DeepShield: An Explainable AI Framework for Multimodal Deepfake Detection

by

Md. Melon Mia
ID: CSE2202026017

Md. Josim Miah
ID: CSE2002020026

Milton Bardhan Pongkaz
ID: CSE2203027068

Somaiya Binta Hasem
ID: CSE2203027149

Supervised by
Prof. Bulbul Ahamed
Head, Department of CSE

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Science in Computer Science and Engineering



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SONARGAON UNIVERSITY (SU)

September 2026

APPROVAL

The thesis titled " DeepShield: An Explainable AI Framework for Multimodal Deepfake Detection" submitted by Md. Melon Mia (CSE2202026017), Md. Josim Miah (CSE2002020026) Milton Bardhan Pongkaz (CSE2203027068) Somaiya Binta Hasem (CSE2203027149) to the Department of Computer Science and Engineering, Sonargaon University (SU), has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering and approved as to its style and contents.

Board of Examiners

Prof. Bulbul Ahamed **Supervisor**
Professor & Head
Department of Computer Science and Engineering
Sonargaon University (SU)

[Examiner Name and Signature] **Examiner 1**

Department of Computer Science and Engineering
Sonargaon University (SU)

[Examiner Name and Signature] — **Examiner 2**
Department of Computer Science and Engineering
Sonargaon University (SU)

[Examiner Name and Signature] **Examiner 3**
Department of Computer Science and Engineering
Sonargaon University (SU)

DECLARATION

We hereby declare that the work presented in this thesis is the outcome of the investigation performed by us under the supervision of **Prof. Bulbul Ahamed, Head, Department of Computer Science and Engineering, Sonargaon University**, Dhaka, Bangladesh. We reaffirm that no part of this thesis has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned:

(Prof. Bulbul Ahamed)
Supervisor

Signature

Md. Melon Mia
CSE2202026017

Md. Josim Miah
CSE2002020026

Milton Bardhan Pongkaz
CSE2203027068

Somaiya Binta Hasem
CSE2203027149

ABSTRACT

Deepfakes increasingly combine manipulated facial appearance, synthetic or replaced speech, temporal inconsistencies, audio-visual synchronization artifacts, and context-level anomalies. A detector that relies on only one cue can therefore miss manipulations that preserve that cue while altering another. This thesis presents DeepShield, an explainable, multimodal, reliability-aware deepfake forensic framework that integrates Face, Voice, Temporal, Lip-Sync, and Background evidence through mask-aware learned attention fusion. The Face branch combines Swin Transformer, official GenConViT-ED, GenConViT-VAE, and frequency-domain evidence; the Voice branch uses MFCC and Log-Mel features with a CNN-LSTM; the Temporal branch models ordered face sequences using BiLSTM and temporal attention; the Lip-Sync branch analyzes aligned mouth-audio evidence; and the Background branch uses CLIP and DINOv2 representations over original and face-suppressed frames. Validation-only temperature scaling and a locked decision threshold are used before a single frozen internal TEST evaluation. The project-wide eligible primary registry contained 254,228 media items. Earlier leakage-safe and broad visual stages used 177,584 selected samples and 180,206 visual sequences respectively, while the final tightly aligned multimodal canonical experiment used 21,544 A/V cases split into 17,234 TRAIN, 2,159 validation, and 2,151 frozen TEST cases. The final frozen internal TEST achieved ROC-AUC 0.998277, accuracy 0.992562, precision 1.000000, recall 0.992385, F1-score 0.996178, expected calibration error 0.007261, and Brier score 0.005121. The results demonstrate strong performance on the locked aligned evaluation cohort while the thesis deliberately limits claims to this internal experimental setting. DeepShield is positioned as an automated forensic screening framework rather than legal proof, with explicit modality masks, uncertainty-aware decisions, and explainability outputs designed to support human review.

Keywords: DeepShield, Swin Transformer, Generative Convolutional Vision Transformer- Encoder Decoder (GenConViT-ED), Generative Convolutional Vision Transformer- Variable Auto Encoder (GenConViT-VAE), Mel-Frequency Cepstral Coefficients (MFCC), Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM), Bidirectional Long Short-Term Memory (BiLSTM).

ACKNOWLEDGMENT

At the outset, we would like to express our sincere gratitude to the Almighty Allah for granting us the strength, patience, and wisdom to successfully complete this project within the scheduled time. Without His blessings, this accomplishment would not have been possible.

We feel deeply honored and fortunate to work under the kind guidance and supervision of **Prof. Bulbul Ahamed**, Professor and Head, Department of Computer Science and Engineering, Sonargaon University. His continuous support, valuable advice, constructive feedback, and constant encouragement played a crucial role in the successful completion of this project.

We are also sincerely thankful to all the respected teachers of the Department of Computer Science and Engineering, Sonargaon University, for sharing their knowledge, guidance, and inspiration throughout our academic journey, which greatly contributed to our learning and development.

We would like to express our profound gratitude to the Military Institute of Science and Technology (MIST) for providing us with access to their High-Performance Computing (HPC) resources. The availability of the RTX 5090 GPU, 64 GB RAM, and large-scale storage played a vital role in implementing various stages of our thesis, including Swin, official GenConViT, audio preprocessing, and the multimodal pipeline.

Finally, we would like to express our deepest gratitude and heartfelt love to our parents for their unconditional support, encouragement, patience, and endless love. Their motivation and sacrifices inspired us to move forward with confidence and achieve our goals.

LIST OF ABBREVIATIONS

Abbreviation	Full Form
AI	Artificial Intelligence
AUC	Area Under the Curve
AV	Audio-Visual
BiLSTM	Bidirectional Long Short-Term Memory
CLIP	Contrastive Language-Image Pre-training
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DFDC	Deepfake Detection Challenge
DINOv2	DIstillation with NO Labels (Version2)
DL	Deep Learning
ECE	Expected Calibration Error
ED	Encoder Decoder
FN	False Negative
FP	False Positive
GenConViT	Generative Convolutional Vision Transformer
GPU	Graphics Processing Unit
HPC	High-Performance Computing
LSTM	Long Short-Term Memory
MTCNN	Multi-task Cascaded Convolutional Networks
ROC	Receiver Operating Characteristic
ROI	Region of Interest
ROC-AUC	Receiver Operating Characteristic - Area Under Curve
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
VAE	Variable Auto Encoder
VAL	Validation Split
XAI	Explainable Artificial Intelligence

TABLE OF CONTENTS

DECLARATION iii
ABSTRACT iv
ACKNOWLEDGMENT v
LIST OF ABBREVIATIONS vi

CHAPTER 1 - INTRODUCTION 1
1.1 Background 1
1.2 Motivation 2
1.3 Problem Statement 3
1.4 Objectives 3
1.5 Research Questions 4
1.6 Main Contributions 4
1.7 Scope and Delimitations 5
1.8 Significance of the Study 5
1.9 Research Design Principles 6
1.10 Forensic Requirements and Reliability Perspective 6
1.11 Application Scenarios and Intended Users 7
1.12 Terminology and Evidence Levels 7
1.13 Organization of the Thesis 8

CHAPTER 2 - LITERATURE REVIEW AND TECHNICAL BACKGROUND 9
2.1 Deepfake Detection Landscape 9
2.2 Benchmark Datasets and Experimental Protocols 9
2.3 ConvNeXt as a Spatial Representation Backbone 10
2.4 Temporal Modeling with Recurrent Networks 10
2.5 Multimodal Deepfake Detection 11
2.6 DINOv2 and Self-Supervised Visual Features 11
2.7 Frequency-Domain Deepfake Detection 12
2.8 Explainable AI for Digital Forensics 12
2.9 Audio Deepfake and Voice Forensics 12
2.10 Lip-Synchronization as Cross-Modal Evidence 13
2.11 Background and Contextual Evidence 13
2.12 Multimodal Fusion Strategies 14
2.13 Calibration, Thresholding, and Reliability 14
2.14 Human-Centered Explainability in Deepfake Forensics 14
2.15 Research Gap 15

CHAPTER 3 - DATASET, DATA ENGINEERING AND PREPROCESSING 16
3.1 Dataset Strategy 16
3.2 Historical Leakage-Safe Selection 17
3.3 Broad Historical Visual Stage 17
3.4 Final Aligned Multimodal Cohort 18
3.5 Computational Environment 18
3.6 Face Preprocessing with MTCNN 19
3.7 Image and Video Sample Construction 20
3.8 Audio and Multimodal Data Considerations 21
3.9 Reproducibility and Data Governance 21

3.10 Label Semantics and Ground-Truth Contracts 22
3.11 Canonical Case IDs, Provenance, and Traceability 22
3.12 Identity-Safe Split Construction 22
3.13 Modality Availability and Missing-Data Handling 23
3.14 Frame Decoding, Face Quality, and Visual Quality Control 23
3.15 Audio Preparation and Feature Quality Control 23
3.16 Lip-Sync Window Preparation 24
3.17 Background Context Construction 24
3.18 Class Imbalance and Sampling Considerations 25
3.19 Artifact Versioning, Checkpoints, and Audit Trail 25
3.20 Preprocessing Pipeline Summary 25

CHAPTER 4 - DEEPSHIELD SYSTEM ARCHITECTURE 27
4.1 Evolution from Proposal to Final Locked Configuration 27
4.2 Face Detection 27
4.3 Historical ConvNeXt-Tiny Image Branch 28
4.4 Historical Video Feature Sequence 28
4.5 Historical Bidirectional LSTM 29
4.6 Historical Temporal Attention 29
4.7 Final Temporal Preprocessing 29
4.8 Final Audio Feature Extraction 29
4.9 Final Lip-Sync Evidence 30
4.10 Final Background Context Frames 30
4.11 Explainability with Grad-CAM 30
4.12 Final Face Ensemble Architecture 30
4.13 Swin Transformer Representation 31
4.14 GenConViT-ED and GenConViT-VAE Subbranches 31
4.15 Frequency-Domain Evidence Pathway 31
4.16 Voice CNN-LSTM Architecture 32
4.17 Final Temporal Branch Architecture 32
4.18 Lip-Sync Architecture and Cross-Modal Consistency 32
4.19 Background Architecture with CLIP and DINOv2 33
4.20 Common Projection, Mask-Aware Fusion, and Decision Head 33

CHAPTER 5 - PROPOSED DEEPSHIELD METHODOLOGY 34
5.1 End-to-End Methodological Overview 34
5.2 Face Branch 35
5.3 Face Standardization with MTCNN 36
5.4 Deterministic Face-Frame Sampling 36
5.5 Swin Transformer Subbranch Method 36
5.6 GenConViT-ED Subbranch Method 37
5.7 GenConViT-VAE Subbranch Method 37
5.8 Frequency Evidence Extraction 38
5.9 Learned Face-Subbranch Fusion 38
5.10 Voice Branch 38
5.11 Audio Standardization and Feature Extraction 38
5.12 Voice CNN-LSTM Training and Inference 39
5.13 Temporal Branch 40
5.13.1 Ordered Temporal Sequence Construction 40
5.14 Lip-Sync Branch 40

5.14.1 Lip-Sync Alignment and Offset Evaluation 40
5.15 Background Branch 41
5.15.1 CLIP and DINOv2 Context Encoding 41
5.16 Common Projection and Mask-Aware Attention Fusion....42
5.16.1 Modality Masks and Attention Normalization 42
5.17 Calibration, Reliability, and Decision....42
5.17.1 Validation-Only Threshold and Temperature Locking....42
5.18 Explainability and Forensic Reporting 43
5.19 End-to-End Inference Sequence 44
5.20 Methodological Reproducibility and Failure Handling 44
5.21 Software Toolchain and Responsibility of Each Tool 45
5.22 Branch Tensor Interfaces and Feature Storage 46
5.23 Stage-by-Stage Quality Gates 47
5.24 Computational Sequencing and Resource-Aware Execution 48
5.25 Failure Recovery, Resume Logic, and Non-Destructive Processing....48
5.26 Final Forensic Output Contract....49

CHAPTER 6 - TRAINING, VALIDATION AND ABLATION 50
6.1 Experimental Environment....50
6.2 Training and Model-Selection Protocol....50
6.3 Historical ConvNeXt Image Model Training 50
6.4 Historical Visual-Temporal Model Training 50
6.5 Final Aligned Branch Training Scope 51
6.6 Historical Visual-Temporal Model Selection 52
6.7 Final Fusion Temperature Calibration 52
6.8 Evaluation Metrics 53
6.9 Validation Results....53
6.10 Five-Branch Fusion....54
6.11 Historical Confidence Intervals 54
6.12 Explainability Procedure....55
6.13 Standalone Ablation and Generalization Experiments 55
6.14 Internal Validity Controls 55
6.15 Ablation Study 56
6.16 Validation-Only Calibration 57
6.17 Optimization Logic and General Training Strategy....57
6.18 Class Imbalance and Metric-Aware Training Interpretation 57
6.19 Checkpointing, Early Stopping, and Run Management....58
6.20 Test Lock, Calibration Lock, and Experimental Finalization....58
6.21 Statistical Interpretation of the Final Evaluation 59
6.22 Training and Validation Reproducibility Summary 59

CHAPTER 7 - RESULTS, ANALYSIS AND DISCUSSION 60
7.1 Final Canonical Status 60
7.2 Final Frozen Internal TEST Performance....61
7.3 Confusion Matrix Interpretation 62
7.4 Calibration Quality....62
7.5 Branch-Level Interpretation....63
7.6 Historical Spatial-Temporal Design Interpretation....63
7.7 Model Selection and Threshold Behavior....64
7.8 Temperature Calibration 64

7.9 Explainability Output....64
7.10 Auxiliary Frequency and DINOv2 Findings....64
7.11 Comparison with Related Research....65
7.12 Reliability and Forensic Interpretation....65
7.13 Threats to Validity....65
7.14 Internal Validity and Scope of Claims....66
7.15 Forensic Use and Ethical Interpretation....66
7.16 Detailed Error Profile....66
7.17 Metric Interpretation under Strong Class Imbalance....67
7.18 Branch Interaction and Ablation Interpretation....67
7.19 Operational Meaning of the Threshold and Calibrated Score....67
7.20 Answers to the Research Questions....68
7.21 Deployment and Human-Review Workflow....68
7.22 Results Chapter Summary....68

CHAPTER 8 - CONCLUSION, LIMITATIONS AND FUTURE WORK....70
8.1 Conclusion....70
8.2 Limitations....70
8.3 Future Work....71
8.4 Final Research Statement....71
8.5 Research Contributions Revisited....72
8.6 Practical and Academic Implications....72
8.7 Prioritized Future Research Roadmap....72
8.8 Closing Perspective....73

REFERENCES.... 74

LIST OF TABLES

Table 3.1: Primary eligible dataset registry 16
Table 3.2: Historical leakage-safe selection 17
Table 3.3: Final aligned multimodal split 18
Table 3.4: Computing resources used for large-scale implementation 18
Table 5.1: End-to-end DeepShield methodological sequence 35
Table 5.2: Branch input and feature contracts 43
Table 5.3: Final branch roles, evidence, and availability contract 45
Table 6.1: Validation branch results 53
Table 6.2: Five-branch validation ablation 56
Table 7.1: Final frozen TEST metrics 61
Table 7.2: Confusion matrix of final frozen TEST 62
Table 7.3: Calibration and decision configuration 63

LIST OF FIGURES

Figure 1.1:	RealVideo Frame……..	1
Figure 1.2:	Fake Video Frame……..…	1
Figure 3.1:	Dataset scope across DeepShield pipeline stages……..	17
Figure 3.2:	Historical broad visual evidence extraction using Swin Transformer and GenConViT variants over the 180,206-sequence visual development Cohort……	18
Figure 3.3:	Fake video full frame…….	19
Figure 3.4:	MTCNNface crop……..	19
Figure 3.5:	MTCNN Face crop-1…….	20
Figure 3.6:	MTCNN Face crop-2…….	20
Figure 3.7:	MTCNN Face crop-3…….	21
Figure 3.8:	MTCNNFace crop-24……	21
Figure 3.9:	Background full frame……..	24
Figure 3.10	Background face suppressed…….	24
Figure 4.1:	Historical video-level feature sequence and temporal-attention pathway……..	28
Figure 5.1:	DeepShield final canonical multimodal architecture……	34
Figure 5.2:	GPU-based Swin Transformer and GenConViT visual evidence extraction during the historical visual development stage……..	37
Figure 5.3:	Resume-safe audiovisual audio extraction for the aligned DeepShield cohort…….	39
Figure 5.4:	Figure 5.4: Voice-branch contract verification showing MFCC and Log-Mel feature dimensions and locked CNN-LSTM compatibility…	39
Figure 5.5 :	Lip sync integration preflight……	41
Figure 5.6:	Reproducible experimental workflow……..	43
Figure 5.7:	Canonical visual-crop recovery and branch-contract verification used in the DeepShield quality-control pipeline……	47
Figure 6.1:	Completion audit of historical Swin and GenConViT visual evidence extraction across the development splits…….	51
Figure 6.2:	Frozen Voice and Lip-Sync evidence integration over the final aligned audiovisual cohort……	52
Figure 6.3:	Best validation AUC by branch/fusion……..	54
Figure 6.4:	Five-branch validation ablation…….	56
Figure 6.5 :	Final experiment-lock audit confirming calibration, frozen TEST, XAI, forensic-report, and canonical pipeline status……..	58
Figure 7.1:	DeepShield canonical final-pass verification for the 2,151-case frozen TEST evaluation……	60
Figure 7.2**:**	Final frozen internal TEST performance……..	61
Figure 7.3:	Read-only verification of the final frozen TEST cohort and locked DeepShield performance metrics……..	62

CHAPTER 1

INTRODUCTION

1.1 Background

Deepfake technology refers to the use of machine learning and generative models to synthesize or manipulate visual and auditory media so that the resulting content appears authentic. Early systems were often characterized by obvious blending boundaries, unstable lighting, or temporal flicker, but rapid improvements in generative adversarial networks, autoencoders, diffusion models, and Transformer-based architectures have reduced many of these artifacts. Contemporary surveys describe deepfake detection as a continually evolving contest between generation and forensic analysis, where performance on a familiar benchmark is not equivalent to robustness in real-world conditions [1], [2]. The problem is particularly important in social media, digital evidence, identity verification, journalism, financial security, and online communication, because manipulated media can be used to imitate a person or alter the meaning of an event.



Fig: 1.1 Real Video Frame



Fig: 1.2 Fake Video Frame

Figure 1.1 and 1.2 : Representative REAL and FAKE video frames from the DeepShield evaluation pipeline.

Deepfake detection is therefore not only a pattern-recognition problem; it is also a problem of evidence quality, generalization, and trust. A detector may produce a high score because it has learned compression traces, acquisition differences, or dataset-specific artifacts rather than manipulation evidence. Benchmark projects such as DFDC, FaceForensics++, Celeb-DF++, and DeepfakeBench were developed in part to make systematic comparison possible and to expose the gap between in-dataset and cross-dataset behavior [3]-[6]. These resources have pushed the field toward stronger experimental protocols, but they also show

that dataset design, split policy, source identity, and preprocessing consistency are as important as the classifier architecture itself.

DeepShield was developed in this context. The work began with a broad multimodal research proposal covering face, voice, temporal consistency, contextual information, frequency-domain clues, and explainability. During implementation, the pipeline was repeatedly verified against data leakage, preprocessing inconsistencies, compute constraints, and model behavior. The final canonical system integrates Face, Voice, Temporal, Lip-Sync, and Background evidence through mask-aware learned attention fusion. The Face branch combines Swin Transformer, official GenConViT-ED, GenConViT-VAE, and frequency-domain evidence; the Background branch uses CLIP and DINOv2 representations; and the historical ConvNeXt-BiLSTM pathway remains documented as an earlier visual-temporal stage. This evolution is treated as part of the research process rather than hidden. Auxiliary standalone experiments with DINOv2 and frequency-domain methods remain relevant to ablation and generalization analysis while their representations also contribute to the final Background and Face branches respectively.

1.2 Motivation

The principal motivation of this research is the practical difficulty of identifying manipulated media when visual quality is high and when the media originate from heterogeneous sources. Human observers may notice severe artifacts, but subtle face swaps, reenactments, and identity manipulations can preserve realistic texture and motion. Single-frame detectors can learn useful spatial evidence, yet a video contains additional information: identity should remain stable, motion should be coherent, and local frame-level evidence should agree over time. Research on temporal identity consistency and spatiotemporal modeling demonstrates that sequence information can reveal anomalies that isolated frames may not expose [8]-[12].

A second motivation is the need for transparent decision making. In digital forensics, a binary label without supporting evidence can be difficult to interpret. Explainability does not transform a machine-learning output into legal proof, but it can reveal whether the detector is focusing on plausible facial regions or irrelevant backgrounds. Explainable digital-forensics research emphasizes the importance of human-interpretable evidence, especially when automated systems may influence investigative decisions [15], [16], [20], [21]. DeepShield therefore integrates Grad-CAM as part of the reporting workflow and separates model probability from the stronger notion of certainty.

A third motivation is resource-aware reproducibility. The research team had access to MIST high-performance computing resources with a 32 GB GPU, 64 GB RAM, a 2 TB SSD, a 16-core CPU, 64 MB L3 cache, and a stated clock speed of 4.5 GHz. This enabled large-scale preprocessing and training, but compute availability alone does not guarantee a reliable experiment. The final pipeline was designed to use deterministic frame

selection, versioned data processing, validation-based model selection, temperature calibration, locked testing, and resumable workflows so that a long training run would not need to be redesigned after expensive computation had already begun.

1.3 Problem Statement

Deepfake detectors face several connected problems. First, the distribution of manipulations changes over time. A detector trained on one manipulation family may fail when the generator, codec, camera, demographic distribution, or platform compression changes. Second, video-level decisions require a principled way to combine evidence across frames. Naive averaging may ignore the fact that some frames are more informative than others, while selecting only a single frame can miss temporal anomalies. Third, highly confident softmax scores can be misinterpreted as calibrated probabilities even when the model is overconfident. Finally, explainability and data governance are often treated as optional additions rather than core elements of the forensic pipeline.

The specific problem addressed by this thesis is therefore how to construct a practical, reproducible deepfake video detector that combines strong spatial representation with temporal reasoning, validates the decision process on a large locked internal test population, reports uncertainty more responsibly through calibration, and provides visual explanations. The broader DeepShield concept is audio-visual and multi-person in scope; earlier locked quantitative evidence concerned the face/video pathway, while the final canonical quantitative evidence in this thesis concerns the five-branch multimodal fusion. Rather than inventing missing branch metrics, the thesis explicitly distinguishes between historical components, individually evaluated branches, and the final implemented-and-evaluated fusion.

1.4 Objectives

The primary aim is to develop and evaluate a reliability-aware explainable deepfake detection framework whose final locked detector integrates Face, Voice, Temporal, Lip-Sync, and Background evidence through mask-aware learned attention fusion.

