Federal University of Alagoas

Computing Institute

Postgraduate Coordination in Informatics

Master's Thesis

A Comprehensive Framework for Atrial Fibrillation Classification: From ECG Images to Multimodal Analysis

Master's Candidate

Rafael Monteiro Laranjeira

Advisor

Thiago Damasceno Cordeiro, Dr.

Co-Advisor

Marco Antonio Gutierrez, Dr.

Maceió, AL

January 31 - 2025

RAFAEL MONTEIRO LARANJEIRA

A COMPREHENSIVE FRAMEWORK FOR ATRIAL FIBRILLATION CLASSIFICATION: FROM ECG IMAGES TO MULTIMODAL ANALYSIS

Master thesis presented as a partial requirement to obtain a Master Degree by the Master's in Informatics Program of the Computing institute at Federal University Of Alagoas
Advisor: Thiago Damasceno Cordeiro, Dr.

MACEIÓ, AL
JANUARY - 2025

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO
Av. Lourival Melo Mota, S/N, Tabuleiro do Martins, Maceió - AL, 57.072-970
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO (PROPEP)
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Folha de Aprovação

RAFAEL MONTEIRO LARANJEIRA

UM FRAMEWORK ABRANGENTE PARA CLASSIFICAÇÃO DA FIBRILAÇÃO ATRIAL: DAS IMAGENS DE ECG À ANÁLISE MULTIMODAL

A COMPREHENSIVE FRAMEWORK FOR ATRIAL FIBRILLATION CLASSIFICATION: FROM ECG IMAGES TO MULTIMODAL ANALYSIS

Dissertação submetida ao corpo docente do Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas e aprovada em 31 de janeiro de 2025.

Banca Examinadora:

Prof. Dr. THIAGO DAMASCENO CORDEIRO
UFAL – PPGI- Instituto de Computação
Orientador

Prof. Dr. BRUNO ALMEIDA PIMENTEL
UFAL – PPGI- Instituto de Computação
Examinador Interno

Prof. Dr. MARCO ANTONIO GUTIERREZ
USP-Universidade de São Paulo
Coorientador

Prof. Dr. ESTELA RIBEIRO
USP–Faculdade de Medicina da Universidade de São Paulo.
Examinador Externo

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária: Girlaine da Silva Santos – CRB-4 – 1127

L318c Laranjeira, Rafael Monteiro.
A comprehensive framework for atrial fibrillation classification: from ecg images to multimodal analysis / Rafael Monteiro Laranjeira. – 2025.
166 f. : il.

Orientador: Thiago Damasceno Cordeiro.
Dissertação (Mestrado em Informática.) - Universidade Federal de Alagoas, Instituto de Informática. Programa de Pós-Graduação em Informática. Maceió, 2025.

Apêndices: f. 101-149.
Bibliografia: f. 150-166.

1. Fibrilação atrial. 2. Machine Learning. 3. Eletrocardiografia. 4. Redes neurais (Computação). I. Título.

CDU: 004.81: 616.12-073.97

Faithless is he that says farewell when the road darkens.

J.R.R. Tolkien

Acknowledgments

I'm grateful to Prof. Thiago Cordeiro for his ongoing support, friendship, patience, guidance, motivation, and inspiration. I also would like to express my gratitude to everyone from InCor who supported me with ideas, advice and for providing the database used in this work.

I also wish to express my gratitude to the members of my master's thesis committee for their considerable contributions to this research. Their time, effort, and valuable insights have had a significant impact on the quality of this work, and their constructive feedback and guided support have been essential in the development of this research. I am especially grateful to Professors Bruno Pimentel, Estela Ribeiro and Marco Gutierrez for their dedication and encouragement, which have motivated me to improve this work.

The accomplishment of this endeavor would not have been feasible without the unwavering support of my parents and family members, who maintained their faith in me even during periods of adversity.

I would like to acknowledge my classmates, with whom I had the privilege of studying. Their companionship and support during challenging moments were invaluable throughout this journey. My sincere gratitude goes to Gustavo Miranda, whose encouragement and friendship have been constants since the very beginning of this path. I am also extremely grateful to Gustavo Cabral for being an incredibly encouraging friend for many years, sharing the academic journey at UFAL and it is a joy to now have him as a work colleague as well.

I want to take this opportunity to express my heartfelt gratitude to Giovana for being an invaluable partner, offering boundless support and encouragement through every success and hardship. Your presence has been a great source of strength, inspiration and motivation. Thank you for standing by me through these challenging times and for always believing in me, regardless of the circumstances.

I am deeply grateful to everyone who has supported me in one way or another during this journey. I would like to express my special thanks to Álvaro, Roger, Sofia, Artur, Eduardo, Laryssa, Rodrigo, Valério, João Arthur, Augusto, Roberto, Kevin, Hugo Tallys, Audrey, and Maria Júlia for their enduring support, constant motivation, and the countless moments of laughter and joy we have shared.

List of Figures

List of Tables

Contents

Resumo

Esta pesquisa apresenta um framework abrangente para a detecção automatizada de fibrilação atrial (FA) que preenche a lacuna entre a prática clínica e as técnicas avançadas de aprendizado de máquina. Introduzimos um fluxo de processamento que transforma imagens padrão de exames de ECG de 12 derivações em múltiplas representações complementares, permitindo uma investigação sistemática de diferentes abordagens para a detecção de FA. O framework processa imagens brutas de ECG para extrair dados da derivação II, que são então transformados em três modalidades distintas: imagens processadas, séries temporais e espectrogramas. Cada modalidade é então analisada usando arquiteturas de redes neurais especializadas que são otimizadas para suas características específicas.

A investigação envolve dois cenários experimentais: uma comparação balanceada entre FA e ritmos normais e um cenário clinicamente realista que mantém as distribuições naturais das classes (FA e não FA), usando uma base privada (InCor-DB) e uma base pública (Zheng-DB). No cenário balanceado, a abordagem multimodal alcançou um F1-score de $99,28\% \pm 0,02\%$, enquanto as modalidades individuais alcançaram consistentemente valores acima de $98.36\% \pm 0.17\%$. No cenário clinicamente realista, onde os casos de FA representaram $8,45\%$ dos dados, a robustez do framework foi demonstrada com a abordagem multimodal alcançando um F1 score de $88,59\%$. A validação externa usando o conjunto de dados Zheng-DB confirmou a generalização do framework, com a abordagem multimodal mantendo um bom desempenho (F1 score de $98,13 \pm 0,36\%$) em condições balanceadas.

Uma parte importante da nossa abordagem é o mecanismo de fusão ponderada que combina recursos de cada modalidade, usando pesos aprendidos para determinar como as diferentes representações contribuem para a análise final. Nossos experimentos mostram que esses pesos se adaptam à complexidade da tarefa, mantendo contribuições equilibradas entre as modalidades (imagem: $0,3382 \pm 0,0025$, espectrograma: $0,3450 \pm 0,0033$, série temporal: $0,3167 \pm 0,0045$) para discriminar a FA de ritmos normais, ao mesmo tempo em que mostra especialização (série temporal: $0,5025 \pm 0,1252$) para discriminar a FA de várias arritmias. Esse comportamento adaptativo demonstra a capacidade do mecanismo de otimizar o uso de recursos com base em desafios de classificação específicos, contribuindo para um desempenho consistente em diferentes cenários clínicos.

Esta pesquisa contribui para o campo da análise automatizada de ECG, fornecendo evidências empíricas para a eficácia de diferentes representações de dados na detecção de FA. Também foi observado que não é muito comum encontrar trabalhos que abordem espectrogramas, imagens e séries temporais simultaneamente. Esse estudo busca mostrar a viabilidade dessa combinação dessas entradas e o relacionamento entre elas. A capacidade do framework de processar imagens de ECG padrão o torna compatível com os ambientes onde apenas o formato de imagem está disponível, o que pode facilitar sua adoção em ambientes com menos recursos e médicos disponíveis.

Abstract

This research presents a comprehensive framework for automated atrial fibrillation (AF) classification that bridges the gap between clinical practice and advanced machine learning techniques. We introduce a pipeline that transforms standard 12-lead electrocardiogram (ECG) examination images into multiple complementary representations, enabling a systematic investigation of different approaches to AF classification. The framework processes raw ECG images to extract Lead II data, which is then transformed into three distinct modalities: processed images, time series, and spectrograms. Each modality is then analyzed using specialized neural network architectures that are optimized for their specific characteristics.

The investigation involves two experimental scenarios: a balanced comparison between AF and normal rhythms and a clinically realistic scenario that maintains the natural class distributions (AF and non-AF), using a private dataset (InCor-DB) and a public dataset (Zheng-DB). In the balanced scenario, the multimodal approach achieved an F1-score of $99.28\% \pm 0.02\%$, while individual modalities consistently reached values above $98.36\% \pm 0.17\%$. In the clinically realistic scenario, where AF cases accounted for 8.45% of the data, the robustness of the framework was demonstrated, with the multimodal approach achieving an F1-score of 88.59%. External validation using the Zheng-DB dataset confirmed the generalization of the framework, with the multimodal approach maintaining strong performance (F1-score of $98.13 \pm 0.36\%$) under balanced conditions.

An important part of our approach is the weighted fusion mechanism that combines features from each modality, using learned weights to determine how different representations contribute to the final analysis. Our experiments show that these weights adapt accordingly to the task complexity, maintaining balanced contributions across modalities (image: 0.3382 ± 0.0025, spectrogram: 0.3450 ± 0.0033, time series: 0.3167 ± 0.0045) to discriminate AF from normal rhythms, while showing strong specialization (time series: 0.5025 ± 0.1252) to discriminate AF from various arrhythmias. This adaptive behavior demonstrates the mechanism's ability to optimize feature usage based on specific classification challenges, contributing to robust performance in different clinical scenarios.

This research contributes to the field of automated ECG analysis by providing empirical evidence for the effectiveness of different data representations in AF detection. It was also

observed that it is not very common to find studies that address spectrograms, images and time series simultaneously. This study seeks to show the feasibility of combining these inputs and the relationship between them. The framework's ability to process standard ECG images makes it compatible with environments where only the image format is available, which could facilitate its adoption in environments with fewer resources and doctors available.

Chapter 1

Introduction

1.1 Cardiovascular Diseases: A Global Health Challenge

1.1.1 Epidemiology and Impact

Cardiovascular diseases (CVDs) encompass a broad category of conditions that affect the heart and blood vessels, with notable examples including coronary artery disease, stroke, heart failure, and AF. These conditions collectively represent a significant global health burden, contributing substantially to morbidity, mortality, and healthcare costs worldwide.

Globally, CVDs have emerged as the leading cause of death over the past few decades, surpassing other major causes such as infectious diseases, cancer, and respiratory ailments [TAA+23]. The prevalence of CVDs is particularly pronounced in high-income countries, where sedentary lifestyles, unhealthy dietary habits, and aging populations contribute to their increasing incidence. However, CVDs are not limited to affluent nations; they also pose a considerable health threat in low- and middle-income countries, where access to healthcare services and preventive measures may be limited.

In Brazil, the burden of CVDs is substantial, reflecting trends observed in other parts of the world. Brazil mirrors this global trend, grappling with a significant burden of CVD, which accounts for nearly 30% of all deaths and surpasses infectious diseases in terms of mortality rates. This stark reality translates into approximately 400,000 deaths annually, positioning CVD as the leading cause of death in the country, as reported by the Brazilian Ministry of Health [bvsnd]. Factors such as urbanization, lifestyle changes, and an aging

population have contributed to the rising prevalence of CVD risk factors, including hypertension, diabetes, obesity, and smoking.

The ECG is a low-cost, non-invasive test that records the heart's electrical activity over a short period (approximately 10 seconds). It can be recorded using 12 leads, which combine the position of electrodes located on the limbs and the front of the chest. The ECG signals' format enables the identification of various arrhythmias, heart muscle, valve, or artery problems.

1.1.2 The Rise of Atrial Fibrillation

AF, a common type of arrhythmia characterized by irregular heart rhythms, represents a significant subset of CVDs. AF is associated with an increased risk of stroke, heart failure, and other cardiovascular complications, making its early detection and management crucial for reducing morbidity and mortality rates. In Brazil, as in other parts of the world, AF prevalence rates are on the rise, reflecting demographic shifts and changes in lifestyle factors. Notably, the prevalence of AF in Brazil mirrors that of high-income countries, with estimates indicating its impact on 5 to 7 million individuals and prevalence rates ranging from 2.5% to 3.3% [Fav21].

1.1.3 Current Clinical Challenges

Given the growing burden of CVDs and AF, there is a pressing need for innovative approaches to prevention, diagnosis, and treatment. Advances in medical technology, particularly in the field of artificial intelligence (AI), hold promise for improving risk stratification, facilitating early detection, and optimizing treatment strategies for CVDs, including AF. By leveraging these technologies and integrating multidisciplinary approaches, researchers and healthcare providers can work towards mitigating the impact of CVDs and improving outcomes for affected individuals in Brazil and beyond.

1.2 Evolution of Machine Learning in ECG Analysis

1.2.1 Traditional Approaches

To optimize the diagnostic pathway within medical facilities providing remote ECG reporting services, the development of computer algorithms stands out as a promising avenue. These algorithms can be designed to efficiently categorize ECG signals into two distinct groups: those exhibiting normal cardiac electrical activity and those manifesting alterations, thereby streamlining the diagnostic process and ensuring timely intervention when warranted. Moreover, the versatility of such algorithms can be expanded to encompass advanced classifiers capable of precisely delineating the specific type of anomaly depicted on the ECG trace.

In particular, these classifiers can differentiate among various types of arrhythmias, including but not limited to AF, ventricular tachycardia, and bradyarrhythmias. Such granular discrimination empowers clinicians to promptly identify and address these aberrations in cardiac rhythm. Additionally, through adequate training, these algorithms can discern subtle indicators of acute myocardial infarction, such as ST-segment elevation or depression, facilitating expedited diagnosis and intervention in cases of acute coronary syndrome.

1.2.2 Deep Learning Architectures

Against the backdrop of an ever-evolving healthcare landscape, ML and deep learning (DL) algorithms emerge as indispensable assets in disease diagnosis, leveraging a multitude of sources and formats of ECG signals. Renowned for their adeptness in deciphering intricate patterns, these algorithms play an important role in early disease detection, particularly in the realm of cardiac ailments such as AF [MSY+21]. However, it is important to note that prevailing studies predominantly rely on unidimensional ECG signals for AF detection, despite ECG exams often being available in image format [DRM+23].

The application of deep learning techniques, particularly Convolutional Neural Networks (CNNs), has transformed the field of ECG analysis. By integrating computer vision approaches, these techniques have enhanced the accuracy and efficiency of cardiac disease detection. This integration has led to considerable advancements in automated diagnosis and

interpretation of heart conditions. For instance, the traditional interpretation of ECGs relies on clinical expertise, which can be susceptible to errors due to fatigue. Knowing that, the work of Mahmud et al. [MBI+23] addresses this issue by employing CNNs with transfer learning models to analyze both two-dimensional ECG images and one-dimensional heartbeat signals. The study trains CNNs and transfer learning models on one-dimensional heartbeat signals using ensemble techniques. Furthermore, CNNs and transfer learning models have been constructed for 2D heartbeat images, with ensemble methods employed to combine model outputs. The combined approach yielded an accuracy of 94% for ECG signal classification and 93% for ECG image classification, indicating a notable improvement in diagnostic precision. These findings underscore the efficacy of integrating 2D-CNNs, transfer learning, and ensemble methods for ECG data classification, exemplifying their potential in early cardiovascular disease detection.

1.2.3 Recent Advances in AF Detection

Over the past decades, diverse methodologies have been implemented to detect AF in ECG exams [MSY+21]. Wang et al. [WYL+23] introduced a network model with a multi-head self-attention mechanism. This model demonstrates the ability to process large amounts of data simultaneously, thereby accelerating the training process while achieving remarkable performance results. The use of multi-head self-attention allows the network to dynamically focus on different segments of the input data, facilitating improved learning and representation of intricate patterns within the dataset. However, it still has its limitations. There is no mathematical model that can be explained, namely, it is difficult to explain the features extracted from this model. Also, the division of the signals is done using traditional label division methods.

In a complementary vein, Zihlmann et al. [ZPT17] proposed two different architectures of deep neural networks aimed at evaluating AF cases within a large dataset. The first architecture consists of a CNN, while the second architecture integrates convolutional layers with long short-term memory (LSTM) layers. The effectiveness of both architectures in accurately classifying instances of AF underscores their utility in exploiting the hierarchical features present in ECG signals. The second architecture obtained an F1 score of 82.1% on the Physionet hidden challenge testing set. By integrating both spatial and temporal depen-

dencies, these architectures demonstrate robust performance in detecting subtle variations indicative of the presence of AF, demonstrating their potential for clinical use.

Moreover, Almalchy et al. [AAP20] introduce a deep learning approach for automated ECG diagnosis, focusing on AF. The methodology employs a D-CNN with transfer learning and a multiclass SVM classifier to automate the identification of ECG patterns. The proposed method uses frozen initial layers of the D-CNN to retain general feature extraction and fine-tunes the final layers for AF-specific characteristics. The study contrasts the impact of data augmentation on model performance. The model achieved 99.21% accuracy without data augmentation, demonstrating the efficacy of transfer learning.

In a recent contribution to the field, Ping et al. [PCW+20] put forth an innovative hybrid deep learning architecture, designated as 8CSL, which combines an 8-layer CNN with shortcut connections and a single LSTM layer for AF detection in ECG signals. The model exhibits effective feature extraction capabilities while addressing long-term dependencies in the data through its distinctive architectural design. The framework was evaluated using the Computing in Cardiology Challenge 2017 dataset, attaining optimal performance with a 10-second segment length (F1 score of 89.55%). The incorporation of eight shortcut connections enhanced data transmission efficiency, suggesting potential utility in clinical AF detection applications.

Nurmaini et al. [NTD+20] tackle the challenge of accurately detecting AF using single-lead ECGs. Diagnosing AF is often difficult due to overlapping features with other rhythms and signal noise. The authors propose a novel method that combines discrete wavelet transform (DWT) with 1D-CNNs to classify ECGs into three categories: normal sinus rhythm (NSR), AF, and non-AF. The model, tested on three public datasets and one from an Indonesian hospital, achieved notable results, including 99.98% accuracy in two-class classification and 99.17% accuracy in three-class classification. These results suggest that the model could significantly improve AF diagnosis in both clinical and self-monitoring settings.

To add to this, Liaqat et al. [LDZ+20] proposed a study aiming to enhance the detection of AF. The authors developed an ECG signal-processing framework that leverages machine learning and deep learning to identify AF episodes. Their experiments indicate that LSTM networks outperform other models, achieving approximately 10% better accuracy than traditional machine learning methods like support vector machines and logistic regression. This

approach aims to aid clinicians in accurately diagnosing AF, potentially reducing errors and lowering fatality rates associated with the condition.

In the pursuit of more efficient methods for detecting AF, Ma et al. [MZC+20] introduce an automatic AF detection method, CNN-LSTM, which employs deep learning to analyze ECG data. The model uses CNNs and LSTMs to extract features from ECG signals. Tested on the MIT-BIH Atrial Fibrillation Database, it achieved 97.21% classification accuracy. The CNN-LSTM approach detects AF onsets and classifies them well, making it a good solution for automatic AF classification. It is more accurate and uses fewer resources than traditional ECG classification methods, making it suitable for wearable ECG monitoring. Future research will make the network more adaptable and perform better.

1.3 Multimodal Analysis: A Complete Approach

1.3.1 Integration of Multiple Data Representations

The progress in deep learning has highlighted the growing significance of multimodal analysis in medical signal processing, especially in ECG interpretation. The ECG can be represented in different formats and each of them may offer complementary information that can be used to improve the model's performance and enhance the overall analysis, as demonstrated by Reyna et al. [RWK+24]. Multimodal frameworks leverage diverse data representations and domain-specific features, enhancing both the robustness and diagnostic precision of analytical models.

1.3.2 Advantages of Multimodal Frameworks

In healthcare, the multimodal approach has allowed for interesting results, aligned with the growing amount of data available in these settings. Some works related to this area demonstrate the potential of using multiple data modalities to improve diagnosis such as the works by Teoh et al. [TDZ+24], Carrillo et al. [CPMCS+22] and Satayeva et.al [Sat23].

In ECG analysis, this multimodal approach is particularly beneficial as it enables the integration of insights across various domains—including time-series, frequency-domain, and wavelet transformations. Each domain offers distinct perspectives on cardiac function,

contributing to a more comprehensive assessment of cardiovascular health. There are various works in the literature that use the multimodal approach like [Mob, EBE24].

Evaluating the efficacy of different data modalities is essential, given that clinical environments may have limited access to certain data formats, reinforcing the need for adaptable approaches. While time-series data remains the preferred format for ECG analysis, there are cases where only image-based representations are available due to equipment or export limitations. Facilitating ECG classification from images not only increases accessibility but also extends the applicability of diagnostic tools across diverse healthcare settings, ensuring that models remain practical and adaptable in various clinical scenarios.

1.3.3 Recent Work With Different Modalities

The use of spectrograms in ECG classification has demonstrated promising potential in recent studies. For instance, employing denoising techniques alongside Fourier transformation-based spectrograms has achieved high classification accuracy, with one study reporting an impressive 99.06% accuracy rate, thereby surpassing traditional raw signal approaches [SNP22]. These findings highlight the advantages of spectrograms in ECG analysis, underscoring their capacity to capture nuanced temporal and spectral information that may be less accessible in one-dimensional formats.

Additionally, CNNs offer a powerful framework for feature extraction, particularly suited to image-based data, where spatial patterns and structural characteristics can be identified. This capability is especially valuable for detecting complex waveform patterns or morphological features pertinent to conditions such as AF , which might not be readily apparent in raw time-series data. Recent comparative studies further emphasize the potential of time-frequency domain representations, such as scalograms and spectrograms, in ECG classification tasks. For example, one study achieved an 82.30% accuracy rate in predicting obstructive sleep apnea using ECG-derived scalograms and spectrograms, illustrating the utility of these visual representations in enhancing diagnostic accuracy [NE21].

Similarly, Bui et al. [BHP+23] introduced the TSRNet framework. This framework operates by exploiting multimodal information from both the time series and time-frequency domains through a restoration-based approach. By fusing information from different domains, TSRNet is able to capture comprehensive characteristics of ECG signals, enabling the

identification of anomalous patterns with unprecedented accuracy. TSRNet's architecture exhibits promising performance metrics, suggesting its potential for broad clinical applicability across a spectrum of cardiac diagnoses.

Additionally, Aldughayfiq et al. [AAJH23] created a hybrid deep learning framework that integrates 1D CNNs and bidirectional long short-term memory (BiLSTM) architectures for the detection of AF using photoplethysmogram (PPG) signals. Their approach is distinctive in that it integrates both ECG and PPG signals as multi-featured time series data, thereby addressing the limitations of traditional ECG-based methods while leveraging the accessibility of PPG monitoring. The proposed model exhibited robust performance, achieving 95% accuracy in AF classification, with precision, recall, and F1 scores of 88%, 85%, and 84%, respectively. Their work contributes to the under-explored transmissive PPG signal analysis area for AF detection, offering a promising non-invasive diagnostic approach.

Zhou et al. [ZF24] presents an enhanced method for automatically classifying heart arrhythmias from ECG signals. The principal innovation is the conversion of one-dimensional ECG data into multiple image types, employing techniques such as Recurrence Plot (RP), Gramian Angular Field (GAF), and Markov Transition Field (MTF), and subsequent analysis through a CNN model augmented with frequency channel attention (FCA). This multimodal approach demonstrated 99.6% accuracy in identifying five types of arrhythmias, exhibiting superior performance compared to previous methods.

1.4 Digital Transformation in ECG Analysis

1.4.1 Challenges in ECG Digitization

The digitization of ECG signals represents a significant advancement in the field of healthcare. This is particularly true given that many ECG machines rely on proprietary formats or export data solely as images or PDFs, which can impede data interoperability and comprehensive analysis. This issue is of particular concern in resource-limited regions, where a shortage of medical professionals highlights the necessity for efficient diagnostic tools and telemedicine solutions. The capacity to convert these restricted formats into standardized digital signals could significantly enhance remote diagnostic capabilities, facilitat-

ing automated analysis and supporting healthcare providers in underserved areas. Recent research has concentrated on developing accessible tools for ECG digitization, with several innovative approaches emerging in the literature to address this pressing need.

1.4.2 Modern Solutions for Signal Processing

In particular, Baydoun et al. [BSAH+19] introduce a MATLAB-based tool and algorithm for digitizing ECG signals from printed or scanned formats, enabling the use of historic ECG data in modern machine learning applications. The tool detects, extracts, and digitizes ECG data by employing image processing techniques, achieving over 95% accuracy in matching standard ECG parameters such as PR, QRS, QT, and RR intervals. This user-friendly tool supports cardiologists and researchers, offering a valuable resource for incorporating rare and historical ECG records into machine learning algorithms and enhancing diagnostic and prognostic evaluations for cardiovascular disease.

Moreover, Wu et al. [WPL+22] present a significant advancement with a fully automated online tool for digitizing 12-lead ECGs from scanned paper formats, facilitating the integration of previously unusable records into deep learning projects. The tool employs automated anchor point detection and a dynamic morphological algorithm to accurately segment and extract ECG signals. Validation on 515 ECGs, including printed, scanned, and re-digitized versions, demonstrated a 99% correlation with the original digital ECGs after the exclusion of signals with lead overlap and up to 97% correlation in specific configurations without exclusions. The tool eliminates the need for manual segmentation, making it a reliable and user-friendly solution for the large-scale digitization of paper ECG repositories for clinical and research applications.

Comparatively, Oliveira et al. [dOMW+21] developed a comprehensive algorithm for processing PDF and digital signal data from multiple sources and formats, addressing the critical need for improved accessibility in cardiac diagnostic tools. Using a sophisticated combination of CNNs and image processing methods, the algorithm efficiently extracts individual leads from PDF documents and converts them to a digital format suitable for subsequent classification tasks. The implementation of a CNN based on the ResNet architecture for classification demonstrates the algorithm's versatility and effectiveness in real-world clinical scenarios.

1.5 Research Objectives

1.5.1 Main Objective

The objective of this research is to develop and validate a comprehensive framework for automated AF classification through multimodal deep learning approaches, while also comparing the single-modality models in the process. The framework introduces a neural network architecture that processes and integrates multiple ECG representations—images, spectrograms, and time series data—to achieve robust and interpretable AF classification across diverse clinical scenarios.

The foundation of this work rests on three key aspects. First, a specialized pipeline is presented that transforms standard 12-lead ECG images into complementary modalities, with the aim of addressing the common challenge of varied ECG data formats in clinical settings. Second, a trainable weighted fusion mechanism is implemented that dynamically combines information from different modalities, allowing the model to adapt to varying signal qualities and characteristics.

The proposed framework goes beyond traditional single-modality approaches by leveraging the complementary strengths of different ECG representations. Specifically, the image modality captures spatial and morphological features, the spectrogram analysis reveals frequency-domain characteristics, and the time series data preserves temporal relationships. This multimodal integration aims to achieve a more complete representation of cardiac signals while maintaining robustness across different data formats and clinical conditions.

This research utilizes systematic evaluation in both balanced and realistic clinical scenarios to investigate the contributions of different ECG representations to AF classification. It also explores whether the combination of these representations through trainable weighted fusion can enhance diagnostic performance. The framework's development prioritizes technical performance metrics and practical considerations, such as clinical applicability.

To achieve these goals, this research addresses several fundamental questions about multimodal ECG analysis in the context of AF.

1.5.2 Research Questions

The primary investigation centers on three main aspects of multimodal AF classification:

What information does each modality capture that others miss?

The unique contributions of each modality in AF classification are examined: What distinct information does each representation capture, and how do these perspectives complement each other in the diagnostic process? Understanding these relationships can be very helpful for optimizing the fusion mechanism and ensuring that each modality contributes meaningfully to the final classification.

How do different class distributions affect each modality's performance?

An investigation regarding the impact of the class distribution on model performance is conducted: How do varying proportions of AF cases affect each modality's reliability, and how does the fusion mechanism adapt to these variations? This question is particularly relevant given the inherent class imbalance in real-world cardiac diagnostics, where AF cases typically represent a minority of observations.

What are the computational trade-offs between single-modality and multimodal approaches?

The practical implications of the multimodal integration are analyzed: What are the computational trade-offs between single-modality and multimodal approaches, and how do these affect real-world clinical applications? This investigation includes assessments of processing time, resource requirements, and the balance between model complexity and performance gains.

1.6 Structure

The study is organized into seven main chapters that address the challenges and opportunities in ECG analysis for AF classification using single-modality and multimodal approaches.

Chapter 1 introduces the global health challenge of cardiovascular diseases, with particular emphasis on AF. It presents the epidemiological context and current clinical challenges, followed by an overview of how machine learning has evolved in ECG analysis. The chapter concludes by establishing the research objectives and questions that guide this investigation.

Chapter 2 provides comprehensive background information, covering fundamental concepts of ECGs, AF characteristics, and Lead II-based detection methods. It also reviews relevant ECG databases and preprocessing techniques essential for the analysis.

Chapter 3 details the methodology, including database description, preprocessing pipelines, and data quality control measures. It presents two main experiments: AF versus Normal classification and AF versus Non-AF classification. The chapter elaborates on the data splitting strategy and model architecture designs for both experiments.

Chapter 4 presents the results of Experiment 1 (AF vs Normal Classification), comparing individual modality performances across image, spectrogram, and time series approaches. It includes detailed analyses using multiple random seeds and evaluates the effectiveness of multimodal fusion strategies.

Chapter 5 focuses on Experiment 2 (AF vs Non-AF Classification), examining the model's ability to distinguish AF from other cardiac conditions. This chapter follows a similar analytical structure to Chapter 4, evaluating individual modalities and multimodal fusion performance in this new scenario.

Chapter 6 highlights the discussion of the findings, analyzing performance metrics and clinical relevance, exploring modality fusion dynamics, and addressing generalization capabilities. It also outlines methodological insights gained and suggests directions for future research.

Chapter 7 concludes with a summary of the key achievements of the work presented, addressing the significant findings and contributions made throughout all previous chapters. Additionally, the discussion addresses the limitations encountered during the research and acknowledges factors that may have affected the results. Finally, the chapter outlines potential areas for future work, suggesting directions for further research and improvements that could enhance the overall study.

Chapter 2

Background

2.1 Introduction

This chapter provides an overview of the essential background knowledge, covering the clinical and technical aspects that inform the study. The chapter begins with a description of ECGs and their role in diagnosing AF. It then explores preprocessing techniques for ECG signals, focusing on methods such as signal processing, image preparation, and spectrogram generation. These steps ensure data quality and consistency across modalities.

A brief introduction to deep learning sets the stage for discussions on loss functions, including binary cross-entropy and focal loss, which are important for handling imbalanced datasets. This is followed by an overview of various neural network architectures, such as multilayer perceptrons, CNNs, LSTMs and inception networks. Each architecture is described in terms of its components, advantages, and specific applications to ECG analysis.

The subsequent discussion introduces multimodal neural networks, emphasizing their capacity to integrate features from multiple data representations.

The objective of this chapter is to present the necessary context for the methods and experiments described in subsequent sections, while also establishing a link between clinical relevance and computational strategies.

2.2 Electrocardiogram

The ECG signal serves as a temporal-voltage representation depicting the cardiac electrical activity over consecutive moments, derived from readings obtained via an array of electrodes, termed leads. This diagnostic modality holds significant importance in the identification and characterization of various cardiac pathologies, including myocardial infarction, ischemia, arrhythmias, and cardiomyopathies [GGS17].

A typical ECG signal manifests five primary components: the P wave, signifying atrial depolarization; the QRS complex, representing ventricular depolarization; and the ST segment, T wave, and U wave, symbolizing ventricular repolarization. Electrocardiographic analysis conventionally emphasizes the QRS complex, comprising the Q, R, and S waves, with the U wave often overlooked due to its minimal amplitude and infrequent detection in standard studies [GGS17].

Beyond these entities, ECG interpretation involves the examination of various segments and intervals. Segments delineate the inter-wave intervals, while intervals encapsulate one or more complete waves. Fundamental segments encompass the PR, ST, and TP intervals, corresponding to atrial repolarization, ventricular repolarization, and a resting state between beats, respectively. Routine interval measurements comprise the PR, QRS, QT, and RR intervals, with the RR interval commonly utilized for instantaneous heart rate determination.

The 12 standard leads are subdivided into six limb leads and six precordial leads (Fig. 2.2), each capturing cardiac activity from distinct anatomical perspectives. Limb leads (I, II, III, aVR, aVL, aVF) acquire signals from the extremities, providing a frontal plane view, while precordial leads (V1, V2, V3, V4, V5, V6) obtain signals from the anterior thorax, offering a horizontal plane perspective. This spatial divergence enables a comprehensive three-dimensional visualization of atrial and ventricular depolarization and repolarization dynamics [GGS17] (Fig. 2.1).

2.2.1 ECG Databases

A number of freely accessible biomedical research databases provide a wide range of digitized recordings of diverse physiological signals, with ECGs representing one of the most prominent categories. These repositories have become indispensable for a broad range

of scientific studies, supporting investigations into various cardiac conditions, multi-lead ECGs, and datasets recorded in diverse environments. Among these, PhysioNet stands out as a major source of open-access biomedical data, known for its minimal restrictions on usage [GAG+00]. PhysioNet was created through collaboration among researchers from multiple American institutions with the support of the NIH. It has significantly advanced the field by offering datasets encompassing a wide array of physiological signals. These include recordings from both healthy individuals and patients with conditions such as arrhythmias, neurological disorders, sleep disorders, and age-related physiological changes.

While PhysioNet is undoubtedly a valuable resource, it represents only one facet of the broader landscape of biomedical databases. A significant challenge arises from the fact that datasets of this nature are frequently reused extensively, which may result in limitations in data diversity and novelty. Additionally, a significant number of PhysioNet databases concentrate on single-lead recordings, which may impede the accuracy of diagnostic procedures for certain diseases that necessitate multi-lead analyses. Similarly, the presence of class imbalances in these datasets, resulting from the unequal prevalence of different cardiac disorders, can diminish the efficacy of machine learning models and give rise to biased predictions. Furthermore, datasets frequently lack sufficient representation of simultaneous or overlapping cardiac conditions, limiting their applicability in real-world scenarios where such overlaps are prevalent.

The importance of high-quality, diverse datasets has been highlighted by recent research. For instance, Ribeiro et al. [RRP+20] introduced a deep neural network trained on over 2 million labeled ECG exams collected by the Telehealth Network of Minas Gerais in Brazil. This large and diverse dataset, spanning 811 counties, allowed the model to achieve excellent performance. However, even with such a dataset, limitations like class imbalance and reduced data diversity significantly impacted the classifier's statistical significance and real-world applicability. Similarly, Weimann et al. [WC21] leveraged transfer learning by pre-training a CNN using the Icentia11k dataset [TAC+19], which contained over 2 billion labeled beats from 11,000 patients. Fine-tuning this network with PhysioNet's AF dataset resulted in a 6.57% performance improvement over CNNs trained solely on the smaller AF dataset, emphasizing the importance of leveraging diverse and expansive datasets for robust performance.

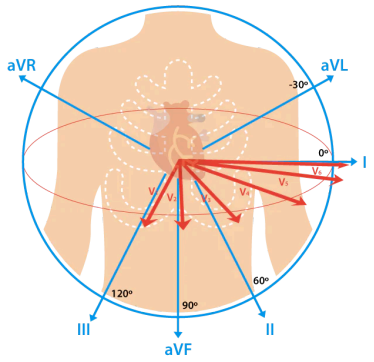

Figure 2.1: Illustration of the horizontal and vertical electrical planes. From [CS].

While PhysioNet and other large databases have made significant contributions to biomedical research, a broader exploration of datasets from diverse sources and formats is invaluable. These databases often reflect socioeconomic and geographical inequalities in access to medical resources, resulting in the underrepresentation of specific populations. The incorporation of datasets from regions with limited access to advanced medical care can provide a more comprehensive understanding of global health challenges and enhance the equity of AI-based healthcare solutions.

In order to progress, it is indispensable to prioritize datasets that encompass greater diversity. The datasets used in this work offer great data diversity and are good choices in this regard. This should be done not only in terms of geography and demographics but also in terms of recording conditions, disease prevalence, and data formats. In order to circumvent the aforementioned limitations, it is essential to expand the scope beyond well-known repositories like PhysioNet to include local, institution-specific, or crowd-sourced databases. These efforts will not only enhance the robustness of AI models but also address disparities in healthcare, paving the way for more equitable and impactful biomedical research.

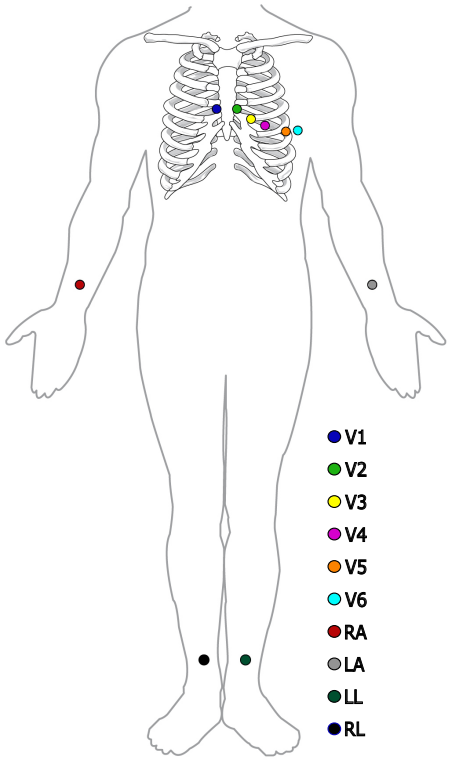

Figure 2.2: The commonly used 12-lead system. Adapted from [Ope13].

2.3 Atrial Fibrillation

AF is a common cardiac arrhythmia that is identifiable by distinct features on an ECG. Key characteristics include an irregular rhythm, absence of P waves, and variable timing of QRS complexes.

The irregular rhythm is the signature of AF, caused by disorganized electrical activity in the atria, leading to inconsistent intervals between R waves. In normal sinus rhythm, these intervals are predictable. Additionally, P waves, which indicate atrial depolarization, are missing in AF and replaced by low-amplitude fibrillatory waves that create an irregular baseline.

While the shape of QRS complexes remains largely unchanged, their timing is irregular due to unpredictable atrial conduction. This combination of features is crucial for diagnosing AF and differentiating it from other arrhythmias, guiding clinicians in risk assessment and treatment decisions. A substantial body of research has highlighted the importance of these specific ECG characteristics in the diagnosis of AF, particularly in light of its significant risk factors, such as stroke and heart failure, demonstrating the necessity for precise detection in clinical practice[CCT+23] [NGJ24].

2.4 Lead II and Atrial Fibrillation Detection

Lead II (DII) is especially beneficial for examining AF due to its alignment with the heart's electrical axis, which allows for enhanced visualization of atrial activity and fibrillatory waves. It captures distinctive characteristics that are pivotal for the diagnosis of AF, including the absence of P waves and an irregular RR interval. These characteristics are more pronounced in Lead II than in other leads, thereby facilitating improved signal clarity for the detection of atrial disorganization.

Furthermore, the longitudinal perspective of Lead II allows for the observation of atrial depolarization abnormalities, providing a reliable means of detecting fibrillatory waves and baseline irregularities, which are hallmarks of AF. It has been demonstrated in studies that the robust signal quality of Lead II makes a significant contribution to automated detection methods, increasing diagnostic sensitivity and specificity in clinical settings [Pea20, Pea19].

The prominence of this lead in AF studies highlights its essential role in both manual interpretation and machine learning algorithms designed for arrhythmia classification [Gea21, Zea20].

2.5 Recent Advances in AF Detection

Over the past decades, diverse methodologies have been implemented to detect AF in ECG exams [MSY+21]. Wang et al. [WYL+23] introduced a network model with a multi-head self-attention mechanism. This model demonstrates the ability to process large amounts of data simultaneously, thereby accelerating the training process while achieving remarkable performance results. The use of multi-head self-attention allows the network to dynamically focus on different segments of the input data, facilitating improved learning and representation of intricate patterns within the dataset. However, it still has its limitations. There is no mathematical model that can be explained, namely, it is difficult to explain the features extracted from this model. Also, the division of the signals is done using traditional label division methods.

In a complementary vein, Zihlmann et al. [ZPT17] proposed two different architectures of deep neural networks aimed at evaluating AF cases within a large dataset. The first architecture consists of a CNN, while the second architecture integrates convolutional layers with long short-term memory (LSTM) layers. The effectiveness of both architectures in accurately classifying instances of AF underscores their utility in exploiting the hierarchical features present in ECG signals. The second architecture obtained an F1 score of 82.1% on the Physionet hidden challenge testing set. By integrating both spatial and temporal dependencies, these architectures demonstrate robust performance in detecting subtle variations indicative of the presence of AF, demonstrating their potential for clinical use.

Moreover, Almalchy et al. [AAP20] introduce a deep learning approach for automated ECG diagnosis, focusing on AF. The methodology employs a D-CNN with transfer learning and a multiclass SVM classifier to automate the identification of ECG patterns. The proposed method uses frozen initial layers of the D-CNN to retain general feature extraction and fine-tunes the final layers for AF-specific characteristics. The study contrasts the impact of data augmentation on model performance. The model achieved 99.21% accuracy without data

augmentation, demonstrating the efficacy of transfer learning.

In a recent contribution to the field, Ping et al. [PCW+20] put forth an innovative hybrid deep learning architecture, designated as 8CSL, which combines an 8-layer CNN with shortcut connections and a single LSTM layer for AF detection in ECG signals. The model exhibits effective feature extraction capabilities while addressing long-term dependencies in the data through its distinctive architectural design. The framework was evaluated using the Computing in Cardiology Challenge 2017 dataset, attaining optimal performance with a 10-second segment length (F1 score of 89.55%). The incorporation of eight shortcut connections enhanced data transmission efficiency, suggesting potential utility in clinical AF detection applications.

