Leveraging State-Space Models for Temporal Analysis in Deepfake Detection

by

Affshafee Rahman
22341024

Diniya Tahrin Bhuiyan
22341046

Shami Islam Khan
22301186

Salim Miah
23241051

MD Shohbat Ahsan
22341029

A thesis report submitted to the
Department of Computer Science and Engineering
Brac University

Department of Computer Science and Engineering
Brac University
October 2025.

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Students' Full Names & Signatures:

Affshafee Rahman
22341024

Diniya Tahrin Bhuiyan
22341046

Shami Islam Khan
22301186

Salim Miah
23241051

MD Shohbat Ahsan
22341029

Approval

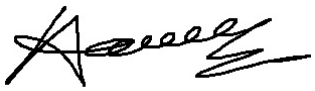
The thesis titled "Leveraging State-Space Models for Temporal Analysis in Deepfake Detection" submitted by

1. Affshafee Rahman (22341024)
2. Diniya Tahrin Bhuiyan (22341046)
3. Shami Islam Khan (22301186)
4. Salim Miah (23241051)
5. MD Shohbat Ahsan (22341029)

of Summer 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in Summer 2025.

Examining Committee:

Supervisor:
(Member)



Dr. Amitabha Chakrabarty
Professor
Department of Computer
Science and Engineering
BRAC University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Associate Professor and Chairperson
Department of Computer Science and Engineering
BRAC University

Abstract

Deepfake technologies, especially those based on lip-sync forgeries, present an advanced threat to integrity in digital media as they produce seamless audiovisual forgeries that are hard to detect. Transformer-based models show promise, but are resource-heavy and fail to generalize against forgeries created by modern, generative methods. This thesis addresses these issues by proposing an efficient novel framework for the detection of lip-sync forgeries that is based on State-Space Models (SSMs). We propose a dual-stream architecture using parallel Mamba blocks to independently model in the temporal domain the visual dynamics associated with lip movements and the audio dynamics based on audio spectrograms. Both streams use a lightweight MobileNetV3-Small backbone for spatial feature extraction and are configured with an optimal state dimension of 160, discovered through a two-stage ablation study. The resulting temporal feature vectors are fused and a classification is performed using a small MLP head. Trained on the high-quality AV Lips dataset, the Mamba based model proposed achieves a new state of the art accuracy of 94.60% and an AUC of 99.12%, while having an exceptionally low number of parameters, at 2.48 million. In addition, the model achieves robust generalization, emphasizing its potential as a powerful and deployable solution for audio-visual deepfake detection.

Keywords: Deepfake Detection, Lip-Sync Forgery, State-Space Models, Mamba, Audio-Visual Synchronization, Temporal Modeling, Multimodal Deep Learning, Lightweight Architecture, Cross-Dataset Generalization, Media Forensics

Acknowledgement

We would like to express our heartfelt gratitude to our supervisor, Dr. Amitabha Chakrabarty, for without his splendid assistance this research would not have been able to fulfill its potential. His advice proved to have such a relevance at every step, as well as his suggestion of the excellent and new architectures which could be examined, in particular that of making use of the State-Space Models (Mamba). This suggestion proved so vital to the direction of this work, both as regards the success and the matter of this research. We shall not forget also how approachable and industrious he proved to be whenever we contacted him, either in his office or through emails.

We are equally grateful to his research assistant, Azwad Aziz, for without his considerable assistance this research work would not have been completed. His splendidly down-to-earth style of conversation, resourcefulness and never failing desire to help were simply great. He always went far beyond what was due to give us the appropriate and necessary technical help and useful suggestions whenever we hit a technical hurdle.

This thesis on "Leveraging State-Space Models for Temporal Analysis in Deepfake Detection" would never have been able to fulfil its completeness without the cooperation and help which both these illustrious individuals were kind enough to give to us. We are indeed fortunate in being able to have had such painstaking and knowledgeable people supervise our research area.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter serves as the essential context for the whole thesis, introducing the phenomenon of the rapid evolution of deepfake technology as both a highly nuanced and constantly evolving danger to the information ecosystems of the world. It presents the main problem not merely as a question of technique, but as a very serious concern for the social standing of trust, personal security, and the integrity of information. The primary purposes of the chapter are to provide the background for the technological developments that render hyper-realistic synthetic forms of media, and to state formally the problem which poses itself as the motivating factor for this research elaborate the significant, interrelated problems of poor generalization of models, and a high computational cost, which inhibit the state of the art on the one hand. The justification of the research is stated with respect to the context of a "technological arms race" and an argument is proposed for the handling of the solutions for the problems that have been identified, for the purpose of the transition of forensic tools from being curiosities of the academic world which are of none but academic interest, to tools which can be deployed in an effective manner a vouchsafe trustworthiness. Lastly, this chapter gives a brief outline of the mode of research and the scope of validity of this thesis, which also supplies the reader with a delineation of the course of investigation for the next chapters.

1.1 Background

The term "deepfake" emerged into the public vocabulary in late 2017, thanks to a user of the social media platform Reddit who employed deep learning algorithms to swap the faces of celebrities over the faces of people found in pornographic videos. Although image and video manipulation by digital means were not a new discovery, often requiring a considerable amount of human labor and technical proficiency using dedicated packages such as Adobe Photoshop, this was nonetheless an important test case. The arrival of powerful frameworks for deep learning, in particular deep neural networks and their special variant, the Convolutional Neural Network (CNN) lent the computer architecture the technical basis whereby the face swapping feature for video was realized with unprecedented realism and on a scale difficult to imagine before this time.

The technology quickly changed from being a curiosity in obscure online venues to being utilized widely. Developers took notice of the public interest and created such

mobile applications as FaceApp and FaceLab that allowed the end-user the ability to create entertaining and innocuous deepfakes with a few clicks. However, the democritization of the ability to generate synthetic media has made abundantly clear its potential for unscrupulousness. The lowered barrier to entry to the production of believable fabrications has opened a new avenue for disinformation, fraud, and harm to reputations and proper behavior. A glaring example of the threat was apparent in 2022 during the Russian invasion of Ukraine, when a deep fake of Ukrainian President Volodymyr Zelenskyy surrendering to Russian forces was posted online with the idea being to confuse and demoralize the Ukrainian people.

Lip-syncing forgeries, one of the several types of methods of manipulation, represent a particularly malevolent domain of forgery because they preserve the identity of the imagified person, while altering their lip-movement so that it coincides with a faked or altered piece of audio material; whereas the full-face forgery because it alters the identity completely, is a simple matter to identify by the human observer and the usual machines that detect forgery.

The manipulation necessary for lip-sync forgeries occurs primarily in the movement of the mouth and jaw in order to align the visual representation of speech with the target audio signal. The artifacts produced are often subtle and transitory, and lack the blatant visual inconsistencies that characterize more rudimentary forgeries. Detection poses a specific challenge since the forgery does not employ a false identity but rather complicates the temporal consistency of two coincident modalities. The social peril involved in this technology is tremendous. Malicious actors may weaponize lip-sync forgeries to create fraudulent video calls, false confessions, political disinformation, non-consensual material, and thereby undermine the perceived reality of all digital video evidence.

1.2 Rationale of the Study or Motivation

The increasing sophistication of deepfake technology demands a fundamental transformation in methods of detection. This phenomenon can be understood in a broader context as a "technological arms race" in which the producers of synthetic media and the forensic detectors are engaged in a constant process of advancement and counter-advancement. As generative technologies advance, the fakes they generate become increasingly seamless, thus excising the tell-tale artifacts that were exploited by earlier detection methods.

This arms race features a very critical asymmetry that puts a different burden on the detection community. A bad actor needs to create a single forgery that succeeds in getting past a detector to achieve his objective. A detection system must succeed, however, consistently against an infinite and ever expanding universe of manipulation techniques, including techniques which it has never seen before. This asymmetry demands detection models which are not only accurate on known data, but also models which are very robust and generalizable to new or novel attack techniques. The deficiencies of the classical algorithms have made it necessary to turn to Deep Learning, not simply as an option, but as a necessity. Deep Learning models, especially because of the new efficient architectures, are only suited to this

task, because they can learn from data very complex, high dimensional, often non-intuitive, patterns.

1.3 Problem Statement

Despite the rapid advancement of deep learning-based detection, the field faces a number of important interrelated problems that inhibit the development of truly reliable and deployable systems. This thesis aims to tackle the problems of poor generalization across datasets, the great computational cost of state of the art architectures, and the need to validate models against current high quality forgeries. To tackle these issues head on, we present a novel, efficient, dual stream architecture based on State-Space-Models (Mamba), which is in theory much better suited for modelling the long time-series common to audio-visual data, whilst requiring a smaller computational load.

1.3.1 Architectural Limitations of Transformers in Temporal Modeling

For more than five years now, the transformer architecture has been central in the sequence modeling paradigm and has achieved SOTA results in many areas. The key innovation of the transformer is the self-attention mechanism, which allows long-range dependencies to be modeled by attending over the relevance of each token in the sequence in terms of their relevance to each other in the sequence. The self-attention mechanism is however also the chief bottleneck in the architecture, as the computational and memory requirements of self-attention are both quadratic with the length of the sequence i.e. $O(N^2)$, where N is the number of tokens. This quadratic scaling is the reason that processing long sequences, especially those found in high frame rate video or high resolution audio is prohibitively expensive and memory intensive.

These blocks (blockages) have necessitated the formation of new architectures which are more effective, in particular State-Space Models (SSM), a new architecture SSM called Mamba has appeared with promise as a potential alternative. The Mamba-style models are optimised for sequentially processing series of linear complexity, $O(N)$ with constant time inference. This puts them in a potentially superior stance in principle to the other architectures in terms of the modelling of continuous streams of long data mass such as found in audio, video and the like. And thus appealing in overcoming the computing restrictions of complex transformer architectures, the Mamba gives a toe-hold towards the build of powerful but deployable models for real-time audio-visual analysis.

1.3.2 The Generalization Crisis and the Need for Modern Benchmarks

A major problem for many modern deepfake detection models is their inability to generalize to previously unseen data; this is often described as the "generalization crisis." Models that achieve high accuracy on older, established benchmarks are

typically found to collapse catastrophically in performance on new, high-fidelity forgeries or unbounded data from the "real world." This brittle performance is often a reflection of the datasets with which these models have been trained, which may be created with outdated methods for producing forgeries or the datasets exhibit little diversity.

To fill this significant gap in research, the present thesis has used the **AV Lips dataset** for its training and evaluation corpus. The AV Lips dataset was created specifically for research into lip-sync forgery detection using state of the art lipsync generators and producing high-quality forgery examples that challenge existing detection methods. The temporal inconsistency of the audio and video streams is exploited in the dataset, creating a modern and highly relevant benchmark dataset for this specific problem domain. Furthermore, this is particularly important given the scarcity of high quality, publicly available datasets available containing both audio and video in perfect synchronisation, the importance of AV Lips is clear. Its availability is crucial in enabling the research as presented here. By training on this modern dataset, and testing for generalization on a less recent and more imbalanced dataset (FakeAVCeleb), this work uses a robust evaluation protocol designed for a more realistic evaluation of model robustness.

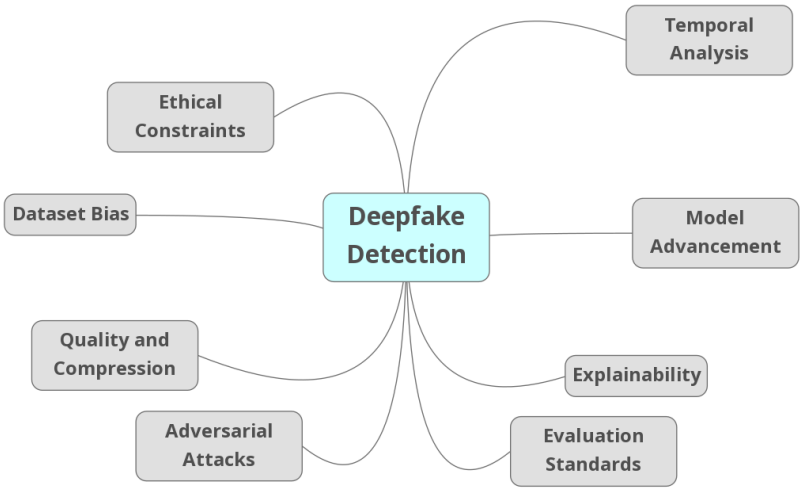


Figure 1.1: A simple mindmap of the big hurdles in the domain of deepfake detection.

1.4 Objective

The main aim of this work is to design, implement and validate a novel and efficient dual stream Mamba based architecture for the detection of lip-sync forgeries. This work will show that State-Space Models can in fact outperform well known architectures such as Transformers in this difficult time-based task and exhibit a significantly more desirable performance-efficiency profile.

To achieve this, the specific aims of this research are:

1. To develop a Mamba system for detection of audio-visual forgery, this represents one of the first applications of this architecture to the lip-sync detection problem.

2. To perform a wide ranging empirical analysis aimed at showing a superior performance-efficiency trade-off compared to existing temporal modelling architectures with the aim of proving the applicability of the proposed model for use in real world conditions where resources are limited.

3. To undertake an in-depth probing analysis of the generalisation properties of the model through a cross-dataset evaluation protocol aimed at proving its robustness to data from domains and forgery techniques which differ between the training and the test sources.

4. To benchmark the proposed architecture against a state-of-the-art lightweight detection scheme in order to position its performance and to back-up its validity as a solution deployable in the real world.

1.5 Methodology in Brief

The research methodology focuses on the design and empirical evaluation of a new "Dual-Stream Mamba Fusion Network." The system consists of two parallel streams, one visual stream and one audio stream, which process their respective data streams by first extracting spatial features utilising a MobileNetV3-Small backbone before being fed into Mamba Blocks which have a state dimension of 160 to model the temporal dynamics. This final configuration was established from a 2 stage ablation study that first established the best backbone in terms of efficiency before the Mamba dimensions were fine tuned to give maximum performance. The final temporal feature vectors from both streams are concatenated and fed to a simple Multi-Layer Perceptron (MLP) classification head to outputs whether a lip-sync forgery is present or not.

1.6 Scopes and Challenges

This research has specific boundaries and anticipates several challenges. The research is limited specifically to discovering AI-based facial manipulation in video data, focusing solely on the lip-syncing variant of deepfakes. It will not include audio-only forgeries or detection of manipulation in still images.
The successful execution of this research anticipates three primary challenges:

1. **Data Acquisition**: The obtaining of audio-visual data obtained in sufficiently great quantities, varieties and modern and representing the data of the latest generation of techniques presents an inherent and vital problem.

2. **Computational Cost**: The training of complex spatio temporal models for subsequent evaluation is enormously expensive from the computational side, and requires vast resources of GPU, etc.

3. **The Generalization Problem**: The main problem is that of constructing one model only, which will give robust generalization over the whole, fast varying field of deep fake techniques, is still in the end the main and indeed the most difficult problem of the whole field.

1.7 Key Learnings and Insights

The rise of deepfake technology is a complicated and fluid threat to information ecosystems all over the world. It has been established in this introduction, that while deep learning might offer the best opportunity for forensic detection, the current state-of-the-art is in a critical state because of a particular collection of interrelated problems, which are principally documented by the failure of models to be generalizable and prohibitive computational cost. This thesis is directly inspired by these problems, which are pressing in the research landscape. The next chapter will go into a detailed and critical examination of the existing literature, which will form the basis for a new and improved solution.

Chapter 2

Literature Review

This chapter presents a comprehensive and critical review of the existing literature on deepfake detection. It moves beyond a simple enumeration of prior work to provide a systematic and analytical narrative of the field's evolution, charting the progression of detection paradigms in direct response to the increasing sophistication of forgery techniques. The objectives are threefold: first, to establish a clear technical foundation by defining the core types of facial manipulation and the fundamental deep learning architectures employed for detection; second, to trace the architectural trajectory from early CNN-based methods to modern Transformer-based systems; and third, to synthesize these findings to precisely identify and detail the prevailing research gaps that the current state-of-the-art fails to address. This review serves as the essential scholarly foundation for the thesis. By deconstructing the strengths and, more importantly, the weaknesses of existing methods, it provides the necessary context and justification for the novel Mamba-based framework proposed in this thesis.

2.1 Preliminaries

Before diving into the research, we need a shared language. This section defines the core ideas and technologies that form the backbone of deepfake detection. Think of this as a toolkit overview; understanding these components is crucial for grasping the strategic choices discussed later.

2.1.1 Types of Facial Manipulation

Not all deepfakes are the same. Digital face forgeries come in several flavors, each with its own telltale signs and detection challenges:

Identity Swap (Face Swap): This is the classic deepfake most people imagine. It works by taking one person's face and plastering it onto another person's body in a video. The forger's goal is a seamless blend, where the target's expressions and movements remain, but their identity is completely replaced by the source's [30].

Attribute Manipulation: Here, the forger isn't replacing the whole face, just tweaking specific features. It's like digital plastic surgery changing someone's perceived age or gender, adding or removing a beard, or altering their skin tone [6].

The person is still recognizable, but details are off.

Expression Swap (Face Re-enactment/Puppeteering): In this technique, the target person's identity stays the same, but their expressions, head motions, and even speech movements are controlled by a source video. It's a form of digital puppetry, making someone appear to say or feel things they never did [44].

Entire Face Synthesis: This is the most extreme form, generating a completely new, photorealistic human face from nothing. This face doesn't belong to any real person. Technologies like StyleGAN [17] are famous for this, creating vast galleries of convincing but entirely fictional people.

2.1.2 Core Deep Learning Architectures

The battle against deepfakes is fought with different AI architectures, each acting as a specialized tool for a specific job.