- To create a reproducible data pipeline capable of handling heterogeneous public and custom media while preserving source-group and identity information.
- To standardize face-centric preprocessing with MTCNN and deterministic frame selection.
- To train and evaluate visual face representations, including the historical ConvNeXt-Tiny pathway and the final Swin, GenConViT-ED, GenConViT-VAE, and frequency-domain Face subbranches.
- To construct ordered face sequences and model temporal evidence using BiLSTM and temporal attention within the final multimodal framework.
- To select branch and fusion checkpoints using validation evidence and to keep the internal TEST set frozen from model selection.

- To apply temperature scaling and a validation-derived decision threshold for more disciplined probability reporting.
- To generate Grad-CAM explanations for visual inspection of influential facial regions.
- To evaluate frequency-domain and DINOv2-based representations both within their final Face/Background roles and through standalone ablation/generalization analysis.
- To report accuracy, precision, recall, F1-score, ROC-AUC, calibration metrics, confusion-matrix counts, and the locked TEST population for the final canonical experiment.
- To document limitations related to cross-dataset generalization, multi-person completeness, and real-world deployment.

1.5 Research Questions

1. Can the final Face branch, combining Swin Transformer, GenConViT-ED, GenConViT-VAE, and frequency-domain evidence, provide sufficiently discriminative facial representations for deepfake classification within the aligned cohort?
2. Does sequence modeling with BiLSTM and temporal attention provide useful complementary evidence when combined with the other DeepShield modalities?
3. Can validation-only model selection, thresholding, and temperature scaling reduce the risk of test-set tuning and improve interpretability of model outputs?
4. What level of performance does the final locked five-branch multimodal fusion achieve on the frozen internal TEST population, and how should its discrimination, calibration, and error counts be interpreted?
5. How do the Face, Voice, Temporal, Lip-Sync, and Background branches contribute individually and under ablation to the final multimodal decision?

1.6 Main Contributions

The first part is a practical research workflow that treats data handling and clean data splits as core parts of the detector. The project grew from a simple draft into a large data processing and training system. This shift needed steady sample IDs, exact processing steps, error management, and a strict wall between testing and validation choices. The final thesis keeps this engineering proof instead of showing the model as a simple, standalone network.

The second part is the multimodal design developed from the historical visual and time-based pathway. ConvNeXt updated the earlier image-processing stage using ideas from modern Transformer models [7], while the final Face branch combines Swin Transformer, GenConViT-ED, GenConViT-VAE, and frequency-domain evidence. Temporal BiLSTM attention, Voice CNN-LSTM evidence, Lip-Sync evidence, and CLIP/DINOv2 Background evidence are projected into a common space and combined by mask-aware learned attention. This matches other research showing that temporal and multimodal clues can complement strong spatial evidence [8]-[16].

The third part is a reliable and careful evaluation method. The earlier visual-temporal pathway used a validation-derived threshold of 0.531 and a temperature value of 1.397722, while the final five-branch multimodal fusion uses the locked decision threshold 0.89 and validation-fitted temperature 1.1680239439. The final frozen internal TEST contains 2,151 aligned cases and reports ROC-AUC, accuracy, precision, recall, F1-score, ECE, Brier score, and exact confusion-matrix counts. This matters because high scores can look overly perfect unless the test group size, class balance, calibration, and limitations are explained clearly.

The fourth part is about clear explanations and honest reporting. Grad-CAM is used for face-region inspection, while modality and subbranch contributions are retained in the final forensic reporting workflow. The thesis also records how the first multimodal plan evolved through historical visual-temporal experiments into the completed five-branch fusion. This honesty prevents historical branch results from being confused with the final canonical multimodal result, and it leaves a clear path for future external validation.

1.7 Scope and Delimitations

The quantitative core of the thesis is binary real-versus-fake classification over an aligned multimodal cohort, with the final reported metrics belonging to the five-branch fusion. MTCNN is retained for face preprocessing. The final Face branch combines Swin Transformer, GenConViT-ED, GenConViT-VAE, and frequency-domain evidence; the Voice branch uses MFCC and Log-Mel CNN-LSTM modeling; the Temporal branch uses ordered face evidence with BiLSTM attention; the Lip-Sync branch uses aligned mouth-audio evidence; and the Background branch uses CLIP and DINOv2 context representations. Model and fusion selection are based on validation evidence, and the frozen TEST population is not used to tune the architecture or threshold. The broader DeepShield name reflects a framework intended for audio-visual, multi-person, reliability-aware forensic analysis. The supplied final result evidence provides a final multimodal fusion score and branch-level validation/ablation evidence, but it does not provide a separate final multi-person tracking metric. Multi-person tracking is therefore discussed as an architectural extension and limitation rather than assigned an invented quantitative result. The study also does not claim that the historical visual-temporal ROC-AUC of 99.9815% or the final multimodal ROC-AUC of 0.998277 guarantees equivalent performance on unseen public platforms or future deepfake generators.

1.8 Significance of the Study

The significance of DeepShield lies in treating deepfake detection as a forensic evidence-integration problem rather than as a single image-classification task. Contemporary manipulated media can preserve realistic facial texture while altering speech, temporal behavior, synchronization, or surrounding context. A detector that depends on one evidence source can therefore fail when that source remains visually convincing. The five-branch design used in the final aligned experiment is important because it allows

facial, vocal, temporal, lip-synchronization, and background cues to be analyzed independently and then combined under an explicit fusion rule. This approach follows the broader direction of multimodal deepfake research, where complementary evidence can reduce dependence on any one artifact family [13]-[16].

The study is also significant from a research-engineering perspective. Large deepfake projects frequently fail because of data leakage, inconsistent preprocessing, ambiguous audio labels, changing dataset versions, or repeated tuning on the test set. DeepShield addresses these risks through a canonical registry, identity-aware splits, deterministic processing, branch-specific availability masks, validation-only calibration, and a single frozen final TEST evaluation. The contribution is therefore not limited to the final accuracy value; it includes a traceable experimental process that makes the reported result more interpretable and defensible.

A further significance is the emphasis on reliability and explainability. Deepfake predictions may influence moderation, journalism, identity verification, or forensic screening, but a high model score is not equivalent to legal proof. The thesis therefore separates discrimination, calibration, and explanation. ROC-AUC measures ranking quality, threshold-based metrics describe operational decisions, ECE and Brier score summarize calibration behavior, and XAI outputs provide supporting visual evidence rather than certainty. This layered interpretation is aligned with reliability-oriented and explainable digital-forensics research [20]-[22].

1.9 Research Design Principles

The final research design was guided by five principles: evidence diversity, leakage prevention, reproducibility, conservative claim scope, and auditability. Evidence diversity motivated the use of multiple branches so that the model could observe different forensic signals. Leakage prevention required source and identity information to be preserved when constructing splits. Reproducibility required deterministic manifests, fixed preprocessing rules, versioned artifacts, and locked checkpoints. Conservative claim scope required the thesis to report only metrics produced by the documented final experiment. Auditability required that intermediate artifacts and final decisions remain traceable to a case identifier and branch-level evidence.

These principles influenced model engineering as much as model selection. For example, the system does not automatically interpret an unavailable modality as evidence of manipulation. Instead, availability masks identify which branches are present for each case. Likewise, the final TEST set is not used to choose checkpoints, fusion weights, calibration temperature, or decision threshold. The system design therefore treats methodological discipline as part of the architecture rather than as a separate reporting step.

1.10 Forensic Requirements and Reliability Perspective

A forensic screening system must satisfy requirements that are broader than raw classification performance. It should make clear what evidence was available, what preprocessing was applied, what model version produced the decision, and how uncertain the decision may be. DeepShield addresses these requirements through explicit modality masks, case-level identifiers, calibration, and explainability outputs. In practical use, the framework is intended to support an analyst by prioritizing suspicious media and exposing the evidence that contributed to the model output.

Reliability is particularly important because deepfake inputs can be degraded by compression, low resolution, occlusion, background noise, desynchronization, or missing audio. A model may also encounter generators that are absent from the training distribution. For these reasons, the thesis deliberately avoids describing the internal score as universal accuracy. The final frozen result demonstrates performance on the aligned internal cohort, while reliability analysis, limitations, and proposed external evaluation define what remains necessary before deployment in a broader operational environment [22].

1.11 Application Scenarios and Intended Users

The DeepShield framework is relevant to several application scenarios. In social-media moderation, it can be used as a screening layer that flags high-risk videos for human review. In digital journalism, the model can contribute a technical signal during source verification, especially when the authenticity of a circulated video is uncertain. In identity-protection settings, branch-level evidence can support investigation of impersonation, face replacement, synthetic speech, or mismatched audiovisual content. In an academic or laboratory environment, the modular architecture also provides a platform for controlled ablation studies in which individual branches can be removed or replaced without redesigning the full pipeline.

The intended user is therefore not an end user who should accept a binary label without context. The more appropriate users are researchers, analysts, moderators, or forensic practitioners who can interpret the output alongside provenance, metadata, source history, and independent verification. This human-in-the-loop positioning reduces the risk that a probabilistic model output is treated as definitive proof.

1.12 Terminology and Evidence Levels

Throughout this thesis, the terms sample, case, frame, crop, sequence, branch, and fusion have distinct meanings. A media item is an original image, audio file, or video registered in the project-wide dataset. A canonical case is an aligned unit used by the final multimodal experiment. A frame is an image decoded from a video, while a face crop is a standardized facial region extracted from that frame. A sequence is an ordered set of frame-level or feature-level observations. A branch is a model pathway responsible for

one evidence family, and fusion is the learned process that combines available branch representations into a case-level decision.

The thesis also distinguishes historical evidence from final canonical evidence. Historical stages document important large-scale experiments, preprocessing runs, and visual-model development, but they do not define the final multimodal score. Final canonical evidence refers specifically to the 21,544 aligned A/V cohort and its locked 2,151-case TEST split. Making this distinction explicit prevents earlier 14,991-video visual-temporal experiments from being confused with the final five-branch evaluation.

1.13 Organization of the Thesis

The remainder of the thesis is organized to move from scientific context to experimental evidence. Chapter 2 reviews deepfake detection literature and the technical ideas that motivate spatial, temporal, audio, contextual, multimodal, explainable, and reliability-aware analysis. Chapter 3 documents the dataset ecosystem, canonical data engineering, preprocessing, split construction, and governance. Chapter 4 describes the architecture and explains the role of the principal models and evidence branches. Chapter 5 presents the end-to-end methodology in operational sequence, including feature extraction, branch inference, fusion, calibration, and reporting.

Chapter 6 describes training, validation, ablation, model selection, and experimental controls. Chapter 7 presents the final frozen TEST results and interprets discrimination, calibration, confusion counts, branch behavior, validity threats, and operational meaning. Chapter 8 concludes the thesis, summarizes limitations, identifies future research priorities, and restates the contribution at the level justified by the evidence. References follow the academic chapters; defense-preparation questions and answers are intentionally excluded from the thesis body.

CHAPTER 2

LITERATURE REVIEW AND TECHNICAL BACKGROUND

2.1 Deepfake Detection Landscape

Deepfake detection research has developed from artifact-specific classifiers into a broad field covering spatial, temporal, frequency, biological, semantic, and multimodal evidence. Surveys emphasize that early detectors often relied on visible inconsistencies such as blending, color mismatch, or warping, whereas newer systems increasingly use deep representation learning and Transformer-inspired architectures [1], [2]. The central challenge is that deepfake generation evolves rapidly, so a detector can become obsolete when it depends on artifacts associated with a particular generation process.

The literature also distinguishes between in-dataset performance and generalization. In-dataset evaluation measures how well a detector performs on data sampled from the same benchmark distribution used during training and validation. Cross-dataset evaluation is more difficult because it introduces new identities, codecs, compression profiles, image statistics, and generation methods. DeepfakeBench was created to make evaluation more reproducible across datasets and detector families [6]. This broader benchmarking perspective is important when interpreting DeepShield: its locked internal metrics are strong, but they must be separated from claims about universal generalization.

Another trend is movement from frame classification to video reasoning. A manipulated face may be locally plausible in an individual frame while revealing subtle identity changes, motion irregularities, or inconsistent appearance across time. Temporal methods therefore attempt to model relationships among frames rather than treating all observations as independent. The final DeepShield video branch follows this direction by using a sequence model with attention.

2.2 Benchmark Datasets and Experimental Protocols

DFDC introduced large-scale manipulated video data and helped establish deepfake detection as a benchmark problem [3]. FaceForensics++ provides a controlled framework for studying several manipulation families and compression settings [4]. Celeb-DF++ extends the challenge toward more recent and more generalizable forensic evaluation [5]. These resources complement one another because no single dataset captures all combinations of identity, recording condition, manipulation method, and post-processing.

Benchmark diversity creates a methodological difficulty: directly merging datasets can cause a model to learn dataset identity rather than manipulation evidence. For example, one dataset may use different resolution, frame rate, face cropping, or compression than another. A robust pipeline therefore needs dataset-aware sampling and leakage-safe split construction while preventing dataset name from becoming a model feature. DeepShield

follows this principle at the data-engineering level by maintaining canonical source identifiers and by separating model selection from final testing.

The historical broad visual TEST population contained 14,991 videos. This number is large enough that its historical point estimates could be accompanied by relatively narrow confidence intervals, but sample size alone does not guarantee independence. If multiple derived clips share a source identity or generator, effective statistical diversity may be lower than the raw clip count. The thesis therefore treats those historical confidence intervals as uncertainty around that measured visual population rather than as a claim about the final 2,151-case multimodal TEST cohort or every possible deployment domain.

2.3 ConvNeXt as a Spatial Representation Backbone

ConvNeXt was proposed as a modernized convolutional architecture that incorporates design lessons associated with the Transformer era while preserving the inductive structure of convolution [7]. The architecture revisits stage ratios, patch-like stem design, large-kernel depthwise convolution, normalization choices, and inverted bottleneck structure. The result is a convolutional network that can compete with contemporary hierarchical Transformers on image-recognition tasks.

For deepfake detection, a strong spatial encoder must capture local texture irregularities, boundary artifacts, subtle changes around facial components, and broader semantic appearance. ConvNeXt-Tiny offers a practical balance between representational capacity and computational cost. In DeepShield, it is used not only as the image classifier but also as a feature generator for the video branch. This shared representation reduces the conceptual gap between single-frame and sequence-level analysis: each sampled face is encoded consistently, and the temporal model reasons over a sequence of features in the same embedding space.

Using a pretrained image backbone also supports transfer learning. Rather than learning all low-level filters from scratch, the model can adapt existing visual representations to the real-versus-fake task. This is especially valuable when the training corpus contains heterogeneous sources because the early representation is not tied exclusively to one deepfake dataset. Nevertheless, transfer learning does not automatically solve domain shift, and the final evaluation must still distinguish internal and external generalization.

2.4 Temporal Modeling with Recurrent Networks

Long short-term memory networks were designed to model sequential dependencies while reducing the vanishing-gradient problems of conventional recurrent networks. A bidirectional LSTM processes a sequence in both forward and backward directions, allowing each time step to be represented using context from earlier and later observations. Deepfake video studies have combined convolutional encoders with

LSTM or hybrid Transformer structures to capture temporal evidence beyond individual images [8], [9].

Temporal identity consistency has become a specific deepfake research direction. Identity Consistency Transformer and TI2Net analyze whether a person remains semantically and visually consistent across time [10], [11]. Spatiotemporal dropout Transformer methods likewise show that temporal modeling can help identify manipulated videos [12]. These works support the design principle used in DeepShield: spatial embeddings are treated as a time series and then summarized by a learned sequence model.

The decision to use four deterministic frames per video represents a practical trade-off. More frames may provide richer motion information but increase storage, I/O, GPU memory, and training time. Very sparse sampling risks missing short manipulations. The chosen design emphasizes reproducibility and computational efficiency. Because the exact frame positions are deterministic, repeated runs use the same temporal evidence rather than introducing sampling noise between experiments.

2.5 Multimodal Deepfake Detection

Multimodal deepfake detection investigates relationships among visual, audio, and sometimes linguistic information. AVTENet uses audio-visual Transformer-based ensemble reasoning, while M2TR explores multi-modal and multi-scale representations [13], [14]. Recent explainable multimodal research further argues that cross-modal feature analysis and dynamic weighting can improve both detection and interpretability [15], [16]. The underlying rationale is that a manipulation may preserve one modality while disrupting another; for example, a visually convincing face may not match the acoustic or temporal structure of the speech.

The initial DeepShield proposal adopted this multimodal philosophy and considered facial, voice, temporal, frequency, and contextual branches. The final locked result set now reports the completed five-branch Face, Voice, Temporal, Lip-Sync, and Background fusion rather than only the earlier visual-temporal pathway. This distinction is important. Literature motivates multimodal fusion, while the reported DeepShield validation, ablation, calibration, and frozen TEST evidence supports the specific final multimodal implementation documented in Chapters 5-7. The final thesis therefore uses multimodal studies to motivate the framework while reserving numerical claims for components and fusion results supported by the supplied result evidence.

2.6 DINOv2 and Self-Supervised Visual Features

DINOv2 demonstrates that strong visual representations can be learned without conventional supervised class labels and transferred to a range of downstream tasks [17]. Such representations are attractive in forensic research because they may encode semantic and structural information that differs from a classifier trained directly on dataset labels. DeepShield evaluated DINOv2 as an auxiliary ablation and generalization direction.

The final status records DINOv2 as a representation used in the canonical Background branch and also as an independently evaluated background/ablation representation. This result is methodologically useful: an ablation can inform model choice even when a standalone variant is not the complete system. Reporting the standalone experiment as auxiliary prevents a common thesis error in which every explored model is presented as if it alone produced the final score.

2.7 Frequency-Domain Deepfake Detection

Synthetic image generation can introduce spectral patterns that are difficult to observe directly in pixel space. Frequency-aware deepfake detection therefore studies transformed representations to expose generation traces and improve generalization [18]. Earlier work on frequency-aware clues similarly showed that spatial artifacts are not the only useful evidence in face forgery detection [19].

Frequency analysis is part of the broader DeepShield research design and contributes frequency-domain evidence within the final Face branch while also being evaluated as a standalone branch for ablation. This means its role in the thesis is both component-level and comparative. It supports discussion of alternative evidence spaces and future robustness work, but the historical 99.3730% visual-temporal locked accuracy is not attributed to frequency fusion and is not the final multimodal score.

2.8 Explainable AI for Digital Forensics

Explainability is especially important in forensic contexts because stakeholders may need to inspect the basis of an automated decision. Grad-CAM produces a localization map by relating gradients of a target prediction to feature maps in a convolutional network. Deepfake-specific explainability studies have used Grad-CAM and related techniques to show where a model concentrates its attention [20]. Broader digital-forensics research argues that interpretability can reduce distrust and support more responsible use of AI evidence [21].

DeepShield implements Grad-CAM for the visual model. The output is interpreted as an influence visualization, not as proof that the highlighted pixels were manipulated. A convincing heatmap can still reflect a spurious feature, and a diffuse heatmap does not necessarily invalidate a prediction. Accordingly, explainability is treated as a supporting evidence layer that should be inspected alongside the prediction, model reliability, and input quality.

2.9 Audio Deepfake and Voice Forensics

Audio deepfakes include text-to-speech synthesis, voice conversion, speaker cloning, and editing that changes linguistic content while attempting to preserve a natural acoustic signal. From a forensic perspective, audio can contain clues that are independent of facial appearance. Spectral envelopes, prosody, pitch behavior, phase characteristics, vocoder

artifacts, and temporal dynamics may differ between natural and synthetic speech. In a multimodal video, the audio track must therefore be labeled according to its own authenticity contract rather than inheriting the visual label automatically.

DeepShield represents speech using MFCC and Log-Mel features before sequence classification. MFCCs compress the spectral envelope into a compact representation that has long been useful in speech processing, while Log-Mel features preserve a richer time-frequency map. A CNN-LSTM architecture can learn local spectral patterns with convolution and longer temporal dependencies with recurrent layers. In the final framework, voice evidence is treated as a separate branch so that genuine speech paired with a manipulated face does not become mislabeled training data.

2.10 Lip-Synchronization as Cross-Modal Evidence

Lip-synchronization analysis examines whether visible mouth motion is temporally compatible with the acoustic speech signal. A deepfake may contain a realistic face and a realistic voice yet still reveal inconsistency in the timing of phonetic events and mouth shapes. This makes lip-sync evidence conceptually different from both visual deepfake classification and voice-deepfake classification: it evaluates a relationship between modalities rather than the authenticity of either modality in isolation.

A robust lip-sync pipeline requires aligned temporal windows, reliable face or mouth localization, and a consistent audio timeline. The final DeepShield manifest records window counts, offsets, and availability so that synchronization evidence can be masked when valid windows do not exist. This design avoids forcing the fusion model to interpret missing synchronization data as evidence of FAKE. The branch is therefore best understood as a complementary cross-modal consistency signal whose usefulness depends on speech activity, face visibility, and alignment quality.

2.11 Background and Contextual Evidence

Deepfake research often emphasizes the manipulated face, but the surrounding scene can also carry useful information. Full-frame context may reveal blending inconsistencies, camera or compression patterns, semantic mismatches, altered body regions, or correlations between foreground manipulation and background appearance. At the same time, background evidence creates a risk of dataset shortcuts: a classifier may learn source-specific scenery, borders, resolution, or capture conditions instead of manipulation evidence.

DeepShield therefore separates contextual analysis from face analysis and includes both original full frames and face-suppressed variants. CLIP offers semantically rich visual representations learned through image-text supervision, whereas DINOv2 provides strong self-supervised visual features [17]. Comparing full-background and face-suppressed variants helps determine whether contextual representations retain discriminative information after the face is removed. Such ablations are important because strong

context performance should be interpreted carefully rather than automatically treated as causal manipulation evidence.

2.12 Multimodal Fusion Strategies

Multimodal fusion can be performed at several levels. Early fusion combines raw or low-level features before specialized processing, late fusion combines independent prediction scores, and intermediate fusion combines learned branch embeddings. DeepShield follows an intermediate learned-fusion strategy because the branches have different input types and different internal architectures. Each branch produces a representation that can be projected into a common space before case-level attention is computed.

A central challenge is that modalities are not always simultaneously available. Some videos may lack valid speech, some frames may not contain a detectable face, and some lip-sync windows may be unreliable. A naive fusion model can learn misleading patterns if missing values are replaced by a fixed FAKE score. Mask-aware attention addresses this problem by excluding unavailable branches from the normalization step. The final representation therefore depends on the evidence actually present for the case, which is more defensible than treating absence itself as manipulation.

2.13 Calibration, Thresholding, and Reliability

Classification scores serve two different purposes: ranking and decision making. ROC-AUC evaluates whether positive and negative cases are ranked correctly across many thresholds, while accuracy, precision, recall, and F1 depend on a specific operating threshold. A model can have excellent ROC-AUC but still produce probabilities that are too extreme or poorly aligned with observed outcomes. Calibration is therefore a separate reliability problem.