Nurmaini et al. [NTD+20] tackle the challenge of accurately detecting AF using single-lead ECGs. Diagnosing AF is often difficult due to overlapping features with other rhythms and signal noise. The authors propose a novel method that combines discrete wavelet transform (DWT) with 1D-CNNs to classify ECGs into three categories: normal sinus rhythm (NSR), AF, and non-AF. The model, tested on three public datasets and one from an Indonesian hospital, achieved notable results, including 99.98% accuracy in two-class classification and 99.17% accuracy in three-class classification. These results suggest that the model could significantly improve AF diagnosis in both clinical and self-monitoring settings.

To add to this, Liaqat et al. [LDZ+20] proposed a study aiming to enhance the detection of AF. The authors developed an ECG signal-processing framework that leverages machine learning and deep learning to identify AF episodes. Their experiments indicate that LSTM networks outperform other models, achieving approximately 10% better accuracy than traditional machine learning methods like support vector machines and logistic regression. This approach aims to aid clinicians in accurately diagnosing AF, potentially reducing errors and lowering fatality rates associated with the condition.

In the pursuit of more efficient methods for detecting AF, Ma et al. [MZC+20] introduce an automatic AF detection method, CNN-LSTM, which employs deep learning to analyze ECG data. The model uses CNNs and LSTMs to extract features from ECG signals. Tested on the MIT-BIH Atrial Fibrillation Database, it achieved 97.21% classification accuracy. The CNN-LSTM approach detects AF onsets and classifies them well, making it a good solution for automatic AF classification. It is more accurate and uses fewer resources than

traditional ECG classification methods, making it suitable for wearable ECG monitoring. Future research will make the network more adaptable and perform better.

2.6 Preprocessing Techniques

2.6.1 Signal Processing

In the field of biomedical signal analysis, particularly in the context of ECG data, the extraction of meaningful information requires the application of processing methodologies that are capable of handling the nuances and complexities inherent in such data. The intrinsic complexity and variability of physiological signals necessitate comprehensive preprocessing methodologies to enhance signal quality and ensure reliable interpretation. In order to effectively prepare the data for analysis, multiple techniques are often combined. These methods are utilized in numerous research efforts to address the challenges encountered in biomedical signal processing [SRM18, SWM12, LLW20].

Signal standardization represents an essential preliminary step in ECG analysis. The process typically entails adjusting the sampling frequency to ensure consistency across varying data sources, as ECG signals often showcase variability in temporal resolution depending on the acquisition equipment and protocols employed. This standardization is important for maintaining uniform temporal characteristics, which enables accurate comparison across datasets and promotes both time-domain and frequency-domain analyses [SRM18]. The consistency in sampling rate is particularly valuable for preserving the intricate details of cardiac waveforms and their temporal relationships.

Baseline correction represents another essential aspect of signal processing in ECG analysis. It is not uncommon for physiological signals to display baseline wander, a low-frequency artifact that is primarily attributable to patient movement and respiratory patterns. Such interference has the potential to obscure critical diagnostic features, such as P and T waves, potentially compromising the accuracy of arrhythmia detection. To eliminate baseline drift while maintaining the signal's morphological integrity, various filtering techniques, including moving average and high-pass filters, are employed. This has been demonstrated in previous studies, for example, by Chawla et al. [CS15] and Clifford et al. [CAM05].

These methods allow for the separation of clinically significant dynamic components from unwanted baseline variations.

The application of noise reduction techniques is of considerable importance in the enhancement of signal quality. ECG signals are susceptible to a variety of forms of interference, including high-frequency noise from electrical sources, muscle activity, and environmental factors. Advanced filtering methods, such as the Savitzky-Golay filter, are effective in improving the signal-to-noise ratio while preserving essential waveform characteristics [SG64, MS12]. These approaches are of particular note for their role in maintaining the fidelity of key cardiac features, such as the QRS complex and T-wave morphology, which are fundamental to diagnostic interpretation.

Amplitude standardization addresses the variations in signal magnitude that arise from different sources, including electrode placement, patient-specific factors, and equipment variations. The process involves normalizing signal amplitudes to a consistent range while maintaining the relative proportions of salient features, such as R-wave peaks and T-wave amplitudes [LLW20, SRM18]. This standardization is of particular importance in the context of machine learning applications, where input consistency has a significant impact on model performance and reliability.

In practical implementations, a variety of software tools and libraries offer robust capabilities for signal processing. These tools facilitate the efficient application of filtering techniques, baseline correction, and amplitude normalization. The systematic implementation of these processing steps ensures that ECG signals are optimally prepared for subsequent analysis, whether for clinical interpretation or research purposes. An understanding of the principles, methods, and limitations of signal processing is fundamental to its effective application in biomedical signal analysis.

2.6.2 Image Preprocessing

In the realm of medical image analysis, the extraction of significant information necessitates the employment of advanced methodologies. The combination of multiple strategies is often required to effectively obtain data in the desired format. A considerable body of research use these techniques individually or in conjunction to address the challenges inherent in medical image analysis. [NMOHK22, WSN23, HCD+20, YLLK20].

Thresholding is a crucial technique in medical image analysis that segments pixels into foreground and background based on intensity thresholds.Thresholding is a process that categorizes pixels into foreground and background based on a predefined intensity threshold. In grayscale images, pixel intensities range from 0 (black) to 255 (white). These intensities are compared to the threshold value, with values above assigned to the foreground and values below to the background. There are various methods available for selecting the appropriate threshold value. Global Thresholding applies a single threshold value to the entire image, which is effective for images with distinct intensity differences. On the other hand, Adaptive Thresholding adjusts threshold values locally within the image, which accommodates variations in illumination or object intensities. Otsu's Method automatically determines the threshold that maximizes the variance between foreground and background pixels, making it suitable for images with complex intensity distributions.

Thresholding simplifies images into a binary format, highlighting objects of interest based on intensity differences. However, its effectiveness is influenced by factors such as image contrast, illumination, and overlapping objects. Inaccurate threshold selection can lead to misclassification of pixels, necessitating the use of more sophisticated techniques or pre-processing steps. It is pivotal to comprehend the principles, methods, and limitations of thresholding for its practical application in medical image analysis.

Filtering techniques are essential for noise reduction and feature enhancement in medical images. Edge detection filters are crucial for accurately identifying anatomical structures, which is vital in applications like tumor detection and delineation.

Morphological operations are a set of image processing techniques that allow for the modification of the shape and structure of objects within medical images. These operations are especially beneficial for segmenting blood vessels or outlining organ boundaries. They rely on two main components: binary images and structuring elements. Binary images categorize pixels as either black (0) or white (1), while structuring elements serve as templates for shape manipulation.

Dilation is one of the fundamental operations that enlarges objects within an image by overlaying the structuring element on each pixel. When any part of the element intersects with a white pixel in the original image, the pixel and its neighbors are replaced with white, thickening objects and occasionally merging nearby ones. In contrast, erosion shrinks objects

by replacing pixels and their neighbors with black if the whole structuring element coincides with white pixels in the original image. This process removes protrusions, shrinks objects, and can separate touching objects.

By combining dilation and erosion, more complex transformations can be achieved. For instance, opening, which involves erosion followed by dilation, removes small objects while preserving larger ones. Meanwhile, closing, achieved through dilation followed by erosion, fills small holes within objects while maintaining their overall shape.

While these operations can be useful for various image processing tasks, it is important to choose the parameters carefully to achieve the desired results. It is also essential to note that these techniques have limitations. Dilation, for example, can merge nearby objects, while erosion may inadvertently break thin structures.

In medical image analysis, precise object delineation is crucial for tasks such as tumor assessment, organ segmentation, and anatomical landmark identification. Contour detection methods are essential tools for accurately localizing and tracking structures within medical images. Contour detection methods systematically analyze pixel intensity variations within medical images to identify object boundaries. These methods involve algorithms that traverse the image, identifying regions of significant intensity changes or gradients that correspond to object edges. Techniques such as thresholding, edge detection, active contour models, and deep learning-based approaches are commonly employed to delineate contours [NFMS23].

In practical implementations, OpenCV [Its15] provides robust tools for preprocessing medical images. For instance, `cv2.threshold()` is commonly used for global thresholding, while `cv2.adaptiveThreshold()` facilitates adaptive thresholding. For edge detection, functions like `cv2.Canny()` effectively highlight edges by detecting intensity gradients. Noise reduction can be accomplished with `cv2.GaussianBlur()` or `cv2.medianBlur()`, which smooth images while preserving edges. Morphological operations such as dilation and erosion can be implemented using `cv2.dilate()` and `cv2.erode()`, with structuring elements created through `cv2.getStructuringElement()`. Furthermore, contour detection can be performed using `cv2.findContours()` to extract object boundaries, which are subsequently processed for precise analysis. These OpenCV functions streamline preprocessing, enabling

researchers to efficiently extract and refine critical features in medical images.

2.6.3 Spectrograms

In signal processing, a spectrogram is a visual representation of a signal's frequency characteristics as they change over time. It is widely used in various fields, including audio processing, speech recognition, radar technology, and vibration analysis, due to its analytical capabilities. Spectrograms are important in medical imaging, including Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and ultrasound imaging. They allow for the examination and visualization of tissue attributes, anatomical structures, and physiological signals such as ECG, Electroencephalography (EEG), and Electromyography (EMG) [SKG+21, ZHX+21, LQD+22, RSKS22, CRS22].

A spectrogram is derived by subjecting signal segments to Fourier transform operations across temporal spans. This process transforms signals from the temporal domain to the frequency domain, revealing the distribution of signal energy across different frequencies. The resulting spectrogram is a two-dimensional plot that shows spectral components at specific time-frequency intersections. The horizontal axis represents temporal progression, while the vertical axis represents frequency specification. The spectrogram's color or shading intensity corresponds to the spectral constituents' magnitude [OS10].

The Short-Time Fourier Transform (STFT) is a popular method for generating spectrograms. It divides the signal into overlapping short segments and computes the Fourier transform of each segment to produce a time-frequency portrayal. An alternative approach is to use the Wavelet Transform. This method decomposes the signal into wavelet coefficients across various scales and positions, resulting in a time-frequency representation similar to the STFT but with variable time and frequency resolutions.

In STFT, window functions are essential for segmenting the signal into shorter frames. Window functions, such as the Hamming, Hanning, or Blackman window, are applied to each segment to reduce spectral leakage and produce smoother frequency representations. These window functions taper the signal at the edges of each frame, minimizing distortion and artifacts caused by abrupt transitions. The selection of a window function can significantly affect the resolution and frequency response of the spectrogram, making it a crucial factor in spectrogram analysis and interpretation. Equation 2.1 and Equation 2.2 display the STFT

equation and its magnitude (spectrogram), respectively.

$$X[m, \omega] = \sum_{n=-\infty}^{\infty} x[n] \cdot w[n - m] \cdot e^{-j\omega n}, \tag{2.1}$$

where:

- $X[m, \omega]$ represents the STFT at time index m and frequency ω.
- $x[n]$ is the input signal.
- $w[n - m]$ is the window function centered at time index m.
- $e^{-j\omega n}$ is the complex exponential term representing the frequency component.

$$S[m, \omega] = |X[m, \omega]|^2, \tag{2.2}$$

where:

- $S[m, \omega]$ denotes the magnitude squared of the STFT at time index m and frequency ω.

2.7 Multilayer Perceptrons

Multilayer Perceptrons (MLPs) are a fundamental class of artificial neural networks that represent the building blocks of modern deep learning architectures [LBH15, GBC16]. These networks are designed to learn complex mappings between input data X and output Y through multiple layers of interconnected neurons, making them versatile tools for both classification and regression tasks. A visual representation of a MLP can be seen in Fig. 2.3.

2.7.1 Architectural Components

The MLP architecture consists of several key components that work together to enable effective function approximation:

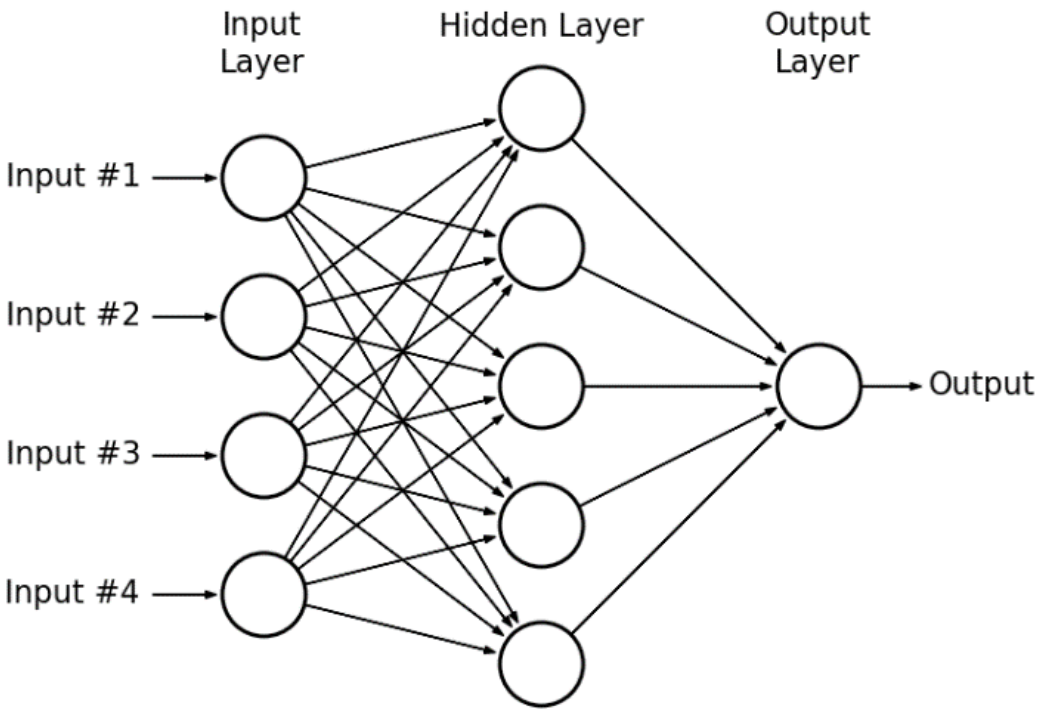

Figure 2.3: Example Multilayer Perceptron Network. From [MNZS15].

Neural Units

Each neuron processes its inputs using a weighted sum, followed by a nonlinear activation function. Mathematically, this can be expressed as:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right), \tag{2.3}$$

where:

- w_i are the weights associated with the inputs,
- x_i represent the input values,
- b is the bias term, and
- f denotes the activation function.

Activation Functions

Different activation functions serve specific purposes in the network:

- **Sigmoid:** Maps outputs to [0,1], suitable for binary classification
- **Hyperbolic Tangent:** Provides symmetric activation in [-1,1]
- **ReLU:** Offers computational efficiency and helps prevent vanishing gradients [NH10]
- **Leaky ReLU:** Addresses the "dying ReLU" problem by allowing small negative gradients
- **PReLU:** Parametric ReLU introduces learnable negative slopes, enabling the activation to adapt to the data [HZRS15].

2.7.2 Implementation Advantages

MLPs offer several key benefits for pattern recognition tasks:

- **Universal Approximation:** Ability to approximate any continuous function given sufficient capacity
- **Adaptive Learning:** Automatic adjustment of weights through backpropagation
- **Nonlinear Modeling:** Capacity to capture complex, nonlinear relationships in data

2.7.3 Optimization and Regularization

Modern MLP implementations incorporate several techniques to improve training and generalization:

- **Optimization Algorithms:** Advanced optimizers like Adam and RMSprop for efficient training [KB14]
- **Regularization Methods:** Dropout [SHK+14] and weight penalties to prevent overfitting
- **Batch Normalization:** Stabilizes training and accelerates convergence [IS15]

2.7.4 Applications in ECG Analysis

MLPs serve as crucial components in deep learning architectures for ECG analysis, particularly in the final classification stages where they integrate features extracted by specialized layers (CNN, LSTM) to make diagnostic predictions.

2.8 Introduction to Deep Learning

Deep learning has transformed the field of artificial intelligence by enabling machines to learn complex patterns and representations from raw data. It is a subset of machine learning that utilizes hierarchical structures, commonly referred to as deep neural networks, to automatically identify intricate features across multiple levels of abstraction [LBH15]. Unlike traditional machine learning models, which heavily rely on manually engineered features, deep learning models can process vast amounts of unstructured data, including images, audio, and text, to generate meaningful insights.

A fundamental concept in deep learning is neural networks, which are computational systems inspired by the structural and functional organization of biological brains. These systems consist of interconnected layers of artificial neurons, each capable of performing simple mathematical operations such as weighted summation and non-linear transformation. By stacking multiple layers of neurons, deep learning models can capture both low-level features (e.g., edges in images) and high-level semantic representations (e.g., object categories) [GBC16].

Several key advancements have contributed to the growth and development of deep learning. The availability of extensive datasets, enhanced computational resources (e.g., GPUs and TPUs), and innovative algorithmic techniques, such as backpropagation [RHW86], have made it feasible to train deep networks with millions of parameters. Furthermore, additional architectural enhancements, including CNNs [LBBH98] and RNNs [HS97a], have significantly expanded the range of applications for deep learning techniques.

2.9 Loss Functions

In the field of machine learning, loss functions play a central role, providing a means of guiding algorithms by measuring the differences between predicted and actual outcomes. These functions facilitate the iterative improvement of models by minimizing errors. In classification tasks, particularly those involving multiple classes or imbalanced datasets, the selection of a loss function can have a significant impact on a model's overall effectiveness and its ability to generalize [GBC16, Bis06].

Loss functions serve to quantify the degree of correspondence between a model's predictions and the actual values, thereby providing a numerical score that is to be minimized during the training phase. The selection of an appropriate loss function is of substantial importance, as it influences the manner in which the model behaves and performs across a range of tasks. For example, cross-entropy loss has been demonstrated to be highly effective in classification tasks due to its capacity to address problems with a greater degree of complexity than alternative approaches such as quadratic loss [JC17, JS21].

2.9.1 Binary Cross-Entropy Loss

Binary cross-entropy loss is commonly used for binary classification tasks. It evaluates the performance of models predicting probabilities between 0 and 1, making it an effective choice for binary outcomes like determining the presence of specific objects in images. Some advantages are that it is well-suited for probability distribution outputs, provides clear probabilistic interpretation of predictions and penalizes confident yet incorrect predictions more heavily than less confident ones [MMZ23, RFPLDGR18].

Binary cross-entropy loss can be represented mathematically as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right],$$

where:

- N is the number of samples,
- $y_i \in \{0, 1\}$ is the true label for the i-th sample,
- $\hat{y}_i$ is the predicted probability for the positive class (i.e., $\hat{y}_i \in [0, 1]$),

- $\log(\cdot)$ is the natural logarithm.

2.9.2 Focal Loss

Focal loss is a variant of cross-entropy loss designed to address class imbalance in classification tasks, introduced by Lin et al. [Lin17]. It reduces the dominance of easy examples in the loss calculation, ensuring greater focus on harder-to-classify samples. The focal loss function assigns higher importance to misclassified examples from minority classes, enhancing model focus on challenging examples and improving performance on imbalanced datasets.

Focal loss is defined as:

$$\mathcal{L}_{\text{Focal}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\alpha(1-\hat{y}_i)^{\gamma} y_i \log(\hat{y}_i) + (1-\alpha)\hat{y}_i^{\gamma}(1-y_i)\log(1-\hat{y}_i)\right],$$

where:

- $\alpha \in [0, 1]$ is a weighting factor to balance positive and negative classes,
- $\gamma \geq 0$ is the focusing parameter that adjusts the rate at which easy examples are down-weighted,
- The remaining terms are as defined in the binary cross-entropy equation.

2.10 Convolutional Neural Networks

CNNs represent a fundamental class of deep learning architectures that have demonstrated exceptional performance in image processing and pattern recognition tasks [RKHD23, WLF+21]. These networks are distinguished by their use of the convolution operation to automatically learn and extract hierarchical features from input data, making them very effective for tasks involving spatial or temporal patterns. The convolution process is exhibited on Fig. 2.4.

2.10.1 Architectural Components

The CNN architecture comprises two primary components that work together to enable efficient feature extraction and representation learning:

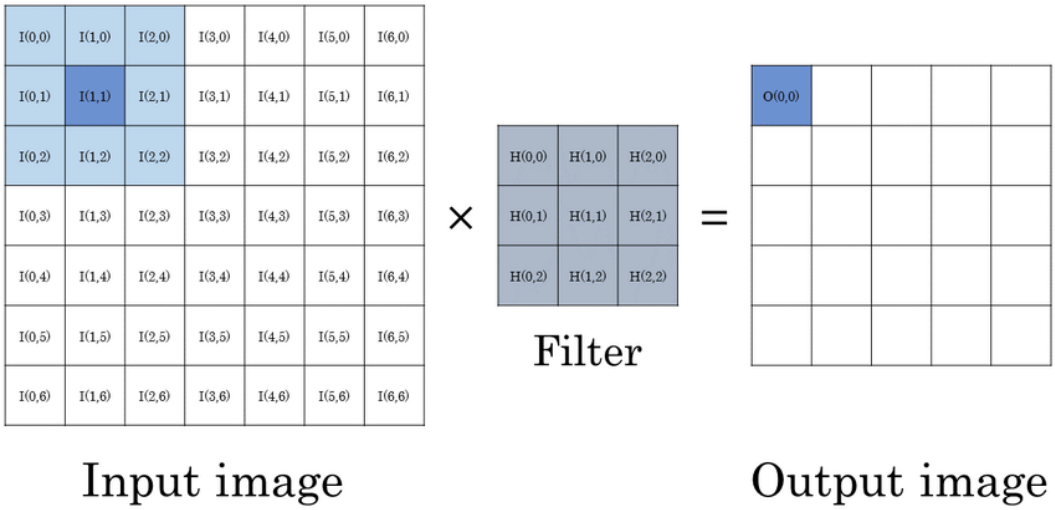

Figure 2.4: Convolution Operation on a 7x7 Matrix with a 3x3 Kernel. From [BLZ+18].

Convolutional Layers

The core building block of CNNs is the convolution operation, where learned filters (kernels) are applied across the input data to detect specific patterns. For an input x and kernel w, the convolution operation at position (i, j) is defined as:

$$(x * w)i, j = \sum m \sum_{n} x_{i+m,j+n} \cdot w_{m,n} \tag{2.4}$$

where $*$ denotes the convolution operation, and the summation is performed over the kernel dimensions.

Pooling Layers

After the convolutional layers, pooling layers are applied to reduce the spatial dimensions of the feature maps while retaining essential information. The two most common pooling methods are max pooling and average pooling. Max pooling selects the highest value within each defined spatial region, capturing the most prominent features, while average pooling computes the mean value within each region, providing a more generalized representation of the features. These operations help improve computational efficiency and reduce overfitting by downsampling the feature maps while maintaining their most relevant characteristics.

2.10.2 Implementation Advantages

CNNs offer several key benefits for pattern recognition tasks:

- **Feature Learning:** Automatic extraction of relevant features from raw input data
- **Parameter Sharing:** Efficient use of parameters through weight reuse across spatial locations
- **Translation Invariance:** Robust detection of patterns regardless of their position in the input

2.10.3 Applications in ECG Analysis

CNNs have demonstrated remarkable effectiveness in ECG analysis, particularly for tasks involving morphological pattern recognition and local feature extraction [YYZM, AKSH, RHT+]. Their ability to automatically learn hierarchical representations makes them well-suited for analyzing ECG signals, where both fine-grained details and broader patterns are important for accurate diagnosis.

2.11 Residual Neural Networks

Residual Neural Networks (ResNets) represent a significant advancement in deep learning architectures that effectively address the degradation problem in very deep networks [HZRS16, XWW+24]. These networks introduce skip connections that enable the training of deeper architectures while maintaining or improving performance. The basic residual block structure is illustrated in Fig. 2.5.

2.11.1 Architectural Components

The ResNet architecture introduces two key components that work together to enable effective training of very deep networks:

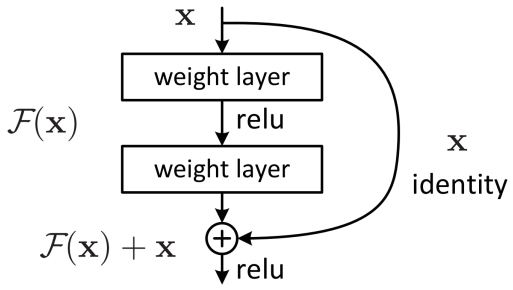

Figure 2.5: Residual Block Architecture Showing the Skip Connection. From [HZRS16].

Residual Blocks

The fundamental innovation of ResNets is the residual block, which incorporates a skip connection that bypasses one or more layers. For an input x, the output of a residual block is defined as:

$$H(x) = F(x) + x \tag{2.5}$$

where $F(x)$ represents the residual mapping to be learned and x is the identity mapping through the skip connection.

Skip Connections

Skip connections, also known as shortcut connections, facilitate gradient flow during backpropagation. These connections can be:

- **Identity Mappings:** Direct connections that add the input to the layer output
- **Projection Mappings:** Connections that use 1×1 convolutions to match dimensions when needed

2.11.2 Implementation Advantages

ResNets provide several crucial benefits for deep learning applications:

- **Gradient Flow:** Enhanced gradient propagation through deep networks, mitigating vanishing gradients
- **Optimization:** Easier optimization through residual learning, allowing for much deeper architectures

- **Degradation Mitigation:** Effective resolution of the degradation problem in very deep networks

2.11.3 Applications in ECG Analysis

ResNets have proven particularly effective in ECG analysis, where deep architectures are often necessary to capture complex temporal patterns [SGS+22, ALPP24, QZZ+24, HCPSLBV24]. The ability of ResNets to maintain gradient flow through deep architectures makes them especially suitable for analyzing long-term ECG recordings, where both local and global temporal features are crucial for accurate diagnosis and classification.

2.12 Long Short-Term Memory Networks

LSTM networks, introduced by [HS97b], represent a specialized class of RNNs designed to effectively model long-term dependencies in sequential data. These networks overcome the vanishing gradient problem that plagues traditional RNNs through a sophisticated gating mechanism, making them particularly effective for time series analysis and sequence modeling tasks [CWL+23]. A visual representation of the LSTM architecture can be seen in Fig. 2.6.

2.12.1 Architectural Components

The LSTM architecture has several key components that work together to control information flow through the network:

Gating Mechanism

The LSTM cell employs three gates to regulate information flow:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \; i_t \quad = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \; o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.6}$$

where f_t, i_t, and o_t represent the forget, input, and output gates respectively, controlling the flow of information through the cell.

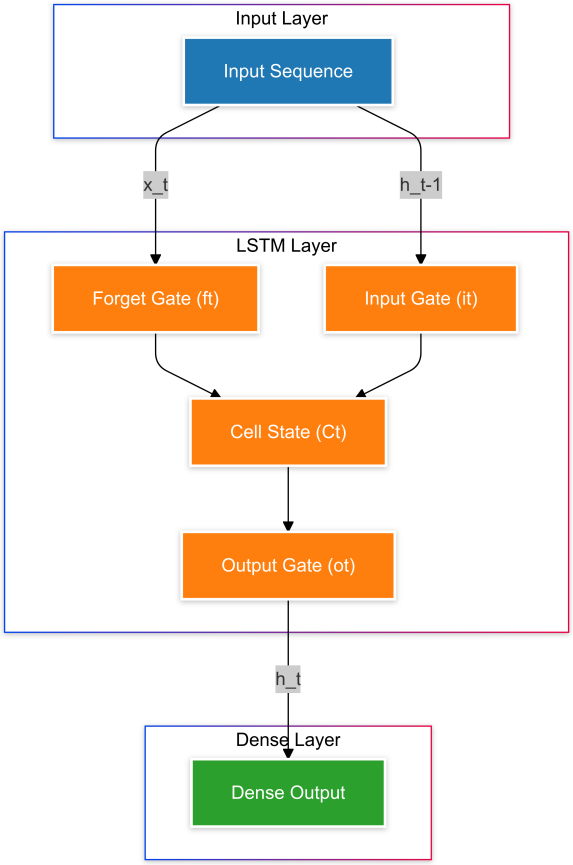

Figure 2.6: Example LSTM Network.

Memory Cell

The cell state serves as the network's memory, updated through carefully controlled operations:

$$\tilde{C}t = \tanh(W_C \cdot [ht-1, x_t] + b_C)\ C_t \quad = f_t * C_{t-1} + i_t * \tilde{C}_t\ h_t = o_t * \tanh(C_t) \tag{2.7}$$

where C_t represents the cell state and h_t the hidden state output.

2.12.2 Implementation Advantages

The LSTM architecture offers several key benefits:

- **Long-term Dependency Modeling:** Effective capture of relationships across extended sequences
- **Gradient Control:** Mitigation of vanishing gradient problems through gated information flow

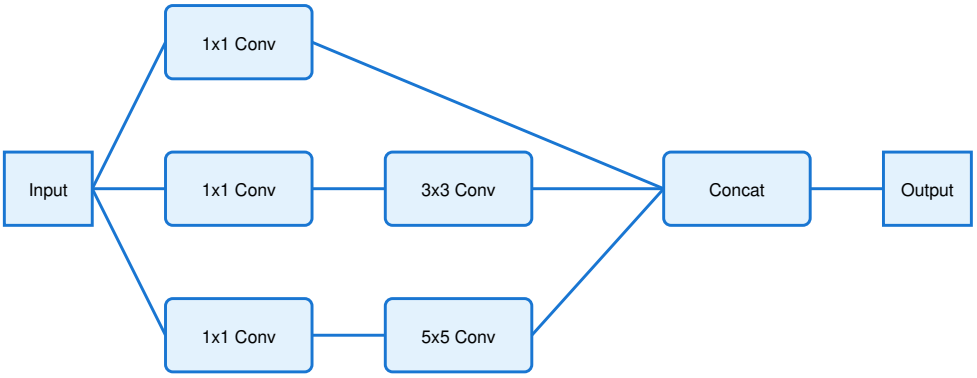

Figure 2.7: Architecture of an Inception Module showing parallel convolution paths.

- **Selective Memory:** Ability to learn which information to retain or discard

2.12.3 Applications in ECG Analysis

Recent studies have demonstrated the effectiveness of LSTM networks in ECG classification tasks [PNM23, SPH+23, SMS23]. Their ability to model temporal dependencies makes them particularly suitable for analyzing time series ECG data, where both short-term morphological features and long-term rhythm patterns are crucial for accurate diagnosis.

2.13 Inception Neural Networks

Inception networks, introduced by [SLJ+15], represent a specialized class of CNNs designed to achieve superior performance in image analysis tasks while maintaining computational efficiency. These networks are notable for their ability to process features at multiple scales simultaneously, making them appropriate for complex pattern recognition tasks in medical imaging and signal processing. The architecture of an inception layer can be seen in Fig. 2.7.

2.13.1 Architectural Components

The Inception architecture comprises several key components that work together to enable efficient multi-scale feature extraction:

Multi-Scale Convolutions

Each Inception module processes input data through parallel convolutional paths with varying filter sizes (1×1, 3×3, and 5×5), enabling the network to capture both fine-grained details and broader patterns simultaneously. A dimensionality reduction layer using 1×1 convolutions precedes larger filters to maintain computational efficiency.

Feature Integration

The outputs from parallel convolution paths are combined through concatenation along the channel dimension:

$$O_{\text{Inception}} = \text{Concat}\big(C_{1\times1}, C_{3\times3}, C_{5\times5}, P_{\text{max}}\big) \tag{2.8}$$

where $C_{n\times n}$ represents the output of each convolutional path and P_{max} denotes the max-pooling branch output.

2.13.2 Implementation Advantages

The Inception architecture offers several key benefits:

- **Multi-scale Feature Extraction:** Concurrent processing at different scales enables comprehensive feature capture
- **Computational Efficiency:** Strategic use of 1×1 convolutions reduces computational overhead
- **Adaptive Feature Learning:** The network automatically learns to emphasize the most relevant spatial scales

2.13.3 Applications in ECG Analysis

Recent studies have demonstrated the effectiveness of Inception networks in ECG classification tasks [PPC+23, TAK+23, CSC+23]. Their ability to capture both localized abnormalities and global patterns makes them particularly suitable for analyzing spectrograms, where both temporal and frequency domain features are crucial for accurate diagnosis [SWSS20].

2.14 Multimodal Neural Networks

Multimodal neural networks (MNNs) have emerged as a powerful approach for integrating heterogeneous data sources, enabling simultaneous processing of diverse input modalities such as images, time-series signals, and spectrograms [NKK+11, BAM18]. An example MNN can be seen in Fig. 2.8.

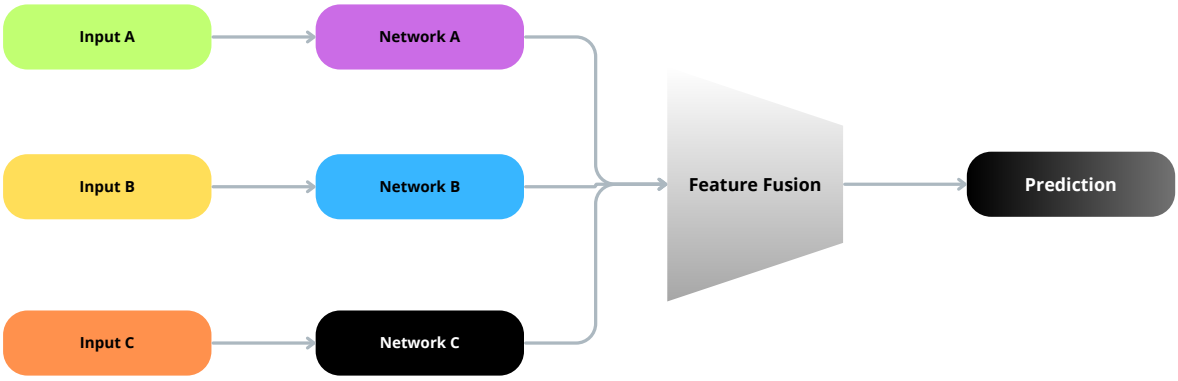

Figure 2.8: Example of a Multimodal Neural Network Architecture.

2.14.1 Architectural Components

The architecture of an MNN typically comprises three essential components:

Modality-Specific Feature Extraction

Each input modality undergoes specialized processing through dedicated neural network branches. For image data, CNNs are commonly employed, while time-series data often utilizes architectures such as LSTM networks or specialized 1D CNNs.

Feature Fusion Mechanisms

The integration of features from different modalities represents a critical component of MNN design. Three primary fusion strategies exist:

- **Early Fusion:** Combines raw inputs before feature extraction, trading computational efficiency for potential information loss
- **Late Fusion:** Merges independently processed modalities at the decision level, preserving modality independence and allowing flexibility but potentially missing cross-modal interactions
- **Intermediate Fusion:** Combines features at an intermediate stage, balancing computational efficiency with cross-modal interaction modeling

Contemporary approaches often implement trainable weighted fusion [VPP14], where modality importance is learned during training. This is important because the model can adjust the importance of each modality based on the specific task, learned weights can help mitigate the impact of noisy or less informative modalities and the learned weights can provide insights into which modalities are most important for the task. These advantages can be seen in the works by Huang et al. [HPZ+20], Mohsen et al. [MAEHS22], Teoh et al. [TDZ+24], Stahlschmidt et al. [SUS22], Zhou et al. [ZXZ22] and Su et al. [SHLC20]. The overall idea of the trainable weighted fusion can be seen in Eq. 2.9.

$$\mathbf{F}\text{fused} = \sum m = 1^{M} \alpha_m \cdot \mathbf{F}_m, \quad \text{where } \alpha_m = \text{softmax}(W_m \mathbf{F}_m + b_m) \tag{2.9}$$

Here, ($\mathbf{F}_m$) represents the feature vector for modality (m), and (α_m) denotes its learned importance weight.

2.14.2 Implementation Challenges

MNNs face several key challenges:

- **Modality Imbalance:** Different modalities may dominate due to varying dimensionality or signal quality, necessitating careful weighting strategies

- **Computational Complexity:** Processing multiple modalities increases computational demands, requiring efficient architecture design
- **Cross-Modal Alignment:** Ensuring proper temporal or spatial alignment between modalities remains challenging

2.14.3 Applications in ECG Analysis

In healthcare applications, these networks have demonstrated particular promise by combining multiple diagnostic inputs for more robust predictions [YTA+18, AK20, SLZ+23, DPC24].

Chapter 3

Methodology

3.1 Introduction

The classification of AF through the use of AI involves several steps, including data preprocessing, model design, and evaluation. This chapter describes the methodology developed for this study, utilizing a structured approach that promotes accuracy and consistency in the classification of ECGs. The methods outlined in this chapter are designed to handle multimodal data, including images, spectrograms, and time series, while addressing challenges such as data imbalance and model validation.

The chapter introduces two primary experiments: one comparing AF with normal ECGs and another distinguishing AF from non-AF cases. Each experiment involves specific steps for data preparation, model development, and training, which are described in detail.

The methodology also includes strategies for splitting the dataset, maintaining balance across training, validation, and test sets, and designing neural network architectures tailored to each data modality. These processes constitute a systematic workflow that aligns with the study's objectives and establishes a framework for subsequent chapters that focus on results and analysis. As illustrated in Fig. 3.1, the proposed methodology is presented as a diagram, offering a graphical representation of its framework.

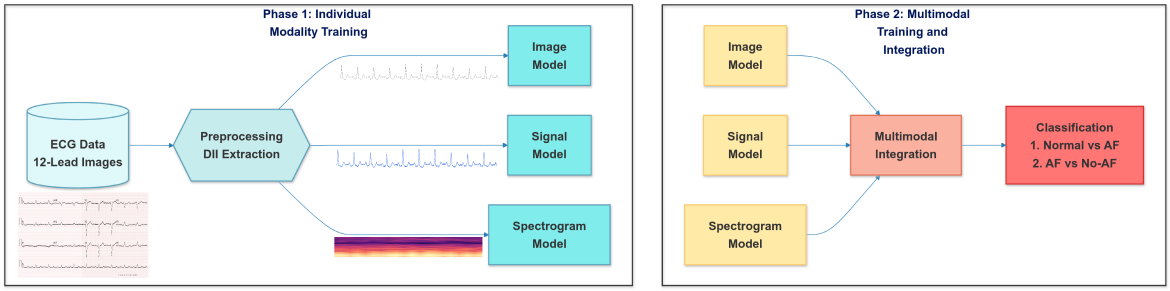

Figure 3.1: The proposed methodology for ECG-based AF classification comprises two phases. In the first phase, individual modality training is employed, which involves the processing of raw ECG data through preprocessing and DII lead extraction. This generates three parallel inputs, namely the ECG image, processed signal, and spectrogram. Each of these inputs is then processed by a dedicated model. In the second phase, multimodal training and integration are carried out. This involves retraining the individual models and then going through a multimodal integration layer, enabling the final classification between normal ECG and AF, as well as AF and non AF cases in different experiments.

3.2 Database Description and Preprocessing

3.2.1 ECG Data Characteristics

This study uses a private, image-based dataset of 12-lead ECG examinations, called InCor-DB, collected between 2017 and 2020 at a specialized tertiary referral hospital in Brazil [DRM$^+$23, DSR$^+$21]. The dataset includes over 100,000 in PNG format, each with a resolution of 3,122 x 1,671 pixels, accompanied by detailed diagnostic reports to ensure accurate labeling and analysis. This fully anonymized private dataset from the Instituto do Coração - HC FMUSP complied with all pertinent ethical regulations and received approval Institutional Review Board (IRB) under registration number CAAE 45070821.3.0000.0068. A representative example of an ECG exam is shown in Figure 3.2. While the dataset includes demographic information, this study focuses solely on the ECG data and associated diagnostic reports for classification tasks.

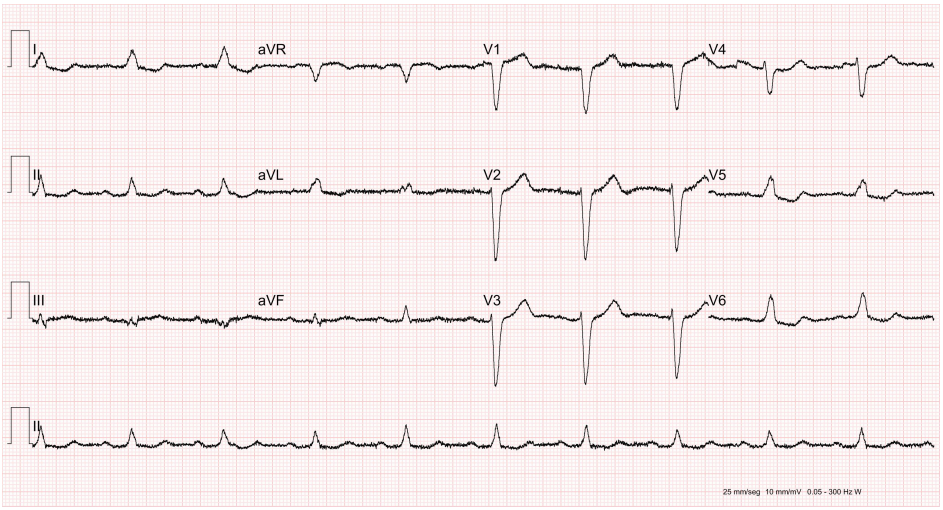

Figure 3.2: Example exam from InCor-DB.

The initial dataset contained ECG examinations with 52 distinct clinical cardiac diagnoses. Following the approach outlined in [MZC$^+$20], the focus was narrowed to AF and normal rhythm classifications. An important initial screening step involved the identification and removal of examinations with overlapping leads for the first experiment, which could compromise signal quality and interpretation. Figure 3.3 illustrates an example of such overlap, demonstrating the necessity of this quality control measure. After this screening process and removal of examinations with undefined diagnoses, these exams were identified for fur-

ther analysis.

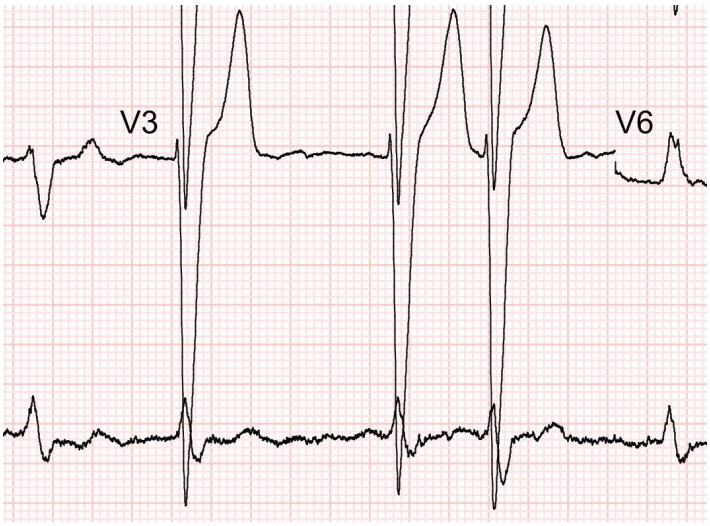

Figure 3.3: Example overlap found in an exam.