Convolutional Neural Networks (CNNs): CNNs were the first major weapon in the deepfake detection arsenal. Their power comes from layers that act like a series of filters, learning to spot patterns from simple edges and textures up to complex facial parts. This makes them naturals for finding **local, textural glitches** unnatural blending around the edges of a forged face or strange pixel patterns that shouldn't be there [30]. You'll often see familiar CNN designs like ResNet, VGG, and Xception used as the starting point for detectors.

Transformers: Originally built for language, Transformers (and their vision counterpart, ViT) brought a new superpower to the fight: global attention. Their "self-attention" mechanism lets them weigh the importance of all parts of an input relative to each other. This allows a Transformer to understand the **overall context** of a face like checking if the eyes, nose, and mouth are in a natural spatial arrangement or how video frames connect over time [5]. This helps fix a key CNN blind spot: a narrow, localized focus.

Graph Neural Networks (GNNs): GNNs take a different approach, modeling data as a web of connected nodes. In deepfake detection, an image can be represented as a graph where nodes might be key facial landmarks (e.g., corners of the eyes, tip of the nose) or distinct facial regions. The edges then represent the spatial and structural relationships between these nodes [13]. GNNs are powerful for analyzing the **integrity of these structural relationships**, detecting forgeries where the underlying geometry of the face has been subtly distorted in a way that is inconsistent with a real human face [25].

2.2 Review of Existing Research

The story of deepfake detection isn't a static list of methods; it's a dynamic narrative of adaptation. The field has moved through clear phases, with each new wave of models rising to meet the challenge of more convincing fakes. This section follows

that story, showing the shift from hunting for simple, local mistakes to reasoning about complex, global inconsistencies.

2.2.1 Detecting Local and Frequency-Domain Artifacts with CNNs

The first successful detectors were built on CNNs. The thinking was straightforward: the process of creating or altering a face must leave behind tiny, low-level traces. While a human might miss these, a deep network could learn to spot them. Early models often used pre-trained CNN backbones like **XceptionNet** or **ResNet** as powerful feature-spotters, training them to find spatial glitches like unnatural blending, pixel inconsistencies, or mismatched compression levels between the forged region and the background [30].

Researchers quickly sought to enhance the performance of these models by guiding them toward more discriminative features. One effective strategy involved preprocessing images to explicitly highlight potential forgery traces. For example, the work by Rafique et al. demonstrated the use of **Error Level Analysis (ELA)**, a technique that identifies areas of an image with different compression levels. By feeding these ELA maps, rather than the raw RGB images, into a CNN, the model could more easily focus on the compression mismatches characteristic of a simple "copy-paste" forgery.

A significant advancement in this theme was the shift from the spatial domain to the frequency domain. The rationale was that generative processes often introduce unnatural high-frequency noise or disrupt the natural frequency spectrum of an image. The **Frequency-Aware Attentional Feature Fusion** model exemplifies this approach by using the **Discrete Cosine Transform (DCT)** to convert image patches into their frequency representations [20]. This allows the model to learn to spot tell-tale frequency-domain artifacts that are not apparent in the spatial domain. An even more nuanced approach focuses on the unique **noise signatures** inherent to digital images. The **TruFor** framework leverages a modified **Noiseprint** algorithm, which extracts a fingerprint-like noise pattern characteristic of a specific camera sensor or generative model. Similarly, the **NoiseDF** model uses a Siamese Network to compare the noise patterns extracted from the face with those from the background, flagging inconsistencies as evidence of manipulation [24].

This first generation of CNN-based detectors proved deep learning could work for forensics. Their success was rooted in targeting the "unnatural fingerprints" of the digital creation process itself. However, this approach has an inherent vulnerability: as generative models become more advanced, they learn to better replicate the statistical properties of real images, thereby erasing or masking the very artifacts these detectors are trained to find. This limitation necessitated a move toward detecting more fundamental inconsistencies.

2.2.2 The Integration of Transformers

As forgeries became more seamless and local artifacts less reliable, the focus of the research community shifted. The new challenge was no longer just to detect "bad pixels" but to identify inconsistencies in the **global context** of a face or the **temporal flow** of a video. A forgery might have perfect local texture but exhibit an unnatural relationship between the eyes and the chin, or it might feature inconsistent head motion or blinking patterns over time. This evolution in the problem demanded a new class of models capable of long-range, contextual reasoning a role perfectly suited for the Transformer architecture.

The first step in this direction was the development of **hybrid architectures**. In these models, a CNN backbone is first used to extract rich, local features, and its output is then fed into a Transformer encoder. The Transformer's self-attention mechanism can then analyze the relationships between these features across the entire image, capturing the global context that the CNN alone would miss. The work by Heo et al., which adds a Transformer encoder after an EfficientNet feature extractor [5], is a prime example of this effective hybrid strategy.

For video detection, the challenge expanded to the temporal dimension. Several models were designed specifically to capture temporal disturbances. The **MRE-Net** framework [18], for instance, employs a multi-rate sampling strategy, processing video frames at different temporal rates to learn both short-term (e.g., frame-to-frame jitter) and long-term (e.g., inconsistent patterns over several seconds) inconsistencies. Other methods, like **iCap-sNet-TSF**, integrate classical computer vision techniques, using the Lucas-Kanade optical flow algorithm to explicitly calculate motion between frames and detect unnatural flow patterns, before feeding this information into a deep learning architecture [16].

2.3 Review of all existing state-of-the-art detection methods

The landscape of deepfake detection is characterized by a diverse array of models and methodologies, constantly evolving to counter increasingly sophisticated forgery techniques. These approaches often build upon foundational deep learning architectures, adapting them to identify subtle inconsistencies that betray manipulated media.

2.3.1 Foundational and Hybrid Convolutional Neural Network (CNN) Approaches

Many early and foundational deepfake detection methods leverage Convolutional Neural Networks (CNNs) to extract features from images and videos. **MesoNet** [27] [23], for instance, was proposed by Afchar et al. to classify deepfake videos by analyzing mesoscopic noise features, arguing that low-layer noise degrades with compression while high-layer semantic features are difficult to use for detection. Similarly, **XceptionNet** [22] [15] [21] [23], a widely adopted CNN architecture, has

demonstrated impressive performance by directly taking the whole face as input to learn global feature representations. However, traditional CNNs are often limited in capturing global feature relationships and show poor generalization across diverse datasets.

Other CNN-based approaches include **VGG16** and **ResNet**, which have been used as backbones for extracting features [21] [3] [5] [16]. Bonettini et al. studied combinations of different pre-trained CNN models, starting with **EfficientNetB4** as a backbone, utilizing end-to-end and Siamese training methods [4]. Later, Deng et al. chose **EfficientNet-V2** as a baseline network for deepfake detection, optimizing its structure to balance speed and accuracy. In attempts to improve generalization, some methods incorporate attention mechanisms, such as Dang et al.'s approach that adds an attention mechanism to a baseline network to detect entire video frames. In attempts to improve generalization, some methods incorporate attention mechanisms, such as Dang et al.'s approach that adds an attention mechanism to a baseline network to detect entire video frames.

2.3.2 Capsule Networks and Their Enhancements

Capsule Networks (CapsNets), introduced as an alternative to CNNs, aim to overcome the limitations of CNNs in learning complex spatial relationships by using vector neurons to represent facial features and considering positional relationships between them [27] [15] [5]. Nguyen et al. first applied CapsNet to deepfake video detection to learn detailed information about facial pose and extract richer spatial features [16]. Building on this, the **iCapsNet–TSF** method combines an improved CapsNet with an optical flow algorithm to make comprehensive decisions based on temporal and spatial features of facial images [16]. The iCapsNet in this method optimizes the dynamic routing algorithm (iDRA) to address shortcomings in the original algorithm, such as distinguishing noise capsules by calculating cosine similarity between low-layer capsules [16]. Ilyas et al. proposed **E-Cap Net** [11], an efficient-capsule network that introduces a low-cost max-feature-map (MFM) activation function in each primary capsule to suppress low activation neurons, making the model light and robust.

2.3.3 Vision Transformers and Hybrid Architectures

The success of **Vision Transformers (ViTs)** in natural language processing has led to their exploration in computer vision, including deepfake detection, for their ability to learn global features and long-range dependencies [37].

One notable parameter-efficient tuning approach is **DeepFake-Adapter** [37], which adapts large pre-trained ViTs for deepfake detection. This method addresses the issue of existing methods overfitting to low-level forgery patterns by leveraging generalizable high-level semantics from ViTs [37]. It introduces lightweight dual-level adapter modules: Globally-aware Bottleneck Adapters (GBA) inserted in parallel to MLP layers for global low-level features, and Locally-aware Spatial Adapters (LSA) that interact with ViT features via cross-attention to capture local low-level forgeries [37]. This dual-level adaptation helps exploit better generalizable forgery

representations by interacting high-level semantics with global and local low-level forgeries.

Deep Convolutional Pooling Transformer is another hybrid model that jointly leverages convolutional pooling and re-attention with a stack of CNNs to extract local features, followed by an enhanced deep Transformer to enrich global feature learning and analyze relations between image feature patches [23]. This method uniquely emphasizes the importance of extracting keyframes from compressed videos, as they carry complete information without loss during reconstruction, boosting detection performance.

The **Self-Supervised Graph Transformer** [14] framework introduces a self–supervised pre-training model that uses a contrastive learning framework to extract high-level visual representations of facial landmarks, aiming for robustness against post-processing perturbations like compression or blur. It integrates **Graph Convolutional Networks (GCNs)** and Transformer architectures to capture complex dependencies among different image regions, modeling both local relationships (via GCNs) and global interdependencies (via Transformers).

The **Interpretable Spatial-Temporal Video Transformer (ISTVT)** [28] proposes decomposing the self-attention mechanism into spatial and temporal components, coupled with an innovative self-subtract mechanism to effectively detect temporal artifacts. ISTVT also provides interpretability by visualizing class-discriminative heatmaps for spatial and temporal self-attentions separately.

Thumbnail Layout (TALL) [38] is a strategy that transforms video clips into pre-defined image layouts, preserving spatiotemporal dependencies and converting temporal modeling into spatial modeling. An enhanced version, TALL++, introduces a Graph Reasoning Block (GRB) to enhance interactions between semantic regions and a Semantic Consistency (SC) loss to improve generalization by ensuring semantic coherence between consecutive frames.

FreqFaceNet [41] is a Transformer-based architecture that utilizes group spatial and channel-level self-attention to efficiently capture the global context of the image while reducing computational costs. It focuses on facial regions like eyes, nose, eyebrows, cheeks, and lips when classifying images as real or fake.

For resource-constrained scenarios, the S**hallow Vision Transformer** is proposed [21], using an attention mechanism with a multi-head attention module to highlight important sections of deepfake images. This model is designed to be lightweight, with significantly fewer parameters and FLOPs compared to baseline ViTs, yet it maintains high recognition accuracy even with less training data.

2.3.4 Frequency Domain and Noise-Based Methods

Some deepfake detection methods delve into the frequency domain to uncover forgery clues. **FDS_2D** is a multi-branch network that rethinks magnitude and phase features, arguing that both spectra contain different image information and that relying

on only one can be disturbed by noise [26]. FDS_2D separates spectral information into three categories: magnitude spectrum, phase spectrum, and the relationship between the two, designing independent modules for feature extraction from each. It also incorporates a **Multiple Cross Attention (MCA)** mechanism for information interaction between branches.

F3-Net [4] is another method that considers frequency domain information, dividing its architecture into two branches: one to extract frequency domain information and another to process spatial domain information after high and low frequency separation and inverse Fourier transform.

Regarding **noise characteristics**, the **NoiseDF model** proposes a noise-based deepfake detection approach from a digital forensics perspective [24]. It extracts face-background pairs using a Siamese network, assuming background areas remain authentic while faces are manipulated, and analyzes noise patterns. A **Multi-Head Relative-Interaction method** with depth-wise separable convolutions is devised to justify the level of interaction and similarity between face and background noise features [24]. Previous attempts with PRNU noise for deepfake detection were generally inconclusive [24].

The **Artifacts-Disentangled Adversarial Learning (ADAL)** [15] framework aims to disentangle artifacts from deepfake videos to improve detection and provide visual evidence. It employs an adversarial learning strategy to extract artifacts, helping to avoid overfitting, and includes a **Multi-scale Feature Separator (MFS)** to precisely separate artifacts from irrelevant information at different levels.

2.3.5 Identity, Semantic, and Physiological Feature-Based Approaches

A key insight in advanced deepfake detection is that each person has unique characteristics that synthetic generators struggle to reproduce consistently [7]. The **Implicit Identity Driven (IID)** framework focuses on identity inconsistencies, designing **explicit identity contrast (EIC)** loss to pull real samples closer to their explicit identities and push fake samples away, and **implicit identity exploration (IIE)** loss to guide fake faces with known target identities to have small intra-class and large inter-class distances [7]. This framework aims to explore "fake-invariant features" that are robust to manipulation and post-processing.

The **ID-unaware Deepfake Detection Model** argues that "Implicit Identity Leakage"—the mistaken learning of identity representation by binary classifiers—is a stumbling block to generalization [8]. To mitigate this, it uses an **Artifact Detection Module (ADM)**, an anchor-based detector that focuses on local artifact areas with multi-scale anchors, reducing attention to global identity information [8]. This model also uses a Multi-scale Facial Swap method to generate fake images with ground truth artifact area positions for training.

POI-Forensics (Person-of-Interest Deepfake Detector) proposes a multi-modal (audio-visual) analysis using a contrastive learning approach. It trains networks so

that learned representations characterize temporal segments of the same identity closely but keep different identities far apart [7]. At test time, it computes similarity indices between features from the analyzed video and a set of pristine reference videos of the POI [7]. This method trains exclusively on real videos, aiming for high generalization by detecting anomalous behavior in manipulated content [7].

2.3.6 Multi-Modal (Audio-Visual) Integration

Recognizing that both audio and visual modalities can be forged, multi-modal approaches combine information from both streams for enhanced detection. **AVoiD-DF (Audio-Visual Joint Learning for Detecting Deepfake)** proposes a framework [27] that jointly learns audio-visual inconsistency at temporal-spatial levels. It includes a **Temporal-Spatial Encoder (TSE)** for feature embedding and a **Multi-Modal Joint-Decoder (MMD)** for jointly learning multi-modal interaction and fusion [27]. This framework addresses the limitations of uni-modal detection by explicitly considering situations where audio, visual, or both are falsified.

Earlier multi-modal efforts, such as Emotions Don't Lie and MDS, extract and analyze similarity or dissimilarity between audio and visual modalities [27]. However, these often use audio as an extra supervisory signal without acknowledging that audio can also be forged.

2.3.7 Temporal and Spatiotemporal Consistency Detection

Deepfake videos often exhibit subtle temporal inconsistencies that are not easily caught by frame-based detectors. **MRE-Net (Multi-Rate Excitation Network)** aims to effectively excite dynamic spatial-temporal inconsistency from multiple rates [18]. It comprises a **Bipartite Group Sampling (BGS)** strategy to divide video into multiple bipartite groups with different rates, covering various face motion dynamic evolution. It also introduces a **Momentary Inconsistency Excitation (MIE)** module to encode spatial artifacts and intra-group short-term temporal inconsistency, and a Longstanding Inconsistency Excitation (LIE) module to perceive inter-group long-term temporal dynamics.

Other methods focused on temporal aspects include those using optical flow to extract inter-frame correlations [16] [18], two-stream RNNs to mine time information [16], and the "Snippet" unit (**Intra-Snippet Inconsistency Module (Intra-SIM)** and **Inter-Snippet Interaction Module (Inter-SIM)**) to study local motion information and establish inconsistent dynamic modeling frameworks.

2.3.8 Generalization and Interpretability Strategies

Addressing the critical challenge of **generalization capability** to unseen data, several models propose specific strategies. The **Artifacts-Disentangled Adversarial Learning (ADAL)** framework (mentioned previously) is designed to avoid overfitting by disentangling artifacts, improving transferability across manipulation types and datasets [15].

The **ID-unaware Deepfake Detection Model** (also mentioned previously) directly addresses the Implicit Identity Leakage phenomenon to improve generalization by focusing on local artifacts rather than global identity [8].

For **interpretability**, the **Self-Supervised Graph Transformer** framework introduces a graph Transformer relevancy map to pinpoint manipulated regions and explain the model's decision-making by highlighting the importance of individual regions [14]. Similarly, the **Interpretable Spatial-Temporal Video Transformer (ISTVT)** (also mentioned previously) can visualize class-discriminative heatmaps for its decomposed spatial and temporal self-attentions [28], providing insights into which spatial or temporal information is extracted and used by the model. **TruFor** [10] aims for trustworthiness by leveraging external uncertainty quantification to design a confidence map from anomaly localization heatmaps.

While most models only provide a binary classification, a few have made significant strides in incorporating interpretability.

TruFor: This framework moves beyond a simple binary output by providing an integrity score supplemented by an **anomaly and confidence map**. This gives the user more evidence to make an informed decision about the media's authenticity [10].

Self-supervised Graph Transformer: This model, proposed by Khormali and Yuan, incorporates a graph Transformer relevancy map. This map is generated from the model's output activation map and is designed to transparently highlight the subtle details the model believes have been manipulated, while ignoring irrelevant information [14].

Frequency-Aware Attentional Feature Fusion: The model from Tian et al. uses a Gradient-weighted Class Activation Map (Grad-CAM) during its evaluation. This provides a visual aid, in the form of a heatmap, that highlights the specific regions within the source image that the model considered most important in making its classification [20].

Interpretable Spatial-Temporal Video Transformer (ISTVT): The model from Zhao et al. introduces explainability based on Layer-wise Relevance Propagation. This allows the model to visualize the d**iscriminative and salient areas**, effectively showing the user which parts of the video contributed most to the deepfake detection [28].

DFGNN: This model enhances interpretability through its graph-based structure. The construction of the graph, where image patches are nodes, allows for a more transparent analysis of the relationships the model uses to detect forgeries [13].

