INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Atrial Flutter Classification with Active Learning in Imbalanced Learning Context

Ana Mercês Soares dos Reis Moreira

Master in Computer Engineering

Supervisor:
PhD Ana Maria de Almeida, Associate Professor with Aggregation,
Iscte – Instituto Universitário de Lisboa

Supervisor:
PhD Luís Miguel Martins Nunes, Assistant Professor,
Iscte – Instituto Universitário de Lisboa

October, 2025

[This page is intentionally left blank.]

TECHNOLOGY
AND ARCHITECTURE

Department of Information Science and Technology

Atrial Flutter Classification with Active Learning in Imbalanced Learning Context

Ana Mercês Soares dos Reis Moreira

Master in Computer Engineering

Supervisor:
PhD Ana Maria de Almeida, Associate Professor with Aggregation,
Iscte – Instituto Universitário de Lisboa

Supervisor:
PhD Luís Miguel Martins Nunes, Assistant Professor,
Iscte – Instituto Universitário de Lisboa

October, 2025

[This page is intentionally left blank.]

To the reader

[This page is intentionally left blank.]

Acknowledgment

I would like to express my gratitude to my two supervisors, Ana Maria de Almeida and Luís Nunes, for their continuous guidance and valuable insights throughout this work. Their expertise and support were fundamental to the completion of this dissertation.

I am also greatly thankful to the cardiology specialist Dr. Luís Rosário, for his assistance at the early stages of this research, particularly in clarifying medical concepts and explaining how to interpret ECG related-data. I extend my thanks to Fundação para a Ciência e a Tecnologia, I.P. (FCT), whose partial support through the ISTAR Projects UIDB/04466/2025 and UIDP/04466/2025 was appreciated and highly valued.

Finally, I wish to thank my parents, for their constant care and concern, and to my sister, Raquel, whose encouragement continually inspires me to do better.

[This page is intentionally left blank.]

Resumo

Os dados desequilibrados continuam a ser um dos principais desafios em aprendizagem de máquina, sobretudo em contextos médicos, onde a classe minoritária está frequentemente sub-representada. Na classificação de arritmias por eletrocardiogramas, a deteção de distúrbios raros como o *Flutter Atrial* é dificultada pelo forte desequilíbrio entre classes. Esta dissertação avalia a eficácia de combinar algoritmos de Aprendizagem Ativa com geração de dados sintéticos para melhorar a deteção de *Flutter Atrial* sob condições de desequilíbrio.

O conjunto de dados foi analisado para compreender o papel de cada variável na representação do sinal de ECG. Os dois algoritmos de Aprendizagem Ativa - *Certainty-Based Active Learning* (CBAL) e *Fisher Matrix Information Pairwise Selection* (FIMPS) - foram implementados com um classificador XGBoost. O desequilíbrio foi atenuado através da geração de amostras sintéticas da classe minoritária com uma *Generative Adversarial Network* (GAN). Ambos os cenários, com e sem dados gerados pela GAN, foram avaliados por métricas padrão: *accuracy*, *precision*, *recall*, *F1-Score* e *ROC-AUC.*

Os resultados mostram que, embora métodos tradicionais de reamostragem como *Oversampling*, *Undersampling* e *SMOTE* apresentem desempenho ligeiramente superior, a inclusão de dados gerados pela GAN aumentou a sensibilidade dos algoritmos de Aprendizagem Ativa, sobretudo para a classe minoritária. O FIMPS superou o CBAL em *recall* e robustez, sendo mais adequado a dados ruidosos e sobrepostos como os sinais de ECG.

Este estudo demonstra o potencial de combinar Aprendizagem Ativa e geração sintética para mitigar o desequilíbrio em dados médicos, salientando a importância de uma extração criteriosa de variáveis e definição de classes.

Palavras Chave: *Aprendizagem com Desequilíbrio*; *Aprendizagem Ativa*; *Flutter Atrial*; *Classificação*; *Eletrocardiograma*; *Generative Adversarial Networks*

[This page is intentionally left blank.]

Abstract

Imbalanced data remains one of the major challenges in machine learning, particularly in medical domains where the minority class is often underrepresented in critical contexts. In electrocardiogram-based arrhythmia classification, the detection of rare cardiac disorders such as Atrial Flutter is frequently constrained by a strong class imbalance. This dissertation investigates the effectiveness of combining Active Learning (AL) algorithms with synthetic data generation to enhance Atrial Flutter detection under imbalanced conditions.

The dataset underwent analysis and interpretation to understand the role of each feature in representing the ECG signal. Two Active Learning algorithms - Certainty-Based Active Learning (CBAL) and Fisher Matrix Information Pairwise Selection (FIMPS) - were implemented using an XGBoost classifier. The imbalance in the dataset was mitigated through the generation of synthetic minority samples using a Generative Adversarial Network (Generative Adversarial Network (GAN)). Both scenarios-with and without GAN-generated data-were evaluated using standard performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

The results indicate that although traditional resampling techniques such as Oversampling, Undersampling, and SMOTE achieved slightly higher overall performance, the integration of GAN-based data improved the sensitivity of Active Learning algorithms, particularly for the minority class. The FIMPS algorithm consistently outperformed CBAL in recall and overall robustness, suggesting that it is better suited for highly overlapping and noisy data distributions such as Electrocardiogram (ECG) signals.

This study demonstrates the potential of combining Active Learning with synthetic data generation to address severe imbalance in medical datasets, while emphasizing the importance of careful feature extraction and class definition to enhance future clinical applicability.

KEYWORDS: *Imbalanced Learning*; *Active Learning*; *Atrial Flutter*; *Classification*; *Electrocardiogram*; *Generative Adversarial Networks*

[This page is intentionally left blank.]

Contents

Acknowledgment iii

Resumo v

Abstract vii

List of Figures xi

List of Tables xiii

List of Acronyms xv

Chapter 1. Introduction 1
1.1. Motivation 1
1.2. Research Questions 2
1.3. Methodology 3
1.4. Document Structure 4

Chapter 2. Context 7
2.1. Electrocardiogram 8
2.1.1. ECG's Leads 8
2.1.2. Electrocardiogram Interpretation 9
2.2. Atrial Flutter 10
2.3. Active Learning Strategies 12
2.3.1. Certainty-Based Active Learning 13
2.3.2. FIM-Based Pairwise Selection 14

Chapter 3. Literature Review 15
3.1. Research Methodology 15
3.2. PRISMA 2020 Flow Diagram's Results 16
3.3. Background Concepts 19
3.4. Previous Work on Imbalanced and Active Learning Techniques 22
3.5. Previous Work on Imbalanced Learning in ECG-based Arrhythmia Detection 26
3.6. Conclusion 29

Chapter 4. Methodology 31
4.1. Dataset 31
4.2. Class Distribution 32

4.3. Preprocessing Data 35
4.3.1. ECG Signal Noise Cleaning 35
4.3.2. Feature Extraction 36
4.3.3. Data Preparation 38
4.4. Experiences 40
4.4.1. Dataset split & Classifier 40
4.4.2. Techniques & Performance Metrics 41
4.4.3. Generative Adversarial Network 42

Chapter 5. Results 45
5.1. Undersampling 45
5.2. Oversampling 48
5.3. SMOTE-Tomek 52
5.4. Isolation Forest 56
5.5. CBAL 59
5.6. FIMPS 62

Chapter 6. Conclusion 65

Bibliography 71

List of Figures

Figure 2.1 Normal ECG waveform 9
Figure 2.2 Atrial Flutter Waveform 12

Figure 3.1 PRISMA 2020 Flow Diagram Results 17
Figure 3.2 Distribution of documents per publication year 18
Figure 3.3 Distribution of documents per type 18

