TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
PULCHOWK CAMPUS

A
PROJECT PROPOSAL
ON
MULTI-DOMAIN FEATURE FUSION DEEPFAKE DETECTION

SUBMITTED BY:
SANDESH KUIKEL (PUL078BCT076)
SHREYA UPRETY (PUL078BCT086)
SUBASH KANDEL (PUL078BCT091)

SUBMITTED TO:
DEPARTMENT OF ELECTRONICS & COMPUTER ENGINEERING

August, 2025

Acknowledgments

We would like to express our sincere gratitude to our project supervisor, Prof. Dr. Subarna Shakya, for their invaluable guidance, expertise, and continuous support throughout this project. Their insights and feedback have been instrumental in shaping our research direction and technical approach.

We also extend our appreciation to the Department of Electronics and Computer Engineering at Pulchowk Campus, Institute of Engineering, for providing us with the necessary resources and academic environment to conduct this research.

Our thanks go to the open-source community for developing and maintaining the tools and frameworks that made this project possible, especially the contributors of PyTorch, XceptionNet, and DINO.

Finally, we acknowledge the researchers whose prior work in deepfake detection has laid the foundation for our project and inspired our multi-modal approach.

Contents

List of Figures

List of Abbreviations

AFSL	Adversarial Feature Similarity Learning
AI	Artificial Intelligence
AUC	Area Under the Curve
BCE	Binary Cross Entropy
CDDB	Continual Deepfake Detection Benchmark
CNN	Convolutional Neural Network
CVPR	Conference on Computer Vision and Pattern Recognition
DCT	Discrete Cosine Transform
DFDC	Deepfake Detection Challenge
DINOv2	Distillation with No labels version 2
FF++	FaceForensics++
FFT	Fast Fourier Transform
GAN	Generative Adversarial Network
HFE	High-Frequency Enhancement
JPEG	Joint Photographic Experts Group
MLP	Multi-Layer Perceptron
MTCNN	Multi-task Cascaded Convolutional Network
PGD	Projected Gradient Descent
ReLU	Rectified Linear Unit
RGB	Red Green Blue
ROC	Receiver Operating Characteristic
ViT	Vision Transformer

1. Introduction

1.1 Background

The proliferation of synthetic media, particularly deepfakes, has emerged as a significant technological and societal challenge. Deepfakes, AI-generated or manipulated media depicting individuals saying or doing things they never did, have evolved rapidly in sophistication since their emergence in 2017. Recent advancements in generative AI, including models like Sora, Midjourney, and advanced GANs, have dramatically improved the quality and accessibility of synthetic media creation tools, making fake content increasingly difficult to distinguish from authentic media.

The potential for misuse of deepfake technology extends across multiple domains, including:

1. Misinformation and political manipulation
2. Non-consensual intimate imagery
3. Financial fraud through voice and video impersonation
4. Identity theft and social engineering
5. Undermining of media authenticity and public trust

As deepfake technology becomes more sophisticated, traditional detection methods that rely on single-modality analysis (e.g., only spatial features) have shown decreasing effectiveness against state-of-the-art forgeries. According to recent benchmarks, single-modality detectors experience a significant performance drop when confronted with new deepfake generation methods not seen during training, with accuracy decreasing from over 95% on in-distribution data to below 70% on cross-dataset evaluations [14].

1.2 Problem statements

Despite significant research efforts, current deepfake detection systems face several critical challenges:

1. **Limited Cross-Domain Generalization**: Existing detectors perform well on familiar deepfake types but struggle significantly with novel generation methods not repre-

sented in their training data. This ”reality gap” represents a fundamental challenge for practical deployment.

2. **Modality-Specific Vulnerabilities**: Single-modality detectors (e.g., those using only RGB pixel analysis) can be easily defeated by adversarial attacks targeting their specific detection mechanisms, such as recompression, noise addition, or other perturbations.

3. **Rapid Evolution of Generation Methods**: The accelerating pace of deepfake technology development means detection systems quickly become outdated as new forgery techniques emerge that can bypass existing detection signatures.

4. **Real-World Deployment Constraints**: Many academic deepfake detectors demonstrate high performance in controlled environments but suffer significant degradation when applied to ”in-the-wild” media with varying quality, compression levels, and unknown provenance.

These challenges highlight the urgent need for a more robust, adaptable approach to deepfake detection that can maintain effectiveness across diverse forgery techniques and real-world scenarios.

1.3 Objectives

1. Develop an image deepfake detection framework that fuses spatial, frequency, and semantic features for more reliable authenticity analysis.

2. Evaluate the system’s performance with particular emphasis on cross-dataset generalization and robustness to common adversarial manipulations.

1.4 Scope

This project will focus on the following scope:

- **Media Types**: Our primary focus will be on face-centric deepfake detection in static images and video frames. While audio deepfakes represent an important related challenge, they fall outside the scope of this project.

- **Detection Approach**: We will develop a binary classification system (real/fake) with confidence scoring, rather than attempting forgery localization or attribution to specific generation methods.

- **Datasets**: We will utilize publicly available deepfake datasets including FaceForensics++, Deepfake Detection Challenge, Celeb-DF, DFDC, and DeepfakeEval-2024 for comprehensive evaluation across generation methods.

- **Modalities**: Our multi-modal approach will incorporate:
  - Spatial features via XceptionNet CNN architecture
  - Frequency domain analysis through DCT/FFT transformations.
  - Semantic features leveraging DINOv2 pre-trained representations
- **Performance Metrics**: We will evaluate our system using standard classification metrics (accuracy, F1-score) with particular emphasis on cross-dataset performance and robustness to common adversarial manipulations.
- **Implementation**: The system will be developed using PyTorch, with deployment considerations for both research and potential real-world applications.

The project will not attempt to address all forms of media manipulation, nor will it focus on creating a production-ready deployment system. Rather, it will emphasize research innovation in multi-modal feature fusion and cross-domain generalization for enhanced deepfake detection.

2. Literature Review

2.1 Related Theory

2.1.1 Deepfake Generalization Challenge

The primary theoretical challenge in deepfake detection is *generalization*. Early detectors often achieved high performance by overfitting to artifacts specific to a particular generation method or dataset, such as the unique upsampling pattern of a certain GAN. This approach is brittle, as the detector fails when confronted with novel forgery techniques because it has learned superficial ”tells” rather than the fundamental properties of a manipulated image.