These efforts represent a critical shift from not just detecting deepfakes, but also explaining how they are detected, a necessary step for building trustworthy forensic tools.

2.3.9 Patch-Based and Feature Restoration Methods

Some approaches focus on analyzing images at a finer granularity or restoring degraded features. A **patch-based approach** is proposed to detect deepfakes in occluded images, using a multi-path decision that combines reasoning on the entire face, concatenation of feature vectors from face patches, and a majority vote from individual patch classifications [19]. This method also incorporates occlusion removal and assigns weights to facial patches to improve accuracy.

DF-UDetector [12] transfers the problem of deepfake detection to the feature space, proposing a novel restoring model to detect deepfakes by adopting a feature extractor combined with a feature transforming module to capture and transform the feature map, arguing that direct image restoration is insufficient for various degradations.

2.4 Research Gaps

Looking across all the research, a clear picture emerges: despite moving fast, the field of deepfake detection is stuck on a set of tough, interconnected problems. These research gaps are the biggest hurdles standing in the way of building robust, reliable tools that can be used in the real world. The next sections will walk through these gaps, starting with the well-known issues of generalization and explainability before moving to trickier challenges about data practices, future threats, and practical deployment.

2.5 The Emergence of State-Space Models for Deepfake Detection

Just as the field is struggling with the heavy computational costs of Transformers, a new kind of architecture called the State-Space Model (SSM) has stepped into the spotlight [32]. The Mamba architecture, a recent and powerful implementation of an SSM, has shown it can match Transformer performance on many sequence tasks while only needing linear-time complexity [32]. This makes it a perfect candidate for efficient deepfake detection that doesn't require a supercomputer.

2.5.1 Mamba for Audio Deepfake Detection

The application of Mamba to deepfake detection has been particularly prominent in the audio domain. Models like Fake-Mamba [46] and RawBMamba [29] have been proposed as direct, efficient alternatives to Transformer and Conformer-based architectures for speech deepfake detection. These models typically replace the computationally expensive multi-head self-attention mechanism with a bidirectional Mamba block. The bidirectional approach is critical, as it allows the model to process information from both past and future contexts within an audio sequence, which is essential for capturing the global dependencies needed to spot subtle forgery artifacts [29]. Fake-Mamba, for instance, integrates a pre-trained XLSR front-end with a bidirectional Mamba encoder to capture both local and global artifacts, achieving

state-of-the-art results on several benchmarks while maintaining real-time inference speeds [46]. This line of research demonstrates that Mamba's input-dependent selection mechanism is highly effective at identifying the subtle, anomalous cues in synthetic speech with significantly reduced computational overhead compared to self-attention.

2.5.2 Mamba for Visual Deepfake Detection

While still an emerging area, Mamba is also being adapted for visual deepfake detection. The WMamba framework is a notable example, combining wavelet analysis with the Mamba architecture for face forgery detection [45]. The core idea is that wavelet transforms are effective at exposing subtle, high-frequency artifacts and unnatural contours that are often characteristic of manipulated images. However, these artifacts can be slender, fine-grained, and globally distributed across the face. WMamba leverages the Mamba architecture's ability to efficiently model long-range spatial relationships with linear complexity. This allows the model to analyze fine-grained features from small image patches and detect inconsistencies across the entire face, a task for which Transformers are powerful but computationally expensive. This approach showcases how Mamba's architectural strengths can be tailored to the unique characteristics of visual forgery artifacts.

2.6 Research Gaps

Our deep dive into the literature makes it clear that despite quick progress, deepfake detection is held back by a collection of stubborn and intertwined challenges. These research gaps are the main barriers stopping us from creating strong, dependable, and ready-to-use forensic tools.

2.6.1 The Generalization Crisis and Dataset Limitations

The most significant failing of current state-of-the-art (SOTA) deepfake detectors is their inability to generalize. Many models that report near-perfect accuracy are, in reality, "brittle." They are trained and validated on the same datasets (intra-dataset validation), often using well-established but aging benchmarks like FaceForensics++ (FF++). When these high-performing models are tested on data from a different source (cross-dataset validation) or on uncurated media from the internet so-called "in the wild" data their performance often collapses dramatically. This failure is a direct consequence of the datasets themselves. A critical limitation in the field is the lack of publicly available global datasets that provide a fair and diverse representation of both forgery techniques and real-world media. Existing datasets frequently lack diversity in demographics (ethnicity, gender, age), pose, lighting conditions, and background context [30]. This can lead to significant model bias, where a detector performs well for one demographic but poorly for another, a critical issue with severe consequences if deployed in real-world applications like law enforcement. Furthermore, many popular benchmarks were created with what are now outdated forgery techniques, meaning models trained on them are unprepared for the artifacts produced by newer generative architectures like Diffusion Models [40]. The

over-reliance on "old" datasets like FF++ for benchmarking, while useful for comparability, hinders true progress as these datasets no longer represent a sufficient challenge for modern detectors.

2.6.2 The Explainability Deficit and the "Black Box" Dilemma

A fundamental barrier to the adoption of deepfake detectors in real-world, high-stakes applications is their lack of transparency. The vast majority of deep learning models operate as "black boxes". They take an image or video as input and produce a binary output real or fake with a confidence score, but provide no justification for their decision. For a journalist trying to verify a source, a law enforcement officer investigating a crime, or a court assessing evidence, a simple "fake" label is insufficient. The inability to answer the question, "Why is this considered a fake?" or "Which part of this media was manipulated?" erodes trust and makes these powerful tools practically unusable where accountability is paramount. Addressing this "black box" problem is a crucial area of research. While most models only provide a binary classification, a few, such as ISTVT with its visual heatmaps [28], have made significant strides in incorporating interpretability. However, such models remain the exception, and the development of trustworthy and accountable forensic tools is contingent on advancing this frontier of Explainable AI (XAI).

2.6.3 The Data-Centric Divide: Diversity, Modernity, and Ethics

Beyond the general issue of dataset limitations lies a more profound, tripartite crisis in the data that forms the bedrock of the entire field: a crisis of diversity, modernity, and ethics. This is not merely a need for "more data," but a fundamental flaw in the nature of the data being used.

A significant number of widely used datasets are heavily skewed towards specific demographics. For example, the popular Celeb-DF (v2) [2] dataset is comprised of over 88% of subjects which are Caucasian. This lack of diversity in age, gender, and ethnicity is a critical research gap because it directly leads to the development of models that are inherently biased. A model trained on such imbalanced data may perform unfairly across different populations, a severe issue with profound ethical consequences if deployed in sensitive applications like legal verification or law enforcement. The creation of specialized datasets like the Gender Balanced Deepfake Dataset is a direct acknowledgment of this gap, but the problem remains pervasive across the field.

Compounding the diversity issue is the field's over-reliance on outdated benchmarks. There is a widespread academic practice of benchmarking new models against "old" datasets like FaceForensics++ to ensure comparability. This practice, however, creates a false sense of progress. As argued by the creators of the Deepfake-Eval-2024 benchmark, these legacy datasets were generated with techniques that have since been superseded and no longer reflect the quality or characteristics of modern, "in the wild" deepfakes [40]. This leads to inflated performance metrics that do not translate to real-world efficacy and ultimately hinders the development of truly ro-

bust detectors capable of handling contemporary threats.

Finally, a critical and often-overlooked research gap lies in the ethical framework for data curation. Datasets like WildDeepfake and Celeb-DF are sourced from the internet, often with little to no evidence of subject consent or consideration for privacy. This practice raises serious ethical questions and legal risks. The discovery of private medical images within the large-scale LAION-5B dataset serves as a stark warning of the severe potential for harm when data is scraped without rigorous oversight. The absence of a standardized, ethical framework for dataset creation and stewardship is a major gap that undermines the trustworthiness of the entire research field. The difficulty and ethical complexity of sourcing diverse, modern, and consented data is what drives researchers to rely on easily accessible but flawed and outdated benchmarks. This reliance, in turn, perpetuates the development of biased, non-generalizable models. The core research gap is therefore not just a need for better data, but for sustainable and ethical methodologies for data curation that can keep pace with the rapid evolution of forgeries.

2.6.4 The Oncoming Wave of Generative Models: The Diffusion Model Challenge

The vast majority of current deepfake detectors have been developed and trained in an ecosystem dominated by Generative Adversarial Networks (GANs). However, a new class of generative models, Diffusion Models (DMs), has emerged, capable of producing hyper-realistic content with fundamentally different underlying processes [36]. This technological shift creates a critical, forward-looking research gap, as the detection community appears largely unprepared for this next wave of forgery techniques.

The core of the problem lies in the different artifact signatures produced by these models. GANs are known to introduce specific, often high-frequency, artifacts such as grid-like patterns, which many detectors are implicitly or explicitly trained to identify. Diffusion Models, which operate by iteratively denoising an image from a random state, do not necessarily produce these same artifacts. On the contrary, research indicates that DMs produce fewer and more subtle detectable artifacts, making them inherently harder to detect with existing methods [35]. Consequently, state-of-the-art detectors trained on GAN-based forgeries exhibit a significant performance drop, often failing entirely, when tested on images generated by DMs [35]. This suggests that much of the field may be focused on "artifact detection" rather than true "forgery detection." That is, models are learning the signature of the tool (a specific GAN architecture) rather than the fundamental evidence of manipulation (e.g., the blending of source and target identities). When a new class of models with a different signature appears, these specialized detectors become obsolete. A true research gap therefore lies in developing methods that can generalize across generative paradigms, not just across datasets from the same paradigm [36]. This requires moving away from low-level artifact detection towards higher-level semantic, physical, or behavioral inconsistency detection that is agnostic to the specific generation method.

2.6.5 The Scalability and Deployment Barrier

A significant and growing chasm exists between the architectures that achieve state-of-the-art performance in academic literature and those that are practical for real-world deployment. This scalability barrier is a critical engineering research gap that prevents the translation of powerful research into practical tools.

This situation suggests a potential misalignment of research incentives. The academic focus on achieving marginal SOTA improvements on established benchmarks can lead to a "gamified" environment where building ever-larger models to gain a fractional accuracy improvement is prioritized over the engineering challenge of operationalizing these models. The gap is not just in creating lightweight models, but in developing new architectural paradigms and training techniques, such as knowledge distillation and structured pruning, that can break the unfavorable correlation between performance and efficiency. A significant research direction is the development of "efficiency-aware" design principles and benchmarks that reward not just raw accuracy, but performance-per-FLOP or accuracy-per-parameter, shifting the research focus towards creating tools that are both effective and deployable.

2.6.6 The Multi-Modal and Multi-Domain Blind Spot

The majority of deepfake detection research is narrowly focused on detecting facial manipulation in the visual domain of videos. This creates significant blind spots for increasingly common and sophisticated forgeries that span multiple modalities or domains.

Modern deepfakes, particularly talking-head videos, often involve manipulation of both the audio and visual streams to create a seamless forgery, such as perfectly synchronized lip movements. Uni-modal detectors that only analyze video frames are blind to audio-visual inconsistencies. A major research gap exists in developing robust multi-modal detection systems that can jointly analyze audio-visual streams to detect subtle desynchronization or inconsistencies that single-modality detectors would miss [43]. Furthermore, the intense focus on facial manipulation leaves other forgery types under-researched. This includes full-body deepfakes, where a person's entire body and movements are synthesized, and the manipulation of non-human objects or scenes [42]. This research gap is exacerbated by a lack of large-scale, diverse, multi-modal deepfake datasets. While datasets like FakeAVCeleb and the new ILLUSION dataset are emerging, they are not as established or widely used as video-only benchmarks, slowing progress in this critical area.

The focus on facial video reflects a reactive posture in the research community, targeting the most prominent threat after it has emerged. The true research gap is the need for a more proactive and holistic approach to media forensics. This involves developing generalized frameworks that can ingest and analyze multiple data streams (video, audio, text, metadata) and apply forensic principles that are not tied to a specific object (like a face) but to the fundamental physics and statistics of media capture and transmission. This represents a shift from content-based detection to physics-based or sensor-based detection.

2.6.7 The Evaluation Standardization Deficit

The methods used to evaluate and compare deepfake detectors aren't standardized or rigorous enough, making it hard to know what the true state of the art is or if a new model is actually useful in the real world. This deficit is creating a "reproducibility crisis" and blocking real progress.

As highlighted in recent surveys, there is an absence of a structured and uniform approach to evaluation. Researchers use different datasets, pre-processing steps, and training protocols, which leads to results that are not directly comparable and can often be over-inflated. Many publications report high performance based only on intra-dataset evaluation. However, as shown empirically in this thesis and in other studies using "in the wild" data, performance collapses dramatically in cross-dataset scenarios [40]. The failure to consistently report these more challenging generalization metrics is a major gap in evaluation practice. Moreover, the majority of studies rely almost exclusively on Accuracy and AUC. While useful, these metrics do not tell the whole story. Other metrics like Precision, Recall, F1-Score, and Equal Error Rate (EER) provide different and valuable insights into a model's performance, especially with imbalanced datasets. The failure to report a wider suite of metrics can obscure a model's weaknesses.

This evaluation deficit creates an environment where the goal can become topping a specific benchmark's leaderboard rather than creating a genuinely robust tool. This can misdirect research efforts towards brittle, overfit solutions. The deeper research gap is the need for a cultural shift in the community towards a more holistic, transparent, and challenging evaluation paradigm. This could involve shared, blind test sets and standardized reporting templates, similar to the NIST evaluations for biometrics, which would standardize evaluation and drive research towards true generalization.

2.7 Summary of Key Findings

Looking across the entire body of deepfake detection research, we see a field bursting with new ideas but held back by some deep-rooted problems. A few big-picture takeaways emerge that capture where we are today and hint at where we need to go next.

There is a clear evolutionary path in the design of detection architectures, progressing from foundational CNNs to complex Transformer-based systems. This trajectory was a necessary response to the escalating sophistication of forgery techniques. Despite these architectural advances, the gap between model performance on controlled benchmarks and performance in real-world scenarios remains the single greatest challenge confronting the field. This generalization gap underscores that achieving robustness to unseen manipulation techniques is a largely unsolved problem.

For deepfake detection to transition from a purely academic exercise to a trusted, deployable technology, interpretability is not a feature but a requirement. This review finds that the overwhelming majority of SOTA models remain opaque "black boxes." Finally, progress across the entire field is fundamentally bottlenecked by the

quality and scope of available training and testing data. This ‘dataset dilemma’ not only fuels the generalization crisis but also slows the development of next-generation detectors

Chapter 3

Requirements, Impacts and Constraints

This chapter serves as the critical bridge between theoretical analysis covering the literature and practical engineering to produce a deployable solution. It aims to formally articulate the real world context that the proposed deepfake detection system must operate within, going from abstract to formal specification and considerations. The goals of this chapter are to systematically outline the eventual technical requirements, the enormous societal and environmental implications, the key ethical issues that have to be considered, as well as the standards and project management plan that governs the development process. This parameterization is essential to ensure that the research is not only technically sound but also responsible, relevant and practically based.

3.1 Final Specifications and Requirements

The main aim of this project is to design a deepfake detection model that is both highly accurate and computationally efficient, with a view to being able to work successfully in real-world environments. Primarily, the work aims to create a framework that can achieve high detection accuracy of current lip-sync forgery while retaining a lightweight structure suitable for operation in realistic domains. The system must be designed to conform to various important specifications in order to fulfill this aim.

The architecture has to be fundamentally lightweight if it is to be implemented on resource-constrained devices or one of the real-time systems. This is accomplished by using the Mamba architecture, which has linear time complexity rather than the quadratically scaling Transformer architecture. The model will need to have high performance (target >99% AUC) on the principal AV Lips dataset and must show robust generalization on the cross-domain FakeAVCeleb dataset. One of the main specifications is that the model size should be very small (target <3 million parameters). The implementation will make use of standard software tools including Python and libraries for deep learning such as PyTorch or TensorFlow, with GPU support necessary during the training phase. The project will require access to established benchmark deepfake datasets (AV Lips, FakeAVCeleb) for the appropriate evaluation of detection performance and generalization.

3.2 Societal Impact

Deepfakes represent a serious and growing threat to modern society and present a pressing need for effective and deployable detection methodologies. This project directly addresses some major civil evils prompted by the malicious implementation of this technology. Firstly, it addresses the propagation of disinformation and "fake news" where the use of lip-sync forgeries can produce false political speeches or news clips to influence the masses and erode confidence in credible media sources. A readily apparent and dangerous application of such forgeries is in the political arena. Manipulation of this type can be used as a weapon against a candidates character, or it can be used to alter the outcome of electoral process. This model is intended to serve as a confirmation/verification tool for the media apparatus as well as for election monitors. Further, the project seeks to provide solutions to the terrible creation of non-consensual material. This particular evil means to work extreme psychological harm. On an even greater social scale the advent of hyper-realistic forgeries serves to add to cultural bankruptcy of confidence. That is, as citizens begin to doubt the legitimacy of all visual media, they breed cynicism which robs the value of evidence of its efficacy. Finally, the work addresses safety and legal abuses associated with hostile takeovers of fake videos as they relate to fan incitement to violence and the effective use of fake videos to mislead police officers. Ultimately, this work aids in constructing a better and more responsible digital future in which the proper institutions are equipped with effective tools to combat this visual evidence disinformation.

3.3 Environmental Impact

This project proceeds from the viewpoint of developing a lightweight and efficient model architecture, thereby minimizing the environmental costs resulting from the training and deployment of deep learning models. It makes the project focused on Mamba (linear time complexity) as opposed to the Transformer (quadratic complexity) architecture in line with the tenets of green computing. It would not be good practice to develop bulky and log-deep system deepfake detection systems requiring extensive computational power, thereby engendering higher energy usage. This project calls for a model that is (9.54 MB) markedly less than the farther-reaching and less desirable options available (32 MB etc.). This would mean less energy in training and, of more importance, lower energy requirements in real-time inference. A more affordable AI solution is proposed.