Temperature scaling is a post-hoc calibration method that rescales logits using a scalar fitted on validation data. The class ranking is preserved, but probability magnitudes can become better aligned with the observed distribution. DeepShield fits the final temperature only on validation data and locks the decision threshold before frozen TEST evaluation. The final report includes ECE and Brier score in addition to discrimination metrics, which enables the thesis to discuss whether high predictive performance is accompanied by useful probability behavior [22].

2.14 Human-Centered Explainability in Deepfake Forensics

Explainability in deepfake forensics should answer a narrower question than causal proof: what evidence influenced the model output, and is that evidence plausible enough to support further review? For face-based convolutional pathways, Grad-CAM can highlight spatial regions associated with a prediction. For multimodal systems, branch- or subbranch-attention weights can show how much relative influence each available

evidence source received during fusion. These forms of explanation operate at different levels and should not be interpreted as interchangeable.

Human-centered use requires caution. A visually convincing heatmap can still reflect a spurious correlation, and a high attention weight does not prove that a modality contains manipulation. DeepShield therefore treats explanations as audit aids rather than ground-truth localization. This position is consistent with digital-forensics literature that emphasizes interpretability while warning against replacing expert reasoning with an opaque automated decision [20], [21].

2.15 Research Gap

The literature demonstrates strong progress in spatial detection, temporal identity analysis, multimodal fusion, explainability, and reliability, yet practical student research faces an additional gap between ambitious architectures and reproducible implementation. A framework can contain many sophisticated components but still produce weak scientific evidence if the dataset split leaks identities, the preprocessing differs between training and inference, or the final test set influences model selection.

DeepShield addresses this implementation gap by emphasizing a locked experiment. The final model is not the largest combination of all proposed branches. Instead, it is the configuration that completed training, validation, calibration, and large-scale internal testing under a reproducible pipeline. The thesis therefore prioritizes evidence quality over architectural complexity and treats unvalidated branches as future work rather than as completed contributions.

CHAPTER 3

DATASET, DATA ENGINEERING AND PREPROCESSING

3.1 Dataset Strategy

DeepShield was designed around a heterogeneous dataset ecosystem rather than a single benchmark. The project registry included FaceForensics++, FakeAVCeleb, DFDC, DeeperForensics, OpenForensics, Celeb-DF++, DFD, LAV-DF and a custom personal mobile-phone collection Self_Dataset. These datasets differ in media type, manipulation process, scale, compression, identity distribution, and annotation structure. A unified training system therefore cannot safely assume that all folders encode labels in the same way.

The data-engineering strategy separates raw datasets from manifests and processed derivatives. Raw files are treated as immutable. Dataset-specific adapters convert original structures into a canonical registry containing sample identity, source group, media type, authenticity label, visual label, audio label, manipulation metadata when available, and split status. This approach is more robust than physically merging all real files into one folder and all fake files into another because it preserves provenance and enables dataset-aware analysis.

Public datasets provide scale and manipulation diversity, while custom mobile-phone data provide local recording conditions, different people, voices, locations, camera positions, and real-world acquisition variation. The personal collection was conceived as complementary evidence rather than a replacement for public benchmarks. In particular, local mobile recordings can reveal whether the preprocessing and classifier are sensitive to backgrounds, codecs, or camera characteristics that differ from academic datasets.

Dataset	Eligible Media
Celeb-DF-v2	6,529
DFD	3,068
DFDC	19,909
DeeperForensics-1.0	59,475
FaceForensics++	6,000
FakeAVCeleb	21,544
LAV-DF	136,304
Self Dataset	1,399
TOTAL	254,228

Table 3.1: Primary eligible dataset registry

3.2 Historical Leakage-Safe Selection

An earlier leakage-safe selection stage retained 177,584 samples from the broad registry. Identity grouping and dataset-aware split logic were used to reduce cross-split identity leakage. This stage is historically important because it documents the broader data preparation effort, but it is not identical to the final aligned multimodal experiment.

Split	Historical Leakage-Safe Samples
TRAIN	125,886
VAL	26,612
TEST	25,086
TOTAL	177,584

Table 3.2: Historical leakage-safe selection

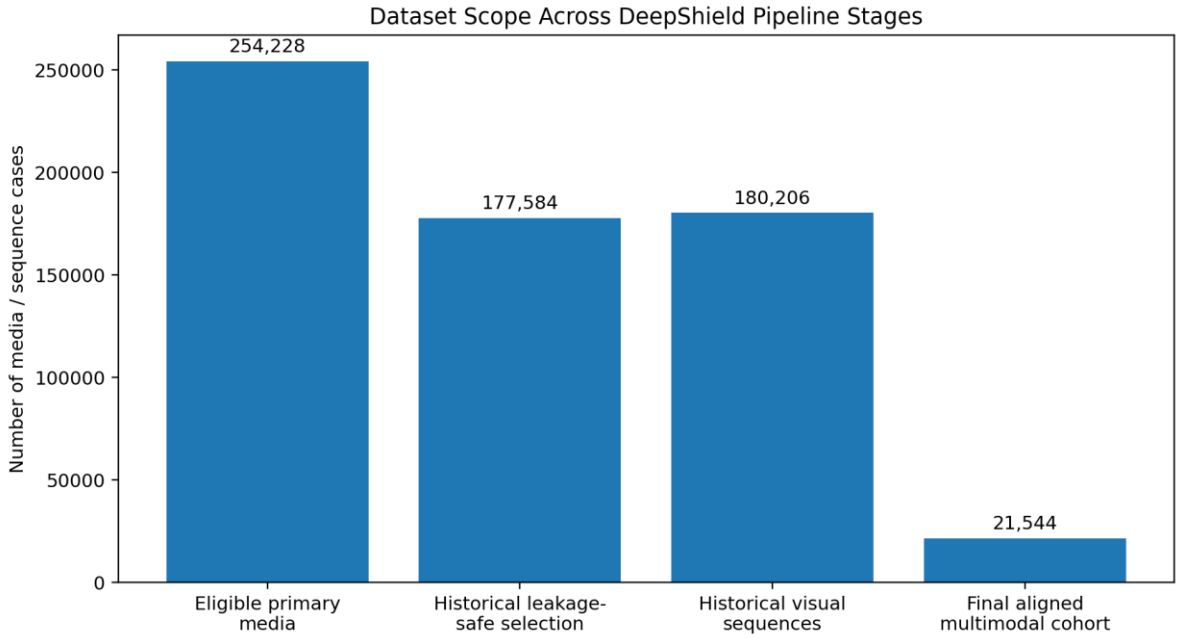


Figure 3.1: Dataset scope across DeepShield pipeline stages

3.3 Broad Historical Visual Stage

A later historical visual manifest contained 180,206 visual sequences and 716,253 usable frame rows. The split included 150,223 TRAIN sequences, 14,992 validation sequences, and 14,991 TEST sequences. Up to four deterministic face frames were stored per sequence. Historical Stage08A1 evidence extraction confirmed that Swin, GenConViT-ED, and GenConViT-VAE embeddings were produced across this broader visual corpus.

These values show large earlier visual experimentation and must not be silently replaced by the smaller final aligned cohort.

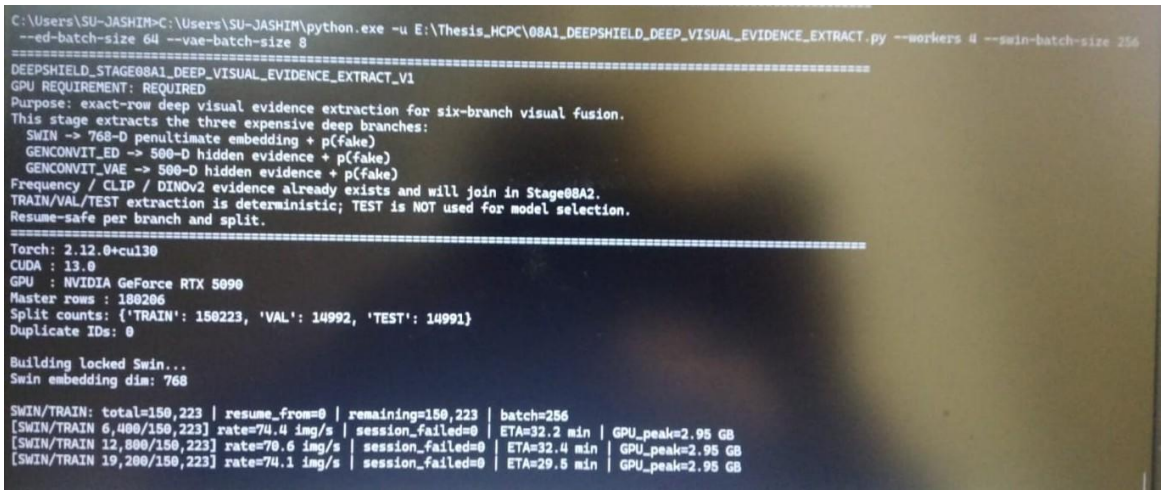


Figure 3.2: Historical broad visual evidence extraction using Swin Transformer and GenConViT variants over the 180,206-sequence visual development cohort.

3.4 Final Aligned Multimodal Cohort

The final canonical multimodal experiment used 21,544 FakeAVCeleb-aligned A/V cases. The identity-safe split was 17,234 TRAIN, 2,159 VAL, and 2,151 frozen TEST. All 21,544 cases were retained at the master level, while each branch used its own availability contract rather than dropping the entire case when one modality was missing.

Split	Final Aligned Cases
TRAIN	17,234
VAL	2,159
TEST	2,151
TOTAL	21,544

Table 3.3: Final aligned multimodal split

3.5 Computational Environment

Resource	Final Research Environment
HCPC source	MIST high-performance computing access
GPU memory	32 GB
System RAM	64 GB
Storage	2 TB SSD
CPU	16 cores
L3 cache	64 MB
Stated clock speed	4.5 GHz
Primary software stack	Python, PyTorch ecosystem, OpenCV/FFmpeg-compatible media processing

Table 3.4: Computing resources used for large-scale implementation

High-performance computing was important because face extraction, repeated image decoding, feature generation, temporal model training, and large-scale inference are I/O- and compute-intensive. The available 32 GB GPU memory also reduced pressure to use extremely small batches during final training. Nevertheless, the pipeline was designed with checkpointing and resume support because long-running jobs may still be interrupted by scheduler, storage, or software issues.

Reproducibility depends on more than hardware. The implementation records environment versions, configuration files, random seeds, dataset manifests, and locked checkpoints. This reduces the risk that an apparently identical rerun silently uses a different library version, data split, or threshold.

3.6 Face Preprocessing with MTCNN

MTCNN face detection was used to create standardized face crops. The final visual manifest contained 77,538 crop rows covering 19,391 unique A/V IDs. 19,371 IDs had exactly four available face crops. The current Swin and GenConViT training pipeline did not consume all four crops simultaneously: one crop per sample per epoch was selected deterministically in an epoch-dependent manner, while validation used a deterministic middle path. This distinction prevents the stored-frame count from being misreported as simultaneous model input.

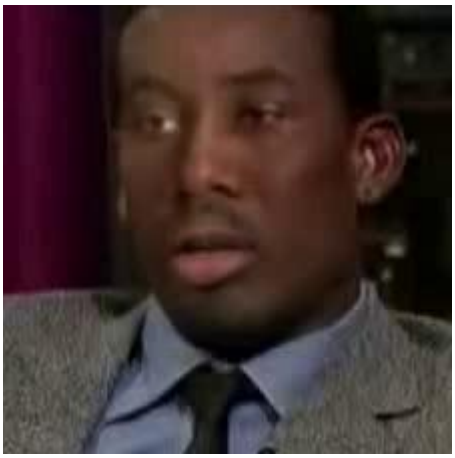


Fig: 3.3: Fake video full frame



Fig: 3.4: MTCNN face crop

Figure 3.3 and 3.4: Example of MTCNN-based face localization and face-crop construction in the DeepShield preprocessing pipeline

MTCNN performs cascaded face detection and alignment using a sequence of neural-network stages. In DeepShield, it is retained as the approved face-preprocessing component because earlier experimentation established a stable workflow around MTCNN-derived crops. The objective is not to compare face detectors but to provide a

consistent input geometry for downstream classification. Consistency between training and inference is critical. A crop produced with a different margin, alignment rule, or resizing strategy can shift the distribution of facial texture and background context. Earlier prototype experiments demonstrated that preprocessing choices can change model scores even when the source frame is identical. The final data pipeline therefore treats preprocessing parameters as versioned configuration rather than ad-hoc code embedded in an inference notebook. For multi-person media, a single-face detector call can return several candidates. A production-grade multi-person framework should maintain tracks for every relevant face, but the final locked quantitative result sheet does not separately report multi-person tracking performance. Consequently, this thesis describes the multi-person requirement at framework level while limiting numerical claims to the evaluated face/video classification pipeline.

3.7 Image and Video Sample Construction

Image samples are constructed from accepted MTCNN face crops and are used by ConvNeXt-Tiny to learn spatial real-versus-fake evidence. Video samples are represented by an ordered sequence of four face-derived embeddings. Each embedding corresponds to one deterministic frame and is generated by the ConvNeXt visual encoder. This design avoids re-learning an unrelated temporal feature space and enables the video model to reason over consistent spatial features.

The separation between image and video branches also supports analysis. The image model measures what can be learned from appearance alone, while the video model asks whether ordered spatial features provide additional discriminative information. The historical ConvNeXt image/video pathway remains an essential part of the implementation record, while the final result evidence reports the locked five-branch multimodal fusion and its branch-level validation/ablation metrics. The historical image model also generated features consumed by the earlier temporal pathway, whereas the final Face and Temporal branches are documented separately in Chapters 4-6.



Fig: 3.5: MTCNN Face crop-1



Fig: 3.6:MTCNN Face crop-2

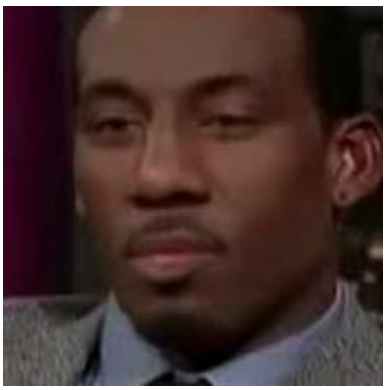


Fig: 3.7: MTCNN Face crop-3

Fig: 3.8:MTCNN Face crop-4



Figure 3.5-8: MTCNN-derived standardized face crop from a single video.

3.8 Audio and Multimodal Data Considerations

The DeepShield research concept includes voice analysis because a deepfake video may contain authentic audio, synthetic audio, or an audio-video mismatch. However, a visually fake video does not automatically contain fake speech. Public face-manipulation datasets often preserve the original audio, so assigning audio=fake from the video label would corrupt an audio classifier. A correct multimodal registry therefore maintains separate visual and audio ground-truth fields.

The uploaded non-training draft described MFCC, spectrogram, and CNN-LSTM voice processing as part of the proposal-aligned architecture. Those design ideas remain relevant to the framework, and multimodal literature supports cross-modal analysis [13]-[16]. The final locked result evidence now includes a trained Voice branch, branch-level validation/ablation evidence, and a final five-branch multimodal fusion result. For scientific integrity, the final book reports only the supplied audio and fusion evidence and does not invent unsupported external voice-deepfake generalization claims. The audio branch is therefore documented as an implemented component whose future work should extend evaluation to additional verified fake-audio datasets and unseen voice-generation methods.

3.9 Reproducibility and Data Governance

Raw data are not modified by training notebooks. Processed faces, embeddings, logs, checkpoints, predictions, and reports are written to separate versioned locations. This design protects against accidental overwriting and enables a failed stage to be rerun without re-downloading or re-labeling the raw datasets. It also supports auditability: a prediction can be traced back to a registered source path and preprocessing configuration.
Custom data require additional governance. Participant identities should be stored using pseudonymous identifiers in machine-learning files, while consent records remain separate. When data contain several people or voices, annotations should distinguish the

manipulated subject from bystanders and unrelated speakers. These requirements are especially important for future multi-person and audio-visual extensions of DeepShield.

3.10 Label Semantics and Ground-Truth Contracts

A multimodal deepfake dataset requires more than one binary label because visual authenticity, audio authenticity, and the global case label may not always be identical. A face-swap video can contain genuine original speech; a synthetic voice can be paired with an unmodified face; and a synchronization manipulation can create an audiovisual inconsistency without changing the appearance of individual frames. DeepShield therefore treats label semantics as a data-engineering contract rather than as an implicit property of the folder name.

The canonical registry preserves global target, visual target, audio target, source dataset, manipulation metadata when available, and modality availability. Branch training uses the label that corresponds to the evidence seen by that branch. This prevents a Voice model from learning incorrect targets merely because the associated video is visually fake. The same principle supports clearer error analysis because a case can be examined according to which modality is manipulated and which branch was available.

3.11 Canonical Case IDs, Provenance, and Traceability

Every derived artifact should remain traceable to an original media source. DeepShield therefore uses stable identifiers that link a canonical case to its dataset, source path, preprocessing outputs, split assignment, and branch artifacts. The practical advantage is that a suspicious prediction can be traced backward from the final forensic report to the frames, face crops, audio features, and manifest rows that contributed to it.

Traceability also reduces the risk of accidental duplication. Large datasets can contain copies, derivatives, or multiple manipulated versions of the same source identity. If identifiers are regenerated independently at each stage, two related items may appear unrelated during split construction. A stable registry allows grouping rules to be applied before downstream preprocessing and supports reproducible reruns when the file system or notebook session changes.

3.12 Identity-Safe Split Construction

Random file-level splitting is not sufficient when multiple clips share a subject, source video, or manipulation lineage. If closely related items appear in both training and test sets, a model can exploit identity or source cues and produce an optimistic estimate. DeepShield therefore uses dataset-aware and identity-aware grouping where the available metadata permit such control. The historical 177,584-sample selection and the final aligned cohort are documented separately because they represent different experimental scopes.

The final aligned split contains 17,234 TRAIN, 2,159 VAL, and 2,151 frozen TEST cases. TRAIN is used for parameter learning, VAL is used for checkpoint selection, ablation comparison, threshold selection, and calibration, and TEST is reserved for a single final evaluation. This separation is one of the most important controls in the thesis because it limits test-driven iteration and makes the final result easier to interpret scientifically.

3.13 Modality Availability and Missing-Data Handling

Missing modality handling is a first-class part of the data design. A case can remain valid at the master level even when a particular branch cannot produce evidence. For example, a video may have no usable speech segment, a face may be temporarily occluded, or a lip-sync window may fail alignment. Dropping every such case would shrink the dataset and could bias the retained population toward easy examples.

DeepShield instead records branch-level availability masks. During fusion, an unavailable branch is masked rather than replaced with a FAKE probability. This distinction is essential: absence of evidence is not evidence of manipulation. The mask-aware contract also makes branch coverage measurable, allowing the thesis to report how many cases contributed to each component and to investigate whether missingness correlates with error patterns.

3.14 Frame Decoding, Face Quality, and Visual Quality Control

Video preprocessing begins before the classifier sees an image. Frames must decode successfully, have a valid timestamp, contain sufficient visual information, and produce a usable face region when the face branch is required. Quality-control logic therefore records video-open failures, face-detection failures, and accepted crops rather than silently substituting arbitrary images. Deterministic sampling ensures that reruns select the same temporal positions unless the preprocessing version is deliberately changed.

Face quality is especially important for a multi-model visual branch. Swin, GenConViT, frequency analysis, and temporal processing should receive evidence derived from the same canonical case without uncontrolled crop differences. A versioned MTCNN configuration provides consistent detection and alignment, while stored crop metadata makes it possible to audit how many frames were available and which one was used in a particular training epoch or validation pass.

3.15 Audio Preparation and Feature Quality Control

Audio preparation standardizes waveform properties before feature extraction. In the documented DeepShield frontend, audio is converted to a 16 kHz representation and transformed into MFCC and Log-Mel sequences using a fixed window, hop length, FFT size, and Mel filter-bank configuration. Feature-wise normalization is then applied at the

utterance level. These steps create a stable numerical interface for the Voice CNN-LSTM and reduce variation caused by inconsistent source encodings.

Quality control must also distinguish silent, extremely short, corrupt, or unavailable audio from valid speech. A valid container track does not necessarily contain useful voice evidence. The pipeline therefore treats audio extraction, feature generation, and label validity as separate checks. This is particularly important in deepfake research because many visual datasets retain genuine speech even when the face is manipulated.

3.16 Lip-Sync Window Preparation

Lip-sync preprocessing aligns two temporal streams: mouth-region visual observations and the corresponding audio segment. The canonical manifest stores the number of valid windows, temporal offsets, and availability state so that the branch can evaluate synchronization without assuming that every video contains continuous visible speech. Multiple offsets are useful for characterizing whether the best audiovisual correspondence occurs at the expected alignment or at a shifted position.

The preprocessing stage also protects the fusion model from invalid cross-modal evidence. If the active speaker is not visible, the mouth is occluded, or the audio segment does not contain speech, a lip-sync score may be meaningless. Recording those conditions as availability information is preferable to creating a numerical score that appears valid but is based on poor evidence.

3.17 Background Context Construction

The Background branch uses full-scene evidence rather than face crops. Exactly four original frames per aligned case were constructed for the final background workflow, producing 86,176 original frame rows. For contextual ablation, a corresponding face-suppressed version was created so that the influence of non-facial scene information could be studied separately. This design supports a direct distinction between foreground facial evidence and wider environmental context.



Fig: 3.9: Background full frame



Fig: 3.10: Background face suppressed

Figure 3.9-10: Matched original full frame and face-suppressed frame constructed for contextual Background evidence.

Context construction requires careful bookkeeping because the frame used for background analysis should correspond to the same case and temporal neighborhood used elsewhere in the pipeline. The original and face-suppressed variants therefore remain linked by case ID and frame index. When CLIP or DINOv2 features are extracted, the resulting embeddings can be compared without ambiguity about which visual content generated them.

3.18 Class Imbalance and Sampling Considerations

The final aligned cohort is strongly imbalanced toward the FAKE class, particularly in the TEST split. Class imbalance affects how accuracy should be interpreted and can also influence optimization if the training sampler repeatedly exposes the model to one class more than the other. DeepShield therefore reports precision, recall, F1, ROC-AUC, confusion counts, and calibration metrics in addition to accuracy. These complementary metrics make the result more informative than a single overall percentage.

Sampling strategy must also preserve the natural relationship between case-level labels and branch availability. Aggressive oversampling can distort the joint distribution of modalities, while naive undersampling can discard rare but important REAL cases. The thesis therefore emphasizes transparent split composition and branch-level counts rather than claiming that one balancing rule is universally optimal.