Figure 4.1 Class Distribution 33
Figure 4.2 Sex Distribution on Atrial Flutter 34
Figure 4.3 Age Distribution on Atrial Flutter 35
Figure 4.4 Low-pass Filter on ECG 36
Figure 4.5 High-pass Filter on ECG 36
Figure 4.6 Band-pass Filter on ECG 37
Figure 4.7 Leads II, III, aVF and V1 of Atrial Flutter Record 38
Figure 4.8 Atrial Flutter Record's ECG Lead II 39
Figure 4.9 Histograms 40
Figure 4.10 Histograms per Class 41
Figure 4.11 Functional Diagram of a GAN model 43

Figure 5.1 Scenario 1 - Undersampling Confusion Matrix 45
Figure 5.2 Scenario 1 - F1-Score in Undersampling 46
Figure 5.3 Scenario 2 - Confusion Matrix of Undersampling 47
Figure 5.4 Scenario 2 - F1-Score in Undersampling 48
Figure 5.5 Scenario 1 - Oversampling Confusion Matrix 49
Figure 5.6 Scenario 1 - F1-Score in Oversampling 50
Figure 5.7 Scenario 2 - Confusion Matrix of Oversampling 51
Figure 5.8 Scenario 2 - F1-Score in Oversampling 52
Figure 5.9 Scenario 1 - SMOTE-Tomek Confusion Matrix 53
Figure 5.10 Scenario 1 - F1-Score in SMOTE-Tomek 54
Figure 5.11 Scenario 2 - Confusion Matrix of SMOTE-Tomek 55
Figure 5.12 Scenario 2 - F1-Score in SMOTE-Tomek 56

Figure 5.13 Scenario 1 - Isolation Forest Confusion Matrix 56
Figure 5.14 Scenario 1 - Isolation Forest Performance Metrics 57
Figure 5.15 Scenario 2 - Confusion Matrix of Isolation Forest 58
Figure 5.16 Scenario 2 - Performance Metrics of Isolation Forest 59
Figure 5.17 Scenario 1 - CBAL Confusion Matrix 60
Figure 5.18 Scenario 1 - CBAL Performance Metrics 60
Figure 5.19 Scenario 2 - Confusion Matrix of CBAL 61
Figure 5.20 Scenario 2 - Performance Metrics of CBAL 61
Figure 5.21 Scenario 1 - FIMPS Confusion Matrix 62
Figure 5.22 Scenario 1 - FIMPS Performance Metrics 63
Figure 5.23 Scenario 2 - Confusion Matrix of FIMPS 64
Figure 5.24 Scenario 2 - Performance Metrics of FIMPS 64

List of Tables

Table 4.1	Class Distribution by Sex	34
Table 4.2	Class Distribution by Sex and Age	35
Table 4.3	Class Distribution With Missing Values	39
Table 4.4	Data split summary - Original Data	41
Table 4.5	Data split summary - Data with GAN samples	44
Table 5.1	Scenario 1 - Undersampling Performance Metrics per Ratio	46
Table 5.2	Scenario 2 - Undersampling Performance Metrics per Ratio	47
Table 5.3	Scenario 1 - Oversampling Performance Metrics per Ratio	49
Table 5.4	Scenario 2 - Oversampling Performance Metrics per Ratio	51
Table 5.5	Scenario 1 - SMOTE-Tomek Performance Metrics per Ratio	53
Table 5.6	Scenario 2 - SMOTE-Tomek Performance Metrics per Ratio	55

[This page is intentionally left blank.]

List of Acronyms

AL: Active Learning

AFL: Atrial Flutter

CBAL: Certainty-Based Active Learning

DSRM: Design Science Research Methodology

ECG: Electrocardiogram

FIMPS: Fisher Matrix Information Pairwise Selection

GAN: Generative Adversarial Network

HRV: Heart Rate Variability

IL: Imbalanced Learning

kNN: K-Nearest Neighbors

ML: Machine Learning

QBC: Query-By-Committee

SLR: Systematic Literature Review

SMOTE: Synthetic Minority Oversampling Technique

SNOMED CT: Systematized Nomenclature of Medicine - Clinical Terms

SVM: Support Vector Machine

[This page is intentionally left blank.]

CHAPTER 1

Introduction

Although Machine Learning (ML) has significantly advanced science, technology and transforming the world daily, it also comes with inherent limitations. One of the key challenges when working with data is understanding its distribution and the factors influencing model outcomes. The way data is structured directly impacts how algorithms learn and generalize, making data representativeness a critical aspect of machine learning models [1]. In real-world scenarios, data is rarely perfectly balanced where some classes are overrepresented, while others are significantly diminished.

Data imbalance has become a major concern in ML and data science, particularly in classification problems where traditional algorithms tend to favor majority classes, often at the expense of minority ones. While these underrepresented classes may contain fewer instances, they often hold critical information that should not be ignored. In some cases, the very existence of a minority class reflects an important pattern in the dataset, providing essential insights into rare but significant events [2].

The issue of imbalanced data has gained increasing attention due to its severe impact on both research and real-world applications [3]. When models fail to properly account for minority classes, their predictions become unreliable and biased, leading to serious consequences in high-stakes fields. For instance, in medical diagnosis, an imbalanced dataset could cause a model to overlook positive cancer cases, simply because negative test results dominate the data [1]. Similarly, in fraud detection, an algorithm trained on an imbalanced dataset might systematically ignore fraudulent transactions, increasing security vulnerabilities.

These failures demonstrate the urgent need for improved methodologies that can effectively handle data imbalance and ensure fair and accurate model predictions.

1.1. Motivation

The challenge of balancing class distributions in classification problems remains an open research issue. Even though numerous algorithms have been developed to address data imbalance, having multiple solutions to a single problem often leads to better performance and flexibility [3]. This dissertation explores not only traditional resampling methods, widely applied in Imbalanced

Learning (IL) contexts, but also advanced Active Learning (AL) with high potential in imbalanced contexts. The dissertation focuses on an Electrocardiogram (ECG) based dataset to evaluate the ability of these algorithms to detect and classify the rarely mentioned Atrial Flutter (AFL) heart condition from waveform-derived features. Moreover, it aims to determine their comparative effectiveness in handling severe class imbalance.

The ultimate goal of this dissertation is to enhance adaptability, fairness, and accuracy in IL scenarios, ensuring that critical minority class information is not overlooked. By assessing alternative techniques and their impact on data imbalance, this dissertation contributes to the broader goal of developing more reliable, generalizable and ethically sound machine learning models.

1.2. Research Questions

Overall, the dissertation aims to explore the intersection between the challenges of Imbalanced Learning and the application of AL techniques, with the goal of addressing the issues posed by skewed data distributions in classification tasks. Moreover, during the classification process, the variables also raise questions regarding their impact in identifying successfully minority class instances. Therefore, this section presents two groups of research questions - given their nature and areas of interest - to guide the study and achieve the objectives.

The first group covers questions related to “To what extent do techniques help with class imbalance?”, i.e, which algorithms, including AL, prove to be a solution to the problem of class representation?

Having set the tone, the research questions for Group 1 are as follows:

RQ1. "How effective are the Certainty-Based Active Learning (CBAL) and Fisher Matrix Information Pairwise Selection (FIMPS) Active Learning algorithms in improving classifier performance on severely imbalanced ECG datasets, compared to traditional sampling?”

This question aims to ascertain the contributions of AL application before the classification of Atrial Flutter. It compares the outcomes with the traditional sampling methods and focuses on quantifying the extent to which Active Learning can reduce or not the impact of severe class imbalance.