3.4 Ethical Issues

The ethics of the development and use of the system will need to be worked through diligently across a number of important dimensions. One essential question with the ethics will be bias in the data and the model itself. It will be important that the model flows through to performance across a range of demographic groupings fairly and without bias because it is known that datasets used for training introduce bias and will not perform from a fair and equal detection process. Although this project uses the best datasets available, it knows that the field presents a wider

challenge of little diversity across datasets. Another crucial aspect of the ethics surrounding this project will be that of potential misusage. There is always the risk that results from any detection process will be reverse engineered by any malicious actors and that a far more effective deepfake will be created that will not be able to be detected by computer vision models which save their best performances for these so called true deepfakes. It will be the duty of the project practitioners then to have ensured that the system has been tested for fairness to a high level, that where limitations are exhibited they forthwith are disclosed and from a lengthy period of time that systems will be distributed with measures surrounding against misusage or misinterpretation.

3.5 Standards

Although formal standards across the deepfake detection domain have not yet reached a commonly adopted formal standard, the project will strive to maintain acceptable standards of AI best practices and computer vision development. An important aspect to consider is that of explainability and transparency to users of model output, so that models show performance beyond simple, binary predictions. Data privacy principles will be upheld, so that where known data-processing frameworks apply, such as GDPR provisions, these will need to be adhered to, especially where data about users has been or is likely to be collected. Furthermore, to maintain scientific integrity and uplift collaborative openness, the project will abide by stringent guidelines in respect of reproducibility, so that model architecture, training procedure and evaluation metrics are documented to the fullest extent. The evaluation of the model performance will be performed against established and widely accepted evaluation metrics from the field of computer vision, such as precision, recall, F1-score, Area Under Curve (AUC) performance, etc., so that the performance of the model can be accurately assessed, and compared to other protest methods at the frontline of performance within this field.

3.6 Project Management Plan

Project Schedule (14 weeks total):

- Weeks 1–4: Literature review, problem refinement, and baseline model selection.
- Weeks 5–8: Model development, implementation of dual-stream architecture, and integration of Mamba blocks.
- Weeks 9–11: Evaluation and testing using benchmark datasets.
- Week 12: Optimization and real-time testing.
- Weeks 13–14: Report writing, documentation, and presentation.

Budget:

- Use of free/open-source frameworks (TensorFlow, OpenCV, PyTorch).

- Cloud computing (GPU access via Google Colab, RunPod, Jarvislabs.ai, or university HPC): estimated $20 to $40 for training.
- No hardware or licensing costs are anticipated.

Resource Management:

- Team coordination through GitHub, Google Drive, and Discord.
- Tasks and deadlines managed using tools like Google Sheets.
- Weekly progress evaluations and collaborative code review.

3.7 Risk Management

A structured risk management plan is essential for navigating the technical and ethical complexities of this AI project.

3.7.1 Model & Performance Risk

The primary risk is that the proposed Mamba-based model may underperform or fail to generalize effectively.

This risk is mitigated by using a rigorous, multi-dataset evaluation methodology. By benchmarking performance on both a quality training dataset (AV Lips) and an unseen, difficult cross domain dataset (FakeAVCeleb), we can robustly assess the generalization performance of our Mamba-based model implementations. Continuous benchmarking against established baselines (CNN, LSTM, Transformer) and a SOTA benchmark provides a clear measure of efficacy.

3.7.2 Data Risk

This project uses public datasets which may contain latent bias and/or characteristics impairing model performance or fairness. These shortcomings are mitigated by using AV Lips, which is a modern dataset that was produced specifically to alleviate shortcomings of the older benchmark datasets. We understand the limitations and imbalance of the FakeAVCeleb dataset used here, and we will use the appropriate metrics (such as F1-score, when appropriate) for a more nuanced analysis. A robust data governance approach is committed to with appropriate explanations of data sources and characteristics.

3.7.3 Operational & Scalability Risk

A common failing in academic models is their prohibitive cost in terms of computation, which makes them worthless for deployment.
This risk is one of the principal aims of the project. The exceedingly simple possibility of the Mamba architecture, with its linear time complexity compared to a Transformer is the principal counter to this. We will quantify this by measuring the size and inference speed, making sure that the final model is such that it meets the criteria of being light in weight and efficient.

3.7.4 Ethical Risk

The model could demonstrate biases, or be misused, if released without protection. We will mitigate this by making the results as transparent as possible. We will document the limitations of the model, and any possible biases, as well as their source if possible. Any public release of the model or code would be accompanied by a policy of responsible use, encouraging correct use and discouraging mis-use.

3.8 Economic Analysis

Although this research project does not directly have a commercial aspect, an economic argument is still important for justifying the chosen research direction, especially in terms of a possible return on investment in terms of a technological and social point of view. As mentioned in the further details of the project cost ratio, mainly the non-monetary costs will be large, compared to the costs for the commercial application of the new model. The costs caused by the time spent by the students on the development and analysis of the project are the large ones. The only monetary costs will be the relatively small accounts for the computational expenses, mainly the costs for GPU time used for training the deep learning models. One of the large economic advantages for this project is the active use of the Mamba architecture, which is seen as a very computationally inexpensive architecture compared to the Transformer architecture that is used in the baseline situation. This will help considerably from a training cost point of view, but also more importantly for running costs if the model should ever be used commercially on a larger scale.

The societal benefits are dual: on the one hand, it alters mainstream thinking about deepfakes, and on the other, achieves a practical solution to the task of detecting them. It is clear that deepfake detection will interest both businesses and government agencies, and this paper has shown how this aim can be achieved within an efficient architecture. From an economic perspective, this project is worthwhile because it leads to a cheap, quick and effective product which obviously has a lower threshold for incorporation into existing systems for social media moderation and also for journalistic verification systems. Thus, the economic significance of this research lies in the fact that in its ideology of stressing efficiency it becomes more probable that the technology can be installed practically, and thus that it can fulfill its societal significance.

Chapter 4

Proposed Methodology

This chapter presents the core technical contribution of the thesis, detailing the design and rationale for the proposed "Dual-Stream Mamba Fusion Network." It marks the transition from identifying the problems in the existing literature to the systematic engineering of a targeted solution. The chapter will provide a comprehensive architectural overview, followed by a detailed breakdown of the visual and audio processing streams. It will explain the function of the MobileNetV3-Small backbones for feature extraction, the role of the Mamba blocks in modeling temporal dynamics, and the final modality fusion and classification mechanism. This detailed specification provides the blueprint for the model validated in the subsequent experimental chapters.

4.1 Design Process or Methodology Overview

The challenge of creating an efficient deepfake detection solution is located in a dynamic, adversarial context that is often deemed a "technological arms race". The increasing sophistication of counterfeit content creates the imperative for a design philosophy that considers not only classification accuracy, but also the inter-related issues of generalisation, efficiency and scalability. The methodology of this thesis is predicated in response to this demand.

The meta-analysis undertaken in the literature review established the quadratic complexity of transformer architectures as the critical bottleneck in the analysis of long resampling audio-visual sequences, presenting a scalability barrier for practical application. This research directed the choice of State-Space Models (Mamba) as core architectural alternative due to their quadratic complexity and successfully established utility in the capture of long-range dependencies. The design process thus described is not an incrementally improved version of existing models, but a strategic change of tack to a more efficient paradigm. The aim is to resolve the contradiction between performance and efficiency by an engineering feat in the formation of a dual-stream framework that is both efficacious and efficient. This change in perspective is the essential step that is needed to produce a robust forensic solution that is genuinely usable in the environments in which it is most needed.

The proposed framework comprises three sequential stages:

1. **Problem and Baseline Analysis:** This initial stage involves a thorough architectural analysis of dominant Transformer-based models to systematically identify their computational bottlenecks. For this reason a data driven approach will be of assistance in justifying the choice of Mamba architecture to a more efficient architecture to be adopted for the aims of temporal modelling.

2. **Development Cycle:** Based on the analysis, this stage involves the design and implementation of the novel "Dual-Stream Mamba Fusion Network." The model is then trained and evaluated on the high-quality AV Lips dataset to establish its baseline performance.

3. **Final Evaluation:** This final stage applies a rigorous, multi-dataset evaluation protocol. The model's generalization is tested on the unseen FakeAVCeleb dataset, and its performance is benchmarked against traditional baselines (CNN, LSTM, Transformer) and a SOTA Anomaly Detection [9] model to provide a comprehensive assessment of its capabilities.

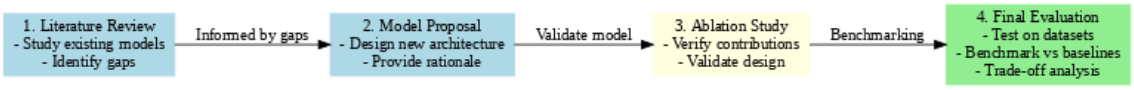


Figure 4.1: A study pipeline for this thesis.

4.2 Architectural Overview

To address the identified research gaps, this thesis proposes a novel "Dual-Stream Mamba Fusion Network," a deep learning framework designed specifically for the efficient and accurate detection of lip-sync forgeries. The architectural blueprint is illustrated in Figure 4.2.

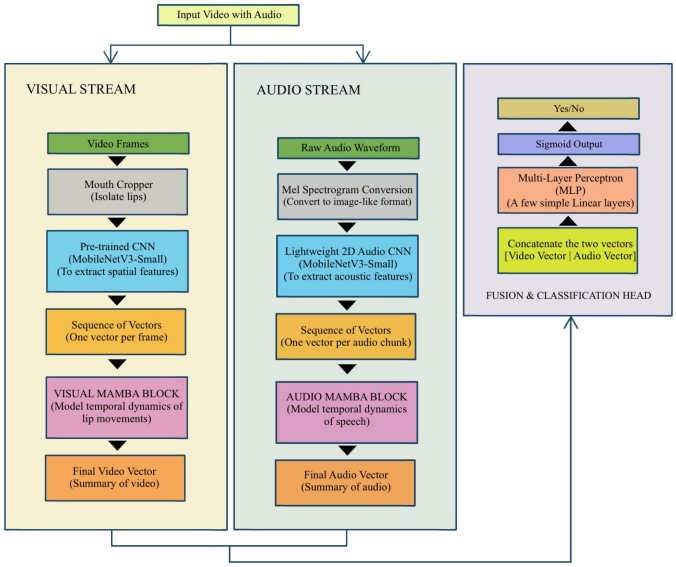


Figure 4.2: Proposed architecture of our Dual-Stream Mamba Fusion Network

The central architectural approach is grounded in parallel, modality-dependent processing, followed by late fusion. The network architecture consists of two parallel

streams, one for visual data and one for audio data, that process their input in parallel. Each stream must extract a compact temporal feature representation describing the dynamic evolution of its modality over time. The visual stream models the dynamics of lip movements, while the audio stream models the dynamics of speech. By processing these independently before fusion, the model is forced to learn robust modality-dependent temporal features. The final feature vector output of each stream is then concatenated and passed to a simple classification head, which makes the final determination of whether temporal consistency exists between the audio and visual streams.

4.3 The Visual Stream: From Mouth Cropping to Temporal Encoding

The visual stream processes the video component of the input. Its main task is to analyze the sequence comprising of the images so as to capture the evolving dynamic nature of the speakers lips. This is done through a multi-stage processing pipeline, which isolates a region of the face, extracts powerful spatial features from each image of the moving sequence and subsequently models the temporal evolutionary function of the dynamics of these features to produce a single, concatenated feature vector encoding the visual dynamics of speech.

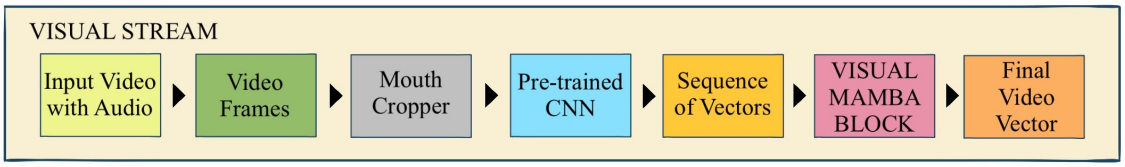


Figure 4.3: Simplified flow diagram of the Visual Stream.

4.3.1 Spatial Feature Extraction with MobileNetV3-Small

The visual stream begins with a targeted preprocessing pipeline designed to focus the model's attention on the region of interest. For each frame of the input video, a face detection algorithm is first applied. Once the face is localized, a tight bounding box is cropped around the mouth region. This step is crucial as it isolates the part of the face directly involved in speech production and most likely to contain manipulation artifacts, while discarding irrelevant background information.

The sequence of cropped mouth images is then fed into a pre-trained CNN backbone for spatial feature extraction. The chosen backbone is **MobileNetV3-Small**, a selection validated by the detailed ablation study mentioned in Chapter 5. This choice is important to the framework's goal of maximizing efficiency without compromising on performance. MobileNetV3 is an architecture designed explicitly for high efficiency on resource-constrained devices [1]. Its efficiency stems from key architectural innovations including depthwise separable convolutions, inverted residual blocks with linear bottlenecks, and the integration of a Squeeze-and-Excitation module to adaptively recalibrate channel-wise feature responses.

4.3.2 Modeling Lip Dynamics with the Visual Mamba Block

The MobileNetV3 backbone processes each mouth crop independently, producing a fixed-size feature vector for each frame. This results in a sequence of spatial feature vectors, where each vector represents the state of the lips at a specific point in time. This sequence is then passed to the **Visual Mamba Block**.

The function of this block is to model the long-range temporal dependencies within the sequence of lip movements. Drawing inspiration from the bidirectional scanning mechanism used in Vision Mamba (Vim) [39], this block processes the entire sequence of frame-level features to capture the complete trajectory of lip motion. It learns to model the movement of the lips, the opening and closing of the lips over time, thus producing a representative vector that gives a summary of the dynamic aspects of the visual aspects of speech. Hence, it is able to detect unnatural and inconsistent movements which would not be apparently indicated by a separate analysis of each frame by itself.

4.4 The Audio Stream: Acoustic Feature Processing

Running simultaneously with the visual stream, the audio stream's goal is to analyze the corresponding audio waveform. Here, the objective is to change the raw one-dimensional audio input to rich two-dimensional data, and then to synthesize the appropriate time structures. This imitates the course of the visual pathway, which presents a simplified feature vector describing the acoustic dynamics of the speech for later comparison with the visual one above.

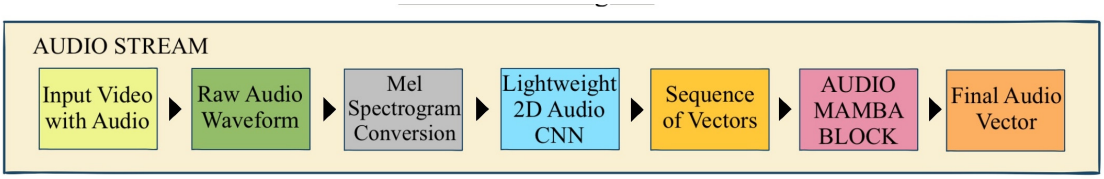


Figure 4.4: Simplified pipeline of the Audio Stream.

4.4.1 Mel Spectrogram Representation

The audio stream processes the raw audio waveform that comes with the video. The first operation is to convert this 1D time-series signal into a 2D representation which is suitable for processing with a CNN. This is accomplished by creating the **Mel spectrogram** of the audio. A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. The Mel scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another. To use a Mel spectrogram formats the audio into an image-like format in which the x-axis represents time, the y-axis represents frequency (on the Mel scale), with the intensity of each pixel corresponding to the amplitude of a particular frequency at any particular time [33]. This is a standard and highly effective representation for

deep learning-based audio tasks.

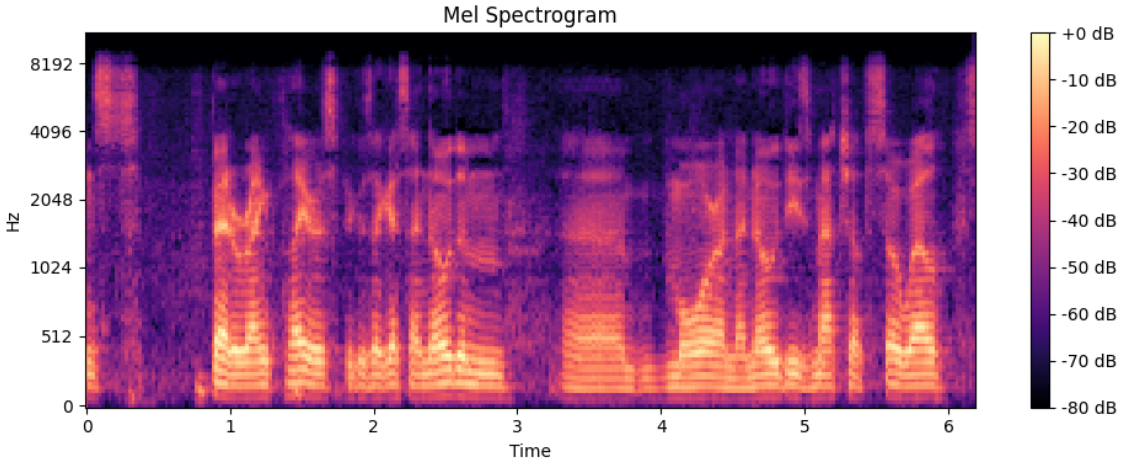


Figure 4.5: Mel spectrogram representation of a sample audio clip from the AV Lips dataset. The horizontal axis represents time (in seconds), the vertical axis represents Mel-frequency bands (ranging from 0 to 8 kHz), and the color intensity indicates the magnitude of frequency components (in dB), with warmer colors representing higher energy. The spectrogram captures the temporal-spectral characteristics of speech, revealing formant structures and temporal variations that are essential for audio-visual synchronization detection.