Additionally, for the purpose of external validation, a publicly available dataset [ZZD+20] corresponding to 10,646 12-lead ECGs exams was utilized. The signals were in time series format but were converted to images to mirror the process required by the private dataset. This dataset, referred to as Zheng-DB in this study, was used under similar constraints, focusing only on AF and Normal rhythm diagnoses for the first experiment and using only a subset of the exams. For the second experiment, all exams with proper diagnosis were selected. The dataset comprises eleven distinct cardiac rhythm classifications, with Sinus Bradycardia being the most prevalent (3,889 cases, 36.53%), followed by Sinus Rhythm (1,826 cases, 17.15%) and Atrial Fibrillation (1,780 cases, 16.72%). Less common rhythms include Sinus Tachycardia (1,568 cases, 14.73%), Supraventricular Tachycardia (587 cases, 5.51%), Atrial Flutter (445 cases, 4.18%), and Sinus Irregularity (399 cases, 3.75%). The dataset also contains relatively rare conditions such as Atrial Tachycardia (121 cases, 1.14%), Atrioventricular Node Reentrant Tachycardia (16 cases, 0.15%), Atrioventricular Reentrant Tachycardia (8 cases, 0.07%), and Sinus Atrium to Atrial Wandering Rhythm (7 cases, 0.07%).

3.2.2 ECG Signal Extraction and Preprocessing Pipeline

The extraction of the DII lead signal from ECG examinations required the implementation of a specialized, multi-stage digital signal processing pipeline to ensure the accurate digitization and signal quality. The process began with the segmentation of the ECG image, through which the extended DII lead region was isolated. This process relies primarily on vertical pixel density analysis and strategic signal point identification. Then, it was followed by a series of preprocessing steps, the purpose of which was to enhance signal fidelity and remove artifacts.

The process begins with an analysis of the pixel density distribution across the image. The system examines the vertical distribution of signal pixels, creating a density profile that reveals distinct patterns of ECG lead placement. This analysis identifies areas of high pixel concentration, corresponding to the ECG signal traces, and areas of low density, representing the spaces between leads. All the values used on this process were selected after carefully testing with various examples and adjusting edge cases.

Signal separation is achieved through peak detection algorithms that identify regions of maximum pixel density. The system requires a minimum threshold of 150 pixels and maintains a separation distance of 300 pixels between peaks to ensure accurate lead identification. Between these peaks, the algorithm locates points of minimum pixel density that serve as optimal cut points for lead separation.

The actual signal extraction process is performed on a column-by-column basis across the image. For each vertical column, the system identifies all signal pixels and calculates their average position, determining the center of the signal at that particular x-coordinate. This approach produces a series of coordinate pairs that trace the signal's path through the image.

To improve accuracy and eliminate potential artifacts, the extracted signal is subjected to cluster analysis using the DBSCAN algorithm. This statistical approach identifies the primary signal cluster while removing outlier points that may have been incorrectly captured during the initial extraction. The clustering process uses proximity parameters of 50 pixels to define the neighborhood and requires a minimum of 5 points to form a cluster, ensuring signal continuity and removing noise.

The initial image preprocessing employed bilateral filtering with carefully tuned param-

eters that were chosen empirically (kernel size = 9, $\sigma_{\text{color}} = 15$, $\sigma_{\text{space}} = 15$) to reduce noise while preserving edge information critical for signal detection. Subsequently, the filtered image was subjected to adaptive thresholding using the Otsu method, which resulted in the generation of a binary representation, effectively separating the ECG trace from the background grid. A morphological dilation operation with a 2×2 elliptical kernel was applied to ensure signal continuity.

The extracted coordinates underwent a series of quality control procedures. A validation step was employed whose purpose was to ensure signal continuity, and this was achieved by examining the temporal relationship between adjacent points. The algorithm was designed to address potential discontinuities in the signal by analyzing point density and spatial distribution. Signals that did not meet the requisite quality thresholds were identified for subsequent manual review or reprocessing.

The final stage of the pipeline involved signal normalization and standardization. The extracted coordinates were transformed into a time series format, with amplitude values normalized to a consistent scale. This extraction and preprocessing pipeline guaranteed the conservation of essential diagnostic characteristics while reducing the impact of artifacts and noise. The resulting digital signals retained a high degree of fidelity to the original ECG traces, providing a reliable foundation for subsequent analysis and classification tasks. Figure 3.4 illustrates the key stages of this process, from the original ECG image to the final extracted and processed signal.

3.3 Data Cleaning and Quality Control

3.3.1 Signal Quality Enhancement

The following preprocessing pipelines were implemented to ensure consistent signal quality across all examinations and each modality underwent specific quality control measures:

Figure 3.4: ECG signal extraction pipeline showing the progression from raw image to processed signal: (a) original ECG image with grid, (b) binary thresholding for signal isolation, (c) contour detection, and (d) final cleaned and normalized signal.

3.3.2 Modality-Specific Processing

Time Series Representation

The time series processing begins with the extraction of Lead II from the raw ECG data, chosen for its diagnostic value in AF detection. All signals undergo frequency standardization to 300 Hz to ensure uniform temporal resolution in relation to the InCor-DB dataset. Baseline wander is eliminated using a moving average filter with a 201-sample window, which appeared to be sufficient for the analysis and did not impact the signal integrity. The processing then continues with noise suppression through Savitzky-Golay filtering (window length = 15, polynomial order = 3) to preserve smooth signal characteristics. These values

were tested and provided good results while not being too costly. The processing concludes with amplitude normalization for consistent range and length standardization to 3010 samples through cropping or padding. This value was calculated based on the average metrics of the time series data.

Image Representation

The image processing pipeline standardizes all ECG traces to 80x320 pixel dimensions while maintaining high-resolution quality at 300 DPI. These dimensions were selected in order to keep the aspect ratio of the original image and to minimize the computational cost. Images undergo grayscale conversion with intensity normalization to the 0-1 range. Grid lines and axes are removed to reduce noise, and signal traces are optimized to achieve maximum contrast.

Spectrogram Generation

Spectrograms are generated using STFT with a Hann window function, utilizing a window length of 512 samples and 75% overlap. 512 samples offer a good balance between time and frequency resolution. 75% overlap ensures that successive windows of the signal share a significant portion of data, minimizing the loss of information at window boundaries in order to capture transient events in the ECG [JJLJ20]. The frequency range is restricted to 0.5-50 Hz, in order to focus on the most relevant frequencies and remove noise. The contrast is enhanced through percentile-based normalization. The final spectrograms are converted to grayscale, normalized for intensity, and standardized to 80x320 pixel dimensions, following the same methodology of the image representation.

3.3.3 Dataset Partitioning and Balance Control

Two experimental configurations were implemented:

Experiment 1 (AF vs Normal):

- Class balance achieved through majority class downsampling.

- 5-fold cross-validation across four different random seeds to ensure robustness. This decision was motivated by observations of seed sensitivity in preliminary experiments, where some models' performance showed bigger variation across different random initializations. The combination of multiple seeds with k-fold cross-validation provides a more comprehensive assessment of model stability and performance reliability, particularly important given the stochastic nature of neural network training. This approach helps mitigate the risk of drawing conclusions from potentially fortunate or unfortunate random splits and initializations.

Experiment 2 (AF vs Non-AF):

- Preservation of natural class distribution.
- Implementation of class weights to handle imbalance.
- Inclusion of previously excluded overlap cases.
- Single-seed 5-fold cross-validation implementation.

The tables 3.1 and 3.2 provide a detailed comparison of the InCor-DB and Zheng-DB datasets.

Table 3.1: Description of the datasets utilized for experiment 1.

Dataset	**AF**	**Normal**	**Total**
Incor-DB [DRM$^+$23]	8,447	8,447	16,894
Zheng-DB [ZZD$^+$20]	413	1,366	1,779

Table 3.2: Description of the datasets utilized for experiment 2.

Dataset	**AF**	**Non AF**	**Total**
Incor-DB [DRM$^+$23]	9,061	91,367	100,428
Zheng-DB [ZZD$^+$20]	1,780	8,866	10,646

3.4 Data Splitting Strategy (Training/Validation/Test)

The data splitting strategy employs the `StratifiedGroupKFold` from scikit-learn combined with a custom train-validation split function to maintain class distribution and prevent patient data leakage across training, validation, and test sets.

3.4.1 Details of Splitting Methodology

`StratifiedGroupKFold` ensures appropriate data partitioning through three key mechanisms: stratification to preserve class distribution across all subsets, patient-level grouping to prevent data leakage by keeping all records from the same patient within a single split, and consistent random shuffling using a fixed seed for reproducible partitioning across runs.

3.4.2 Stratified Group K-Fold Implementation

The implementation combines `StratifiedKFold` for maintaining similar class distribution across splits with `GroupKFold` to prevent patient data leakage. Using a fixed random state, `StratifiedGroupKFold` divides the data into five folds while ensuring no patient's data appears in both training and test sets for any fold. This patient-level grouping is essential in ECG classification to prevent the model from memorizing patient-specific patterns rather than learning generalizable features.

3.4.3 Cross-Validation Strategy Details

The cross-validation process uses the `StratifiedGroupKFold` function to create five folds while maintaining class distribution and patient grouping integrity. Within each fold, a customized function further divides the training set into training and validation subsets, ensuring no patient overlap. The data for each fold is then serialized and saved in TFRecord format across all modalities (images, spectrograms, time series and multimodal) to enable efficient storage and retrieval.

3.5 Model Architecture Design (Experiment 1)

This section presents specialized deep learning architectures optimized for ECG analysis across three distinct modalities and the multimodal architecture. Each architecture was designed to capture modality-specific features while trying to maintain computational efficiency. The following sections detail the architectural design for each modality in Experiment 1, focusing on Normal vs. AF classification.

3.5.1 Image-Based Architecture

The image modality architecture implements a residual network design optimized for ECG-based AF detection using image data. The network balances feature extraction capability with computational efficiency through four sequential convolutional blocks, each utilizing a dual-path structure. This design enables effective learning of both detailed ECG wave morphologies and broader rhythm patterns, while maintaining reasonable computational requirements for future multimodal integration.

Main Path:

- **Conv2D layer:** Kernel size (3,3), He initialization, and L2 regularization ($\lambda = 0.001$)
- **Batch Normalization:** Momentum = 0.99, followed by ReLU activation
- **Second Conv2D layer:** Maintaining the same number of filters
- Additional **Batch Normalization**

Residual Path:

- **1x1 Convolution:** For dimensionality matching
- **Addition:** Combined with the main path output
- Final **ReLU activation**

The filter progression follows a systematic expansion pattern (16→32→64→128) across blocks, designed to manage feature space growth. Spatial reduction is achieved through max-pooling operations (2×2) after each block, complemented by dropout (rate = 0.3) for regularization.

Classification Head:

- **Dual-Stream Pooling:** Global Average Pooling and Global Max Pooling streams, followed by concatenation
- **Dense Layer:** 128 units, L2 regularization ($\lambda = 0.01$), batch normalization, and ReLU activation
- **Dropout:** Rate = 0.3
- **Dense Layer:** 64 units, with identical regularization
- Final **Sigmoid activation layer**

A visual representation of the image-based architecture is provided in Fig. 3.5.

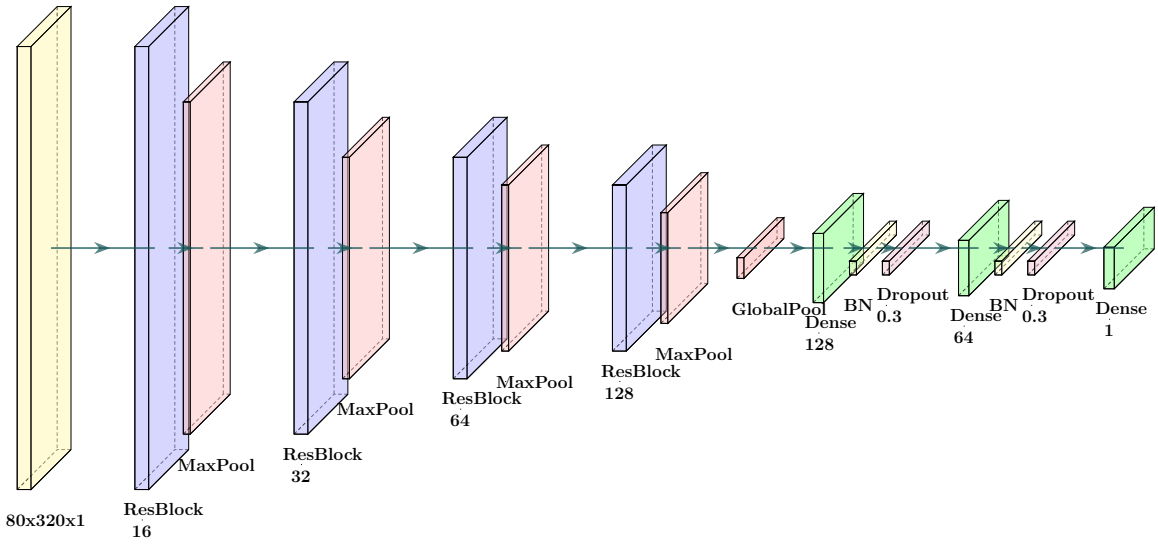

Figure 3.5: Visual representation of the image architecture for the first experiment.

The image model has been optimized for the processing of two-dimensional representations of ECG data. The four-block residual network design, with its systematic filter expansion (16→32→64→128), enables effective hierarchical feature learning - capturing both subtle wave morphologies and broader rhythmic patterns in the ECG signal.

Each block's dual-path structure plays an important role in ECG analysis. The main path, with its two consecutive 3×3 convolutional layers and batch normalization, allows for detailed feature extraction, while the residual path's 1×1 convolution preserves essential signal

characteristics through identity mapping. This architecture is promising for ECG interpretation, as it maintains fidelity to the original signal characteristics while allowing the network to learn increasingly complex feature representations.

The model's spatial reduction strategy, combining max pooling and average pooling operations after each block, is specifically tailored for ECG signal characteristics. This dual-stream pooling approach ensures that both critical peak information (such as R waves in the QRS complex) and overall wave morphology are preserved effectively. The dropout regularization (rate = 0.3) between blocks helps prevent overfitting while maintaining the network's ability to learn robust feature representations.

The classification head's design, with its parallel global average and max pooling streams followed by dense layers (128 and 64 units), enables the model to synthesize both local and global ECG features effectively. This architecture is particularly successful in capturing the multi-scale nature of ECG diagnostically relevant features, from individual wave components to rhythm-level patterns.

3.5.2 Spectrogram-Based Architecture

The spectrogram architecture implements an Inception-inspired design specifically for the analysis of time-frequency representations of ECG signals for AF classification using spectrogram data. The network uses three sequential Inception modules, each of which simultaneously processes the input at multiple scales via parallel convolutional paths. This multi-scale approach captures both localized frequency transitions and broader spectro-temporal patterns characteristic of arrhythmias, while maintaining computational efficiency through strategic filter expansion (32→64→128) and bottleneck layers.

Initial Feature Extraction:

- Input normalization via **Batch Normalization**
- **7x7 Convolution:** 32 filters

Core Architecture: Composed of three Inception modules, each containing:

- **1x1 Convolution Path**
- **Reduced 3x3 Convolution Path:** 1x1 → 3x3 convolutions

- **Reduced 5x5 Convolution Path:** 1x1 → 5x5 convolutions
- **Max Pooling Path:** With a 1x1 projection

The filter progression (32→64→128) promotes hierarchical feature learning. Each path incorporates:

- Independent **Batch Normalization**
- **ReLU Activation**
- Path concatenation for feature fusion

Classification Head:

- **Global Average Pooling**
- **Dense Layer:** 256 units, batch normalization, ReLU activation, and dropout (rate = 0.5)
- Final **Sigmoid activation layer**

A visual representation of the spectrogram-based architecture is provided in Fig. 3.6.

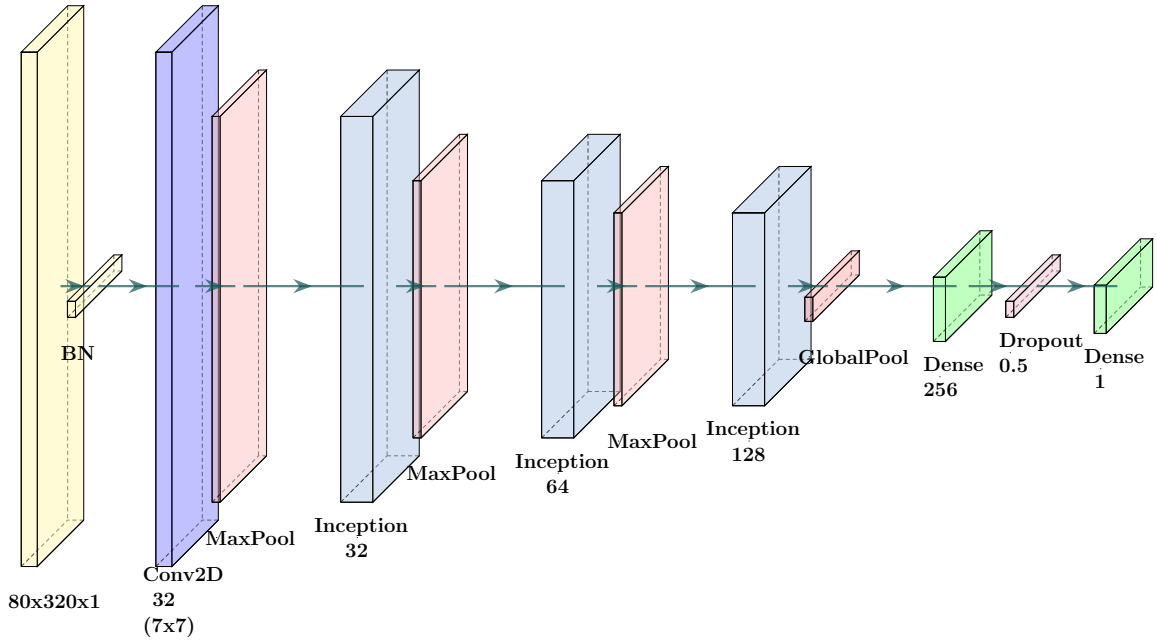

Figure 3.6: Visual representation of the spectrogram architecture for the first experiment.

The spectrogram model adopts an Inception-inspired architecture optimized for analyzing time-frequency representations of ECG signals. The initial feature extraction begins with batch normalization for input standardization, followed by a 7×7 convolutional layer with 32 filters that establishes basic feature representations. This initial wide receptive field is useful for capturing broad spectro-temporal patterns in ECG data.

The core architecture consists of three sequential Inception modules with a progressive filter expansion (32→64→128) that enables hierarchical feature learning. Each Inception module implements four parallel paths that are particularly advantageous for spectrogram analysis. The 1×1 convolution path provides efficient channel-wise feature transformation, while the reduced 3×3 and 5×5 convolution paths, each preceded by 1×1 bottleneck layers, capture features at different scales. The max-pooling path, complemented by a 1×1 projection, helps preserve salient frequency components. This multi-scale approach is appropriate for ECG spectrograms because it simultaneously processes both localized frequency transitions and broader spectro-temporal patterns characteristic of different cardiac conditions. Each pathway incorporates independent batch normalization and ReLU activation, promoting stable and effective feature learning. The subsequent concatenation of pathway outputs enables the network to effectively synthesize multiscale spectral information. This is particularly important for conditions such as AF, which exhibits distinct patterns across different time-frequency scales.

The classification head uses global average pooling to distill the learned spectro-temporal features, followed by a dense layer with 256 units that integrates this information. The inclusion of batch normalization, ReLU activation, and dropout (rate = 0.5) in this final stage helps maintain robust feature representations while preventing overfitting. This architecture can be highly effective at capturing the complex spectro-temporal patterns that characterize different cardiac rhythms in the frequency domain.

3.5.3 Time Series Architecture

The time series architecture implements a hybrid ConvLSTM design that combines convolutional and recurrent elements for sequential ECG analysis in AF detection. The network processes ECG data through two main stages: initial temporal feature extraction using strided 1D convolutions (32 and 64 filters), followed by dual bidirectional LSTM layers (64 and 32

units) to capture complex rhythm patterns. This architecture utilizes convolutional layers to capture local ECG morphologies while using bidirectional LSTMs to analyze temporal relationships across different time scales, making it effective at detecting rhythm irregularities while maintaining computational efficiency for multimodal integration.

Temporal Feature Extraction:

- **Conv1D Layer:** 32 filters, kernel size 3, stride = 2
- **Batch Normalization:** Followed by ReLU activation
- **Second Conv1D Layer:** 64 filters, stride = 2
- Additional **Batch Normalization**

Sequential Processing:

- **Bidirectional LSTM Layer:** 64 units, retaining sequences
- **Dropout Layer:** Rate = 0.3
- **Second Bidirectional LSTM Layer:** 32 units
- Additional **Dropout Layer:** Rate = 0.3

Classification Head:

- **Dense Layer:** 32 units with ReLU activation
- **Dropout Layer:** Rate = 0.2
- **Dense Layer:** 16 units
- Final **Sigmoid activation layer**

A visual representation of the time series architecture is provided in Fig. 3.7.

The time series architecture implements a hybrid ConvLSTM design specifically optimized for sequential ECG data analysis. The temporal feature extraction begins with two consecutive Conv1D layers (32 and 64 filters, respectively) using a stride of 2, which enables efficient downsampling while preserving critical ECG morphological features. This initial

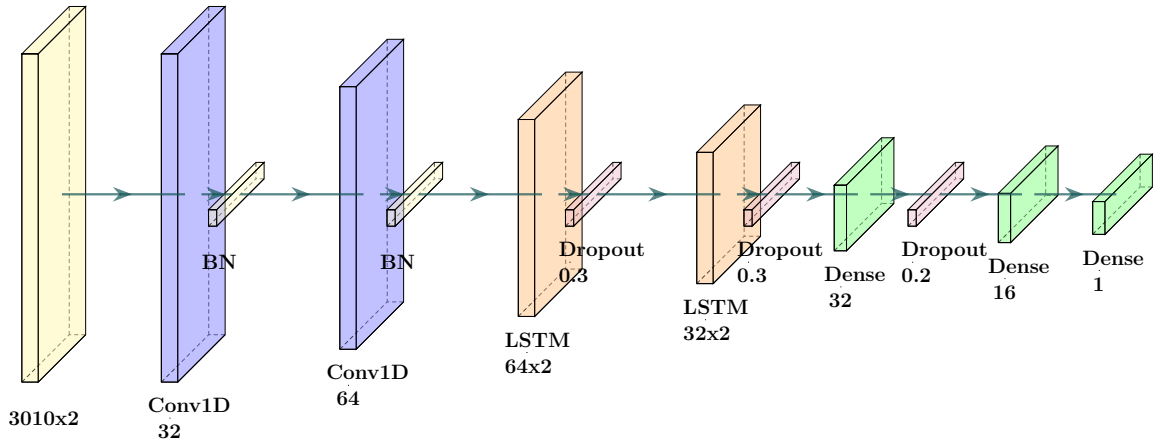

Figure 3.7: Visual representation of the time series architecture for the first experiment.

stage, complemented by batch normalization and ReLU activation, is particularly useful for capturing localized wave components such as P-waves, QRS complexes, and T-waves, while reducing sequence length for computational efficiency.

The sequential processing stage employs a dual Bidirectional LSTM structure, which proves helpful in capturing complex temporal dependencies in ECG signals. The first Bidirectional LSTM layer, with 64 units and sequence retention, enables the model to analyze both forward and backward temporal relationships in the signal. This bidirectional analysis is important for understanding how different ECG components relate to each other over time, especially for identifying rhythm irregularities characteristic of cardiac conditions. The second Bidirectional LSTM layer, with 32 units, further refines these temporal representations. Strategic dropout layers (rate = 0.3) between LSTM components help maintain robust feature learning while preventing overfitting.

The classification head uses a progressive dimensionality reduction through two dense layers (32 and 16 units) with ReLU activation. This gradual reduction, combined with a final dropout layer (rate = 0.2), enables the model to synthesize the learned temporal features effectively. This architecture is particularly adept at capturing both short-term wave morphologies and longer-term rhythm patterns in ECG signals, making it suitable for detecting conditions that manifest through temporal irregularities in the cardiac cycle.

This design strikes an effective balance between feature extraction capability and computational efficiency, making it particularly well-suited for integration into a broader multimodal framework while maintaining high performance in ECG sequence analysis.

3.5.4 Multimodal Fusion Architecture

The multimodal fusion architecture combines features from three ECG representations - images, spectrograms, and time series - through an adaptive weighting mechanism. This fusion approach allows the model to automatically determine the optimal contribution of each modality during training, potentially leading to more robust atrial fibrillation detection.

The fusion process works through a custom trainable layer that implements a weighted sum of features, using features obtained from the layers before the classification head. Initially, the weights are set equally across all three modalities (approximately 0.33 each). During training, these weights evolve through backpropagation to emphasize the most informative modalities for the classification task. The weights are processed through a softmax function to ensure they always sum to 1, making the contribution of each modality interpretable as a percentage of the final decision. This approach allows for resource-efficient integration since it requires learning only three additional parameters while potentially capturing the complementary strengths of each modality.

Trainable Weighted Fusion:

Introduced as a custom Tensorflow layer to perform dynamic weighting of the extracted features from each modality. Uses a softmax-normalized weight vector, initialized with equal weights ($[0.33, 0.33, 0.34]$), that adapt during training through backpropagation.

The fusion mechanism combines features through a weighted sum, where each modality's contribution is determined by its learned importance:

$$\text{Fused Features} = w_{\text{image}} \cdot x_{\text{image}} + w_{\text{spec}} \cdot x_{\text{spec}} + w_{\text{ts}} \cdot x_{\text{ts}} \tag{3.1}$$

This can be generalized for any number of modalities as:

$$\text{Fused Features} = \sum_{i=1}^{n} w_i \cdot x_i \tag{3.2}$$

where w_i represents the learned weight for modality i, and x_i is the feature vector from that modality. The weights are normalized using softmax:

$$w_i = \frac{\exp(w_i^{\text{raw}})}{\sum_{j=1}^{n} \exp(w_j^{\text{raw}})} \tag{3.3}$$

Here, w_i^{raw} represents the trainable weight factor before softmax normalization, ensuring all weights sum to 1 and remain positive.

Classification Head:

- **Dense Layers:** Two fully connected layers with 128 and 64 units, respectively, each incorporating:
  - **Batch Normalization** for improved convergence.
  - **ReLU Activation** for non-linearity.
  - **Dropout Regularization** to prevent overfitting (dropout rates: 0.6 and 0.4).
- **Output Layer:** A single neuron with a sigmoid activation function for binary classification.

This fusion-based architecture is designed to leverage complementary features from each modality while maintaining robustness to modality-specific features. A schematic of the complete architecture is illustrated in Fig. 3.8.

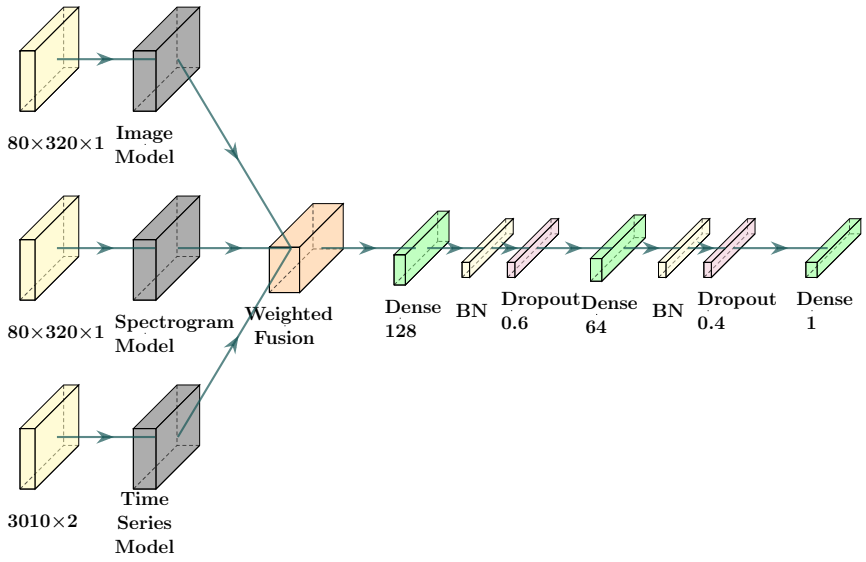

Figure 3.8: Visual representation of the multimodal architecture for the first experiment.

3.5.5 Training Configuration and Pipeline

The tests were conducted using the following software and hardware setup:

Software and CUDA Environment

The models were implemented in Python 3.12.6. The development environment utilized several key libraries, including TensorFlow 2.17.0, Keras 3.5.0, Scikit-learn 1.5.1, NumPy 1.26.4, SciPy 1.14.1, and Pandas 2.2.2. For visualization and analysis, the environment incorporated Matplotlib 3.9.2 and Seaborn 0.13.2, along with interpretability tools Shap 0.46.0 and Lime 0.2.0.1.

The tests were executed within a CUDA 12 environment, utilizing NVIDIA cuDNN 8.9.7.29, NVIDIA CUDA Toolkit 12.3, and NVIDIA TensorRT 10.4.0 for GPU acceleration and optimization.

Hardware Specifications

The training infrastructure consisted of an NVIDIA GeForce RTX 4070 GPU, paired with an AMD Ryzen 5 7600 CPU and 32 GB of DDR5 RAM operating at 6000 MT/s.

Main Training Configuration

The `CosineDecayRestarts` scheduler from TensorFlow was employed for both the single-modality models and the multimodal model, with an initial learning rate of 1×10^{-3}. The `first_decay_steps` parameter was set to 4000 for the single-modality models and 2000 for the multimodal model. These values were determined through manual tuning, based on an analysis of the models' convergence behavior.

The optimizer selected for training was the `AdamW` optimizer, a variant of the Adam optimizer that includes decoupled weight decay for improved generalization. The optimizer was configured with the following parameters:

- **Learning rate**: Controlled by a `CosineDecayRestarts` scheduler.
- **Weight decay**: 0.001
- **Gradient clipping norm (`clipnorm`)**: 1.0

- **Beta_1**: 0.9
- **Beta_2**: 0.999
- **Epsilon**: 1×10^{-7}
- **AMSGrad**: Enabled (`amsgrad=True`)

Additionally, the optimizer was wrapped with the `LossScaleOptimizer` from TensorFlow's mixed-precision training API to dynamically scale the loss during backpropagation. This configuration promotes numerical stability and takes full advantage of mixed-precision computation for enhanced performance on modern GPUs. The chosen batch size for training was 128 for all the models.

The binary cross-entropy loss function was utilized in this experiment, as it is a widely used and practical choice for binary classification tasks. Its ability to measure the divergence between predicted probabilities and true class labels makes it a good choice for distinguishing between normal ECG and AF in this case.

Since the dataset for this experiment is balanced, there is no need for the use class weights during training.

Training Pipeline

The training, validation, and test datasets were loaded from TFRecord files using customized functions tailored for each modality. Training was conducted for a maximum of 100 epochs with early stopping implemented to prevent overfitting.

The following metrics were monitored during training to evaluate the model's performance:

- **Binary Accuracy**: Measures the overall accuracy of the predictions.
- **Precision-Recall AUC (PR AUC)**: The area under the Precision-Recall curve.
- **ROC AUC**: The area under the Receiver Operating Characteristic curve.
- **Precision**: The ratio of true positives to predicted positives.
- **Recall**: The ratio of true positives to actual positives.

- **Specificity**: The ratio of true negatives to all negative outcomes.
- **F1 Score**: The harmonic mean of precision and recall.
- **True Positives (TP)**: The number of correctly predicted positive samples.
- **False Positives (FP)**: The number of incorrectly predicted positive samples.
- **True Negatives (TN)**: The number of correctly predicted negative samples.
- **False Negatives (FN)**: The number of incorrectly predicted negative samples.

Early Stopping and Model Checkpointing

For all modalities, early stopping was employed with patience of 8 epochs, focusing on the validation F1 score metric. The model checkpointing was also saved considering the validation F1 score metric.

3.6 Model Architecture Design (Experiment 2)

This section presents specialized deep learning architectures optimized for ECG analysis in three different modalities and the multimodal architecture. Each architecture was designed to capture modality-specific features while maintaining computational efficiency. The following sections detail the architecture design for each modality in Experiment 2, focusing on AF vs. Non AF classification. The complexity of the problem increased significantly, necessitating changes to the architectures and strategies to adapt to this new complexity. The architectural differences between the experiments primarily reflect an iterative optimization process throughout the study period. While alternative architectures were evaluated, the practical constraints of the research schedule and the numerous remaining experimental components precluded revisiting and re-running the entire set of Experiment 1 experiments with updated architectures. This pragmatic approach allowed the research to progress while incorporating architectural improvements where feasible.

3.6.1 Image-Based Architecture

The image model implements a deep CNN with residual connections and progressive feature extraction, optimized for processing ECG image data with efficient parameter utilization.

Input Processing:

- **Gaussian Noise Layer:** Noise factor 0.014 for robustness
- **Batch Normalization:** Momentum 0.951 for stable training
- **Initial Conv2D:** 16 filters with 3x3 kernels, maintaining spatial dimensions

Feature Extraction Backbone:

- **Progressive Filter Expansion:** Starting at 16 filters, multiplying by 1.5 across 4 blocks
- **Residual Blocks:** Each containing dual Conv2D layers with skip connections
- **Mixed Pooling:** Weighted combination of max and average pooling
- **Regularization:** L2 regularization (0.000225) and progressive dropout (0.2, 0.1, 0.2, 0.2)

Global Feature Integration:

- **Parallel Pooling:** Combined global average and max pooling
- **Dense Layers:** Two stages with 32 and 96 units
- **Dropout:** Rate 0.2 for each dense layer
- **Output:** Single unit with sigmoid activation and bias initialization -0.2

A visual representation of the image model architecture is provided in Fig. 3.9.

The image model has been optimized for the processing of two-dimensional representations of ECG data. The progressive filter expansion across four blocks (initiating with 16 filters and multiplying by 1.5) enables the construction of increasingly intricate feature representations, which is significant given that ECG images contain both low-level features

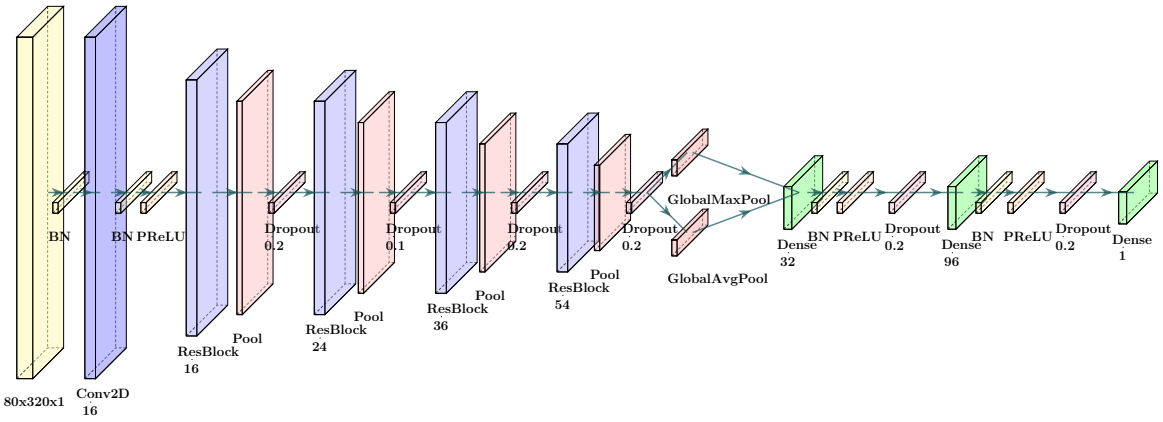

Figure 3.9: Visual representation of the image model architecture showcasing the progressive feature extraction and residual connections for the second experiment.

(such as individual wave shapes) and high-level patterns (like rhythm variations across multiple beats). The mixed pooling approach in the image model, combining max and average pooling with learned weights, is particularly effective for ECG analysis. Max pooling helps capture important peak values (like R peaks in the QRS complex), while average pooling helps maintain information about the overall morphology of the waves. The residual connections are also valuable because they help preserve information about the original signal characteristics while allowing the network to learn additional features, which is important because both the basic wave shapes and their subtle variations can be diagnostically significant.

3.6.2 Spectrogram-Based Architecture

The inception-inspired model employs a dual-path architecture that processes features at multiple scales simultaneously, combining local and global pattern recognition capabilities.

Input Processing:

- **Initial Normalization:** Batch normalization and Gaussian noise (0.015)
- **Feature Extraction:** 48 filters with 7x7 kernels
- **Regularization:** Combined L1 (1e-6) and L2 (5.0966e-6) regularization

Dual-Path Network:

- **Local Path:** Two residual blocks with 3x3 kernels (64, 128 filters)
- **Global Path:** Two residual blocks with 7x7 kernels (64, 128 filters)
- **Path Dropout:** Graduated rates (0.3, 0.2) for each path

Feature Integration:

- **Path Fusion:** Concatenation of local and global features
- **Context Module:** 192-unit dense layer with global average pooling
- **Final Extraction:** 256 filters with 3x3 kernels
- **Classification Head:** Progressive dense layers (192, 128, 64, 1)

A visual representation of the spectrogram architecture is provided in Fig. 3.10.

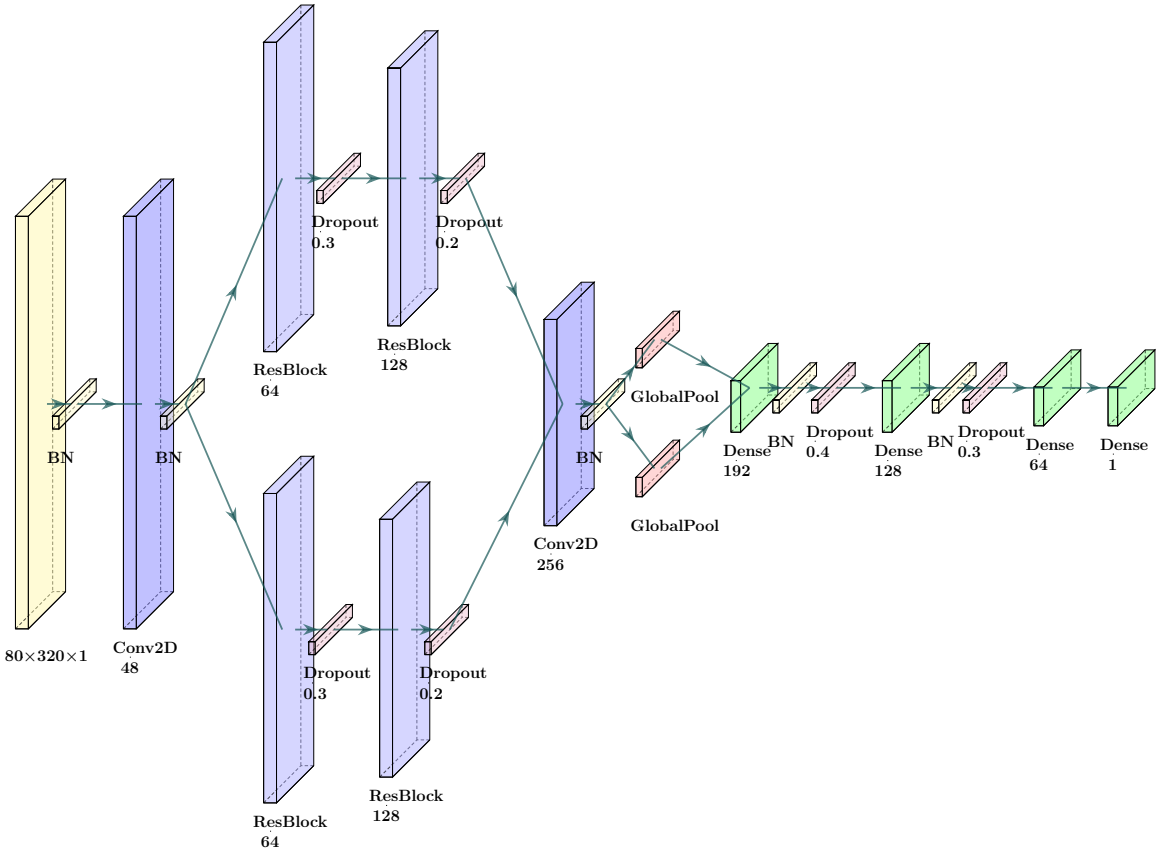

Figure 3.10: Visual representation of the spectrogram architecture highlighting the dual-path processing strategy for the second experiment.

The inception-inspired model is particularly well-suited for processing spectrograms of ECG signals. The dual-path architecture with different kernel sizes (3×3 and 7×7) is specifically advantageous in this context, as spectrograms of ECG data include both fine-grained frequency details and broader spectro-temporal patterns. The local path, comprising 3×3 kernels, is capable of capturing precise frequency transitions and local spectral features. In contrast, the global path, utilizing 7×7 kernels, is adept at capturing wider temporal and frequency relationships. This is of great importance as a variety of cardiac conditions manifest as specific patterns in the frequency domain over time. The context module in the inception model, incorporating global average pooling and a dense layer, enables the model to comprehend the overall spectral distribution of the signal. This is especially beneficial for detecting conditions that affect the overall frequency composition of the ECG, such as AF, which typically exhibits distinguishing frequency patterns. The progressive dense layers then promote the integration of these different levels of spectral information for final classification.

3.6.3 Time Series Architecture

The Time series model implements a memory-efficient hybrid architecture combining convolutional and recurrent elements for processing sequential data with spatial correlations.
Efficient Downsampling:

- **Initial Conv1D:** 48 filters with stride 4 for aggressive downsampling
- **Batch Normalization:** Momentum 0.99 for training stability
- **Activation:** PReLU with shared axes for adaptive learning

Feature Extraction:

- **Separable Convolutions:** Three blocks (64, 128, 256 filters)
- **Progressive Pooling:** MaxPooling1D with size 2
- **Regularization:** L2 regularization with graduated constraints
- **Dropout:** Variable rates (0.1, 0.2, 0.1) across blocks

Temporal Processing:

- **LSTM Layer:** 64 units with sequence retention
- **Attention Mechanism:** Dense-based attention for temporal weighting
- **Global Features:** Combined average and max pooling
- **Dense Integration:** 64 units with PReLU activation

A visual representation of the signal architecture is provided in Fig. 3.11.

