly impossible to tell the difference. They use that cloned authority to authorize transfers or steal secrets, representing a massive jump in capability. Ultimately, this shifts the entire battlefield: we aren't just defending technical vulnerabilities anymore; we are trying to protect against the exploitation of human trust itself, with deepfakes being the most visible face of that threat.

Generative AI is seriously challenging the old-school ways we handle cybersecurity, which usually just focus on spotting malware, patching network holes, or dealing with structured attacks. Even though we have foundational guidebooks like the NIST AI Risk Management Framework, the literature provided does not really explain how these are supposed to adapt to the specific, unique problem that GenAI creates.

2.3 Detection and Technical Challenges

2.3.1 Frequency Analysis and Artifacts

A massive shift in the strategy for digital forensics has been moving away from just analyzing images in the standard pixel space and instead interrogating them in the frequency domain. It is really similar to how a sound engineer looks at a waveform to spot tiny distortions that the ear cannot even pick up on its own. The point is to find those unintentional and systematic fingerprints left by the generative process—sort of like spectral echoes of the algorithm itself.

For years, the frequency domain was a rich area of interest for GAN-generated fakes. The up-sampling operations inherent in most GAN architectures often imprinted a tell-tale artifact: a "regular grid" pattern, clearly visible in the Discrete Fourier Transform spectrum

of a synthetic image. This became a well-established vulnerability, a spectral signature that many early and effective detectors learned to exploit with ruthless efficiency [11, 12, 13].

This is precisely where Diffusion Models changed the game. A frequency analysis of DM-generated images reveals a startling resemblance to real photographs. They are conspicuously free of the tell-tale grid-like artifacts that plagued GANs, a key reason why GAN-trained detectors are so blind to them [13]. It seemed, for a moment, that diffusion models had managed to perfectly mimic the frequency profile of reality. But if we dig a little deeper, we eventually discover a new, much more subtle weak spot. It turns out that in a variety of different Diffusion Models, there is this consistent pattern where they systematically underestimate high frequencies [13]. The fine-grained, noisy details that characterize authentic high-resolution images are consistently muted. The prevailing hypothesis links this flaw directly to the training objectives common in diffusion models, such as L_{simple}. These objectives are optimized for perceptual quality—what looks good to the human eye—rather than mathematically precise reproduction of high-frequency information. This high-frequency detail is primarily synthesized in the final, most delicate steps of the denoising process (where the timestep $t \to 0$). As visual heatmaps of the spectral error demonstrate, it is in these last few steps that the model struggles most, down-weighting their importance in favor of overall coherence and thereby leaving a faint but detectable spectral gap [13].

2.3.2 Generalization and Model Failure

The state-of-the-art detectors from yesterday are suddenly rendered obsolete by new generative techniques, leaving us in a never-ending game of cat and mouse. Therefore, the approach is to create systems that can generalize and reliably detect modifications from unseen generating families rather than just building strong detectors. This is the main issue.

Nowhere is this failure of generalization more starkly illustrated than in the recent collision between legacy detectors and the new wave of Diffusion Models (DMs). An extensive experimental study reveals a critical vulnerability: forensic tools meticulously trained on a training set of Generative Adversarial Network (GAN) artifacts suffer a catastrophic performance collapse when confronted with images from DMs. The Area Under the Receiver Operating Characteristic Curve (AUROC), a common indicator of classifier performance, decreases by an average of 15.2% in this case, according to the data [13]. However, there is more to the story than that.

But what is amazing about this failure is how asymmetrical it is. An intriguing image appears when the training schedule is inverted. A detector re-trained on a corpus of DM-generated images not only achieves near-perfect detection rates on fakes from other Diffusion Models but also generalizes surprisingly well to GAN-generated content, a capability that simply does not exist in the opposite direction [13]. This one-way generalization strongly suggests a fundamental difference in the artifacts produced by these two genera-

tive families.

The most plausible hypothesis, grounded in this evidence, is that images generated by Diffusion Models possess fewer, or at least far more subtle, "family-specific artifacts" than their GAN-based predecessors [13]. They are, in essence, cleaner forgeries. This makes them inherently more difficult to detect with legacy tools but, paradoxically, a superior source of training data for a truly robust, generalized detector. The very subtlety of their tells forces a model to learn more fundamental and universal indicators of artificiality. This points directly to the underlying architectural philosophies that birthed these artifacts in the first place.

A review of the current literature highlights several key bottlenecks that stand in the way of widespread success. Perhaps the most significant is the lack of generalization and robustness; models trained on pristine academic datasets often degrade when exposed to newer, diverse, and chaotic "in-the-wild" deepfakes [5, 4, 14]. Then there is the issue of scalability. We simply aren't keeping pace with the overwhelming volume of content or the rapid evolution of generative tech [14]. This is made worse by "domain shift," where models fail because the input data varies slightly from what they studied—think diffusion-generated videos, or audio tracks featuring background music or non-English speech [14].

2.3.3 Real World Detection

When we are dealing with the high-stakes world of deepfake detection, relying on simple accuracy is not just inadequate, it is dangerous. It creates a false sense of security. In any real world scenario, genuine content is always going to vastly outnumber the malicious fakes. So, for a classifier working in that kind of lopsided environment, hitting 99.9% accuracy is easy—all it really has to do is label everything as "real." That's a useless outcome that fails the security mandate, even if it looks impressive in a report.

We must look at the practical fallout of the different error types. A false positive, where real media gets wrongly flagged as a deepfake, risks unwarranted censorship and kills trust in the system. A false negative, on the flip side, is a direct security failure—a malicious deepfake slips through the net, spreading disinformation and poisoning the discourse. The societal cost for those two errors is totally different. This implies we need more nuanced metrics built for this kind of imbalance. Instead of accuracy, the field needs to prioritize Precision, which answers: "Of the things we flagged, what proportion were actually fake?" and Recall (or Sensitivity), which asks, "Of all the actual fakes out there, how many did we catch?". The F1-Score gives a harmonic mean of these two, offering a more balanced assessment than accuracy alone [7].

The specialized literature has settled on even tougher measures to really test these systems. We have the Area Under the Receiver Operating Characteristic Curve (AUROC), which essentially evaluates how well a classifier separates the real from the fake across every possible decision threshold—giving us a holistic view of its actual power. But when

it comes to actual deployment? The Probability of Detection at a Fixed False Alarm Rate is arguably the metric that really matters. It gets straight to the operational limits of social media platforms, where a flood of false positives would just kill the user experience. By measuring Recall at a rigid, low false alarm rate—think 1%—we get a real-world benchmark for utility when the price of flagging authentic content is too high [13]. The problem is, those perfect lab scores and curves often fall apart completely when they collide with the messy, unpredictable reality of deepfakes online.

The ultimate crucible for any detection model is not the sterile, controlled environment of an academic dataset, but the chaotic digital ecosystem of the open internet. Media shared on social platforms is compressed, re-encoded, screenshotted, and adorned with modifications that were never part of the detector's original training data. The true test of a detector is its resilience in this "in-the-wild" context.

Research confirms that this transition is brutal. Even the most advanced models see their performance drop significantly when they run into common, real-world noise. Just adding something simple like background music to a fake audio clip is tied to a steep 17.94% drop in accuracy and a massive 26.12% spike in false negatives—meaning those fakes often slide right by without getting caught [14]. It's a similar story with images. Text overlays, which are on every meme and social post out there, are linked to a 9% dip in accuracy and a 10.5% drop in F1-score, even for fine-tuned models [14]. The alarming part is that these are not even purposeful attacks; they are just the standard messiness of the internet. Other common issues, like "video re-encoding attacks" or just the presence of "low quality videos," are also known to completely compromise detector performance [15].

This shines a light on a major weak point in how we currently evaluate performance: our datasets are just too limited. For the most part, deepfake datasets are manufactured synthetically; researchers take a narrow set of manipulation tricks and apply them to a small, often pretty rigid library of source videos [15]. The real trouble is that this approach does not even scratch the surface of the chaotic mix of content, unusual generation styles, and random artifacts we stumble across on the web. It creates a disconnect between the sanitized data the model was trained on and the complex reality it must survive in. That gap right there is what accounts for a massive chunk of the performance crash we are seeing [14].

The catastrophic failure of GAN-trained detectors on Diffusion Models, and the performance collapse of modern classifiers when faced with simple text overlays, are not isolated incidents; they are symptoms of a foundational mismatch between our sterile training environments and the chaotic digital wild. Therefore, the path forward must be a dual-pronged assault on the problem. We must continue to develop more sophisticated and generalizable architectures, models capable of learning from subtler, more universal artifacts, such as the high-frequency suppression characteristic of Diffusion Models. At the same time—and just as urgently—we've got to put serious resources into building bigger, grittier "in-the-wild" benchmark datasets. The only way we are ever going to build

defenses that work in the real world, rather than just looking good on paper, is if we train and test these models against the messy reality of the threat as it exists.

One of the most critical takeaways lately is the glaring performance gap between academic setups and real-world data. We have these top-tier models that achieve high accuracy on benchmark datasets like FF++, but they often have a severe performance degradation when tested against contemporary, "in-the-wild" deepfakes. When the Deepfake-Eval-2024 benchmark dropped, it revealed a precipitous decline in performance for leading open-source models, which confirms that academic benchmarks are not representative of the actual threats we face now. This gap just goes to show that many state-of-the-art models have overfit to the specific artifacts of older generation methods, failing to learn the deeper, generalizable traces of manipulation found in modern deepfakes [14].

If we had to pinpoint the most critical failure, it is model generalization. A lot of detectors spot forgeries made with methods they know but fail significantly in "cross-forgery" and "cross-dataset" tests [4]. They are overfitting to specific artifacts in the training data rather than learning what a deepfake looks like fundamentally. The statistics here paint a pretty worrying picture. Chandra et al. [14] demonstrated that even state-of-the-art open-source models took a massive hit in performance, with AUC scores plummeting by as much as 50% when tested against the Deepfake-Eval-2024 benchmark. This essentially confirms that our current academic datasets are outdated; they simply aren't representative of the sophisticated deepfakes currently circulating online.

2.4 Strategic Response and Future Directions

2.4.1 Defensive Strategies

The literature examinates the forward-looking technical, organizational, and policy-level moves we can make to fight back against GenAI threats. It reviews the main countermeasures people are building right now, while being honest about the fact that they aren't perfect, and we still have a lot of research gaps to fill in this constant cat-and-mouse game.

On the technical front, the defense really relies on three main approaches: detection, proactive labeling, and verification. The most common method is essentially using AI to catch AI, where developers build specialized deep learning classifiers—usually Convolutional Neural Networks (CNNs)—that are trained to spot those tiny, unusual artifacts or statistical glitches that show up in synthetic media. These tools are built to give a simple yes-or-no answer on whether content is real or fake [3]. Beyond just detection, there is a push for "digital watermarking," which is a more proactive move where we hide an invisible, robust identifier inside the file right when it is created. Think of it as cryptographic proof of where the file came from, if that watermark is missing or broken, it is a red flag that someone altered the media. Finally, researchers are looking at blockchain to verify authenticity by leveraging its decentralized, unchangeable nature. The idea is to store cryptographic hashes of the original files on a ledger to create a tamper-proof history,

letting anyone verify a file just by checking if its current hash matches what is stored on the chain [16]

But technology isn't enough on its own, dealing with GenAI risks also demands strong organizational policies and better regulations, with a huge focus on the human side of things. Since social engineering attacks are getting so sophisticated, regular people are effectively becoming the final line of defense. That makes user education—like cybersecurity awareness training—critical because a well-informed user is way more likely to spot suspicious content like AI-driven vishing or complex phishing scams. Zooming out to the big picture, there is a growing consensus that we need comprehensive regulation. Right now, the AI market is a bit of a gray area without a solid regulatory framework, which leaves massive gaps in accountability and ethical oversight regarding potential misuse. Researchers and policymakers are urgently calling for new rules to fix this, ensuring that as we build and deploy AI, we are sticking to principles of safety, fairness, and accountability [1].

2.4.2 Policy, Ethics, and Future Directions

Beyond technical constraints, we face some serious ethical dilemmas. There is a bit of a paradox here: to build a decent "hunter," we need prey. Researchers effectively must generate harmful fake content just to gather enough training data for their defensive systems [1]. It's a necessary evil, but an irregular one. Finally, the "black box" nature of these tools is a major problem. Many deep learning models can flag a file as fake but offer no explanation as to why, which is a dealbreaker in digital forensics, where explaining your findings is critical [3].

Complicating matters even further is the stubborn issue of adversarial robustness and evasion. It feels like a constant back-and-forth: as detection methods improve, the Malicious actors are right there developing new ways to overcome them. One particular threat is adversarial attacks, where attackers inject tiny, often imperceptible, perturbations into a deepfake video to confuse the model [2]. If an attacker knows how a detector operates, they can hand-craft inputs specifically designed to break it. This reveals a major crack in our current defenses and makes developing adversarial models a top priority.

To remain effective, future detection systems must evolve. We must stop relying solely on facial analysis and start considering a much wider range of signals and content types. Deepfakes are becoming multi-modal, messing with both audio and video, like in the DFDC dataset [17]. Consequently, we need methods that check if the audio matches the video, looking for things like phoneme-viseme mismatches, checking if the lip movements match the sounds [18, 3]. We also need to get better at detecting non-face content. The same tech making fake faces is making fake art and satellite imagery. There is some hope here; Yan et al. [4] found that a CLIP-large model trained only on face deepfakes could surprisingly detect non-face AI content from the GenImage dataset. This suggests that models can learn universal "fingerprints" of forgery that apply to more than just faces.