The colored map in Figure 4.5 illustrates a corresponding Mel spectrogram generated from an audio sample in the AV Lips dataset. Here, we can see the rich temporal-spectral structure captured by the Mel-frequency representation, where the horizontal axis represents progression of time, and the vertical axis corresponds to the 80 Mel-frequency bins within the relevant frequency range of human hearing. The intensity of color in the spectrogram encodes the magnitude of spectral energy at each time-frequency band, with brighter regions indicating higher acoustic energy.

The visible patterns in the spectrogram correspond to the phonetic form of speech, including formants (resonant frequencies of the vocal tract), fricatives (high-frequency noise bands), and voiced segments (periodic structures with harmonic characteristics). The time-frequency domains thus provide a compact and rich information input to the audio stream of the proposed architecture. By transforming the raw audio waveforms into this two-dimensional representation, the Mel spectrogram allows the MobileNetV3-Small backbone to process the audio feature extraction in a manner analogous to that of image processing, where convolutional operations may examine the data to identify local spectral features and temporal transitions. These extracted spatial features are then handled by the Mamba block to process the sequence of continuous features in order to model the long range temporal features across the entire audio clip and allow for the subtle asynchronies between audio and visual modalities which characterize lip-sync forgeries.

4.4.2 Acoustic Feature Extraction with MobileNetV3-Small

The sequence of Mel spectrogram chunks is processed by a lightweight 2D CNN to extract acoustic features. For this task, a pre-trained MobileNetV3-Small was selected as the backbone, a choice strongly supported by the ablation study in Chapter 5. The use of a consistent, highly efficient architecture across both streams ensures a balanced and lightweight final model. The powerful performance of MobileNetV3-Small on image-like data, combined with its minimal computational footprint, made it the optimal choice for extracting features from the spectrograms in our final model configuration.

4.4.3 Modeling Speech Dynamics with the Audio Mamba Block

Similar to the visual stream, the MobileNetV3-Small backbone outputs a sequence of feature vectors, each representing a short time segment of the audio. This sequence is then fed into the Audio Mamba Block. This block's function is analogous to its visual counterpart: it models the temporal evolution of the acoustic features over the entire clip. By processing the sequence of acoustic vectors, it learns the characteristic patterns and flow of natural speech. This process is inspired by the architecture of Audio Mamba (AuM), which has proven effective at modeling long-range dependencies in audio spectrograms [31]. The final output is a single vector that summarizes the temporal dynamics of the audio signal.

4.5 Modality Fusion and Classification Head

The last step in the pipeline involves merging the information coming from the two parallel streams and making a classification. The fusion mechanism is simple yet effective, whereby the last summary vector of the Visual Mamba Block and the last summary vector of the Audio Mamba Block are concatenated together to form one larger summary vector. This vector now contains a rich presentation of the visual lip movement and the acoustic representation of the human speech, with a temporal awareness.

This fused vector is then fed into a simple classification head, which all of information compression to a **Multi-Layer Perceptron (MLP)** with some linear layers. The MLP now learns to map the pattern in the fused feature space to that of a final classification. The final layer of the MLP uses a sigmoid activation function in which the output is restrict to an output between 0 and 1. This output is interpreted as the probability that the video is a forgery, with an output closer to the value of 1 being a high probability of being a forgery, whilst, on the other hand, an output closer to the value of 0 being a high probability of being a real video. Therefore, it is intentionally low complexity of the fusion and classification components that allows the model performance to be dictated by the faultless quality of the temporal features being learnt in the Mamba blocks.

4.6 Datasets

The evaluation protocol embraces a multi-dataset approach in order to assess both model performance under controlled, benchmark conditions and model robustness in difficult, real-world situations. In particular this is intended to experimentally test the 'generalisation crisis', a perilous failing of current detectors whereby models that perform well on familiar data frequently underperform on unfamiliar data.

4.6.1 Training and Baseline Evaluation Dataset: AV Lips

The primary dataset for training and initial evaluation is the AV Lips dataset. This is a modern, high-quality dataset created specifically to address the challenge of detecting lip-sync forgeries [34] that are becoming increasingly challenging to detect. This dataset was developed by using state-of-the art lip generators such as Wav2Lip to produce forgeries that were visually seamless and contained none of the artifacts that can be seen in older deepfake datasets [34].

Figure 4.6: Sample frames of a real and forged clip from the AV Lips dataset [34].

The core focus of AV Lips is on the subtle temporal inconsistencies that arise between the manipulated visual lip movements and the corresponding audio track [34]. This makes it very suitable as a benchmark for testing the capability of the proposed model to measure fine temporal dynamics present in the data, extending from simple artifact recognition to a more subtle and sophisticated class of audio/visual coherence measurement.

4.6.2 Generalization Testing Dataset: FakeAVCeleb

To rigorously assess the model's ability to generalize to unseen data from a different domain, the FakeAVCeleb dataset was used for cross-dataset evaluation.

FakeAVCeleb is a large-scale, multi-modal deepfake dataset created from videos in the VoxCeleb2 corpus [6]. Unlike AV Lips, which focuses on a specific type of forgery, FakeAVCeleb contains manipulations created with a variety of techniques, including face-swapping (Faceswap, FSGAN), audio-driven facial reenactment (Wav2Lip), and synthetic voice cloning (RTVC) [6].

This dataset is characterized by its purposeful diversity across demographics (ethnicity, gender, age), and also significant class imbalance, consisting of 500 real videos and 19,500 fake ones. The evaluations provided by testing the models trained on

(a) Frame from a real video in FakeAVCeleb.

(b) The corresponding fake video's frame from FakeAVCeleb.

Figure 4.7: Sample frames of a real and fake video from FakeAVCeleb.

AV Lips against this different and more biased data distribution offers a severe test of the robustness and resilience of the models to domain shift.

Chapter 5

Result Analysis

In this chapter, the proposed Mamba-based framework is put through a thorough empirical validation and statistical analysis process. The objective of this chapter is to assess the model performances against established benchmarks and a state-of-the-art benchmark across a variety of datasets. The chapter describes the experimental design employed, along with a justification for the datasets chosen, and the metrics used for evaluation. The results of this analysis are compared and the analytical outlook is achieved in respect of intra-dataset accuracy, cross-dataset generalisation and the critical efficiency/performance trade-off. Finally, the results are discussed in detail with emphasis on their implications for the field, and acknowledgements made of the study's limitations.

5.1 Performance Evaluation

In this section, we outline the framework for empirical validation of the deepfake-detection models created for this study. This encompasses the methodology of validation, with a conscious choice of datasets and performance metrics. The emphasis is on getting insight into model performance, the ability to generalize, and the practical usefulness of the models in providing a transparent and reproducible framework for comparing deep learning models in regard to deepfake detection purposes.

5.1.1 Dataset Selection and Rationale

In the evaluation protocol, we implement a multi-dataset strategy designed to evaluate both the efficacy of the model in controlled, benchmark conditions and the robustness of the model in difficult, challenging real-world environments. This was designed for the purpose of rigorously testing the "generalization crisis", which is a major problem with current detectors, since models that do well on known data often drop severely when confronted with strange new data.

Dataset	Real Videos	Fake Videos	Used for Training	Used for Testing	Total Videos
AV Lips	3,397	4,207	4,000 (2,000 per class)	Remaining subset	7,604
FakeAV Celeb	500	19,500	N/A (Cross-dataset test only)	3,500 (500 real, 3,000 fake)	20,000

Table 5.1: Dataset statistics showing the distribution of real and fake videos used for training and testing.

Due to the computer resource constraints of this work, with limited time availability and GPU resources, a **balanced subset of each class of 2,000 videos** (totaling 4,000 videos) was randomly sampled from the AV Lips dataset for use in training. This ensures that the binary classification remains balanced while at the same time being computably feasible with the resources available. The remaining videos employed were the videos in the AV Lips dataset, used for validation and intra-dataset testing.

For cross-dataset evaluation on FakeAV Celeb, a subset of **500 real videos and 3,000 fake videos** was selected to evaluate generalization performance. This sampling strategy was adopted to manage the evaluation time while maintaining representation of both classes, though the class imbalance reflects the original distribution of the FakeAV Celeb dataset, which contains significantly more fake videos than real ones.

5.1.2 Evaluation Metrics

Model performance was quantified using two standard binary classification metrics: Accuracy (ACC) and Area Under the Receiver Operating Characteristic Curve (AUC). Accuracy measures the proportion of correct predictions, while AUC provides a more robust measure of a model's ability to distinguish between classes across all classification thresholds. To analyze the performance-efficiency trade-off, we also measure the model size in millions of parameters, computational cost (FLOPs), latency, and throughput.

5.1.3 Standardized Experimental Environment

In order to obtain a fair, unbiased comparative analysis, all experiments were performed in a standardized environment, with common training hyperparameters (e.g., optimizer, learning rate, batch size, number of epochs between updates) across all models. This methodological control was required in isolation of the architectural design of each model as the main independent variable exerting performance influence. Thus keeping the training protocol constant, the differences in accuracy, generalization or efficiency that are recorded can be more readily attributed to the strengths and weaknesses of the various models' architectures.

5.2 Comparative Analysis of Baseline Architectures

This section presents the core experimental results, comparing our optimized Mamba-based model (V1d) against three distinct baselines. The models are evaluated on detection performance, generalization, and computational efficiency to provide a comprehensive assessment of their capabilities.

5.2.1 Benchmark Model Selection and Rationale

A cohort of deep learning architectures was selected for this final comparative study to situate the performance of our proposed model within the broader landscape of

temporal and lightweight deepfake detectors.

Mamba Baseline (V1d): This is the final, optimized model proposed in this thesis, resulting from the two-stage ablation study.

LSTM Baseline: To create a fair comparison with traditional recurrent architectures, this baseline was constructed by taking the V1d model and replacing the Mamba blocks in both the audio and visual streams with standard LSTM layers. All other components, including the MobileNetV3-Small backbones, remained identical.

Transformer Baseline: Similarly, this baseline was created by replacing the Mamba blocks in the V1d architecture with Transformer encoder blocks. Each stream utilized 4 Transformer blocks, creating a powerful but computationally heavy alternative for temporal modeling.

SSVFAD Benchmark: To benchmark against the state-of-the-art in lightweight detection, we included the Self-Supervised Video Forensics for Anomaly Detection (SSVFAD) model. For brevity, this is referred to as SSVFAD. This model was chosen as it is one of the lightest and most recently developed multi-modal deepfake detection methods, providing a stringent and relevant benchmark for our efficiency-focused approach.

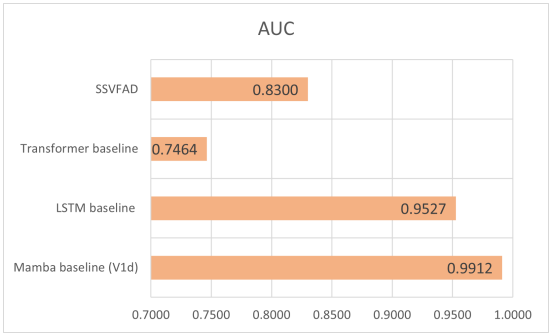
Model	**Parameters (M)**
Mamba baseline (V1d)	2.476
LSTM baseline	3.452
Transformer baseline	8.6001
SSVFAD	0.0192

Table 5.2: Parameter comparison of the chosen set of architectures for this study.

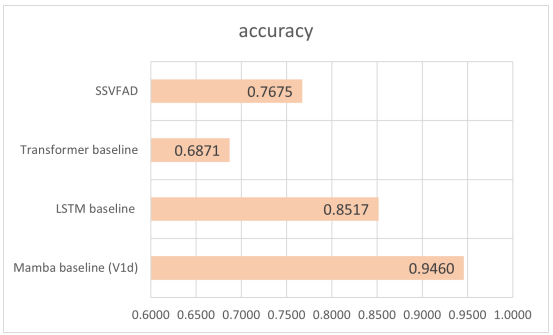
5.2.2 Intra-Dataset Performance on AV Lips

From results in Figure 5.1 and Table 5.3, the proposed **Mamba baseline (V1d) shows a widely superior result** when trained on and tested with the AV Lips dataset. Results of an **accuracy of 94.60%** and an **AUC of 99.12%** significantly surpassed the results of all other models. Results for the second best accuracy of the models were that the LSTM baseline was able to achieve an accuracy of 85.17%.

Notably, the **Transformer baseline, despite being over 3.4 times larger than the Mamba model, performed poorly**, achieving only 68.71% accuracy. This result challenges the assumption that larger models are inherently better and suggests that the Transformer's self-attention mechanism may be less suited for this specific temporal task compared to Mamba's selective state-space design. Our model also vastly outperformed the ultra-lightweight SSVFAD benchmark. These findings empirically validate that the Mamba architecture provides the most effective framework for capturing the nuances of lip-sync forgeries within this dataset.



(a) AUC scores



(b) Accuracy scores

Figure 5.1: The performance metric of the four models when trained and tested on the AV Lips dataset. The Mamba baseline (V1d) achieves the highest scores in both AUC and accuracy.

Model	Accuracy	AUC	Precision	Recall	F1
Mamba baseline (V1d)	**0.9460**	**0.9912**	**0.9815**	**0.9628**	**0.9720**
LSTM baseline	0.8517	0.9527	0.8943	0.8836	0.8889
Transformer baseline	0.6871	0.7464	0.6900	0.6900	0.6900
SSVFAD	0.7675	0.8300	0.7238	0.6947	0.7091

Table 5.3: Performance comparison of baseline models across key metrics when trained and tested on AV Lips. Bold numbers indicate the best performance for each metric.

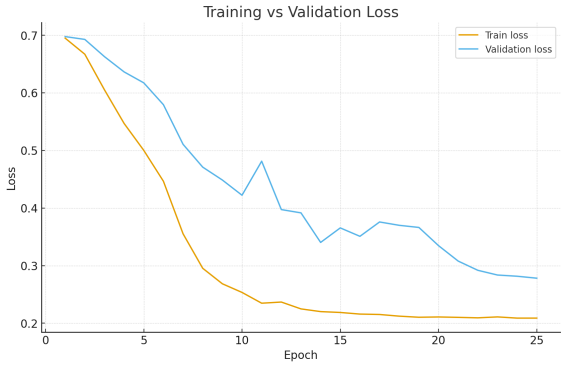
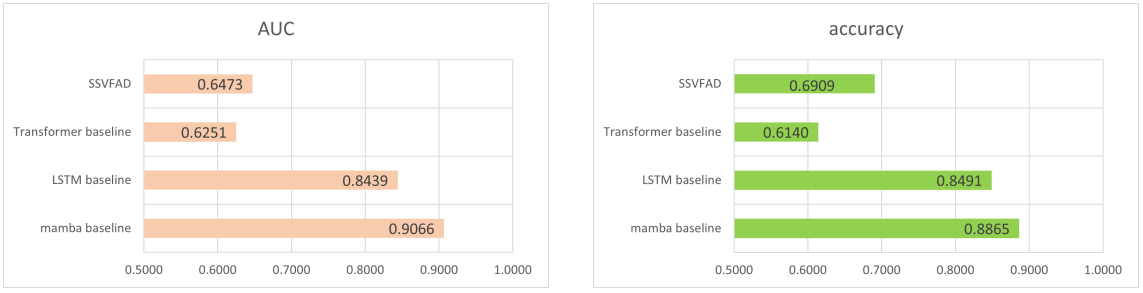


Figure 5.2: Training and validation loss curves of the Mamba baseline (V1d) model during training on the AV Lips dataset. The model demonstrates stable convergence with minimal overfitting, as evidenced by the close alignment between training and validation loss curves over 25 epochs.

The training dynamics of the Mamba baseline (V1d) model are shown in Figure 5.2. The loss curves converge steadily and systematically through training, with the validation plateaus closely following the training loss. This close relationship indicates that the model has well learnt generalizable features from the AV Lips dataset, without overfitting. This is an important quality in providing good performance in cross-dataset. Convergence occurred at around epoch 20, after which the training and validation losses plateaued. This indicates that the model has reached the full convergence potential for the given data.

5.2.3 Cross-Dataset Generalization Performance on FakeAV Celeb

The cross-dataset evaluation, where models trained on AV Lips were tested on the unseen FakeAVCeleb dataset, provides a critical measure of real-world robustness. The results, shown in Figure 5.3 and Table 5.4, reinforce the superiority of the Mamba-based approach.



(a) AUC scores

(b) Accuracy scores

Figure 5.3: The performance metric of the four models when trained on the AV Lips dataset, but evaluated on the FakeAV Celeb dataset.

The **Mamba baseline (V1d) outperformed again, achieving a maximum accuracy of 88.65%** and an **AUC of 90.66%**. All models suffered from the predictable reduction in performance from the domain shift, however the Mamba model's loss in performance was the least aggressive here, demonstrating a remarkable ability to generalise to different distributions of data and methods of forgery. The other models lost more performance, with the accuracy of the Transformer baseline falling to 61.40%. This remarkable generalisation performance seems to suggest that the Mamba architecture is learning more foundational and transferable features of audio-visual desynchronisation, and therefore will serve as a more useful tool in the area of "in-the-wild" deepfake detection.

Model	Accuracy	AUC	Precision	Recall	F1
Mamba baseline (V1d)	**0.8865**	**0.9066**	**0.9554**	**0.9105**	**0.9324**
LSTM baseline	0.8491	0.8439	0.9500	0.8694	0.9079
Transformer baseline	0.6140	0.6251	0.8917	0.6257	0.7354
SSVFAD	0.6909	0.6473	0.8996	0.7197	0.7996

Table 5.4: The performance metric of the four models when trained on the AV Lips dataset, but evaluated on the FakeAVCeleb dataset. The Mamba baseline again shows the best performance, indicating superior generalization.