RQ2. "How does the application of a Generative Adversarial Network (GAN) in the minority class impact on the performance of classification models and how does it influence Active Learning methods?"

This question addresses the impact of using a GAN to generate synthetic data for the minority class in imbalanced datasets. It explores how this approach affects the overall performance of classification models and analyses the influence it has with AL strategies.

Furthermore, the role of ECG features in designing the classification pipeline is also highly relevant, as they provide the essential information that enables the model to recognize and differentiate minority class instances. Hence, Group 2 unfolds in the following research questions:

RQ3. "Does the integration of ECG signal analysis (data understanding) lead to a measurable improvement in AFL detection in the context of class imbalance?"

This question investigates the importance of data preparation stages prior to the modeling phase. It emphasizes how feature selection can influence model performance, especially in imbalanced classification scenarios. Additionally, it highlights the potential negative impact of including multiple features without clear criteria, which may introduce noise and reduce the model's effectiveness.

RQ4. "What are the distinguishing ECG signal characteristics to identify AFL, and how can these features be exploited to enhance model performance in highly imbalanced classification scenarios?"

The final question focuses on the iterative process of analyzing, selecting and testing several times the most relevant ECG signal features for accurately detecting AFL. Moreover, it highlights the selected features to include in the dataset, having always in mind that the essential signal's information is covered but without exaggerating the number of variables to include.

1.3. Methodology

Considering the nature of this dissertation, the Design Science Research Methodology (DSRM) was selected as the methodological approach to define the guidelines and development phases for this study. The choice was influenced by the work of Venable et al. (2017) who presented the Information Systems Development Methodology Framework created by Avison and Fitzgerald (2006) and adapted to the specific needs of DSRM. Authors tailored this framework to align with the characteristics of Design Science Research and analyzed various DSRM methodologies using the seven elements of the Comparison Framework to help researchers choose the best methodology according to the characteristics of their research.

The results of each DSRM were compiled into a table that facilitated comparison, proving to be instrumental in selecting the best methodology. Ultimately,

it was concluded that DSRM is the most suitable one. It consists in a well-structured and iterative methodology specifically designed for research focused on scientific objectives, making it particularly appropriate for this case.

The DSRM is a six-step process including: Identify Problem & Motivation, Define Objectives of a Solution, Design & Development, Demonstration, Evaluation, and Communication. As mentioned earlier, DSRM is a repetitive framework, allowing researchers to revisit and modify earlier phases during the research process when necessary. Such revisits may occur during the Evaluation or Communication steps, leading back to the Define Objectives or Design & Development stages. This may happen for several reasons, including a change in the direction of research, results that do not align with expectations or other adjustments that seem fit. Besides, DSRM defines the objectives of a solution, evaluates and extensively adapts it to a daily use unlike Systems Development Research Methodology and Design Science Research Process Model which do not consider the solution's longevity and applicability [4].

Furthermore, DSRM differs from traditional research methodologies by placing strong emphasis on Evaluation. Instead of proposing a theoretical model, DSRM requires researchers to demonstrate and measure how well the artifact solves the identified problem. This aligns with the goals of this dissertation, which involve testing and validating alternative approaches to address data imbalance in machine learning.

By adopting the DSRM framework, the dissertation guarantees a structure, iterative and outcome-oriented research process leading to an academically rigorous and practically relevant artifact.

1.4. Document Structure

This document is composed of six chapters: Introduction, Context, Systematic Literature Review, Methodology, Results and Conclusion. The structure adopted so far will be applied to the rest of the document, where the chapter is subdivided into several sections to address related topics and keep the content well organized and easy to find.

Chapter 2 presents the technical concepts related to ECG, its interpretation, Atrial Flutter (AFL) and the Active Learning algorithms to be employed in the experiments.

Then Chapter 3 provides information regarding the Literature Review, the research methodology and the framework used to select literature - PRISMA 2020 statement. Moreover, the background concepts related to the Imbalance Learning Problem and Active Learning for Imbalanced Data are presented in Section 3.3. Section 3.4 is focused on the related work; state-of-the-art is analyzed and some of theAL techniques encountered presented.

Chapter 4 presents the procedures involved in the Exploratory Data Analysis finishing with a section of contextualization of the Experiments phase that will follow in the Results Chapter. Section 4.4 will address the model's design and employed configurations. Chapter 5 presents the outcomes from the experiences made and the document will finish with the Conclusions in Chapter 6 where the research questions will be answered and the Future Work will be displayed.

[This page is intentionally left blank.]

CHAPTER 2

Context

Considering the purpose of the present research, the first step in the process is to identify a suitable dataset. *PhysioNet* emerges as an interesting and promising source of information for exploring the challenges of imbalanced data. As the website specializes in healthcare and clinical data, these fields often involve datasets with sparse results and limited representativeness. In fact, literature consistently highlights that imbalanced data is a frequent issue in medical records, whether in medical tests or final diagnoses [3]. Therefore, choosing a dataset of this nature seems both accurate and fitting for the study.

PhysioNet houses an extensive medical database, requiring careful investigation and thoughtful consideration when selecting the right dataset. However, after a thorough evaluation, the "A large scale 12-lead electrocardiogram database for arrhythmia study" [5] dataset conveys to be a flexible choice for further analysis and exploration.

Given the limited background in healthcare and ECG-related topics, the involvement of a medical specialist became essential to ensure the technical soundness of the study. To that end, Luís Rosário, M.D., a renowned cardiologist, joined the research team, bringing valuable expertise and clinical insights. His contribution not only lends credibility but also confidence to the initial stages of the project where most decisions regarding the data assessment need to be employed.

Following a comprehensive briefing of the research objectives, Dr. Rosário recommended focusing on a specific and relatively uncommon heart condition to define as the minority class: AFL, a supraventricular tachycardia. Thereby, the study revolves around Atrial Flutter - a severe arrhythmia that, while less prevalent than Atrial Fibrillation, is equally impactful and alarming.

Besides, Dr. Rosário provided a detailed explanation of the electrocardiogram exam and principal components. He outlined the various waves that compose the ECG signal, as well as the segments and intervals that are particularly relevant during data analysis of the recordings and how to interpret them. Furthermore, with a stronger emphasis on Atrial Flutter, he appointed the distinct characteristics of the arrhythmia in case and the essential aspects to observe when analyzing the corresponding ECGs. Concepts such as the

sawtooth pattern and cardiac conduction were introduced as well and referenced as an important part of the AFL condition full comprehension, which will be further addressed later in the chapter.

In conclusion, this chapter will introduce the reader to key concepts in order to contextualize the data. Structured into three sections, the chapter begins explaining what an electrocardiogram exam entails and the principles behind its interpretation. Afterwards, the focus shifts to the heart condition AFL, outlining its clinical relevance, impact on health and how it can be identified in an ECG recording exam – with particular emphasis on its leads. Finally, the Active Learning algorithms to be implemented and tested in the experiments' phase will be presented.

2.1. Electrocardiogram

The electrocardiogram, commonly referred to as an ECG by the medical community, is one of the most frequently used non-invasive diagnostic tests for assessing cardiac conditions. It records the heart's electrical activity painlessly over a short period of time, providing useful observation into the heart's rhythm and overall condition. During the procedure, each contraction of the heart muscle is triggered by small electric impulses generated by the heart itself which are detected by electrodes placed on the patient's body. The patient lies down with 10 adhesive electrodes applied in different parts of the body – six on the thorax and other four on the limbs (normally wrists and ankles). These electrodes are connected to an electrocardiograph which captures the heart's electric signals and prints them, producing the ECG trace [6] [7].