Our project is founded on the premise that robust detection requires moving beyond a single perspective. We hypothesize that by analyzing a potential fake from multiple complementary points of view: spatial, frequency, and semantic, we can build a model that learns a more fundamental and generalizable concept of authenticity.

2.1.2 Foundations of Feature Extraction

To construct this multi-faceted view, we leverage three distinct theoretical paradigms. Each is designed to uncover different classes of forgery artifacts that may be missed by the others.

- **Spatial Domain (Local Artifact Analysis):** This approach concerns the pixel-level representation of an image. *Convolutional Neural Networks (CNNs)*, like XceptionNet, excel here by learning a hierarchy of features. Their theoretical strength is in identifying local, low-level artifacts such as pixel grid mismatches, unnatural blending at forgery boundaries, or inconsistencies in skin texture.

- **Frequency Domain (Global Inconsistency Detection):** This paradigm provides a global perspective by applying transformations like the *Discrete Cosine Transform (DCT)* or *Fast Fourier Transform (FFT)*. The theory is that the deepfake generation pipeline often introduces a subtle, high-frequency periodicity or disrupts the natural frequency statistics of a real photo. These global artifacts become apparent in the *frequency spectrum*, even if they are invisible in the spatial domain.

- **Semantic Domain (High-Level Coherence Verification):** This domain addresses the high-level meaning and context of the image. Our use of DINOv2 is grounded in the theory of *self-supervised learning* on *Vision Transformers (ViTs).* A ViT's self-attention mechanism captures global context, allowing it to verify high-level coherence. A deepfake may have flawless pixels but feature inconsistent lighting or unnatural facial geometry. A semantic model is designed to detect such incongruities.

2.1.3 Theoretical Framework for Multi-Modal Fusion

The three feature extraction paradigms are not redundant but complementary, each with its own blind spots. Simply concatenating these features would be suboptimal, as it fails to account for the varying importance of each modality for different inputs.

The theoretical foundation for our fusion strategy rests on *attention mechanisms*. Attention provides a mathematical framework for the model to learn a dynamic, context-dependent policy for integrating information. It effectively learns to ask, ”For this specific image, is the most telling clue in the spatial, frequency, or semantic features?” By calculating attention weights, our fusion module can dynamically amplify the most relevant modality, creating a synergistic effect where the combined model is more robust than the sum of its parts.

2.2 Related Works

Deepfake detection focuses on identifying manipulated facial media using AI. This review covers CNN-based methods, frequency-domain analysis, transformer models, cross-dataset generalization, robustness, and benchmarks, aligning with our project's scope: binary classification via spatial (XceptionNet), frequency (DCT/FFT), and semantic (DINOv2) features on datasets like FaceForensics++ [18], DFDC [4], and Celeb-DF [14].

2.2.1 CNN-Based Detection Methods

CNNs form the backbone for deepfake detection. Rössler et al. [18] established XceptionNet as a baseline, achieving 99.3% accuracy on FaceForensics++ high-quality fakes. Afchar et al. [1] introduced MesoNet, a network designed to detect mesoscopic details, achieving approximately 98% accuracy on DeepFake videos.

Recent advances include the work of Huang et al. [7], who proposed an Implicit Identity Driven CNN using Xception to improve generalization via identity losses. Yan et al. [24] modified Xception for disentanglement, boosting AUC from 0.672 to 0.811 on Celeb-DF. In a similar vein, Wang et al. [23] proposed AltFreezing for 3D spatiotemporal CNNs, enhancing cross-dataset performance. More recently, Sun et al. [20] augmented EfficientNet-B4 with DiffusionFake, increasing AUC on unseen data by approximately 10% on Celeb-DF. Lipianina-Honcharenko et al. [15] found Xception achieved around 87.7% accuracy on forgery tasks.

These methods consistently achieve over 90% intra-dataset accuracy, supporting our use of Xception for extracting spatial features.

2.2.2 Frequency-Domain Analysis

Frequency artifacts often aid detection, especially in compressed media. Le and Woo [11] used ADD for frequency-attention distillation, improving the detection of low-quality fakes. Gao et al. [6] proposed HFE, which outperformed existing methods on compressed DFDC/FF++ with an AUC greater than 85%. Li et al. [13] introduced FreqBlender for data augmentation, yielding an 88–90% cross-dataset AUC.

Furthermore, Tan et al. [21] developed FreqNet, utilizing FFT and phase spectra, which improved performance by approximately 9.8% across various GANs. These findings align with our decision to use a DCT/FFT modality to capture subtle manipulation artifacts.

2.2.3 Transformer and Vision-Language Models

Transformers leverage pre-trained representations for robust detection. Dong et al. [5] proposed the Identity Consistency Transformer with ViT, which generalizes well across degradations. Chen et al. [3] adapted CLIP-ViT with GFF, achieving state-of-the-art results on several benchmarks.

Self-supervised ViTs like DINO, as explored by Nguyen et al. [16], show better generalization when fine-tuned. This supports our choice of DINOv2-Base for semantic feature extraction. Hybrid models such as HCiT [9] combine CNNs and ViTs for superior results on FaceForensics++ and DFDC. Petmezas et al. [17] integrated a CNN-LSTM-Transformer with 3DMM, achieving state-of-the-art performance on multiple datasets using pristine data. CAST, proposed by Thakre et al. [22], fuses CNN features via cross-attention, reporting a 99.49% AUC on FaceForensics++ and 93.31% cross-dataset.

2.2.4 Cross-Dataset Generalization and Robustness

Generalization remains a key challenge. Brodaric & Struc (2024) analyzed the reliance of detectors on dataset-specific cues. The methods proposed by Yan et al. [24], Wang et al. [23], and Sun et al. [20] have shown improved generalization through techniques like disentanglement and augmentation.

For robustness, Khan et al. [10] used Adversarial Feature Similarity Learning (AFSL), maintaining over 80% AUC under PGD attacks. Chen et al. [2] applied adversarial self-supervision, while Sun et al. [19] used dual contrastive learning to handle domain shifts. These works inform our project's emphasis on cross-dataset metrics and adversarial robustness.