2.4.3 Possible Solutions

Tackling the deepfake threat requires a sweeping, cooperative effort that bridges the often-disconnected worlds of technology, policy, and ethics. It is becoming increasingly apparent that tech is not a singular fix. We need a strategy that is far more holistic. One of the most significant paths forward involves a rethinking of media traceability. We need the capacity to record and verify the entire lifespan of a video or image to measure exactly how much it has been altered from its original state. This is where emerging concepts like Blockchain Distributed Ledger Technology could be a game-changer. As suggested by Khan et al. [16], if we can create trustworthy, immutable records of a file's origin, a transparent history log, it becomes much easier to spot manipulated media and, crucially, ensures accountability by tracing those changes back to where they started.

Another major hurdle is the need for transparency in detection methods. A detection system must be trustworthy, particularly when it comes to legal and investigative contexts where the stakes are high. It is not enough for a system to simply flag a video as fake; it needs to be understandable. This brings us to the concept of Explainable AI. To be admissible in court or useful in forensics, we must move beyond "black box" models. As Nguyen et al. [3] point out, the system should allow experts to see exactly how and why a determination was made, ensuring that the outcome can be verified and explained rather than just accepted on blind faith.

However, counting only on reactive detection—trying to catch the fakes after they have already been posted on the internet—is a losing battle. A much stronger long-term solution would be to verify media rights at the moment of creation, rather than struggling to flag it later. This changes the script from reactive detection to proactive verification. If we can establish reliable data provenance—maybe by embedding secure digital watermarks the instant a photo is taken or by keeping trusted online ledgers—we create a verifiable record of where a file came from. This kind of approach would make it exponentially harder for untraceable, synthetic content to gain any real credibility in the first place.

Finally, we cannot ignore the human element. Straightforward prescriptions and education are just as mandatory as the software solutions. There must be robust laws and ethical guidelines defining what is acceptable when creating and sharing digital content. On the flip side, the public needs to be empowered to think critically about what they see on the internet. If we can empower the public to spot the subtle tells of manipulation, we might see a significant drop in how far this harmful material spreads [1]. Ultimately, the way forward isn't going to be straightforward. It demands a complex but essential partnership between technologists, policymakers, and educators to construct a digital environment that is tough enough to withstand the growing flood of manipulated media.

2.5 Conclusion

This review charts the rapid and concerning evolution of generative AI within cybersecurity, illuminating a field that has moved far beyond traditional threat models. At its core, we are dealing with a persistent dual-use dilemma; the technology itself is fundamentally neutral, yet its impact is dictated entirely by its application, creating a tension between creative innovation and the potential for malicious social manipulation [1]. This shift has created a new reality where the authenticity of digital content can no longer be taken for granted, posing a significant challenge to the fabric of our digital society.

The literature paints a picture of a field locked in a relentless "arms race" between increasingly sophisticated generative models and the detection techniques trying to catch them. While defensive methods have certainly grown in sophistication—moving from spotting simple artifacts to using complex deep learning—they are consistently struggling to keep up. There is clear evidence of a performance drop-off; detectors that work well in the lab often crumble when evaluated on newer, higher-quality benchmarks and chaotic "in-the-wild" data [2, 4, 14]. This really highlights a critical gap in robustness. Current models just do not have the generalization needed for the real world, and they often fail because of domain shifts—whether that is due to the emergence of new diffusion architectures or environmental factors like background noise in audio [14].

Ultimately, if we look at the trajectory, it is clear that the evolution of generative models is still moving faster than our ability to defend against them. We are in urgent need of solutions that are not just new but genuinely adaptive and holistic. Fixing this is not simply a matter of shipping better code; we need a strategy that considers the bigger picture. We certainly require detection models that are accurate and scalable enough to handle real-time threats, but these must operate in tandem with robust policy, legal frameworks, and a public that understands what they are looking at. The limitations identified in this review—particularly the way current detectors stumble when facing realistic threats—provide a strong justification for further empirical research. This work essentially serves as a stepping stone, laying the groundwork for the next generation of resilient tools required to secure our shared digital future.

Chapter 3

Methodology

This chapter outlines the operational plan of our research. It maps out the complete technical pipeline detailing how we selected the dataset, the architectural tools, and the tested models. Crucially, in a study focusing on synthetic media, we must be transparent about our own tools. Generative AI was utilized during the drafting phase strictly to augment phrasing and facilitate brainstorming; it served as a linguistic assistant, not a source of analytical content. The conclusions, interpretation of data, and intellectual ownership of the work remain with us. For the sake of reproducibility and transparency, the full code for this project is publicly available on our GitHub repository (link)

3.1 Research Design and Data

We structured this study around a direct, adversarial comparison. Rather than simply training a single model to detect deepfakes, the research design sets up a "battle of architectures." We are effectively pitting the established veterans of computer vision, Convolutional Neural Networks (CNNs), against the emerging dominant paradigm of Vision Transformers (ViTs), and finally, against the raw capabilities of a Multimodal Large Language Model. The goal is not only to find which model has the highest accuracy, but also to understand the trade-offs between computational weight, inference speed, and detection fidelity.

When it comes to the dataset, we sourced the "Deepfake and Real Images" dataset curated by Manjil Karki from Kaggle (link). It is a balanced repository, which is critical since an uneven dataset in a binary classification task is a recipe for biased results. The "Real" images come from the FFHQ (Flickr-Faces-HQ) dataset, providing a solid baseline of organic human diversity in terms of lighting, age, and ethnicity. The "Fake" counterpart consists of faces generated by StyleGAN. These examples can be particularly hard to spot because StyleGAN does not just copy features, it creates textures that look normal to the naked eye. To keep our testing reliable, we split this data into training, validation, and testing sets before running any code. This way, we avoided any data leaks that might

Figure 3.1: Data distribution

change the performance of our results.

3.2 Preprocessing Pipeline

The standardization pipeline made before inputting any data into the neural networks helped to ensure uniformity of input data and allow better convergence rates. Firstly, all raw images were downsampled at the fixed resolution of 224×224 pixels. This was because this is the dimension of the input required by the pre-trained ImageNet backbones integrated into our transfer learning protocol. Following resizing, the intensity values of the pixels were normalized. We scaled the raw pixel data from the standard 0–255 range to a normalized 0–1 float range, subsequently applying the standard ImageNet mean and standard deviation normalization. This step is critical for centering the data, which aids the optimizer in navigating the loss landscape more efficiently. In addition, we resorted to data augmentation during the learning phase to minimize the incidence of overfitting, which is a major issue concerning the Vision Transformer model without any of the generalized inductive biases associated with CNNs. Among other techniques, aligned horizontal flips and rotating images slightly skewed the chances for vanilla memorization of an arrangement of pixels while instructing the model in structural feature learning only.

3.3 Architectural Differences

We chose three specific models to get a full picture of the strengths of each. First, we have ResNet-50, as established baseline architecture for this study due to its proven stability. It uses "skip connections" to solve the vanishing gradient problem, allowing it to go 50 layers deep without losing the signal. It is our baseline; if a new, complex model fails to beat ResNet, it is probably not worth the computational cost. Then we have EfficientNet-B0. The philosophy here is different. Instead of just making the network deeper, EfficientNet scales width, depth, and resolution uniformly. We included this to answer a practical question: can we run deepfake detection on a mobile phone? EfficientNet is lightweight

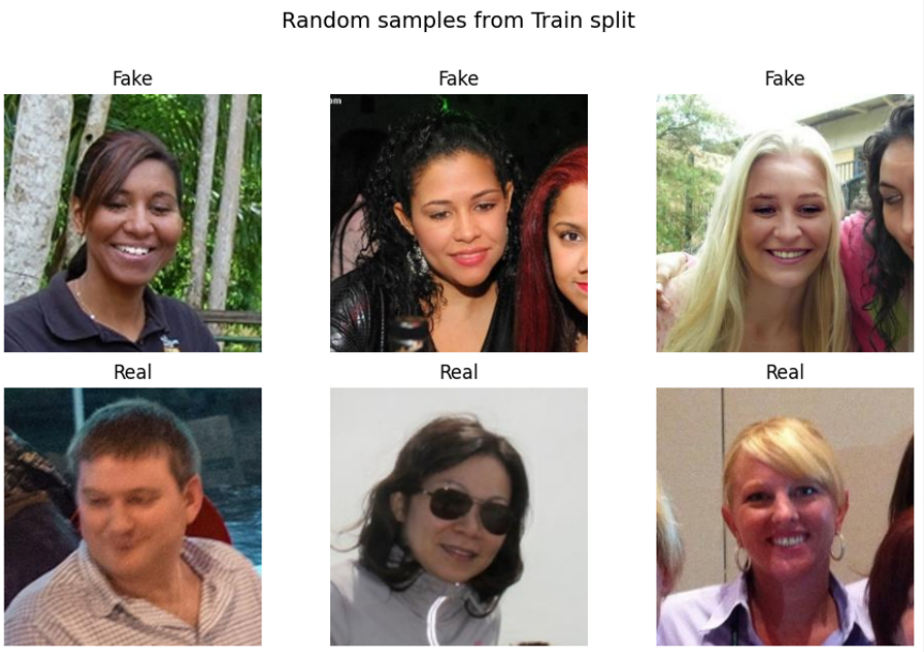

Figure 3.2: Random sample of images from the training split

and fast, making it the proxy for "edge deployment." The third contender is the Vision Transformer (ViT). This is a total departure from CNNs. It divides the image into patches and processes them using self-attention, similar to how LLMs process words. The theory is that while CNNs are great at local details (like the edge of an eye), ViT sees the global context, potentially spotting structural inconsistencies—like mismatched ears—that a CNN might miss.

Table 3.1: Model Architecture Comparison

Model	**Type**	**Why we chose it**
ResNet-50	CNN (Residual)	The industry standard benchmark. Reliable and proven.
EfficientNet-B0	CNN (Scaled)	To test if high accuracy is possible with low computational cost.
Vision Transformer	Transformer	To see if “global attention” catches artifacts that CNNs miss.

3.4 Training and Evaluation

We built the training environment in PyTorch, using the AdamW optimizer. We picked AdamW over standard SGD because it handles weight decay better, which helps prevent the models from getting too comfortable with the training data. We set the learning rate at 1e-4 but added a scheduler to drop that rate if the validation loss stopped improving—essentially telling the model to "slow down and look closer" when it got stuck. We ran this for 20 epochs with a batch size of 32, limited largely by the VRAM we had available. We also used "early stopping" with a patience of 5 epochs. If the model did not get better for 5 rounds, we stop the training to save time and prevent overfitting. For evaluating the models, we looked past simple accuracy. In cybersecurity, accuracy can be a vanity metric. We focused heavily on Precision and Recall. We need to know: is the model flagging real photos as fake (a false positive nightmare for users), or is it letting fakes through (a security failure)? So for this study the F1-Score will be a great metric since it give us a combination of the precision and recall into a single value indicating the model's reliability.

3.5 Research Limitations

While the primary goal of the research pertained to examining deep learning models for detecting deepfakes, this research study does not lack flaws in its blindness either. State of Affairs: Being transparent on the limitations within this research allows for a better understanding of this paper's results and their current situation.

Perhaps the greatest limitation of the project is the complete reliance on static images. Deepfakes, by definition, are a multimedia problem. They are video, often audio, and sometimes even three-dimensional. By limiting the scope of the project to static image recognition, the temporal and audio cues necessary to give the problem a holistic investigation are being disregarded.

Thus, while these results indicate these models are quite acute at least when viewing a single image, it does not necessarily reveal what these models could do when faced with, say, a live video stream where, you know, time is a factor. And then, of course, there is the problematic aspect of all this which, quite frankly, needs to be addressed. As objective as we may strive to be, these results, by necessity, are going to be grounded firmly in the data sets that we choose to work with. When these sets lack any kind of demographic representation, what these models learn could very well depend on the subject's ethnicity and/or sex. Essentially, by working from pre-existing data sets, what these models learn is steeped in the problems inherent in those sets and making these detection models work equitably for all is, quite frankly, still a huge open question.

Chapter 4

Research Findings

4.1 Model comparison and metrics

The results here tell a pretty classic story in machine learning: you have these models with massive theoretical power running headfirst into the messy reality of unseen data. When we break down the performance across Training, Validation, and Testing, we start to see exactly how the different architectures—ResNet, EfficientNet, and Vision Transformer (ViT)—handle the deepfake detection problem differently.