3.19 Artifact Versioning, Checkpoints, and Audit Trail

A reproducible multimodal experiment produces many derived artifacts: manifests, face crops, audio feature arrays, temporal sequences, embeddings, model checkpoints, validation predictions, calibration parameters, fusion outputs, XAI reports, and final tables. DeepShield stores these outputs in versioned locations so that one pipeline stage can be rerun without overwriting the evidence produced by another version. The version identifier becomes part of the experimental record.

This audit trail is especially important when the project evolves over time. Earlier ConvNeXt-based visual-temporal experiments and later five-branch canonical experiments must remain distinguishable even when they share source datasets or preprocessing code. Versioned artifacts allow the thesis to preserve historical evidence while clearly identifying which files support the final result.

3.20 Preprocessing Pipeline Summary

The complete preprocessing pipeline can therefore be summarized as a controlled transformation from heterogeneous media to aligned branch-ready evidence. Raw media

are registered without destructive modification; source and identity metadata are preserved; split assignments are frozen; frames and audio are decoded under deterministic rules; face, temporal, voice, lip-sync, and background evidence are constructed independently; and every output remains linked to the canonical case ID.

The main methodological objective of this stage is not to maximize the number of derived files. It is to guarantee that the models in later chapters receive consistent inputs with known provenance and valid labels. The final 21,544-case aligned cohort is smaller than the historical broad registry because multimodal alignment imposes stricter requirements, but that tighter cohort is the appropriate basis for the final five-branch experiment.

CHAPTER 4

DEEPSHIELD SYSTEM ARCHITECTURE

4.1 Evolution from Proposal to Final Locked Configuration

The initial research proposal explored a broad ensemble containing Transformer-based visual models, frequency analysis, voice forensics, temporal consistency, background analysis, lip-synchronization, multimodal fusion, and explainability. Such an architecture is attractive conceptually because different detectors can compensate for one another. During implementation, however, every additional branch increases data-label requirements, training cost, integration risk, and the possibility that a weak branch degrades rather than improves the final decision.

The final training evidence shows that the project converged on a five-branch locked configuration. MTCNN performs face preprocessing; the Face branch combines Swin Transformer, official GenConViT-ED, GenConViT-VAE, and frequency-domain evidence; the Voice branch uses MFCC and Log-Mel CNN-LSTM modeling; the Temporal branch uses ordered face evidence with BiLSTM attention; the Lip-Sync branch uses aligned mouth-audio evidence; and the Background branch uses CLIP and DINOv2 representations. Mask-aware learned attention combines the available branch tokens before validation-only calibration and the final locked decision. The earlier ConvNeXt-BiLSTM visual-temporal pathway and standalone frequency/DINOv2 experiments remain documented as historical and ablation evidence rather than being confused with the final canonical fusion.

This architectural evolution is scientifically meaningful. A thesis should describe what was actually trained and evaluated, not only what was proposed at the beginning. The proposal remains useful for explaining the research direction and the unimplemented extensions, while the locked result sheet defines the quantitative claims.

4.2 Face Detection

The input image or sampled video frame is passed through MTCNN to obtain a face region. Standardized cropping reduces irrelevant scene content and provides a consistent scale for the spatial classifier. Face detection also acts as a quality gate: a frame without a usable face cannot contribute the same type of evidence as a clear frontal face.

Face-centric preprocessing is particularly important for a detector that uses only four frames. If one of those frames fails detection, the effective temporal evidence is reduced. A robust implementation therefore records detection outcomes and can exclude unusable frames rather than silently replacing them with unrelated content.

4.3 Historical ConvNeXt-Tiny Image Branch

ConvNeXt-Tiny maps an aligned facial crop to a deep feature representation and an image-level authenticity prediction. The architecture is well suited to this role because it combines modern convolutional design with efficient hierarchical feature extraction [7]. Features from earlier stages capture local texture and edges, while deeper stages integrate larger spatial context.

The image branch has two roles in the final system. First, it can classify single facial images. Second, its embeddings become the frame representations used by the video model. This dual use encourages consistency between image and video processing and avoids the computational cost of training a completely separate visual encoder for the temporal branch.

4.4 Historical Video Feature Sequence

For each video, four deterministic frame positions are selected. After face preprocessing, the ConvNeXt encoder generates one embedding per accepted frame. The ordered embedding matrix can be written conceptually as X = [x1, x2, x3, x4], where each xi is a learned visual representation of the corresponding face crop.
This representation separates spatial and temporal learning. ConvNeXt is responsible for describing what each frame looks like; the sequence model is responsible for learning how those representations behave together. This decomposition is common in CNN-recurrent approaches to video deepfake detection [8], [9].

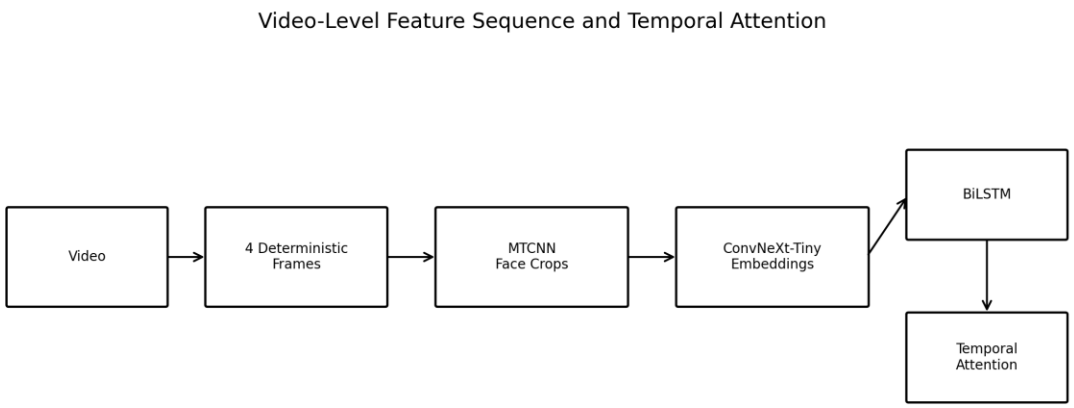


Figure 4.1: Historical video-level feature sequence and temporal-attention pathway.

4.5 Historical Bidirectional LSTM

The BiLSTM reads the ordered ConvNeXt embeddings in forward and reverse directions. The forward state summarizes evidence from earlier frames, while the backward state summarizes evidence from later frames. Concatenating these states produces a context-aware representation for each sequence position. This is useful when a suspicious frame should be interpreted relative to the appearance before and after it.

Temporal identity methods in the literature provide strong motivation for this strategy [10], [11]. The final DeepShield implementation does not claim to reproduce those architectures; instead, it adopts the broader principle that identity and appearance consistency over time contain forensic information.

4.6 Historical Temporal Attention

Temporal attention assigns an importance coefficient to each BiLSTM output. If h_t is the hidden representation at time t, a scoring function produces e_t, and normalized weights are obtained using a softmax operation. The video representation is then a weighted sum of the hidden states. Conceptually, frames with stronger discriminative evidence receive more influence on the final representation.

The attention layer does not guarantee that the most visually obvious manipulation receives the highest weight. Its parameters are optimized for classification performance. Consequently, attention should be analyzed as part of model behavior rather than treated as an objective localization label.

$$\alpha_t = \exp(e_t) / \Sigma_j \exp(e_j), \qquad v = \Sigma_t \alpha_t h_t$$

4.7 Final Temporal Preprocessing

Temporal modeling required at least two ordered face crops and could use up to four evenly spaced frames. The final aligned temporal availability included 19,391 IDs, with 15,523 TRAIN cases and 1,948 validation cases used during temporal model development. Ordered sequence construction preserved frame order so that the BiLSTM could model motion and identity consistency across time.

4.8 Final Audio Feature Extraction

Audio was standardized to 16 kHz. The documented frontend used pre-emphasis 0.97, a 400-sample window, 160-sample hop, 512-point FFT, 64 Mel bands, and 20 MFCC coefficients. MFCC arrays had shape [T,20] and Log-Mel arrays [T,64], followed by per-utterance feature-wise normalization. The Voice branch was trained from these precomputed feature arrays, not directly from raw waveforms.

4.9 Final Lip-Sync Evidence

Lip-Sync evidence was generated from aligned mouth and audio windows. The final aligned manifest covered all 21,544 master cases and included availability masks, window counts, offset statistics, and multiple temporal offsets. The design intentionally kept Lip-Sync evidence separate from Voice evidence so that synchronization mismatches could contribute independently during fusion.

4.10 Final Background Context Frames

The Background branch used exactly four original full video frames per aligned case, producing 86,176 original frame rows. Each original frame also had a face-suppressed counterpart for contextual ablation. These were not face crops: they preserved the scene and surrounding context. CLIP ViT-B/32 and DINOv2 ViT-S/14 embeddings were extracted from the full-frame evidence.

4.11 Explainability with Grad-CAM

Grad-CAM generates a class-specific heatmap from convolutional feature maps and gradients. In DeepShield, it is used to visualize which facial regions influence the image-level prediction. Examples of plausible focus areas include boundaries around eyes, mouth, skin texture, or blending regions, but the method itself does not label a region as manipulated.

The explainability layer is most useful when combined with input inspection. If a model highlights only a background border or compression block, the prediction may depend on a shortcut. If it consistently attends to relevant face regions, confidence in the plausibility of the learned evidence increases, although the explanation still remains qualitative [20], [21].

4.12 Final Face Ensemble Architecture

The final Face branch is an ensemble rather than a single backbone. Its purpose is to combine complementary spatial representations before the global multimodal fusion stage. Swin Transformer contributes hierarchical self-attention features, GenConViT-ED and GenConViT-VAE contribute generative-convolutional-transformer evidence with different latent reconstruction mechanisms, and the Frequency component contributes spectral information. These subbranch outputs are projected and combined by learned face-level attention to produce one canonical Face token.

The ensemble design reflects a central deepfake-detection challenge: no single representation is guaranteed to capture every manipulation artifact. Pixel-space texture, hierarchical semantic structure, latent reconstruction behavior, and frequency-domain traces can respond differently to compression, blending, synthesis, and identity manipulation. Combining them at the Face-branch level allows the global fusion model to

receive one consolidated visual representation while still preserving subbranch attention weights for analysis.

4.13 Swin Transformer Representation

Swin Transformer uses a hierarchical vision architecture in which self-attention is computed within local windows and successive blocks shift the window partition to enable cross-window communication. This design reduces the computational burden of global self-attention while maintaining multi-scale visual representation [23]. For deepfake detection, hierarchical attention is useful because manipulation evidence may occur at different spatial scales: local texture around eyes or mouth, boundary inconsistencies around the face, and broader structural relationships across the crop.

Within DeepShield, Swin is not treated as an isolated final classifier. Its role is to produce a strong subbranch representation that complements GenConViT and frequency evidence. The high validation ROC-AUC reported for the Swin subbranch supports its usefulness on the aligned validation population, while the thesis avoids interpreting that value as independent proof of generalization to unseen domains.

4.14 GenConViT-ED and GenConViT-VAE Subbranches

GenConViT combines convolutional and Transformer-style visual representation with generative latent modeling. The published architecture uses ConvNeXt and Swin components for feature extraction and integrates autoencoder or variational-autoencoder pathways to model latent structure in deepfake video detection [24]. DeepShield retains the distinction between the ED and VAE variants because the two generative pathways can emphasize different reconstruction and latent-distribution behavior.

The ED variant can be interpreted as reconstruction-aware feature learning with a deterministic latent pathway, whereas the VAE variant introduces a probabilistic latent representation. In the final Face ensemble, both are treated as evidence generators rather than as complete independent systems. Their embeddings are combined with Swin and frequency evidence, allowing the face-level attention mechanism to assign relative importance based on the current case.

4.15 Frequency-Domain Evidence Pathway

Deepfake generation can introduce artifacts that are subtle in pixel space but more visible after a frequency transform. Upsampling operations, synthesis pipelines, compression, and blending can affect spectral energy distribution or produce periodic patterns. Frequency-aware deepfake studies therefore examine transformed representations as a complementary evidence domain [18], [19].

In DeepShield, the Frequency component is deliberately weaker than the strongest visual subbranches when considered alone, but this does not make it useless. A branch can

contribute complementary information even when its standalone ROC-AUC is lower. The correct interpretation is empirical: its value should be determined through ablation and fusion behavior rather than assumed from theory. This is why the thesis reports the standalone validation result and retains the component within the documented Face-branch design.

4.16 Voice CNN-LSTM Architecture

The Voice branch receives two synchronized acoustic feature sequences: MFCC and Log-Mel representations. Convolutional layers model local time-frequency patterns such as short spectral transitions, while recurrent layers model how those patterns evolve over a longer utterance. The final locked VoiceCNNLSTM produces a fixed-dimensional embedding together with a branch-level probability of manipulation.

A key architectural decision is to keep the Voice branch independent from the visual branch until the fusion stage. This prevents visual labels from contaminating the audio objective and allows genuine speech to coexist with a manipulated face. It also means that the voice embedding can contribute only when valid audio evidence exists. When audio is unavailable, the global modality mask prevents the fusion attention from treating the missing branch as a negative or positive signal.

4.17 Final Temporal Branch Architecture

The final Temporal branch models ordered facial evidence rather than treating sampled frames as an unordered set. After face preprocessing, frame-level representations are arranged in temporal order and processed by a bidirectional recurrent model with attention. The forward and backward hidden states allow each position to be interpreted using context from both earlier and later frames, while the attention mechanism forms a weighted case-level temporal representation.

This branch addresses a different question from the Face ensemble. The Face branch asks whether individual visual observations contain manipulation evidence; the Temporal branch asks whether the sequence behaves consistently over time. Its standalone validation AUC is lower than the principal Face models, so the final thesis characterizes Temporal as supporting evidence rather than the dominant signal. That distinction is important because a multimodal architecture should not imply that every branch contributes equally.

4.18 Lip-Sync Architecture and Cross-Modal Consistency

The Lip-Sync branch operates on paired mouth-motion and audio windows. Its architecture is designed to produce a consistency representation rather than a conventional visual-only or audio-only deepfake score. Temporal offsets are part of the evidence because a synthetic or edited sequence may preserve plausible speech and plausible mouth motion while shifting their correspondence.

The final fusion experiment shows that the lip-sync branch did not improve validation AUC relative to the no-Lip-Sync ablation in the recorded run. This negative or neutral result remains scientifically useful. It indicates that synchronization evidence is data-dependent and that a branch should be retained or weighted based on measured behavior rather than architectural intuition alone.

4.19 Background Architecture with CLIP and DINOv2

The Background branch encodes full-frame contextual evidence using CLIP and DINOv2. CLIP learns visual representations through image-text supervision and transfers to many downstream visual tasks [25], while DINOv2 learns robust self-supervised visual features [17]. These representation families are attractive for context analysis because they encode semantic and structural information that differs from a face-specific classifier.

DeepShield uses both original and face-suppressed frames to investigate whether contextual discrimination depends primarily on the visible face. The combined background pathway is therefore both a detection branch and an ablation mechanism. Strong contextual performance must be interpreted with caution because it can reflect legitimate manipulation cues or dataset-specific shortcuts. The masked-background comparison helps expose that distinction.

4.20 Common Projection, Mask-Aware Fusion, and Decision Head

Because the five branches produce embeddings with different dimensions and statistical properties, a common projection layer maps each available branch representation into a shared fusion space. Learned attention scores are then computed over the available tokens. Before normalization, unavailable branches are masked so that they receive no attention probability. This creates a case-specific evidence mixture rather than forcing every video to use a fixed branch combination.

The fused representation is passed to the final decision head, which produces a case-level logit. Calibration is applied after the fusion checkpoint is locked, and the validation-derived decision threshold is then fixed before frozen TEST evaluation. This separation between representation learning, evidence fusion, probability calibration, and decision thresholding makes the architecture easier to audit and reduces the risk that operational settings are mixed into model training.

CHAPTER 5

PROPOSED DEEPSHIELD METHODOLOGY

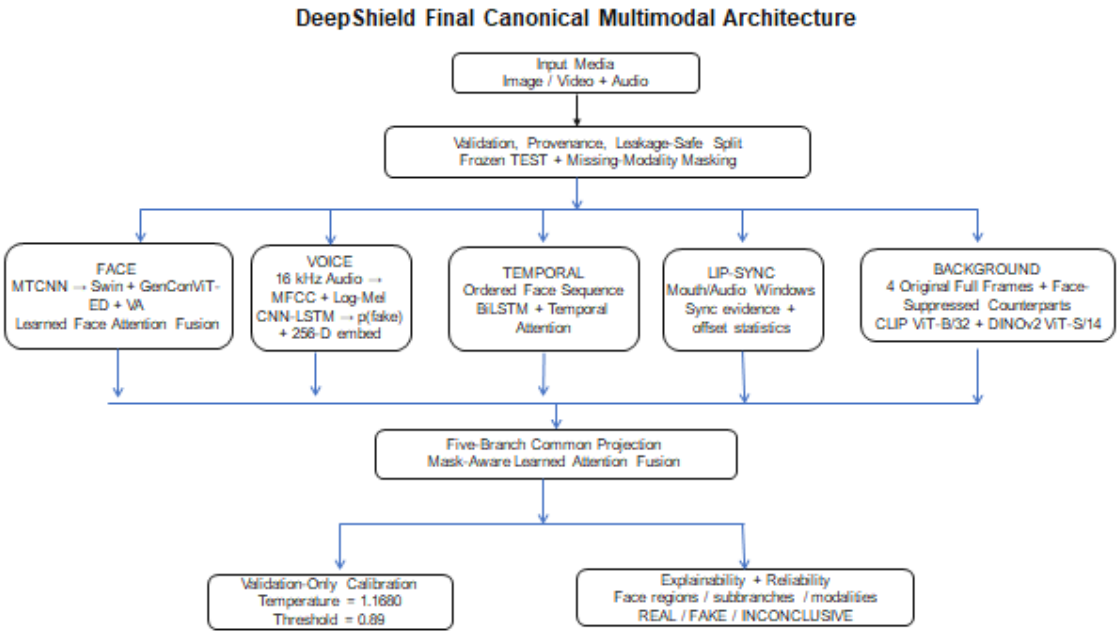


Figure 5.1: DeepShield final canonical multimodal architecture

5.1 End-to-End Methodological Overview

The final methodology is organized as a sequence of controlled transformations from registered media to a calibrated case-level forensic decision. First, raw media are mapped to canonical IDs and frozen TRAIN/VAL/TEST assignments. Second, branch-specific preprocessing creates face crops, audio features, ordered temporal sequences, lip-sync windows, and full-frame background evidence. Third, each branch produces an embedding and a local score under its own label and availability contract. Fourth, available branch embeddings are projected into a common space and fused with learned mask-aware attention. Fifth, the fusion checkpoint is selected using validation evidence, calibration is fitted on validation logits, the decision threshold is locked, and the final frozen TEST is evaluated once.

This sequence is important because it prevents methodological shortcuts. The model does not infer branch availability from a zero-filled vector, the audio target is not copied from the visual target, and the final TEST does not participate in checkpoint or threshold selection. Each step produces an artifact that can be audited independently. The result is a

methodology in which data preparation, representation learning, fusion, calibration, and reporting are connected but remain distinguishable experimental stages.

Stage	Primary input	Core operation	Primary output
1. Registry and split	Heterogeneous media	Canonical IDs, labels, provenance, frozen split	Aligned case registry
2. Face evidence	Video/image frames	MTCNN crops, visual subbranch preparation	Face inputs and embeddings
3. Voice evidence	Audio track	16 kHz standardization, MFCC and Log-Mel extraction	Voice feature sequences
4. Temporal evidence	Ordered face observations	Sequence construction and BiLSTM attention	Temporal embedding
5. Lip-Sync evidence	Mouth + audio windows	Alignment and offset analysis	Lip-Sync embedding
6. Background evidence	Full frames	Original/masked context encoding with CLIP/DINOv2	Background embedding
7. Fusion	Available branch embeddings	Projection, modality mask, learned attention	Fusion logit
8. Reliability and report	VAL/TEST logits	Temperature scaling, locked threshold, XAI	Calibrated forensic report

Table 5.1: End-to-end DeepShield methodological sequence

5.2 Face Branch

The Face branch combines four complementary subbranches. Swin Transformer contributes hierarchical self-attention features. Official GenConViT-ED contributes useful generative-convolutional-vision-transformer evidence, while GenConViT-VAE contributes a generative reconstruction-aware feature. The Frequency component encodes spectral evidence from a representative face frame. A learned face-attention fusion combines the subbranch embeddings into one canonical Face token and preserves subbranch attention weights for explainability.

5.3 Face Standardization with MTCNN

The Face methodology begins with standardized localization and alignment. MTCNN applies a cascaded multi-task design for face detection and landmark estimation [26]. In DeepShield, the detector is used as a preprocessing tool rather than as a trainable part of the final classifier. Its purpose is to reduce geometric variation so that downstream models receive a consistent facial region instead of an uncontrolled mixture of full-frame backgrounds and face scales.

The preprocessing configuration is kept stable across training, validation, and inference. A change in crop margin, resize rule, landmark alignment, or selected face can shift the distribution seen by the visual models. For multi-person frames, a production system would need persistent identity tracking, but the final quantitative experiment is interpreted at the evaluated case level and does not claim a separate tracking metric.

5.4 Deterministic Face-Frame Sampling

The final aligned visual manifest stores multiple face crops per case, but the training and validation pathways do not necessarily consume every crop simultaneously. Deterministic selection rules are used so that repeated experiments can reproduce which temporal evidence was available. Epoch-dependent deterministic selection can expose the visual model to different stored crops over training while avoiding uncontrolled random behavior that would make exact reruns difficult.

For validation, deterministic middle-path or otherwise fixed selection is preferred because validation metrics should not change merely because a different random frame was sampled. Temporal modeling uses ordered sequences and therefore preserves frame order explicitly. These distinctions are documented so that the number of stored crops is not confused with the number of images presented to a particular model in one forward pass.

5.5 Swin Transformer Subbranch Method

The Swin subbranch receives standardized face evidence and transforms the image into a hierarchical representation using shifted-window self-attention [23]. Local attention allows the model to examine texture and structure within windows, while shifted partitions enable information exchange across neighboring regions. For deepfake detection, this creates a useful balance between local artifact sensitivity and larger-scale facial structure.

During model development, checkpoint selection is based on validation behavior rather than final TEST performance. The resulting Swin embedding is retained as one component of the Face ensemble, not as the final global decision. This methodological separation allows the thesis to report a strong Swin validation result while still evaluating whether other visual evidence and non-visual modalities improve the complete forensic system.

5.6 GenConViT-ED Subbranch Method

The GenConViT-ED pathway is used to capture deepfake evidence through a combination of discriminative visual representation and reconstruction-aware latent modeling [24]. The encoder-decoder mechanism encourages the model to represent structure that is useful not only for class separation but also for reconstructive consistency. In practice, this can make the representation sensitive to synthesis artifacts that are not identical to the features emphasized by a purely discriminative backbone.