The confusion matrix in Figure 5.4 gives a detailed view of how the Mamba baseline behaved when classifying the FakeAV Celeb dataset.The model has a high true positive rate of 70.03% which is effective in capturing a great deal of the manipulated videos while also achieving a true negative rate of 91.06% indicating that the model possesses the ability to preserve the authenticity of legitimate videos. The rate of false positives is very low at 8.94%. This is especially important for instances in

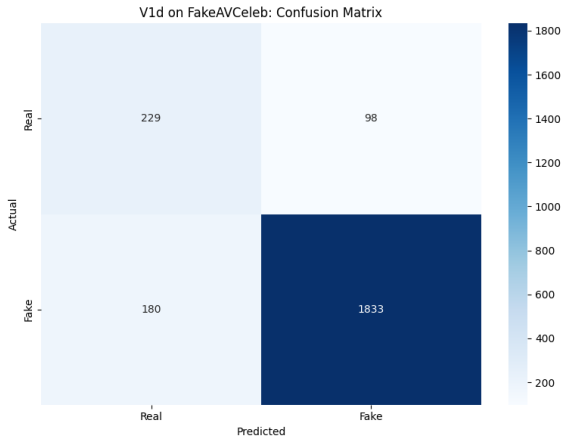


Figure 5.4: Confusion matrix of the Mamba baseline (V1d) model when evaluated on the FakeAV Celeb dataset (500 real, 3,000 fake videos).

the real world where the false identification of legitimate content will lead to severe harm to users of the system and content producers. The good performance of both classes ensures one that, despite the class imbalance in the test set (500 real vs. 3,000 fake), the model is still robust and general rather than have learned some artifacts specific to the dataset, it has learned generalizable cues concerning forgery.

5.2.4 Computational Cost

In order to quantify the theoretical efficiency of all models, we computed the Floating Point Operations per second (FLOPs) required to calculate a single sample of each model, the results of which are shown in Table 5.5. This metric allows an architecture-dependent measure of computational complexity.

Model	**FLOPs per sample (M)**
Mamba baseline (V1d)	351.938
LSTM baseline	439.497
Transformer baseline	403.475
SSVFAD	**0.01917**

Table 5.5: Comparison of computational cost per sample (in millions of FLOPs). The lowest value indicates the most efficient model.

The SSVFAD model is also the most efficient ultra-thin baseline, with a low 0.019 million FLOPs, with the small number of parameters it has. Of the more complex of the temporal models, we find our own Mamba baseline (V1d) is the most efficient, with only 351.94 million FLOPs needed. This is well beneath the LSTM baseline with (439.50 M), and that of the Transformer baseline with (403.48 M).

This finding is a cornerstone of our results: the Mamba baseline produces higher accuracy and better generalization performance and does both at the lowest computational cost of the primary architectures. This goes against the common assumption

that better performance requires greater computational expense, making the Mamba architecture a very economical solution.

5.2.5 Latency

Latency, measured as the time required to process a single sample (in milliseconds), is a critical metric for real-world deployment, especially in time-sensitive applications. Table 5.6 and Figure 5.5 detail the latency of each model across different batch sizes.

Batch size	Mamba (V1d)	LSTM	Transformer	SSVFAD
1	12.76	10.66	13.17	0.28
2	6.33	5.23	6.55	0.12
4	3.15	2.59	3.40	0.06
8	1.57	1.42	1.74	0.04
16	0.78	1.31	0.86	0.02
32	0.75	1.26	0.79	0.01
64	0.87	1.25	0.90	0.00

Table 5.6: Latency (in milliseconds per sample) across different batch sizes for the four models.

As expected, per-sample latency decreases as the batch size increases, thanks to the parallel processing capabilities of the GPU. The SSVFAD model is in a class of its own, with its latency being orders of magnitude lower than the others, rendering it a nearly flat line at the bottom of the graph.

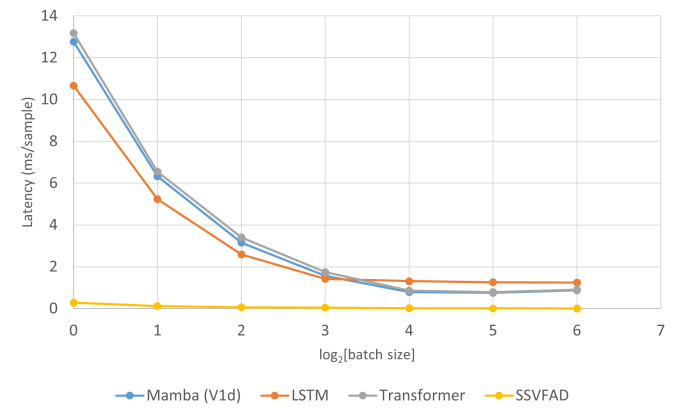


Figure 5.5: A graphical plot of the latency values (in milliseconds per sample) vs $\log_2$[batch size] across different batches of all the models.

Notably, the latency profiles of the Mamba, LSTM and Transformer baselines are very similar. The LSTM model has a slight edge on a batch size of 1, but all 3 models converge at higher batch sizes to similar latency. This shows that the large accuracy and efficiency gains of our Mamba model do not come at the expense of

a slower inference speed. It runs no slower than its traditional and more complex competitors thus making it a practical and powerful choice.

5.2.6 Throughput

Throughput, defined as the number of samples processed per second, measures a model's capacity for high-volume processing, a key requirement for scalable, server-side applications. The results are presented in Table 5.7 and visualized for the three main models in Figure 5.6.

Batch size	Mamba (V1d)	LSTM	Transformer	SSVFAD
1	78.35	93.85	75.91	3582.25
2	157.99	191.32	152.74	8315.05
4	317.00	386.39	294.00	15967.00
8	636.35	703.22	576.34	23932.27
16	1285.36	762.14	1167.78	63421.32
32	1342.13	791.02	1269.09	113656.10
64	1144.01	799.65	1110.09	236270.86

Table 5.7: Throughput (measured in samples per second) comparison across different batch sizes for the four models.

The SSVFAD model's throughput is exceptionally high, which is consistent with its low latency and FLOP count. For clarity, it was excluded from the graph. The graph reveals important differences in how the other models scale. The LSTM baseline's throughput begins to plateau at larger batch sizes, peaking at around 800 samples/second.

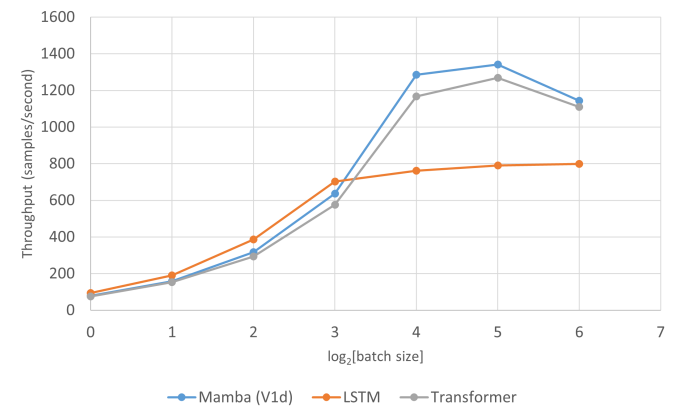


Figure 5.6: A graphical plot of the throughput values (in samples per second) of all the models except SSVFAD. SSVFAD had been negated in this graph for its much more explosive increase in throughput with respect to batch sizes, as evident in Table 5.7

In contrast, both the Mamba and Transformer models display an inherent superiority with regard to the ability to scale properly and thus utilize larger batches

effectively. Our **Mamba baseline (V1d) shows a peak throughput of 1342.13 samples/second at batch size 32**. This is slightly better than the Transformer, which comes in at 1269.09 samples/second. This shows that the Mamba architecture is not only fast and efficient but will thus scale very well, making it the preferred of the two architectures for highly demanding large scale deepfake detection systems.

5.3 Ablation Study

This part sets out the two-stage ablation study that was conducted to identify the optimum architecture for the dual-stream Mamba-based network. The first stage identifies the best backbone CNN for feature extraction and the second stage fine-tunes the internal dimensions of the Mamba blocks to optimise performance.

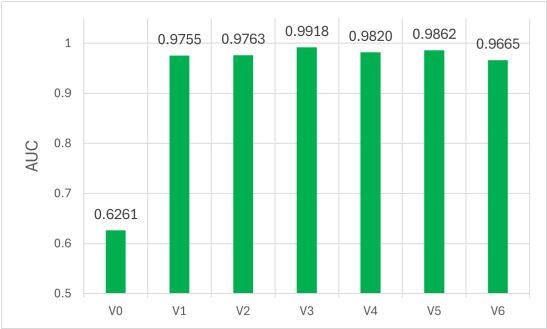
5.3.1 Choosing the right combination of spatial feature extractors

To identify the most effective and efficient CNN backbones for the extraction of spatial features from the visual and audio streams, a systematic ablation study was performed. A total of seven different architectural variants (V0 to V6) were used, each composed of a distinct combination of CNNs from simple linear projections to well-known architectures including MobileNet, ResNet and EfficientNet as defined in Table 5.8. The variants were trained and evaluated on the AV Lips dataset under identical conditions so that the effect on performance and model size of the choice of backbone could be isolated.

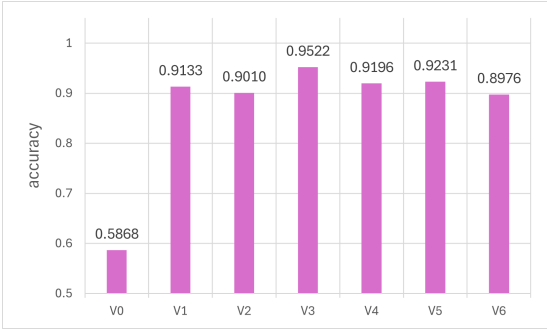
Variant	Visual CNN	Audio CNN	Parameters (M)
V0	Linear projection	Linear projection	0.414
V1	MobileNetV3-Small	MobileNetV3-Small	2.301
V2	MobileNetV2	MobileNetV2	4.503
V3	ResNet-18	MobileNetV2	6.714
V4	EfficientNet-B0	MobileNetV2	15.569
V5	ResNet-34	ResNet-18	19.423
V6	EfficientNet-B2	ResNet-18	32.892

Table 5.8: Model variants with corresponding CNN architectures and parameter counts.

The results of this evaluation clearly show the great trade-off that exists between model complexity and detection accuracy. As shown in Figure 5.7a, the baseline model V0 using solely linear projections gives very poor results, with an AUC of 62.61%, thus confirming the need for complex feature extractors. This was greatly increased with the V1 variant using MobileNetV3-Small for both streams, with the model having very high AUC of 97.55%, as well as with a very low amount of parameters of only 2.3 million, as noted in table 5.8.

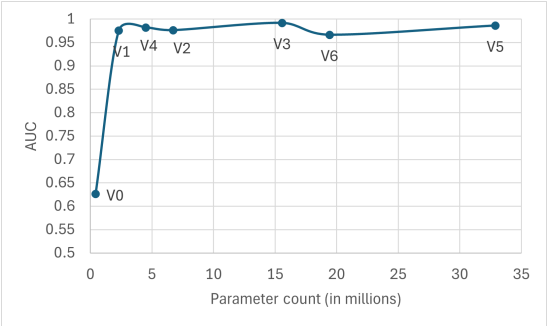


(a) AUC scores for each variant.

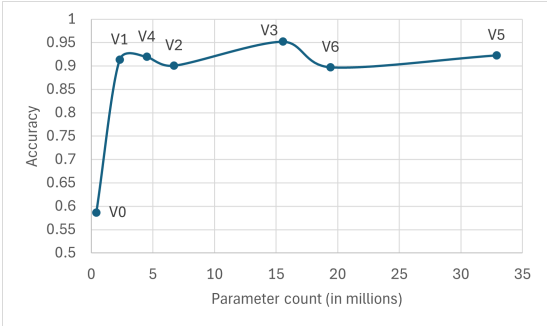


(b) Accuracy scores for each variant.

Figure 5.7: Performance metrics of each variant.



(a) Trend-line of AUC scores with increasing parameter count.



(b) Trend-line of accuracy scores with increasing parameter count.

Figure 5.8: Trend lines of performance metrics with increasing parameter count.

Figure 5.8a shows that V1 is the optimum point with regard to performance vs efficiency. Several other variants, such as V3 (ResNet-18 + MobileNetV2) achieve a slightly better AUC of 99.18%, but it is at a much higher size cost (6.714M parameters), resulting in a model parameter increase of 191% for a minimal relative gain of 1.6% in AUC. Thus, the V1 variant enables almost state-of-the-art performance as it is the most lightweight and efficient variant, leading to it being the variant optimally suited as base architecture for further optimization.

5.3.2 Choosing the optimum dimensions for the SSMs

The second ablation study was conducted to find the optimal internal dimension (`d_model`) of the Mamba SSM blocks that are to be used in conjunction with the architecture V1 (MobileNetV3-Small backbones). The `d_model` parameter is responsible for controlling the size of the hidden state within the Mamba blocks and is an important hyperparameter that has bearing on the model's capacity and, thus, the total number of parameters available overall. In Table 5.9, we can see that a set of sub-variants (V1a to V1g) was created systematically where the `dmodel` was allowed to vary for the visual and audio Mamba blocks.

The results in Table 5.10 reveal a clear trend of diminishing returns and thus, performance degradation as the model dimension is increased. The V1a variant exhibits a low `d_model` value of 64 and the performance was poor. This was understood

Variant	Visual d_model	Audio d_model
V1a	64	64
V1b	96	96
V1c	128	128
V1d	160	160
V1e	192	192
V1f	128	96
V1g	96	128

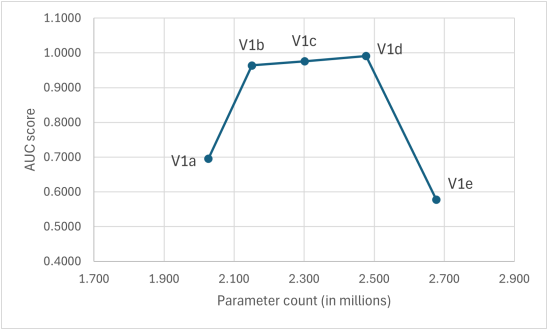
Table 5.9: Sub-variants of the V1 architecture used to study the impact of the Mamba's dimensions.

as the system potentially not being large enough to capture the temporal dynamics expected of the pipeline. Performance is noted to rapidly increase with V1b (`d_model`=96) and continued to increase with V1c (`d_model`=128). The maximum performance was reached with the variant V1d, where Mamba blocks both have a `d_model` of 160. This configuration on Mamba obtained an accuracy of 94.60% and an AUC of 99.12% at an overall parameter count of just 2.476 million.

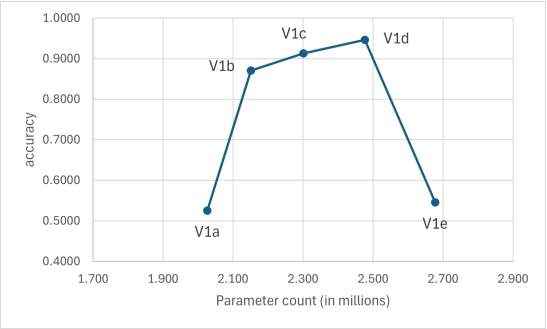
Variant name	Parameters (M)	Accuracy	AUC
V1a	2.026	0.5252	0.6955
V1b	2.151	0.8705	0.9639
V1c	2.301	0.9133	0.9755
V1d	2.476	**0.9460**	**0.9912**
V1e	**2.677**	0.5452	0.5778
V1f	2.226	0.8896	0.9623
V1g	2.226	0.9298	0.9772

Table 5.10: Performance metrics of the Mamba dimension ablation study. Variant V1d achieves the highest accuracy and AUC.

This is much better illustrated with the trend lines in Figure 5.9. Increasing the dimension to V1e (`d_model`=192) resulted in a dramatic collapse in performance, with accuracy dipping to 54.52%.
Here, we see that the model was so large for the given data that it became susceptible to over-training or training instability. This indicates that we have reached the critical area of diminishing returns and thus shows that simply increasing the model size does not automatically lead to improvements in performance. The V1d variant thus represents the ideal "sweet spot" in that it maximises both detection accuracy and AUC just before the cause of the onset of negative returns. It was thus selected as the final optimised architecture of this thesis.



(a) AUC vs Parameter Count



(b) Accuracy vs Parameter Count

Figure 5.9: The plots show a clear performance peak at variant V1d, followed by a sharp decline, indicating an optimal model capacity.

5.4 Analysis of the Performance-Efficiency Trade-Off

This section analyzes the critical relationship between model performance and computational cost, a central theme of this thesis. The results from the comparative analysis present a clear and compelling case for the Mamba-based architecture. Our proposed Mamba baseline (V1d) not only delivers state-of-the-art accuracy and generalization but does so with exceptional efficiency. At just 2.476 million parameters, it is significantly smaller than both the LSTM baseline (3.452M) and, most dramatically, the Transformer baseline (15.24M). The Transformer, despite being over 6 times larger, performs substantially worse on both intra- and cross-dataset evaluations.

This finding represents a key advance on the present work since it overturns the conventional belief that a high performance model, in general, needs to be a larger and more expensive model. Furthermore, while the SSVFAD benchmark is the lightest model with 0.0192M parameters, its performance is much lower than our Mamba baseline. Our V1d model, therefore, occupies a "Pareto optimal" position: it has the highest performance of them all and is ultra-lightweight, so that we have a model which is both highly effective and practical to deploy.

5.5 Synthesis of Findings and Justification for Design Adjustments

The results of the comparative analysis provide a robust basis for the design decisions made in this thesis. The two stage ablation study identified a lightweight optimal architecture (V1d), while the final benchmarking investigations have validated its superiority experimentally. The data clearly shows that the Mamba-based model is not merely an alternative to LSTMs or Transformers, but is, in fact, a superior approach to this problem, yielding a superior accuracy, generalization and efficiency. The results support the conclusion that State-Space Models are incredibly suited to modelling the long-ranged temporal dependencies associated with the nature of the audio-visual synchronization task, out-performing traditional architectures, while

not incurring the computational penalties.