Table 4.1: Model Performance Metrics across Phases

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	EER	Threshold
Training Phase							
EfficientNet	0.9909	0.9959	0.9858	0.9908	0.9998	0.0070	0.2744
ResNet	0.9949	0.9912	0.9987	0.9949	0.9999	0.0043	0.7329
ViT	0.9659	0.9385	0.9971	0.9669	0.9987	0.0197	0.9097
Validation Phase							
EfficientNet	0.9796	0.9849	0.9742	0.9795	0.9979	0.0192	0.3245
ResNet	0.9826	0.9708	0.9951	0.9828	0.9988	0.0152	0.8711
ViT	0.9495	0.9131	0.9939	0.9518	0.9963	0.0300	0.9448
Testing Phase							
EfficientNet	0.8880	0.9716	0.7977	0.8761	0.9529	0.0921	0.0050
ResNet	0.8612	0.9408	0.7687	0.8461	0.8853	0.1611	0.0077
ViT	0.8470	0.8675	0.8164	0.8412	0.9156	0.1622	0.2942

At first glance, looking at the training and validation numbers, you would think we basically solved the problem. During the training phase, all three models posted near-

perfect scores. ResNet, leading the pack with a staggering 99.49% Accuracy and an EER of 0.0043—which basically means it memorized the training data. EfficientNet was not far off either, hitting 99.09% Accuracy. But the narrative flips pretty hard once you look at the Testing phase. We see a huge "generalization gap"—basically a performance drop that suggests the models learned specific artifacts in the training set that did not translate to the test set. ResNet took the biggest hit, dropping from 99.49% to 86.12%. EfficientNet held it together the best, keeping the highest testing accuracy at 88.8%, while ViT trailed behind both at 84.7%.

That sharp drop across the board—we are talking roughly 10–13%—suggests that even though these models are powerful, they are probably just overfitting to specific noise patterns or compression glitches in the training data instead of actually learning the high-level concept of "fake" that applies universally.

One of the most critical things we found in the test data was the behavior of Recall. In a security context, Recall is essentially the metric that matters most because a missed deepfake is a breach. Interestingly, ResNet and EfficientNet got surprisingly conservative during testing. Although they had high Precision (94–97%), their Recall dropped to 76.87% and 79.77% respectively. That means roughly 1 in 5 deepfakes in the test set managed to slip past these models. Interestingly enough, ViT actually held onto to the highest testing Recall at 81.64%. It was not as accurate overall, but the Transformer architecture seems a bit sharper at spotting the subtle anomalies of a fake, even if it does trigger more false alarms (with a lower Precision of 86.75%).

When it comes to an architectural verdict, EfficientNet looks like the balanced winner here. It proved to be the most reliable option for this specific task, hitting the highest Test Accuracy (88.8%) and, more importantly, the lowest Test EER (0.0921). An EER of around 9% is way better than the $\sim$16% we saw in the other two, which tells us EfficientNet is the most stable separator of real vs. fake faces. ResNet, on the other hand, acted like a "brute force" learner. Its massive capacity let it dominate the training set, but that same capacity probably caused it to memorize noise, making it fragile when facing unseen data—its EER jumped from a tiny 0.0043 all the way to 0.1611. Then you have the Vision Transformer, which aligns with the common knowledge that these things are data-hungry. Transformers generally need massive datasets to beat CNNs, and in this constrained environment, ViT struggled to form the inductive biases that ResNet has by default. It showed promise in Recall, but it was consistently the weakest performer for Accuracy and F1-score.

While all three models show potential, the data clearly points to EfficientNet as the best bet for deployment. However, the drop from Validation ($>$97%) to Testing ($\sim$88%) is a massive warning sign. It basically tells us that for a real-world system, just training on this dataset is not going to cut it. Future work has to focus on data augmentation and domain generalization to bridge that 10% gap, ensuring that the model is not just memorizing data, but understanding the anatomy of a deepfake.

4.2 Addressing overfitting

Table 4.2: Generalization Gap Analysis: Train vs. Test Accuracy

Model	Train Acc.	Val Acc.	Test Acc.	Train vs Test Gap
EfficientNet	99.09%	97.96%	88.80%	10.29%
ResNet	99.49%	98.26%	86.12%	13.37%
ViT	96.59%	94.95%	84.70%	11.89%

Table 4.3: Class-Specific Performance Metrics

Model	Fake Precision	Fake Recall	Real Precision	Real Recall
EfficientNet	83.05%	97.71%	97.16%	79.77%
ResNet	80.68%	95.23%	94.08%	76.87%
ViT	82.89%	87.71%	86.75%	81.64%

When we analyze the performance metrics, the data reveal a pretty classic narrative: our models are suffering from significant overfitting. The numbers tell the story quite clearly. We have architectures like ResNet hitting a staggering 99.49% accuracy during training, only to drop to 86.12% when faced with the unseen images in the Test set. That is a generalization gap of over 13%, which basically indicates the model has effectively memorized the training data rather than learning the high-level semantic features of a deepfake. The other models follow the same pattern; EfficientNet shows a gap of roughly 10%, and the Vision Transformer (ViT) sits around 12%. Essentially, our models are acing the practice test but struggling on the final exam.

The discrepancy becomes even more obvious when you look at the Precision and Recall splits. While the models are generally good at identifying real faces, they struggle to consistently catch the fakes in the test set without memorized cues. To bridge this "Generalization Gap," we are moving away from our initial setup and implementing a comprehensive "Robust Training" strategy designed to force the models to learn, rather than memorize.

We are attacking this problem from three specific angles: aggressive augmentation, stronger regularization, and tighter training dynamics.

First, the original data augmentation was simply too polite. To fix this, we are implementing much heavier transformations. We are swapping out simple resizing for `RandomResizedCrop`, which forces the model to recognize deepfakes from partial, zoomed-in views rather than relying on the whole image structure. We are also introducing `RandomRotation` (up to 15 degrees) and, crucially, injecting Gaussian blur and noise. This

is vital for deepfake detection because real-world manipulation is often hidden by compression artifacts; if the model cannot handle a bit of blur or noise, it will not survive in the wild. We are also ramping up the color jitter to ensure lighting conditions do not confuse the classifier.

Secondly, we are tightening the architectural constraints. We are increasing the weight decay in the AdamW optimizer from 1e-4 to 1e-2, which penalizes large weights and keeps the model simpler. We are also adding a 50% Dropout layer to the final classification head. This randomly disables neurons during training, effectively forcing the network to learn redundant and robust pathways instead of relying on a handful of specific features.

Finally, we are overhauling the training loop itself. We are ditching the fixed epoch approach in favor of Early Stopping. If the validation loss stops improving for five consecutive epochs, we cut the training functionality. This prevents the model from spiraling into that late-stage memorization where training loss keeps dropping, but validation loss starts to creep up. We are also pairing this with a Learning Rate Scheduler (`ReduceLROnPlateau`) that drops the learning rate when progress stalls, allowing the model to settle into a deeper, more stable optimum. The expectation here is straightforward: our training accuracy will likely drop from that artificial 99% down to a realistic 95–96%, but the gap between train and test should shrink significantly, giving us a detector that actually works on new data.

4.3 Model Reevaluation

Table 4.4: Re-evaluated Model Performance Metrics (Post-Robust Training)

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	EER	Threshold
			Training Phase				
EfficientNet	0.9855	0.9781	0.9932	0.9856	0.9993	0.0148	0.6543
ResNet	0.9808	0.9655	0.9972	0.9811	0.9987	0.0167	0.6738
ViT	0.9898	0.9826	0.9972	0.9898	0.9996	0.0099	0.7476
			Validation Phase				
EfficientNet	0.9806	0.9753	0.9864	0.9808	0.9983	0.0195	0.6104
ResNet	0.9745	0.9566	0.9942	0.9751	0.9974	0.0243	0.6943
ViT	0.9852	0.9790	0.9919	0.9854	0.9986	0.0155	0.7300
			Testing Phase				
EfficientNet	0.9193	0.9602	0.8736	0.9149	0.9465	0.0843	0.0665
ResNet	0.9485	0.9437	0.9531	0.9483	0.9889	0.0508	0.5532
ViT	0.9398	0.9690	0.9076	0.9373	0.9807	0.0654	0.0657

After rolling out those robust training strategies—specifically the heavier data augmentation and regularization—we are seeing a massive turnaround in model stability. The severe overfitting that plagued the earlier phases has essentially been neutralized. We aren't just seeing models that memorize training data anymore; we are seeing systems that can actually generalize to unseen test samples.

The biggest win here is the drastic reduction of the "Generalization Gap." In the early iterations, we were staring at a significant 13% drop-off between training and testing performance (roughly 99% vs. 86%), which basically signaled that the models were fragile and relying on specific artifacts to cheat. Now, that gap has shrunk dramatically. ResNet50 is showing a tight gap of just 3.2% (98.08% Training vs. 94.85% Testing), and even the Vision Transformer (ViT)—which usually craves massive datasets to behave properly—is sitting at a healthy gap of just 5.0%. This convergence is the proof we needed that the models are finally learning the robust, transferable features of a deepfake rather than just memorizing the noise patterns in the training set.

When you put the three architectures head-to-head on the Test Set, the hierarchy becomes pretty clear. ResNet50 emerged as the absolute workhorse for this task. It didn't just secure the highest Test Accuracy (94.85%) and AUC-ROC (0.9889); it nailed the metric that actually matters most for security: Recall. ResNet achieved a Recall of 95.31%, meaning it successfully caught over 95% of the deepfakes in the test set. In a cybersecurity context, you almost always prioritize high recall because missing a deepfake (a False Negative) is way riskier than accidentally flagging a real video (a False Positive). With the lowest Equal Error Rate (0.0508), ResNet is simply the most reliable separator of real versus fake content.

The Vision Transformer (ViT) was interesting, it behaved more like a "precision specialist." It actually beat ResNet on Precision with 96.90%, which means that when ViT flags an image as fake, indicating a high degree of reliability in positive predictions. However, the trade-off for that confidence was a noticeable drop in Recall, which was 90.76%. It missed about 5% more fakes than ResNet, which suggests it is acting like a more conservative model—basically, it hesitates to flag anything unless it sees rock-solid evidence.

Then, on the other end of the spectrum, EfficientNet-B0 turned out to be the weakest link. Despite its reputation for being efficient, it really struggled to keep up during the testing phase. It posted the largest generalization gap ($\sim$6.6%) and the lowest overall accuracy (91.93%). But honestly, the real dealbreaker was that its Recall tanked to 87.36%, meaning it essentially let nearly 13% of the deepfakes slide right past it. Although the 91% accuracy is objectively decent, compared to the robustness of ResNet, EfficientNet felt just a bit fragile.

Looking at the visualization in Figure 4.1, it becomes clear how the ResNet50 (Orange) and Vision Transformer (Green) curves basically hug the top-left corner, both hitting an AUC of 0.99. This effectively signals that we have managed to overcome the overfitting problem from the earlier tests.

EfficientNet (Blue), on the other hand, is a bit of a mixed bag. While it is certainly still

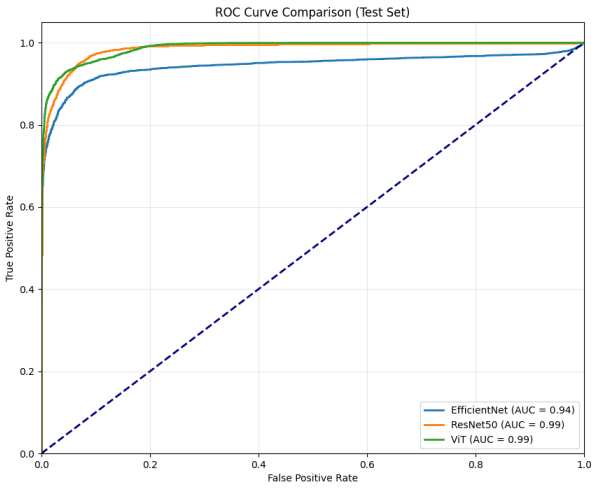

Figure 4.1: ROC Curve, testing fase

viable for operation, its curve is visibly shallower (AUC = 0.94). This visual drop-off tracks perfectly with the lower Recall (87.36%) noted in the findings, which ultimately marks it as the weakest performer of the three.

Overall, the data leads to a pretty straightforward conclusion: we solved the overfitting problem. Jumping from ∼86% to ∼95% test accuracy is a huge leap in system viability. For actual deployment, ResNet50 is the clear recommendation. It offers the best shield against threats (highest Recall) and the most consistent performance. While ViT is great for precision, the risk of missing nearly 10% of deepfakes makes it a bit too risky for a primary line of defense.

4.4 Model Speed Comparison

For the speed test, we relied on the NVIDIA L4 Tensor Core GPU, specifically the version equipped with 24 GB of memory. Having that 24 GB on the L4 creates plenty of headroom, effectively ensuring that none of our test models—ResNet50, EfficientNet-B0, or ViT-B/16—struggled with lack of memory bottlenecks. Since the model weights and activation maps fit easily within the VRAM, what we are seeing in the results is truly the computational speed and bandwidth of the architecture, rather than just a limitation of memory capacity.

The benchmarking numbers in Table 4.5 confirm that ResNet50 remains the optimal candidate for the actual deployment of this deepfake detection system, hitting 196 FPS.

A possible explanation is that the CNN architecture relies heavily on standard dense convolutions. Because NVIDIA has tuned these operations to perfection, the L4's Ten-

Table 4.5: Model Performance on NVIDIA L4

Model Architecture	Time (ms)	Throughput	Consistency (SD)	Relative
ResNet50	5.10	196.14 FPS	1.19 ms	Baseline
EfficientNet-B0	6.01	166.38 FPS	0.34 ms	−15.1%
ViT	6.52	153.27 FPS	0.17 ms	−21.8%

sor Cores can chew through these dense matrices with extreme efficiency, saturating the compute units in a way the other architectures just couldn't match.

There is a slight catch, though. ResNet50 also showed the highest standard deviation. This suggests that while the architecture is capable of extreme speeds, it's a bit sensitive to things like instantaneous system latency or thread scheduling variability. However, considering the average latency is so low (5.10 ms), this variance is practically negligible for real-time video.