2.2.5 Benchmarks and Datasets

Standardized datasets are crucial for evaluation. FaceForensics++ [18] provides a foundational benchmark with 1,000 pristine videos. The Deepfake Detection Challenge (DFDC) dataset [4] offers approximately 100,000 diverse videos. Celeb-DF [14] presents a challenge with more realistic fakes, while DeeperForensics-1.0 [8] adds various real-world distortions.

DeepfakeBench, introduced by Yan et al. [24], provides PyTorch implementations for approximately 36 detectors on these datasets, which will be instrumental in our evaluation process.

2.2.6 Open-Source Implementations

Many researchers release their code in PyTorch. DeepfakeBench [24] includes pretrained weights. CDDB [12] offers a framework for continual learning. The availability of these resources facilitates our PyTorch-based innovation in feature fusion.

3. Proposed Methodology

Our approach to deepfake detection is built upon a multi-modal framework designed to fuse complementary features from the spatial, frequency, and semantic domains. By integrating these diverse signals, the system aims for a more robust and generalizable analysis of image authenticity. The methodology is structured into the following key stages.

3.1 Data Collection and Preprocessing

A diverse set of public datasets will be used for training and comprehensive evaluation, ensuring exposure to various generation techniques and quality levels.

- **Core Datasets**:
  - FaceForensics++ (FF++): A foundational dataset containing thousands of videos manipulated by four distinct methods.
  - Deepfake Detection Challenge (DFDC): A large-scale dataset featuring a wide variety of deepfakes with real-world augmentations.
  - Celeb-DF: Known for its high-quality, visually convincing deepfake videos that challenge modern detectors.
  - DeepfakeEval-2024: Provides current examples of state-of-the-art deepfakes to test against the latest generation methods.
- **Preprocessing Pipeline**:
  - Face detection and alignment using MTCNN (Multi-task Cascaded Convolutional Networks) to isolate facial regions.
  - Cropping and resizing faces to a consistent input resolution , with the final dimensions optimized during experimentation.
  - Normalization of pixel values using XceptionNet statistics to match the pre-trained models' expected input distribution.
- **Data Augmentation**: To fulfill the objective of robustness, we will apply augmentations that simulate common post-processing manipulations:
  - JPEG compression with varying quality factors.

  - Gaussian blur and noise injection.
  - Adjustments in brightness, contrast, and saturation.
  - Standard geometric transformations like random cropping and horizontal flipping.

3.2 Multi-Domain Feature Extraction

Our framework's core innovation lies in extracting features from three orthogonal domains to create a comprehensive representation of the input image.

3.2.1 Spatial Feature Extraction

- **Architecture**: **XceptionNet**, as specified, known for its effectiveness in capturing fine-grained spatial artifacts common in deepfakes.
- **Implementation**: We will use an ImageNet pre-trained XceptionNet model, removing the final classification layer to use it as a feature extractor.
- **Feature Dimension**: A **2048-dimensional** feature vector will be extracted from the global average pooling layer.
- **Objective**: To capture pixel-level evidence of manipulation, including inconsistent textures, sharpening artifacts, and unnatural blending boundaries.

3.2.2 Frequency Domain Analysis

- **Transformations**: Both the Discrete Cosine Transform (DCT) and Fast Fourier Transform (FFT) will be applied to the image.
- **Feature Representation**: Statistical features (e.g., mean, variance, kurtosis) will be computed from the coefficients of the DCT and the amplitude spectrum of the FFT. These will be concatenated into a single feature vector.
- **Objective**: To identify high-frequency artifacts or unnatural periodic patterns introduced by generative models (e.g., upsampling checkerboard patterns) that are often imperceptible in the spatial domain.

3.2.3 Semantic Feature Extraction

- **Architecture**: **DINOv2-Base**, a powerful pre-trained Vision Transformer that excels at capturing high-level semantic information.
- **Implementation**: The pre-trained DINOv2-Base visual encoder will be used with its parameters frozen to act as a robust semantic feature extractor.

- **Feature Dimension**: A **768-dimensional** embedding will be obtained from the model's output.
- **Objective**: To detect logical or contextual inconsistencies in the image, such as unnatural facial expressions, inconsistent lighting on different parts of the face, or subtle geometric distortions.

3.3 Feature Fusion and Classification

The extracted features from each modality are intelligently combined and processed to make a final prediction.

- **Feature Concatenation**: The spatial (2048-dim), frequency (N-dim), and semantic (768-dim) feature vectors are first concatenated.
- **Fusion Mechanism**: A simple concatenation followed by a Multi-Layer Perceptron (MLP) will serve as the baseline. More advanced techniques like cross-attention may be explored to allow the model to dynamically weigh the importance of each modality for a given input.
- **Classifier Architecture**: A Multi-Layer Perceptron (MLP) with dropout for regularization will process the fused feature vector.
  - Input Layer: Accepts the concatenated feature vector (e.g., $2048 + 768 + N_{freq}$ units).
  - Hidden Layers: One or more hidden layers with ReLU activation to learn non-linear relationships between the features.
  - Output Layer: A single neuron with a Sigmoid activation function to produce a probability score between 0 (real) and 1 (fake).
- **Output**: The model outputs a confidence score, providing insight into the certainty of its prediction.

3.4 Training Strategy

The model will be trained with a focus on achieving the primary objective of cross-dataset generalization.

- **Loss Function**: Binary Cross-Entropy (BCE) loss, suitable for binary classification tasks.

- **Optimization**: The AdamW optimizer will be used for its effective handling of weight decay.
- **Learning Rate Schedule**: A cosine annealing schedule will be employed to adjust the learning rate during training, helping the model converge to a more robust minimum.
- **Validation and Early Stopping**: Performance will be monitored on a hold-out validation set, and early stopping will be used to prevent overfitting and select the best-performing model checkpoint.

3.5 Evaluation Protocol

We will implement a comprehensive evaluation framework to assess both in-distribution and cross-domain performance:

- **Standard Metrics**: Accuracy, precision, recall, F1-score, AUC-ROC
- **Cross-Dataset Evaluation**: Train on one dataset (e.g., FF++) and test on another (e.g., Celeb-DF) to measure generalization
- **Robustness Analysis**: Evaluate performance under various perturbations:
  - Compression (multiple JPEG quality levels)
  - Noise addition (Gaussian, salt-and-pepper)
- **Ablation Studies**: Systematically evaluate the contribution of each modality and fusion component

This methodology integrates multiple complementary approaches to create a robust, adaptive deepfake detection system capable of maintaining performance across diverse forgery techniques and real-world conditions.