The ECG is notably effective in evaluating cardiac arrhythmias, but it also plays a role in identifying other sorts of cardiovascular abnormalities such as cardiomyopathy and pericarditis, proving to be a useful tool for both diagnosing and screening in clinical practice.

2.1.1. ECG's Leads

Although the ECG uses 10 conductive electrodes, it records 12 leads, resulting in 12 distinct waveforms on the ECG paper. While each electrode detects electrical activity, the combination of signals from specific electrode pairs allows for the calculation of the 12 leads. These are known as leads I, II, III, aVR, aVL and aVF and six precordial leads (V1-V6). The last group of leads, positioned on the chest, provide views of the septal, anterior and lateral areas of the heart. Meanwhile, the standard limb leads (I, II, and III) and the augmented limb leads (aVR, aVL, aVF) capture the heart's electrical activity between the limbs [8]. Together, these sensors offer multiple perspectives of one heart event, enhancing diagnostic accuracy.

2.1.2. Electrocardiogram Interpretation

Each leas is represented as a separate line on the ECG trace, reflecting different anatomical perspectives of the heart. Within each lead, the electrical signal is expressed through a series of waves – notably the P, Q, R, S and T waves. Occasionally, the U wave may appear, although it is less frequently observed. These waveforms take a determining role into the functional state of the heart and serve as the foundation for the diagnosing phase [8]. Furthermore, the ECG waveform is composed of segments and intervals, each corresponding to an explicit electrical event within the cardiac cycle.

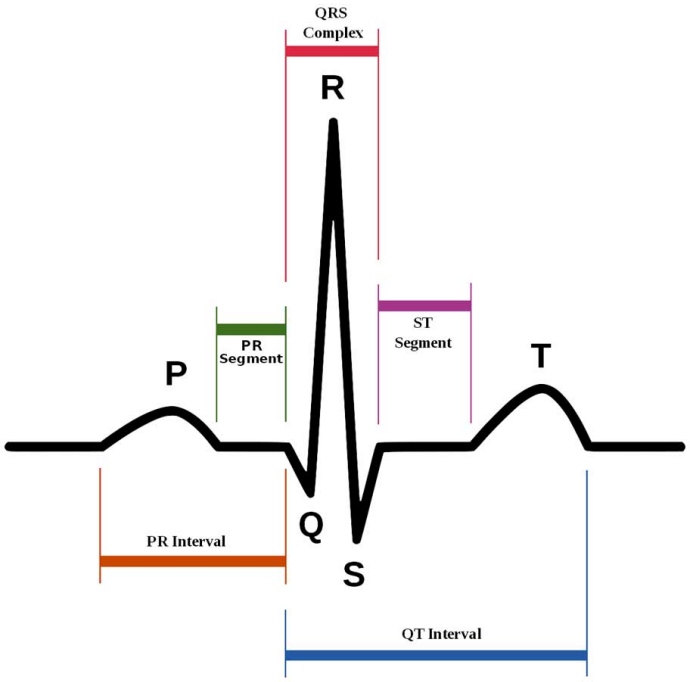


FIGURE 2.1. Normal ECG waveform

Figure 2.1 demonstrates an ECG waveform of a normal cardiac event, starting with atrial depolarization - the contraction of the atria (upper chambers of the heart). This movement is represented by the P-wave, a small and rounded deflection on the trace. The appearance of the P-wave marks the beginning of the heartbeat under normal physiological conditions, commonly called by the medical community as normal sinus rhythm.

Meanwhile, the PR interval has already started, as this interval represents the time from the onset of the P-wave to the start of the QRS complex. It reflects the time the electrical impulse takes to travel from the atria to the ventricles (lower chambers of the heart), as well including the conduction through the atrioventricular (AV). Normally, the PR interval ranges in duration between 0.12 and 0.20 ms.

The heartbeat follows with the ventricular depolarization, represented on the ECG by the QRS complex. It corresponds to the contraction of the ventricles and as the name implies, the complex begins with the Q-wave - a small negative deflection. Followed by a sharp upward positive deflection known as the R-wave, it finishes with the S-wave, another negative deflection - typically more pronounced than the Q-wave [9].

Usually, the QRS complex lasts between 0.06 and 0.10 seconds and its amplitude falls within the range of 1 to 2 mV, although this can vary depending on the heart's size and condition.

After the QRS complex, the ST segment begins. This segment represents the ventricular repolarization rest period – the phase when the ventricles recover after contraction. A normal ST segment should appear as a flat and horizontal line without significant elevation or depression.

Finally, it comes the T wave reflecting ventricular repolarization and indicating that the heart muscle is preparing for the next contraction (heartbeat). The T-wave generally has an amplitude between 0.1 and 0.5 mV, though it may vary between individuals. In terms of duration, it is commonly longer than both the P-wave and the QRS complex [8].

As an additional note, it is relevant to highlight that the direction of the waveforms deflections on the ECG is directly related to the orientation of the electrical activity in each lead. When the electrical impulse travels towards a certain lead, it produces a positive deflection, whereas when it moves away, a negative deflection is drawn. This principle helps explain the morphology and direction of the peaks seen on the ECG trace (mainly between II, III and V1 leads) and indicates the underlying electrical behavior of the heart if an abnormality happens to occur.

2.2. Atrial Flutter

Although Atrial Fibrillation is the most common atrial arrhythmia worldwide, Atrial Flutter (AFL) remains the second most prevalent condition within the same category. Despite being considerably less frequent, its clinical significance should not be underestimated. Unlike Atrial Fibrillation, which affects millions globally, Atrial Flutter occurs under more specific conditions of the heart, making it rarer in the general population.

In [10] a study is developed to assess the incidence of Atrial Flutter. The study comprised a population of 100 000 individuals, and in total, 181 Atrial Flutter diagnoses were encountered. On another attempt, occurrence dropped significantly in younger populations where, among 100 000 people under the age of 50, only five cases were identified with the AFL condition. If only considering people over 80 years old, the incidence of AFL would rise substantially to 587 cases, in a world of 100 000 people. These statistics suggest a strong

correlation between the heart condition and advancing age and predominantly in male patients. This effect will be verified once again in Section 4.2 , when analyzing the employed dataset. Clinical records further reveal that 16% of the diagnosed individuals also present heart failure, 12% have chronic obstructive pulmonary disease and only 1.7% of the patients do not show predisposing risk factors associated.

AFL represents a deviation from the heart's natural rhythm, which under healthy conditions is regulated by a highly coordinated electrical conduction system. In fact, the system coordinates the contractions of the upper chambers (atria) and the lower chambers (ventricles). In the case of Atrial Flutter, however, the atria behave in a different manner, and a new circular faulty pathway assumes the cycle, often due to damaged or scarred heart tissue. The abnormal circuit causes the atria to contract at an accelerated and regular pace – typically 250 and 350 beats per minute (bpm) [11].

As a result, the atrial contractions outnumber the ventricular and the atrioventricular conduction ratio can vary (e.g., 2:1, 3:1, or higher), depending on how many atrial impulses are successfully transmitted through the AV node. Typically, the ventricular rate is a fraction of the atrial rate. For instance, in a case of 2:1 conduction, meaning the ventricles beat at half the frequency of the atria, if the atria rate is at 300 bpm, the ventricular frequency will be 150 bpm.

Some individuals with diagnosed AFL, may face symptoms such as palpitations, shortness of breath, fatigue and chest tightness. As for the causes, factors like coronary artery disease, high blood pressure and heart valve disorders can increase the risk of developing AFL. Age is also a determinant factor - the older the patient is, the higher the likelihood of developing cardiovascular diseases. Both the symptoms and causes of Atrial Flutter are similar to Atrial Fibrillation, so the most reliable tool to detect which condition the patient is dealing with, is by performing an ECG exam.