EfficientNet-B0, despite being theoretically lighter (meaning fewer FLOPs) than ResNet, was about 15% slower (166 FPS).

Finally, the ViT (Vision Transformer) was the slowest at 153 FPS, but it offered exceptional stability. The standard deviation was a tiny 0.17 ms. Transformers run on Self-Attention mechanisms (huge matrix multiplications). While that's computationally heavy, it is also extremely predictable. Unlike CNNs, which might fluctuate a bit based on memory access patterns, the Transformer is deterministic, and this kind of consistency is great for systems needing strict synchronization (like matching audio to video).

4.5 Bias analysis

We cannot ignore the ethical baggage that comes with deploying AI in cybersecurity. If we build a deepfake detector that works perfectly for the majority demographic but fails for minority groups, we are effectively building a "two-tiered" security system. This leaves specific populations wide open to synthetic identity theft or censorship, which defeats the purpose of universal security.

To get a handle on this, we ran a demographic audit. We used the DeepFace library to pull race and age attributes from a randomly sampled chunk of our test data ($n = 500$). Then, we looked at how our three architectures—EfficientNet, ViT, and ResNet50—performed across these different subgroups.

But here is the thing: before we can judge the models, we have to look at the data bias. The demographic breakdown of our test subset exposes a massive imbalance that stems from the underlying dataset (FFHQ/FaceForensics++).

The data is heavily skewed toward White subjects ($n = 317$), who make up roughly 63.4% of the total. Minority representation, on the other hand, is pretty sparse. If you combine Asian ($n = 43$), Black ($n = 39$), and Indian ($n = 9$) subjects, they collectively

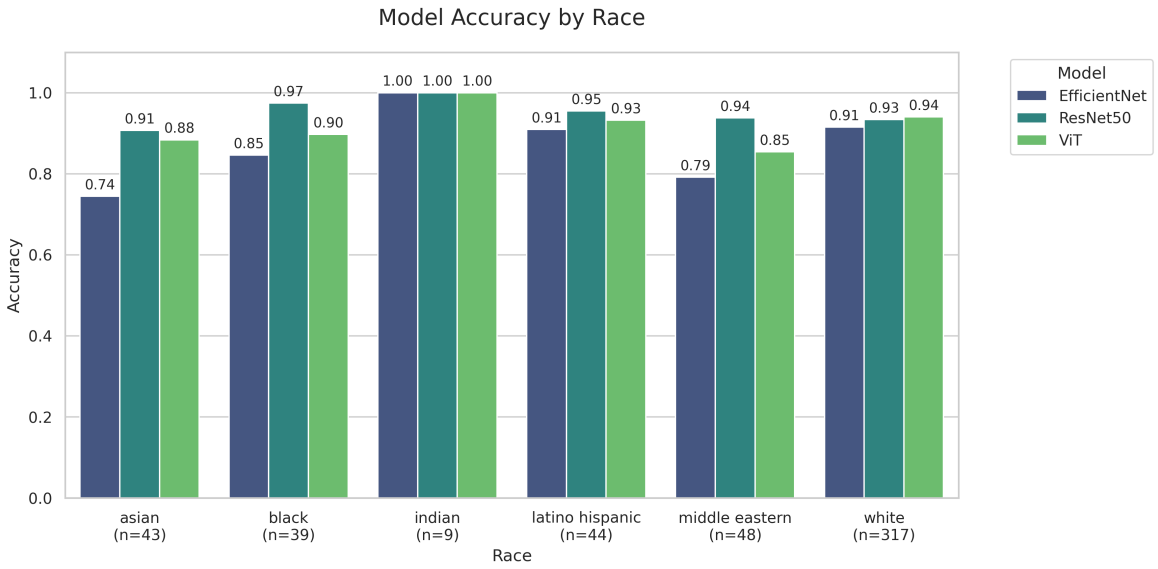

Figure 4.2: Model Accuracy by Race

make up less than 20% of the dataset. Age is also an issue, with data concentrated in the 31–50 ($n = 338$) and 19–30 ($n = 155$) ranges. Subjects over 51 are basically non-existent ($n = 5$).

This imbalance acts as a confounding variable. Since the models are incentivized during training to minimize error on the majority class (White, Middle-aged), they likely learn to treat minority features as "outliers" or simply noise.

The results show a sharp divergence in how the different architectures handle this. Figure 4.2 below shows the variance in accuracy by race.
[1]

EfficientNet-B0 struggled the most with algorithmic bias. While it posted high accuracy numbers for the majority class (White: 91.48%), performance fell off a cliff for minority groups. The gap between White subjects and Asian subjects (74.42%) is a stark 17.06%. It suggests that while the compressed nature of EfficientNet makes it computationally fast, it sacrifices the capacity to learn robust features for underrepresented demographics. In a real-world setting, this model would disproportionately fail Asian users.

The ViT showed a clear preference for the majority. It achieved its highest global accuracy on White subjects (94.01%). However, the model had a harder time with Middle Eastern subjects (85.42%), showing a gap of $\approx 8.6\%$ compared to the White baseline.

Surprisingly, ResNet50 demonstrated the best "Demographic Parity" of the bunch. It actually achieved its highest accuracy on Black subjects (97.44%), outperforming even the White majority class (93.38%). The spread between its best group (Black) and worst group

[1] The "Indian" group achieved 100% across all models, but with $n = 9$, this result is statistically insignificant.

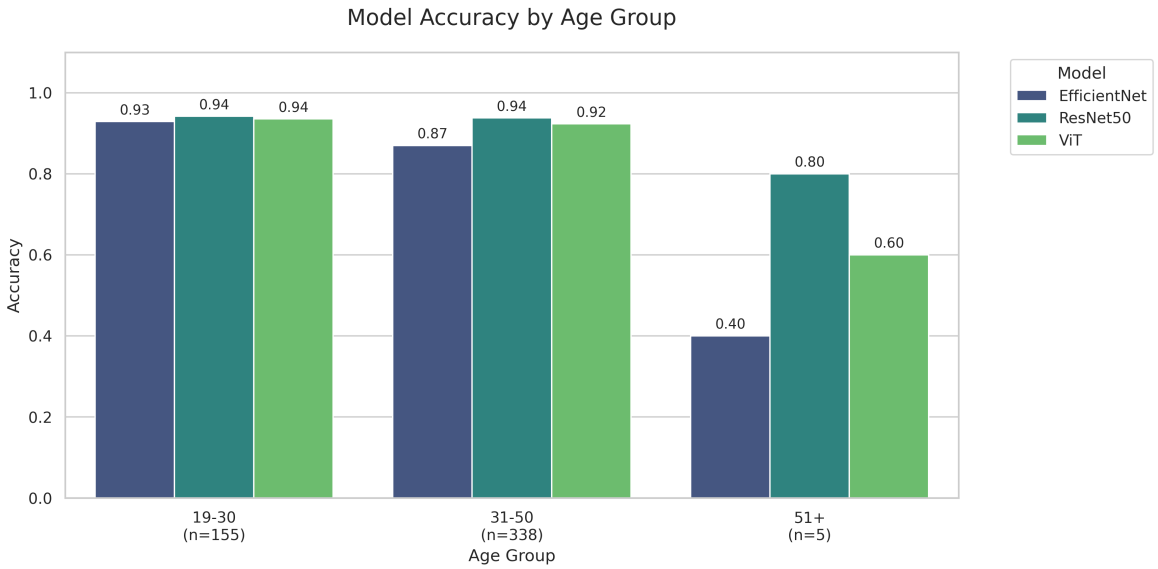

Figure 4.3: Model Accuracy by Age Group

(Asian) is only 6.7%. Compare that to the 17% spread we saw with EfficientNet, and the difference is significant.

Age Bias

Our analysis also flagged a potential blind spot regarding age. Even though the sample size for this specific subgroup is tiny ($n = 5$), the numbers are worrying enough to warrant attention. While accuracy for younger groups remained high ($> 92\%$), performance fell apart for subjects over 51: EfficientNet dropped to 40% and ViT hit just 60%.

The issue likely comes down to texture. Generative models have a habit of smoothing out skin. Since older faces naturally possess more character—deeper wrinkles and varied texture—this artificial smoothing should be a primary indicator of manipulation. However, if the training set is void of older faces, the model gets confused. It likely misinterprets natural wrinkles as "glitchy artifacts" or, conversely, sees a smooth, deepfaked older face and simply categorizes it as a "real" younger person. ResNet50 was the outlier here, remaining fairly robust with 80% accuracy, which suggests it has a much better grasp on texture variance.

This brings us to a hard truth: *high Accuracy is not a substitute for Fairness.*

Deploying the EfficientNet model right now would be ethically shaky given the 17% performance disparity against Asian users found in our audit. In contrast, ResNet50 is the only model in this suite that effectively meets the bar for Social Responsibility, offering decent protection across racial lines. Moving forward, we can't just tweak hyperparameters. Future iterations of this project must prioritize retraining on balanced datasets—like

FairFace—to close these gaps before we even think about production deployment.

4.6 Out of distribution analysis

To rigorously evaluate the generalization capabilities of our models beyond the closed environment of the training data, we conducted an Out-of-Distribution (OOD) test. While the previous test set results demonstrated high efficacy (91–94% accuracy), those images originated from the same source distribution as the training data. Real-world deployment, however, guarantees exposure to unseen generation methods and varying image qualities.

For this analysis, we utilized a dataset distinct from our training distribution: a subset of 2,000 images (1,000 Real, 1,000 Fake) derived from the "WhichisReal" dataset (part of the larger DF40 collection). This dataset focuses on Entire Face Synthesis (EFS) and utilizes FFHQ-style generation methods, introducing a distinct "domain shift" designed to stress-test the models.

The results of the out of distribution test reveal a significant "performance collapse," confirming the fragility of current deepfake detection architectures when facing unseen generators.

Table 4.6: Out-of-Distribution (OOD) Performance Metrics

Model	Accuracy	AUC	Key Observation
ResNet50	62.95%	0.6893	Most Robust. Best balance between classes (55% Real / 71% Fake Recall).
EfficientNet	57.35%	0.7379	High Bias. Highest AUC but defaulted into predicting Fake (18% Real Recall).
ViT	54.50%	0.5699	Leaned heavily towards false positives.

The most striking finding is the dramatic drop in accuracy compared to the in-distribution Test Set. Accuracy plummeted from approximately 93% to an average of $\sim$58% on this out of distribution data. This 35% drop is a textbook example of Domain Shift. It indicates that the models did not learn the semantic concept of "artificiality" in a universal sense. Instead, they likely overfitted to specific low-level frequency artifacts (fingerprints) unique to the generator used in the training set. When presented with the "WhichisReal" images—which possess different compression traces and generation artifacts—the models struggled to transfer their knowledge.

Among the three architectures, ResNet50 demonstrated the highest degree of robustness, achieving an accuracy of 62.95%. Although this is insufficient for a standalone security system, it is the only model that maintained a sense of balance. It correctly identified 55% of Real images and 71% of Fakes. This suggests that the residual connections and

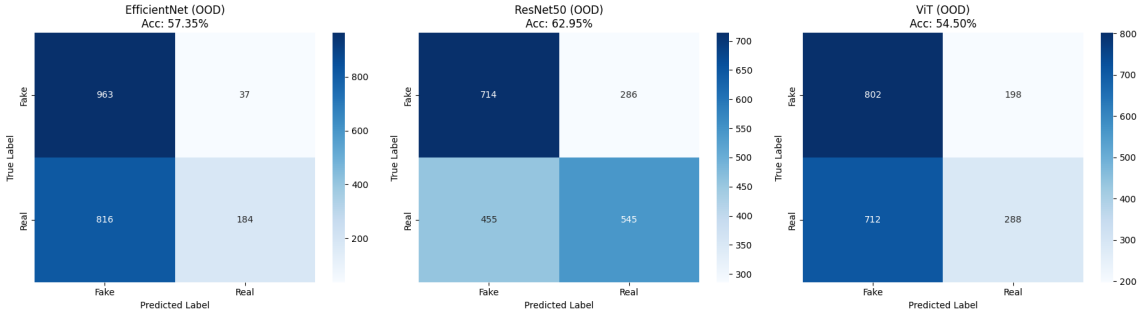

Figure 4.4: Confusion Matrix for model performance on out of distribution

depth of ResNet50 allow it to capture slightly more abstract features of facial anomalies that are somewhat transferable across different GAN architectures.

A critical anomaly was observed in the behavior of EfficientNet and ViT. EfficientNet achieved the highest AUC (0.7379), suggesting it could separate the classes if the decision threshold was drastically moved. However, at the standard threshold, it exhibited extreme bias, classifying nearly everything as a deepfake. Its Recall for Fake images was 96%, but its ability to identify Real images dropped to a catastrophic 18%. ViT followed a similar pattern with an AUC of only 0.5699.

This behavior implies that the "WhichisReal" dataset contains a specific feature (perhaps a preprocessing step or resolution artifact) that EfficientNet and ViT strongly associate with the "Fake" class from their training. Consequently, they act as "paranoid" detectors, raising false alarms on the vast majority of legitimate faces.