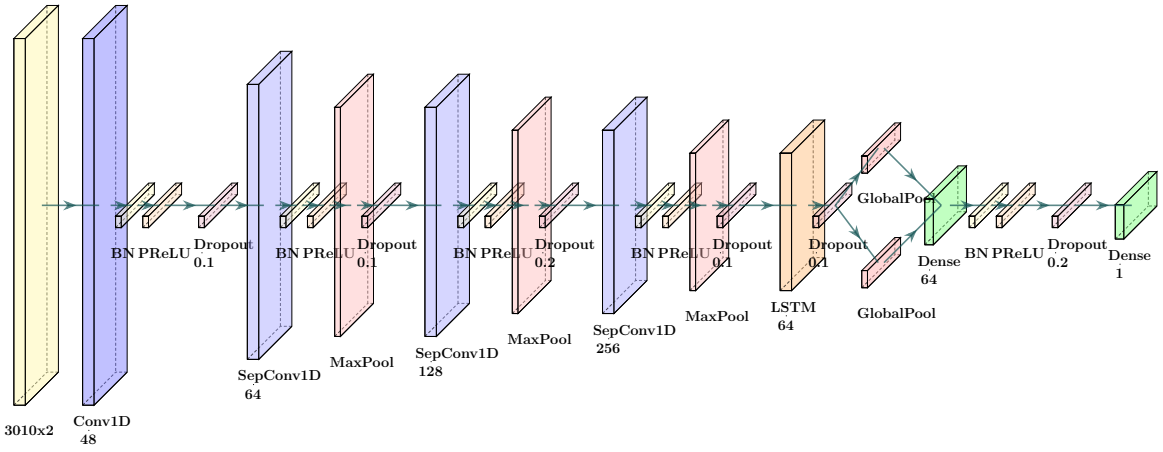

Figure 3.11: Visual representation of the signal architecture showing the hybrid temporal-spatial processing pipeline for the second experiment.

The ConvLSTM model has been developed with the specific purpose of processing time series ECG data, and its architectural design addresses a number of significant challenges typically encountered in ECG processing. The initial convolutional layers with aggressive downsampling are of significance as they are capable of capturing local patterns in the ECG signal, including P-waves, QRS complexes, and T-waves, while simultaneously reducing the sequence length, thus facilitating more efficient subsequent processing. The separable convolutions are particularly effective in this context, given that ECG patterns often have a hierarchical structure, comprising smaller patterns (such as R peaks) that combine to form larger patterns (such as QRS complexes). The use of separable convolutions enables the model to efficiently learn these patterns at varying scales while utilizing fewer parameters. This is a valuable consideration, as a multimodal model will be explored with greater depth at a later stage. The subsequent LSTM layer is crucial for capturing longer-term dependencies

in the ECG signal. This is crucial because cardiac conditions often manifest in the relationships between different parts of the signal. For example, the manner in which T-waves relate to previous QRS complexes, or how rhythm patterns evolve over time, can be indicative of underlying cardiac issues. The attention mechanism further enhances this by allowing the model to focus on the most relevant parts of the sequence when making predictions, which is particularly useful for detecting irregular events or anomalies in the ECG.

3.6.4 Multimodal Fusion Architecture

The multimodal fusion architecture integrates diverse data modalities—images, spectrograms, and time series—through a trainable weighted fusion mechanism. This design enables the model to dynamically learn the optimal contribution of each modality during training, thereby improving classification performance across heterogeneous inputs. To enhance computational efficiency, the architecture incorporates streamlined layers, building on the optimizations implemented in the prior experiment but keeping the overall architecture unchanged. Additionally, hyperparameter tuning was conducted using Keras Tuner with Bayesian Optimization, promoting a more robust and well-calibrated model for the increased complexity of this experiment. The same implementation of the trainable weighted fusion layer from Experiment 1 was used here.

Classification Head:

- **Dense Layers:** Two fully connected layers with 96 and 48 units respectively, each incorporating:
  - **Batch Normalization** with momentum 0.95 for stable training.
  - **ReLU Activation** for non-linearity.
  - **L2 Regularization** ($\lambda = 1e - 4$) and He initialization for robust learning.
  - **Bias-free design** to reduce parameters.
  - **Dropout Regularization** (rate = 0.3) on the first dense layer.
- **Output Layer:** A single neuron with sigmoid activation and calibrated negative bias (-0.2) for binary classification.

This fusion-based architecture is designed to leverage complementary features from each modality while maintaining computational efficiency through careful parameter management and regularization strategies. A schematic of the complete architecture is illustrated in Fig. 3.12.

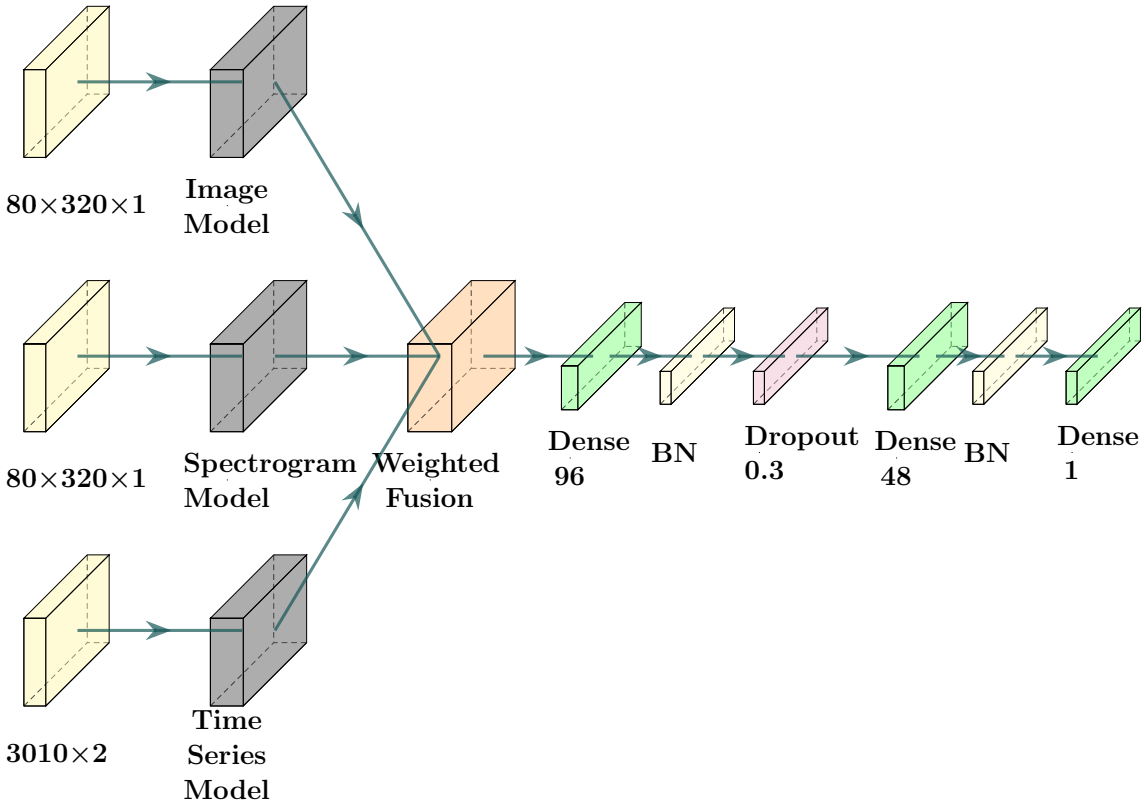

Figure 3.12: Visual representation of the multimodal architecture for the second experiment.

3.6.5 Training Configuration and Pipeline

Software and CUDA Environment

The software configuration was the same for the first experiment.

Hardware Specifications

The hardware configuration was the same for the first experiment.

Main Training Configuration

For the single-modality models, no learning rate scheduler was applied during this experiment. Instead, an initial learning rate of 1.4281×10^{-4} was used. To aid the training process in overcoming local minima, the `ReduceLROnPlateau` callback was employed, dynamically reducing the learning rate when the monitored metric plateaued.

For the learning rate schedule selected in the multimodal model training, the following configuration was implemented to address the increased complexity and class imbalance challenges:

- **Initial learning rate**: 6×10^{-3}
- **Warmup steps**: 15 epochs (approximately 15,045 steps)
- **First decay steps**: 20 epochs (approximately 20,060 steps)
- **Cycle multiplier (`t_mul`)**: 2.0
- **Magnitude multiplier (`m_mul`)**: 0.85
- **Minimum learning rate (`alpha`)**: 1×10^{-4}

The schedule implements a two-phase approach, as outlined in the following section. The initial warmup phase employs a linear increase from zero to the maximum learning rate, which is crucial for stabilizing the early training process across the multiple modality branches. Subsequently, a cosine decay with restarts phase is initiated, where the learning rate follows a cyclical pattern, marked by gradually increasing periods and decreasing magnitudes.

This design choice was motivated by a number of factors. The higher initial learning rate, in comparison to the single-modality models' constant rate of 1.4281×10^{-4}, provides a more effective means of exploring the weight space, which is of particular importance for the optimization of the fusion layer parameters. The warmup phase serves to attenuate the potential for instability that may arise from the higher learning rate, while the subsequent cosine decay with restarts enables navigation of the complex loss landscape that is characteristic of multimodal architectures and class imbalance.

The selection of specific parameters, such as the cycle multiplier and magnitude multiplier, was guided by the necessity to achieve a balance between exploration and exploitation in the training process. The gradual increase in cycle length (controlled by `t_mul`=2.0) provides progressively longer periods for fine-tuning, while the decay in restart magnitude (`m_mul`=0.85) ensures a smooth transition from exploration to exploitation phases.

In contrast, the single-modality models employed a more straightforward approach, utilizing a fixed learning rate without a scheduled adjustment. This strategy proved sufficient for their less complex architectures and more straightforward optimization landscapes.

The optimizer selected for training was the `AdamW` optimizer, a variant of the Adam optimizer that includes decoupled weight decay for improved generalization. The optimizer was configured with the following parameters:

- **Learning rate**: 1.4281×10^{-4}.
- **Weight decay**: 2.8326×10^{4} and 1×10^{-4} for single-modality and multimodal, respectively
- **Gradient clipping norm (`clipnorm`)**: 1.0
- **Beta_1**: 0.9
- **Beta_2**: 0.999
- **Epsilon**: 1×10^{-7}
- **AMSGrad**: Enabled (`amsgrad=True`)

Additionally, the optimizer was wrapped with the `LossScaleOptimizer` from TensorFlow's mixed-precision training API to dynamically scale the loss during backpropagation. This configuration promotes numerical stability and takes full advantage of mixed-precision computation for enhanced performance on modern GPUs. The chosen batch size for training was 128 for all the single-modality models. It was reduced to 64 for the multimodal test due to the increase in complexity for both the model's architecture and the dataset size.

In order to address the significant class imbalance present within the dataset, focal loss was selected as the loss function for the purposes of this experiment. Focal loss is particularly well-suited for addressing the challenges posed by imbalanced datasets. It places a greater emphasis on training samples that are more difficult to classify, effectively down-weighting the contribution of well-classified examples. This helps to offset the dominance of the majority class and ensures that the model dedicates sufficient attention to the minority class.

The focal loss was configured with an α parameter derived from the class weights to balance the importance of the positive class (AF) relative to the negative class (Non AF). Based on the dataset, the normalized α value for the positive class was calculated as approximately 0.916. The γ parameter, which controls the rate at which easy examples are down-weighted, was set to 3.0 for the single-modality models and 2.0 for the multimodal model, emphasizing difficult samples during training.

Additional techniques were applied to improve model performance and robustness:

- **Label Smoothing**: A value of 0.1 was used to prevent overconfidence in predictions by softly distributing the label probabilities.
- **Reduction Method**: The `sum_over_batch_size` reduction was employed to ensure loss values were aggregated consistently across batches.

Training Pipeline

The training, validation, and test datasets were loaded from TFRecord files using customized functions tailored for each modality. Training was conducted for a maximum of 150 epochs, and early stopping was implemented to prevent overfitting. This increase in the maximum epochs was due to the increased complexity of the problem, which allowed the model more time to learn, if necessary.

The following metrics were monitored during training to evaluate the model's performance:

- **Binary Accuracy**: Measures the overall accuracy of the predictions.
- **Precision-Recall AUC (PR AUC)**: The area under the Precision-Recall curve.

- **ROC AUC**: The area under the Receiver Operating Characteristic curve.
- **Precision**: The ratio of true positives to predicted positives.
- **Recall**: The ratio of true positives to actual positives.
- **Specificity**: The ratio of true negatives to all negative outcomes.
- **F1 Score**: The harmonic mean of precision and recall.
- **True Positives (TP)**: The number of correctly predicted positive samples.
- **False Positives (FP)**: The number of incorrectly predicted positive samples.
- **True Negatives (TN)**: The number of correctly predicted negative samples.
- **False Negatives (FN)**: The number of incorrectly predicted negative samples.

Early Stopping and Model Checkpointing

In the case of the single-modality models, early stopping was implemented with a patience of 15 epochs, with the focus being on the validation F1 score metric. However, in the case of the multimodal model, early stopping was implemented with a patience of 20 epochs, with the same focus on the validation F1 score metric, due to the increased difficulty in converging. All models used a model checkpointing callback, with the same focus on the validation F1 score metric.

Chapter 4

Experiment 1 - AF vs Normal Classification

4.1 Introduction

This chapter presents the key results of the AF classification models across different modalities. Four distinct approaches are evaluated: image-based analysis, spectrogram analysis, time series analysis, and a multimodal fusion approach. Each model was tested with four different random seeds (42, 73, 99, 122) to ensure robust evaluation and assess the stability of the results. The individual details can be found in the appendices.

4.2 Dataset Overview

This experiment utilized a 5-fold Stratified Group K-Fold cross-validation strategy, repeated across four different random seeds. The dataset statistics are summarized in Table 4.1.

Table 4.1: Dataset statistics across random seeds for experiment 1.

Seed	Split	Samples/Fold	Normal (%)	AF (%)	Class Ratio
42	train	10836 ± 26	49.91 ± 0.35	50.09 ± 0.35	1.004 ± 0.014
	val	2680 ± 26	50.38 ± 1.39	49.62 ± 1.39	0.986 ± 0.056
	test	3379 ± 1	50.00 ± 0.00	50.00 ± 0.00	1.000 ± 0.000
73	train	10833 ± 21	49.77 ± 0.31	50.23 ± 0.31	1.009 ± 0.012
	val	2682 ± 22	50.93 ± 1.24	49.07 ± 1.24	0.964 ± 0.049
	test	3379 ± 1	50.00 ± 0.00	50.00 ± 0.00	1.000 ± 0.000
99	train	10808 ± 31	50.04 ± 0.15	49.96 ± 0.15	0.998 ± 0.006
	val	2707 ± 32	49.83 ± 0.59	50.17 ± 0.59	1.007 ± 0.024
	test	3379 ± 1	50.00 ± 0.00	50.00 ± 0.00	1.000 ± 0.000
122	train	10826 ± 23	49.99 ± 0.21	50.01 ± 0.21	1.000 ± 0.009
	val	2689 ± 22	50.04 ± 0.87	49.96 ± 0.87	0.999 ± 0.035
	test	3379 ± 1	50.00 ± 0.00	50.00 ± 0.00	1.000 ± 0.000

Values are presented as mean $\pm$ standard deviation across folds.

Class ratio is calculated as AF / Normal.

4.3 Fusion Trainable Weight Analysis

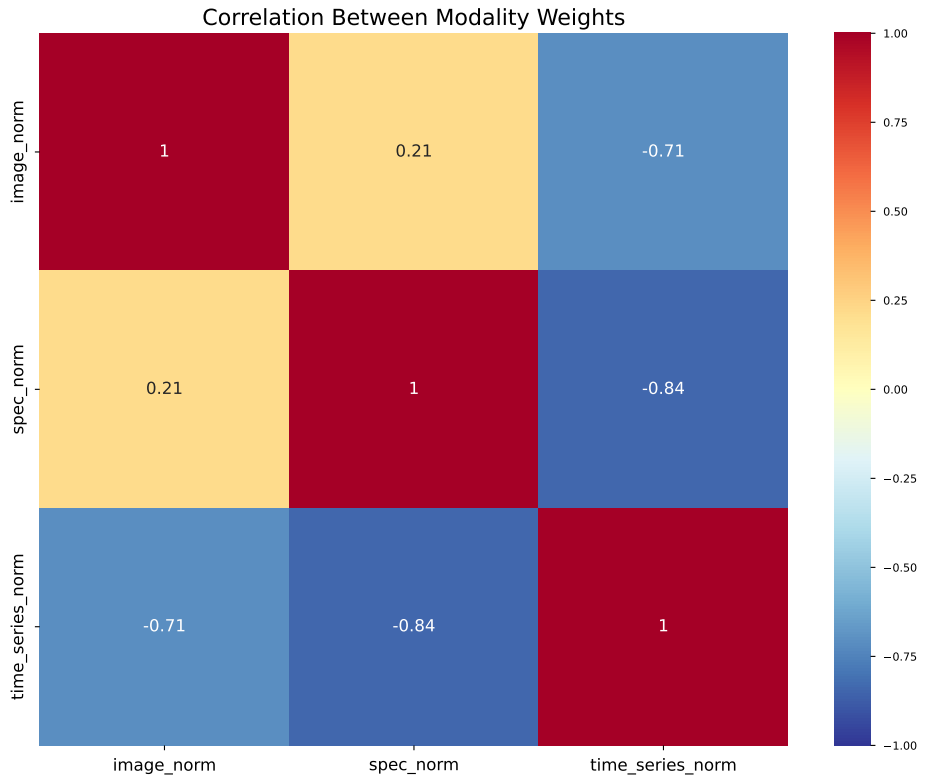

Figure 4.1: Correlation heatmap between all the modalities.

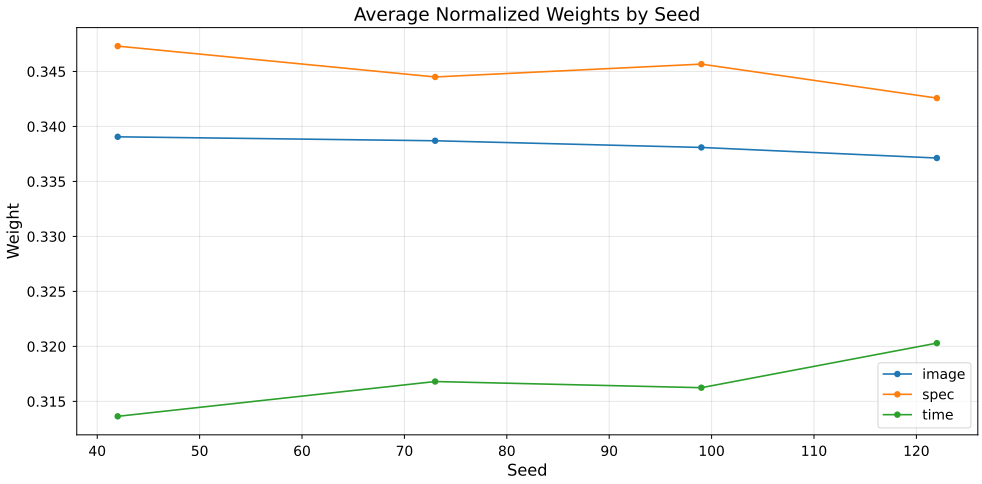

Figure 4.2: Averaged normalized weights by seed.

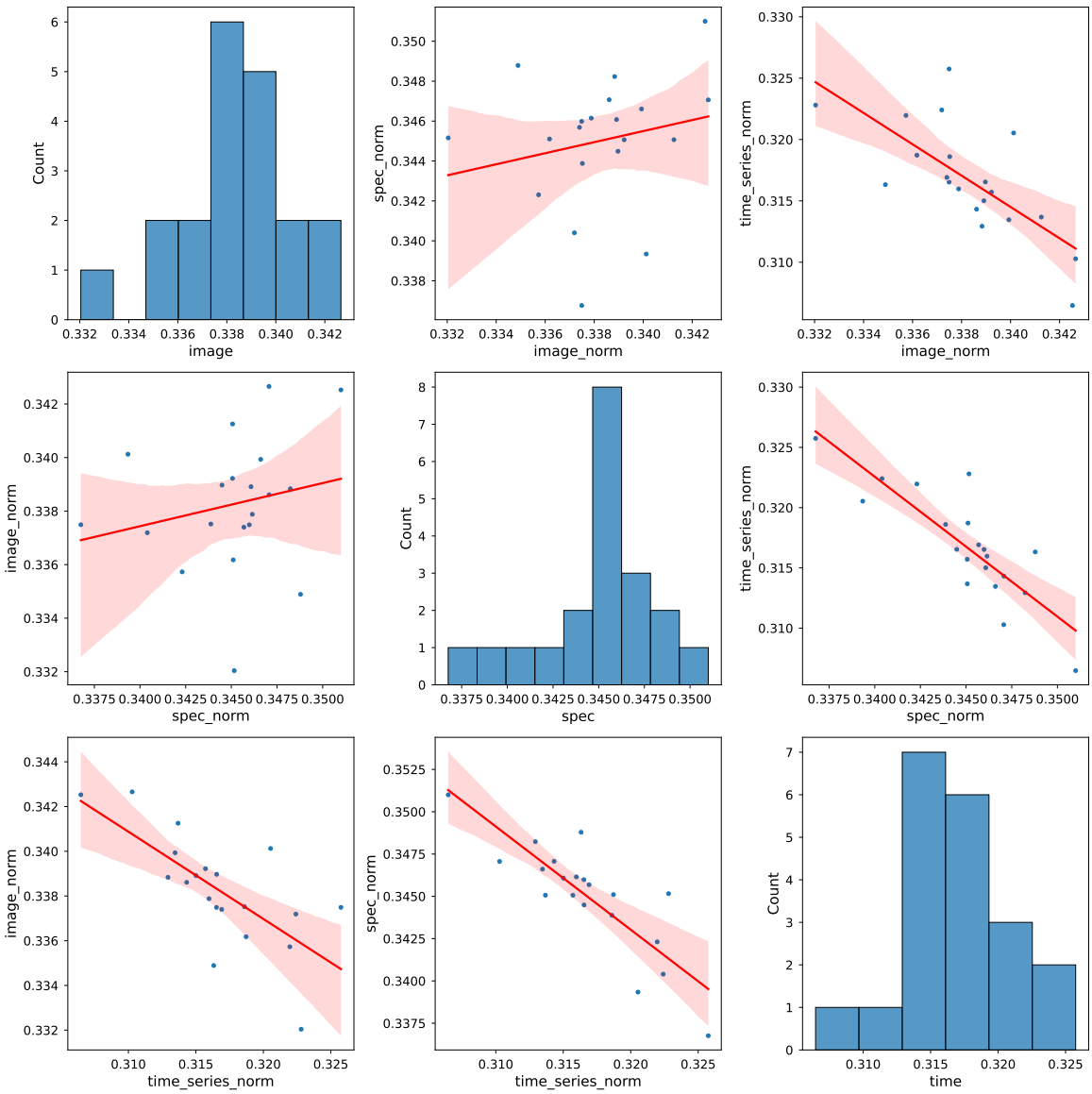

Figure 4.3: Analysis of modality fusion weights across five cross-validation folds and four seeds. The diagonal shows weight distributions for image, spectrogram, and time series modalities. Off-diagonal plots display pairwise correlations between modalities with fitted regression lines (red).

4.4 Cross-modality Comparison

To enable direct comparison between modalities, a table with the performance metrics for each seed is presented.

Table 4.2: Performance metrics across modalities for seed 42.

Metric	Image	Spec	Time Series	Multimodal
Accuracy	0.989	0.981	0.983	0.993
Precision	0.993	0.980	0.982	0.993
Recall	0.985	0.982	0.985	0.993
F1 Score	0.989	0.981	0.983	0.993
Specificity	0.993	0.980	0.982	0.993
Roc Auc	0.999	0.996	0.995	0.998
Pr Auc	0.999	0.995	0.994	0.998

Table 4.3: Performance metrics across modalities for seed 73.

Metric	Image	Spec	Time Series	Multimodal
Accuracy	0.990	0.981	0.980	0.993
Precision	0.992	0.982	0.977	0.992
Recall	0.988	0.979	0.984	0.993
F1 Score	0.990	0.981	0.980	0.993
Specificity	0.992	0.982	0.976	0.992
Roc Auc	0.999	0.997	0.994	0.998
Pr Auc	0.999	0.997	0.992	0.997

Table 4.4: Performance metrics across modalities for seed 99.

Metric	Image	Spec	Time Series	Multimodal
Accuracy	0.991	0.975	0.981	0.993
Precision	0.993	0.976	0.977	0.992
Recall	0.988	0.974	0.986	0.993
F1 Score	0.991	0.975	0.981	0.993
Specificity	0.993	0.976	0.977	0.992
Roc Auc	0.999	0.993	0.995	0.998
Pr Auc	0.998	0.992	0.993	0.997

Table 4.5: Performance metrics across modalities for seed 122.

Metric	Image	Spec	Time Series	Multimodal
Accuracy	0.992	0.978	0.983	0.993
Precision	0.993	0.974	0.980	0.991
Recall	0.991	0.982	0.986	0.994
F1 Score	0.992	0.978	0.983	0.993
Specificity	0.993	0.974	0.980	0.991
Roc Auc	0.999	0.996	0.994	0.997
Pr Auc	0.999	0.996	0.992	0.996

4.5 Key Results Summary

A summary of the primary results of Experiment 1 is presented in this section. Four aspects are covered in the analysis: performance comparison across different modalities, performance versus cost across different modalities, external validation results and fusion weights.

Figure 4.4: F1 score comparison across different modalities in Experiment 1.

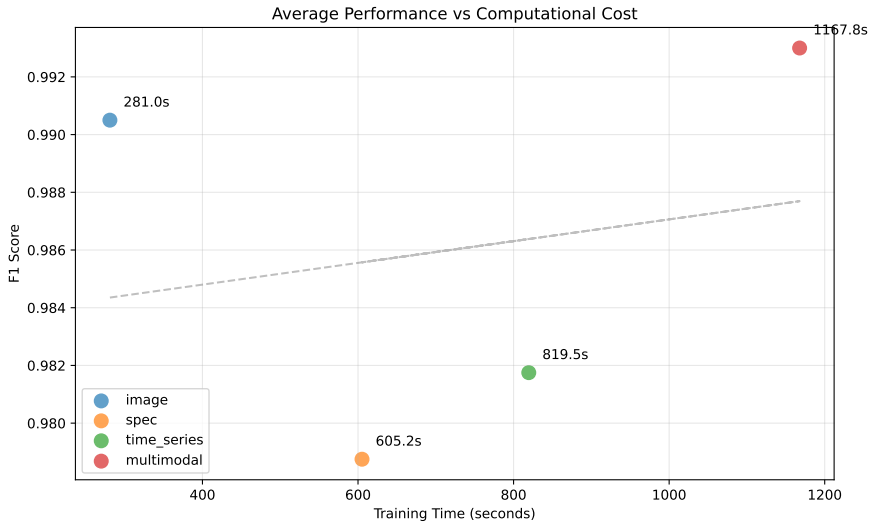

Figure 4.5: Average performance vs cost comparison across different modalities in Experiment 1.

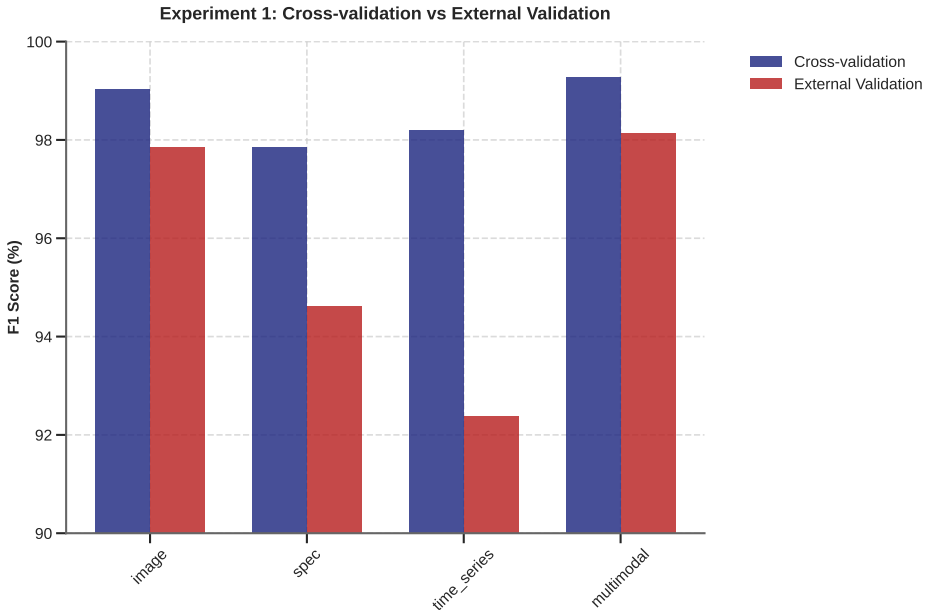

Figure 4.6: Comparison between cross-validation and external validation F1 scores across all modalities in Experiment 1.

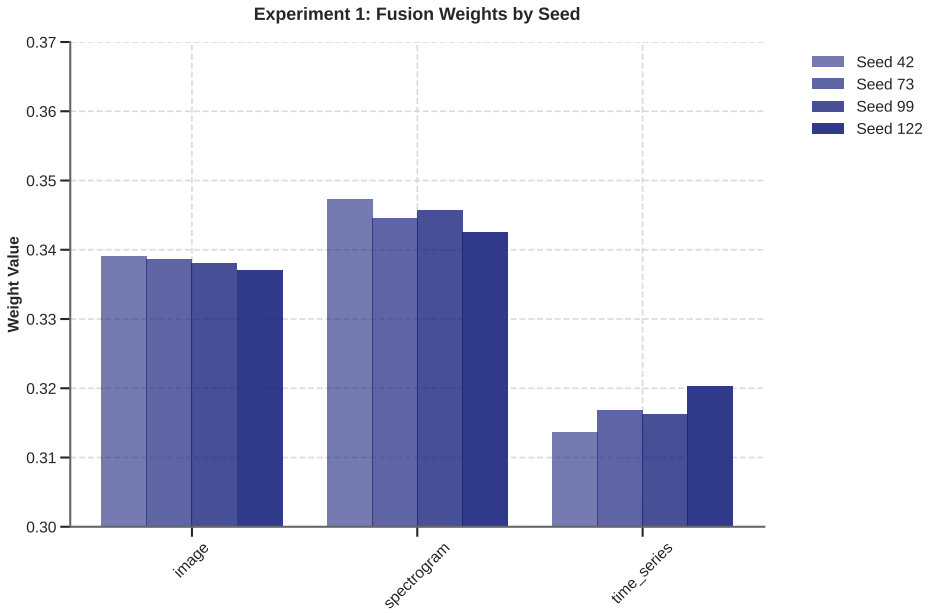

Figure 4.7: Distribution of fusion weights across different seeds in Experiment 1.

Chapter 5

Experiment 2 - AF vs Non AF Classification

5.1 Introduction

This chapter presents the key results of AF classification models across different modalities and using a different class distribution. Four different approaches are evaluated: image-based analysis, spectrogram analysis, time series analysis, and a multimodal fusion approach. For this experiment, each model was tested with only a single seed due to time and computational resources constraints. The individual results can be seen in the appendices.

5.2 Dataset Overview

This experiment utilized a 5-fold Stratified Group K-Fold cross-validation strategy for a single random seed. The dataset statistics are summarized in Table 5.1.

Table 5.1: Dataset Statistics for experiment 2.

Seed	Split	Samples/Fold	Non AF (%)	AF (%)	Class Ratio
42	train	64242 ± 93	91.55 ± 0.05	8.45 ± 0.05	0.092 ± 0.001
	val	16083 ± 93	91.70 ± 0.19	8.30 ± 0.19	0.091 ± 0.002
	test	20081 ± 1	91.59 ± 0.00	8.41 ± 0.00	0.092 ± 0.000

Values are presented as mean $\pm$ standard deviation across folds.

Class ratio is calculated as AF / Non AF.

5.3 Fusion Trainable Weight Analysis

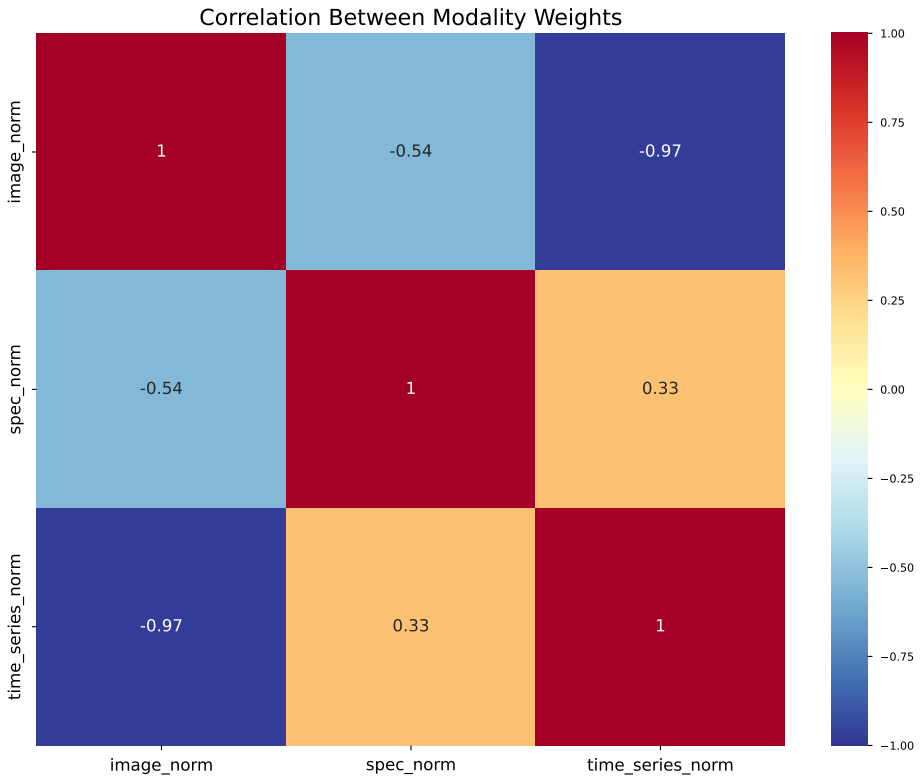

Figure 5.1: Correlation heatmap between all the modalities.

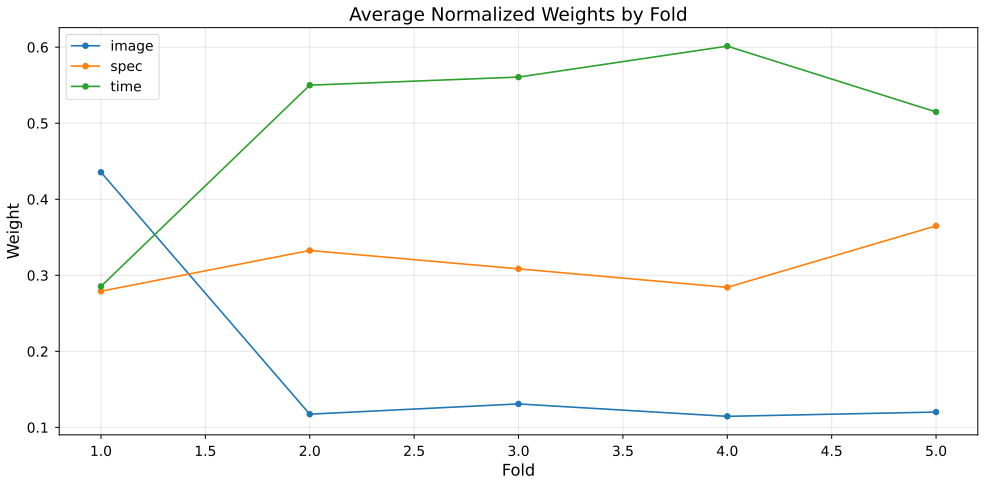

Figure 5.2: Averaged normalized weights by fold.

5.4 Cross-modality Comparison

This section shares the same structure of experiment 1, but with only a single seed.

Table 5.2: Performance metrics across modalities for seed 42 (experiment 2).

Metric	Image	Spec	Time Series	Multimodal
Accuracy	0.976	0.973	0.976	0.978
Precision	0.842	0.824	0.821	0.849
Recall	0.905	0.884	0.939	0.928
F1 Score	0.872	0.853	0.875	0.886
Specificity	0.983	0.981	0.980	0.983
Roc Auc	0.989	0.988	0.992	0.992
Pr Auc	0.924	0.911	0.934	0.933

5.5 Key Results Summary

A summary of the main results of Experiment 2 is displayed in this section. Four aspects are covered in the analysis: performance comparison across different modalities, performance versus cost across different modalities, external validation results and fusion weights.

Figure 5.3: F1 score comparison across different modalities in Experiment 2.

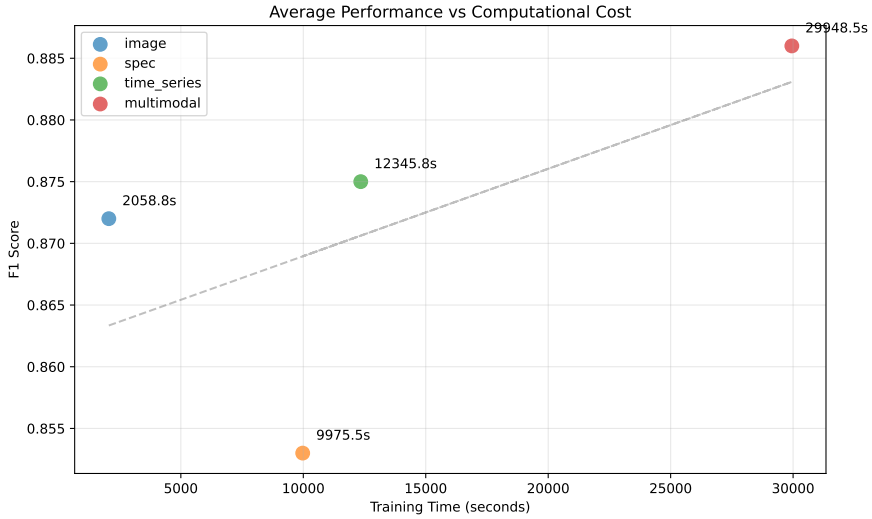

Figure 5.4: Average performance vs cost comparison across different modalities in Experiment 2.

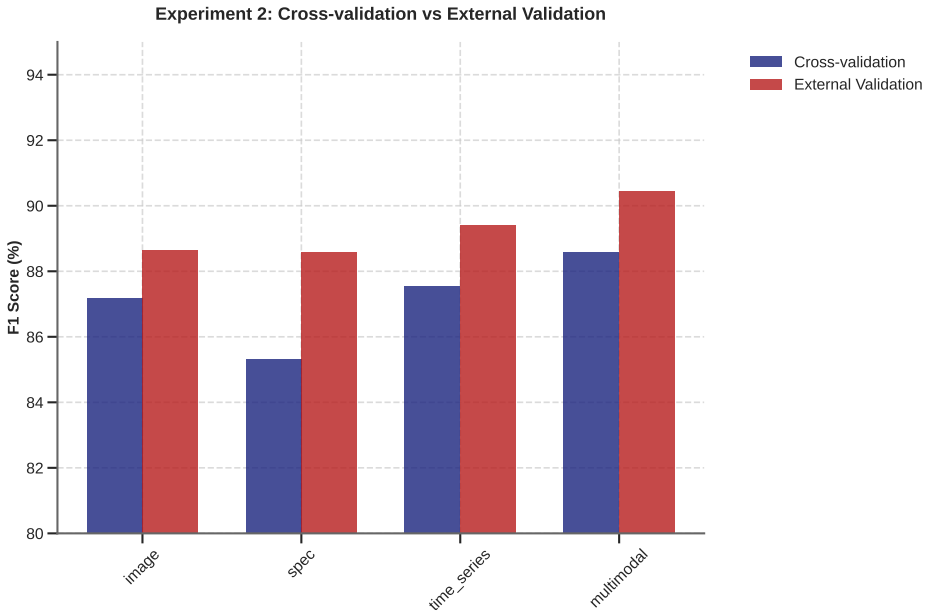

Figure 5.5: Comparison between cross-validation and external validation F1 scores across all modalities in Experiment 2.

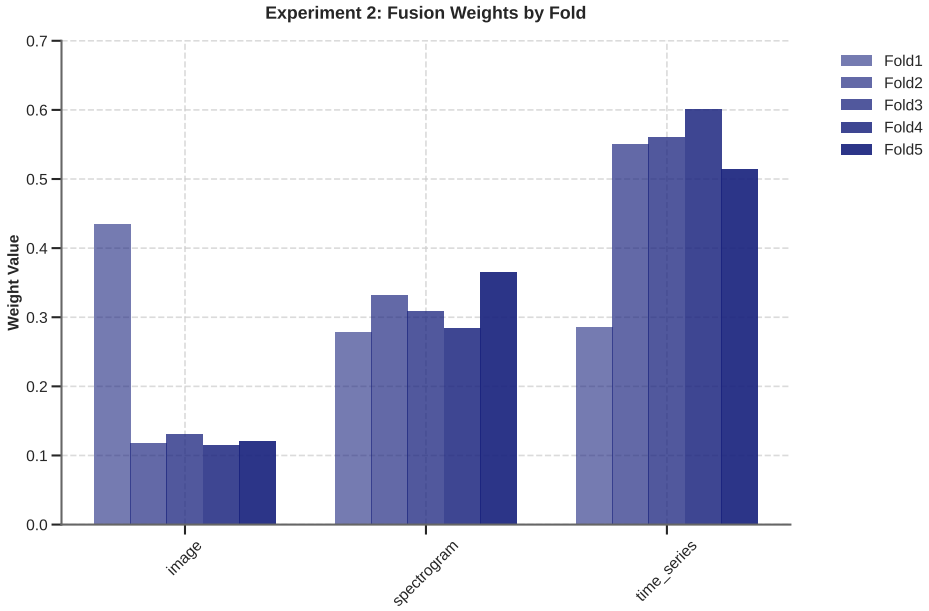

Figure 5.6: Distribution of fusion weights across different folds in Experiment 2.