ECG's Interpretation of Atrial Flutter case

When reading an ECG recording, the following visual indicators are usually associated with Atrial Flutter and can be easily identified within few seconds of analysis:

- Absence of the P-wave, which normally marks the beginning of the cardiac cycle;
- Presence of a *sawtooth* pattern preceding the QRS complex;
- An elevated heart rate, which can be estimated from the ECG grid.

As Figure 2.2 suggests, both first and second criteria are met [11].



FIGURE 2.2. Atrial Flutter Waveform

However, in this study, since the ECGs are stored and displayed digitally, traditional paper-based visualization is impractical. As an alternative, the *Neurokit2* Python library - specifically designed for the interpretation of physiological data - is employed to process the signals. Consequently, the third diagnostic criterion mentioned above (heart rate estimation based on ECG grid paper) is not directly applicable. Nevertheless, the heart rate can be assessed by calculating the RR intervals extracted from the digital ECG signals.

In cases where no distinct P-waves are visible to the naked eye, their mean amplitude is computed to confirm the wave's absence. Usually, in Atrial Flutter, the P-wave is not detected - presents a negative amplitude - or is close to zero. PQ interval is of high importance as well, as it reflects the time taken for the electrical impulse to travel from the atria to the ventricles.

2.3. Active Learning Strategies

As outlined in Chapter 1 and formulated through the research questions, this dissertation seeks to explore strategies capable of improving classification performance in highly imbalanced datasets. Recent studies, highlight Active Learning as a promising paradigm to address class imbalance more effectively.

Instead of just relying on standard training procedures, this work investigates the impact of Active Learning as a key pillar of the experimental setup Literature suggests that AL methods can significantly improve learning efficiency by focusing on the most informative samples, which is particularly beneficial in cases where minority class instances are rare. For this reason, two Active learning algorithms were selected for implementation, both specifically designed to alleviate the challenges posed by imbalanced data.

So, this section is structured in two parts, each introducing one AL strategy. CBAL and FIMPS focus on guiding the training and learning process toward informative and diverse samples from the minority class. The ultimate objective is to enhance the classifier's ability to detect Atrial Flutter accurately, despite the severe disparity in class distribution.

2.3.1. Certainty-Based Active Learning

Fu and Lee (2013) in [12] propose the CBAL algorithm to address the limitations of traditional active learning in scenarios involving highly imbalanced datasets. Usually, conventional active learning methods often rely on global decision boundaries and tend to query samples from the majority class, as these dominate the input space. This results in biased classifiers that perform poorly on the minority class. CBAL tries to mitigates this issue by emphasizing local decision behaviors, allowing it to focus on informative regions in the data that would otherwise be overshadowed by majority instances.

The CBAL is designed to identify informative unlabeled samples for querying by evaluating the importance of each sample within an explored neighborhood. The construction of this neighborhood begins by identifying the *m* nearest labeled samples using the K-Nearest Neighbors (kNN) technique. Instead of defining the neighborhood size in advance, CBAL incrementally expands it using an exponential growth strategy combined with a binary search to efficiently determine the optimal local region in terms of classification certainty. Within each explored neighborhood, the query probability of an unlabeled sample is computed based on the difference between two hypotheses: the current best hypothesis - h_k - and an alternative hypothesis - h'_k that disagrees with h_k on the prediction of that sample. This difference, expressed in the equation quantifies the level of certainty within the neighborhood.

$$G_k = \text{err}(h'_k, S) - \text{err}(h_k, S) \tag{2.1}$$

The larger G_k, the more confident the classifier is in that region. Consequently, a smaller G_k implies higher uncertainty, increasing the likelihood that the corresponding sample will be queried.

Moreover, to guide the exploration process, CBAL adopts two strategies based on a one-step look-ahead (OLA) policy:

- Certainty-Enhanced: When a sample is estimated to lie in a "certain" region, CBAL seeks to maximize G_k, thereby assigning a lower query probability since the classifier is already confident in this area.
- Uncertainty-Enhanced: For samples in "uncertain" regions, the algorithm minimizes G_k, resulting in a higher query probability that increases the chance of querying these samples for labeling.

By focusing on certainty-based local exploration, CBAL prevents majority samples from dominating the querying process and effectively discovers informative minority samples near the decision boundary. The algorithm further integrates the theoretical framework of IWAL-ERM (Importance Weighted Active Learning with Empirical Risk Minimization), ensuring a polynomial improvement in label complexity over passive learning.

In practice, CBAL has demonstrated high macro-average accuracy across multiple real-world and synthetic datasets, indicating balanced performance between majority and minority classes. Additionally, it minimizes the risk of overfitting during the querying process and achieves efficient neighborhood exploration, typically requiring only a few iterations. These properties make CBAL a practical and scalable solution for real-world active learning applications in contexts of class imbalance, offering a strong balance between theoretical soundness and empirical effectiveness.

2.3.2. FIM-Based Pairwise Selection

FIMPS, presented in [13] was created to overcome the ineffectiveness of traditional active learning when dealing with imbalanced datasets. In such scenarios, active learning tends to select redundant majority instances, causing the classifier to become biased and the decision boundary to skew toward the minority class. FIMPS addresses this issue by integrating a Fisher Information Matrix-based uncertainty measure with a pairwise sampling strategy that preserves class balance during the querying process.

The algorithm begins with training an initial classifier using a variant of the Support Vector Machine (SVM) designed for imbalanced data, which is a granular SVM with repetitive undersampling. At each iteration,FIMPS evaluates the informativeness of unlabeled samples based on how much they would increase the model's Fisher Information (i.e., reduce parameter uncertainty), thus identifying the most uncertain and informative instance to query.

To control computational cost, the algorithm applies the Learning from Small Pools technique, selecting candidates from small randomly sampled subsets of the unlabeled pool.

A key innovation of FIMPS lies in its pairwise selection mechanism. Rather than querying a single instance, the algorithm selects two complementary samples: one from the majority class and one from the minority class. This ensures that the labeled dataset remains balanced as learning progresses. If a majority instance is selected according to the Fisher Information Matrix criterion, a compensatory minority instance with the highest certainty is also chosen, and vice versa.

CHAPTER 3

Literature Review

3.1. Research Methodology

A Systematic Literature Review (SLR) is a widely recognized method used by researchers to gather information about recent breakthroughs and studies on a specific topic. This structured process involves selecting, collecting, filtering, and analyzing research documents relevant to the researcher's study. The types of documents reviewed can vary significantly and may include articles, book chapters, surveys, and conference papers.

Systematic Literature Review (SLR) facilitate the search for research documentation by utilizing appropriate search engines, academic publishing websites, and specific criteria, such as keywords, to narrow the search results effectively.

SLR is essential in research as it provides a comprehensive overview of existing work related to a particular topic. This accumulation of knowledge enables other researchers and students to access the latest advancements without having to wade through all the existing literature and repeat the time-consuming selection process. As authors in [14] state, "Fundamentally, knowledge advancement must be built on prior existing work. To push the knowledge frontier, we must know where the frontier is." (p.1) This perspective emphasizes that to advance knowledge, it is crucial to understand what already exists in order to explore new possibilities.

Silva and Neiva (2016) in [15] propose a well-defined SLR guide with specific principles to assist researchers in the review process. Their method includes the following steps:

- Keep in mind the research questions and the goals of the intended study.

- Define a search query that encompasses the concepts the researcher aims to address.