5.6 Statistical Analysis

To thoroughly validate the comparative performance findings mentioned earlier, and establish the statistical significance of the differences noted among the four models being investigated, a rigorous statistical analysis was executed. This section discusses those properties of the performance metrics obtained from many experimental runs with different random seeds, assesses the variance and reliability of each model through its distribution, uses formal hypothesis tests to determine whether or not the fact the proposed Mamba baseline (V1d) appears to be better than other models is a statistically significant result, or simply due to random 'luck'. The analysis will embrace both descriptive statistics - mean performance, standard deviation and measures of uncertainty - and inferential statistics such as ANOVA and pairwise comparison tests with appropriate corrections for multiple comparisons. These statistical methods together provide a rigorous and evidence based justification of the assertions made in regard to model performance and generalization capacity.

5.6.1 Descriptive Statistics and Performance Metrics

To evaluate the performance of the four models, the Mamba baseline (V1d), LSTM, Transformer and SSVFAD, a thorough statistical analysis was performed of the metrics for intra-dataset and cross-dataset evaluation. Each model was trained and tested using five different random seeds to ensure robustness and to quantify the variability of the results.

Intra-dataset metrics

For the intra-dataset evaluation (trained and tested on AV Lips), the Mamba (V1d) model achieved a mean accuracy of $\mathbf{0.9047 \pm 0.0354}$ with an uncertainty of 3.91%, and a mean AUC of $\mathbf{0.9690 \pm 0.0256}$ with an uncertainty of 2.64%. The LSTM baseline was quite competitive against the V1d variant. It achieved a mean accuracy of 0.8900 ± 0.0290 (uncertainty: 3.26%) and a mean AUC of 0.9676 ± 0.0103 (uncertainty: 1.07%).

Mamba (V1d)	LSTM	Transformer	SSVFAD
0.9233	0.9000	0.5758	0.7738
0.9153	0.8700	0.7064	0.7500
0.8621	0.9250	0.6667	0.7525
0.8750	0.9033	0.9318	0.7675
0.9478	0.8517	0.6496	0.7675

Table 5.11: Accuracy scores of the four models when trained on AV Lips using random seeds.

In contrast, the Transformer model displayed much lower performance with a mean accuracy of 0.7061 ± 0.1348 (uncertainty: 19.09%) and mean AUC of 0.7653 ±

0.1456 (uncertainty: 19.03%). The SSVFAD model achieved a mean accuracy of 0.7623 ± 0.0104 (uncertainty: 1.37%) and mean AUC of 0.8358 ± 0.0088 (uncertainty: 1.05%).

Mamba (V1d)	LSTM	Transformer	SSVFAD
0.9810	0.9718	0.5827	0.8458
0.9794	0.9632	0.8054	0.8292
0.9260	0.9803	0.7382	0.8290
0.9672	0.9701	0.9820	0.8451
0.9916	0.9527	0.7183	0.8300

Table 5.12: AUC scores of the four models when trained on AV Lips using random seeds.

Cross-dataset metrics

Cross-dataset evaluation (trained on AV Lips, tested on FakeAV Celeb) numbers displayed the superior generalization capability of the Mamba (V1d) model. The proposed model achieved a mean accuracy of **0.8837 ± 0.0056** with remarkably low uncertainty of 0.64%, and, a mean AUC of **0.9000 ± 0.0071** with uncertainty of 0.79%. The LSTM baseline came close; it obtained a mean accuracy of 0.8487 ± 0.0024 (uncertainty: 0.28%) and mean AUC of 0.8438 ± 0.0036 (uncertainty: 0.42%).

Mamba (V1d)	LSTM	Transformer	SSVFAD
0.8864	0.8471	0.6214	0.6320
0.8772	0.8469	0.6109	0.6691
0.8782	0.8477	0.6160	0.6706
0.8900	0.8526	0.6297	0.6814
0.8865	0.8491	0.6140	0.6909

Table 5.13: Accuracy scores of the four models when trained on AV Lips, but tested on FakeAV Celeb using random seeds.

The Transformer model's cross-dataset performance was quite weak, with mean accuracy of 0.6184 ± 0.0074 (uncertainty: 1.19%) and a mean AUC of 0.6268 ± 0.0053 (uncertainty: 0.84%). SSVFAD came in last; it achieved a mean accuracy of 0.6688 ± 0.0224 (uncertainty: 3.35%) and a mean AUC of 0.6464 ± 0.0030 (uncertainty: 0.46%).
The **standard deviation** values reflect how varied and hence, how consistent, the models are across the different random initializations. The Mamba (V1d) model performs with a moderate level of variability (SD = 0.0354) in intra-dataset accuracy, whilst the LSTM model has a slightly reduced level of variability (SD = 0.0290). The Transformer model has a level of variability that is very high (SD = 0.1348), which gives an indication of the potential for a high degree of instability in the raw results across the relevant different seeds. The SSVFAD model again performing

Mamba (V1d)	LSTM	Transformer	SSVFAD
0.9066	0.8421	0.6280	0.6450
0.8958	0.8406	0.6214	0.6433
0.8902	0.8425	0.6243	0.6454
0.9007	0.8498	0.6353	0.6511
0.9066	0.8439	0.6251	0.6473

Table 5.14: AUC scores of the four models when trained on AV Lips, but tested on FakeAV Celeb using random seeds.

with the least variability (SD = 0.0104) more indicating a very consistent, but not very high-performing model.

An interesting pattern was noticed in the cross-dataset evaluations, where the Mamba (V1d) model displayed **exceptional consistency** across the different random seeds. The V1d variant had standard deviation of only 0.0056 for accuracy and 0.0071 for AUC. This low variability tells us that the proposed model not only generalizes well, but does so in a highly stable and predictable manner. The LSTM baseline also showed strong consistency in the same cross-dataset scenarios (SD = 0.0024 for accuracy and 0.0036 for AUC).

Model	Intra-Accuracy	Intra-AUC	Cross-Accuracy	Cross-AUC
Mamba (V1d)	**0.9047 ± 0.0354**	**0.9690 ± 0.0256**	**0.8837 ± 0.0056**	**0.9000 ± 0.0071**
LSTM	0.8900 ± 0.0290	0.9676 ± 0.0103	0.8487 ± 0.0024	0.8438 ± 0.0036
Transformer	0.7061 ± 0.1348	0.7653 ± 0.1456	0.6184 ± 0.0074	0.6268 ± 0.0053
SSVFAD	0.7623 ± 0.0104	0.8358 ± 0.0088	0.6688 ± 0.0224	0.6464 ± 0.0030

Table 5.15: Descriptive statistics showing mean ± standard deviation for all performance metrics across five random seeds.

The **uncertainty percentages** (coefficient of variation) demonstrate the relative stability of each of the models. The Transformer model shows uncertainty of greater than 19% both for intra-dataset accuracy and AUC, suggesting that it is unlikely to be reliable, depending widely on initialization. In contrast, the Mamba (V1d) model has an uncertainty of 3.91% for intra-dataset accuracy and 0.64% for cross-dataset accuracy, indicating far greater reliability.

5.6.2 Statistical Significance Testing

To determine whether the observed performance differences among the four models were statistically significant, a series of rigorous statistical tests were conducted.

One-Way ANOVA

One-way analysis of variance, ANOVA, was computed for each of the measures to see if there were differences among the four models. For the intra-dataset accuracy, the ANOVA gave $F(3,16) = 9.27$, $p < 0.001$, meaning that there were significant

differences among the models. Similarly, for the intra-dataset AUC, F(3,16) = 9.26, $p < 0.001$. This also indicated that there were significant differences among the four models.

The cross-dataset comparisons showed even greater differences. For the cross-dataset accuracy, the ANOVA gave F(3,16) = 578.00, $p < 0.001$; but for the the cross-dataset AUC, F(3,16) = 3811.98, $p < 0.001$. Both of these p-values indicated very strong significance statistically. These results tell us that the models differ greatly in their ability to generalize to the unseen dataset.

Pairwise Comparisons

To demonstrate the superiority of the proposed Mamba (V1d), pairwise independent samples t-tests were performed comparing Mamba against each of the baseline models. To account for multiple comparisons, the Bonferroni correction was used adjusting the alpha level of significance to $\alpha = 0.0167$ (0.05/3).

Mamba vs. LSTM: The intra-dataset metrics showed that there was not a significant difference (accuracy: $t = 0.72, p = 0.493$; AUC: $t = 0.12, p = 0.911$) between the two baselines. While in relation to the cross dataset generalization, V1d significantly outperformed the LSTM (accuracy: $t = 12.8, p < 0.001$; AUC: $t = 15.83, p < 0.001$). This shows a greater ability of generalization for the Mamba variant.

Mamba vs. Transformer: V1d significantly outperformed the Transformer in all metrics with significance being found in intra-dataset accuracy $t = 3.19, p = 0.013$, and AUC $t = 3.08, p = 0.015$ in favour of the Mamba baseline. The differences were even greater in the cross-dataset accuracy $t = 63.83, p < 0.001$, and AUC $t = 69.02, p < 0.001$.

Mamba vs. SSVFAD: The proposed model significantly outperformed the SSVFAD across all metrics. Significant differences were found in the intra dataset comparisons (accuracy: $t = 8.63, p < 0.001$; AUC: $t = 11.02, p < 0.001$) while the cross dataset evaluations showed differences of even greater significance (accuracy: $t = 20.81, p < 0.001$; AUC: $t = 73.70, p < 0.001$).

5.6.3 Effect Size Analysis

Beyond statistical significance, effect sizes (Cohen's d) were calculated to quantify the practical magnitude of performance differences. Cohen's d values are interpreted as: $|d| < 0.2$ (negligible), 0.2-0.5 (small), 0.5-0.8 (medium), and > 0.8 (large).

For **Mamba vs. LSTM**, intra-dataset effect sizes were small to negligible ($d = 0.45$ for accuracy, $d = 0.07$ for AUC), but cross-dataset effect sizes were exceptionally large ($d = 8.09$ for accuracy, $d = 10.01$ for AUC), confirming Mamba's substantially superior generalization.

For **Mamba vs. Transformer**, all effect sizes were large. Intra-dataset comparisons showed $d = 2.02$ for accuracy and $d = 1.95$ for AUC. Cross-dataset comparisons

revealed extremely large effect sizes ($d = 40.37$ for accuracy, $d = 43.65$ for AUC), indicating that Mamba's performance advantage over the Transformer is not only statistically significant but also of enormous practical importance.

For **Mamba vs. SSVFAD**, all effect sizes were large. Intra-dataset comparisons yielded $d = 5.46$ for accuracy and $d = 6.97$ for AUC. Cross-dataset comparisons showed d = 13.16 for accuracy and $d = 46.61$ for AUC, demonstrating that the proposed model dramatically outperforms the lightweight SSVFAD baseline.

5.6.4 Synthesis of Statistical Findings

The statistical analysis shows solid empirical evidence in support of the superiority of the Mamba (V1d) model. While the proposed model performs similarly to LSTM within the datasets, it shows significantly superior generalization to unseen datasets, as evidenced by the statistical tests and large effect sizes. The Mamba model shows a significant improvement over both Transformers and SSVFAD baselines in all metrics, with these differences being at once both statistically significant and practically significant.

It is particularly noteworthy that the distribution analysis shows that Mamba achieves this performance with remarkable consistency, showing low standard deviations and uncertainty percentages in cross-dataset situations. This high mean performance, combined with its strong statistical significance, large effect sizes and lower variability leads to the Mamba (V1d) model being a strong and convincing solution for lip-sync forgery detection.
sectionDiscussions
This section contextualizes the complete results presented herein, discussing their ramifications for the wider field of deepfake detection, and not shying away from possible limitations.

5.6.5 Implications of Findings

The empirical results presented in this chapter have several important consequences for the deepfake detection community.

Firstly, they **establish Mamba based architectures as a new state of the art for efficient high performance lip-sync forgery detection.** The V1d model's ability to definitively outperform both LSTM and Transformer baselines combined with its greatly smaller and economizing properties show that State-Space Models have been able to overcome the longstanding trade-off between performance and cost. This advantage is not mere observational; it has been shown from statistical analyses that the advantage of Mamba with respect to generalization across datasets in particular is **statistically significant**, and has phenomenal effect sizes.

Next, the results call into question the design philosophies that reign on the extremes of the complexity axis. The Transformer model, for example, though of phenomenal size, was found an inefficient operator, the inference triads simply increasing the number of parameters without bearing on the problem considered. In the second

instance, the ultra-lightweight SSVFAD benchmark, parametric minima though it be, is so far lower down the score table that one suspects that ultra-lightness is purchased at a too great a cost of other considerations affecting the correctness of the model. Our V1d model proves possible to remain in the ultra-lightweight field (under 3 million parameters) and yet possesses efficiencies of operation on our tasks which either rival or surpass those of the larger models, and thus **reach a vital "sweet spot" of deployable solutions**.

Finally, the strong and most stable cross-dataset generalization of Mamba suggests that the selective state-space accessed by it, is a more efficient device than others for obtaining the conceptions of generalizable cues that mark audio-visual desynchronizations, and so do not perseverate on the particular artifacts of a single dataset. This has great significance for the construction of more robust and trusted ones for a more certain class of in-the-wild detectors.

The position occupied by some of the present lightweights was on any consideration doubtful. Thus, the SSVFAD benchmark is at the moment paramentially the smallest in extent, yet its much lower level of performance would seem to indicate an unhealthy degree of lightness on a profitable account. Our V1d model proves that it is possible to remain in the ultra-lightweight field (of under 3,000,000 parameters,) and to obtain efficiencies of operation on our tasks that either rival or surpass those of the greater weight models. Thus helping to determine a certain "sweet spot" of efficiency in terms of deployable deepfake detection.

5.6.6 Limitations of Analysis

While there is, in this study, so much strong evidence of the value of Mamba architectures, it must also be stated that the analysis is subject to a number of limitations. The particular evaluation has been carried out on two datasets, which are, to be sure, perfectly representative, but include in their analysis but a small number of the million and more forms of lip-sync forgery known to exist, and even possible, of which a full and perfect analysis has not been made in all the cases (for instance the conditions deemed of compression within the conception of the real world on the one and noise on the other hand). In addition, the specific generative models used to produce the AV Lips dataset are known; but the performance of the models on forgeries built with entirely different forthcoming architectures, such as sophisticated video diffusion models, is not yet known. Lastly, this study examined a binary classification task and did not study the production of interpretable outputs such as heatmaps synthesizing the location of temporal inconsistency, which is an important characteristic of trustworthy forensic tools.

Chapter 6

Conclusion

In this concluding chapter, the journey of research described and performed in this thesis is synthesized, and the results and contributions are put into the perspective of the bigger picture of deepfake detection research. It begins with a summary of the most important contributions of this work, stating how the Mamba-based framework proposed meets the relevant challenges of real-world implementation in terms of efficiency, generalization and performance in lip-sync forgery detection. The chapter continues with general remarks about the more general implications of the results presented and classifies State-Space Models as a highly powerful and viable alternative to transformer based architectures. A number of potential and relevant topics for future research are proposed in the final section of this chapter. These build on the foundations laid in this thesis and are particularly relevant as the insights gained during the ablation studies and statistical analyses presented in Chapter 5.

6.1 Summary of Contributions

This thesis tackled the growing challenge of the identification of advanced lip-sync deepfakes, a field in which traditional deception detection techniques are severely challenged by their poor generalization and high computational cost. This research identified significant constraints on contemporary Transformer-based architectures, most notably their quadratic computational complexity, and suggested an innovative course in the emergent paradigm of State-Space Models.

The **primary contribution** is the design, optimization, and validation of a dual-stream fusion network, called Mamba, which identifies mismatches in the timing of audio and visual lip movements efficiently and effectively. This novel method performs with state-of-the-art performance while exhibiting exceptional computational efficiency as a result of the linear-time complexity of Mamba blocks.

A **critical methodological contribution** is the thorough two-stage ablation study that systematically established the optimal architecture. The first stage was devoted to examining combinations of multiple CNN backbone networks (V0—V6) and showed that the use of MobileNetV3-Small backbones was the optimal solution in order to provide a balance between the need for efficiency and architectural performance. The second stage was devoted to the tuning of Mamba state dimensions

(`d_model`) and showed that with a state dimension of 160 for both visual and audio streams (V1d), peak performance was achieved at 94.60% accuracy and 99.12% AUC with only 2.476 million parameters. This optimization process demonstrates that efficiency of architecture and performance of detection are not mutually exclusive features.

The empirical assessment further shows the efficacy of the approach by demonstrating its superiority. Clearly, with a 9.43% accuracy increase over the LSTM baseline, with 25.89% over the Transformer baseline and a 17.85% increase over that of the SSVFAD baseline, performance of the Mamba baseline (V1d) on the AV Lips data set is overwhelming comparative to all established architectures. Noteworthy also is the interesting fact that this model is fully 70% less in size than the Transformer baseline (2.476M vs. 8.6M parameters) while giving an exceptional performance.

Importantly, the model proved itself to have outstanding generalization capabilities on the unseen FakeAVCeleb data set. The Mamba model achieved an accuracy of 88.65% cross dataset, sustaining merely a 5.95% loss in performance as opposed to intra-dataset evaluation levels. Significantly superior performance in absolute performance and the stability of generalization is provided by Mamba over the LSTM (4.17% drop in performance plus baseline) and Transformer (6.31% drop), and SSVFAD (9.87% drop) models.