Chapter 6

Discussion

6.1 Comparison with Related Work

The primary objective of this research was to conduct a systematic evaluation of how different modalities contribute to AF detection, rather than pursuing state-of-the-art performance metrics. The study investigated the strengths and complementary aspects of image-based, spectrogram, and time series approaches for detecting AF. It particularly highlighted their potential for integration within a multimodal framework. While the emphasis was on understanding modality-specific contributions rather than benchmarking against top-performing models, it is important to contextualize this work in relation to existing studies. By focusing on comparative evaluations and integration strategies, this work lays a foundation for further advancements in AF detection that capitalize on the strengths of multimodal learning.

An experimental methodology involving two distinct scenarios was used. Strong performance was demonstrated under controlled conditions by the first experiment, while the robustness of the approach was evaluated under more challenging, real-world conditions by the second experiment. In Experiment 2, nearly all available data was utilized without selection or filtering, resulting in a heavily imbalanced dataset where AF cases were represented by only 8.45% of the training data (with a class ratio of 0.092). This imbalance, which mirrors the natural distribution of AF in clinical settings, lends considerable relevance to the findings for practical applications. Despite these challenging conditions, competitive performance was attained with a multimodal approach, with an accuracy of 97.84% and an

F1-score of 88.59%. These results are highlighted by the maintenance of a stratified group 5-fold cross-validation strategy, ensuring robust validation of the findings. The consistent performance across folds, as evidenced by the low standard deviations in the dataset statistics, suggests that the approach is stable and reliable. The summary of the comparisons can be seen in Tables 6.1 and 6.2.

It is worth saying that while some studies have reported higher performance metrics, it should be emphasized that these results are often derived from more carefully curated datasets or different experimental conditions. A minimally processed, imbalanced dataset is focused on by the present study, which provides a more realistic assessment of how these methods might perform in clinical practice. Furthermore, substantial room for improvement is indicated by the results, particularly in handling class imbalance and optimizing the fusion of different modalities. This suggests promising potential for future research while maintaining an emphasis on practical applicability rather than merely optimizing for benchmark performance.

Table 6.1: Comparison of AF Detection Methods - Experiment 1

Method	**Classification Task**	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**F1-Score (%)**
ResNet10 [FA21]	Multi-class (AF, AFL, SVT, ST, SB, Normal)	98.37	–	–	–
	Binary (All arrhythmias vs Normal)	98.55	99.40	94.30	–
Hybrid DSVM [GC22]	Multi-class (Noisy, Other, AF and Normal)	99.27	–	–	95.00
DDNN [CCG+20]	Binary (AF vs Normal)	**99.35**	99.44	99.19	99.06
CNN [HW20]	Binary (AF vs Normal)	99.23	**99.71**	98.66	–
CNN-RNN [APP19]	Binary (AF vs Normal)	–	98.98	96.95	–
CNN-LSTM [MZC+20]	Binary (AF vs Normal)	97.21	97.34	97.08	–
ResNet50 [KLK+24]	Binary (AF vs Normal)	70.50	79.30	–	71.90
Deep CNN [RSAS21]	Binary (AF vs Normal)	95.50	94.50	96.00	–
BiLSTM [RS22]	Binary (AF vs Normal)	98.85	–	–	–
Our Results (Experiment 1)					
Image-based	Binary (AF vs Normal)	99.03 ± 0.12	98.78 ± 0.22	$\mathbf{99.29 \pm 0.03}$	99.03 ± 0.12
Spectrogram	Binary (AF vs Normal)	97.85 ± 0.24	97.91 ± 0.30	97.81 ± 0.32	97.86 ± 0.24
Time Series	Binary (AF vs Normal)	98.19 ± 0.12	98.51 ± 0.07	97.89 ± 0.20	98.20 ± 0.12
Multimodal	Binary (AF vs Normal)	99.28 ± 0.02	99.33 ± 0.05	99.23 ± 0.08	$\mathbf{99.28 \pm 0.02}$

AFL = Atrial Flutter, SVT = Supraventricular Tachycardia, ST = Sinus Tachycardia, SB = Sinus Bradycardia.

Table 6.2: Comparison of AF Detection Methods - Experiment 2

Method	**Classification Task**	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**F1-Score (%)**
MCNN [YZC17]	Binary (AF vs non-AF)	98.18	98.22	98.11	–
HAN-ECG [MAA20]	Binary (AF vs non-AF)	98.81	99.08	98.54	–
2D-CNN [KJ22]	Binary (AF vs non-AF)	95.00	94.00	–	94.00
Explainable AI [JCL+21]	Binary (AF vs non-AF)	99.40	98.20	99.50	–
ECG DETR [HCZ22]	Binary (AF vs non-AF)	99.23	99.23	99.23	99.23
Dual Channel [FCL+21]	Binary (AF vs non-AF)	–	–	–	83.00
GH-MS-CNN [ZMS+21]	Binary (AF vs non-AF)	**99.84**	**99.54**	**99.88**	**99.62**
DeepAware [KPS+22]	Binary (AF vs non-AF)	98.06	97.94	98.39	–
Our Results (Experiment 2)					
Image-based	Binary (AF vs non-AF)	97.59	90.53	84.17	87.17
Spectrogram	Binary (AF vs non-AF)	97.25	88.42	82.42	85.31
Time Series	Binary (AF vs non-AF)	97.58	93.86	82.07	87.54
Multimodal	Binary (AF vs non-AF)	97.84	92.78	84.93	88.59

6.2 Performance Analysis and Clinical Relevance

The present study demonstrates a notable distinction between idealized and realistic clinical scenarios in automated AF detection. In Experiment 1, a controlled environment was maintained with a balanced distribution between AF and normal rhythms. In this environment, all modalities achieved values consistently above 0.97 on the F1 score metric. The multimodal approach achieved particularly high values (F1 score = 0.9928 ± 0.0002), suggesting that different input representations provide complementary information for AF detection. However, it is pertinent to note that these results must be interpreted within the context of the artificial class balance that was created through the implementation of the sampling approach mentioned before.

Experiment 2 presents a more nuanced and clinically relevant picture, with AF comprising only 8.45% of cases. Despite implementing class weights to address this imbalance, we observed a decrease in performance across all modalities, with F1 scores ranging from 0.85 to 0.88. This performance differential indicates the difficulties in maintaining high precision and recall in actual clinical settings. The class weighting strategy, while crucial for model training, was unable to fully bridge the gap between idealized and practical scenarios.

6.3 Modality Fusion Dynamics

The behavior of the fusion mechanism displays patterns of interest across a range of experimental conditions. In Experiment 1, the fusion weights exhibited consistent stability across different seeds (image: 0.3382 ± 0.0025, spectrogram: 0.3450 ± 0.0033, time series: 0.3167 ± 0.0045), suggesting that each modality contributes almost equally to the classification of normal versus AF patterns. This balanced contribution indicates that the features extracted from different representations complement each other consistently in this simpler binary classification task.

However, Experiment 2 revealed different fusion dynamics, with high variance in modality weights (image: 0.1837 ± 0.1409, time series: 0.5025 ± 0.1252). This variability, particularly pronounced in the image and time series weights, suggests that when distinguishing AF from various other cardiac problems, the model adaptively adjusts its feature utilization strat-

egy. The consistent predominance of time series features in the model's feature utilization strategy suggests their particular importance in differentiating AF from other cardiovascular diseases. This may be attributed to their direct representation of temporal patterns that are characteristic of AF.

6.4 Generalization and External Validation

The external validation results offer insights into the generalization capabilities of the models. In Experiment 1, the multimodal approach demonstrated strong performance on the external dataset (F1 score = 0.9813 ± 0.0036), indicating superior consistency compared to single-modality approaches. This suggests that integrating diverse data representations improves the model's ability to capture universal characteristics of AF, enhancing its resilience to dataset-specific variations.

In Experiment 2, the discrepancy between the cross-validation and external validation results became more evident, reflecting the increased complexity introduced by diverse non-AF rhythms. The multimodal approach attained an F1 score of 0.9045 on the external dataset, compared to 0.8859 in cross-validation, suggesting that while performance diminishes in more challenging scenarios, the model maintains adequate generalization capabilities. This somewhat counterintuitive improvement may suggest that the external dataset contains cases that are better characterized across modalities and may be less complex than the cases present in InCor-DB dataset, allowing the fusion mechanism to more effectively exploit complementary features. However, it also raises important questions about dataset characteristics and their impact on model performance that warrant further investigation. The consistency of this improvement across multiple metrics (precision: 0.8576 vs 0.8493; sensitivity: 0.9575 vs 0.9278) further supports the robustness of the approach in handling diverse data sources.

6.5 Methodological Insights

This work provides fundamental understanding of how multimodal fusion mechanisms adapt to different classification challenges in ECG analysis in the context of AF classification. The correlation patterns between modalities and their associated fusion weights provide

a window into the models' decision-making process, highlighting how they exploit different aspects of the ECG signal depending on the task.

In Experiment 1, where the task is to classify normal and AF patterns, we observe a particularly interesting relationship between modalities. The weak positive correlation (0.21) between image and spectrogram representations suggests that these modalities sometimes work together to identify AF features, although largely independently. This complementary relationship makes intuitive sense: images capture the overall morphological patterns of the ECG, while spectrograms reveal the frequency characteristics typical of the irregular rhythms of AF. However, the strong negative correlations of the time series modality with both image (-0.71) and spectrogram (-0.84) representations suggest that the model might choose between two different strategies: either focusing on visual patterns or using direct temporal analysis.

The transition to the more complex scenario of Experiment 2, where AF must be distinguished from several other CVDs, reveals a fundamental shift in the model's approach. The extremely strong negative correlation (-0.97) between image and time series modalities suggests that the model has developed a more specialized strategy, strongly favoring one modality over the other depending on the specific characteristics of each case. This adaptation is reflected in the noticeable variation of the fusion weights, where the time series modality receives a significantly higher average weight (0.5025), but with substantial variation ($\pm$0.1252) across different data splits.

This shift in strategy is further evidenced by the change in the relationship between the spectrogram and time series modalities, from a strong negative correlation in Experiment 1 to a positive correlation (0.33) in Experiment 2. This suggests that when faced with more complex classification tasks, the model often combines frequency analysis with temporal analysis, recognizing that some non-AF conditions require both temporal and frequency domain information for accurate classification. This mirrors clinical reality, where physicians often need to use different diagnostic strategies when evaluating complex cases versus routine screening. Adapting the model to use either predominantly visual or temporal features, rather than both simultaneously, suggests that certain diseases may be more reliably identified by specific modalities.

The variation in fusion weights across folds in Experiment 2 provides additional insight into the model's adaptive behavior. For example, in Fold 1, the image modality receives

an unusually high weight (0.4355) compared to other folds where it ranges from 0.1145 to 0.1309. This substantial variation suggests that the model identifies and adapts to specific characteristics in different subsets of data, possibly responding to variations in signal quality or the presence of specific CVD patterns.

These results have important implications for the design of multimodal ECG analysis systems. The clear difference in fusion strategies between experiments suggests that the optimal model architecture may depend significantly on the specific classification task. While a balanced fusion approach works well for simple binary classification, more complex scenarios benefit from flexible weighting schemes that can adapt to the specific characteristics of each case. This adaptability comes at the cost of increased model complexity and potential instability in weight assignments, but the performance benefits, especially in realistic clinical scenarios, may potentially justify these trade-offs.

Furthermore, the correlation patterns between modalities suggest that future architectural improvements might benefit from explicitly modeling these relationships. For example, the strong negative correlations in certain scenarios may indicate opportunities for designing attention mechanisms that could more effectively switch between different modalities based on input characteristics.

6.6 Signal Processing and Data Quality Considerations

The efficacy of the multimodal approach was supported by the development of a robust signal extraction and preprocessing pipeline for this study. It was found that bilateral filtering with optimized parameters was an effective tool to preserve critical signal information while reducing noise artifacts, and this preprocessing step was essential for maintaining signal fidelity across all modalities.

The chosen approach of combining contour detection and coordinate calculation was effective in handling variations in ECG trace quality. Validation procedures implemented for ensuring signal continuity turned out to be especially valuable in Experiment 2, where significant challenges were posed by the diversity of input data quality. It was proven that the statistical outlier detection using DBSCAN clustering was robust against common artifacts that could have otherwise compromised classification accuracy.

The relationship between signal quality and model performance became evident in the external validation results. Consistent performance metrics observed in the external validation indicate that signals from variable sources were effectively standardized by our preprocessing pipeline, enhancing the model's generalization capabilities. It was also important that this standardization ensured consistency across all three representations of the ECG signal for the multimodal approach.

Chapter 7

Conclusion

7.1 Clinical Impact and Performance Analysis

This research has advanced the field of automated AF classification through the development and validation of a comprehensive multimodal framework, supported by the proposed signal extraction and preprocessing pipeline. The investigation provided some interesting findings, contributing to both theoretical understanding and practical applications in cardiac diagnostics. The implementation of diverse signal processing techniques, including bilateral filtering and adaptive thresholding, ensured high-quality input data across all modalities, establishing a solid foundation for subsequent analysis.

The experimental results demonstrated the effectiveness of combining multiple ECG representations for AF classification. In controlled conditions (Experiment 1), the multimodal approach achieved an F1-score of 0.9928, with a standard deviation of 0.0002, which exceeded the performance of single-modality approaches while maintaining consistency across different random seeds. Of further significance is the result observed in Experiment 2, which utilized a more realistic clinical setting. In this experiment, AF cases constituted merely 8.45% of the dataset. Despite this relatively limited representation, the multimodal framework demonstrated consistent performance, with an F1-score of 0.8859. Additionally, the model showcased notable generalization capabilities, as evidenced by its enhanced performance on external validation sets, with an F1-score of 0.9045.

The analysis of fusion weight dynamics yielded valuable insights into the model's adaptability to varied classification challenges. In the more straightforward task of differentiating

AF from normal rhythms, the model exhibited relatively stable and balanced fusion weights (image: 0.3382 ± 0.0025, spectrogram: 0.3450 ± 0.0033, time series: 0.3167 ± 0.0045) suggests that complementary information is contributed by each modality. However, in the more complex task of differentiating AF from various other cardiac conditions, more specialized strategies were developed by the model, as evidenced by the greater variation in fusion weights across folds and the predominant reliance on temporal features.

The potential for practical clinical application of the framework is demonstrated by its strong performance on external validation data in both experimental scenarios. This generalization capability, combined with the model's ability to adapt its feature utilization strategy based on input characteristics, suggests that the approach could be valuable in diverse healthcare settings where ECG data may be available in different formats.

7.2 Outlook

Several limitations and opportunities for future research have emerged from this work. Firstly, while the multimodal approach demonstrated superior performance, the computational demands of processing multiple representations could be further optimized. Secondly, the potential benefits from developing more sophisticated fusion mechanisms that can adapt to specific input characteristics are suggested by the variation in fusion weights across different scenarios. Finally, further investigation into the relationship between dataset characteristics and model performance is called for by the improved performance on external validation in Experiment 2.

The signal extraction pipeline, while effective, could be further refined to handle an even broader range of ECG recording qualities and formats. Future work might explore advanced denoising techniques, automated quality evaluation methods, and adaptive preprocessing parameters that adjust based on input signal characteristics. The development of more sophisticated signal validation metrics could also enhance the reliability of the extraction process, particularly for large-scale clinical applications.

Future research directions could substantially extend and enhance this work in several important ways. First, the implementation of sophisticated data augmentation techniques could help address the class imbalance challenges encountered in Experiment 2. These tech-

niques may include synthetic data generation specific to ECG signals, controlled perturbation methods that preserve clinically relevant features, and adversarial augmentation approaches that could enhance model robustness.

The framework's generalization capabilities could be further validated through extensive testing on multiple external datasets from diverse healthcare institutions. This multi-center validation would be particularly valuable for assessing the model's performance across different patient populations, varying ECG recording equipment, and distinct clinical protocols. A comprehensive validation would provide substantial evidence for the practical utility of the framework in various healthcare settings.

The preprocessing pipeline needs optimization in three key areas: implementing efficient signal processing techniques, leveraging parallel processing strategies, and reducing input dimensionality while preserving diagnostic information. Furthermore, the model currently learns non-diagnostic features along image and signal boundaries, which detracts from its ability to detect meaningful patterns. Adding boundary controls during preprocessing would help focus the model's learning on relevant features.

Advanced fusion mechanisms could be developed to better exploit the complementary nature of different modalities. These could include attention-based fusion approaches that dynamically weight different representations based on input quality, hierarchical fusion strategies that combine features at multiple levels of abstraction, and adaptive fusion mechanisms that adjust their behavior based on the specific characteristics of each case.

The framework could be extended beyond AF classification to identify other cardiac conditions. This expansion would involve adapting the architecture to handle multiple classification tasks simultaneously, potentially incorporating additional modalities specific to certain conditions, and developing hierarchical classification strategies that mirror clinical diagnostic processes.

The relationship between dataset characteristics and model performance warrants further investigation, particularly given the interesting dynamics observed in external validation. This research could involve detailed analysis of feature distributions across datasets, investigation of domain adaptation techniques, and development of methods to quantify and account for dataset bias.

Real-time processing capabilities could be explored to enable immediate diagnostic feed-

back in clinical settings. This would require optimization of the processing pipeline, investigation of streaming data handling techniques, and development of efficient methods for continuous monitoring applications. Model pruning and quantization could also be employed.

Finally, the integration of clinical metadata and patient history information could be investigated to provide more contextualized predictions. For instance, this would involve the development of methods to combine structured clinical data with ECG representations.

In conclusion, this research has demonstrated the effectiveness of a multimodal approach to AF classification and has provided valuable insights into the dynamics of modality fusion in cardiac signal analysis. Its potential for improving the accuracy and reliability of automated cardiac diagnostics in clinical settings is suggested by the framework's robust performance across different scenarios and generalization capabilities. Moreover, these future directions would not only enhance the current framework but also contribute to the broader field of automated cardiac diagnostics using ECGs, potentially improving patient care through more accurate and reliable detection of cardiac conditions.

Appendices

Appendix A

Individual Modality Performance for Experiment 1

This section presents a comprehensive analysis of the experimental results obtained across multiple modalities and random seeds. Model performance and reliability are evaluated in depth by the study, utilizing a range of complementary visualization techniques.

The analysis is primarily supported by the confusion matrix, which presents both absolute counts and percentage distributions. Both cross-validation and external validation results are included in the presentation of these matrices, with color gradients indicating prediction frequencies. Both quantitative assessment and intuitive visual interpretation of the model's classification behavior are facilitated by this dual representation.

To assess model stability, performance across multiple cross-validation folds has been analyzed. The progression of key metrics (F1 score, precision, and recall) throughout the training process is illustrated by the learning curves. The statistical significance of the results is assessed and potential optimization issues like overfitting or underfitting are identified by these curves, which include 95% confidence intervals.

A.0.1 Image Modality

Seed 42 Analysis

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.143 (0.046–0.241)	0.141
Accuracy	0.989 (0.984–0.993)	0.990
Pr Auc	0.999 (0.998–0.999)	0.996
Roc Auc	0.999 (0.998–1.000)	0.999
Precision	0.993 (0.988–0.997)	0.976
Recall	0.985 (0.973–0.996)	0.981
Tp	1663.4 (1643.2–1683.6)	405.0
Fp	12.4 (4.1–20.7)	10.0
Tn	1677.0 (1669.3–1684.7)	1356.0
Fn	26.0 (6.3–45.7)	8.0
F1 Score	0.989 (0.984–0.993)	0.978
Specificity	0.993 (0.988–0.998)	0.993

Table A.1: Image Model Performance Metrics for seed 42.

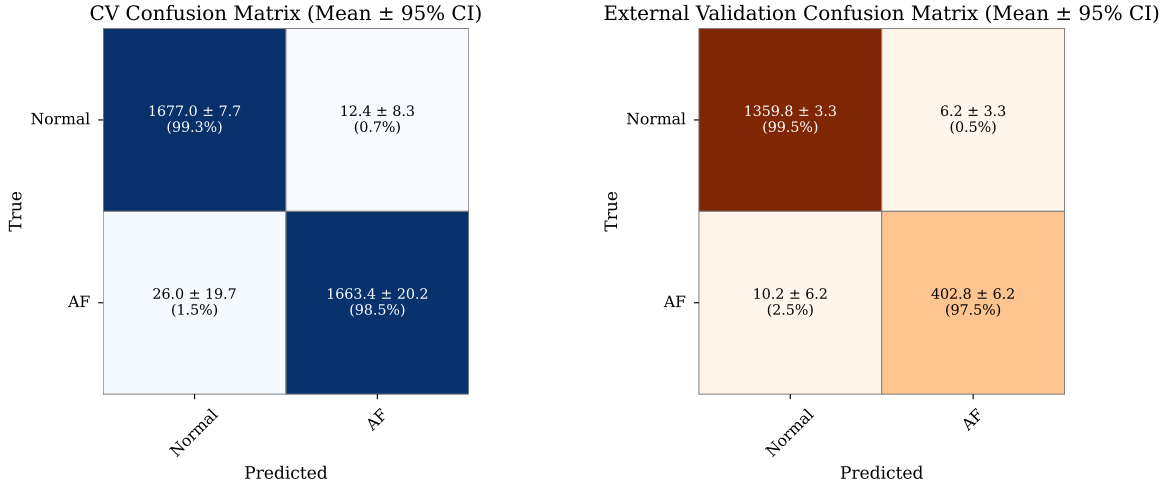

Figure A.1: Averaged confusion matrices displaying the classification performance of the image model with seed 42.

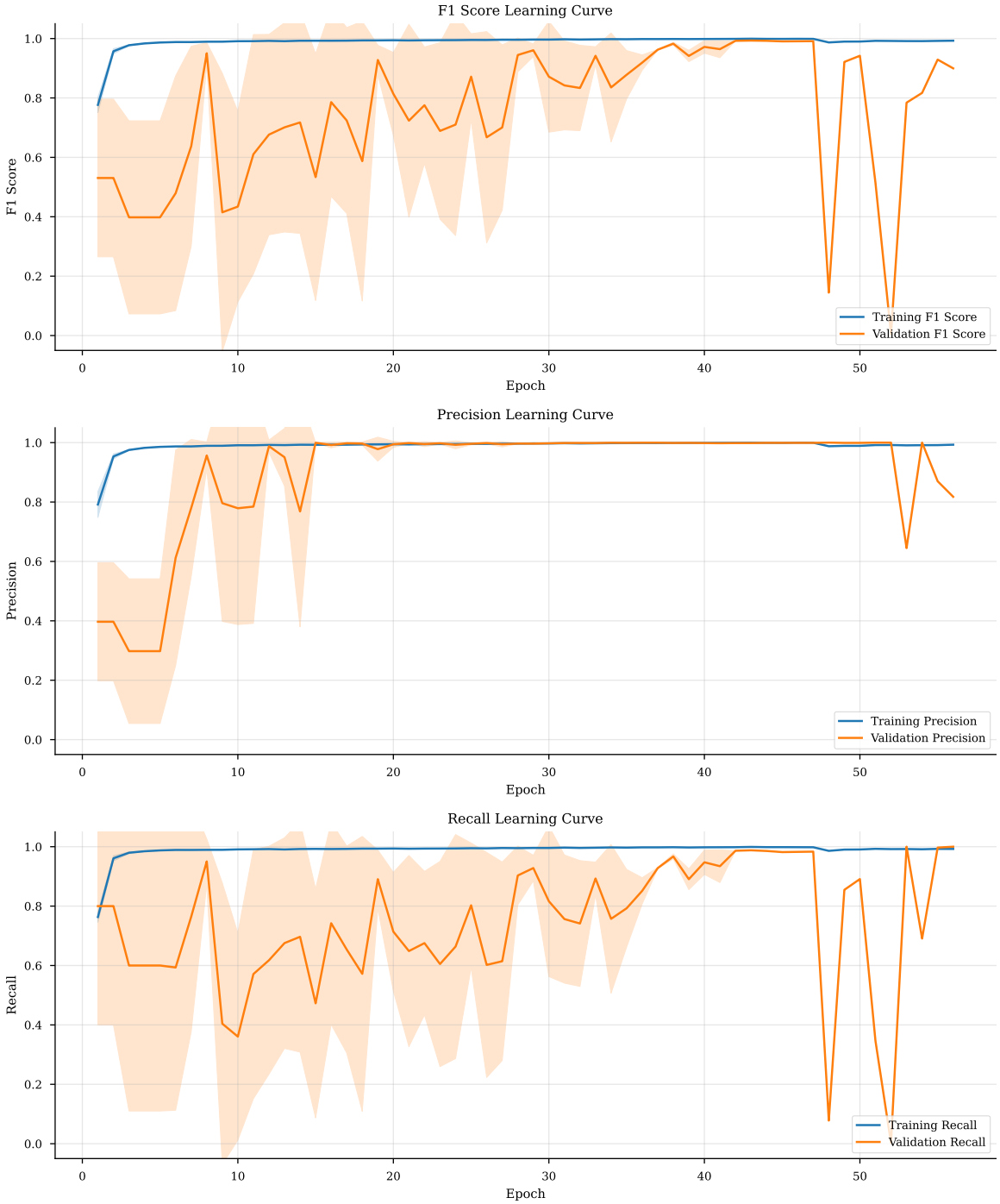

Figure A.2: Learning curves showing the evolution of model performance metrics during training for the image approach with seed 42.

Seed 73 Analysis

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.108 (0.056–0.160)	0.098
Accuracy	0.990 (0.988–0.993)	0.990
Pr Auc	0.999 (0.998–0.999)	0.998
Roc Auc	0.999 (0.998–0.999)	0.999
Precision	0.992 (0.990–0.995)	0.990
Recall	0.988 (0.982–0.994)	0.966
Tp	1668.8 (1658.9–1678.7)	399.0
Fp	12.8 (8.7–16.9)	4.0
Tn	1676.6 (1671.9–1681.3)	1362.0
Fn	20.6 (10.4–30.8)	14.0
F1 Score	0.990 (0.987–0.993)	0.978
Specificity	0.992 (0.990–0.995)	0.997

Table A.2: Image Model Performance Metrics for seed 73.

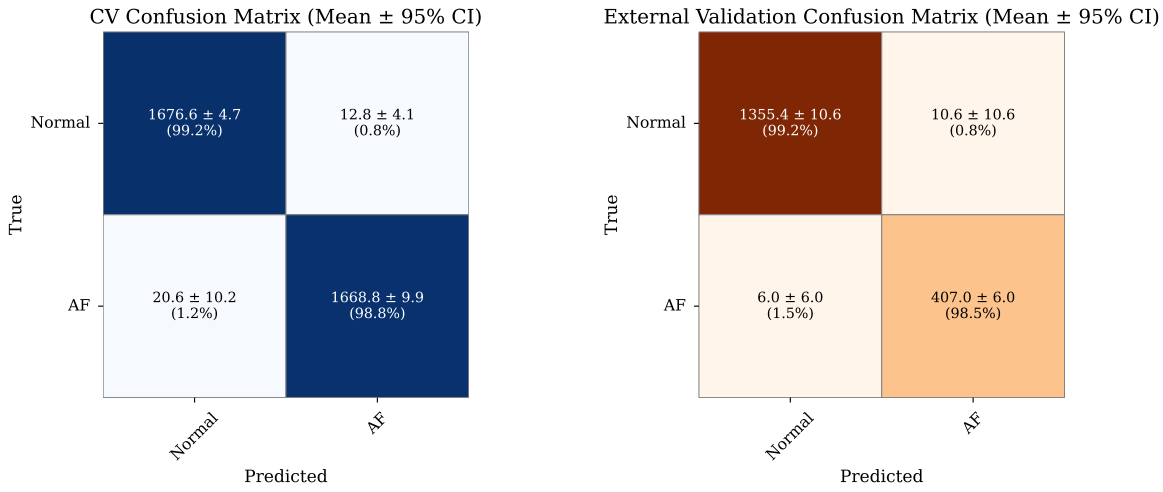

Figure A.3: Averaged confusion matrices displaying the classification performance of the image model with seed 73.

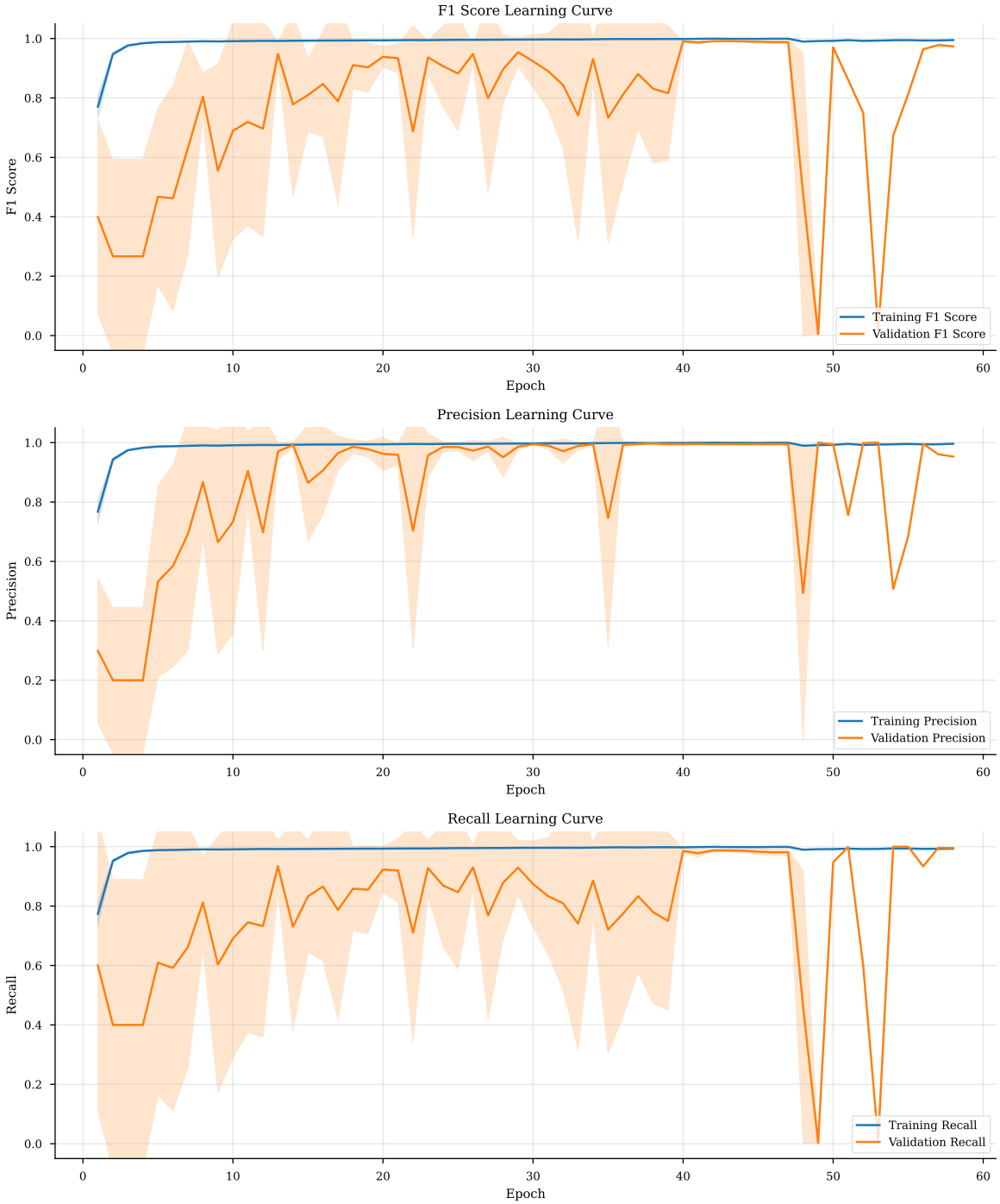

Figure A.4: Learning curves showing the evolution of model performance metrics during training for the image approach with seed 73.

Seed 99 Analysis

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.092 (0.054–0.129)	0.093
Accuracy	0.991 (0.987–0.995)	0.994
Pr Auc	0.998 (0.997–0.999)	0.999
Roc Auc	0.999 (0.998–0.999)	1.000
Precision	0.993 (0.989–0.997)	0.986
Recall	0.988 (0.981–0.995)	0.990
Tp	1669.0 (1657.0–1681.0)	409.0
Fp	11.4 (4.3–18.5)	6.0
Tn	1678.0 (1671.0–1685.0)	1360.0
Fn	20.4 (8.4–32.4)	4.0
F1 Score	0.991 (0.987–0.995)	0.988
Specificity	0.993 (0.989–0.997)	0.996

Table A.3: Image Model Performance Metrics for seed 99.

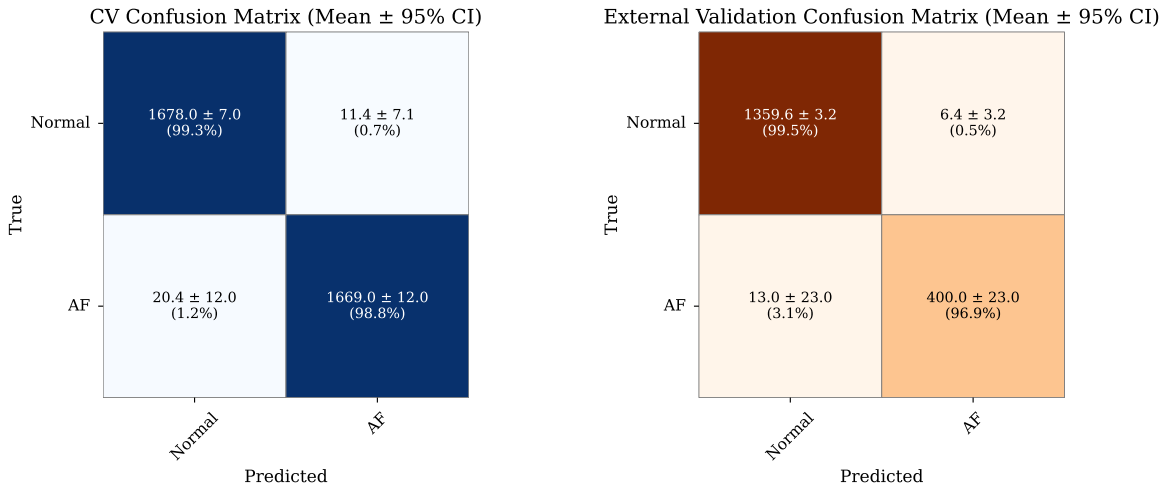

Figure A.5: Averaged confusion matrices displaying the classification performance of the image model with seed 99.

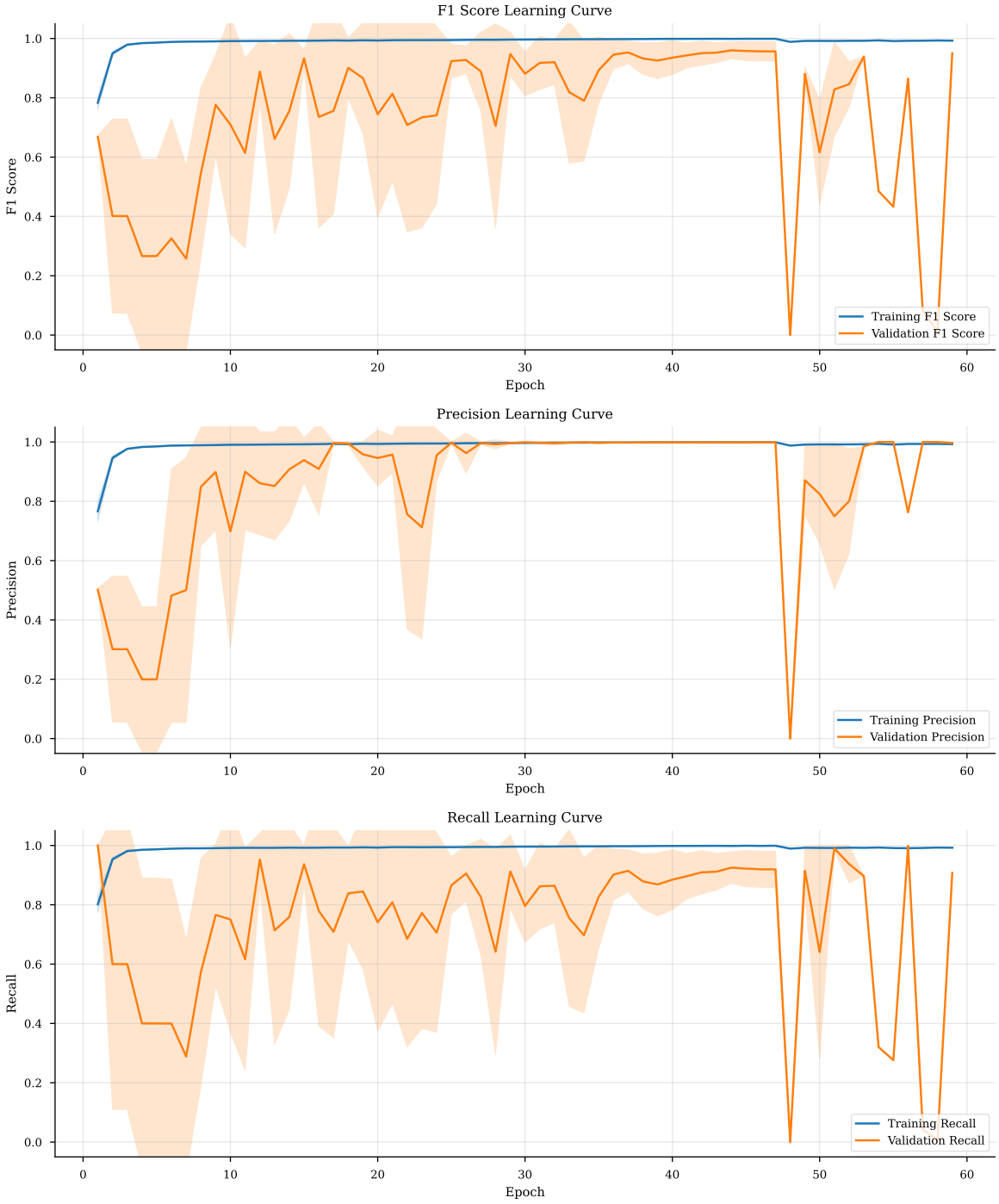

Figure A.6: Learning curves showing the evolution of model performance metrics during training for the image approach with seed 99.

Seed 122 Analysis

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.092 (0.030–0.153)	0.093
Accuracy	0.992 (0.987–0.997)	0.985
Pr Auc	0.999 (0.998–1.000)	0.998
Roc Auc	0.999 (0.998–0.999)	0.999
Precision	0.993 (0.985–1.001)	0.941
Recall	0.991 (0.986–0.996)	0.998
Tp	1674.0 (1665.0–1683.0)	412.0
Fp	11.6 (-2.5–25.7)	26.0
Tn	1677.8 (1664.3–1691.3)	1340.0
Fn	15.4 (6.6–24.2)	1.0
F1 Score	0.992 (0.987–0.997)	0.968
Specificity	0.993 (0.985–1.001)	0.981

Table A.4: Image Model Performance Metrics for seed 122.

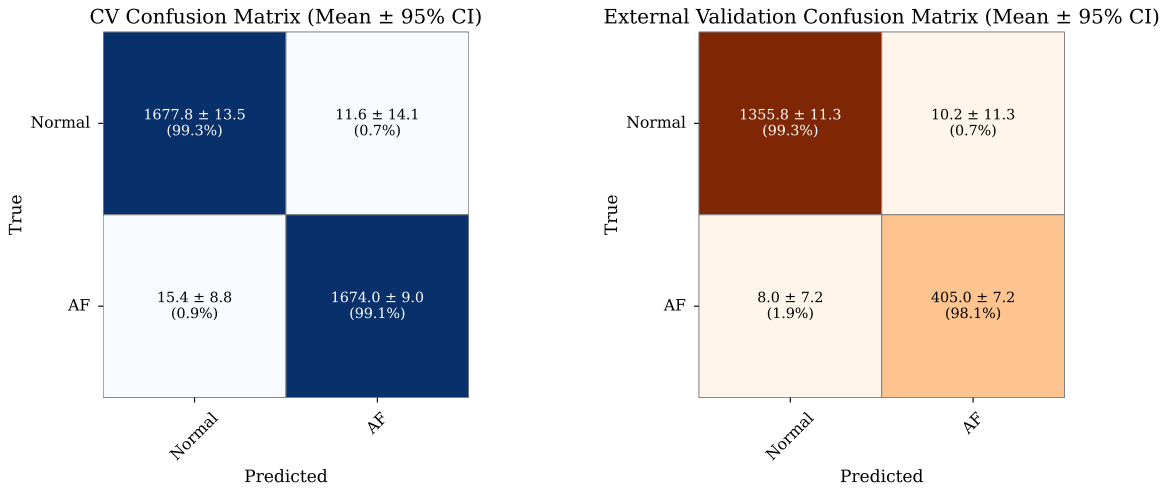

Figure A.7: Averaged confusion matrices displaying the classification performance of the image model with seed 122.