4. Proposed System design

Our deepfake detection system follows a modular architecture designed to extract, process, and fuse multi-modal features for robust classification. The following system design diagrams illustrate the key components and their interactions.

4.1 System Architecture

The overall system architecture comprises five main layers: Input Layer, Feature Extraction Layer, Fusion Layer, Classification Layer, and Output Layer.

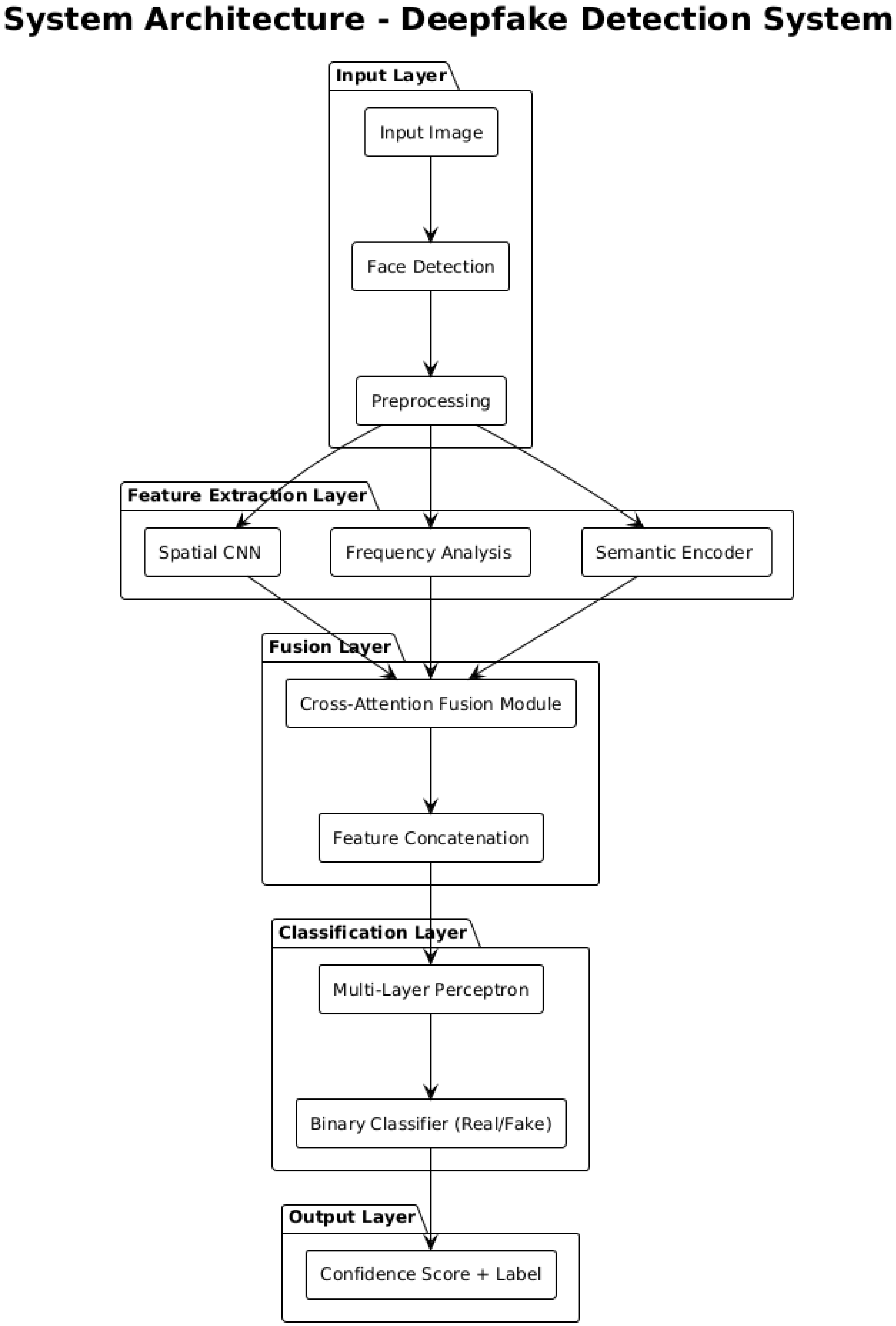

Figure 4.1: System Architecture - Deepfake Detection System

As shown in Figure 4.1, the system processes input images through a multi-stage pipeline:

- **Input Layer**: Handles input image reception, face detection, and preprocessing
- **Feature Extraction Layer**: Processes images through three parallel pathways for spatial features (Spatial CNN), frequency features (Frequency Analysis), and semantic features (Semantic Encoder)
- **Fusion Layer**: Combines multi-modal features using cross-attention mechanisms and feature concatenation
- **Classification Layer**: Processes fused features through a multi-layer perceptron for binary classification
- **Output Layer**: Provides confidence scores and classification labels

4.2 Use Case Diagram

The use case diagram illustrates the interactions between external actors and the Deepfake Detection System. It highlights the primary functionalities provided by the system and the roles of the involved actors.