DeepShield uses the official GenConViT implementation lineage for this subbranch and preserves the locked checkpoint as a separate evidence source. Its embedding is later combined with the other Face subbranches. The thesis intentionally avoids reconstructing undocumented optimizer or schedule details from memory; exact run settings remain tied to the archived training configuration and checkpoint artifacts.

5.7 GenConViT-VAE Subbranch Method

The VAE variant introduces probabilistic latent modeling. Instead of representing each input with only one deterministic latent point, a variational formulation models a latent distribution and regularizes its structure. For forensic analysis, this provides a representation that can respond to differences between natural facial data and generated data in a way that complements the ED pathway [24].

The methodological role of GenConViT-VAE is therefore complementary rather than redundant. Its branch score and embedding are retained independently until face-level fusion. If ED and VAE respond differently to a case, learned attention can adjust their relative contribution. This also creates an interpretable subbranch comparison during validation and XAI analysis.

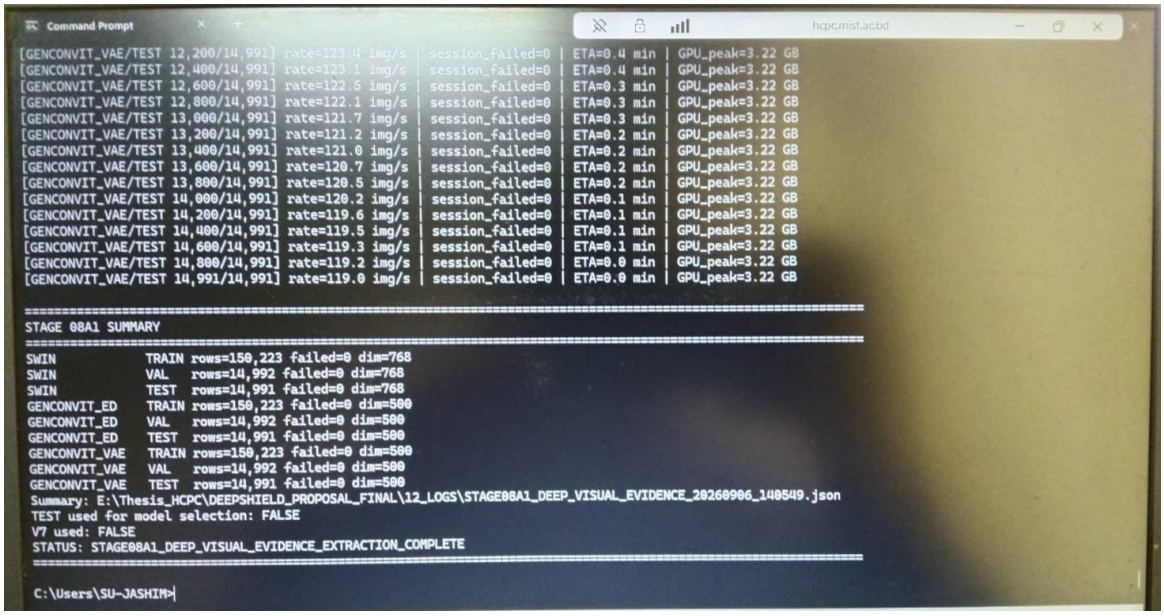


Figure: 5.2 : GPU-based Swin Transformer and GenConViT visual evidence extraction during the historical visual development stage.

5.8 Frequency Evidence Extraction

Frequency evidence is created from a representative facial observation using a deterministic transform pipeline. The objective is to expose spectral patterns that may be difficult to recognize directly in RGB space. Frequency-aware studies show that manipulated imagery can contain characteristic energy distributions or periodic traces associated with synthesis and resampling operations [18], [19].

The final methodology does not assume that a frequency representation must outperform modern visual backbones. Instead, the branch is evaluated empirically. Its lower standalone validation AUC is reported, while its possible complementary role is assessed through the Face ensemble and ablations. This is a more defensible methodology than including a frequency branch solely because prior literature suggests that spectral artifacts may exist.

5.9 Learned Face-Subbranch Fusion

The Swin, GenConViT-ED, GenConViT-VAE, and Frequency outputs are not simply averaged. Each subbranch representation is mapped into a compatible face-fusion space, and a learned attention mechanism estimates relative importance for the current case. The weighted representation becomes the canonical Face token passed to the global multimodal fusion stage.

Face-level attention provides two practical benefits. First, it allows a strong visual subbranch to dominate when its evidence is clear without forcing equal contribution from a weaker component. Second, the resulting subbranch weights can be recorded as part of the explanation report. These weights do not prove which model is correct, but they make the ensemble decision more inspectable than an opaque concatenation.

5.10 Voice Branch

The Voice branch receives MFCC and Log-Mel sequences. Convolutional layers learn local time-frequency patterns, followed by recurrent sequence modeling. The locked VoiceCNNLSTM produces a 256-dimensional embedding and a scalar p(fake). Because voice content can remain genuine when the video is visually manipulated, the audio label contract is explicitly aligned to the audio source rather than derived from the visual or global label.

5.11 Audio Standardization and Feature Extraction

The Voice pipeline begins by standardizing audio to a consistent sample rate and then computing MFCC and Log-Mel representations. The documented frontend uses a 16 kHz signal, pre-emphasis, fixed analysis windows and hop length, a fixed FFT size, 64 Mel bands, and 20 MFCC coefficients. MFCC provides a compact representation of the spectral envelope, while Log-Mel preserves a denser time-frequency description.

The two feature streams are normalized at the utterance level so that the classifier is less sensitive to absolute recording scale. Feature extraction is performed before model training and stored as reproducible arrays. This reduces repeated waveform decoding during training, improves consistency across runs, and creates an auditable interface between media preprocessing and the Voice neural network.

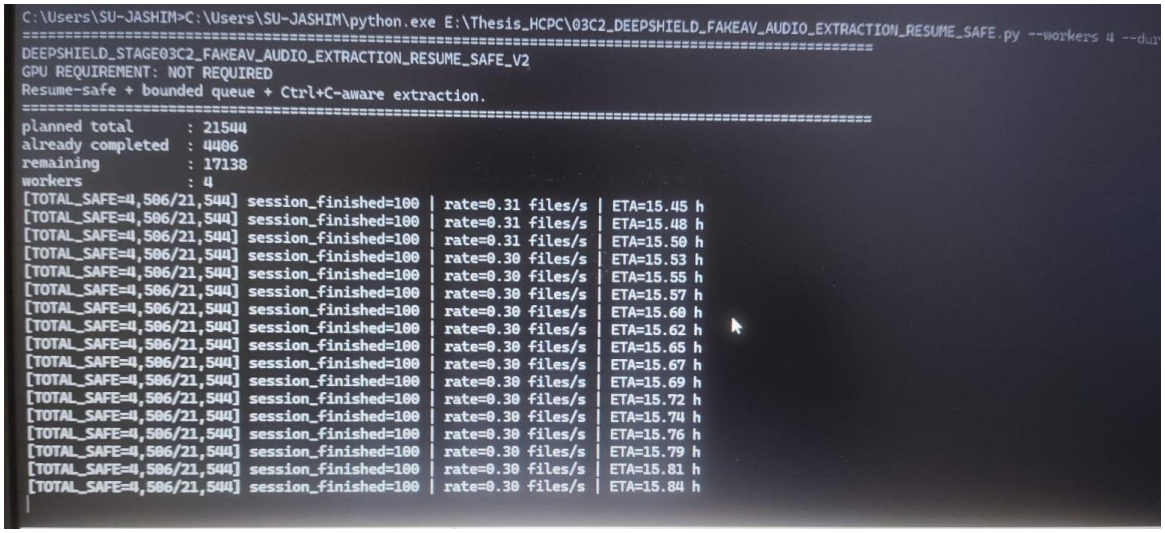


Figure 5.3: Resume-safe audiovisual audio extraction for the aligned DeepShield cohort.

5.12 Voice CNN-LSTM Training and Inference

The VoiceCNNLSTM processes acoustic features in two stages. Convolutional layers learn local motifs in the time-frequency representation, such as short spectral transitions or vocoder-related structures. Recurrent sequence modeling then integrates these local patterns over time, producing an utterance-level representation. The final branch exposes both a fixed-dimensional embedding and a scalar p(fake) for downstream fusion and analysis.



Figure 5.4: Voice-branch contract verification showing MFCC and Log-Mel feature dimensions and locked CNN-LSTM compatibility.

Training uses the audio-label contract rather than the global visual label. This is a methodological safeguard because visually manipulated videos may still contain genuine speech. At inference time, the branch contributes only when the audio feature pipeline marks the case as available. If the audio is missing, corrupt, or not valid under the contract, the global fusion mask excludes the Voice token instead of assigning an artificial score.

5.13 Temporal Branch

The Temporal branch receives ordered face descriptors from two to four available frames. A bidirectional LSTM captures dependencies in both temporal directions, and temporal attention weighs informative frames. The branch is intended as supporting evidence for identity-motion and temporal unusual sign rather than the dominant detector.

5.13.1 Ordered Temporal Sequence Construction

Temporal sequence construction preserves the chronological order of face-derived observations. Each accepted frame is associated with its index or timestamp, and the resulting feature sequence is passed to a bidirectional recurrent network. The minimum sequence requirement prevents the temporal model from pretending to reason about motion when only one observation exists, while the maximum of four frames maintains the efficiency chosen for the final pipeline.

The recurrent output is summarized with temporal attention so that some positions can contribute more strongly than others. This architecture does not guarantee that the highest attention frame corresponds to the visually most obvious artifact; the weight is optimized for classification. For that reason, temporal attention is interpreted as model behavior rather than a ground-truth manipulation localization.

5.14 Lip-Sync Branch

The Lip-Sync branch compares mouth-region dynamics with temporally aligned audio evidence. SyncNet-style dual-encoder reasoning and AV-HuBERT-inspired evidence were used conceptually in the feature construction, while offset statistics and hard-negative temporal shifts help characterize synchronization mismatches. The final fused token contains Lip-Sync evidence and an explicit availability mask.

5.14.1 Lip-Sync Alignment and Offset Evaluation

The Lip-Sync method forms paired visual and acoustic windows around speaking activity and evaluates their temporal correspondence. Multiple candidate offsets allow the system to observe whether the strongest compatibility occurs near the expected alignment or at a displaced position. The manifest records window statistics and availability so that the fusion model receives synchronization evidence only when the alignment procedure produced a valid result.

This branch is evaluated through ablation rather than assumed to be beneficial. In the recorded validation study, removing Lip-Sync did not reduce AUC relative to the full fusion. The thesis therefore presents Lip-Sync as a theoretically meaningful but empirically data-dependent signal. This is an important methodological example of retaining negative evidence in the report instead of rewriting the architecture after seeing the final result.

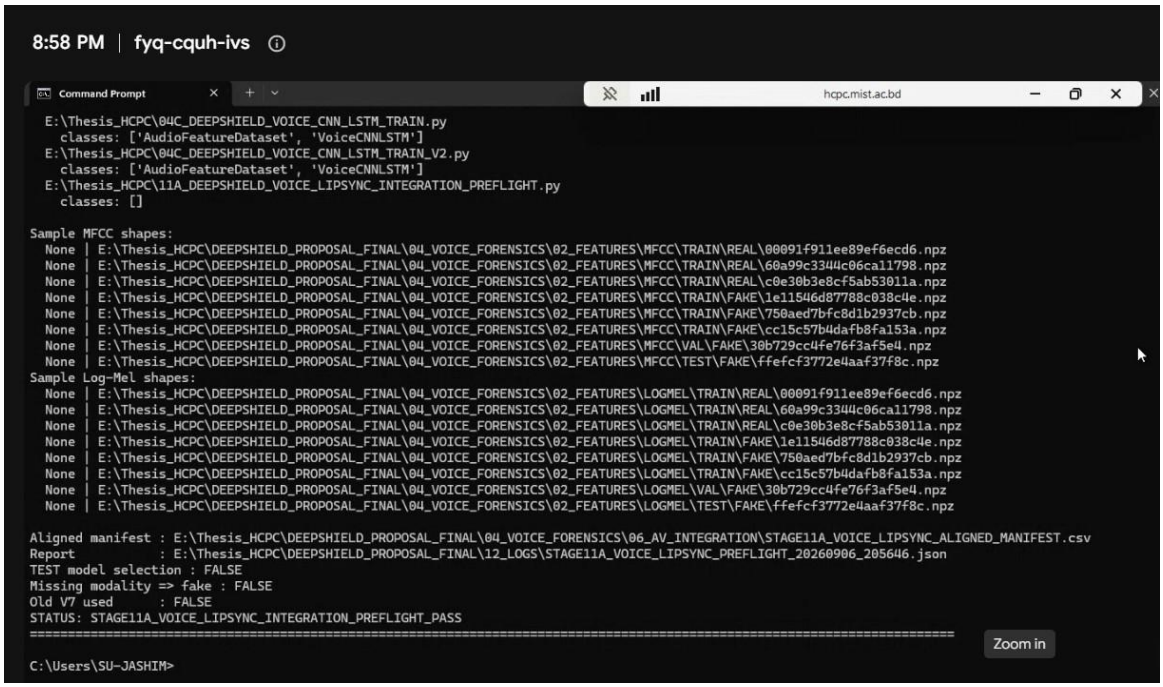


Figure : 5.5: Lip sync integration preflight

5.15 Background Branch

For each case, four original full frames and their face-suppressed versions are embedded. OpenCLIP ViT-B/32 with OpenAI weights provides a 512-dimensional semantic feature, while DINOv2 ViT-S/14 contributes a 384-dimensional self-supervised feature. Cross-frame cosine similarity, variance, and adjacent change statistics are combined with learned projection heads. Separate CLIP, DINOv2, combined, and masked heads enable branch-level ablation and evidence meaning.

5.15.1 CLIP and DINOv2 Context Encoding

For each selected full frame, the Background branch extracts representation using CLIP and DINOv2. CLIP contributes semantically informed features learned from image-text supervision [25], while DINOv2 contributes self-supervised visual structure [17]. Original and face-suppressed frames are processed under consistent indexing so that background-only behavior can be compared against full-context behavior.

The context features are aggregated to case level before global fusion. Because scene context can encode dataset-specific shortcuts, the methodology does not interpret a strong

background score as direct proof of manipulation. Instead, background ablations are used to determine whether context carries stable complementary evidence and whether performance changes when the face is suppressed.

5.16 Common Projection and Mask-Aware Attention Fusion

Each available branch is projected into a common feature space. A modality mask removes unavailable branches from attention rather than replacing them with a fake prior. Learned attention produces a fused feature and modality contribution weights. This model design supports mixed inputs while preserving a consistent final classifier interface.

5.16.1 Modality Masks and Attention Normalization

Each branch produces two outputs relevant to fusion: an embedding and an availability flag. Before attention normalization, logits corresponding to unavailable branches are masked so that their normalized weight becomes zero. Attention is then computed only over the evidence that actually exists for the current case. The fused token is a weighted sum of projected available branch embeddings.

This mechanism is preferable to filling a missing branch with zeros or a fixed class probability because such placeholders can become learnable shortcuts. A zero vector may still have a predictable relationship with a class if missingness is correlated with the target. Explicit masks make the missing-data rule visible and allow the fusion report to explain which modalities contributed to a particular prediction.

5.17 Calibration, Reliability, and Decision

The final fusion logits are calibrated using temperature scaling fitted only on validation data. The locked temperature is 1.1680239439 and the locked decision threshold is 0.89. The framework supports REAL, FAKE, and INCONCLUSIVE outcomes. Reliability reflects evidence availability and calibration quality. it is not a legal or causal certainty measure.

5.17.1 Validation-Only Threshold and Temperature Locking

After model and fusion checkpoints are fixed, the validation set is used to select the operating threshold and fit a scalar calibration temperature. These two operations address different questions. Threshold selection defines how a probability is converted into a class decision, while temperature scaling adjusts the confidence distribution without changing the ranking order of logits.

The final locked values are a temperature of 1.1680239439 and a decision threshold of 0.89. Both are frozen before the 2,151-case TEST evaluation. This procedural order is essential: using TEST labels to refine either value would make the final accuracy, recall, precision, F1, ECE, and Brier score partially test-tuned rather than genuinely held out.

5.18 Explainability and Forensic Reporting

DeepShield reports where suspicious evidence is concentrated and which modalities influenced the decision. Face XAI considers eyes, nose, lips/mouth, forehead/hairline, cheeks/skin, jaw/chin, neck/throat and other normalized crop regions. Additional evidence is reported for background, Voice, Lip-Sync, temporal frames, Face subbranches, and modality attention. The final report interprets these as model-evidence locations rather than proof that a specific pixel or anatomical region was manipulated.

Branch	Input Contract	Representation / Output
Face	MTCNN face crops	Swin 768-D + GenConViT-ED 500-D + GenConViT-VAE 500-D + Frequency → learned Face token
Voice	MFCC [T,20] + Log-Mel [T,64]	256-D embedding + p(fake)
Temporal	2–4 ordered face frames	BiLSTM + temporal-attention token
Lip-Sync	Aligned mouth/audio windows	16-D synchronization evidence + availability
Background	4 original full frames + masked copies	CLIP 512-D + DINOv2 384-D + context statistics
Final Fusion	Five branch tokens + masks	Mask-aware learned attention → calibrated p(fake)

Table 5.2: Branch input and feature contracts

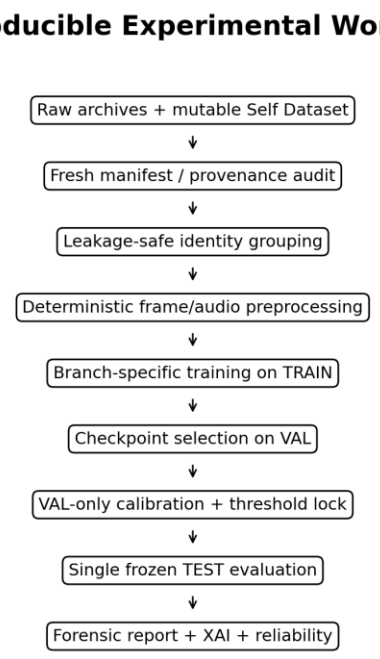


Figure 5.6: Reproducible experimental workflow

5.19 End-to-End Inference Sequence

At inference time, a media case is first resolved to its canonical ID and required preprocessing artifacts. Available Face, Voice, Temporal, Lip-Sync, and Background inputs are loaded or generated under the locked configuration. Each branch produces an embedding and local evidence score. The global fusion module projects embeddings, applies the modality mask, computes attention weights, and generates a fusion logit. The locked temperature converts that logit into a calibrated probability, and the threshold of 0.89 produces the final binary decision.

The final report retains more than the class label. It records the calibrated score, modality availability, branch or subbranch contribution information, XAI outputs where applicable, and the case identifier required for traceability. This report design is important for a forensic screening system because it allows a reviewer to inspect the evidence path instead of receiving only REAL or FAKE.

5.20 Methodological Reproducibility and Failure Handling

A practical multimodal pipeline must define what happens when an intermediate stage fails. DeepShield treats decode errors, missing faces, unavailable speech, insufficient temporal frames, invalid synchronization windows, and branch-model failures as explicit states rather than silently converting them into class evidence. Logs and manifests record these outcomes so that a case can be rerun or audited later.

Reproducibility is strengthened by deterministic sampling, frozen manifests, checkpoint locking, versioned feature directories, and validation-only selection. The methodology also avoids reconstructing unrecorded hyperparameters after the fact. When an exact setting is not present in the archived experiment, the thesis describes the verified architecture and procedure without inventing a numerical value. This principle is particularly important in a final thesis because apparent completeness should not come at the cost of scientific accuracy.

Branch	Evidence family	Representative processing	Availability condition
Face	Facial spatial and spectral evidence	Swin + GenConViT-ED + GenConViT-VAE + Frequency, learned face attention	Usable face evidence
Voice	Acoustic authenticity evidence	MFCC + Log-Mel -> CNN-LSTM	Valid labeled audio features
Temporal	Cross-frame consistency	Ordered face sequence -> BiLSTM + temporal attention	At least required ordered face observations
Lip-Sync	Audio-visual timing consistency	Aligned mouth/audio windows and temporal offsets	Valid speaking/alignment windows
Background	Scene and context evidence	Original and face-suppressed frames -> CLIP/DINOv2	Valid full-frame context
Fusion	Case-level multimodal decision	Common projection + mask-aware learned attention	At least one valid branch

Table 5.3: Final branch roles, evidence, and availability contract

5.21 Software Toolchain and Responsibility of Each Tool

The DeepShield methodology depends on a software toolchain in which each package has a distinct responsibility. Python provides the orchestration layer for manifests, data loading, feature processing, training scripts, audits, and report generation. PyTorch provides tensor operations, automatic differentiation, neural-network modules, checkpoint serialization, and GPU execution for the trainable branches. Torchvision supports standard image transformations and pretrained-model interfaces where appropriate. OpenCV and FFmpeg-compatible media processing are used for video decoding, frame extraction, image conversion, audio stream handling, and practical media validation. Keeping these responsibilities separated reduces the risk that a preprocessing change is hidden inside model code.

Model-specific libraries are also isolated according to their role. The official GenConViT codebase is retained for compatibility with the ED and VAE subbranches rather than being rewritten into an approximate local implementation. A vendored timm 0.6.5 environment is used where required by model compatibility, while OpenCLIP provides

the CLIP visual encoder used by the Background branch. DINOv2 is loaded from a controlled local torch-hub cache so that the model definition and weights do not change unexpectedly between runs. This toolchain design is methodological: reproducibility depends not only on model names but also on the exact software path that creates their features.

The operating environment further includes GPU and system-resource management. The final experiments were executed in the documented Windows HCPC setting with an NVIDIA GeForce RTX 5090 with 32 GB graphics memory, an AMD Ryzen 9 7950X CPU, and 64 GB system RAM. GPU-intensive feature extraction and training jobs are run sequentially to reduce out-of-memory conflicts, while lightweight manifest checks, audits, and report assembly can run independently. This separation makes it easier to recover from a failed stage without repeating unrelated computation.

5.22 Branch Tensor Interfaces and Feature Storage

A multimodal system becomes easier to debug when each branch has a clear tensor interface. The Face pathway receives standardized visual evidence and returns a face-level embedding plus probability and subbranch attention information. The Voice pathway receives time-frequency arrays and returns a fixed-dimensional acoustic embedding plus p(fake). The Temporal pathway receives an ordered sequence of face-derived representations and returns a sequence-level embedding. Lip-Sync returns a cross-modal synchronization representation, while Background returns a context representation derived from full-frame encoders. These interfaces allow the fusion model to treat each branch as a modular evidence producer.