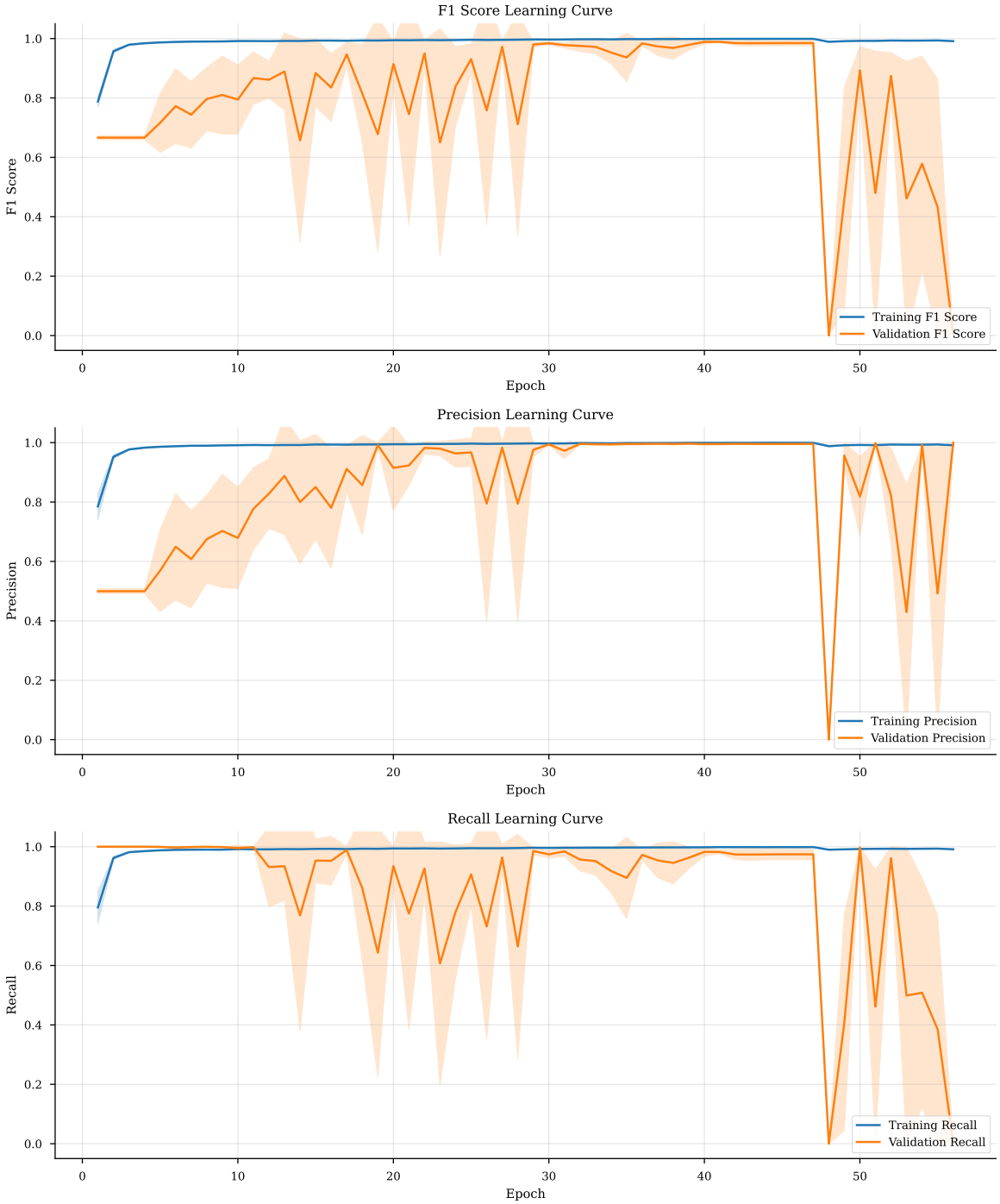

Figure A.8: Learning curves showing the evolution of model performance metrics during training for the image approach with seed 122.

A.0.2 Spectrogram Modality

Seed 42 Analysis

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.064 (0.057–0.071)	0.041
Accuracy	0.981 (0.976–0.985)	0.989
Pr Auc	0.995 (0.994–0.997)	0.991
Roc Auc	0.996 (0.995–0.997)	0.998
Precision	0.980 (0.973–0.987)	0.978
Recall	0.982 (0.977–0.986)	0.973
Tp	1658.4 (1650.9–1665.9)	402.0
Fp	33.8 (21.3–46.3)	9.0
Tn	1655.6 (1643.4–1667.8)	1357.0
Fn	31.0 (23.6–38.4)	11.0
F1 Score	0.981 (0.976–0.985)	0.976
Specificity	0.980 (0.973–0.987)	0.993

Table A.5: Spec Model Performance Metrics for seed 42.

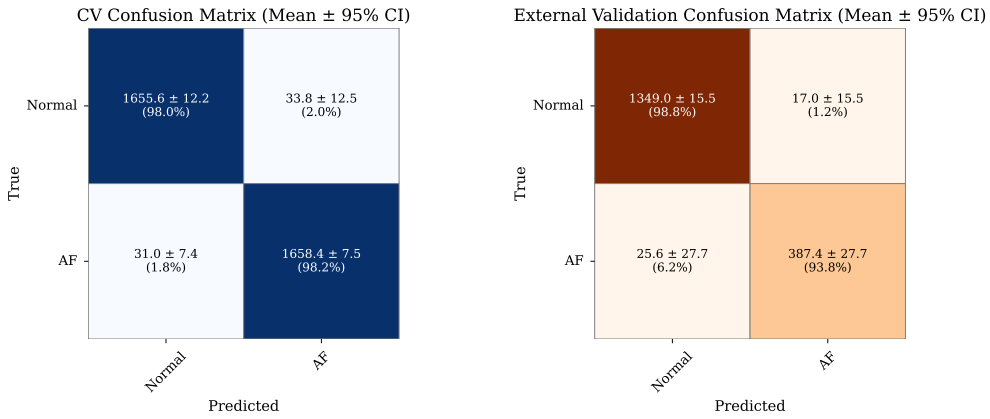

Figure A.9: Averaged confusion matrices displaying the classification performance of the spectrogram model with seed 42.

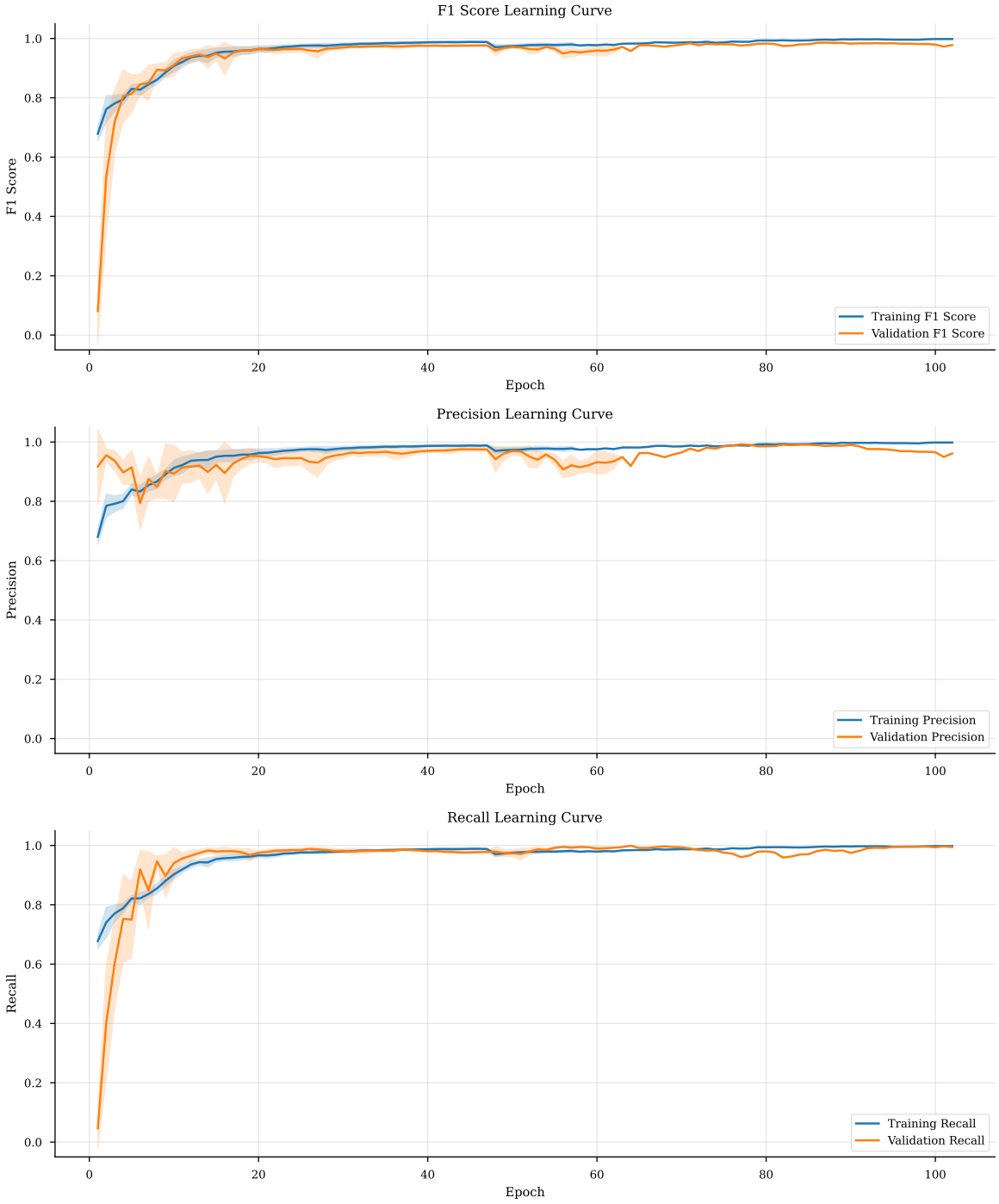

Figure A.10: Learning curves showing the evolution of model performance metrics during training for the spectrogram approach with seed 42.

Seed 73 Analysis

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.059 (0.048–0.069)	0.045
Accuracy	0.981 (0.975–0.986)	0.981
Pr Auc	0.997 (0.995–0.998)	0.995
Roc Auc	0.997 (0.996–0.998)	0.998
Precision	0.982 (0.977–0.987)	0.966
Recall	0.979 (0.973–0.985)	0.954
Tp	1653.8 (1644.4–1663.2)	394.0
Fp	30.0 (21.5–38.5)	14.0
Tn	1659.4 (1651.3–1667.5)	1352.0
Fn	35.6 (25.8–45.4)	19.0
F1 Score	0.981 (0.975–0.986)	0.960
Specificity	0.982 (0.977–0.987)	0.990

Table A.6: Spec Model Performance Metrics for seed 73.

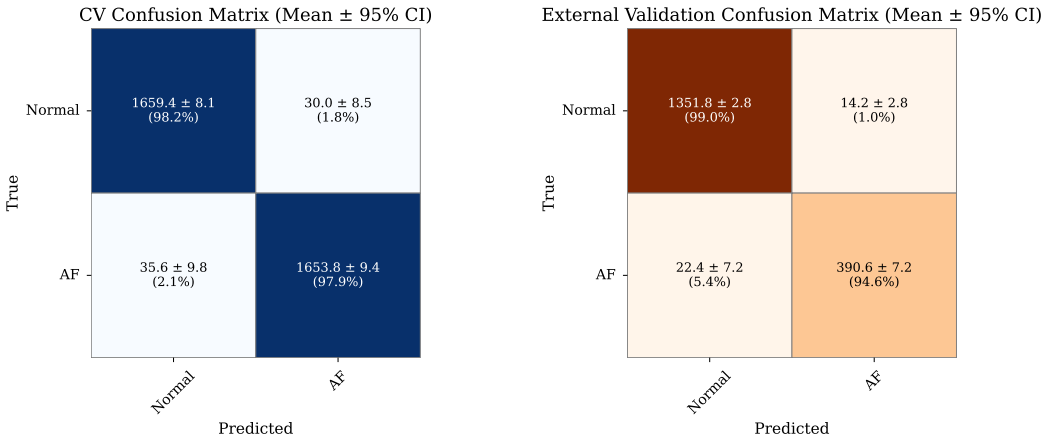

Figure A.11: Averaged confusion matrices displaying the classification performance of the spectrogram model with seed 73.

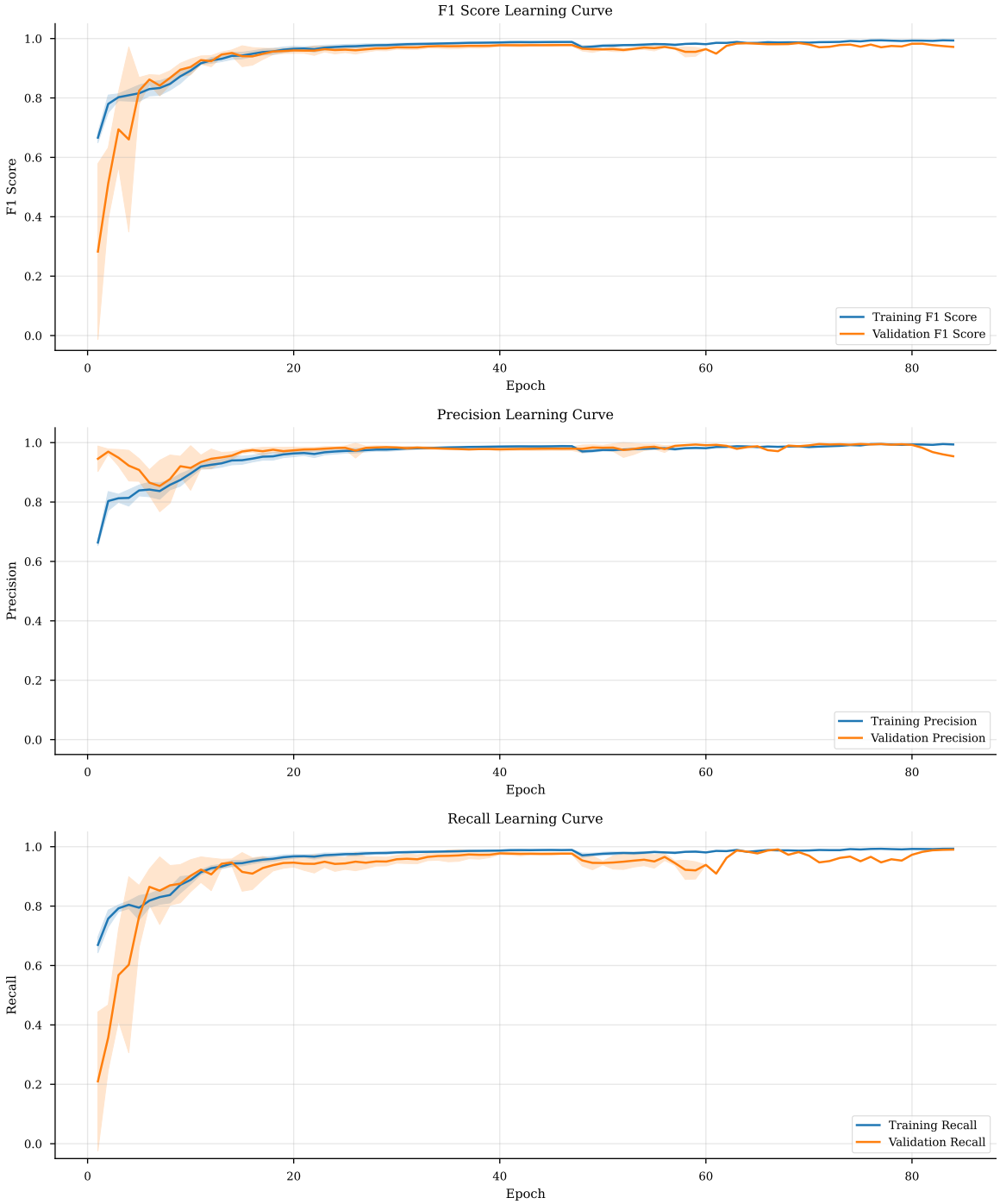

Figure A.12: Learning curves showing the evolution of model performance metrics during training for the spectrogram approach with seed 73.

Seed 99 Analysis

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.089 (0.042–0.135)	0.193
Accuracy	0.975 (0.956–0.994)	0.917
Pr Auc	0.992 (0.986–0.998)	0.917
Roc Auc	0.993 (0.988–0.999)	0.971
Precision	0.976 (0.952–1.000)	0.852
Recall	0.974 (0.961–0.987)	0.780
Tp	1645.8 (1624.3–1667.3)	322.0
Fp	41.0 (-0.8–82.8)	56.0
Tn	1648.4 (1607.1–1689.7)	1310.0
Fn	43.6 (21.6–65.6)	91.0
F1 Score	0.975 (0.957–0.993)	0.814
Specificity	0.976 (0.951–1.000)	0.959

Table A.7: Spec Model Performance Metrics for seed 99.

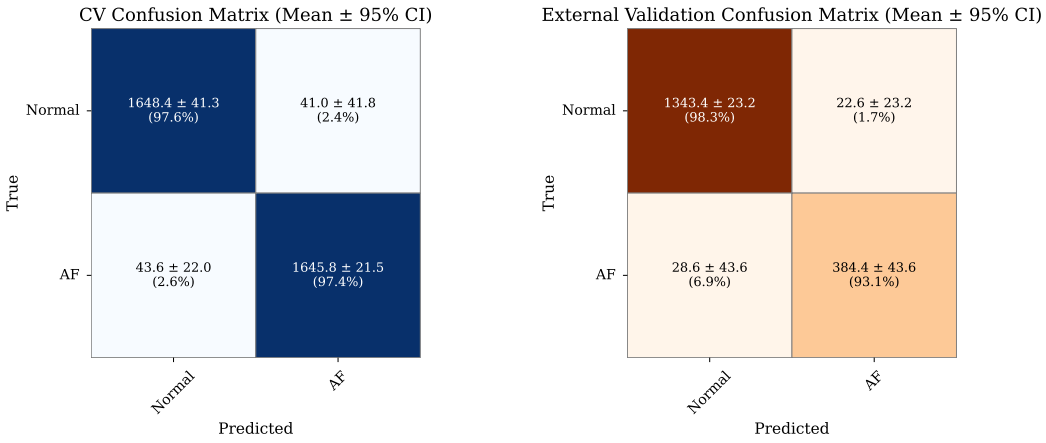

Figure A.13: Averaged confusion matrices displaying the classification performance of the spectrogram model with seed 99.

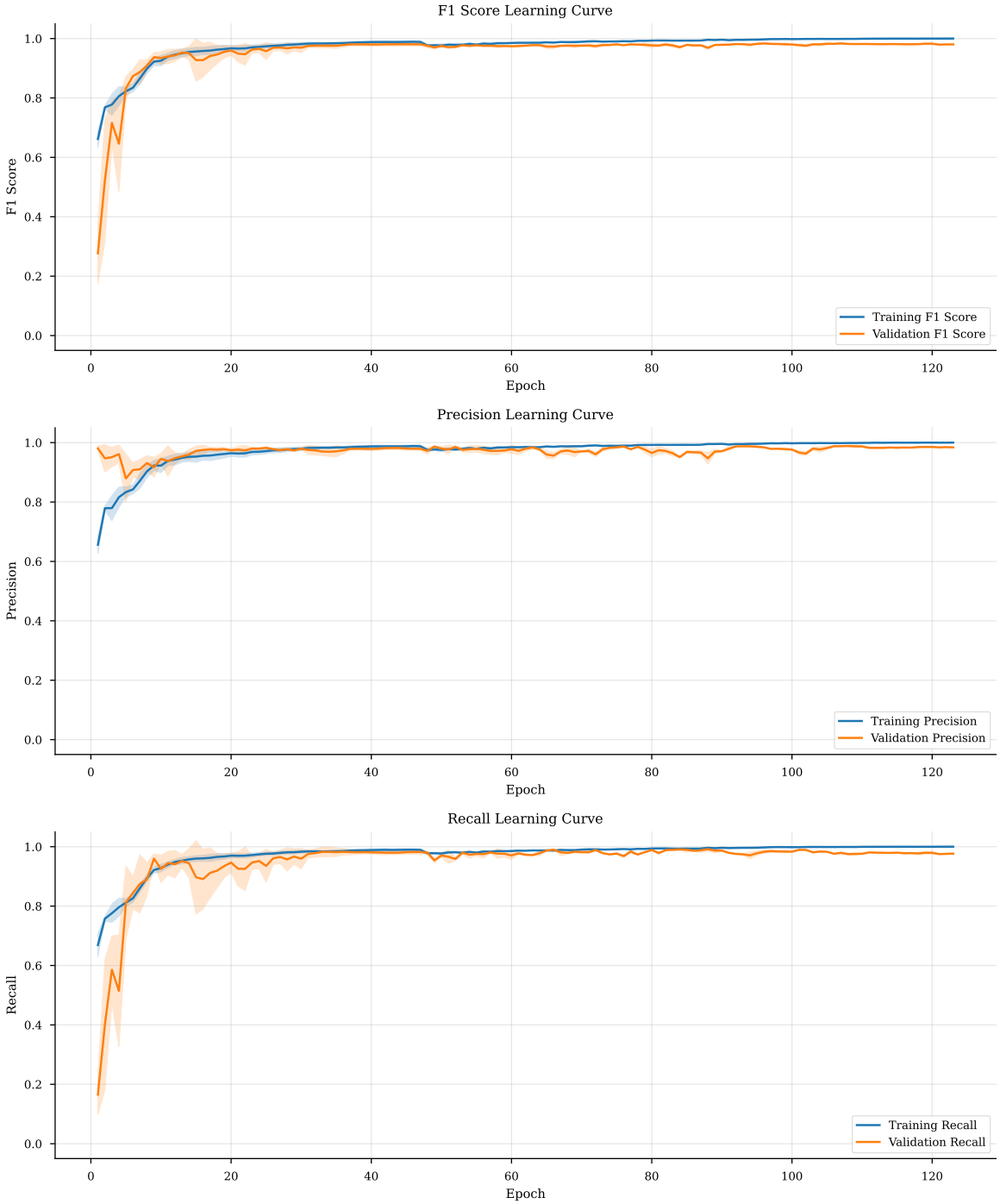

Figure A.14: Learning curves showing the evolution of model performance metrics during training for the spectrogram approach with seed 99.

Seed 122 Analysis

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.069 (0.061–0.078)	0.096
Accuracy	0.978 (0.972–0.983)	0.965
Pr Auc	0.996 (0.995–0.996)	0.980
Roc Auc	0.996 (0.995–0.996)	0.992
Precision	0.974 (0.962–0.986)	0.911
Recall	0.982 (0.976–0.987)	0.942
Tp	1658.4 (1648.9–1667.9)	389.0
Fp	44.0 (23.3–64.7)	38.0
Tn	1645.4 (1625.2–1665.6)	1328.0
Fn	31.0 (22.0–40.0)	24.0
F1 Score	0.978 (0.973–0.983)	0.926
Specificity	0.974 (0.962–0.986)	0.972

Table A.8: Spec Model Performance Metrics for seed 122.

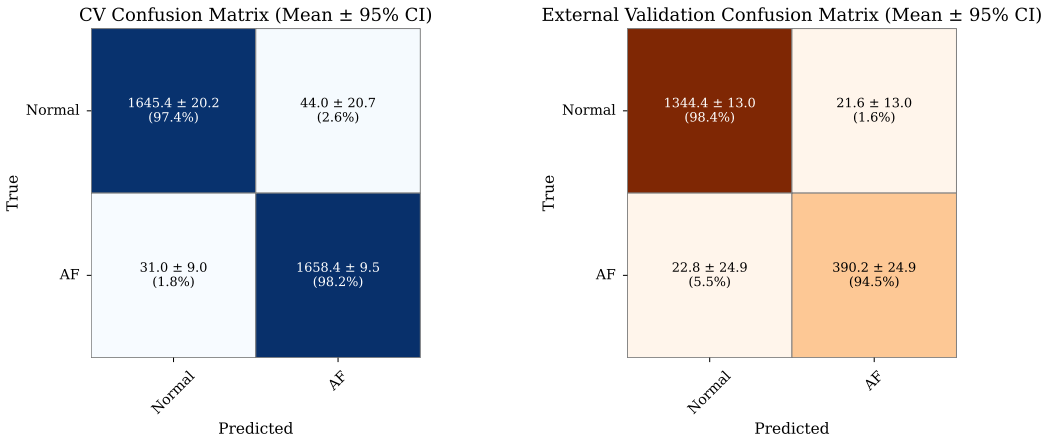

Figure A.15: Averaged confusion matrices displaying the classification performance of the spectrogram model with seed 122.

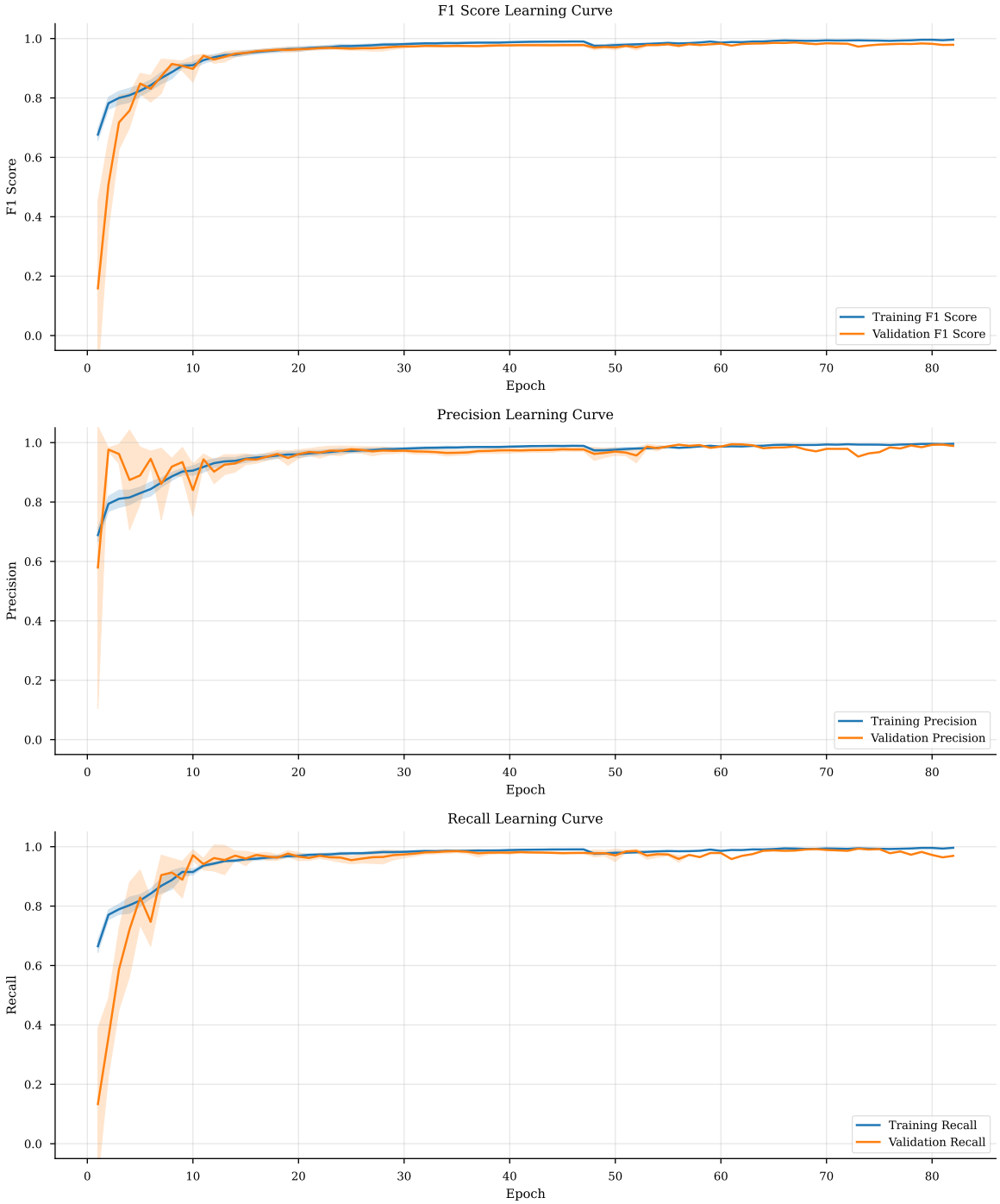

Figure A.16: Learning curves showing the evolution of model performance metrics during training for the spectrogram approach with seed 122.

A.0.3 Time Series Modality

Seed 42 Analysis

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.060 (0.049–0.072)	0.139
Accuracy	0.983 (0.979–0.987)	0.957
Pr Auc	0.994 (0.993–0.994)	0.941
Roc Auc	0.995 (0.994–0.996)	0.982
Precision	0.982 (0.977–0.987)	0.904
Recall	0.985 (0.981–0.988)	0.910
Tp	1663.6 (1657.9–1669.3)	376.0
Fp	31.0 (22.8–39.2)	40.0
Tn	1658.4 (1650.3–1666.5)	1326.0
Fn	25.8 (20.0–31.6)	37.0
F1 Score	0.983 (0.979–0.987)	0.907
Specificity	0.982 (0.977–0.986)	0.971

Table A.9: Time series Model Performance Metrics for seed 42.

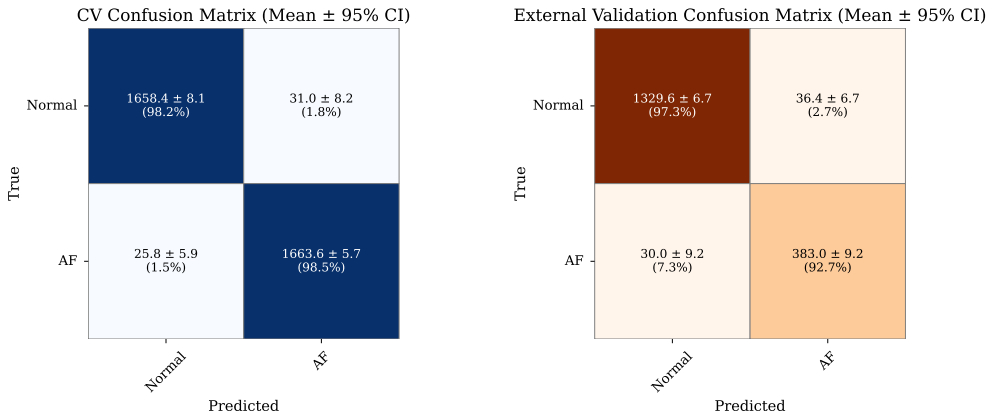

Figure A.17: Averaged confusion matrices displaying the classification performance of the time series model with seed 42.

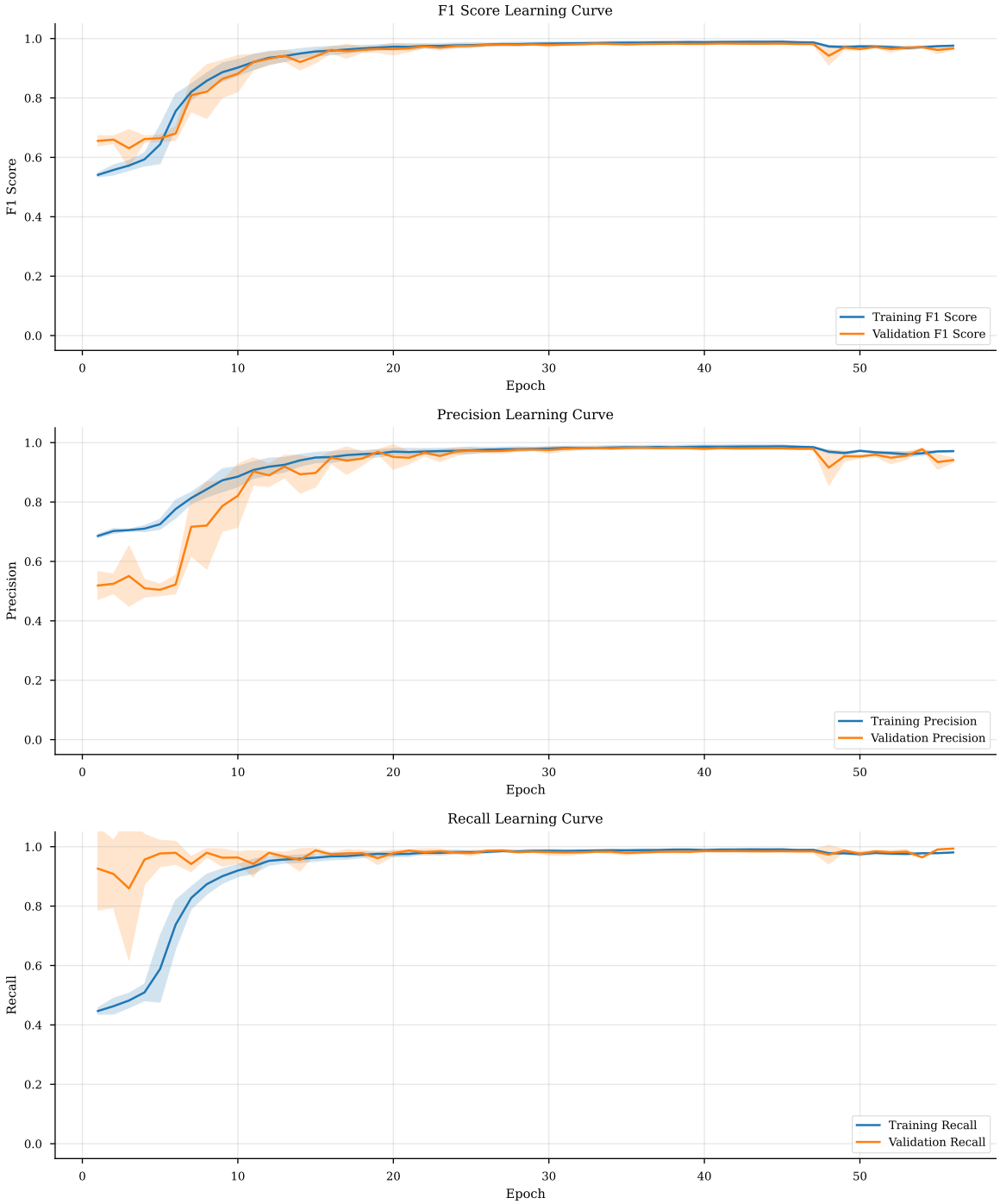

Figure A.18: Learning curves showing the evolution of model performance metrics during training for the time series approach with seed 42.

Seed 73 Analysis

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.071 (0.047–0.094)	0.119
Accuracy	0.980 (0.975–0.985)	0.966
Pr Auc	0.992 (0.988–0.995)	0.941
Roc Auc	0.994 (0.991–0.997)	0.987
Precision	0.977 (0.968–0.985)	0.923
Recall	0.984 (0.982–0.986)	0.932
Tp	1662.4 (1658.0–1666.8)	385.0
Fp	40.0 (25.7–54.3)	32.0
Tn	1649.4 (1634.6–1664.2)	1334.0
Fn	27.0 (23.3–30.7)	28.0
F1 Score	0.980 (0.975–0.985)	0.928
Specificity	0.976 (0.968–0.985)	0.977

Table A.10: Time series Model Performance Metrics for seed 73.

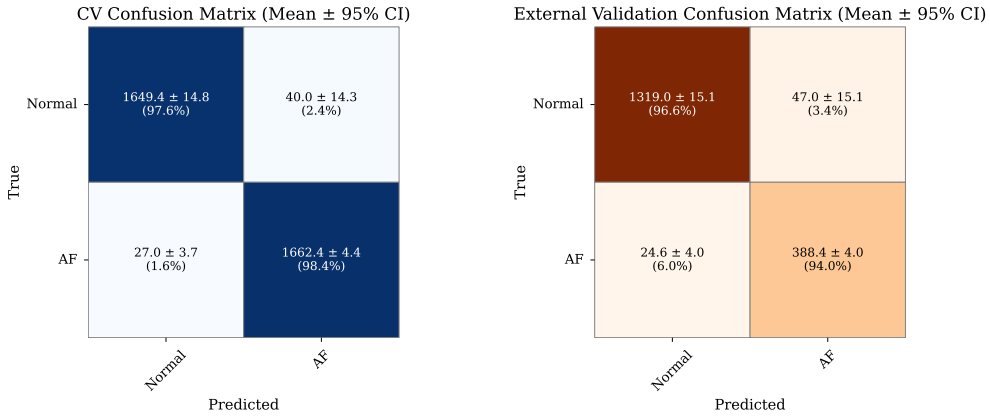

Figure A.19: Averaged confusion matrices displaying the classification performance of the time series model with seed 73.

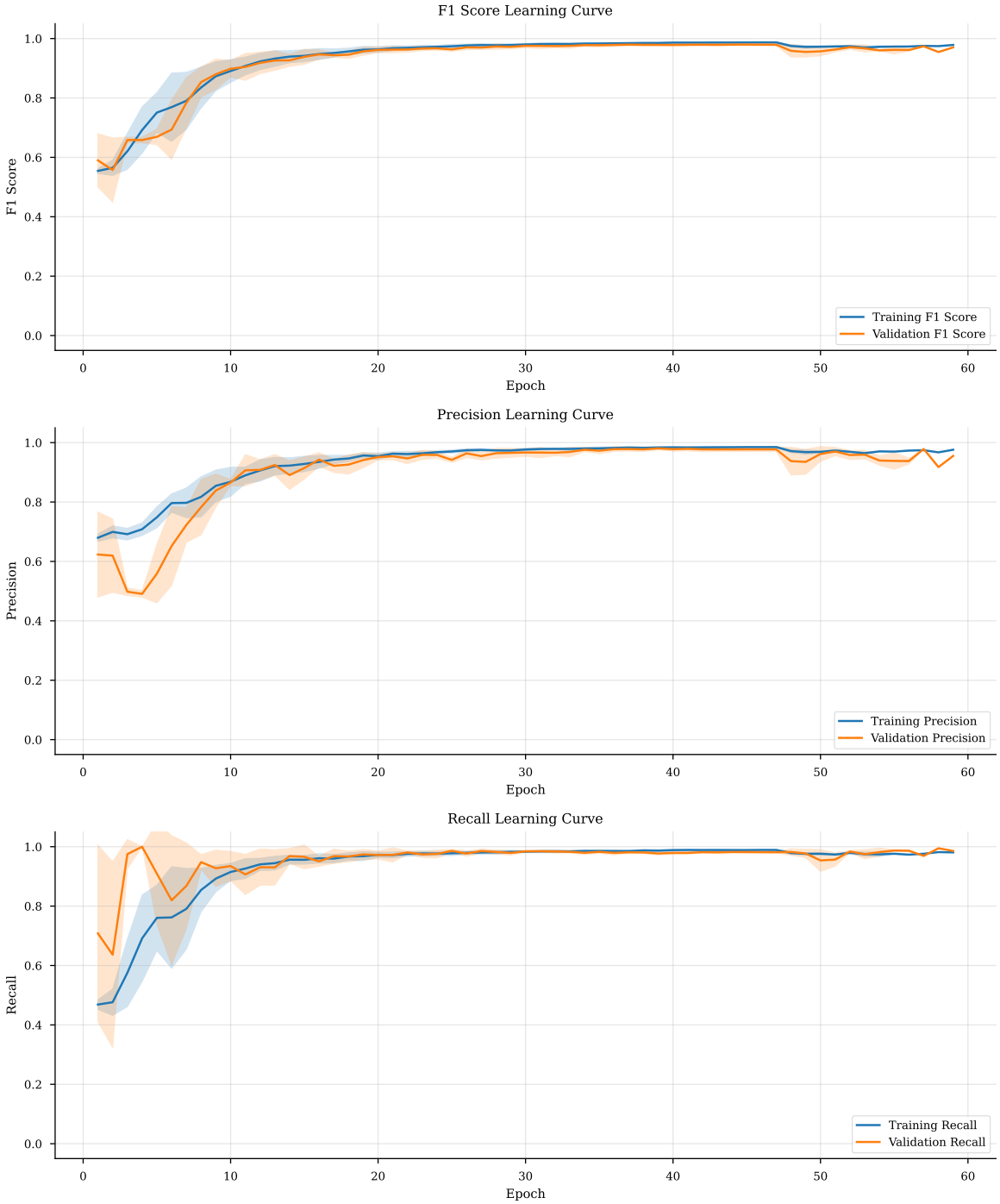

Figure A.20: Learning curves showing the evolution of model performance metrics during training for the time series approach with seed 73.

Seed 99 Analysis

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.066 (0.054–0.078)	0.103
Accuracy	0.981 (0.976–0.987)	0.962
Pr Auc	0.993 (0.992–0.994)	0.962
Roc Auc	0.995 (0.993–0.996)	0.992
Precision	0.977 (0.966–0.988)	0.906
Recall	0.986 (0.979–0.992)	0.932
Tp	1665.0 (1653.1–1676.9)	385.0
Fp	38.6 (19.4–57.8)	40.0
Tn	1650.8 (1631.6–1670.0)	1326.0
Fn	24.4 (13.1–35.7)	28.0
F1 Score	0.981 (0.976–0.987)	0.919
Specificity	0.977 (0.966–0.989)	0.971

Table A.11: Time series Model Performance Metrics for seed 99.

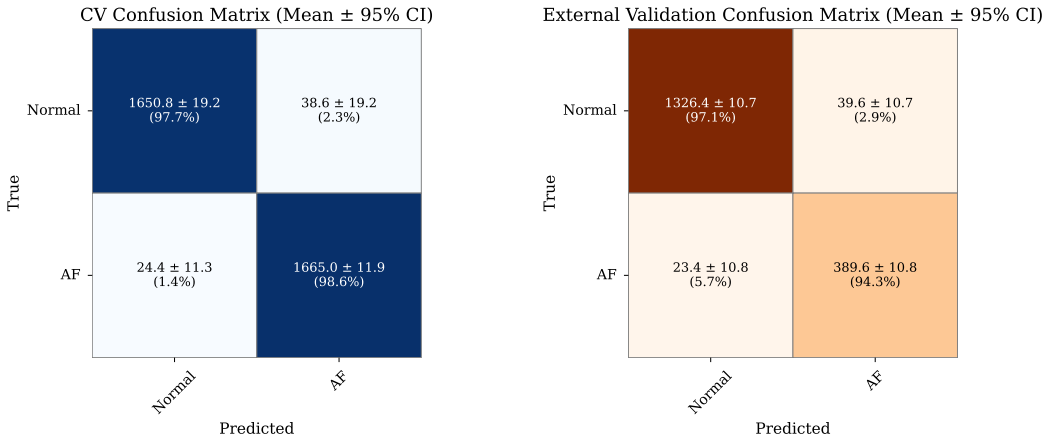

Figure A.21: Averaged confusion matrices displaying the classification performance of the time series model with seed 99.