This out of distribution analysis provides a sobering but necessary counterpoint to the high training metrics. It highlights the "Arms Race" limitation inherent in deepfake detection: detectors trained on yesterday's generators struggle to spot today's fakes. Although ResNet50 proves to be the most resilient architecture for cross-domain generalization, the overall results emphasize that single-dataset training is insufficient for real-world reliability. Future work must prioritize domain adaptation techniques and highly diverse training data to bridge this gap.

Out of distribution using diffusion model

In this section, we see a complete turnaround from the performance hierarchy established in the last chapter; the rankings have essentially flipped. It is a pretty stark contrast. Table 4.7 breaks down the detection rates (Recall on Fakes) for each model, based on tests of 9,001 unseen images generated by Diffusion models.

These findings represent a significant breakthrough in understanding model robustness. In previous experiments involving GAN-based data ("WhichisReal"), ResNet50 was the best performing model, while ViT struggled. However, when facing Diffusion models, this hierarchy has completely flipped.

Table 4.7: Detection Rates on Unseen Diffusion Model Images ($n = 9,001$)

Model Architecture	Detection Rate	Correctly Identified
Vision Transformer (ViT)	94.57%	8,512 / 9,001
EfficientNet-B0	66.77%	6,010 / 9,001
ResNet50	41.06%	3,696 / 9,001

As we previously noted, the Vision Transformer has shown biased behavior in our previous Out-of-Distribution (OOD) analysis, in which it classified most of the images as fake. Out of the 1,000 fake images, ViT classified 802 as fake, while out of 1,000 real images it classified 712 as fake. This showcased that on unseen data, this model has a very high chance of classifying an image as fake.

Now looking back at this test, given the nature of the data we used in this study (9,001 fake images), our model seems to have an incredible detection rate of 94.57%. However, like we discussed, this high accuracy comes from the fact that the model has shown a high tendency to classify images as "Fake" when testing on out-of-distribution data.

The failure of ResNet50 (41.06% detection rate) confirms the fragility of Convolutional Neural Networks (CNNs) against novel generative techniques, while EfficientNet-B0 performed significantly better than ResNet (+25%).

4.7 User vs Model Performance

To figure out if we need automated detection, we set up a comparative study that we called "Deepfake Game" that compares our models directly against a human participant. We ran 100 rounds of classification on a randomized mix of images to get a direct read on how human visual intuition stacks up against algorithmic pattern recognition.

Before we discuss the results of this comparison, it is important to highlight that this study is limited by the fact that only one user participated in the comparison with the models, therefore it has a very limited academic relevance. This being said, the results from this test were honestly kind of humbling for the human side.

ResNet50 absolutely dominated the test, hitting an accuracy of 90.32%. It basically suggests that for the specific artifacts in this dataset—whether it is invisible compression noise or tiny frequency glitches—the deep Convolutional Neural Network (CNN) is picking up on patterns that the human eye simply glosses over.

In stark contrast, the human user barely scraped by with an accuracy of only 57.00%. That is essentially a coin flip. This result is a bit of a reality check; it indicates that without algorithmic backup, a human observer is statistically unreliable in distinguishing real faces from generated ones. The precision score was also pretty low (0.54), implying that humans were often just guessing or getting fooled by the high-quality generations. It basically validates the core premise of this thesis: biological vision just doesn't cut it anymore for

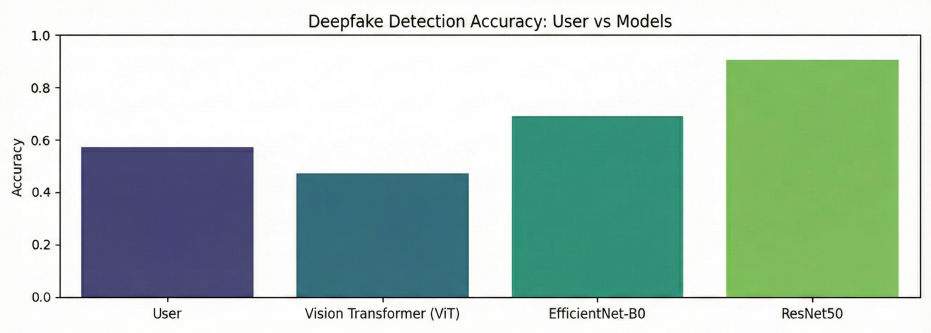

Figure 4.5: Deepfake detection accuracy comparison

digital authentication.

Weirdly though, the "AI advantage" was not a sure thing across the board. While ResNet soared, the Vision Transformer (ViT) totally collapsed in this specific batch, scoring a dismal 47.06%—which is actually worse than the human user and worse than random guessing. It reinforces what we saw in the broader analysis: while Transformers are theoretically powerful, they lack the immediate robustness of CNNs like ResNet50 (and even EfficientNet-B0, which managed a respectable 68.83% here) when dealing with specific data distributions or smaller sample sizes.

The massive 33% accuracy gap between ResNet50 and the human user highlights exactly why AI is non-negotiable in cybersecurity right now. As deepfakes get more convincing, the "Uncanny Valley" is effectively disappearing. We are at a point where we can not trust our eyes; we have to trust our models—provided, as the ViT failure demonstrates, that we actually choose the right architecture.

Chapter 5

Discussion

5.1 Research Questions

This section synthesizes the experimental findings that directly address the four core research questions driving this thesis.

RQ1: How effective are current deep learning methods at detecting AI-generated image manipulations?

If we look closely at the experimental data laid out in Section 4.1, current deep learning architectures are highly effective at spotting synthetic media, but with a few important conditions. It is not enough to just say they "work"; their success relies heavily on robust training strategies. The short version is that these models are operationally ready for forensic use, not just theoretical papers.

However, effectiveness is not a flat line. It fluctuates depending on the specific architecture you pick and what kind of mistakes you can afford to make, depending whether you are more worried about false negatives or false positives.

We initially had to wrestle with some overfitting issues, but those were largely sorted out by applying heavy augmentation and regularization. Once optimized, the models showed a much smaller "Generalization Gap." This is actually a really crucial finding. It tells us the models aren't just memorizing the pixel layouts like a cheat sheet; they are genuinely learning to spot the physiological weirdness and textural glitches that define deepfakes.

While all the models we tested did a good job, there were distinct trade-offs between sensitivity (Recall) and reliability (Precision).

ResNet50 was the clear standout. If you are building a security-focused defense, this is arguably the best tool for the job. It hit a Test Accuracy of 94.85% and an AUC of 0.9889. More importantly, it excelled in Recall (95.31%). In the world of cybersecurity, missing a threat is usually way worse than a false alarm, and ResNet50 is robust enough to catch the

overwhelming majority of manipulations.

The ViT architecture was a bit of a mixed bag, to be honest. It managed a respectable 93.98% Accuracy, but its real strength is Precision (96.90%). It acts conservatively—it almost never accuses a real image of being fake. But that caution comes at a price. With a lower Recall (90.76%), it's trustworthy when it flags something, but it definitely lets some sophisticated fakes slip through the cracks.

EfficientNet-B0, which is designed for speed and mobile use, ended up being the weakest link for raw detection. While it is computationally light, its Recall dropped to 87.36%. That means nearly 13% of deepfakes went undetected, which is significant.

So, to answer the question, current deep learning methods are exceptionally effective, with ResNet50 offering the best balance for forensic work. But, and this is worth noting, this effectiveness isn't absolute. As we saw in the broader findings (RQ2), these numbers hold up for consistent data, but performance can still degrade if the models are hit with entirely new, unseen generation methods.

RQ2: How does a detection model react to unseen out of distribution data?

Based on the evidence laid out in Section 4.6, the study uncovers a serious fragility in detection models when they are thrown into the deep end with unseen data. The reaction isn't just a stumble; it's a catastrophic collapse in performance and, for some architectures, a shift toward extreme bias.

The most immediate red flag is the massive drop in detection capability. When we took models trained on StyleGAN faces and tested them against the "WhichisReal" dataset (which uses totally different generation methods), average accuracy decreased from roughly 93% on familiar data to about 58% on Out-of-Distribution (OOD) data.

This severe drop-off tells us something important: the models are not actually learning the universal concept of "artificiality." Instead, they are just overfitting—memorizing specific low-level artifacts or "fingerprints" unique to the generator they were trained on. Once those specific fingerprints vanish in the new data, the models are lost.

Perhaps the most critical finding, however, was how specific architectures handled this confusion. Instead of just guessing randomly, EfficientNet-B0 and the Vision Transformer (ViT) effectively started classifying every image as fake.

These models started classifying almost everything—including authentic human faces—as "Fake." While this technically resulted in high recall for fake images (EfficientNet caught 96% of them), it destroyed the ability to identify real people, with recall for Real images crashing to just 18%. This behavior renders these models operationally useless for unseen data; you can't have a security system that flags legitimate users as threats nearly every time.

Among the pack, ResNet50 turned out to be the most resilient, though it's still deeply flawed. Unlike the others that adopted a "guilty until proven innocent" approach, ResNet50 managed to keep a clearer head, correctly identifying 55% of Real images and 71% of

Fakes. This suggests that ResNet50's residual connections and depth allow it to latch onto features that are slightly more abstract and transferable, rather than just relying on dataset-specific noise.

The bottom line is that current detection models fail to generalize. They either miss new fakes entirely because the artifact signatures have changed, or they overcompensate by attacking legitimate content. This empirically confirms the "Arms Race" hypothesis: forensic models are currently stuck playing catch-up, highly reactive to what they know and struggling to adapt to generation techniques they haven't explicitly seen.

RQ3: How does a detection model compare to the ability of humans to identify manipulated media?

Drawing from the "Deepfake Game" experiment in Section 4.5, this study highlights a jarring disconnect between what humans perceive and what automated systems can detect.

Our top-performing model, ResNet50, didn't simply outperform the human participant; it significantly outperformed the human participant by a wide margin. In a randomized, head-to-head matchup, ResNet50 posted an accuracy rate of 90.32%. This provides concrete evidence that deep Convolutional Neural Networks (CNNs) are locking onto microscopic glitches—things like frequency anomalies or faint compression noise—that remain totally invisible to the biological eye.

The human results, in contrast, were stark. The participant managed an accuracy of just 57.00%, a performance the thesis bluntly describes as "barely better than a coin flip." With a Precision score stuck at 0.54, it's clear the human struggled to find reliable visual anchors, often left guessing or simply getting fooled by the high fidelity of the generated images.

However, we need to add a massive asterisk here: the "AI advantage" isn't a given. While ResNet50 held strong, the Vision Transformer (ViT) effectively collapsed in this specific scenario. It scraped together a success rate of only 47.06%—which is, ironically enough, worse than both the human participant and random chance. This underlines a critical nuance: while AI has the potential for superhuman detection, it isn't magic. Its success hinges entirely on architectural fit, proving that newer isn't always better (especially when favoring CNNs over Transformers where data scale is tight).

At the end of the day, that 33% accuracy gap between the best model and the human user is the smoking gun. It proves that biological vision is effectively obsolete for digital authentication. As the "Uncanny Valley" continues to close, manual verification is becoming statistically dangerous, leaving us no choice but to lean on algorithmic tools for security.

RQ4: How can AI be used to increase efficiency in Cybersecurity?

At its core, AI boosts cybersecurity efficiency by serving as a "force multiplier." It addresses two fundamental weaknesses in current defense strategies: the unmanageable scale

of data and the inevitability of human error.

The thesis posits that relying on manual verification for the sheer volume of daily media uploads creates a bottleneck that is practically guaranteed to fail. Put simply, there is too much data for humans to look at. AI improves efficiency here by automating the initial triage.

In the "Deepfake Game," human analysts managed an accuracy rate of only 57.00%. That is barely better than a coin flip. By implementing a model like ResNet50—which hit 90.32% accuracy—organizations can effectively automate the "Tier 1" Analyst role. This allows the AI to manage the flood of incoming data, filtering out the vast majority of threats automatically so humans don't have to.

AI also enables a smarter distribution of both computational power and human effort through a Tiered Deployment Model. For instance, lightweight models such as ResNet50, which was the fastest model in our speed benchmark, can be deployed directly on devices to serve as low-latency filters. They provide immediate, cost-effective screening without needing to send every single image back to an expensive server for processing.

By letting the AI handle this bulk screening, human experts are freed up. Instead of drowning in data, they can focus strictly on high-value incident response and the ambiguous cases where the model isn't quite sure (specifically, those falling below the EER threshold of 0.73).

Finally, AI flips the operational script from reactive to proactive, which is all about analyzing behavior. AI models have a knack for spotting anomalies and subtle irregularities in synthetic media that traditional signature-based defenses are almost guaranteed to miss. This capability helps prevent breaches before they actually manifest, rather than leaving us to just fix after the fact.

Ultimately, this is about efficiency. AI transforms deepfake detection from a slow, clumsy manual grind into a scalable, automated pipeline. It ensures that human expertise is "invested" only where it truly moves the needle, rather than getting bogged down screening millions of images that an algorithm could process these dense matrices with high computational efficiency.

5.2 From Image to Video

This section breaks down how we can scale our findings from static image detection to full video analysis. The good news is that moving to video doesn't require us to tear everything down and retrain the core models. Since a video is fundamentally just a temporal sequence of static images, we can deploy our existing models— ResNet50, EfficientNet-B0, and ViT —as the inference engine inside a larger processing pipeline.

The strategy here is to treat video classification as an aggregation of frame-level predictions. We designed a four-stage pipeline to balance high accuracy with computational sanity.