Intermediate features are stored when doing so reduces repeated computation and improves reproducibility. Precomputed face crops, MFCC arrays, Log-Mel arrays, CLIP embeddings, DINOv2 embeddings, and other branch artifacts can be regenerated from the canonical registry, but retaining versioned outputs makes repeated model training faster and provides an audit trail. Each stored artifact must therefore be linked to case ID, preprocessing version, and split assignment. A feature without provenance is not considered equivalent to a reproducible feature even if its tensor shape is correct.

Shape and ordering checks are performed before fusion because a silent interface mismatch can produce scientifically invalid results without raising an obvious software error. Temporal frames must remain ordered, audio arrays must correspond to the correct case, and background frames must not be accidentally substituted for face crops. The final methodology therefore treats interface validation as part of the experiment rather than as routine programming detail.

5.23 Stage-by-Stage Quality Gates

Each major pipeline stage has a quality gate that must be satisfied before the next stage is considered complete. Registry construction checks file existence, source identity, labels, and duplicate or ambiguous records. Media validation checks that files can be opened and decoded. Face preprocessing checks whether a valid face crop is available. Audio preprocessing checks whether the track can be extracted and whether valid speech features can be produced. Temporal preparation checks ordered frame availability. Lip-Sync preparation checks aligned mouth-audio windows. Background preparation checks that the expected original and face-suppressed frame sets exist. These gates prevent later training code from silently operating on malformed inputs.

Training stages have corresponding model-level gates. A branch must produce a valid checkpoint, validation predictions, and expected artifact counts before its outputs are accepted for fusion. Fusion must verify modality masks and case alignment before training. Calibration must receive validation logits from the locked fusion checkpoint. Final TEST inference is permitted only when the preceding locks are satisfied. This staged contract is particularly important in a long-running thesis project because a missing file or partially completed branch can otherwise propagate into the final metrics and become difficult to diagnose later.

Quality gates also support efficient recovery. If a background embedding job fails halfway through, the pipeline can resume from the missing artifact list rather than reprocessing faces, audio, or temporal sequences. If one model checkpoint is corrupted, the failure can be isolated to that branch. The methodological benefit is both computational and scientific: fewer unnecessary reruns mean fewer opportunities for accidental configuration drift.

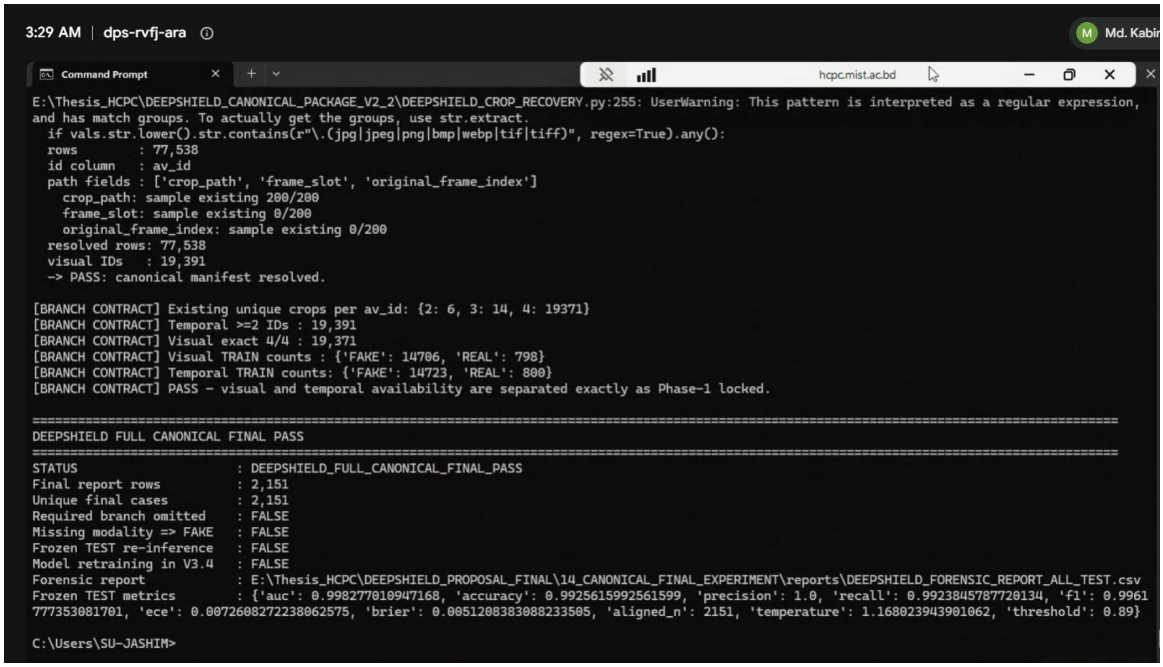


Figure 5.7: Canonical visual-crop recovery and branch-contract verification used in the DeepShield quality-control pipeline.

5.24 Computational Sequencing and Resource-Aware Execution

The order of execution is designed around both dependency and hardware cost. Low-cost registry audits and split validation are completed before expensive GPU work begins. Face crops and other reusable preprocessing artifacts are generated before multiple visual models are trained. Independent branch feature extraction can be scheduled separately, but global fusion is delayed until branch outputs and availability masks are complete. Calibration and final reporting occur only after the fusion checkpoint is locked. This dependency order ensures that expensive computation is not wasted on a dataset or split that later fails an audit.

Resource-aware sequencing also reduces the risk of inconsistent partial runs. Large visual backbones, GenConViT variants, CLIP, DINOv2, and recurrent models can compete for GPU memory if launched simultaneously. The documented workflow therefore executes heavy jobs sequentially and uses checkpoint/resume logic. CPU, storage, and I/O considerations are equally important because repeated decoding of hundreds of thousands of frames can become a bottleneck even when the GPU is available. Precomputation is used where it improves repeatability and reduces repeated media access.

This implementation strategy is part of the methodological contribution because the final model could not be evaluated reliably without a tractable pipeline. Deepfake research at this scale is an engineering system as well as a machine-learning experiment. The thesis therefore documents not only which model is used but also how the sequence of computation protects the validity and recoverability of the experiment.

5.25 Failure Recovery, Resume Logic, and Non-Destructive Processing

All raw datasets are treated as source evidence and are not modified by training notebooks. Derived artifacts are written to separate locations so that a failed experiment cannot corrupt the original media. Resume logic works from manifests, completion markers, or existing artifact checks rather than assuming that a directory is complete merely because it exists. This is particularly important for large feature-extraction stages that may run for many hours and can be interrupted by storage, scheduler, operating-system, or library failures.

Recovery is deterministic wherever possible. When a case is retried, the same frame-selection rule, crop configuration, audio parameters, and model checkpoint should reproduce the same input contract. If a software or preprocessing version changes intentionally, the outputs are written under a new version rather than overwriting the earlier evidence. Historical and final experimental branches can therefore coexist without ambiguity, which is essential in this thesis because the project evolved from broad visual-temporal experiments to the final aligned five-branch design.

Non-destructive processing also supports academic audit. A reviewer can inspect the relationship between the raw source, the derived artifact, and the final prediction. If a result appears unusual, the case can be traced and regenerated without depending on an undocumented notebook state. This makes the methodology stronger than a one-time script that produces a final CSV but cannot explain how each row was obtained.

5.26 Final Forensic Output Contract

The end product of the methodology is a case-level forensic record rather than a bare label. At minimum, the record identifies the canonical case, the frozen split, the calibrated final probability, the locked threshold decision, and the availability of Face, Voice, Temporal, Lip-Sync, and Background evidence. Where available, branch scores, attention weights, and face-region XAI are linked to the same case. This design makes the final output suitable for error analysis, ablation review, and human interpretation.

The output contract also distinguishes model evidence from ground truth. The prediction fields record what DeepShield inferred, while target fields remain separate and are used only for evaluation. This separation is crucial for frozen TEST reporting because it prevents post-processing code from overwriting or deriving targets from predictions. Report-finalization scripts may aggregate or format already generated evidence, but they do not retrain models or rerun the frozen TEST inference.

A complete forensic record therefore answers four questions: what case was evaluated, what evidence was available, what did the locked system predict, and what explanatory information can support review? By preserving these elements together, DeepShield converts a complex multimodal pipeline into an auditable experimental object whose final metrics can be reproduced from case-level outputs.

CHAPTER 6
TRAINING, VALIDATION AND ABLATION

6.1 Experimental Environment

The final experiments were executed on a Windows HCPC environment equipped with an NVIDIA GeForce RTX 5090 with 32 GB graphics memory, AMD Ryzen 9 7950X processor, and 64 GB system memory. The verified software stack included Python 3.12, PyTorch 2.12.0 with CUDA 13.0, Torchvision 0.27.0, FFmpeg, official GenConViT source, a vendored timm 0.6.5 environment for model compatibility, OpenCLIP, and a local DINOv2 torch hub cache. GPU-intensive jobs were executed sequentially to avoid memory conflicts, while lightweight audits and preprocessing checks were run separately.

6.2 Training and Model-Selection Protocol

Training used the TRAIN split only. Validation ROC-AUC was the principal branch-selection signal where applicable. Early stopping and best-checkpoint locking were used to prevent later runs from silently overwriting stronger checkpoints. The TEST split was not used for checkpoint selection. A later finalization script repaired report assembly without retraining or re-running frozen TEST inference.

6.3 Historical ConvNeXt Image Model Training

The ConvNeXt-Tiny image branch is fine-tuned for binary classification using MTCNN face crops. The training objective is cross-entropy over real and fake labels. Transfer learning initializes the backbone from pretrained weights so that optimization begins from a general visual representation rather than random parameters. Data loading is source-aware and uses the frozen training manifest.

The exact optimizer schedule, batch size, and epoch count should be interpreted from the archived training configuration rather than reconstructed from memory. This thesis therefore focuses on the settings that are present in the locked result evidence: branch identities, deterministic sampling rules, model-selection criteria, fusion configuration, threshold, calibration temperature, validation/ablation values, and frozen TEST metrics. This approach avoids inventing hyperparameters that are not contained in the supplied result material.

6.4 Historical Visual-Temporal Model Training

For every training video, four face crops are encoded by ConvNeXt-Tiny into an ordered feature sequence. The BiLSTM processes the sequence bidirectionally, temporal attention aggregates the hidden states, and a classification head produces the video-level logit. Training minimizes binary classification loss at video level.

The use of embeddings rather than raw frames reduces temporal-model complexity. The sequence network does not need to learn low-level edge and texture filters; those are already represented by ConvNeXt. This makes the temporal branch more efficient and focuses its capacity on relationships across sampled frames.

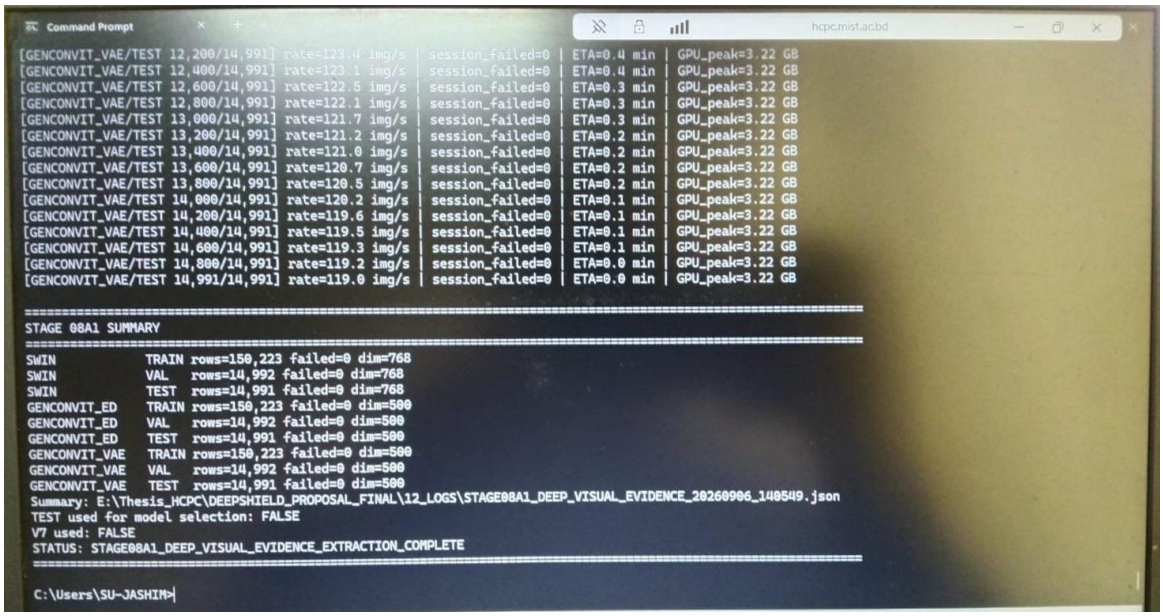


Figure 6.1: Completion audit of historical Swin and GenConViT visual evidence extraction across the development splits.

6.5 Final Aligned Branch Training Scope

The current canonical V3.x Face/Frequency/Temporal adaptation used the aligned cohort rather than the entire 180,206-sequence historical visual corpus. Visual TRAIN included 15,504 usable samples (14,706 FAKE and 798 REAL), Temporal TRAIN included 15,523 samples (14,723 FAKE and 800 REAL), and Voice TRAIN included all 17,234 aligned cases according to the audio-label contract. This branch-specific accounting is essential for reproducibility.

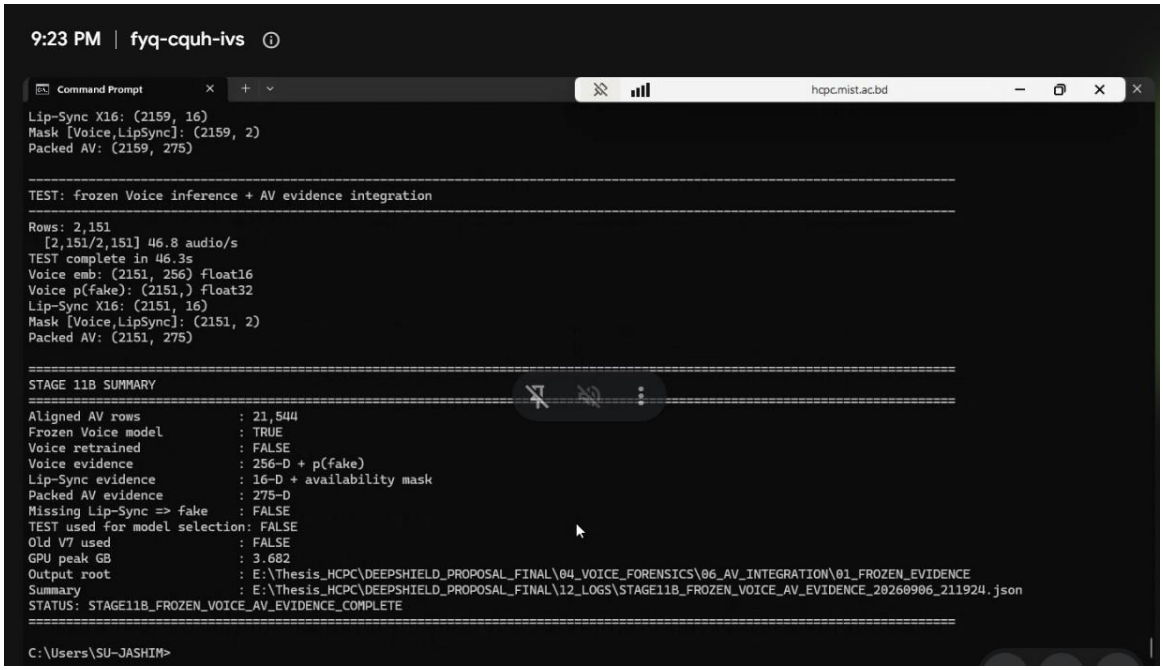


Figure 6.2: Frozen Voice and Lip-Sync evidence integration over the final aligned audiovisual cohort.

6.6 Historical Visual-Temporal Model Selection

The historical visual-temporal video checkpoint was selected by validation ROC-AUC. This criterion evaluates ranking performance and is less dependent on a particular class threshold than accuracy. Selecting by validation rather than test performance is important because repeated test-set comparison would bias the final estimate.

For the historical visual-temporal checkpoint, the decision threshold was fixed at 0.531 using validation information and applied unchanged to its historical internal test set. In the final canonical five-branch fusion, the validation-locked decision threshold is 0.89 and is applied unchanged to the 2,151-case frozen TEST set. The same final 0.89 rule is used to compute accuracy, precision, recall, F1, and the reported confusion matrix.

6.7 Final Fusion Temperature Calibration

Temperature scaling fits a single scalar parameter on validation logits. The learned value in the final system is 1.168024. Because this parameter is estimated without access to final test labels, the calibrated probabilities can be evaluated on the test population without direct label leakage.

Calibration is especially relevant when a model achieves very high discrimination. A nearly perfect ROC-AUC can coexist with poorly calibrated probabilities. Consequently, DeepShield reports the temperature value as part of the locked configuration and interprets probabilities as model outputs rather than absolute certainty.

6.8 Evaluation Metrics

Accuracy measures the fraction of all videos that receive the correct class label. Precision measures the fraction of predicted fake videos that are truly fake. Recall measures the fraction of fake videos that are detected. F1-score is the harmonic mean of precision and recall and is useful when both missed deepfakes and false accusations are important. ROC-AUC measures the probability that a randomly chosen fake sample is ranked above a randomly chosen real sample across all possible thresholds.

Because deepfake detection can have asymmetric costs, no single metric is sufficient. High recall reduces missed manipulated media, while high precision reduces false alarms. ROC-AUC evaluates ranking independently of the chosen threshold, and confidence intervals indicate how precisely the test population supports the estimated performance.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall$$

6.9 Validation Results

Model / Branch	Best Validation ROC-AUC
Swin	0.999001083
GenConViT-ED	0.999312398
GenConViT-VAE	0.999014618
Frequency	0.828129399
Face Fusion	0.999301570
Temporal	0.584323593
Background CLIP	0.979980573
Background DINOv2	0.890706654
Background Combined	0.976270034
Background Masked	0.915220981
Five-Branch Fusion	0.995647226

Table 6.1: Validation branch results

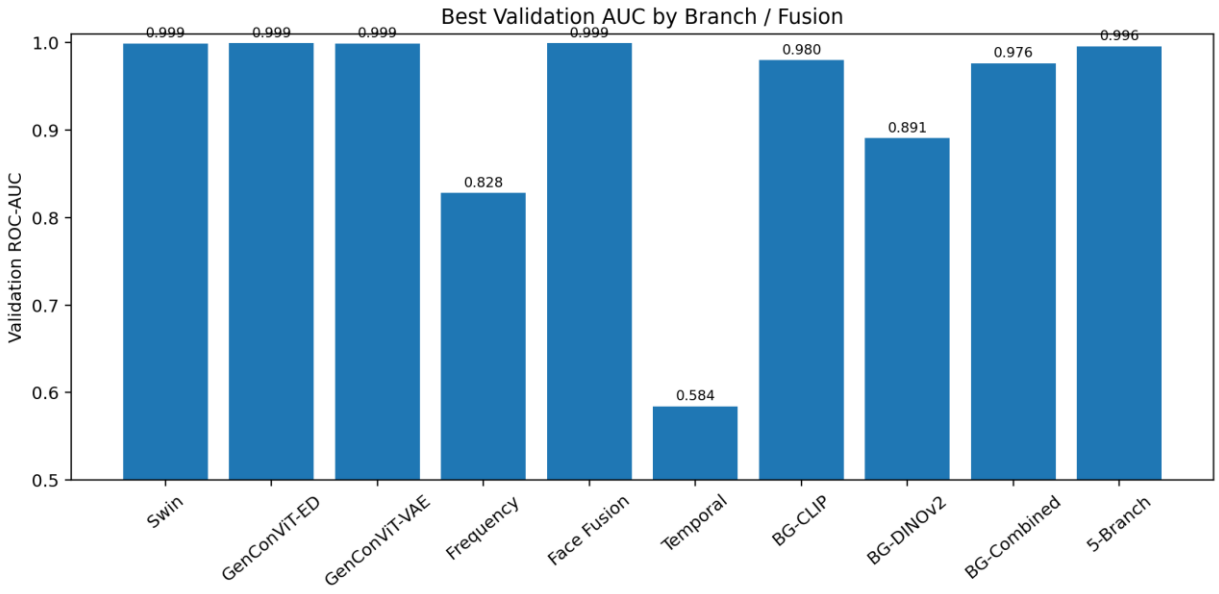


Figure 6.3: Best validation AUC by branch/fusion

The visual Face subbranches and learned Face fusion achieved very high validation separation. The Frequency branch was materially weaker than the deep visual subbranches but still contributed a distinct evidence domain. Temporal standalone AUC was comparatively weak, indicating that the temporal branch should be treated as supporting evidence. The Background branch showed strong CLIP and combined context performance, while DINOv2 and masked-background variants were lower.

6.10 Five-Branch Fusion

The five-branch fusion reached validation ROC-AUC 0.994993 in epoch 1, 0.995557 in epoch 2, and 0.995647 in epoch 3, which was the best value observed in the recorded run. The full fusion is not expected to dominate every individual validation ablation because adding a modality can introduce redundant or noisy evidence while still contributing explanatory or complementary information in other cases.

6.11 Historical Confidence Intervals

The historical visual-temporal result sheet included 95% confidence intervals for accuracy, F1-score, and ROC-AUC. Confidence intervals communicate the variability expected from sampling an evaluated population rather than reporting only a single percentage. The narrow historical intervals were consistent with the 14,991-video historical visual TEST population, whereas the final canonical 2,151-case multimodal report is represented in this thesis by its exact locked metrics, calibration values, and confusion-matrix counts without inventing unreported final confidence intervals.

Confidence intervals should still be interpreted within the evaluation design. If samples are correlated because several videos share identities or manipulation processes, the

nominal interval may not fully represent domain uncertainty. For this reason, the results chapter distinguishes statistical precision on the internal population from generalization to new sources.

6.12 Explainability Procedure

Grad-CAM is generated for representative visual predictions after the classifier is trained. The heatmap is overlaid on the face crop so that an analyst can inspect whether the model emphasizes semantically plausible facial regions. The explanation is stored together with prediction metadata and is not used to change the final locked test labels.

Explainability examples should include correct and incorrect predictions whenever available. Restricting a thesis to successful heatmaps can create selection bias. The purpose of XAI is not only to justify the model but also to reveal possible shortcuts or failure modes.

6.13 Standalone Ablation and Generalization Experiments

Frequency-domain and DINOv2 experiments are categorized as auxiliary ablation/generalization studies. Ablation asks whether a component adds value relative to a simpler configuration. Generalization analysis asks whether a representation behaves robustly when the data distribution changes. These are related but distinct questions.