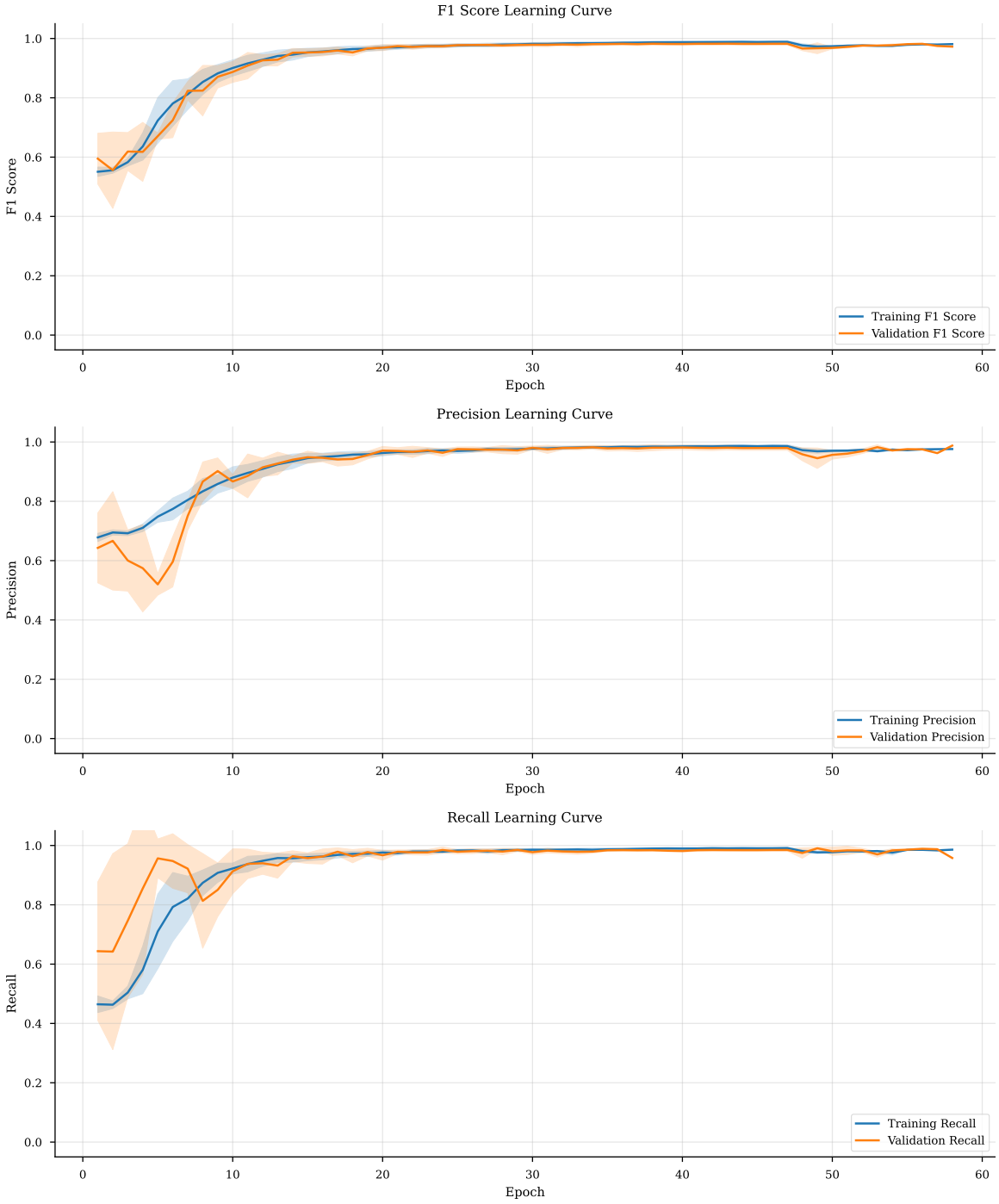

Figure A.22: Learning curves showing the evolution of model performance metrics during training for the time series approach with seed 99.

Seed 122 Analysis

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.067 (0.049–0.086)	0.121
Accuracy	0.983 (0.977–0.988)	0.974
Pr Auc	0.992 (0.990–0.995)	0.948
Roc Auc	0.994 (0.992–0.996)	0.990
Precision	0.980 (0.970–0.989)	0.926
Recall	0.986 (0.981–0.991)	0.966
Tp	1665.6 (1656.9–1674.3)	399.0
Fp	34.2 (17.6–50.8)	32.0
Tn	1655.2 (1638.1–1672.3)	1334.0
Fn	23.8 (15.5–32.1)	14.0
F1 Score	0.983 (0.977–0.988)	0.945
Specificity	0.980 (0.970–0.990)	0.977

Table A.12: Time series Model Performance Metrics for seed 122.

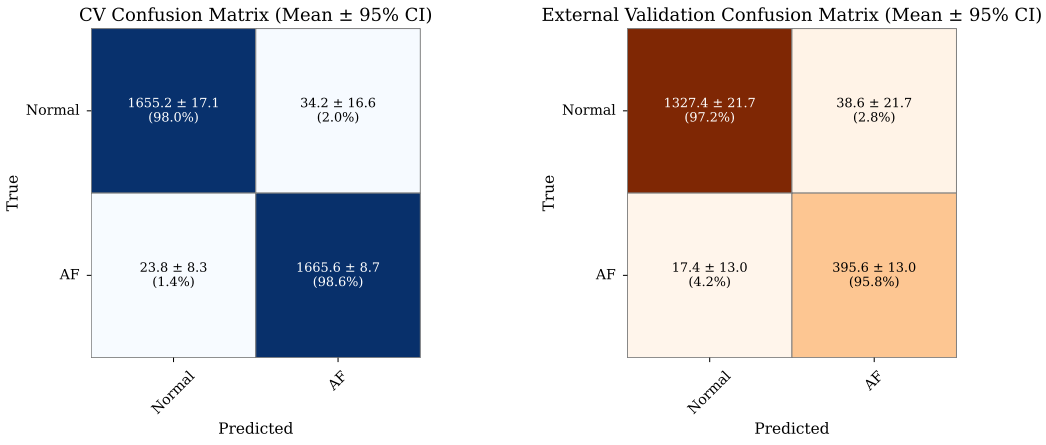

Figure A.23: Averaged confusion matrices displaying the classification performance of the time series model with seed 122.

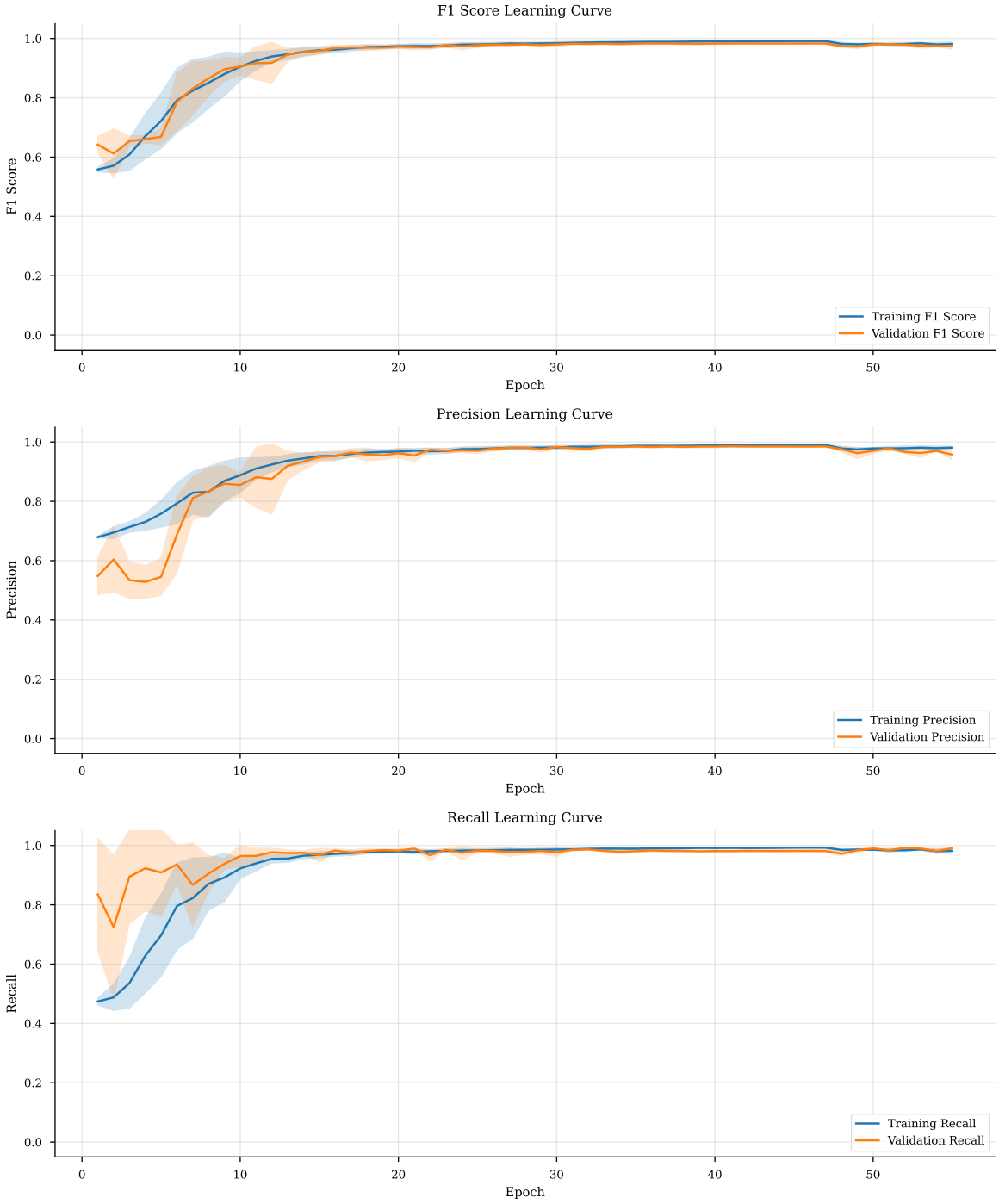

Figure A.24: Learning curves showing the evolution of model performance metrics during training for the time series approach with seed 122.

A.1 Multimodal Modality Performance

A.1.1 Seed 42 Analysis

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.084 (0.075–0.094)	0.069
Accuracy	0.993 (0.991–0.996)	0.996
Pr Auc	0.998 (0.997–0.998)	0.994
Roc Auc	0.998 (0.998–0.999)	0.999
Precision	0.993 (0.988–0.999)	0.988
Recall	0.993 (0.991–0.995)	0.993
Tp	1677.4 (1674.0–1680.8)	410.0
Fp	11.2 (1.6–20.8)	5.0
Tn	1678.2 (1669.1–1687.3)	1361.0
Fn	12.0 (9.2–14.8)	3.0
F1 Score	0.993 (0.991–0.996)	0.990
Specificity	0.993 (0.988–0.999)	0.996

Table A.13: Multimodal Model Performance Metrics for seed 42.

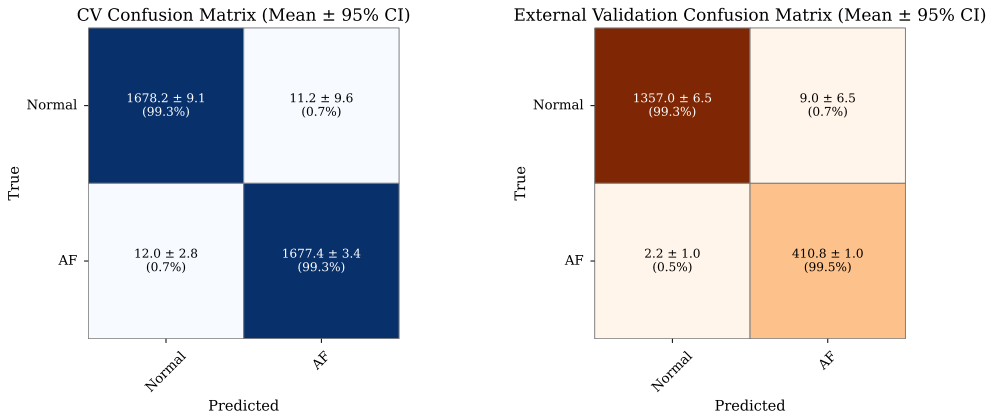

Figure A.25: Averaged confusion matrices displaying the classification performance of the multimodal model with seed 42.

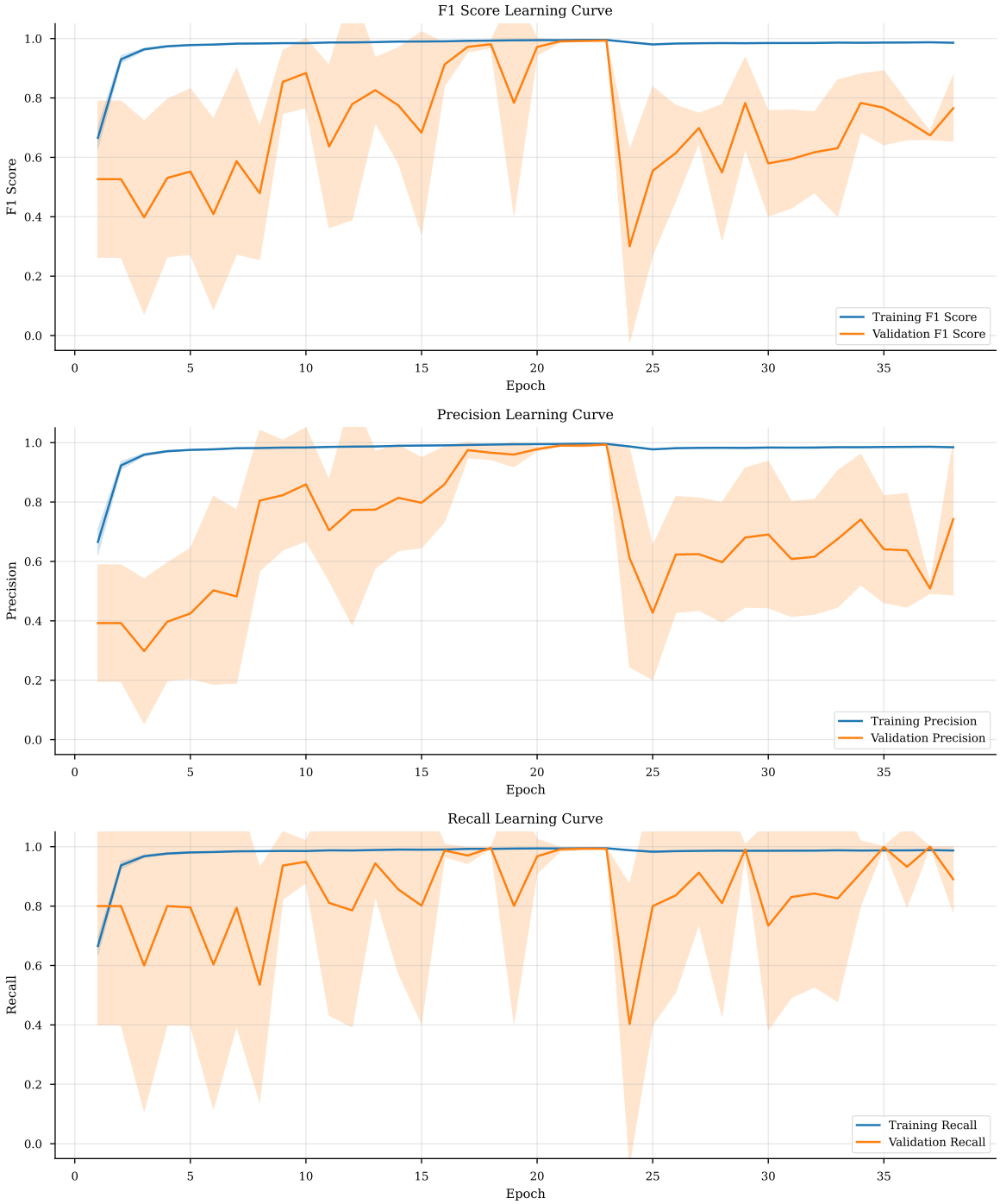

Figure A.26: Learning curves showing the evolution of model performance metrics during training for the multimodal approach with seed 42.

A.1.2 Seed 73 Analysis

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.086 (0.075–0.096)	0.081
Accuracy	0.993 (0.990–0.995)	0.994
Pr Auc	0.997 (0.996–0.999)	0.994
Roc Auc	0.998 (0.997–0.999)	0.999
Precision	0.992 (0.985–1.000)	0.988
Recall	0.993 (0.990–0.996)	0.988
Tp	1677.4 (1671.7–1683.1)	408.0
Fp	13.0 (0.7–25.3)	5.0
Tn	1676.4 (1664.5–1688.3)	1361.0
Fn	12.0 (6.6–17.4)	5.0
F1 Score	0.993 (0.990–0.995)	0.988
Specificity	0.992 (0.985–1.000)	0.996

Table A.14: Multimodal Model Performance Metrics for seed 73.

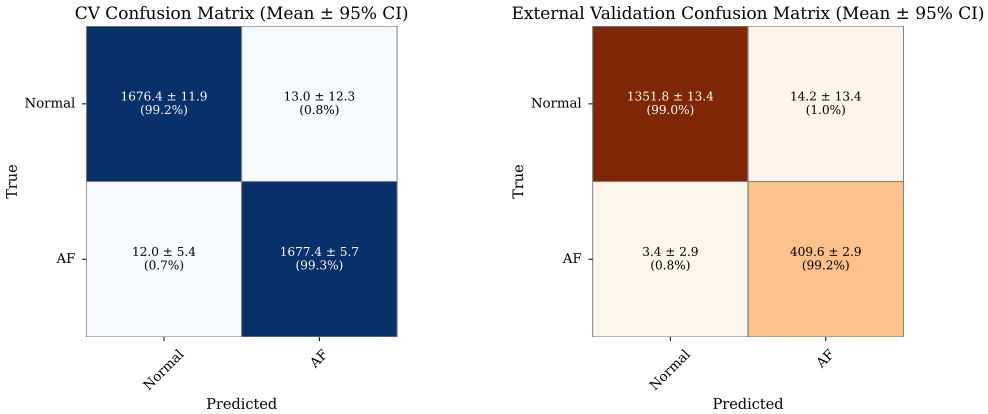

Figure A.27: Averaged confusion matrices displaying the classification performance of the multimodal model with seed 73.

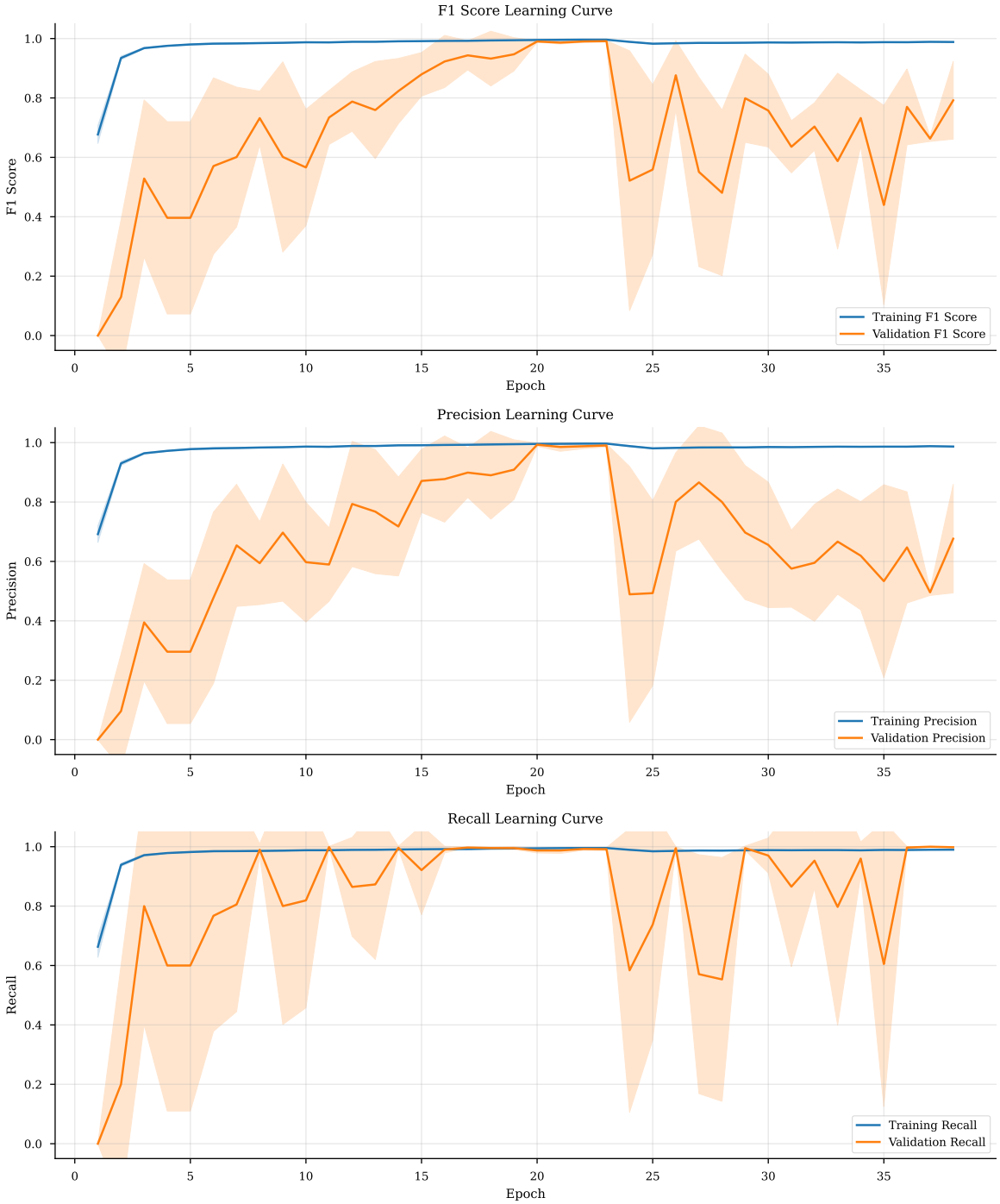

Figure A.28: Learning curves showing the evolution of model performance metrics during training for the multimodal approach with seed 73.

A.1.3 Seed 99 Analysis

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.086 (0.076–0.097)	0.097
Accuracy	0.993 (0.990–0.996)	0.989
Pr Auc	0.997 (0.995–0.999)	0.988
Roc Auc	0.998 (0.997–0.999)	0.998
Precision	0.992 (0.987–0.998)	0.965
Recall	0.993 (0.989–0.997)	0.990
Tp	1678.0 (1670.7–1685.3)	409.0
Fp	13.0 (3.6–22.4)	15.0
Tn	1676.4 (1667.0–1685.8)	1351.0
Fn	11.4 (4.3–18.5)	4.0
F1 Score	0.993 (0.990–0.996)	0.977
Specificity	0.992 (0.987–0.998)	0.989

Table A.15: Multimodal Model Performance Metrics for seed 99.

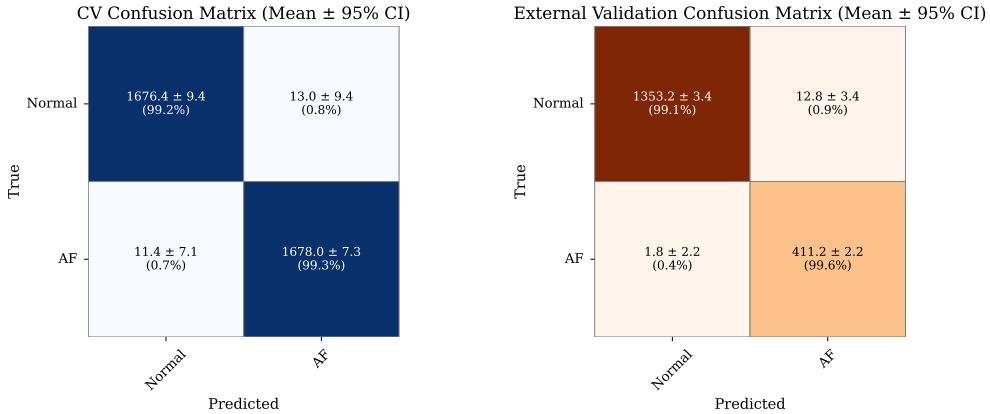

Figure A.29: Averaged confusion matrices displaying the classification performance of the multimodal model with seed 99.

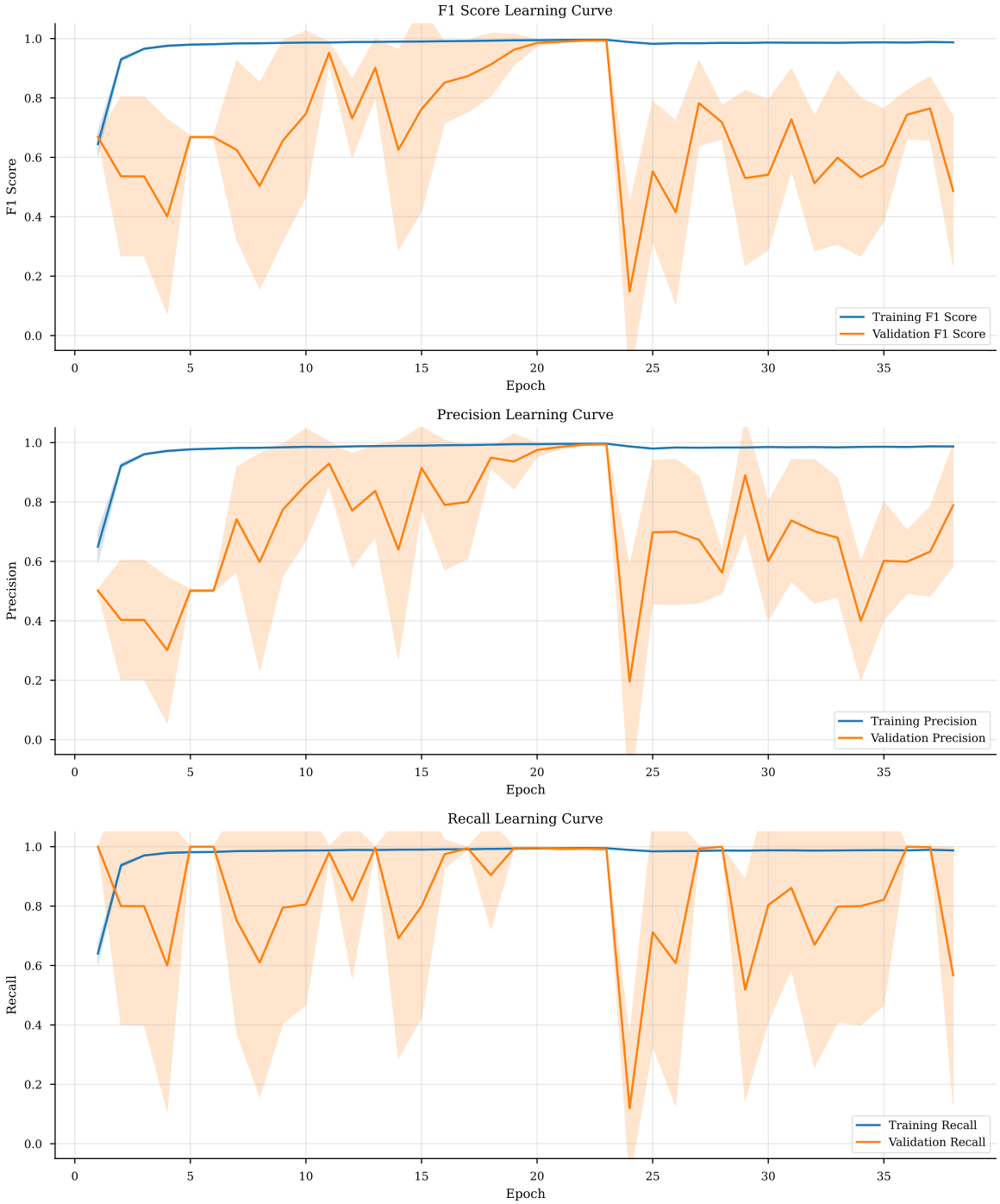

Figure A.30: Learning curves showing the evolution of model performance metrics during training for the multimodal approach with seed 99.

A.1.4 Seed 122 Analysis

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.085 (0.077–0.093)	0.116
Accuracy	0.993 (0.990–0.995)	0.988
Pr Auc	0.996 (0.994–0.998)	0.978
Roc Auc	0.997 (0.996–0.999)	0.997
Precision	0.991 (0.989–0.994)	0.952
Recall	0.994 (0.991–0.998)	0.998
Tp	1679.6 (1673.7–1685.5)	412.0
Fp	15.0 (11.0–19.0)	21.0
Tn	1674.4 (1670.8–1678.0)	1345.0
Fn	9.8 (4.1–15.5)	1.0
F1 Score	0.993 (0.990–0.995)	0.974
Specificity	0.991 (0.989–0.994)	0.985

Table A.16: Multimodal Model Performance Metrics for seed 122.

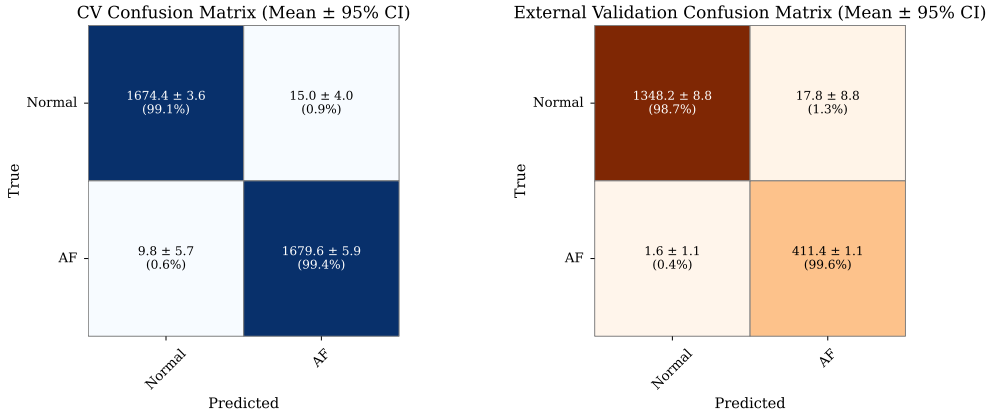

Figure A.31: Averaged confusion matrices displaying the classification performance of the multimodal model with seed 122.

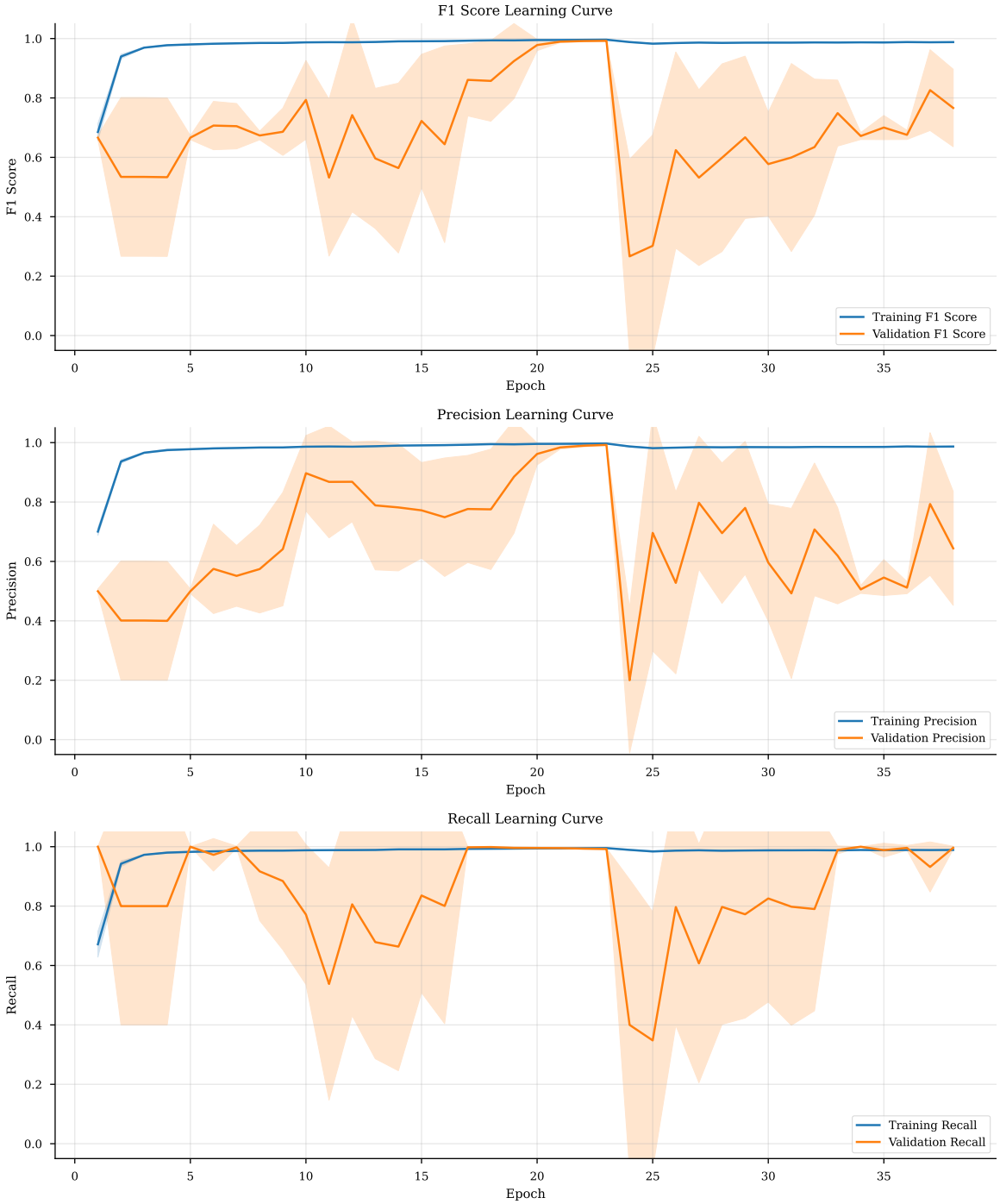

Figure A.32: Learning curves showing the evolution of model performance metrics during training for the multimodal approach with seed 122.

Appendix B

Individual Modality Performance for Experiment 2

B.1 Individual Modality Performance

This section shares the same structure of the experiment 1.

B.1.1 Image Modality

Performance Metrics

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.031 (0.023–0.038)	0.034
Accuracy	0.976 (0.972–0.980)	0.963
Pr Auc	0.924 (0.915–0.934)	0.926
Roc Auc	0.989 (0.986–0.992)	0.988
Precision	0.842 (0.796–0.887)	0.851
Recall	0.905 (0.882–0.928)	0.946
Tp	1640.6 (1598.8–1682.4)	1683.0
Fp	312.8 (200.7–424.9)	295.0
Tn	17960.6 (17847.9–18073.3)	8571.0
Fn	171.6 (129.8–213.4)	97.0
F1 Score	0.872 (0.856–0.887)	0.896
Specificity	0.983 (0.977–0.989)	0.967

Table B.1: Image model performance metrics.

Performance Visualization

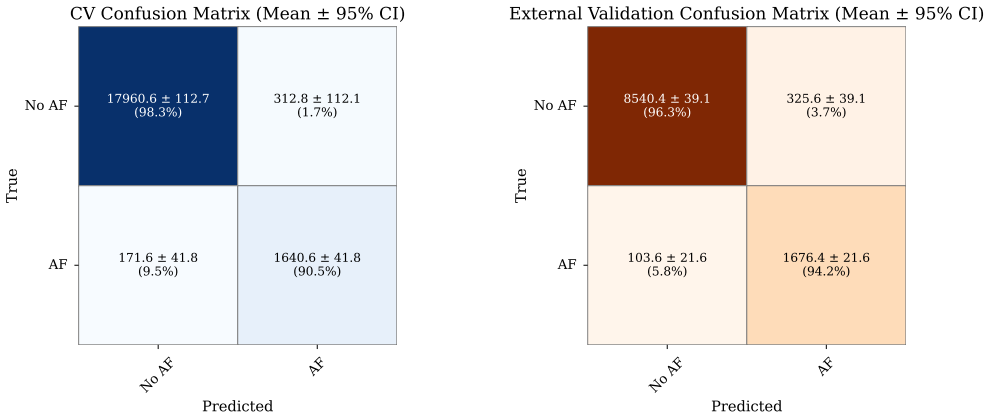

Figure B.1: Image model confusion matrices.

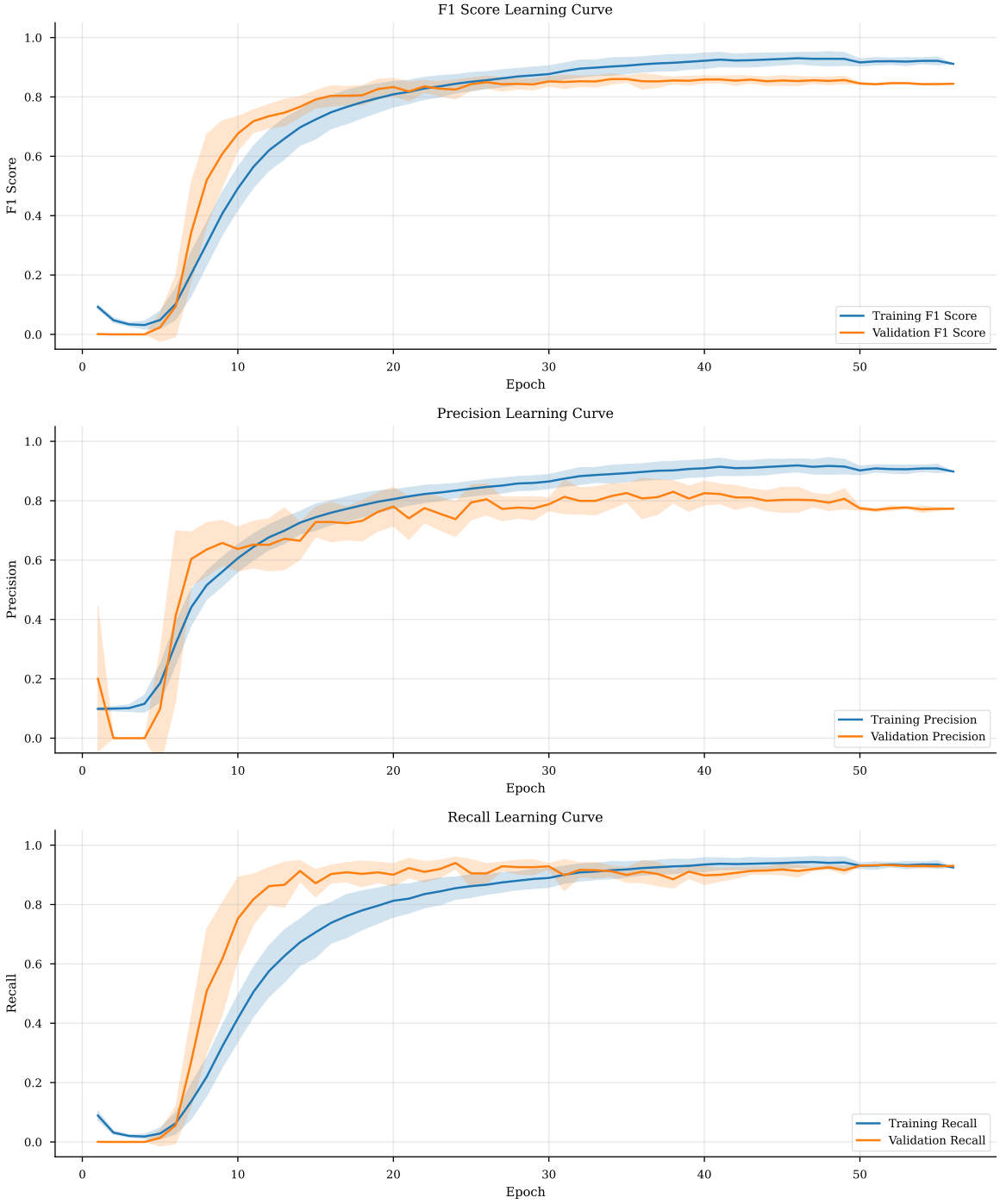

Figure B.2: Image model learning curves.

B.1.2 Spectrogram Modality

Performance Metrics

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.030 (0.028–0.032)	0.039
Accuracy	0.973 (0.970–0.975)	0.958
Pr Auc	0.911 (0.899–0.923)	0.907
Roc Auc	0.988 (0.986–0.989)	0.986
Precision	0.824 (0.806–0.842)	0.827
Recall	0.884 (0.867–0.901)	0.947
Tp	1602.4 (1572.1–1632.7)	1686.0
Fp	342.2 (300.6–384.8)	353.0
Tn	17931.2 (17888.2–17974.2)	8513.0
Fn	209.8 (179.3–240.3)	94.0
F1 Score	0.853 (0.840–0.866)	0.883
Specificity	0.981 (0.979–0.984)	0.960

Table B.2: Spectrogram model performance metrics.

Performance Visualization

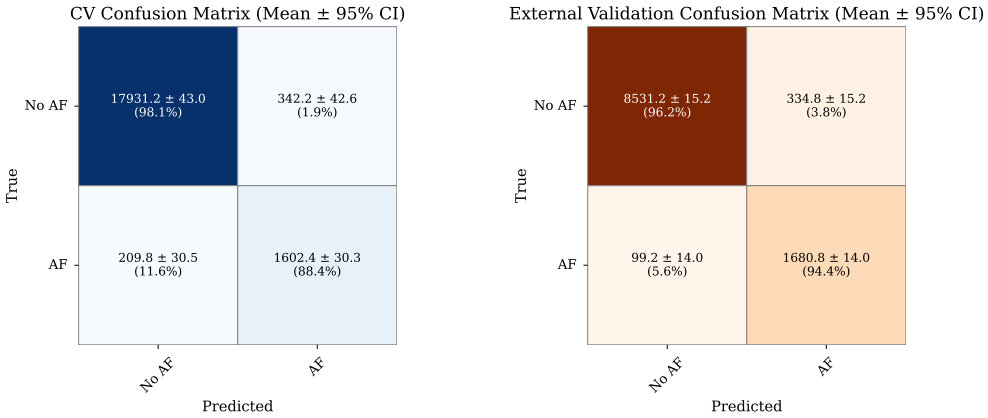

Figure B.3: Spectrogram model confusion matrices.