- Search for documentation in notorious search engines, particularly those favored by the scientific community, such as Google Scholar, Scopus and IEEE Xplore.

- Apply inclusion and exclusion criteria when searching the databases, which can include factors such as the year of publication, document type, or relevance based on the number of citations.

- Analyze the resulting literature, starting with the title, abstract, introduction, and conclusion.

By following this model with well-defined criteria and a strict method, researchers are more likely to obtain trustworthy and valuable documents. This approach increases the researcher's confidence in their results, allowing them to effectively incorporate these findings into their future work.

3.2. PRISMA 2020 Flow Diagram's Results

Conducting a state-of-the-art review is a fundamental step in any research process, as it provides a comprehensive understanding of the current advancements, existing techniques, and possible research gaps in the field. The systematic review serves as the foundation for identifying unexplored areas and potential contributions to a determined topic. To achieve it, formulating a well-defined search query that ensures the retrieval of relevant academic literature is indispensable. Moreover, selecting an appropriate research database is essential to access high-quality and reliable information sources [16].

For the systematic review process, supervisors recommended some repositories, in particular, Scopus and Web of Science. However, when applying the query to the databases, it was found that it does not allow searching by 'Title, Abstract and Keywords' simultaneously in Web of Science. This limitation makes the process more time-consuming and repetitive. Thus, Scopus was chosen as the document repository for conducting a literature search on the topic.

To address the problem, the research questions in Section 1.2 were carefully considered, as they serve as the guidelines for conducting the research. Additionally, the areas of application where the techniques were employed, as identified in the resulting documentation, were also taken into account.

After interpreting the problem and establishing objectives for a potential solution, determining the appropriate keywords is vital to formulating a query that will initiate the article selection process. Hence, IL and AL were identified as fundamental keywords for the search query. Given the broad scope of the problem and the relatively recent nature of the topic, the query was deliberately designed to be general and flexible.

As a result, the final search query was formulated as outlined below:

```
"Imbalanced Learning" AND "Active Learning"
```

To maximize the relevance of the results, the search was further refined using the 'search within' option, limiting the query to the 'Article Title, Abstract, and Keywords' fields. The filtering approach ensured the retrieval of the most relevant documentation related to the selected search terms.

Given the relatively recent focus on the topic, it was anticipated that the number of search results would be limited. The phenomenon is characteristic of emerging research areas, which often have fewer published studies but present substantial opportunities for further exploration. Despite these expectations, the query returned a total of 27 documents, as shown in 3.1.

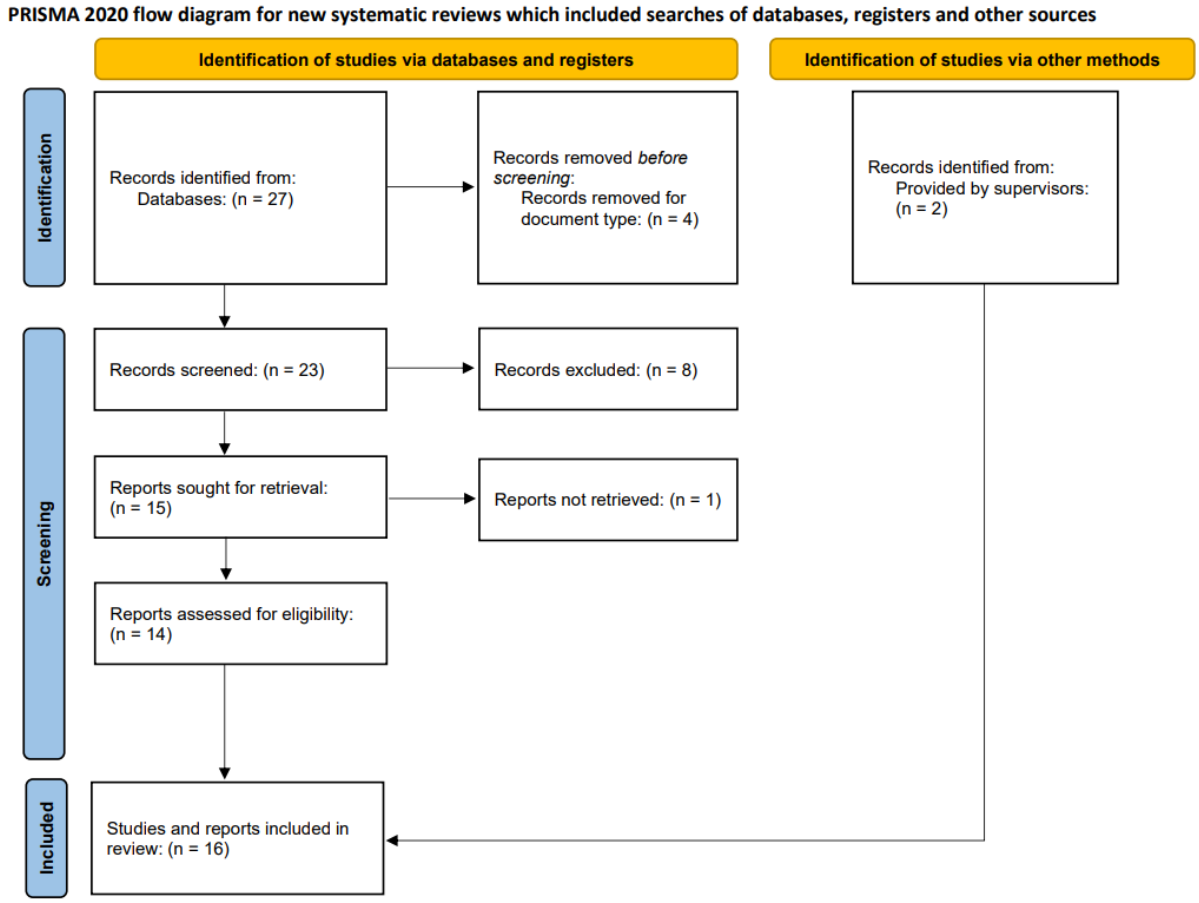


FIGURE 3.1. PRISMA 2020 Flow Diagram Results

After retrieving the query's results, the distribution of the documentation was analyzed based on two key aspects: publication year and document type.

The oldest document dates to 2006, while the majority of the retrieved documents are relatively recent, with a significant increase in publications since 2020. The tendency highlights the growing interest and attention given to the topic in recent years, compared to two decades ago when it was less explored. Despite that, older documents were not excluded, as the initial dataset was already limited in size (see Figure 3.2).

In terms of document type, 13 results were articles, accounting for 48.1% of the total set, followed by eight conference papers, comprising 29.6% (see Figure 3.3). Given the focus of this study on peer-reviewed scientific contributions, only articles, conference papers, book chapters, and short surveys were retained. As a result, the dataset was reduced from 27 to 23 documents, as

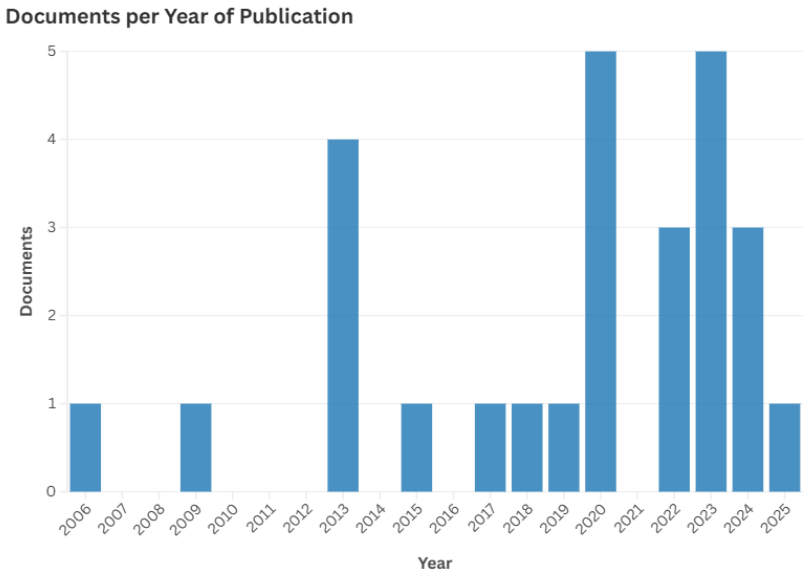


FIGURE 3.2. Distribution of documents per publication year

the flow diagram indicates in the "Records removed before screening" phase (see Figure 3.1).