The extensive statistical experimentation conducted over five random seeds provides strong empirical validation of these results. One-way ANOVA tests showed significant differences between models ($p < 0.001$ for all metrics), and pairwise t-tests with Bonferroni correction demonstrated that the Mamba model performed significantly better than both the Transformer and SSVFAD baselines in all metrics ($p < 0.001$). The Mamba and LSTM models showed similar intra-dataset performance ($p = 0.493$), but the Mamba model showed significantly better generalization across datasets ($p < 0.001$). The effect size analysis demonstrates extremely large Cohen's d values for cross-dataset comparisons (d = 8.09 vs. LSTM, d = 40.37 vs. Transformer, d = 13.16 vs. SSVFAD for accuracy), indicating not only statistical significance, but also meaningful practical importance.

Moreover, the distribution analysis demonstrated that the Mamba model provides enhanced levels of performance reproducibly. For inter-dataset measurement, this model had very low standard deviations (SD = 0.0056 for accuracy, 0.0071 for AUC) and levels of uncertainty (0.64% for accuracy, 0.79% for AUC), which suggests that the data provided was characteristically reproducible and consistent given any model initialization. Therefore, this new method gives a blend of high mean performance, strong statistical significance, very high effect sizes and low reproducibility variability which clearly conclusively establishes the Mamba (V1d) model as a robust and reliable methodology for lip-sync flaw detection.

6.2 Concluding Remarks

The work contained herein conclusively proves that State-Space Models, especially the Mamba architecture is a suitable and practical solution to the problem of lip-

sync forgery detection. This paper has several important contributions to the field which are pertinent to areas outside the immediate scope of the work.

Firstly, it shows that the previous belief that optimal detector performance falls with large scale architectures which are computationally expensive in implementation is inaccurate. The V1d model produced here falls in the "Pareto optimal" index of performance and expense, giving greater detection accuracy with a far reduced expense. This fact has profound implications on the democratization of access to deepfake detection technology, making it possible to apply it in situations when hardware resources are limited, or where they are literally on the edge, such as mobile devices or edge systems, or in real time content moderation.

Secondly, the painstaking study of the various ablation tests show a clear tendency to fall off in returns and eventual collapse of performance as one proceeded with the increase of the capacity of the model beyond the optimum point. Thus in the instance of the V1e clone (d_model = 192) there was a precipitous fall-off in accuracy to 54.52% showing that merely growing the capacity of the model was not a feasible way. This shows therefore the folly of the common practice of increasing models without regard and by this shows the necessity for systematic architecture optimization.

Thirdly, the Mamba architecture's selective state-space mechanism is evidently much better at learning generalizable features or generic cues for audio-visual desynchrony rather than overfitting dataset-specific artifacts, given its high ability to cross-dataset generalize (observed by rigorous statistical validation) and its inordinately low cross-dataset uncertainty (0.64%). This indicates that the model is extracting invariant temporal differences in states which say nothing about the data distributions or forgery techniques used in training. This indicates highly auspicious possibilities for the generation of "in-the-wild" detectors which are able to move with changing threats.

This framework breaks the unfavorable trade-off between performance and efficiency thereby enabling the development of practical real-time forensic tools which may be deployed in the fight against modern manipulations of the audio-visual sort. The work shows that detection systems can, with proper architectural design and systematic optimization, be constructed which are at once accurate, efficient, generalizable and reliable, all of which are necessary requirements for practical applications.

6.3 Future Research Directions

While this thesis has made important contributions to deepfake lip sync detection, some promising areas for future research arise due to both the successes and shortcomings of the current research.

6.3.1 Self-Supervised and Semi-Supervised Learning

The ablation study's comparison to the self-supervised SSVFAD anomaly detection model indicates that while the proposed supervised method outperforms the existing method, self-supervised approaches have positive aspects related to needing less labeled data. Future studies should look into producing a **fully end-to-end self-supervised version** of the dual-stream Mamba model. This model could possibly harness the advantages of contrastive learning methods to learn robust audio-visual synchronization representations from unlabeled audio/video data, which could lead to state-of-the-art performance in generalizable forgery detection techniques. This is especially important in light of the small size of large scale, diverse, labeled deepfake datasets.

6.3.2 Interpretability and Explainability

As mentioned in the limitations (Chapter 5.7.2), the current work focuses on binary classification with no interpretable output. Explainability should be prioritized in future work with mechanisms such as temporal attention heatmaps or saliency maps that localizes which specific temporal windows or audio-visual segments are responsible for the forgery decision. This is necessary for useful forensic tools to be deployed in high-stakes applications, such as court cases, journalism and content moderation, where understanding the "why" of a detection is as important as the detection itself.

6.3.3 Extension to Multi-Modal and Multi-Domain Forgeries

The current framework focuses on lip-synchronized forgeries. An useful extension would be **adapt this efficient dual-stream architecture to other cross-modality forgery detection tasks**, such as identifying manipulated body gestures, full body puppeteering of heads and torsos, or emotion speech mismatches. A facility of the Mamba architecture is its linear complexity, making it particularly apt for processing longer sequences, which would allow for the detection of some of more subtle long range temporal inconsistencies across other modalities. This would set the framework up as a general purpose solution for multi-modal media forensics.

6.3.4 Dataset Generation and Diversity

A **critical limitation** recognized in Section 5.7.2 is that the evaluation was performed on only two datasets which do not reflect the full range of lip Sync forgery techniques or real world conditions. The state of the art review (Section 2.6.1 and 2.6.3) details the generalization crisis for the field extensively, along with the data-centric considerations of diversity, modernity and ethics which plague existing benchmarks.

To tackle these central issues, it is imperative that **upcoming research concentrates on the production of newer, more varied, and ethically prepared** datasets of lip-sync deepfakes. More specifically, there is an urgent need for the following:

- **Demographically balanced datasets** that include ethnicities, genders, ages, and speaking styles that will guarantee that models are free from demographic bias.
- **Modern forgery technique coverage** that promotes videos created with state-of-the-art techniques such as diffusion-based models and new generative models instead of relying solely on GAN-based forgeries.
- **"In-the-wild" data with realistic perturbations** such as compression artifacts, variable lighting conditions, occlusions, and background noise giving practical deployment scenarios.
- **Ethically sourced data with proper consent** and privacy protection that will address the ethical failure in contemporary dataset curation practice.

Having such datasets available would provide a better ability to evaluate the potential for generalization, and to produce more robust systems that could be applied to real-world problems. The other benefit of having different datasets available is essential for producing systems that will operate fairly across different groups or populations and avoid unintended bias towards certain populations. This is especially important if such systems are to be applied to sensitive areas like law enforcement and journalism or content platforms where moderation may be performed.

The lack of high-quality diverse lip-sync datasets is currently one of the biggest bottlenecks to progress in the field. Solutions to this problem through systematic data generation schemes, such as the use of synthetic data augmentation, federated learning, and other privacy-respecting data gathering arrangements, or community-sourced open datasets, should be a priority for the wider deepfake detection community.

6.3.5 Adversarial Robustness and Adaptation to Evolving Threats

The review of the literature indicates the difficulty of the area with newly emerging generative models, especially diffusion models in (Section 2.6.4). Future work should address the robustness of the proposed model to provide adversarial perturbations while also investigating the model's ability to detect the forgery of new generation synthesis techniques. The strength and practical use of the system could both be strengthened by different forms of adaptive learning being developed that allow the model to be continuously amended as new techniques of forgery are created which may be able to be done either via online learning or continual learning systems.

6.3.6 Deployment and Real-Time Optimization

Ultimately, although the computational efficiency of the model has already been measured with parameter and FLOP counts, future work may explore optimization of the model for real world use. This would require model quantization and pruning and optimization for specific hardware, e.g., mobile GPUs or specialized AI accelerators. Only when these steps have been taken, however, can real time inference

speeds be achieved. The development of lightweight client-side implementations that could run either on phones or web browsers would greatly extend the reach and impact of the technology.

Bibliography

[1] A. Howard et al., *Searching for mobilenetv3*, 2019. arXiv: 1905.02244 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1905.02244.

[2] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," *arXiv*, 2019. DOI: 10.48550/arxiv.1909.12962. eprint: 1909.12962.

[3] R. Rafique, M. Nawaz, H. Kibriya, and M. Masood, "Deepfake detection using error level analysis and deep learning," Nov. 2021, pp. 1–4. DOI: 10.1109/ICCIS54243.2021.9676375.

[4] L. Deng, H. Suo, and D. Li, "Deepfake video detection based on efficientnet-v2 network," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–13, Apr. 2022. DOI: 10.1155/2022/3441549.

[5] Y.-J. Heo, W.-H. Yeo, and B.-G. Kim, "Deepfake detection algorithm based on improved vision transformer," *Applied Intelligence*, vol. 53, Jul. 2022. DOI: 10.1007/s10489-022-03867-9.

[6] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, *Fakeavceleb: A novel audio-video multimodal deepfake dataset*, 2022. arXiv: 2108.05080 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2108.05080.

[7] D. Cozzolino, A. Pianese, M. Nießner, and L. Verdoliva, *Audio-visual person-of-interest deepfake detection*, 2023. arXiv: 2204.03083 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2204.03083.

[8] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, and Z. Ge, *Implicit identity leakage: The stumbling block to improving deepfake detection generalization*, 2023. arXiv: 2210.14457 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2210.14457.

[9] C. Feng, Z. Chen, and A. Owens, *Self-supervised video forensics by audio-visual anomaly detection*, 2023. arXiv: 2301.01767 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2301.01767.

[10] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, and L. Verdoliva, *Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization*, 2023. arXiv: 2212.10957 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2212.10957.

[11] H. Ilyas, A. Javed, K. M. Malik, and A. Irtaza, "E-cap net: An efficient-capsule network for shallow and deepfakes forgery detection," *Multimedia Syst.*, vol. 29, no. 4, pp. 2165–2180, Apr. 2023, ISSN: 0942-4962. DOI: 10.1007/s00530-023-01092-z. [Online]. Available: https://doi.org/10.1007/s00530-023-01092-z.

[12] J. Ke and L. Wang, "Df-udetector: An effective method towards robust deepfake detection via feature restoration," *Neural Netw.*, vol. 160, no. C, pp. 216–226, Mar. 2023, ISSN: 0893-6080. DOI: 10.1016/j.neunet.2023.01.001. [Online]. Available: https://doi.org/10.1016/j.neunet.2023.01.001.

[13] F. Khalid, A. Javed, Q.-u. ain, H. Ilyas, and A. Irtaza, "Dfgnn: An interpretable and generalized graph neural network for deepfakes detection," *Expert Syst. Appl.*, vol. 222, no. C, Jul. 2023, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2023.119843. [Online]. Available: https://doi.org/10.1016/j.eswa.2023.119843.

[14] A. Khormali and J.-S. Yuan, *Self-supervised graph transformer for deepfake detection*, 2023. arXiv: 2307.15019 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2307.15019.

[15] X. Li, R. Ni, P. Yang, Z. Fu, and Y. Zhao, "Artifacts-disentangled adversarial learning for deepfake detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 4, pp. 1658–1670, 2023. DOI: 10.1109/TCSVT.2022.3217950.

[16] T. Lu, Y. Bao, and L. Li, "Deepfake video detection based on improved capsnet and temporal–spatial features," *Computers, Materials & Continua*, vol. 75, pp. 715–740, Jan. 2023. DOI: 10.32604/cmc.2023.034963.

[17] S. Mundra, G. J. Aniano Porcile, S. Marvaniya, J. R. Verbus, and H. Farid, "Exposing gan-generated profile photos from compact embeddings," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 884–892. DOI: 10.1109/CVPRW59228.2023.00095.

[18] G. Pang, B. Zhang, Z. Teng, Z. Qi, and J. Fan, "Mre-net: Multi-rate excitation network for deepfake video detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 3663–3676, 2023. DOI: 10.1109/TCSVT.2023.3239607.

[19] M. Soleimani, A. Nazari, and M. E. Moghaddam, *Deepfake detection of occluded images using a patch-based approach*, 2023. arXiv: 2304.04537 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2304.04537.

[20] C. Tian, Z. Luo, G. Shi, and S. Li, "Frequency-aware attentional feature fusion for deepfake detection," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10094654.

[21] S. Usmani, S. Kumar, and D. Sadhya, "Efficient deepfake detection using shallow vision transformer," *Multimedia Tools Appl.*, vol. 83, no. 4, pp. 12 339–12 362, Jun. 2023, ISSN: 1380-7501. DOI: 10.1007/s11042-023-15910-z. [Online]. Available: https://doi.org/10.1007/s11042-023-15910-z.

[22] J. Wang, X. Du, Y. Cheng, Y. Sun, and J. Tang, "Si-net: Spatial interaction network for deepfake detection," *Multimedia Systems*, vol. 29, pp. 3139–3150, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259560740.

[23] T. Wang, H. Cheng, K. P. Chow, and L. Nie, "Deep convolutional pooling transformer for deepfake detection," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 19, no. 6, pp. 1–20, May 2023, ISSN: 1551-6865. DOI: 10.1145/3588574. [Online]. Available: http://dx.doi.org/10.1145/3588574.

[24] T. Wang and K. Chow, "Noise based deepfake detection via multi-head relative-interaction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 14 548–14 556, Jun. 2023. DOI: 10.1609/aaai.v37i12.26701.

[25] Y. Wang, K. Yu, C. Chen, X. Hu, and S. Peng, "Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 7278–7287.

[26] G. Yang, A. Wei, X. Fang, and J. Zhang, "Fds_2d: Rethinking magnitude-phase features for deepfake detection," *Multimedia Systems*, vol. 29, pp. 1–15, Jun. 2023. DOI: 10.1007/s00530-023-01118-6.

[27] W. Yang et al., "Avoid-df: Audio-visual joint learning for detecting deepfake," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2015–2029, 2023. DOI: 10.1109/TIFS.2023.3262148.

[28] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, and J. Tang, "Istvt: Interpretable spatial-temporal video transformer for deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1335–1348, 2023. DOI: 10.1109/TIFS.2023.3239223.

[29] Y. Chen et al., *Rawbmamba: End-to-end bidirectional state space model for audio deepfake detection*, 2024. arXiv: 2406.06086 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2406.06086.

[30] P. Edwards, J.-C. Nebel, D. Greenhill, and X. Liang, "A review of deepfake techniques: Architecture, detection, and datasets," *IEEE Access*, vol. 12, pp. 154 718–154 742, 2024. DOI: 10.1109/ACCESS.2024.3477257.

[31] M. H. Erol, A. Senocak, J. Feng, and J. S. Chung, *Audio mamba: Bidirectional state space model for audio representation learning*, 2024. arXiv: 2406.03344 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2406.03344.

[32] A. Gu and T. Dao, *Mamba: Linear-time sequence modeling with selective state spaces*, 2024. arXiv: 2312.00752 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2312.00752.

[33] H. Liu, X. Liu, Q. Kong, W. Wang, and M. D. Plumbley, *Learning temporal resolution in spectrogram for audio classification*, 2024. arXiv: 2210.01719 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2210.01719.

[34] W. Liu et al., "Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes," in *Advances in Neural Information Processing Systems*, A. Globerson et al., Eds., vol. 37, Curran Associates, Inc., 2024, pp. 91 131–91 155. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/a5a5b0ff87c59172a13342d428b1e033-Paper-Conference.pdf.

[35] G. Pei et al., *Deepfake generation and detection: A benchmark and survey*, 2024. arXiv: 2403.17881 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2403.17881.

[36] J. Ricker, S. Damm, T. Holz, and A. Fischer, *Towards the detection of diffusion model deepfakes*, 2024. arXiv: 2210.14571 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2210.14571.

[37] R. Shao, T. Wu, L. Nie, and Z. Liu, *Deepfake-adapter: Dual-level adapter for deepfake detection*, 2024. arXiv: 2306.00863 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2306.00863.

[38] Y. Xu, J. Liang, L. Sheng, and X.-Y. Zhang, *Learning spatiotemporal inconsistency via thumbnail layout for face deepfake detection*, 2024. arXiv: 2403.10261 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2403.10261.

[39] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, *Vision mamba: Efficient visual representation learning with bidirectional state space model*, 2024. arXiv: 2401.09417 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2401.09417.

[40] N. A. Chandra et al., "Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024," *arXiv*, 2025. DOI: 10.48550/arxiv.2503.02857. eprint: 2503.02857.

[41] V. Gupta, V. Srivastava, A. Yadav, D. K. Vishwakarma, and N. Kumar, "Freqfacenet: An enhanced transformer architecture with dual-order frequency attention for deepfake detection: Freqfacenet: An enhanced transformer architecture with dual-order frequency attention for deepfake detection," *Applied Intelligence*, vol. 55, no. 7, Feb. 2025, ISSN: 0924-669X. DOI: 10.1007/s10489-024-06168-5. [Online]. Available: https://doi.org/10.1007/s10489-024-06168-5.

[42] M. Javed Bhutto, Z. Zhang, F. Dahri, and T. Kumar, "Enhancing multimodal deepfake detection with local–global feature integration and diffusion models," *Signal, Image and Video Processing*, vol. 19, Mar. 2025. DOI: 10.1007/s11760-025-03970-7.

[43] P. Liu, Q. Tao, and J. T. Zhou, *Evolving from single-modal to multi-modal facial deepfake detection: Progress and challenges*, 2025. arXiv: 2406.06965 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2406.06965.

[44] H.-H. Nguyen-Le, V.-T. Tran, D.-T. Nguyen, and N.-A. Le-Khac, *Passive deepfake detection across multi-modalities: A comprehensive survey*, 2025. arXiv: 2411.17911 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2411.17911.

[45] S. Peng et al., *Wmamba: Wavelet-based mamba for face forgery detection*, 2025. arXiv: 2501.09617 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2501.09617.

[46] X. Xuan, Z. Zhu, W. Zhang, Y.-C. Lin, and T. Kinnunen, *Fake-mamba: Real-time speech deepfake detection using bidirectional mamba as self-attention's alternative*, 2025. eprint: arXiv:2508.09294.