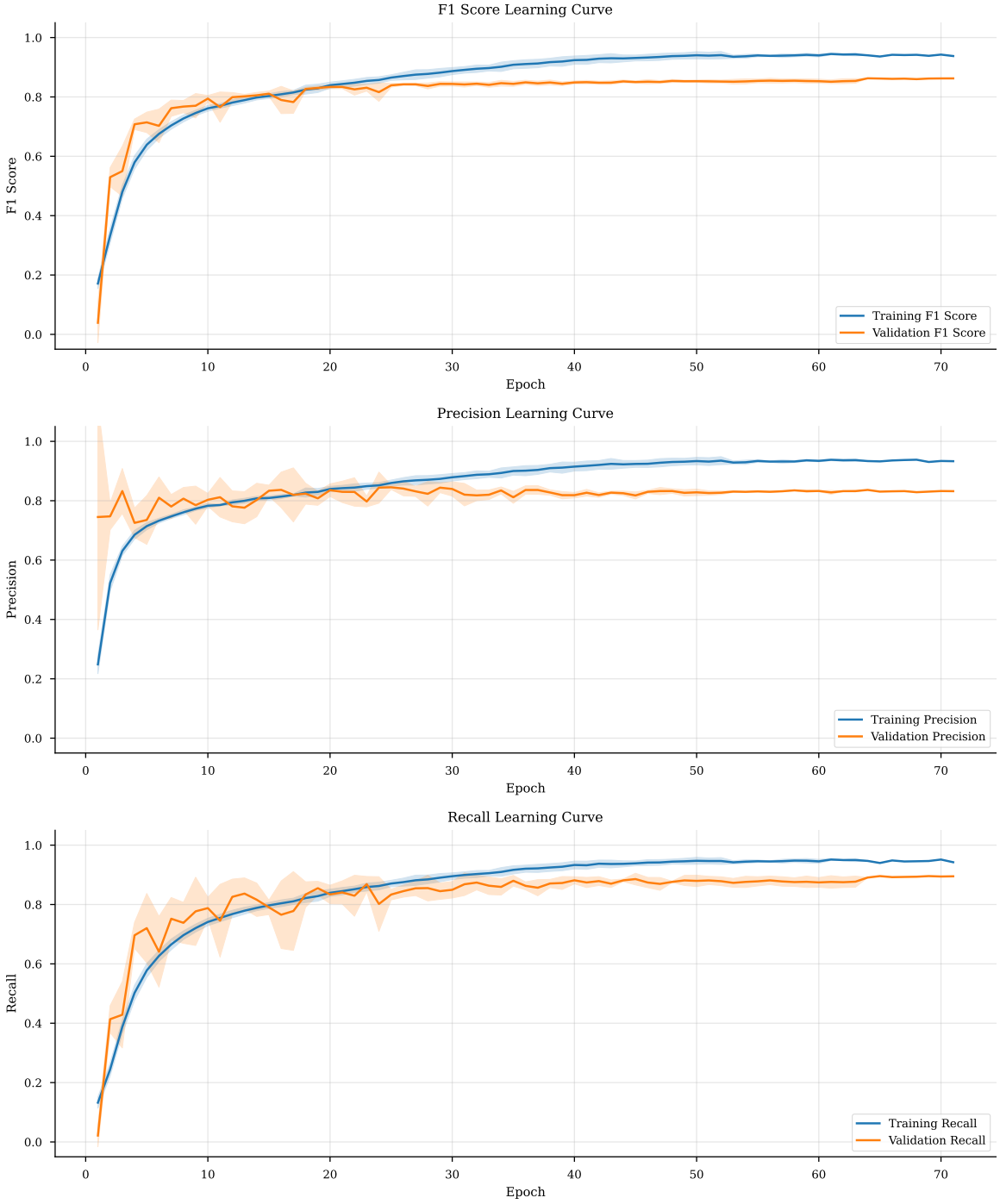

Figure B.4: Spectrogram model learning curves.

B.1.3 Time Series Modality

Performance Metrics

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.012 (0.010–0.013)	0.016
Accuracy	0.976 (0.971–0.981)	0.964
Pr Auc	0.934 (0.920–0.947)	0.919
Roc Auc	0.992 (0.990–0.993)	0.988
Precision	0.821 (0.780–0.861)	0.852
Recall	0.939 (0.924–0.954)	0.947
Tp	1701.0 (1673.6–1728.4)	1685.0
Fp	374.6 (265.0–484.2)	292.0
Tn	17898.8 (17789.0–18008.6)	8574.0
Fn	111.2 (83.9–138.5)	95.0
F1 Score	0.875 (0.853–0.898)	0.897
Specificity	0.980 (0.974–0.985)	0.967

Table B.3: Time series model performance metrics.

Performance Visualization

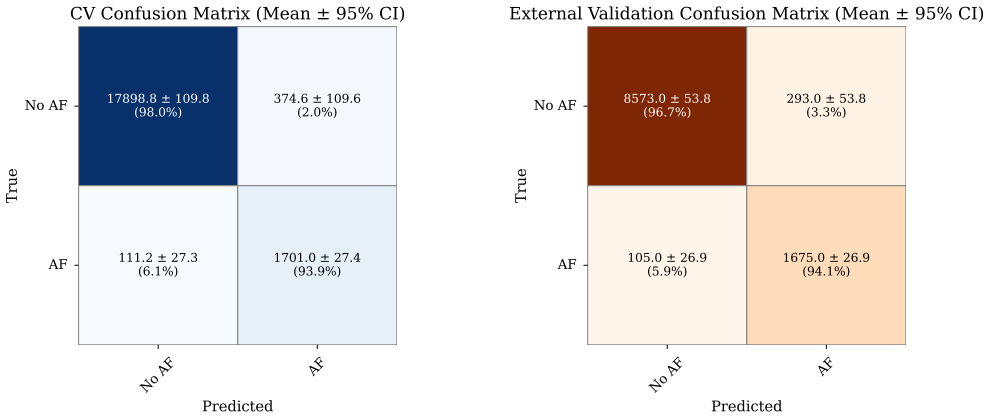

Figure B.5: Time series model confusion matrices.

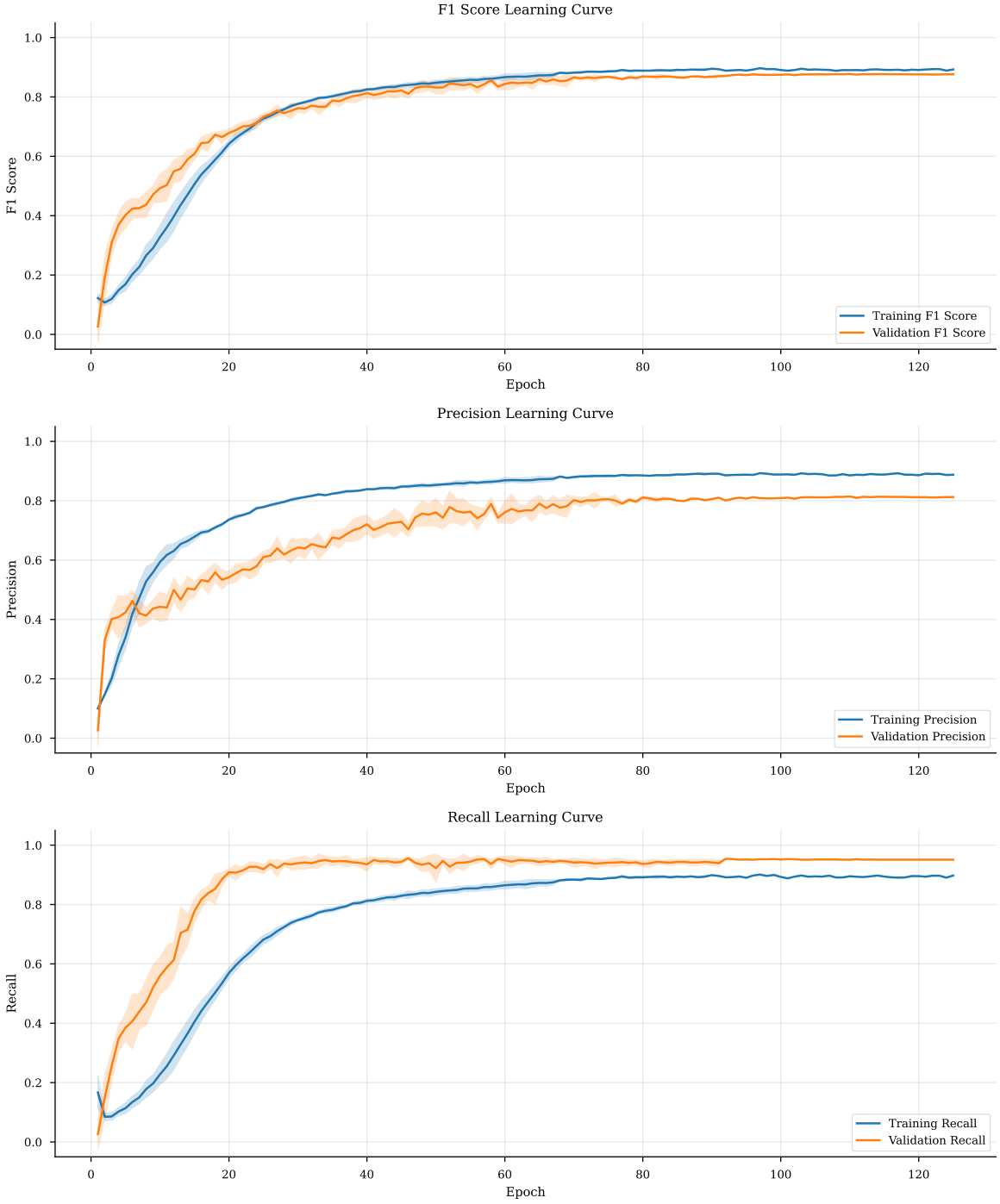

Figure B.6: Time series model learning curves.

B.2 Multimodal Modality Performance

B.2.1 Performance Metrics

Metric	Cross-validation Mean (95% CI)	External Validation Mean
Loss	0.026 (0.020–0.032)	0.046
Accuracy	0.978 (0.974–0.983)	0.958
Pr Auc	0.933 (0.910–0.957)	0.890
Roc Auc	0.992 (0.990–0.994)	0.986
Precision	0.849 (0.807–0.891)	0.836
Recall	0.928 (0.879–0.977)	0.933
Tp	1681.4 (1592.6–1770.2)	1661.0
Fp	302.2 (197.2–407.2)	327.0
Tn	17971.2 (17866.5–18075.9)	8539.0
Fn	130.8 (42.2–219.4)	119.0
F1 Score	0.886 (0.863–0.909)	0.882
Specificity	0.983 (0.978–0.989)	0.963

Table B.4: Multimodal model performance metrics.

B.2.2 Performance Visualization

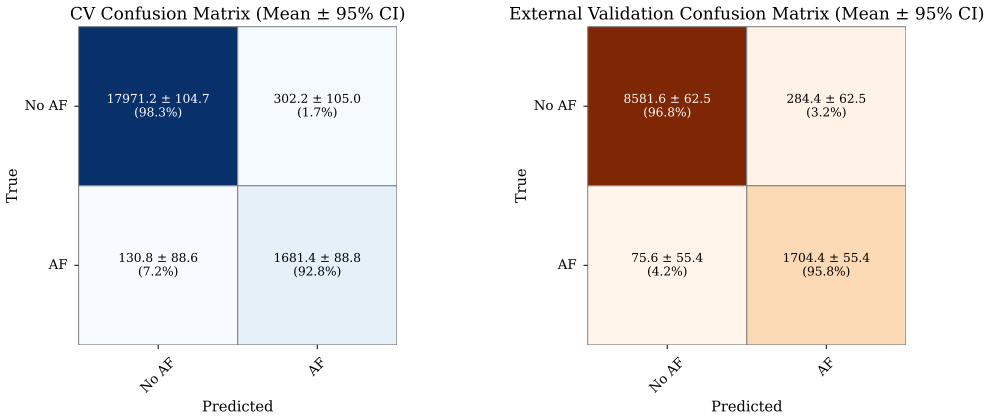

Figure B.7: Multimodal model confusion matrices.

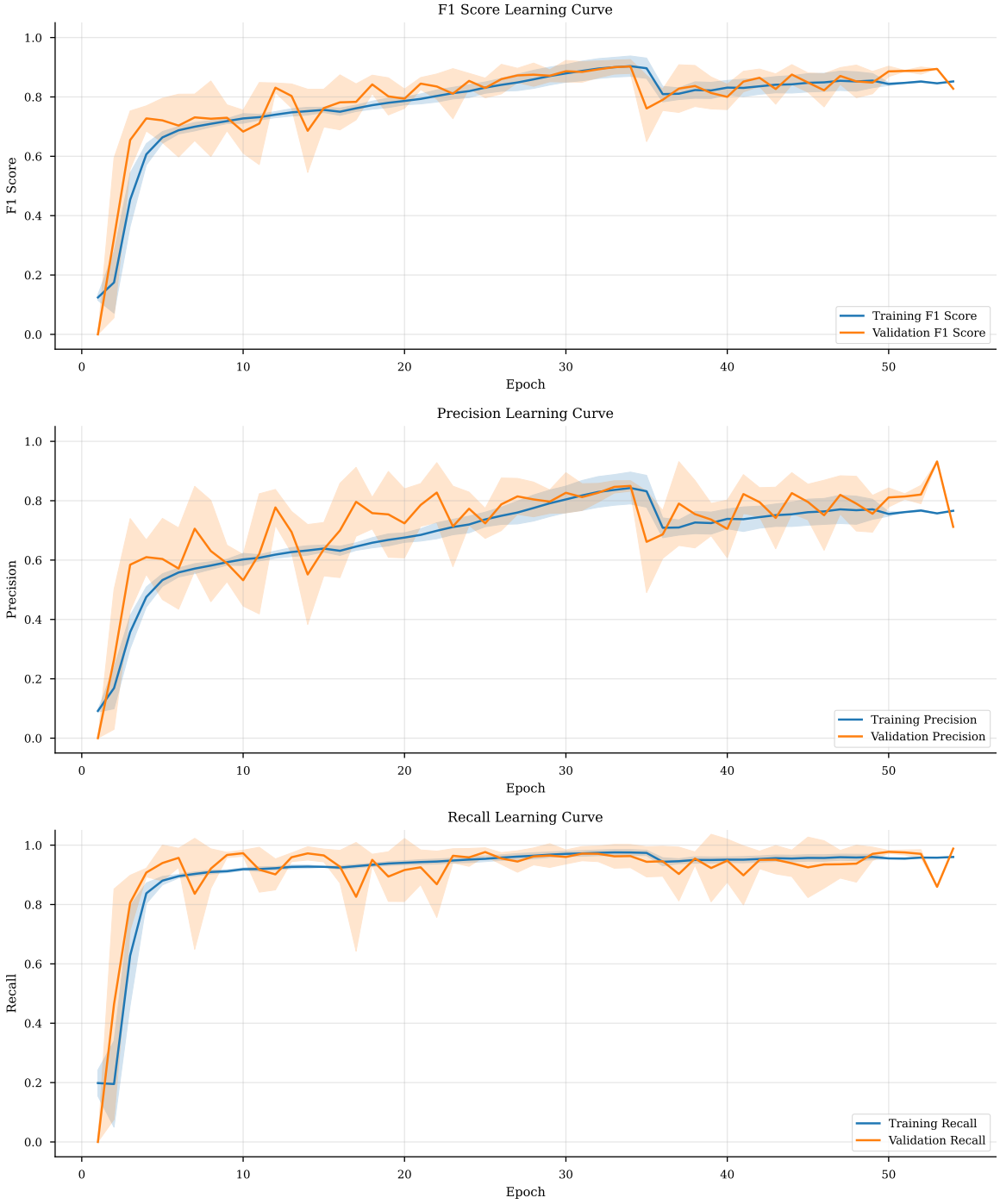

Figure B.8: Multimodal model learning curves.

Bibliography

[AAJH23] Bader Aldughayfiq, Farzeen Ashfaq, NZ Jhanjhi, and Mamoona Humayun. A deep learning approach for atrial fibrillation classification using multi-feature time series data from ecg and ppg. *Diagnostics*, 13(14):2442, 2023.

[AAP20] Mohammed Tali Almalchy, Sarmad Monadel Sabree ALGayar, and Nirvana Popescu. Atrial fibrillation automatic diagnosis based on ecg signal using pretrained deep convolution neural network and svm multiclass model. In *2020 13th International Conference on Communications (COMM)*, pages 197–202. IEEE, 2020.

[AK20] Zeeshan Ahmad and Naimul Mefraz Khan. Multi-level stress assessment using multi-domain fusion of ecg signal. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4518–4521. IEEE, 2020.

[AKSH] Sona Alyounis, Ahsan Khandoker, Cesare Stefanini, and Leontios Hadjileontiadis. A hybrid cnn-lstm model for heart failure detection using raw ecg signals.

[ALPP24] R Anand, S Vijaya Lakshmi, Digvijay Pandey, and Binay Kumar Pandey. An enhanced resnet-50 deep learning model for arrhythmia detection using electrocardiogram biomedical indicators. *Evolving Systems*, 15(1):83–97, 2024.

[APP19] Rasmus S Andersen, Abdolrahman Peimankar, and Sadasivan Puthusserypady. A deep learning approach for real-time detection of atrial fibrillation. *Expert Systems with Applications*, 115:465–473, 2019.

[BAM18] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[BHP+23] Nhat-Tan Bui, Dinh-Hieu Hoang, Thinh Phan, Minh-Triet Tran, Brijesh Patel, Donald Adjeroh, and Ngan Le. Tsrnet: Simple framework for real-time ecg anomaly detection with multimodal time and spectrogram restoration network. *arXiv preprint arXiv:2312.10187*, 2023.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. Focus on Chapter: Maximum Likelihood.

[BLZ+18] Chaim Baskin, Natan Liss, Evgenii Zheltonozhskii, Alex M Bronstein, and Avi Mendelson. Streaming architecture for large-scale quantized neural networks on an fpga-based dataflow platform. In *2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 162–169. IEEE, 2018.

[BSAH+19] Mohammed Baydoun, Lise Safatly, Ossama K Abou Hassan, Hassan Ghaziri, Ali El Hajj, and Hussain Isma'eel. High precision digitization of paper-based ecg records: a step toward machine learning. *IEEE journal of translational engineering in health and medicine*, 7:1–8, 2019.

[bvsnd] Use o coração para vencer as doenças cardiovasculares. Online, n.d.

[CAM05] Gari D Clifford, Francisco Azuaje, and Patrick E McSharry. *Advanced methods and tools for ECG data analysis*. Artech House, 2005.

[CCG+20] Wenjuan Cai, Yundai Chen, Jun Guo, Baoshi Han, Yajun Shi, Lei Ji, Jinliang Wang, Guanglei Zhang, and Jianwen Luo. Accurate detection of atrial fibrillation from 12-lead ecg using deep neural network. *Computers in biology and medicine*, 116:103378, 2020.

[CCT+23] P. A. Chousou, R. Chattopadhyay, V. Tsampasian, V. S. Vassiliou, and P. J. Pugh. Electrocardiographic predictors of atrial fibrillation. *Medical Sciences*, 11(2):30, 2023.

[CPMCS+22] Francisco Carrillo-Perez, Juan Carlos Morales, Daniel Castillo-Secilla, Olivier Gevaert, Ignacio Rojas, and Luis Javier Herrera. Machine-learning-based late fusion on multi-omics and multi-scale data for non-small-cell lung cancer diagnosis. *Journal of Personalized Medicine*, 12(4):601, 2022.

[CRS22] Juan Pablo Gasca Calderón, Diego Fernando Gonzalez Ruiz, and Ruthber Rodríguez Serrezuela. Time-frequency analysis of the emg signal for the identification of hand grasping postures. In *2022 V Congreso Internacional en Inteligencia Ambiental, Ingeniería de Software y Salud Electrónica y Móvil (AmITIC)*, pages 1–8, 2022.

[CS] Cables and Sensors. 12 lead ECG placement guide. `https://www.cablesandsensors.com/pages/12-lead-ecg-placement-guide-with-illustrations`. Accessed: 2023-1-29.

[CS15] Mukesh P Chawla and Archana Sharma. Baseline wander removal from ecg signal: comparative analysis. *Journal of medical engineering ŏ026 technology*, 39(3):152–161, 2015.

[CSC+23] Sanghoon Choi, Hyo-Chang Seo, Min Soo Cho, Segyeong Joo, and Gi-Byoung Nam. Performance improvement of deep learning based multi-class ecg classification model using limited medical dataset. *IEEE Access*, 11:53185–53194, 2023.

[CWL+23] Xun Chen, Chengqi Wang, Yuxin Li, Chao Hu, Qin Wang, and Dupeng Cai. Research on bidirectional lstm recurrent neural network in speech recognition. In *2023 IEEE 6th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, pages 878–883. IEEE, 2023.

[dOMW+21] Jéssica dos Santos de Oliveira, Clément Bernardo Marques, Maria Fernanda Wanderley, Priscilla Koch Wagner, and Walter Martins Filho. Classifying ecg exams of different formats and sources using convolutional networks. In *2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM)*, pages 1–4. IEEE, 2021.

[DPC24] Ítalo Flexa Di Paolo and Adriana Rosa Garcez Castro. Intra-and interpatient ecg heartbeat classification based on multimodal convolutional neural networks with an adaptive attention mechanism. *Applied Sciences*, 14(20):9307, 2024.

[DRM+23] Felipe M Dias, Estela Ribeiro, Ramon A Moreno, Adele H Ribeiro, Nelson Samesima, Carlos A Pastore, Jose E Krieger, and Marco A Gutierrez. Artificial intelligence-driven screening system for rapid image-based classification of 12-lead ecg exams: A promising solution for emergency room prioritization. *Ieee Access*, 2023.

[DSR+21] Felipe M Dias, Nelson Samesima, Adele Ribeiro, Ramon A Moreno, Carlos A Pastore, Jose E Krieger, and Marco A Gutierrez. 2d image-based atrial fibrillation classification. In *2021 Computing in Cardiology (CinC)*, volume 48, pages 1–4. IEEE, 2021.

[EBE24] Alaa Eleyan, Fatih Bayram, and Gülden Eleyan. Spectrogram-based arrhythmia classification using three-channel deep learning model with feature fusion. *Applied Sciences*, 14(21):9936, 2024.

[FA21] Oliver Faust and U Rajendra Acharya. Automated classification of five arrhythmias and normal sinus rhythm based on rr interval signals. *Expert Systems with Applications*, 181:115031, 2021.

[Fav21] Desiderio Favarato. Brazilian population presents prevalence of atrial fibrillation similar to higher income countries, and a low use of anticoagulation therapy, 2021.

[FCL+21] Bo Fang, Junxin Chen, Yu Liu, Wei Wang, Ke Wang, Amit Kumar Singh, and Zhihan Lv. Dual-channel neural network for atrial fibrillation detection from a single lead ecg wave. *IEEE journal of biomedical and health informatics*, 27(5):2296–2305, 2021.

[GAG+00] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-

Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[GC22] Gamal GN Geweid and Jiande DZ Chen. Automatic classification of atrial fibrillation from short single-lead ecg recordings using a hybrid approach of dual support vector machine. *Expert Systems with Applications*, 198:116848, 2022.

[Gea21] Ary L. Goldberger and et al. Multilead electrocardiographic analysis in arrhythmia classification. *Circulation: Arrhythmia and Electrophysiology*, 14:e009981, 2021.

[GGS17] Ary L Goldberger, Zachary D Goldberger, and Alexei Shvilkin. *Clinical electrocardiography: a simplified approach e-book*. Elsevier Health Sciences, 2017.

[HCD+20] Amitava Halder, Saptarshi Chatterjee, Debangshu Dey, Surajit Kole, and Sugata Munshi. An adaptive morphology based segmentation technique for lung nodule detection in thoracic ct image. *Computer Methods and Programs in Biomedicine*, 197:105720, 2020.

[HCPSLBV24] Roberto Holgado-Cuadrado, Carmen Plaza-Seco, Lisandro Lovisolo, and Manuel Blanco-Velasco. A deep and interpretable learning approach for long-term ecg clinical noise classification. *IEEE Transactions on Biomedical Engineering*, 2024.

[HCZ22] Rui Hu, Jie Chen, and Li Zhou. A transformer-based deep neural network for arrhythmia detection using continuous ecg signals. *Computers in Biology and Medicine*, 144:105325, 2022.

[HPZ+20] Shih-Cheng Huang, Anuj Pareek, Roham Zamanian, Imon Banerjee, and Matthew P Lungren. Multimodal fusion with deep neural networks for

leveraging ct imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific reports*, 10(1):22147, 2020.

[HS97a] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[HS97b] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[HW20] Mei-Ling Huang and Yan-Sheng Wu. Classification of atrial fibrillation and normal sinus rhythm based on convolutional neural network. *Biomedical engineering letters*, 10(2):183–193, 2020.

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 1026–1034, 2015.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[IS15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

[Its15] Itseez. Open source computer vision library. `https://github.com/itseez/opencv`, 2015.

[JC17] Katarzyna Janocha and Wojciech Marian Czarnecki. On loss functions for deep neural networks in classification. *arXiv preprint arXiv:1702.05659*, 2017.

[JCL+21] Yong-Yeon Jo, Younghoon Cho, Soo Youn Lee, Joon-myoung Kwon, Kyung-Hee Kim, Ki-Hyun Jeon, Soohyun Cho, Jinsik Park, and Byung-

Hee Oh. Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram. *International journal of cardiology*, 328:104–110, 2021.

[JJLJ20] Hohyub Jeon, Yongchul Jung, Seongjoo Lee, and Yunho Jung. Area-efficient short-time fourier transform processor for time–frequency analysis of non-stationary signals. *Applied Sciences*, 10(20):7208, 2020.

[JS21] Rashi Jaiswal and Brijendra Singh. A comparative study of loss functions for deep neural networks in time series analysis. In *International Conference on Big Data, Machine Learning, and Applications*, pages 147–163. Springer, 2021.

[KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[KJ22] Bartłomiej Król-Józaga. Atrial fibrillation detection using convolutional neural networks on 2-dimensional representation of ecg signal. *Biomedical Signal Processing and Control*, 74:103470, 2022.

[KLK+24] Jiwoong Kim, Sun Jung Lee, Bonggyun Ko, Myungeun Lee, Young-Shin Lee, and Ki Hong Lee. Identification of atrial fibrillation with single-lead mobile ecg during normal sinus rhythm using deep learning. *Journal of Korean Medical Science*, 39(5), 2024.

[KPS+22] Devender Kumar, Abdolrahman Peimankar, Kamal Sharma, Helena Domínguez, Sadasivan Puthusserypady, and Jakob E Bardram. Deepaware: A hybrid deep learning and context-aware heuristics-based model for atrial fibrillation detection. *Computer Methods and Programs in Biomedicine*, 221:106899, 2022.

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[LDZ+20] Sidrah Liaqat, Kia Dashtipour, Adnan Zahid, Khaled Assaleh, Kamran Arshad, and Naeem Ramzan. Detection of atrial fibrillation using a machine learning approach. *Information*, 11(12):549, 2020.

[Lin17] T Lin. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

[LLW20] Yutong Li, Linlin Li, and Ruoxi Wang. A review of feature extraction and classification algorithms for ecg signals. *Journal of Physics: Conference Series*, 1575(1):012015, 2020.

[LQD+22] Chengfan Li, Yueyu Qi, Xuehai Ding, Junjuan Zhao, Tian Sang, and Matthew Lee. A deep learning method approach for sleep stage classification with eeg spectrogram. *International Journal of Environmental Research and Public Health*, 19(10):6322, 2022.

[MAA20] Sajad Mousavi, Fatemeh Afghah, and U Rajendra Acharya. Han-ecg: An interpretable atrial fibrillation detection model using hierarchical attention networks. *Computers in biology and medicine*, 127:104057, 2020.

[MAEHS22] Farida Mohsen, Hazrat Ali, Nady El Hajj, and Zubair Shah. Artificial intelligence-based methods for fusion of electronic health records and imaging data. *Scientific Reports*, 12(1):17981, 2022.

[MBI+23] Tanjim Mahmud, Anik Barua, Dilshad Islam, Mohammad Shahadat Hossain, Rishita Chakma, Koushick Barua, Mahabuba Monju, and Karl Andersson. Ensemble deep learning approach for ecg-based cardiac disease detection: Signal and image analysis. In *2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, pages 70–74. IEEE, 2023.

[MMZ23] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pages 23803–23828. PMLR, 2023.

[MNZS15] Hassan Mohamed, Abdelazim Negm, Mohamed Zahran, and Oliver C Saavedra. Assessment of artificial neural network for bathymetry estimation using high resolution satellite imagery in shallow lakes: Case study el burullus lake. In *International water technology conference*, pages 12–14, 2015.

[Mob] MD Kamrujjaman Mobin. Transforming ecg images to waveforms with u-net-lstm and multilabel classification with resnet-lstm: A multimodal approach by pulseplex.

[MS12] M S Manikandan and K P Soman. Ecg signal processing: A survey. *Proceedings of the International Conference on Advanced Computing and Communication Systems*, pages 151–155, 2012.

[MSY+21] Fatma Murat, Ferhat Sadak, Ozal Yildirim, Muhammed Talo, Ender Murat, Murat Karabatak, Yakup Demir, Ru-San Tan, and U Rajendra Acharya. Review of deep learning-based atrial fibrillation detection studies. *International journal of environmental research and public health*, 18(21):11302, 2021.

[MZC+20] Fengying Ma, Jingyao Zhang, Wei Chen, Wei Liang, and Wenjia Yang. An automatic system for atrial fibrillation by using a cnn-lstm model. *Discrete Dynamics in Nature and Society*, 2020:1–9, 2020.

[NE21] Huseyin Nasifoglu and Osman Erogul. Convolutional neural networks based osa event prediction from ecg scalograms and spectrograms. 2021.

[NFMS23] Vipul Narayan, Mohammad Faiz, Pawan Kumar Mall, and Swapnita Srivastava. A comprehensive review of various approach for medical image segmentation and disease prediction. *Wireless Personal Communications*, 132(3):1819–1848, 2023.

[NGJ24] Z Nesheiwat, A Goyal, and M Jagtap. *Atrial Fibrillation*. StatPearls Publishing, Treasure Island (FL), Updated 2023 Apr 26 edition, 2024. StatPearls [Internet].

[NH10] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[NKK+11] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[NMOHK22] Myat Thet Nyo, F Mebarek-Oudina, Su Su Hlaing, and Nadeem A Khan. Otsu's thresholding technique for mri image brain tumor segmentation. *Multimedia tools and applications*, 81(30):43837–43849, 2022.

[NTD+20] Siti Nurmaini, Alexander Edo Tondas, Annisa Darmawahyuni, Muhammad Naufal Rachmatullah, Radiyati Umi Partan, Firdaus Firdaus, Bambang Tutuko, Ferlita Pratiwi, Andre Herviant Juliano, and Rahmi Khoirani. Robust detection of atrial fibrillation from short-term electrocardiogram using convolutional neural networks. *Future Generation Computer Systems*, 113:304–317, 2020.

[Ope13] OpenStax College. Illustration from anatomy physiology, connexions web site, 2013. This work is licensed under the Creative Commons Attribution 3.0 International License. To view a copy of this license, visit `http://creativecommons.org/licenses/by/3.0/`.

[OS10] A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall, 2010.

[PCW+20] Yongjie Ping, Chao Chen, Lu Wu, Yinglong Wang, and Minglei Shu. Automatic detection of atrial fibrillation based on cnn-lstm and shortcut connection. In *Healthcare*, volume 8, page 139. MDPI, 2020.

[Pea19] Arjun M. Patel and et al. The role of ecg lead placement in atrial fibrillation detection. *Journal of the American College of Cardiology*, 74:2505–2517, 2019.

[Pea20] Pyotr G. Platonov and et al. Electrocardiographic phenotypes of atrial fibrillation: Detection and clinical implications. *European Heart Journal*, 41(12), 2020.

[PNM23] Gyana Ranjan Patra, Manoj Kumar Naik, and Mihir Narayan Mohanty. Ecg signal classification using a cnn-lstm hybrid network. In *2023 2nd International Conference on Ambient Intelligence in Health Care (ICAIHC)*, pages 1–6. IEEE, 2023.

[PPC+23] P Pabitha, R Praveen, Kamma Cheruvu Jayaraja Chandana, S Ponlibarnaa, and AS Aparnaa. A comparative study of deep learning models for ecg signal-based user classification. In *2023 12th International Conference on Advanced Computing (ICoAC)*, pages 1–8. IEEE, 2023.

[QZZ+24] Tianyu Qi, He Zhang, Huijun Zhao, Chong Shen, and Xiaochen Liu. Research on ecg signal classification based on hybrid residual network. *Applied Sciences*, 14(23):11202, 2024.

[RFPLDGR18] Daniel Ramos, Javier Franco-Pedroso, Alicia Lozano-Diez, and Joaquin Gonzalez-Rodriguez. Deconstructing cross-entropy for probabilistic binary classifiers. *Entropy*, 20(3):208, 2018.

[RHT+] Richard Redina, Jakub Hejc, Fabian Theurl, Tomas Novotny, Irena Andrsova, Katerina Hnatkova, Zdenek Starek, Marina Filipenska, Axel Bauer, and Marek Malik. Deep learning end-to-end approach for precise qrs complex delineation using temporal region-based convolutional neural networks.

[RHW86] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[RKHD23] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-time flying object detection with yolov8. *arXiv preprint arXiv:2305.09972*, 2023.

[RRP+20] Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1760, 2020.

[RS22] Jagdeep Rahul and Lakhan Dev Sharma. Artificial intelligence-based approach for atrial fibrillation detection using normalised and short-duration time-frequency ecg. *Biomedical Signal Processing and Control*, 71:103270, 2022.

[RSAS21] Jayroop Ramesh, Zahra Solatidehkordi, Raafat Aburukba, and Assim Sagahyroon. Atrial fibrillation classification with smart wearables using short-term heart rate variability and deep convolutional neural networks. *Sensors*, 21(21):7233, 2021.

[RSKS22] Sneha Rao, Vishwa Mohan Singh, Siddhivinayak Kulkarni, and Vibhor Saran. A spectrogram based novel approach for arrhythmia detection with convolutional neural networks. 2022.

[RWK+24] Matthew A Reyna, James Weigle, Zuzana Koscova, Kiersten Campbell, Kshama Kodthalu Shivashankara, Soheil Saghafi, Sepideh Nikookar, Mohsen Motie-Shirazi, Yashar Kiarashi, Salman Seyedi, et al. Ecg-image-database: A dataset of ecg images with real-world imaging and scanning artifacts; a foundation for computerized ecg image digitization and analysis. *arXiv preprint arXiv:2409.16612*, 2024.

[Sat23] Malika Satayeva. Multimodal neural network for healthcare applications, 2023.

[SG64] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[SGS+22] Nizar Sakli, Haifa Ghabri, Ben Othman Soufiene, Faris A Almalki, Hedi Sakli, Obaid Ali, and Mustapha Najjari. Resnet-50 for 12-lead electrocardiogram automated diagnosis. *Computational Intelligence and Neuroscience*, 2022(1):7617551, 2022.

[SHK+14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[SHLC20] Lang Su, Chuqing Hu, Guofa Li, and Dongpu Cao. Msaf: Multimodal split attention fusion. *arXiv preprint arXiv:2012.07175*, 2020.

[SKG+21] Afshin Shoeibi, Marjane Khodatars, Navid Ghassemi, Mahboobeh Jafari, Parisa Moridian, Roohallah Alizadehsani, Maryam Panahiazar, Fahime Khozeimeh, Assef Zare, Hossein Hosseini-Nejad, et al. Epileptic seizures detection using deep learning techniques: A review. *International Journal of Environmental Research and Public Health*, 18(11):5780, 2021.

[SLJ+15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[SLZ+23] Zhaoyi Sun, Mingquan Lin, Qingqing Zhu, Qianqian Xie, Fei Wang, Zhiyong Lu, and Yifan Peng. A scoping review on multimodal deep learning in biomedical images and texts. *Journal of Biomedical Informatics*, page 104482, 2023.

[SMS23] Mohd Sakib, Suhel Mustajab, and Tamanna Siddiqui. Deep learning-based heartbeat classification of 12-lead ecg time series signal. In *2023 4th International Conference on Data Analytics for Business and Industry (ICDABI)*, pages 273–278. IEEE, 2023.

[SNP22] Muhammad Farhan Safdar, Robert Marek Nowak, and Piotr Pałka. A denoising and fourier transformation-based spectrograms in ecg classification using convolutional neural network. *Sensors*, 22(24):9576, 2022.

[SPH+23] V Saravanan, BD Parameshachari, Abbas Hameed Abdul Hussein, N Shilpa, and Myasar Mundher Adnan. Deep learning techniques based secured biometric authentication and classification using ecg signal. In *2023 International Conference on Integrated Intelligence and Communication Systems (ICIICS)*, pages 1–5. IEEE, 2023.

[SRM18] Upkar Satija, Baranidharan Ramkumar, and M S Manikandan. A review of signal processing techniques for electrocardiogram signal quality assessment. *IEEE Reviews in Biomedical Engineering*, 11:36–52, 2018.

[SUS22] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics*, 23(2):bbab569, 2022.

[SWM12] Kevin T Sweeney, Tomas E Ward, and Se
'an F McLoone. Reducing false arrhythmia alarms in the icu using ecg and context information. *Physiological measurement*, 33(5):825, 2012.

[SWSS20] Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE journal of biomedical and health informatics*, 25(5):1519–1528, 2020.

[TAA+23] Connie W Tsao, Aaron W Aday, Zaid I Almarzooq, Cheryl AM Anderson, Pankaj Arora, Christy L Avery, Carissa M Baker-Smith, Andrea Z Beaton, Amelia K Boehme, Alfred E Buxton, et al. Heart disease and stroke statistics—2023 update: a report from the american heart association. *Circulation*, 147(8):e93–e621, 2023.

[TAC+19] Shawn Tan, Guillaume Androz, Ahmad Chamseddine, Pierre Fecteau, Aaron Courville, Yoshua Bengio, and Joseph Paul Cohen. Icentia11k: An

unsupervised representation learning dataset for arrhythmia subtype discovery. *arXiv preprint arXiv:1910.09570*, 2019.

[TAK+23] Nukala Bhanu Teja, Hridima K Ajay, Rudrakshi Sai Kumar, S Deepa, J Jayapriya, and M Vinay. Deep learning for arrhythmia classification: A comparative study on different deep learning models. In *2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE)*, pages 01–06. IEEE, 2023.

[TDZ+24] Jing Ru Teoh, Jian Dong, Xiaowei Zuo, Khin Wee Lai, Khairunnisa Hasikin, and Xiang Wu. Advancing healthcare through multimodal data fusion: a comprehensive review of techniques and applications. *PeerJ Computer Science*, 10:e2298, 2024.

[VPP14] Aarohi Vora, Chirag N Paunwala, and Mita Paunwala. Improved weight assignment approach for multimodal fusion. In *2014 International conference on circuits, systems, communication and information technology applications (CSCITA)*, pages 70–74. IEEE, 2014.

[WC21] Kuba Weimann and Tim OF Conrad. Transfer learning for ecg classification. *Scientific reports*, 11(1):5251, 2021.

[WLF+21] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling, and Ruigang Yang. Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3239–3259, 2021.

[WPL+22] Huiyi Wu, Kiran Haresh Kumar Patel, Xinyang Li, Bowen Zhang, Christoforos Galazis, Nikesh Bajaj, Arunashis Sau, Xili Shi, Lin Sun, Yanda Tao, et al. A fully-automated paper ecg digitisation algorithm using deep learning. *Scientific Reports*, 12(1):20963, 2022.

[WSN23] K Wisaeng and W Sa-Ngiamvibool. Brain tumor segmentation using fuzzy otsu threshold morphological algorithm. *IAENG International Journal of Applied Mathematics*, 53(2):1–12, 2023.

[WYL+23] Yue Wang, Guanci Yang, Shaobo Li, Yang Li, Ling He, and Dan Liu. Arrhythmia classification algorithm based on multi-head self-attention mechanism. *Biomedical Signal Processing and Control*, 79:104206, 2023.

[XWW+24] Guoping Xu, Xiaxia Wang, Xinglong Wu, Xuesong Leng, and Yongchao Xu. Development of skip connection in deep neural networks for computer vision and medical image analysis: A survey. *arXiv preprint arXiv:2405.01725*, 2024.

[YLLK20] Shoulin Yin, Hang Li, Desheng Liu, and Shahid Karim. Active contour modal based on density-oriented birch clustering method for medical image segmentation. *Multimedia Tools and Applications*, 79:31049–31068, 2020.

[YTA+18] Jingting Yao, S Tridandapani, WF Auffermann, CA Wick, and PT Bhatti. An adaptive seismocardiography (scg)-ecg multimodal framework for cardiac gating using artificial neural networks. *IEEE journal of translational engineering in health and medicine*, 6:1–11, 2018.

[YYZM] Jia Yifan, Cui Yangyang, Yadan Zhang, and Xiang Min. Comparative analysis of 1-d and 2-d deep convolutional neural networks in magnetocardiography classification for coronary artery disease.

[YZC17] Zhenjie Yao, Zhiyong Zhu, and Yixin Chen. Atrial fibrillation detection by multi-scale convolutional neural networks. In *2017 20th international conference on information fusion (Fusion)*, pages 1–6. IEEE, 2017.

[Zea20] Jinglong Zhao and et al. Deep learning algorithms for automated atrial fibrillation detection using ecg signals. *Journal of Clinical Medicine*, 9:782, 2020.

[ZF24] Feiyan Zhou and Duanshu Fang. Multimodal ecg heartbeat classification method based on a convolutional neural network embedded with fca. *Scientific Reports*, 14(1):8804, 2024.

[ZHX+21] Yueying Zhou, Shuo Huang, Ziming Xu, Pengpai Wang, Xia Wu, and Daoqiang Zhang. Cognitive workload recognition using eeg signals and ma-

chine learning: A review. *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

[ZMS+21] Peng Zhang, Chenbin Ma, Yangyang Sun, Guangda Fan, Fan Song, Youdan Feng, and Guanglei Zhang. Global hybrid multi-scale convolutional network for accurate and robust detection of atrial fibrillation using single-lead ecg recordings. *Computers in Biology and Medicine*, 139:104880, 2021.

[ZPT17] Martin Zihlmann, Dmytro Perekrestenko, and Michael Tschannen. Convolutional recurrent neural networks for electrocardiogram classification. In *2017 Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2017.

[ZXZ22] Meng Zhou, Xiaolan Xu, and Yuxuan Zhang. An attention-based multi-scale feature learning network for multimodal medical image fusion. *arXiv preprint arXiv:2212.04661*, 2022.

[ZZD+20] Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific data*, 7(1):48, 2020.