Standalone frequency-domain and DINOv2 experiments are categorized as auxiliary ablation/generalization studies, while frequency-domain evidence also contributes to the final Face branch and DINOv2 contributes to the final Background branch. The thesis therefore uses the standalone experiments to motivate robustness and model-selection rationale without denying their documented roles inside the canonical multimodal architecture. This distinction also enables a fairer interpretation of computational complexity: standalone variants and final fused representations are reported according to their actual experimental roles.

6.14 Internal Validity Controls

Internal validity is protected through deterministic frame sampling, frozen manifests, validation-only model selection, fixed thresholding, validation-only temperature calibration, checkpoint locking, and explicit separation of auxiliary ablation models. These controls reduce the number of hidden degrees of freedom that can inflate test performance.

The training status is recorded as COMPLETE / LOCKED in the final result evidence. This means that the quantitative values in Chapters 6 and 7 should not be modified by further tuning on the same frozen TEST set. Any later architecture revision should create a new experiment version and a new untouched evaluation population.

6.15 Ablation Study

Configuration	Best VAL AUC	Interpretation
Full five-branch	0.995647226	Reference configuration
No Face	0.983641536	Largest degradation; Face is highly influential
No Voice	0.987083926	Voice contributes useful complementary evidence
No Temporal	0.994044571	Small degradation; Temporal standalone is weak
No Lip-Sync	0.997088668	Higher than full on VAL; indicates redundancy/noise in this validation setting
No Background	0.995040303	Small but measurable reduction

Table 6.2: Five-branch validation ablation

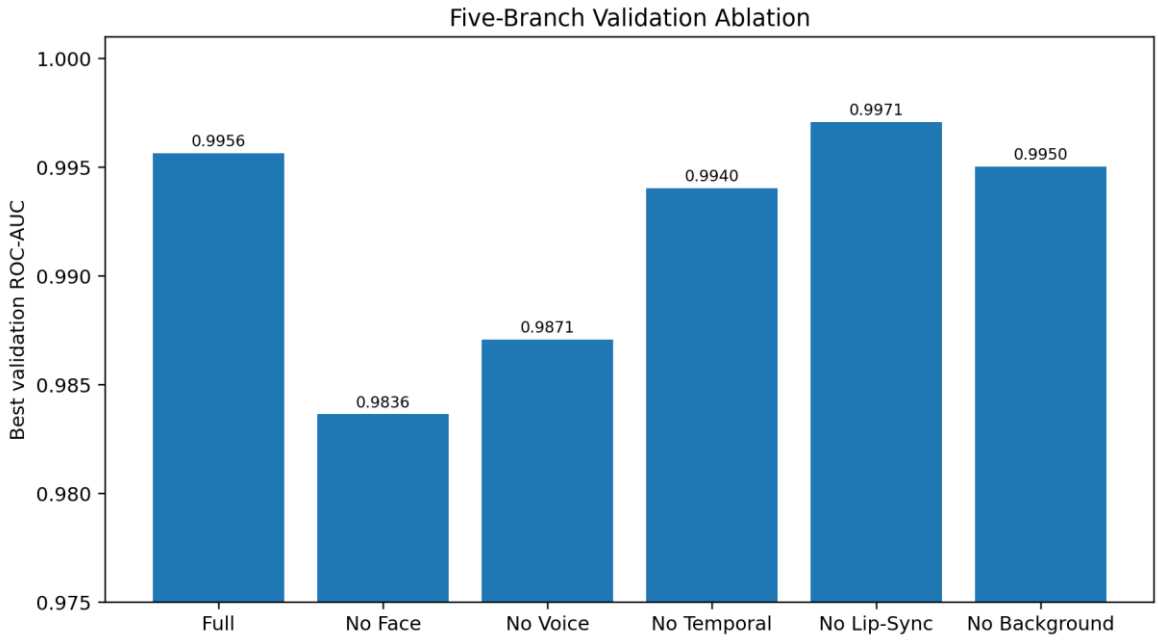


Figure 6.4: Five-branch validation ablation

The no-Lip-Sync ablation slightly exceeded the full fusion on validation. This does not prove that Lip-Sync is useless. it indicates that, for this validation distribution and training state, the branch did not improve discrimination and may have introduced redundant or noisy evidence. Reporting this result is important because ablation should

test the actual contribution of a component rather than automatically justify every proposed branch.

6.16 Validation-Only Calibration

After branch and fusion checkpoints were locked, calibration was fitted on validation data only. Temperature scaling used a single scalar and the decision threshold was locked before the frozen TEST stage. This separates ranking performance (ROC-AUC), decision performance (accuracy, precision, recall, F1), and probability calibration (ECE and Brier score).

6.17 Optimization Logic and General Training Strategy

The training strategy is intentionally branch-aware. Each component learns from the evidence and label contract appropriate to its modality before the global fusion stage is optimized. This reduces the risk that one high-capacity branch dominates early learning and prevents label semantics from being mixed across modalities. Validation ROC-AUC is used as a principal checkpoint-selection signal where appropriate because it evaluates ranking without committing to the final operating threshold.

Optimization settings such as exact learning rate schedules, batch sizes, and epoch counts are tied to the archived run configurations. The thesis does not reconstruct values that are not present in the final evidence package. This makes the methodological record slightly less exhaustive than a fully scripted publication repository, but it is scientifically safer than reporting guessed hyperparameters. The locked checkpoints, validation outputs, and final fusion artifacts define the evaluated system.

6.18 Class Imbalance and Metric-Aware Training Interpretation

The aligned cohort contains substantially more FAKE than REAL cases. Imbalance can produce a misleadingly high accuracy if a classifier simply follows the majority class, which is why training and evaluation must be interpreted through multiple metrics. Precision measures whether predicted FAKE cases are correct, recall measures how many actual FAKE cases are found, and F1 balances the two. ROC-AUC evaluates ranking quality independently of a fixed threshold.

In the final TEST result, precision is 1.0 and no REAL case is falsely classified as FAKE, while 16 FAKE cases are missed. This error profile is more informative than the accuracy alone. During model development, the class composition of branch-specific usable subsets is therefore reported explicitly so that readers can understand the data seen by each component and avoid assuming that all 21,544 cases are present in every branch training batch.

6.19 Checkpointing, Early Stopping, and Run Management

Long multimodal experiments are vulnerable to interruption, accidental overwrite, and selection bias. DeepShield uses best-checkpoint locking so that a later epoch cannot silently replace a stronger validation checkpoint. Early stopping or validation-based selection is applied at the branch or fusion level where relevant. GPU-intensive jobs are executed sequentially in the documented HCPC environment to reduce memory contention, while audits and preprocessing checks can run independently.

Run management also separates training from report finalization. The later finalization stage repaired or assembled report artifacts without retraining models or re-running frozen TEST inference. This distinction matters because changing report code after a final evaluation is not equivalent to changing the model. The thesis therefore records that the final 2,151-case forensic report was assembled from locked inference outputs.

6.20 Test Lock, Calibration Lock, and Experimental Finalization

The final experiment uses a layered lock procedure. First, branch checkpoints are selected from TRAIN and VAL evidence. Second, fusion is trained and its checkpoint is selected without TEST labels. Third, calibration temperature and decision threshold are fitted or selected using validation outputs. Fourth, the TEST split remains frozen until all these choices are complete. Fifth, the final inference is executed and the resulting prediction set is used for the reported metrics and XAI aggregation.

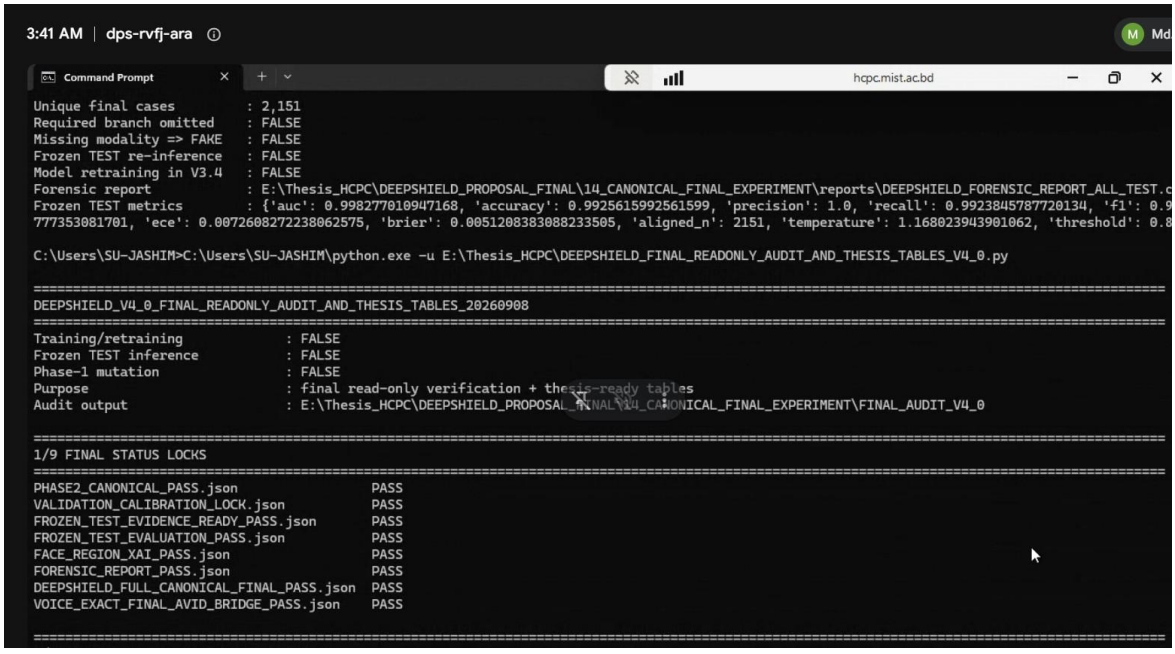


Figure 6.5: Final experiment-lock audit confirming calibration, frozen TEST, XAI, forensic-report, and canonical pipeline status.

This ordering protects the interpretation of the final TEST score. If the threshold were repeatedly changed after inspecting TEST errors, the final accuracy and F1 would become tuned to that population. If calibration were fitted on TEST labels, the ECE and Brier score would no longer represent held-out calibration behavior. The lock procedure therefore provides a methodological basis for calling the final result a frozen internal evaluation.

6.21 Statistical Interpretation of the Final Evaluation

A large sample count can make point estimates numerically stable, but statistical interpretation must still consider class imbalance and case dependence. The final aligned TEST contains 2,151 cases, of which 2,101 are FAKE and 50 are REAL according to the confusion counts. Because the REAL class is much smaller, specificity can change materially with even a small number of false positives, while recall reflects performance across the much larger FAKE population.

Confidence intervals and calibration metrics should therefore be interpreted within the evaluated cohort rather than as a statement about all possible internet media. Cases may share acquisition conditions, identities, or manipulation families even when the split is leakage-controlled. External datasets and unseen generator families are necessary to estimate domain generalization. The internal evaluation answers whether the locked system performs strongly on its defined aligned population; it does not eliminate uncertainty about deployment shift.

6.22 Training and Validation Reproducibility Summary

The training chapter demonstrates that reproducibility is produced by the interaction of data controls and model controls. Frozen split manifests prevent accidental resampling; deterministic preprocessing stabilizes inputs; branch-specific label contracts protect modality semantics; checkpoints preserve selected model states; validation metrics define model selection; ablation tables document comparative evidence; and locked calibration parameters separate probability adjustment from final testing.

Together, these controls make the final result auditable even though the project includes historical stages and multiple model families. The thesis can therefore preserve the earlier ConvNeXt experiments as developmental evidence while identifying the later five-branch aligned experiment as the canonical endpoint. This distinction is essential for avoiding contradictory result claims across chapters.

CHAPTER 7

RESULTS, ANALYSIS AND DISCUSSION

7.1 Final Canonical Status

The final canonical experiment completed with all required branches retained, missing-modality-to-FAKE disabled, and the frozen TEST evaluated only after training, model selection, fusion, and calibration locks. The final forensic report contains 2,151 unique TEST cases. The report-finalization stage did not retrain models and did not rerun frozen TEST inference.

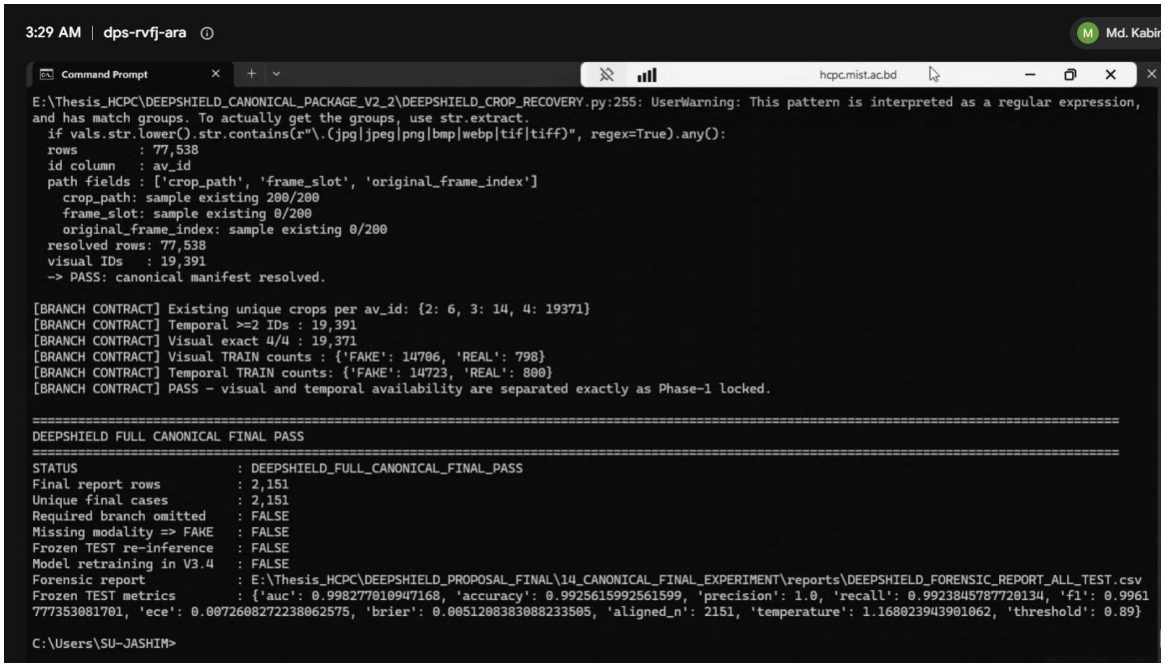


Figure 7.1: DeepShield canonical final-pass verification for the 2,151-case frozen TEST evaluation.

7.2 Final Frozen Internal TEST Performance

Metric	Locked Result
ROC-AUC	0.9982770109
Accuracy	0.9925615993
Precision	1.0000000000
Recall	0.9923845788
F1-score	0.9961777353
Expected Calibration Error	0.0072608272
Brier Score	0.0051208383
Aligned TEST cases	2,151
Temperature	1.1680239439
Decision threshold	0.89

Table 7.1: Final frozen TEST metrics

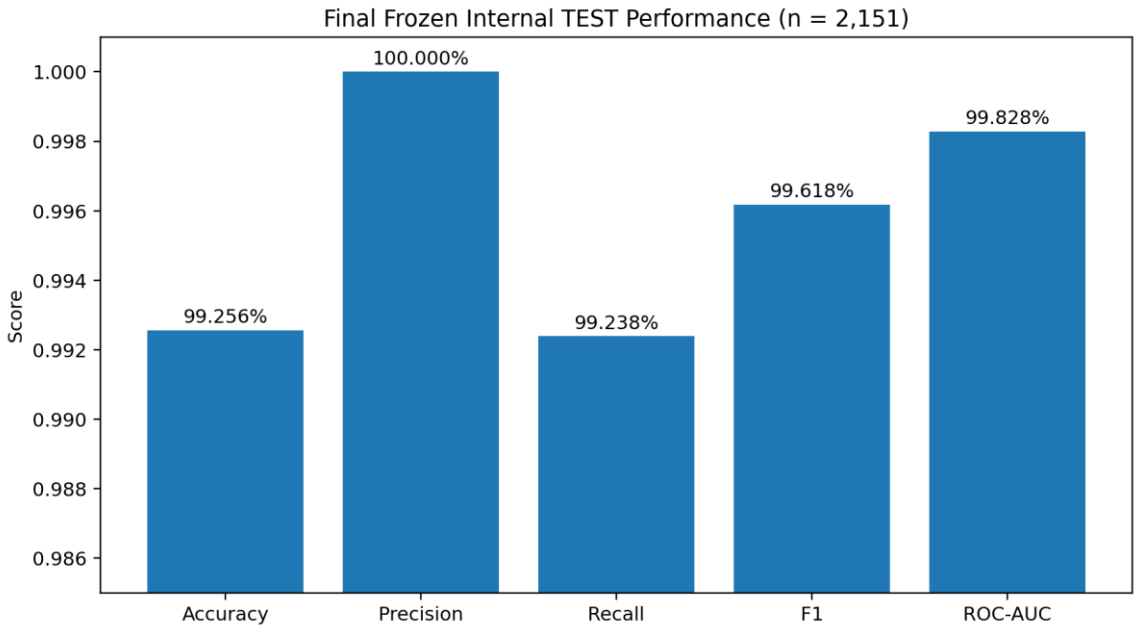


Figure 7.2: Final frozen internal TEST performance

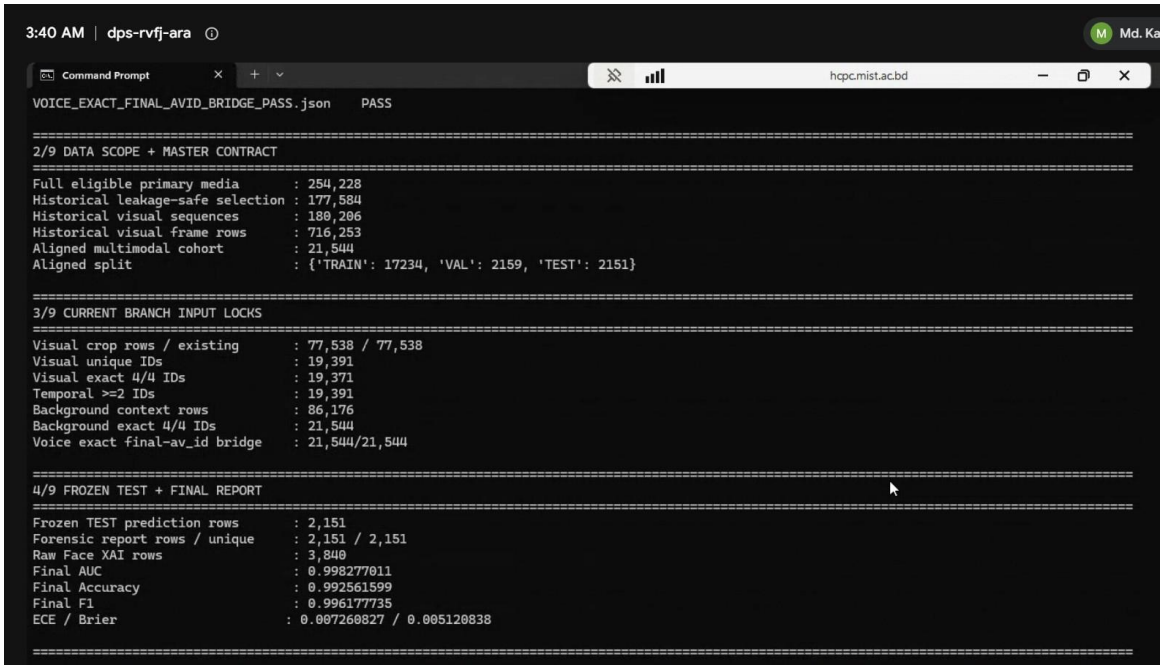


Figure 7.3: Read-only verification of the final frozen TEST cohort and locked DeepShield performance metrics.

7.3 Confusion Matrix Interpretation

The locked metrics correspond to 2,151 final cases. With precision equal to 1.0 and 16 total errors, the resulting confusion counts are TP = 2,085, TN = 50, FP = 0, and FN = 16 for the final global target contract. The TEST set is highly imbalanced toward the FAKE class in this aligned cohort. so ROC-AUC, F1, recall, specificity, calibration, and the underlying split composition are more informative than accuracy alone.

Actual / Predicted	FAKE	REAL
FAKE	2,085	16
REAL	0	50

Table 7.2: Confusion matrix of final frozen TEST

7.4 Calibration Quality

The ECE of 0.007261 and Brier score of 0.005121 indicate that the calibrated probabilities were well aligned with outcomes on the locked internal TEST population. These metrics are particularly useful because an excellent ROC-AUC does not guarantee probability calibration. The locked temperature of 1.1680 moderates the raw fusion logits before the final decision rule is applied.

Parameter	Locked Value	Purpose
Temperature	1.1680239439	Validation-fitted probability calibration
Threshold	0.89	Validation-locked FAKE decision threshold
Missing modality policy	Mask unavailable branch	Prevents unavailable evidence from being treated as FAKE
TEST use	Single frozen evaluation	Prevents test-driven model selection

Table 7.3: Calibration and decision configuration

7.5 Branch-Level Interpretation

The Face branch was the dominant source of useful evidence, consistent with the strong Swin, GenConViT, and Face-fusion validation AUC values. Background context also provided useful discrimination, especially through CLIP. Temporal evidence was weaker in isolation and should so be treated as supportive rather than primary. Voice contributed measurable validation value in the ablation study. Lip-Sync did not improve the reported validation AUC in the full fusion relative to the no-Lip-Sync ablation, so its role should be discussed as complementary evidence whose contribution may depend on data distribution and synchronization quality.

7.6 Historical Spatial-Temporal Design Interpretation

The strong locked performance supports the usefulness of combining ConvNeXt spatial representation with sequence-level reasoning. ConvNeXt provides a high-capacity feature space for each face, while BiLSTM and temporal attention allow the classifier to interpret those features jointly. This is consistent with prior research showing that temporal identity and spatiotemporal representations can complement frame-based detection [8]-[12].

The result does not isolate how much improvement comes from the BiLSTM, attention layer, or ConvNeXt backbone because the supplied final evidence does not include a full numerical ablation table. The thesis therefore avoids claiming a specific percentage gain for each component. A future paper-quality ablation should compare ConvNeXt frame voting, ConvNeXt + LSTM, ConvNeXt + BiLSTM, and ConvNeXt + BiLSTM + temporal attention under the same split.

7.7 Model Selection and Threshold Behavior

Selecting the final fusion checkpoint by validation evidence is appropriate when ranking quality is important and when the class threshold may be adjusted separately. The final validation-locked threshold is 0.89, indicating a deliberately conservative FAKE operating point for the calibrated fusion output.