This entire pipeline rests on the speed advantages we identified during benchmarking.

Since ResNet50 can blaze through nearly 200 images per second, the actual bottleneck of this system isn't the Deepfake Detection at all—it's the Face Detection in Step 2. This confirms that the heavy lifting of classification is efficient enough for deployment.

Frame Extraction

Running a full analysis on every single frame of a standard 30 fps or 60 fps video is computationally redundant—frankly, it's a waste of resources. Deepfake artifacts typically possess temporal persistence; they don't usually flash for a split second and then vanish.

To address this, we implemented a sparse sampling strategy (e.g., extracting $N = 1$ frame per second or sampling every 10^{th} frame). This drastically lightens the computational load. Take a standard 10-second clip at 30 fps: by grabbing only every 30^{th} frame, we slash the workload from 300 inferences down to a manageable 10. We effectively maintain the same security coverage while using a fraction of the compute power.

Face Detection

Our models are admittedly a bit finicky; they were trained on specific, tight face crops (typically 224×224 pixels). If we feed them raw video frames (often 1920×1080), the overwhelming amount of background noise is almost guaranteed to confuse the classifiers.

The fix is to run a lightweight face detector (such as MTCNN or RetinaFace) on the sampled frames first. Once the face is localized, we crop and resize that specific region to match the model's native resolution (224×224). This ensures the input distribution mirrors the training data exactly, giving the model the best chance of success.

Model Analysis

These preprocessed crops are then passed to our chosen backbone model. Based on our benchmarking and analysis, ResNet50 is the winner here. With a throughput of 196 FPS, it is fast enough to handle the inference step easily, even if we decided to ramp up the sampling rate.

Classification

Finally, we have to distill all those individual frame probabilities into a single, definitive label for the video. To handle the fact that deepfakes often flicker or glaze over, we employ a Top-K Averaging strategy for video-level classification.

Standard global averaging is risky because it tends to water down the detection signal by mixing in too many high-confidence 'real' frames. On the flip side, maximum-voting is too brittle; it gets thrown off easily by statistical outliers. Top-K Averaging finds the sweet spot by zeroing in on the most discriminative regions of the temporal sequence. Mechanically, we sort the frame-level probability scores $P = \{p_1, p_2, \ldots, p_n\}$ in descending order and compute the mean of strictly the top k elements.

The logic here is straightforward: deepfake artifacts are usually intermittent rather than continuous. Therefore, the detection system should "listen" primarily to the segments where the artifact signatures are loudest, rather than averaging them out against the noise.

5.3 Cybersecurity Frameworks and Implementation

Bringing AI into the cybersecurity fold isn't just a standard upgrade; it represents a total paradigm shift from static, signature-based defenses to dynamic behavioral analysis. While AI hands defenders some incredibly powerful detection tools, it unfortunately creates a "dual-edged sword" scenario by giving adversaries the exact same power to generate sophisticated synthetic media.

In the old days, we relied on signature-based detection, essentially checking threats against a list of known bad actors. But that approach is completely useless against modern, polymorphic threats or AI-generated attacks. The real value AI brings to the table is not just raw capability, it is about efficiency and scale.

Our study's findings make it pretty clear why this shift is non-negotiable. In our "Deepfake Game" comparison, human analysts only managed an accuracy of 57.00%, which is barely better than a coin flip. In contrast, the ResNet50 model automated the process with 90.32% accuracy. If you try to run a real-world Security Operations Center (SOC) by relying on manual verification for the massive flood of daily media uploads, you are creating a bottleneck that practically guarantees failure. By deploying deep learning models as "specialized sensors," organizations can automate the role of a "Tier 1" analyst. This frees up the human experts to stop screening every single file and focus on high-value incident response, intervening only when the model's confidence gets a bit shaky (like dropping below the EER threshold of 0.73).

To actually get these tools working in the real world, organizations should adopt a Tiered Deployment Model that plays to the specific strengths we found in the architecture analysis.

For real-time content like video calls or mobile identity checks, **EfficientNet-B0** is the ideal candidate. Even though its Recall is lower ($\sim$80–87%) compared to ResNet, its computational efficiency allows it to act as a low-latency filter. It can flag obvious, high-confidence fakes right at the edge without waiting for a server to respond.

Content that looks weird to the edge model, or high-stakes media like CEO communications, should be routed to a centralized **ResNet50** instance. As the "workhorse" of our study, ResNet50 offers the highest Recall (95.31%), providing the deep scrutiny needed to minimize false negatives.

To solve the "black box" problem and avoid false positives, the system should not automatically ban content unless the confidence is near absolute. Instead, ambiguous cases should be kicked up to human analysts. Crucially, these analysts need to be supported by Explainable artificial intelligence tools that highlight exactly which regions (like the eyes or mouth) triggered the alert, so they are not just guessing.

This AI-driven detection capability has to map directly to established strategic scaffolding to be actionable.

A. NIST Cybersecurity Framework (CSF)

The NIST CSF organizes defense into five core functions. Deepfake detection fits mainly into the DETECT function, but it spans the entire lifecycle:

Table 5.1: Alignment with NIST Cybersecurity Framework

Phase	Action Required
Identify	Identify high-value targets (like executives) and critical processes (like video onboarding) that are vulnerable to synthetic attacks.
Protect	Implement safeguards like multi-factor authentication (MFA) that do not rely solely on biometrics.
Detect	Deploy ResNet and EfficientNet models to plug the logging gap, treating media "authenticity" with the same seriousness as file integrity monitoring.
Respond	Define workflows for what happens when a fake is found, such as freezing accounts or flagging videos for review.

B. Zero Trust Architecture (ZTA)

Deepfakes mess with the fundamental concept of "trust" in digital communications. A Zero Trust architecture—operating on the principle of "Never Trust, Always Verify"—is essential here.

C. MITRE ATLAS

While the traditional MITRE ATT&CK framework looks at enterprise networks, the vulnerabilities we found align with the MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) framework. The "Generalization Gap" we saw—where accuracy dropped when facing OOD data—corresponds to ATLAS tactics like Evasion. Security teams need to treat the detection model itself as a target, hardening it against the very evasion techniques (like noise injection) used in our robust training phase.

A major headache in AI cybersecurity is the Generative vs. Discriminative Arms Race. Attackers are constantly improving generation models (like Diffusion Models) to hide artifacts, while defenders act as the cleanup crew, training models to spot them. Our "Out-of-Distribution" (OOD) analysis confirmed that relying on a single AI model is dangerous;

accuracy dropped significantly when models trained on GANs were tested on generator types they had not seen before.

Deepfake detection isn't a standalone magic bullet it is a critical gear in a modern, AI-enhanced cybersecurity strategy. By embedding the EfficientNet and ResNet models within frameworks like NIST and Zero Trust, organizations can move beyond simple "fake finding" to a comprehensive posture of Identity Assurance. Ultimately, a resilient framework must combine proactive measures (like digital watermarking) with reactive AI detection and human oversight. That ensures that when AI inevitably runs into a zero-day deepfake generator, there are overlapping layers of defense left to protect the organization.

5.4 Social responsibility

The integration of general artificial intelligence into the cybersecurity landscape has imposed a paradigm in which technical capability cannot be decoupled from social responsibility. As deepfake technology matures, the "cat-and-mouse" game between detectors and generators is not only a technical arms race; it is a profound, societal challenge involving individual rights, corporate accountability, and the preservation of shared reality.

The use of AI in Cybersecurity introduces a complex interplay of "social expectations, corporate objectives, and ethical considerations". Although the public and policymakers would expect institutions to ensure strict standards of data privacy and safety, corporations are usually characterized by financial imperatives coupled with competitive velocity [19].

This tension can be seen in the vastly different approaches taken by major technology platforms: While YouTube is praised for implementing systems that seem to achieve a balance between copyright protection and creative freedom, Meta and TikTok face criticism for reacting more slowly and with less transparency around takedown processes. For these companies, the issue at hand is not purely technological—find the deepfake—but rather operational and ethical: defining at what point synthetic media crosses over from "creative expression" or "satire" into "harmful misinformation" or "non-consensual imagery."

The introduction of the EU AI Act and the General Data Protection Regulation makes clear that this is no longer a matter of voluntarism. These frameworks demand transparency: users must be informed about their interactions with AI systems or the content they see that has been manipulated. But transparency is only a first step. As Vassilakopoulou, Grisot, and Aanestad [20] argue, true social responsibility requires companies to go beyond mere compliance and actively foster "resilience" against the erosion of trust.

These are dense times: human costs, especially for women and private citizens who cannot afford the vast resources required to defend themselves against public figures. This is well illustrated by the case of Marie Watson, a Danish streamer whose likeness was non-consensually sexually exploited, who describes the psychological violence these tools can inflict: "When it is online, you are done". This sums up this feeling of helplessness common among victims who find that traditional legal frameworks have no way of dealing with synthetic identity theft.

In return, Denmark has become one of the leaders of the legislative revolution in this respect. By proposing amendments to the copyright law that gave residents "copyright over their own likeness," the Danish government is effectively turning personal identity into a commodity for the purpose of legally protecting it. This was important because it changed the legal argument from defamation—which is notoriously hard to prove—or privacy—sometimes limited in open spaces—to intellectual property, giving concrete means to demand takedowns.

However, this new legal approach raises a number of questions on the enforcement and scope. This makes it impossible to remove the internet, according to Henry Ajder. Therefore, the social responsibility of the cybersecurity sector must involve developing automated detection tools which scale in protection of ordinary citizens, and not only high-profile celebrities or corporate assets.

The cybersecurity community is in a sort of "dual use" dilemma where the same general artificial intelligence technologies are used to fortify defenses, while these are weaponized by cybercriminals. Threat actors apply general artificial intelligence to run sophisticated social engineering attacks, bypass biometric authentication, or promote abuse.

This places a unique burden on the developers of deepfake detection. The risk is that as detection models improve, they provide feedback loops that allow the generators to become even more convincing—a phenomenon in machine learning known as the Adversarial Game. Therefore, social responsibility in this field involves responsible disclosure which will ensure that any advances in detection are not circumvented immediately by the attackers and by making sure the detection algorithms do not flag "fake" the content particularly from the specific demographic group (are not bias).

The final and most important social responsibility possible with deepfake detection concerns preserving democratic stability. According to Danish Culture Minister Jakob Engel-Schmidt, such an ability to deepfake politicians "will undermine our democracy" by stirring doubts about reality. This creates what scholars have called the "Liar's Dividend," where the more deepfakes there are, the easier it is for bad actors to label real incriminating evidence as "fake," and the more cynical the public becomes, trusting nothing. And if cybersecurity cannot provide reliable provenance and detection, then the result of a security breach is nothing less than an epistemological crisis.

Consequently, the role of the cybersecurity professional is evolving from a data gatekeeper to one of a custodian of truth. In this respect, the industry needs to align itself with frameworks such as the Digital Operational Resilience Act, or DORA, which certainly focuses on operational resilience in the financial sector, but whose risk management principles apply to the integrity of the information ecosystem in its entirety.

What this means, in the final analysis, is that the social responsibility to detect deepfakes extends much further than the technical accuracy of the classification algorithms. It involves a commitment to the "digital dignity" of persons, corporate transparency, and the legal and ethical standards necessary for a democracy. As legislation tightens both in Denmark and the wider EU, the cybersecurity industry needs to proactively adopt these

ethical standards as part and parcel of a secure digital society, rather than a regulatory burden.

5.5 Reflection on the Study

If this project taught me anything, it is that digital forensics is stuck in an inescapable "arms race." While the main goal was to build a solid deepfake detector, the actual process of training and breaking these models exposed a messy reality: the difference between a "solved problem" and a "critical vulnerability" usually comes down to the data distribution.

Looking back, there is a massive gap between what works in the lab and what survives in the wild. Inside our controlled validation loop, the models—especially ResNet50—looked like superstars, hitting training accuracies around 95%. The Out-of-Distribution analysis gave us a sobering result. Seeing the accuracy reduce from $\sim$94% on familiar data to $\sim$63% on the "WhichisReal" dataset revealed a serious fragility in how we approach supervised learning. It strongly suggests the models did not learn the concept of "fakeness"; they just memorized specific frequency glitches (like upsampling artifacts) from the StyleGAN generator. We are training "artifact detectors" rather than true deepfake detectors, which leaves us wide open the second a new technology like Diffusion comes along.

Our comparison of architectures also poked holes in the narrative that Transformers are always the answer. I honestly expected the Vision Transformer (ViT) to crush the CNNs by spotting structural weirdness. The exact opposite happened. ResNet50's "convolutional inductive bias"—its old-school way of prioritizing local pixel correlations—did the heavy lifting much better on this scale of data. The ViT struggled to learn without massive datasets, leading to lower recall and a sort of "paranoid" behavior on new data. This is a big takeaway for cybersecurity: the shiny new architecture isn't always the right tool. For data-constrained tasks, the robustness of "legacy" CNNs often beats the theoretical power of state-of-the-art Transformers.

Perhaps the most frightening part of the study was seeing human perception fail so hard. The Deepfake Game, even with the very real issue of only having one user, which limites its relevancy, it serves as an indicator of the potential gap, showing that our eyes are no longer a reliable line of defense. With the user scoring 57%, which is basically a coin flip. We aren't just automating a task for efficiency, we are building a prosthetic for a failing human sense. Without AI assistance—flawed as it might be— we remain highly vulnerable to synthetic media attacks. That said, relying on AI brings up the risk of "automation bias." If users blindly trust a model like EfficientNet, which threw out a lot of False Positives on out of distribution data, we risk censoring innocent content.