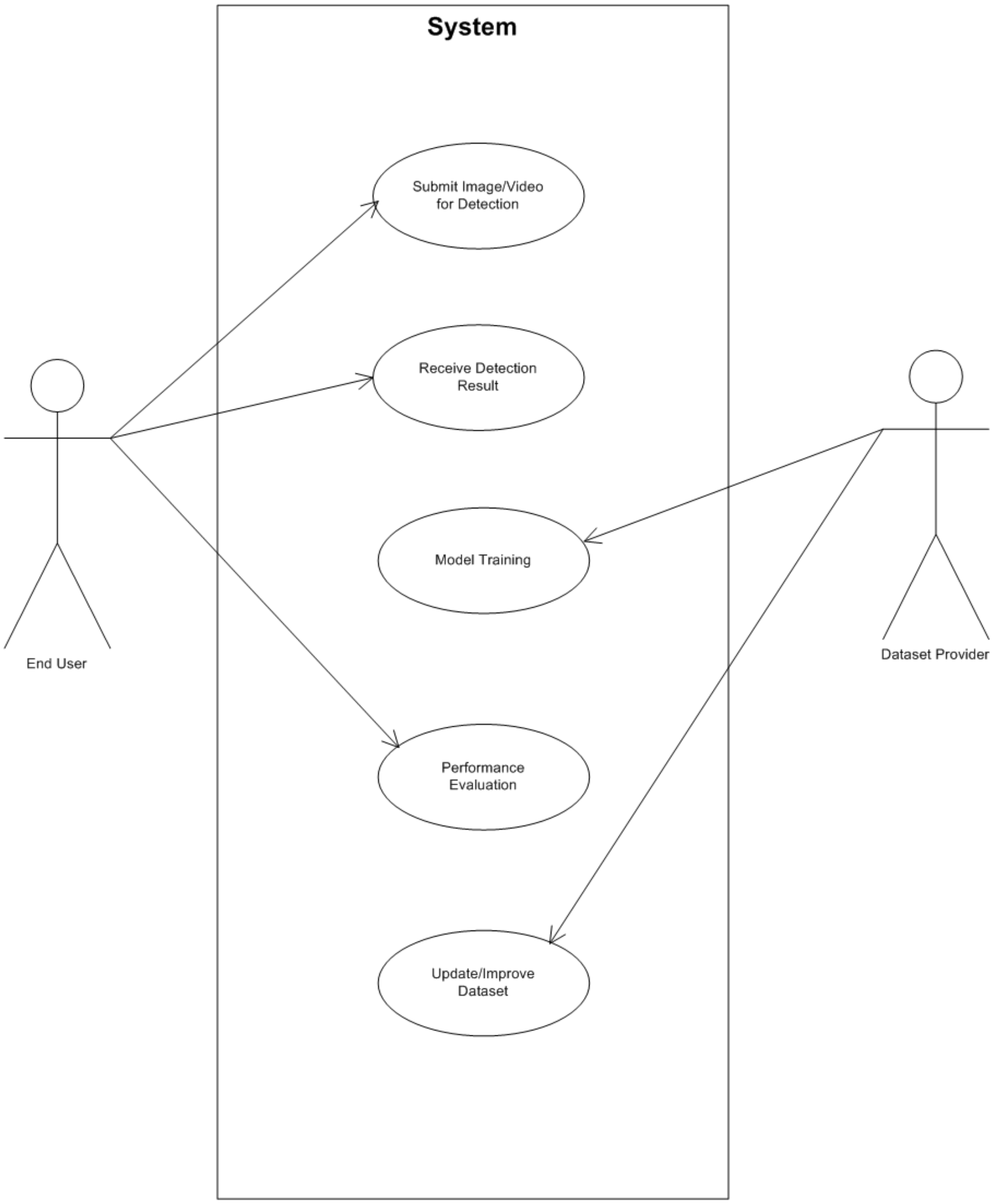

Figure 4.2: Use Case Diagram - Deepfake Detection System

As shown in Figure 4.2, the system supports the following use cases:

- **Submit Image for Detection** – End users upload images or videos for deepfake analysis.
- **Receive Detection Result** – End users receive classification results with a confidence score.
- **Model Training** – Dataset providers supply datasets for training the detection model.
- **Performance Evaluation** – Users/researchers evaluate the system's accuracy and robustness.

- **Update/Improve Dataset** – Dataset providers update datasets to improve system generalization.

4.3 Multi-Domain Feature Extraction Pipeline

The feature extraction pipeline illustrates the parallel processing paths for different modalities.