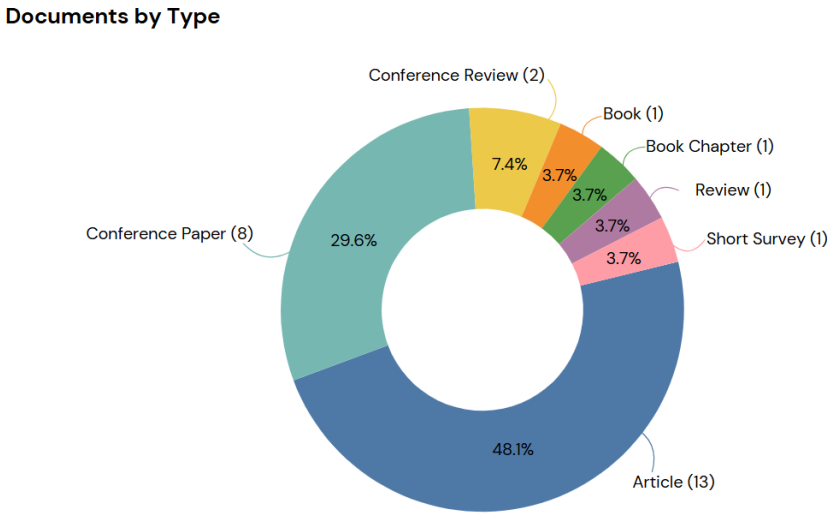


FIGURE 3.3. Distribution of documents per type

Following the PRISMA 2020 guidelines for systematic reviews, the next step involved a screening process of the remaining 23 documents. The process was carried out by reviewing their titles and abstracts to determine the relevance of each document to the research. Literature that did not align with the study's focus was excluded. As a result, eight documents were removed at this iteration.

With 15 remaining documents, the next step was to assess their eligibility, as indicated in the "Reports assessed for eligibility" phase of the flow diagram. A curated list of potential documents was created within Scopus for further verification. At this stage, retrieval of the full-text documents was attempted. However, one document could not be accessed, due to restrictions set by publishers, resulting in 14 documents at the end of the screening process.

Furthermore, the PRISMA model also includes the identification of relevant studies through alternative sources, nominated as "Identification of studies via other methods". In this case, the supervisors provided two additional documents that were not retrieved through the initial search query yet relevant to the research. They were added to the other 14 documents retrieved after via a search query (see Figure 3.1).

Once all the sources were verified, a total of 16 documents were selected for inclusion in the literature review. These documents were thoroughly analyzed to identify existing research gaps in the field of IL, AL techniques, and the areas of application in which these techniques have been implemented.

3.3. Background Concepts

Imbalanced Learning Problem

In machine learning, imbalanced learning refers to scenarios where the distribution of classes in a dataset is highly skewed, meaning that one class has significantly more samples than the minority one. The evident imbalance poses a substantial challenge for standard ML algorithms, which often assume an even distribution of classes and, as a result, tend to favor the majority class, leading to suboptimal classification of the minority class [1] [3].

One of the most critical real-world applications of imbalanced learning is in medical diagnosis. For instance, in mammography exams, the number of non-cancerous cases often exceeds the number of cancerous ones. Consequently, a naive classifier may achieve high accuracy simply by predicting all cases as non-cancerous, but this would be problematic as it would fail to identify patients who require who require other types of medical care [1]. This issue is also prevalent in other areas such as fraud detection and security threat identification where the minority class often contains the most relevant instances for further analysis [3].

The impact of imbalanced data extends beyond classification accuracy. It affects the learning process itself, as the algorithm may struggle to learn the patterns of the minority class due to its limited representation. As a result, it becomes necessary to implement specialized techniques to address this tendency. The state-of-the-art mentions various explored approaches, including data-level methods (resampling strategies such as oversampling the minority class or undersampling the majority class), algorithm-level modifications

(adjusting loss functions to penalize misclassifications differently) and hybrid techniques that combine both strategies [3].

Furthermore, imbalanced learning can be categorized based on the underlying cause of the imbalance, leading to different interpretations. The term "intrinsic imbalance" refers to cases where the imbalance is inherent to the data, to its nature, such as the rarity of fraudulent transactions or network intrusions. In contrast, "extrinsic imbalance" arises from external factors, such as biased data collection methods. Moreover, a relative imbalance occurs when the difference in class distribution is large but not necessarily extreme, while a rare-instance imbalance occurs when the minority class is exceptionally small and underrepresented.

As data complexity increases, particularly with the advent of big data and large-scale systems, the challenges associated with imbalanced learning become even more pronounced. Industries such as finance, healthcare and cybersecurity face pressing needs for intelligent algorithms capable of handling imbalance effectively. As highlighted in [3] current research in the field has categorized solutions into general methods, ensemble learning techniques, imbalanced regression and clustering, long-tail learning, and imbalanced data streams. Within these categories, deeper strategies such as boosting, bagging, cost-sensitive learning and concept drift adaptation continue to evolve to tackle imbalanced learning problems more effectively.

Addressing imbalanced learning is not simply a technical challenge but also an actual problem where misclassifications can have life-changing consequences. Ensuring that intelligent systems recognize and mitigate bias is crucial for the reliability and ethical responsibility of machine learning applications. This discussion sets the stage for the next section, where existing strategies will be examined alongside their real-world applications and effectiveness.

Active Learning for Imbalanced Data

Given that predictive modeling relies heavily on training data with ground-truth labels, the quality and representativeness of these labels significantly impact model performance. In imbalanced learning scenarios, obtaining sufficient labeled data for the minority class is particularly challenging. Labeling data can be costly, time-consuming or require intervention from experts. Active Learning (AL emerges as a promising solution, where in some contexts, the goal is to avoid spending the entire labeling budget on the majority class, ensuring that the minority class is also sufficiently represented.

The work in [17] identify two main perspectives for applying AL on imbalanced datasets: when the labels for all samples are already available and when the labeling process has not yet been conducted. In the first approach, AL is

used to mitigate the adverse effects of class imbalance on the model's performance. In the second one, AL helps select and acquire the most informative data points from both the majority and minority classes during the ongoing labeling process.

Similar views are echoed in [2], who emphasizes that when the learning algorithm can choose the data it learns from, it will perform better with fewer training instances, thus saving computational resources. This idea aligns with the broader challenge in supervised learning: the need for large amounts of labeled data.

While some labels come at little or no cost (e.g., user-generated flags on spam emails), many real-world applications require costly, time-consuming and hard-to-obtain labeled data. That is when AL comes in handy.

Active Learning operates primarily under two selection strategies. Firstly, a stream-based selection which presents unlabeled examples one at a time to the model, deciding whether to query the label for each example. The second, pool-based selection, involves an offline, iterative process where a large set of unlabeled examples is available. In pool-based Active Learning, examples are chosen in multiple iterations, with the process continuing until the budget is exhausted or a stopping criterion is met. Pool-based AL has gained more attention in literature due to its ability to process a larger representation of the problem space at once, which is especially beneficial in imbalanced learning contexts.