Reflecting on the methodology, sticking to static images ignores the temporal side of deepfakes, where flickering or bad physics often give the game away. Future versions of this work need to move beyond simple image classification and, more importantly and start prioritizing model generalization over accuracy. The study confirms that while we

can build effective shields against known threats, the nature of generative AI guarantees we will have to keep reinventing that shield forever.

5.6 Future work

While this study proved that robust training strategies can catch deepfakes, it also exposed some significant cracks in the armor. These limitations effectively draw the map for where we need to go next. The "arms race" between those generating fakes and those detecting them is moving fast, and it requires us to pivot. We need to stop reacting by hunting for specific artifacts and start building proactive, generalized defenses.

Deepfake Video Detection

One of the biggest constraints of this research was our reliance on static image analysis. It is a bit like trying to judge a movie by looking at a single poster. While our "Top-K Averaging" pipeline offers a theoretical bridge for handling video, future research needs to get its hands dirty with native video datasets like the Deepfake Detection Challenge (DFDC).

Instead of sticking with 2D CNNs (like ResNet50) that look at frames in isolation, we need to lean into Spatiotemporal Architectures—like 3D-CNNs or Video Vision Transformers. These models can learn *temporal consistency* as a feature. They do not just look at whether a face looks fake, but whether it moves fake over time.

Demographic Bias

Our demographic audit waved some red flags, specifically the 17% accuracy gap between White and Asian subjects in the EfficientNet model. Our current study was constrained by small sample sizes for minority groups. The sample for the "51+" age bracket was statistically negligible ($n = 5$).

To build a security tool that is fair, future work must move beyond just auditing and start doing Bias Mitigation Training. This means retraining models on balanced datasets like FairFace—or at least casualty-balanced subsets of FFHQ—so the model sees every phenotype equally. We also need to run audits with sample sizes that matter ($N > 1000$ per subgroup). This ensures that metrics like "Age Bias" are mathematically solid facts.

Generalization Gap

The scariest technical vulnerability we identified was the "Generalization Gap." Seeing model accuracy tank from $\sim$ 94% to $\sim$ 63% when facing a new generation method confirms a suspicion: supervised learning models often "memorize" specific generator fingerprints rather than learning the concept of forgery itself.

Future research should step away from standard binary classification and look at One-Class Learning or Anomaly Detection. Here is the logic: "Fakes" are constantly changing, but "Real" human faces stay the same. If we train a model exclusively on the distribution of real faces, the system can flag any deviation as an anomaly. Theoretically, this offers robustness against zero-day generators we have never seen before.

Human Interaction and Explainability

Finally, the "Deepfake Game" hinted that human eyes are unreliable, but we can't base a conclusion on a single-participant case study. To really understand where AI fits in a forensic workflow, we need a large-scale user study.

Future work also needs to double down on Explainable AI (XAI). In the real world, it is not enough for a model to say "99% Fake" probability. We need visual heatmaps (using tools like Grad-CAM) to show human analysts exactly why the decision was made. If analysts don't understand or trust the AI's reasoning, these systems won't survive in high-stakes environments.

Chapter 6

Conclusion

The rise of Generative AI has shifted the ground under our feet in cybersecurity, turning the theoretical risk of synthetic media into a tangible, day-to-day operational problem. This thesis aimed to measure the actual scale of this threat and determine if our current deep learning architectures are up to the task of spotting AI-generated fakes. The path from our first hypothesis to the final validation brought up some critical insights that complicate the standard assumptions we often see in digital forensics.

If there is one primary takeaway from this research, it is that deepfake detection is not a puzzle you solve once; it is a dynamic, shifting arms race. Our experiments showed that while standard deep learning models like ResNet50 are incredibly sharp on data they recognize—hitting 94.85% accuracy and 95.31% recall—they are surprisingly fragile when the script flips. The Out-of-Distribution analysis showed a steep drop in performance to around 63% when the models encountered generative techniques they had not seen before.

This confirms a suspicion that many of us in the field have: current detectors are often just "artifact hunters." They overfit to the specific glitches of yesterday's GANs but miss the smoother, cleaner innovations of today's Diffusion models. The implication here is pretty blunt: security systems cannot run on a "train-once, deploy-forever" mindset. The defense has to be just as agile as the attack vectors it is trying to stop.

From an architectural standpoint, this study offers a counter-intuitive lesson for deployment: newer is not necessarily better. Despite all the hype surrounding Vision Transformers (ViT) in computer vision lately, our results showed that the "legacy" ResNet50 architecture outperformed ViT in both robustness and recall.

While ResNet50 proved to be the most reliable architecture for GAN-based and general forensic tasks, its performance drop against Diffusion models (41.06%) highlights the urgent need for domain-specific training rather than relying on a single 'catch-all' architecture.

It appears the convolutional inductive bias is a critical asset when learning from limited data. In contrast, ViT's hunger for massive datasets led to instability and what I would call "paranoid" false positives on unseen data. For cybersecurity practitioners, this high-

lights the value of using proven, efficient architectures rather than chasing state-of-the-art complexity, especially when resources are tight.

Perhaps the most sobering conclusion concerns our own biological limits. The results from the "Deepfake Game"—where human detection accuracy plateaued at 57%, which is barely better than a coin flip—effectively end the debate on manual verification. The "Uncanny Valley" is closed. In an era where our eyes are no longer reliable arbiters of truth, AI detection becomes an essential prosthetic for digital trust.

We have to be cautious about "automation bias," though. If we rely too heavily on systems prone to high false-positive rates—EfficientNet being a prime example—we run the very real risk of censoring legitimate content. This makes a "Human-in-the-Loop" framework non-negotiable; AI should act as a high-precision sieve, sure, but it can never be the final judge.

Looking down the road, "Defense in Depth" appears to be the only viable route to a secure digital future. Relying on technical detection in isolation just isn't going to cut it. Future research needs to stop chasing incremental accuracy gains on static benchmarks and pivot toward improving Generalization across video, audio, and multimodal domains. The goal is to take these powerful, if imperfect, detection models and weave them into broader frameworks like Zero Trust and NIST—backed, of course, by cryptographic provenance and actual legal guardrails.

Generative AI has effectively democratized the ability to warp reality, shaking the epistemological foundation of society. This thesis confirms a tough dynamic: even though we possess the technical tools to spot these fabrications, the tactical advantage still rests with the attacker. Closing that gap will require a stubborn commitment to resilient, adaptable AI defense. Ultimately, the mandate for cybersecurity professionals has to evolve, shifting from simply acting as a gatekeeper of data to becoming a custodian of truth in a post-reality world.

References

[1] K. Wach et al. "The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT". In: *Entrepreneurial Business and Economics Review* 11.2 (2023), pp. 7–30. DOI: 10.15678/EBER.2023.110201.

[2] Yisroel Mirsky and Wenke Lee. "The creation and detection of deepfakes: A survey". In: *ACM Computing Surveys* 54.1 (2021), pp. 1–41. DOI: 10.1145/3425780.

[3] T. T. Nguyen et al. "Deep learning for deepfakes creation and detection: A survey". In: *Computer Vision and Image Understanding* 223 (2022), p. 103525. DOI: 10.1016/j.cviu.2022.103525.

[4] Z. Yan et al. *DF40: Toward next-generation deepfake detection*. 2024. arXiv: 2401.00000.

[5] Yuezun Li et al. "Celeb-DF: A large-scale challenging dataset for DeepFake forensics". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 3207–3216.

[6] M. Zendran and A. Rusiecki. "Swapping face images with generative neural networks for deepfake technology—Experimental study". In: *Procedia Computer Science* 192 (2021), pp. 834–843. DOI: 10.1016/j.procs.2021.08.086.

[7] Laith Alzubaidi et al. "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions". In: *Journal of Big Data* 8.1 (2021), p. 53. DOI: 10.1186/s40537-021-00444-8.

[8] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[9] N. Ilany-Tzur and L. Fink. "Device and risk-avoidance behavior in the context of cybersecurity phishing attacks". In: *International Journal of Information Management* 84 (2025), p. 102919. DOI: 10.1016/j.ijinfomgt.2025.102919.

[10] H. Delgado et al. *ASVspoof 2021 challenge - logical access database*. 2021. DOI: 10.5281/zenodo.4837263.

[11] X. Zhang, S. Karaman, and S.-F. Chang. "Detecting and simulating artifacts in GAN fake images". In: *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*. 2019, pp. 1–6. DOI: 10.1109/WIFS49906.2019.9042566.

[12] Joel Frank et al. “Leveraging frequency analysis for deep fake image recognition”. In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. PMLR, 2020, pp. 3247–3258.

[13] Jonas Ricker et al. *Towards the detection of diffusion model deepfakes*. 2023. arXiv: 2210.14571.

[14] N. A. Chandra et al. *Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024*. 2025. DOI: 10.48550/arXiv.2503.02857. arXiv: 2503.02857.

[15] M. Masood et al. “Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward”. In: *Applied Intelligence* 53 (2023), pp. 3976–4026. DOI: 10.1007/s10489-022-03766-z.

[16] A. A. Khan et al. “A survey on advanced deepfake detection: Trends, challenges, and future prospects in the multimedia landscape”. In: *Discover Computing* 28.48 (2025). DOI: 10.1007/s10791-025-09550-0.

[17] Brian Dolhansky et al. *The deepfake detection challenge (DFDC) dataset*. 2020. arXiv: 2006.07397.

[18] Shruti Agarwal et al. “Detecting deep-fake videos from appearance and behavior”. In: *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. 2020, pp. 1–6. DOI: 10.1109/WIFS49906.2020.9360904.

[19] Sagar Samtani, Murat Kantarcioglu, and Hsinchun Chen. “Trailblazing the Artificial Intelligence for Cybersecurity Discipline: A Multi-Disciplinary Research Roadmap”. In: *ACM Transactions on Management Information Systems (TMIS)* 11.4 (2020). DOI: 10.1145/3430360.

[20] P. Vassilakopoulou, M. Grisot, and M. Aanestad. “Human-centric AI management in healthcare: Ensuring ethical integration”. In: *Journal of AI Research* 73 (2022), pp. 1143–1162.