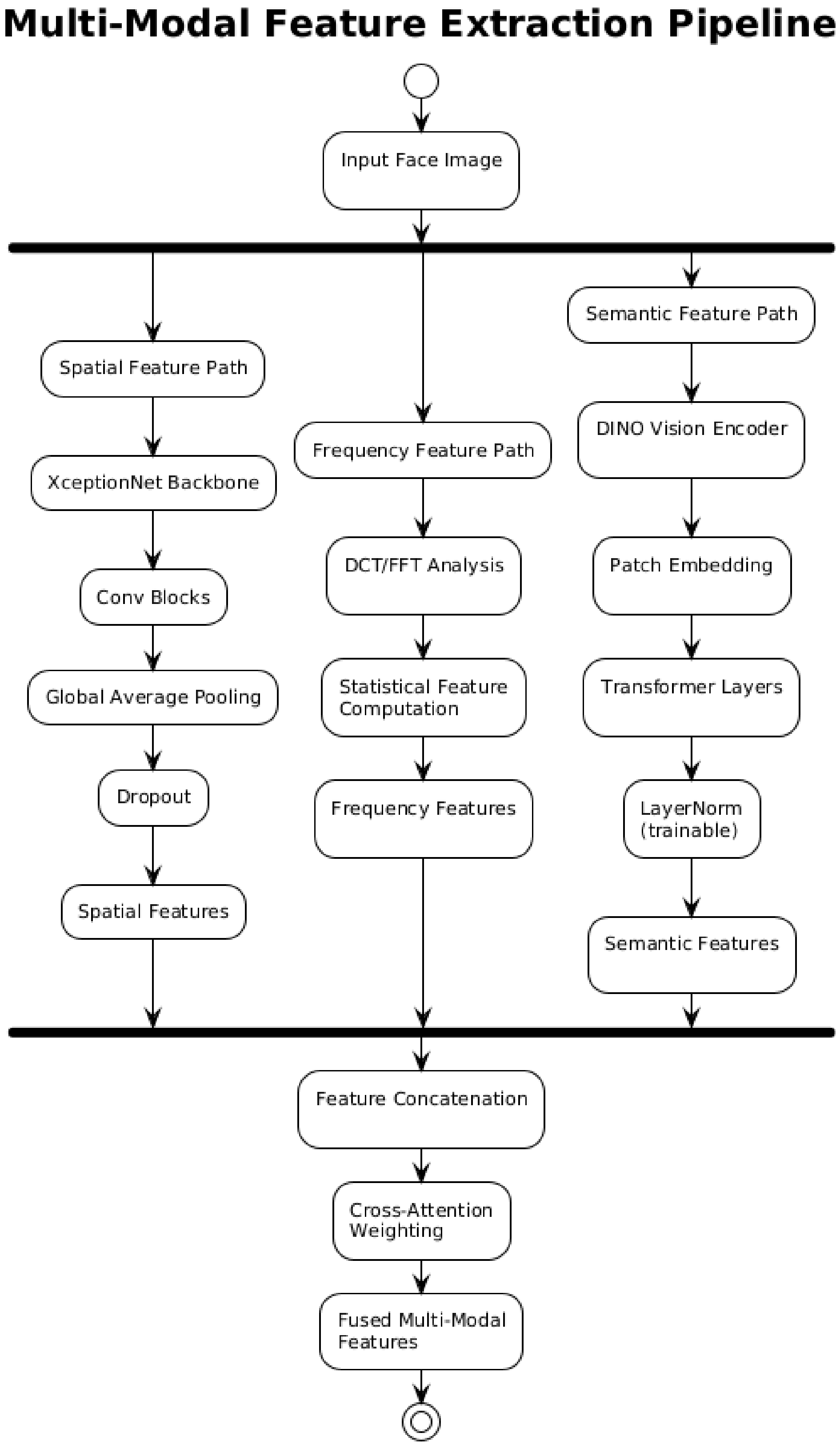

Figure 4.3: Multi-Modal Feature Extraction Pipeline

Figure 4.3 details the feature extraction process:

- Input face images are processed through multiple feature extraction pathways
- Extracted features are combined through feature concatenation
- Cross-attention mechanisms are applied to weight and fuse the features
- The output is a fused multi-modal representation for classification

4.4 Data Flow Diagram

The data flow diagram illustrates the end-to-end processing pipeline from input to classification.

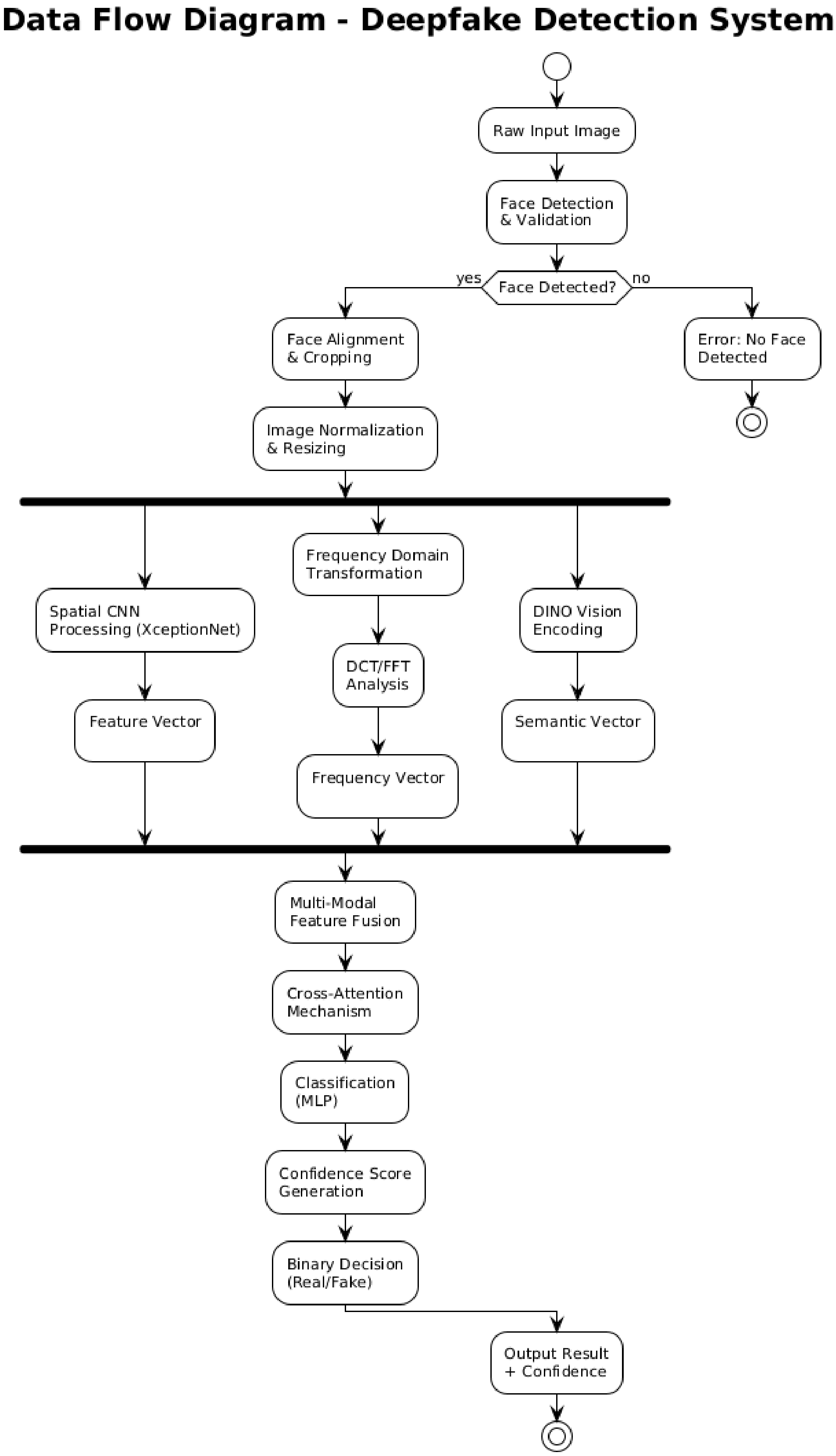
Data Flow Diagram - Deepfake Detection System
Raw Input Image
Face Detection & Validation
Face Detected?
yes
no
Face Alignment & Cropping
Error: No Face Detected
Image Normalization & Resizing
Frequency Domain Transformation
Spatial CNN Processing (XceptionNet)
DINO Vision Encoding
DCT/FFT Analysis
Feature Vector
Semantic Vector
Frequency Vector
Multi-Modal Feature Fusion
Cross-Attention Mechanism
Classification (MLP)
Confidence Score Generation
Binary Decision (Real/Fake)
Output Result + Confidence

Figure 4.4: Data Flow Diagram - Deepfake Detection System

Figure 4.4 outlines the complete data processing flow:

- Raw input images undergo face detection and validation
- Valid faces proceed through image normalization, resizing, alignment, and cropping
- Processed images are transformed to frequency domain and converted to frequency vectors
- Spatial CNN processing extracts spatial feature vectors
- DINOv2 vision encoding extracts semantic vectors
- Multi-modal features are fused using cross-attention mechanisms
- The fused features are classified to generate confidence scores
- The system outputs binary decisions (Real/Fake) with confidence values

5. Timeline

The project will be executed over an 8-month period, organized into three phases. The detailed timeline is illustrated in the Gantt chart.