In deployment, the ideal threshold depends on the cost of errors. A forensic triage tool may prefer higher recall so that suspicious media are not missed, whereas an automated moderation system may prioritize precision to reduce false accusations. The fixed 0.89 threshold is therefore best interpreted as the locked final experimental operating point, not a universal policy.

7.8 Temperature Calibration

The learned final temperature of 1.1680239439 softens logits before probability conversion. Because the value is greater than one, the calibration step generally reduces extreme probability magnitudes. This is compatible with the reliability-aware design goal: a model can retain its class ranking while reporting probabilities that are less overconfident.

The final evidence supplied for this book includes Expected Calibration Error 0.0072608272 and Brier score 0.0051208383. The thesis therefore reports the temperature parameter together with quantitative calibration metrics rather than fabricating missing values. Future evaluation should additionally include reliability diagrams, adaptive calibration error, and external-domain calibration analysis.

7.9 Explainability Output

The final XAI pipeline produced face-region localization and modality/subbranch contributions. Raw Face XAI contained 3,840 rows because some cases had multiple region candidates. the final forensic report aggregated these deterministically to 2,151 unique cases by selecting the strongest case-level region evidence and reranking. This preserves traceability without duplicating final cases.

7.10 Auxiliary Frequency and DINOv2 Findings

The final configuration uses frequency-domain evidence inside the Face branch and DINOv2 inside the Background branch, while standalone frequency and DINOv2 variants are also retained for ablation/generalization interpretation. This outcome is important because the earlier proposal and historical visual-temporal pathway assigned these directions different roles. The locked architecture demonstrates that the team ultimately prioritized a mask-aware five-branch fusion for the production result.

Frequency-aware detection remains valuable for unseen-generator research [18], [19], and DINOv2 remains a strong general-purpose visual representation [17]. Their exclusion from the locked model should not be interpreted as evidence that they are ineffective. It means only that the final reported detector and metrics do not depend on them.

7.11 Comparison with Related Research

Direct numerical comparison between deepfake papers is difficult because datasets, splits, frame sampling, compression levels, and identity protocols differ. The literature reviewed in Chapter 2 shows a broad range of spatial, temporal, and multimodal architectures [6]-[16]. DeepShield contributes a strong internal result, but it would be scientifically incorrect to claim superiority over every published method solely because its internal percentage is higher.

A fair comparison would require running competing models under the same preprocessing, train/validation/test split, and evaluation code. DeepfakeBench provides motivation for this type of standardized comparison [6]. The current thesis therefore uses literature comparison to contextualize design choices rather than to declare state-of-the-art status.

7.12 Reliability and Forensic Interpretation

High internal accuracy can create a temptation to use the detector as a definitive authenticity oracle. This should be avoided. Reliability-oriented deepfake research emphasizes that uncertainty can arise from low-quality inputs, unseen generators, distribution shift, and model overconfidence [22]. DeepShield addresses one aspect through temperature calibration and transparent reporting of the threshold and test scope.

For forensic use, the appropriate interpretation is that the locked model is a high-performing screening system on its evaluated internal population. A final human decision should consider source provenance, metadata, audio evidence, platform context, and independent verification. Grad-CAM may support this review but does not convert a prediction into proof.

7.13 Threats to Validity

The first threat is distribution shift. Internal testing may still share acquisition patterns with training data even when the individual files are held out. External testing on unseen datasets and unseen generators is required to estimate broader generalization. The second threat is source correlation. If many derived videos originate from related sources, the nominal sample size may overstate the number of independent content origins.

The third threat is preprocessing dependence. Face detection and cropping determine which pixels the visual branches observe. A deployment system using a different detector or crop geometry may not reproduce the same score distribution. The fourth threat is multimodal domain coverage: the locked result sheet reports a final five-branch fusion and branch-level validation evidence, but the aligned internal cohort does not establish performance for every unseen voice-cloning method, audiovisual manipulation type, or deployment platform. This limits the claims that can be made about universal multimodal generalization.

The fifth threat is temporal sparsity. Four deterministic frames provide efficiency and reproducibility but may miss brief manipulations. Finally, explainability remains qualitative because no ground-truth explanation metric is reported.

7.14 Internal Validity and Scope of Claims

The reported performance is an internal locked result for the aligned multimodal cohort. The 254,228-item primary registry, 177,584 leakage-safe historical selection, 180,206 historical visual sequences, and 21,544 final aligned multimodal cases are different scopes and should not be conflated. The final V3.x Face/Frequency/Temporal adaptation was performed on usable subsets of the aligned cohort, while the broader historical visual experiments show earlier large-scale processing and model evidence extraction. Accordingly, this thesis does not claim universal cross-dataset generalization or state-of-the-art superiority solely from the internal score.

7.15 Forensic Use and Ethical Interpretation

DeepShield should be treated as an automated deepfake checking system. A model score or attention map is not legal proof that a person created or manipulated media. False positives can harm reputation and trust, while false negatives can allow manipulated content to pass. For practical use, model outputs should be combined with source record checks, metadata examination, source verification, and human expert review.

7.16 Detailed Error Profile

The final confusion matrix contains 2,085 true positives, 50 true negatives, zero false positives, and 16 false negatives. The absence of false positives is consistent with precision equal to 1.0, while the 16 missed FAKE cases determine the recall of approximately 0.992385. This error profile suggests that the selected threshold is conservative with respect to accusing REAL media of being FAKE on the evaluated cohort.

The false negatives are therefore the most important cases for qualitative follow-up. They may correspond to highly realistic manipulations, weak or missing branch evidence, short manipulation intervals, difficult compression, atypical speakers, poor face quality, or

cases in which the branch attention emphasizes non-discriminative evidence. A rigorous future error study should stratify these 16 cases by manipulation type, branch availability, confidence, dataset source, face quality, audio quality, and attention distribution rather than treating them as a homogeneous group.

7.17 Metric Interpretation under Strong Class Imbalance

The TEST set is heavily skewed toward FAKE cases, so accuracy must be read together with class-specific behavior. An accuracy of 0.992562 is strong, but the confusion matrix reveals why it is strong: the system correctly identifies all 50 REAL cases and misses only 16 of 2,101 FAKE cases. Precision, recall, F1, and ROC-AUC therefore provide a more complete description than accuracy alone.

ROC-AUC of 0.998277 indicates near-perfect ranking over the internal cohort across possible thresholds. F1 of 0.996178 shows that the selected threshold maintains a strong balance between precision and recall despite the class skew. The result should nevertheless not be interpreted as proof that the same operating point will behave identically on a balanced external dataset, where the number and diversity of REAL cases may be much greater.

7.18 Branch Interaction and Ablation Interpretation

The validation results show that the Face subbranches are individually strong, Background CLIP also carries substantial signal, Voice contributes useful ablation value, Temporal is weaker in isolation, and Lip-Sync does not improve the reported full-fusion validation AUC in that run. This pattern illustrates why multimodal fusion should not be interpreted as a simple rule that more branches always increase the score.

A learned fusion model can benefit from complementary evidence even when one branch has lower standalone AUC, but it can also be affected by redundancy, noise, missingness, or limited branch calibration. The correct scientific interpretation is therefore case-dependent and ablation-driven. The final architecture demonstrates successful combined performance, while the branch results identify which components deserve further refinement before an external deployment study.

7.19 Operational Meaning of the Threshold and Calibrated Score

The locked threshold of 0.89 is an experimental operating point chosen from validation evidence; it is not a universal definition of deepfake probability. In a different deployment context, the cost of false positives and false negatives may differ. A journalistic verification workflow may tolerate more false alarms in exchange for higher recall, whereas an automated accusation or removal system may require extremely high precision.

The calibrated probability should likewise be understood as a model-based confidence estimate under the learned internal distribution. The low ECE and Brier score indicate good alignment on the frozen internal TEST, but calibration can deteriorate under domain shift. A responsible deployment would therefore retain an uncertainty or abstention policy, monitor input quality, and periodically evaluate calibration on new operating data rather than assuming the internal temperature remains optimal forever.

7.20 Answers to the Research Questions

The final experiment answers the thesis research questions in a more complete form than the earlier visual-only development stages. First, strong facial representations are clearly useful, as shown by the high validation AUC of Swin, GenConViT variants, and Face fusion. Second, temporal modeling provides supporting sequence evidence, although its standalone validation AUC is weaker and does not justify describing it as the dominant source of performance. Third, validation-only checkpoint selection, thresholding, and temperature scaling provide a disciplined separation between model development and final testing.

Fourth, the final locked five-branch system achieves ROC-AUC 0.998277, accuracy 0.992562, precision 1.0, recall 0.992385, and F1 0.996178 on 2,151 frozen TEST cases, with low ECE and Brier score. Fifth, auxiliary and branch-specific ideas should be reported according to measured evidence: strong visual components can be described as primary, weaker or neutral branches can be described as complementary, and historical experiments should remain distinct from the final canonical result.

7.21 Deployment and Human-Review Workflow

A practical DeepShield deployment would begin with automated ingestion and quality checks, followed by branch inference and a calibrated fusion score. Cases with high-confidence FAKE predictions could be prioritized for review, while uncertain cases could be routed to a broader forensic workflow. The reviewer would examine provenance, metadata, source history, branch availability, face XAI, background context, audio evidence, and synchronization behavior rather than relying on the binary decision alone.

Such a workflow is consistent with the thesis position that DeepShield is a screening framework. It can reduce the volume of media requiring immediate manual attention and provide structured technical evidence, but it should not replace independent verification. This human-review layer is especially important when the input comes from an unseen platform, contains multiple speakers or faces, or has undergone heavy post-processing.

7.22 Results Chapter Summary

The final results demonstrate that the aligned five-branch DeepShield pipeline performs very strongly on its frozen internal TEST population. The system combines high ranking

quality, high threshold-based classification performance, zero false positives in the evaluated REAL subset, and low calibration error. At the same time, the ablation and validity analysis show that the evidence is not uniformly strong across every branch and that the dataset is highly imbalanced.

The appropriate conclusion is therefore evidence-based rather than absolute. DeepShield has completed a reproducible multimodal internal evaluation with strong results and traceable model behavior. The next scientific step is not to inflate the internal claim, but to test whether the same locked design, without test-specific tuning, maintains discrimination and calibration across unseen datasets, generator families, compression conditions, and real-world recording environments.

CHAPTER 8

CONCLUSION, LIMITATIONS AND FUTURE WORK

8.1 Conclusion

This thesis developed DeepShield, a five-branch explainable and reliability-aware framework for deepfake forensics. The system integrates Face, Voice, Temporal, Lip-Sync, and Background evidence under explicit modality availability masks. Face analysis combines Swin, GenConViT-ED, GenConViT-VAE, and frequency-domain evidence. Voice uses MFCC and Log-Mel CNN-LSTM modeling. Temporal uses ordered face sequences with BiLSTM attention. Lip-Sync captures audiovisual synchronization evidence. and Background uses CLIP and DINOv2 context features. Learned attention fusion, validation-only calibration, and forensic XAI complete the pipeline.

The final aligned experiment used 21,544 cases with 17,234 TRAIN, 2,159 validation, and 2,151 frozen TEST cases. The locked internal TEST result achieved ROC-AUC 0.998277, accuracy 0.992562, precision 1.0, recall 0.992385, and F1 0.996178, together with low ECE and Brier scores. Ablation showd that Face is the most influential branch, Voice and Background add useful evidence, Temporal is weaker in isolation, and Lip-Sync did not improve validation AUC in the reported run. These findings support the value of multimodal evidence while also showing why each branch should be evaluated empirically rather than assumed to help.

8.2 Limitations

- The strongest reported results are internal; they do not establish equivalent performance on unseen future generators or platforms.
- The supplied locked result set contains a trained Voice branch and a final five-branch multimodal fusion score, but it does not establish separate external voice-deepfake generalization across unseen speech-generation methods.
- No final multi-person tracking metric is reported, even though the broader framework is intended to support multi-person media.
- Only four deterministic frames are used per video, so very short manipulation intervals may be missed.
- Grad-CAM explanations are qualitative and are not validated against ground-truth manipulation masks in the locked result set.

- Calibration is documented through temperature scaling together with ECE 0.0072608272 and Brier score 0.0051208383 on the frozen internal TEST.
- Exact confusion-matrix counts are included for the final frozen TEST, while full external cross-dataset performance was not included in the final result evidence used for this book.

8.3 Future Work

The highest-priority audio extension is broader external verification of the implemented Voice branch using additional samples with trustworthy audio authenticity labels. AASIST, WavLM, or the existing CNN-LSTM architecture could be compared for synthetic speech, voice conversion, and voice cloning. Speaker diarization and active-speaker association would then allow the system to match each voice segment with the correct visible person in multi-person scenes. The existing multimodal fusion should also be stress-tested after the individual branches show stable behavior on external domains.

A second priority is external evaluation. Celeb-DF++, WildDeepfake, and other untouched domains should be used to measure cross-dataset generalization. The final internal threshold and temperature may not transfer directly, so external evaluation should report both the locked operating point and recalibration-free ROC-AUC.

A third direction is richer temporal sampling. Adaptive sampling could select additional frames when uncertainty is high, when the face changes rapidly, or when active speech begins. This would preserve efficiency for easy videos while allocating more computation to difficult cases.

A fourth direction is quantitative explainability. Manipulation masks can support localization metrics, while explanation stability can be tested under compression and small perturbations. Standalone frequency-domain and DINOv2 ablations can also be revisited as robustness studies while preserving their documented roles inside the final Face and Background branches under the same locked evaluation protocol.

Finally, reliability should be expanded beyond the reported temperature scaling, ECE, and Brier score. Future DeepShield versions can include reliability diagrams, adaptive calibration error, ensemble disagreement, out-of-distribution detection, input-quality estimation, and an abstention policy that returns UNCERTAIN when evidence is insufficient. This is particularly important for forensic deployment, where refusing to make an unsupported decision can be safer than producing a confident but incorrect label.

8.4 Final Research Statement

The final contribution of DeepShield is therefore not simply a high percentage. It is a complete research workflow that connects heterogeneous data preparation, reproducible face preprocessing, modern spatial representation, temporal reasoning, validation-only

selection, calibration, locked testing, and explainability. The locked results demonstrate excellent internal discrimination, while the limitations and future-work sections define the evidence still required before the system should be described as a universally reliable multimodal forensic detector.

8.5 Research Contributions Revisited

The completed thesis contributes at three connected levels. At the data level, it develops a provenance-aware workflow that distinguishes historical broad processing from the final aligned cohort and maintains separate modality labels and availability states. At the model level, it implements a modular five-branch architecture combining facial, acoustic, temporal, synchronization, and contextual evidence. At the evaluation level, it uses validation-only model selection and calibration before a single frozen TEST assessment.

These contributions are mutually dependent. A multimodal architecture without correct labels would train on contradictory evidence; a strong branch model without leakage control could produce misleading performance; and a high test score without calibration or claim discipline would be difficult to use responsibly. The thesis therefore presents DeepShield as an integrated research system rather than as one neural network.

8.6 Practical and Academic Implications

Practically, the work shows that a modular multimodal detector can be organized so that missing evidence is handled explicitly and final decisions remain auditable. The model can produce a case-level probability while still exposing which modalities were available and which subbranches contributed. This is useful for forensic screening because it supports a structured review process instead of an unexplained binary output.

Academically, the project demonstrates the value of preserving negative and historical evidence. Temporal or Lip-Sync components that are weaker in one experiment remain informative because they reveal how multimodal assumptions behave under the actual data distribution. Likewise, historical ConvNeXt results remain part of the project record without being confused with the canonical five-branch score. This distinction improves scientific transparency and creates clearer hypotheses for future work.

8.7 Prioritized Future Research Roadmap

Future work should proceed in a controlled sequence. The first priority is external evaluation using untouched datasets and unseen manipulation families while keeping the current final model, threshold, and calibration settings fixed. This will measure true transfer rather than a newly tuned external system. The second priority is to expand the diversity and balance of REAL cases so that specificity and calibration can be estimated more robustly. The third priority is branch-focused improvement, particularly for

Temporal and Lip-Sync evidence, using richer sequences, better active-speaker association, and more diverse synchronization conditions.

The fourth priority is reliability under distribution shift. Future versions should test out-of-distribution detection, ensemble disagreement, input-quality estimation, and an explicit UNCERTAIN outcome. The fifth priority is quantitative explainability, including explanation stability and, where manipulation masks exist, localization metrics. Finally, a deployment study should evaluate latency, resource usage, privacy, human-review workload, and failure modes rather than considering classification accuracy alone.

8.8 Closing Perspective

Deepfake detection will remain a moving target because generation technology, compression pipelines, and distribution channels continue to change. A durable forensic system must therefore be designed for evidence diversity, careful evaluation, and revision rather than for one benchmark score. DeepShield reflects this principle by combining multiple evidence branches while preserving explicit uncertainty, provenance, and experimental boundaries.

The final internal result provides strong evidence that the completed system can distinguish REAL and FAKE cases within the aligned experimental setting. The broader value of the thesis lies in how that result was reached: heterogeneous data were organized, historical and final experiments were separated, branch evidence was evaluated rather than assumed, calibration was locked before testing, and explainability was retained as support for human review. These practices provide a foundation for a more general and externally validated DeepShield system in future research.

REFERENCES

[1] L. Y. Gong and X. J. Li, “A contemporary survey on deepfake detection: Datasets, algorithms, and challenges,” Electronics, vol. 13, no. 3, Art. 585, 2024, doi: 10.3390/electronics13030585.

[2] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu, “Countering malicious deepfakes: Survey, battleground, and horizon,” International Journal of Computer Vision, vol. 130, no. 7, pp. 1678-1734, 2022, doi: 10.1007/s11263-022-01603-4.

[3] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton Ferrer, “The Deepfake Detection Challenge (DFDC) preview dataset,” arXiv:1910.08854, 2019, doi: 10.48550/arXiv.1910.08854.

[4] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” arXiv:1901.08971, 2019, doi: 10.48550/arXiv.1901.08971.

[5] Y. Li, D. Zhu, X. Cui, and S. Lyu, “Celeb-DF++: A large-scale challenging video DeepFake benchmark for generalizable forensics,” arXiv:2507.18015, 2025, doi: 10.48550/arXiv.2507.18015.

[6] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, “DeepfakeBench: A comprehensive benchmark of deepfake detection,” arXiv:2307.01426, 2023, doi: 10.48550/arXiv.2307.01426.

[7] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11976-11986, doi: 10.1109/CVPR52688.2022.01167.

[8] G. Petmezas, V. Vanian, K. Konstantoudakis, E. E. I. Almaloglou, and D. Zarpalas, “Video deepfake detection using a hybrid CNN-LSTM-Transformer model for identity verification,” Multimedia Tools and Applications, vol. 84, pp. 40617-40636, 2025, doi: 10.1007/s11042-024-20548-6.

[9] M. I. Abidin, I. Nurtanio, and A. Achmad, “Deepfake detection in videos using Long Short-Term Memory and CNN ResNext,” ILKOM Jurnal Ilmiah, vol. 14, no. 3, pp. 178-185, 2022, doi: 10.33096/ilkom.v14i3.1254.178-185.

[10] X. Dong et al., “Protecting celebrities from deepfake with identity consistency transformer,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9468-9477, doi: 10.48550/arXiv.2203.01318.

[11] B. Liu, B. Liu, M. Ding, T. Zhu, and X. Yu, “TI2Net: Temporal identity inconsistency network for deepfake detection,” in Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV), 2023, pp. 4690-4699, doi: 10.1109/WACV56688.2023.00468.

[12] D. Zhang, F. Lin, Y. Hua, P. Wang, D. Zeng, and S. Ge, “Deepfake video detection with spatiotemporal dropout transformer,” arXiv:2207.06612, 2022, doi: 10.48550/arXiv.2207.06612.

[13] A. Hashmi, S. A. Shahzad, C.-W. Lin, Y. Tsao, and H.-M. Wang, “AVTENet: A human-cognition-inspired audio-visual transformer-based ensemble network for video deepfake detection,” IEEE Transactions on Cognitive and Developmental Systems, 2025, doi: 10.1109/TCDS.2025.3554477.

[14] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, S.-N. Lim, and Y.-G. Jiang, “M2TR: Multi-modal multi-scale transformers for deepfake detection,” in Proc. 2022 International Conference on Multimedia Retrieval (ICMR), 2022, pp. 615-623, doi: 10.1145/3512527.3531390.

[15] T. Rahman, R. Chakma, and T. Mahmud, “Beyond accuracy: Explainable multimodal deepfake detection through cross-modal feature analysis and dynamic attention weighting,” in Proc. International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN), 2025, pp. 1-6, doi: 10.1109/QPAIN66474.2025.11171737.

[16] S. Armağan, E. Gündoğan, M. Kaya, and R. Alhajj, “An integrated explainable framework for multimodal deepfake detection across image, audio, and video data,” Ain Shams Engineering Journal, vol. 17, no. 7, p. 104233, 2026, doi: 10.1016/j.asej.2026.104233.

[17] M. Oquab et al., “DINOv2: Learning robust visual features without supervision,” Transactions on Machine Learning Research, 2024, doi: 10.48550/arXiv.2304.07193.

[18] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, “Frequency-aware deepfake detection: Improving generalizability through frequency space learning,” arXiv:2403.07240, 2024, doi: 10.48550/arXiv.2403.07240.

[19] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, “Thinking in frequency: Face forgery detection by mining frequency-aware clues,” arXiv:2007.09355, 2020, doi: 10.48550/arXiv.2007.09355.

[20] P. D. Halle, Y. P. Kohakade, A. S. Gurekar, and S. D. Gadkar, “Explainable deepfake detection system for images using XceptionNet, Grad-CAM, and MediaPipe,” Computer Research and Development, vol. 25, no. 11, 2025.

[21] A. A. Solanke, “Explainable digital forensics AI: Towards mitigating distrust in AI-based digital forensics analysis using interpretable models,” Forensic Science International: Digital Investigation, vol. 42, p. 301403, 2022, doi: 10.1016/j.fsidi.2022.301403.

[22] N. Kose, A. Rhodes, U. A. Çiftçi, and İ. Demir, “Is It Certainly a Deepfake? Reliability analysis in detection & generation ecosystem,” in Proc. IEEE/CVF ICCV Workshops, 2025.

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows,” in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012-10022, doi: 10.1109/ICCV48922.2021.00986.

[24] D. W. Deressa, H. Mareen, P. Lambert, S. Atnafu, Z. Akhtar, and G. Van Wallendael, “GenConViT: Deepfake Video Detection Using Generative Convolutional Vision Transformer,” Applied Sciences, vol. 15, no. 12, Art. 6622, 2025, doi: 10.3390/app15126622.

[25] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” in Proc. 38th International Conference on Machine Learning (ICML), PMLR, vol. 139, 2021, pp. 8748-8763.

[26] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint Face Detection and Alignment Using Multi-task Cascaded Convolutional Networks,” IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503, 2016.