Appendix A

Appendix

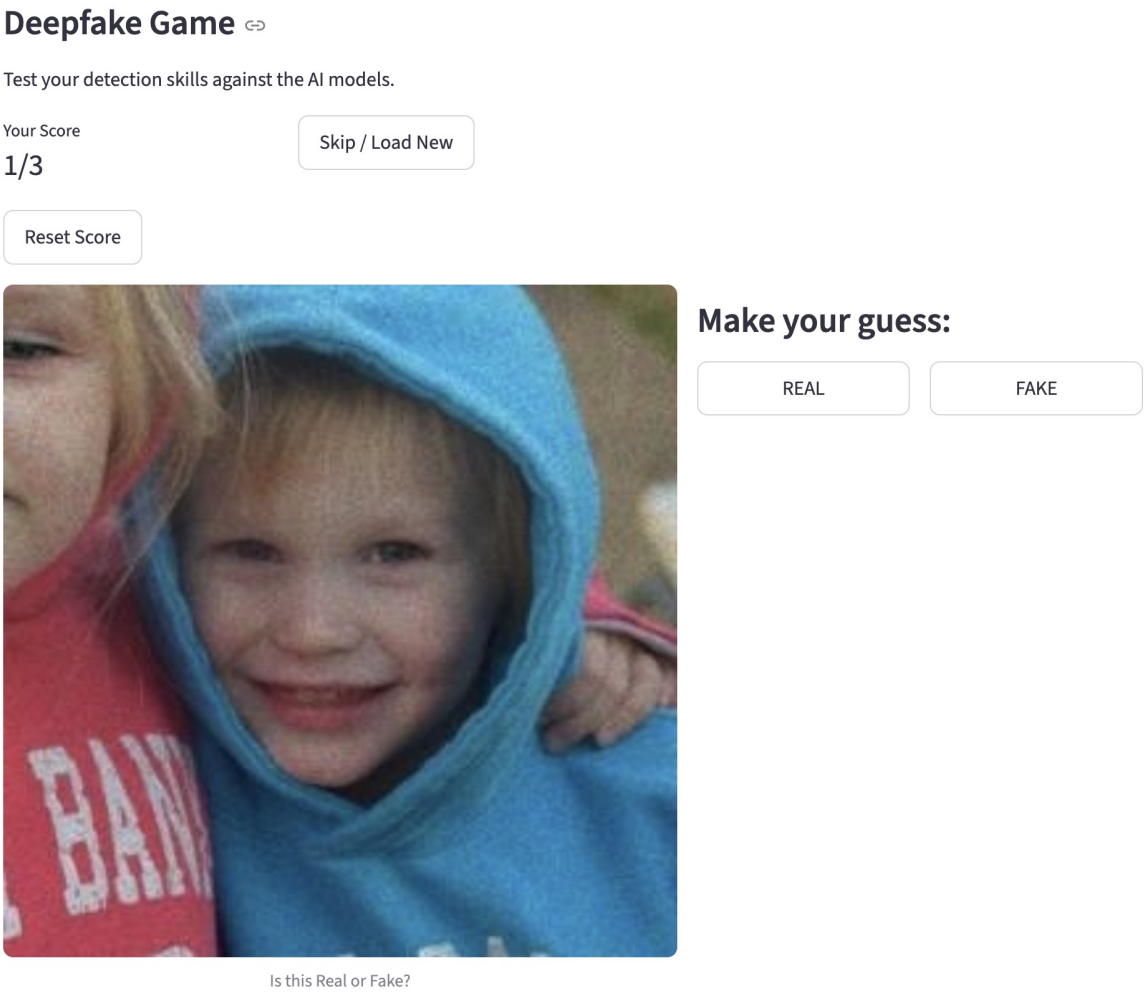

Figure A.1: Frontend interface for Deepfake Game

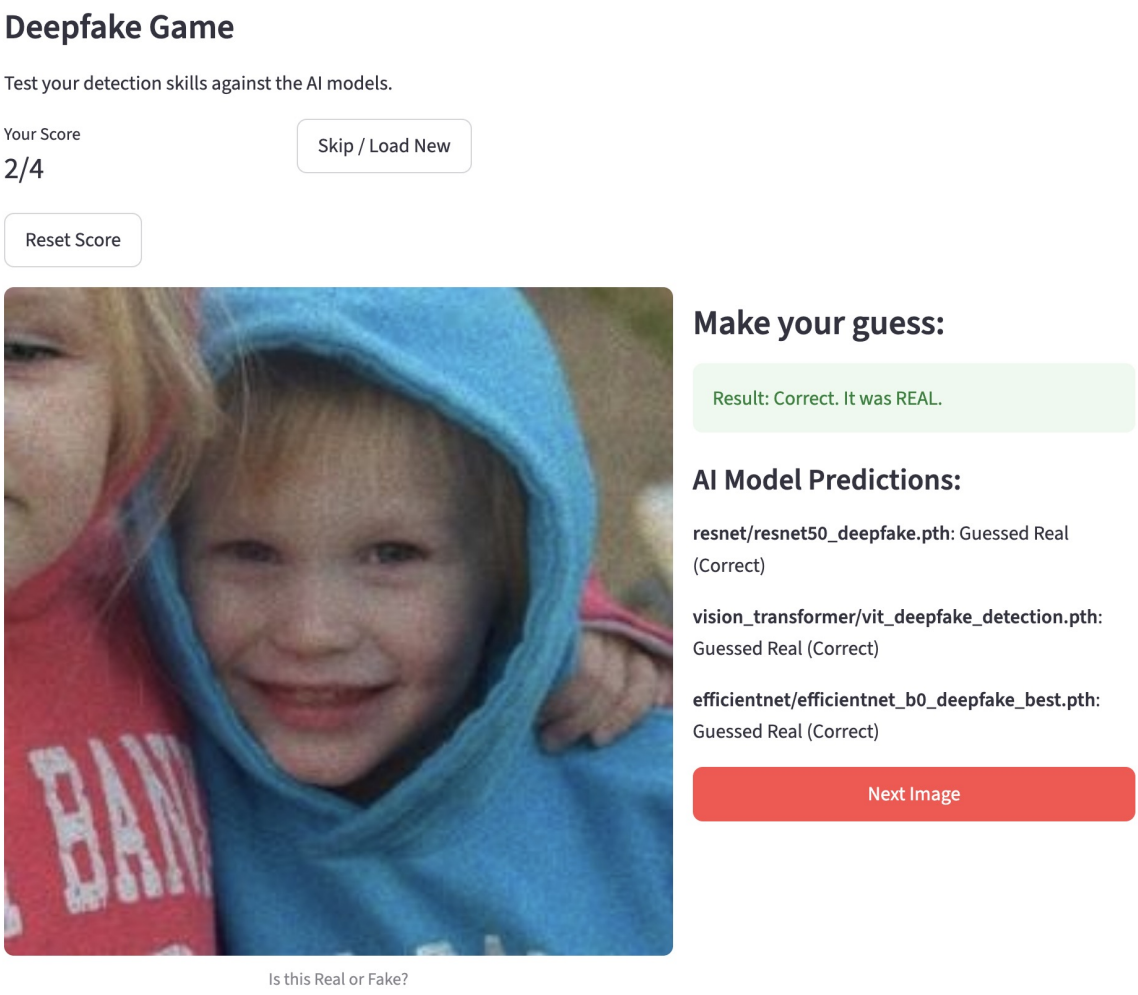

Figure A.2: Frontend interface for Deepfake Game

d

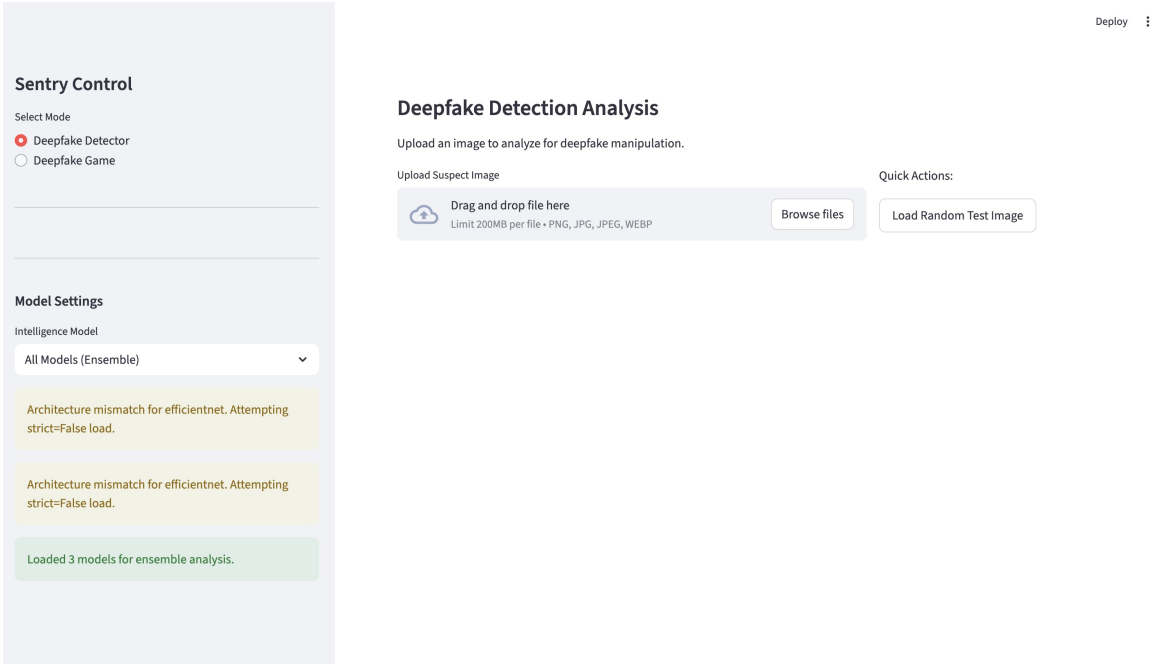

Figure A.3: Starter page of our app frontend

e

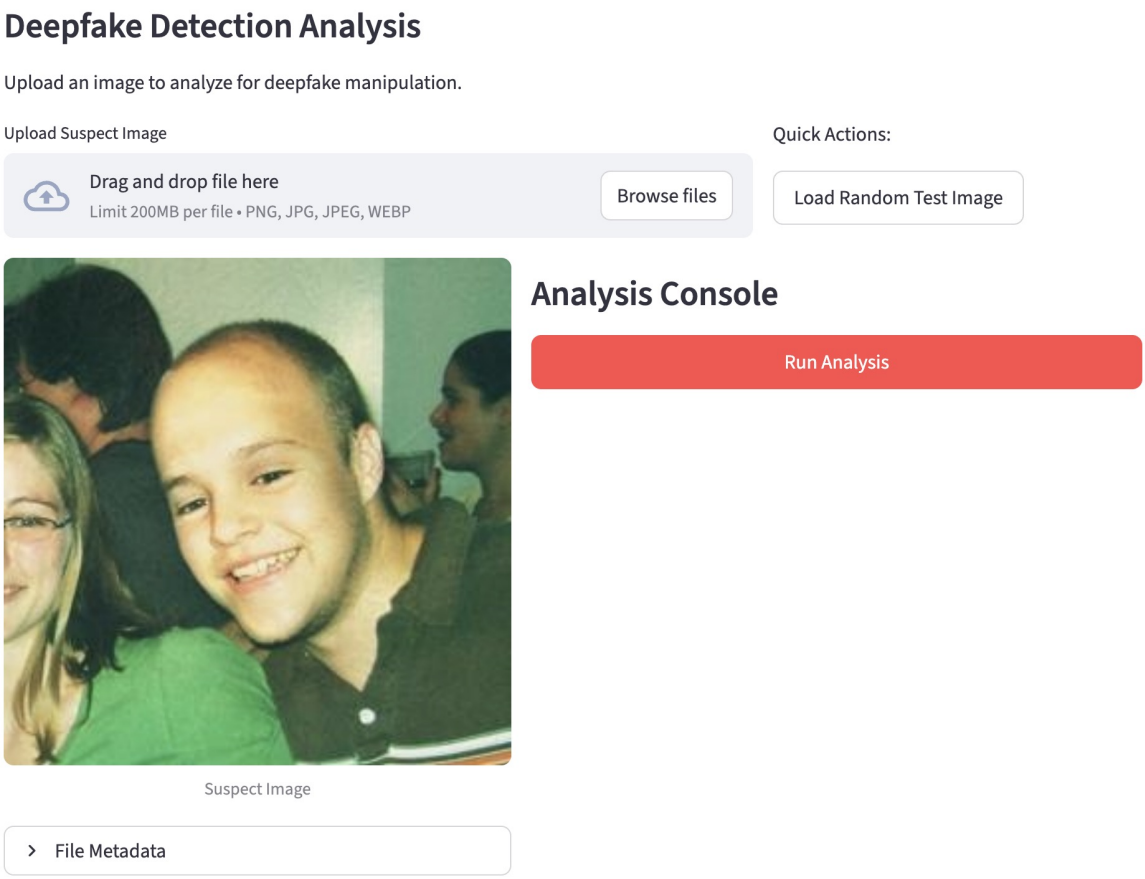

Figure A.4: Frontend interface for Deepfake Detection

f

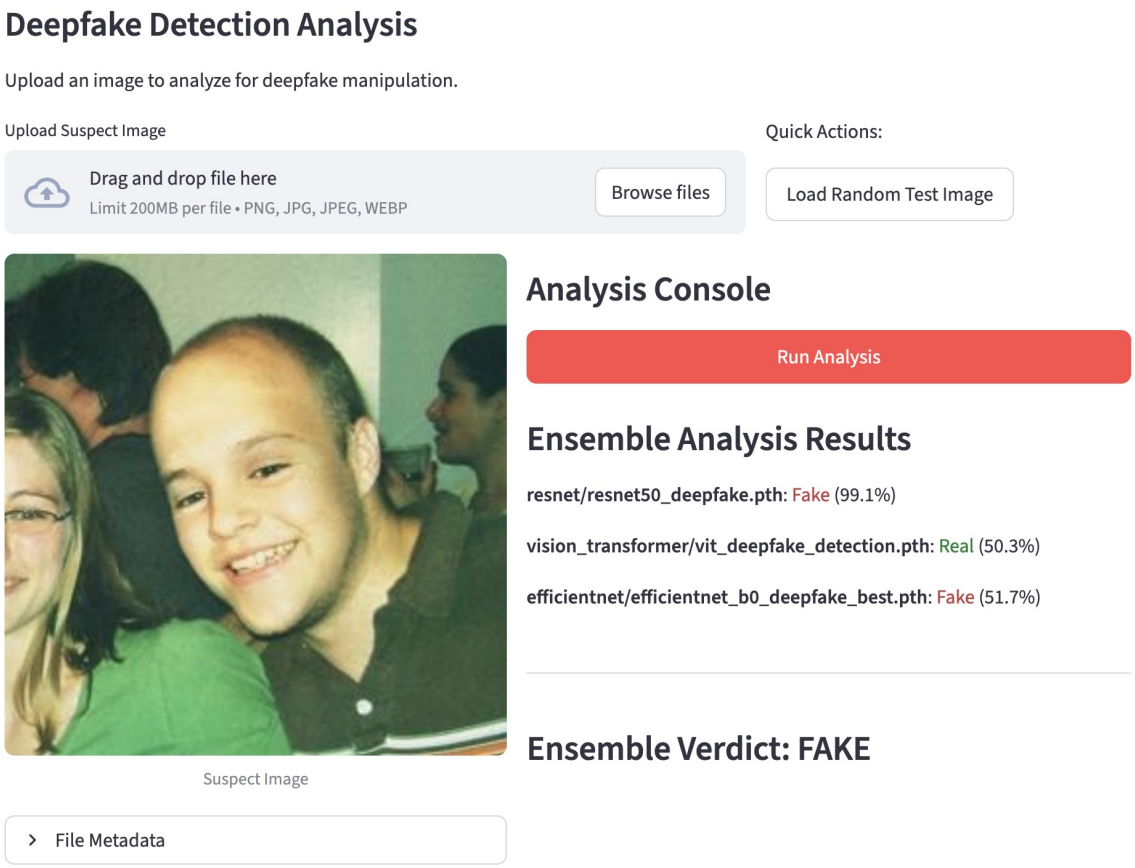

Figure A.5: Frontend interface for Deepfake Detection