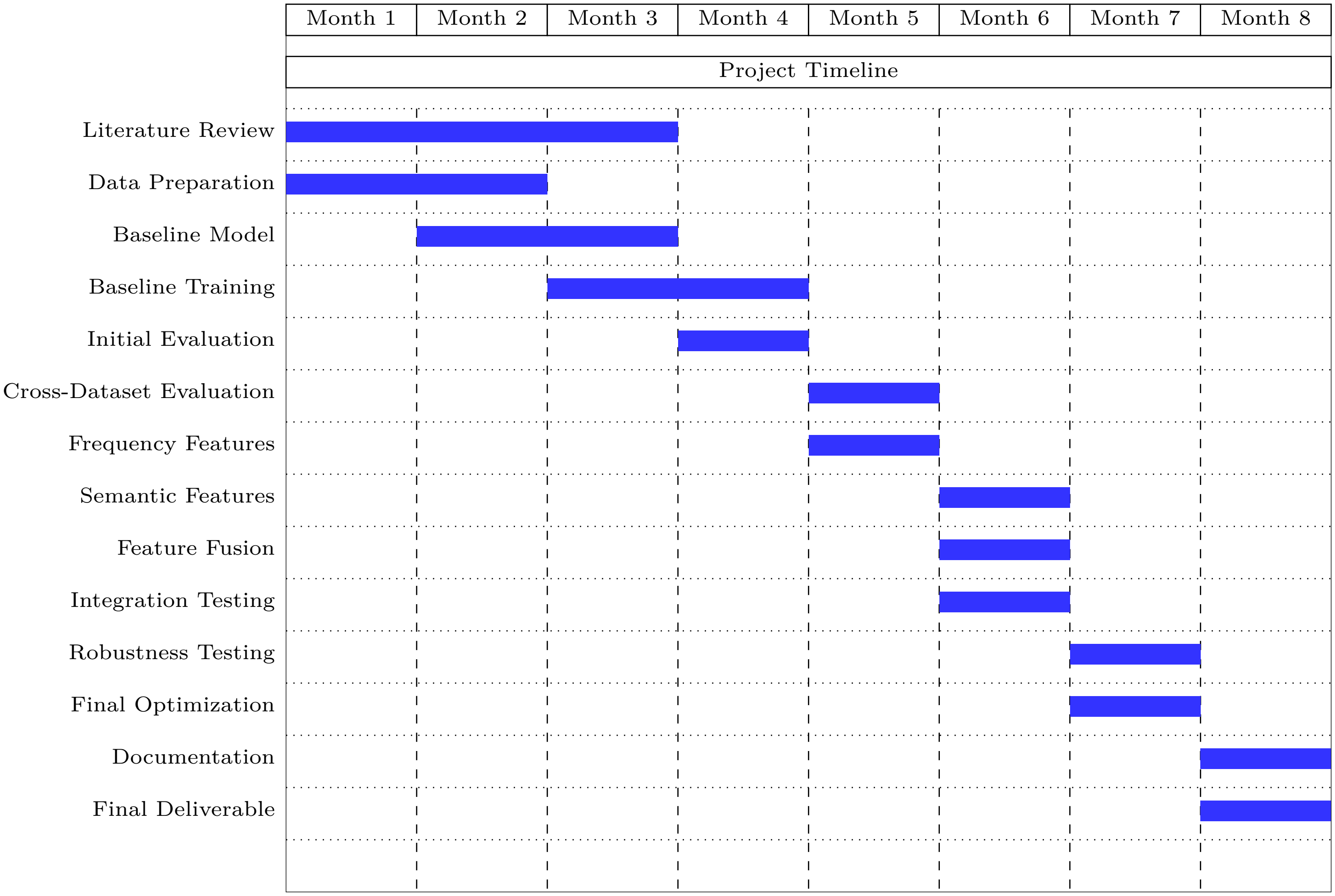

Figure 5.1: Project Timeline - 8 Months Development Plan

5.1 Phase 1: Foundation (Months 1-4)

- **Month 1-3**: Literature Review
  - Conduct comprehensive literature review (spans Months 1–3 in the timeline).
- **Month 1-2**: Data Preparation
  - Configure development environment
  - Set up version control repository
  - Install required libraries and dependencies
  - Acquire and organize datasets
  - Implement data loading pipeline
  - Create dataset metadata tracking system
- **Month 2-3**: Baseline Model
  - Implement XceptionNet spatial feature extractor
  - Create baseline classification head
  - Set up evaluation metrics
- **Month 3-4**: Baseline Training
  - Train baseline model on FF++ dataset
  - Optimize hyperparameters
  - Implement logging and checkpointing
- **Month 3-4**: Initial Evaluation
  - Evaluate baseline on in-distribution test set
  - Conduct preliminary cross-dataset testing
  - Identify performance gaps and limitations
- **Month 4**: Phase 1 Deliverable
  - Document baseline performance
  - Prepare interim report
  - Plan Phase 2 implementation details
  - Complete Phase 1 review and assessment

5.2 Phase 2: Multi-Modal Development (Months 5-6)

- **Month 5**: Cross-Dataset Evaluation
  - Implement comprehensive cross-dataset testing framework
  - Evaluate baseline performance across multiple datasets
  - Document generalization gaps
- **Month 5**: Frequency Features
  - Implement DCT/FFT feature extractors
  - Develop frequency feature processing pipeline
  - Test frequency feature discrimination capability
- **Month 6**: Semantic Features
  - Integrate DINOv2 model
  - Implement semantic feature extraction pipeline
  - Validate semantic feature effectiveness
- **Month 6**: Feature Fusion
  - Implement cross-attention fusion module
  - Develop feature concatenation mechanisms
  - Optimize fusion hyperparameters
- **Month 6**: Integration Testing
  - Create integrated multi-modal pipeline
  - Test end-to-end multi-modal system
  - Perform ablation studies on component contributions

5.3 Phase 3: Optimization and Final Delivery (Months 7-8)

- **Month 7**: Robustness Testing
  - Implement comprehensive robustness test suite
  - Evaluate performance under various perturbations

  - Identify and address vulnerabilities

- **Month 7**: Final Optimization

  - Fine-tune model based on robustness test results
  - Optimize for inference efficiency
  - Conduct final cross-dataset evaluation

- **Month 8**: Documentation

  - Prepare comprehensive technical documentation
  - Create visualizations of results
  - Document system architecture and components

- **Month 8**: Final Deliverable

  - Finalize project report
  - Prepare presentation materials
  - Package code repository with documentation

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018. doi: 10.1109/WIFS.2018.8630761. URL `https://doi.org/10.1109/WIFS.2018.8630761`.

[2] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18689–18698, 2022. URL `https://arxiv.org/abs/2203.12208`.

[3] Yingjian Chen, Lei Zhang, Yakun Niu, Pei Chen, Lei Tan, and Jing Zhou. Guided and fused: Efficient frozen clip-vit with feature guidance and multi-stage feature fusion for generalizable deepfake detection, 2024. URL `https://arxiv.org/abs/2408.13697v1`.

[4] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jiyin Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset, 2020. URL `https://arxiv.org/abs/2006.07397`.

[5] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Fang Wen, and Baining Guo. Protecting celebrities from deepfake with identity consistency transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9468–9478, 2022. URL `https://openaccess.thecvf.com/content/CVPR2022/html/Dong_Protecting_Celebrities_From_DeepFake_With_Identity_Consistency_Transformer_CVPR_2022_paper.html`.

[6] Jie Gao, Zhaoqiang Xia, Gian Luca Marcialis, Chen Dang, Jing Dai, and Xiaoyi Feng. Deepfake detection based on high-frequency enhancement network for highly compressed content. *Expert Systems with Applications*, 2024. doi: 10.1016/j.eswa.2024.123732. URL `https://doi.org/10.1016/j.eswa.2024.123732`.

[7] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*

(CVPR), pages 4490–4499, June 2023. doi: 10.1109/CVPR52729.2023.00436. URL https://openaccess.thecvf.com/content/CVPR2023/papers/Huang_Implicit_Identity_Driven_Deepfake_Face_Swapping_Detection_CVPR_2023_paper.pdf.

[8] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2889–2898, 2020. URL https://arxiv.org/abs/2001.03024.

[9] Bachir Kaddar, Sid Ahmed Fezza, Zahid Akhtar, Wassim Hamidouche, Abdenour Hadid, and Joan Serra-Sagristà. Deepfake detection using spatio-temporal transformer. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20 (11):1–21, 2024. doi: 10.1145/3643030. URL https://doi.org/10.1145/3643030.

[10] Sarwar Khan, Jun-Cheng Chen, Wen-Hung Liao, and Chu-Song Chen. Adversarially robust deepfake detection via adversarial feature similarity learning, 2024. URL https://arxiv.org/abs/2403.08806.

[11] Binh M. Le and Simon S. Woo. Add: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 122–130, 2022. URL https://ojs.aaai.org/index.php/AAAI/article/view/19886.

[12] Chuqiao Li, Zhiwu Huang, Danda Pani Paudel, Yabin Wang, Mohamad Shahbazi, Xiaopeng Hong, and Luc Van Gool. Cddb: A benchmark for continual deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2533–2542, 2023. URL https://arxiv.org/abs/2205.05467.

[13] Hanzhe Li, Jiaran Zhou, Yuezun Li, Baoyuan Wu, Bin Li, and Junyu Dong. Freqblender: Enhancing deepfake detection by blending frequency knowledge, 2024. URL https://arxiv.org/abs/2404.13872.

[14] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/papers/Li_Celeb-DF_A_Large-Scale_Challenging_Dataset_for_DeepFake_Forensics_CVPR_2020_paper.pdf.

[15] Khrystyna Lipianina-Honcharenko, Mykola Telka, and Nazar Melnyk. Comparison of resnet, efficientnet, and xception architectures for deepfake detection. In *Proceedings of the 1st International Workshop on Advanced Applied Information Technologies (AdvAIT-2024)*, 2024. URL https://ceur-ws.org/Vol-3899/paper3.pdf.

[16] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Exploring self-supervised vision transformers for deepfake detection: A comparative analysis, 2024. URL https://arxiv.org/abs/2405.00355.

[17] Georgios Petmezas, Vazgken Vanian, Konstantinos Konstantoudakis, Eleana Almaloglou, and Dimitris Zarpalas. Video deepfake detection using a hybrid cnn-lstm-transformer model for identity verification. *Multimedia Tools and Applications*, 2025. doi: 10.1007/s11042-024-20548-6. URL https://doi.org/10.1007/s11042-024-20548-6.

[18] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. URL https://arxiv.org/abs/1901.08971.

[19] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, and Rongrong Ji. Dual contrastive learning for general face forgery detection, 2021. URL https://arxiv.org/abs/2112.13522.

[20] Ke Sun, Shen Chen, Taiping Yao, Shouhong Ding, Rongrong Ji, and Jinsong Liang. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion, 2024. URL https://arxiv.org/abs/2410.04372.

[21] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space learning, 2024. URL https://arxiv.org/abs/2403.07240.

[22] Aryan Thakre, Omkar Nagwekar, Vedang Talekar, and Aparna Santra Biswas. Cast: Cross-attentive spatio-temporal feature fusion for deepfake detection, 2025. URL https://arxiv.org/abs/2506.21711.

[23] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4129–4138, 2023. URL https://arxiv.org/abs/2307.08317.

[24] Zhiyuan Yan, Yong Zhang, Yubing Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22530–22541, 2023. URL `https://openaccess.thecvf.com/content/ICCV2023/html/Yan_UCF_Uncovering_Common_Features_for_Generalizable_Deepfake_Detection_ICCV_2023_paper.html`.