- **Uncertainty sampling**: Within the broader scope of AL, several techniques are proposed in the state-of-the-art to enhance model performance, particularly in imbalanced settings. One widely-used approach is uncertainty sampling, which aims to select examples from regions where the model is most uncertain, typically near the decision boundary. The logic behind uncertainty sampling is that these regions represent areas where the model is most likely to make mistakes. By incorporating labeled examples from these uncertain regions, the model can refine its decision boundary and, ultimately, improving performance.
- **Query-by-Committee**: The technique applies to where the model relies on an ensemble of models. The ensemble members are used to identify data points where there is a high degree of disagreement among the models. By selecting the highly uncertain instances, the approach ensures that the model is exposed to the most informative examples, enhancing its learning efficiency.
- **Density-Sensitive Active Learning**: Another adaptation is density-sensitive AL, which focuses on the geometry of the problem space to identify rare subregions within the minority class. By concentrating on

these areas, the model can avoid outliers and sparse regions, ensuring focus on the minority class. Methods like information density and pre-clustering are useful in identifying these high-density areas, thus improving the model's ability to generalize across imbalanced data.

- **Support Vector Machines**: A model's most informative examples are those that lie closest to the hyperplane. In an SVM, data points that are located near the margin are less likely to be imbalanced than the full dataset, making these samples particularly valuable for the learning process. By focusing on such examples, Active Learning AL can help address the imbalance issue while also enhancing overall model performance.

Despite the power of these AL techniques, class imbalance presents a unique challenge. The disproportionate number of majority class examples can cause the model to favor the majority class, often leading to poor performance on minority class instances. When labels selected by AL are treated as a random variable, the bias toward majority class examples can exacerbate the imbalance problem, making it more difficult for the model to correctly classify minority class instances.

3.4. Previous Work on Imbalanced and Active Learning Techniques

This section outlines the techniques implemented to address specific contexts where imbalance is present. In addition to explaining each technique, this section will discuss the context in which it was applied, the dataset used, the area of application and the resulting performance. In conclusion, a summary of the findings and insights will be presented.

Rare Classes Exploration and Classification

State-of-the-art introduces a hybrid adaptive technique that integrates uncertainty sampling, likelihood sampling and classifier switching to optimize both rare class discovery and classification. Rather than relying on a fixed query strategy, it dynamically balances likelihood sampling - used to detect potential new classes - and uncertainty sampling, which refines decision boundaries.

A generative-discriminative model pair, where a Gaussian Mixture Model aids in early class discovery and a SVM improves classification as more labels become available. The technique switches between classifiers based on multiclass classification entropy, ensuring that the most effective model is used at different stages of learning [18].

In addition, density weighting is considered to prioritize high-density regions, but the results indicate that this may hamper rare class detection, as

minority classes often occur in low-density areas. The adaptive method differs from traditional AL by dynamically optimizing query selection and classifier choice, making it particularly relevant for imbalanced learning scenarios. While not restricted to a specific type of dataset, its adaptability suggests suitability for applications such as anomaly detection, fraud detection or medical diagnosis, where minority class discovery is critical [18].

CQEnsemble: Cost-sensitive & Query-By-Committee Method

CQEnsemble is a cost-sensitive ensemble approach that leverages Query-By-Committee (QBC) active learning to address the challenge of imbalanced datasets. The algorithm begins by dividing the dataset into multiple subsets, ensuring that the class imbalance is maintained in a controlled way. Two subsets are randomly selected to form a committee of classifiers. These classifiers are trained using the AdaBoost algorithm, with majority class samples being further split into smaller subsets and combined with minority class samples to create balanced training sets. This step ensures that the classifiers learn effectively from both classes [19].

The QBC aspect is applied by using these trained classifiers to classify the remaining subsets. Samples where the classifiers disagree -indicative of uncertainty - are selected as particularly informative and added to a candidate set. The candidate set is then merged with the subsets originally used to train the classifiers, creating a more robust and targeted training set. The final classifier is trained on this set using a cost-sensitive SVM, which explicitly penalizes misclassifications of the minority class.

The authors prove, with various datasets from UCI Machine Learning Repository, the effectiveness of the algorithm for imbalanced datasets, by avoiding traditional oversampling or undersampling methods, which can either introduce noise or lose information. Instead, it focuses on identifying critical samples where the model is uncertain, ensuring better decision boundaries for both classes. Experimental results have demonstrated its superiority in metrics like F-measure, G-mean and AUC compared to methods such as Synthetic Minority Oversampling Technique (SMOTE), AdaBoost and EasyEnsemble, making it a valuable contribution to the state of the art in active learning for imbalanced classification.

Novel Sampling Method for Large Scale and Imbalanced Learning Problem

The proposed algorithm is an active learning method designed to address the challenges of large-scale, imbalanced datasets, particularly in real-world applications such as telecommunications fraud detection. The main objective is to enhance classification performance by iteratively identifying and selecting

the most informative and potentially misclassified samples for labeling and training [20].

Essentially, a small subset of the unlabeled dataset is randomly selected, labeled, and used to form the initial training set. A classifier is then trained on this set and in each subsequent iteration, the remaining unlabeled samples are scored based on their uncertainty and the likelihood of misclassifications. The highest-scoring samples are selected, labeled, and added to the training set. To address the issue of class imbalance, the algorithm integrates two resampling techniques:SMOTE andkNN.

SMOTE generates synthetic samples of the minority class to enhance its representation, ensuring the classifier has sufficient exposure to minority instances. While, KNN-based undersampling selectively retains the most informative majority class samples by identifying those closest to the minority class in feature space. Together, these techniques maintain a balanced training set, preventing the classifier from becoming biased toward the majority class while preserving the most critical data points.

Experimental results demonstrate the innovative method outperformed baseline active learning strategies, including random sampling, uncertainty sampling, and query-by-committee. Using AUC as the primary evaluation metric, the approach exhibited steady improvement over iterations, with SMOTE proving particularly effective in enhancing the detection of minority class instances. Applied to a real-world telecommunications fraud dataset, the algorithm successfully handled extreme class imbalance and delivered superior classification outcomes, especially for the minority class. Overall, the proposed method proved effective, showcasing significant potential for addressing imbalanced classification challenges in practical applications.

Although the focus is on the application of the active learning algorithm to telecommunications fraud detection, using Call Detail Records, the core principles of the algorithm suggest potential applicability to other domains with similar characteristics, namely large-scale and imbalanced datasets. Such flexibility highlights its potential for application across various fields.

Deep Belief Network with Active Learning

In [21] it is proposed an AL algorithm that optimizes a Deep Belief Network for imbalanced data by selecting informative instances around decision boundaries. The selection process balances two key aspects: diversity and uncertainty.

Diversity is achieved by selecting instances that are distinct from each other, while uncertainty is measured by the margin of confidence - the difference between the probabilities of the most likely and second-most likely labels. These

uncertain instances are considered essential for refining the decision boundaries.

The AL algorithm formulates this selection process with a "r" probability value which determines the likelihood of selecting a sample. This allows for the efficient identification of the most informative instances. An optimizing set, which mimics the test set distribution, is then used to guide the model refinement process, minimizing the negative effects of resampling methods like overfitting.

In practice, Deep Belief Network is initially trained on an oversampled dataset to address class imbalance. AL is then applied to focus on instances near decision boundaries, helping to improve the model's performance by optimizing areas with the highest uncertainty. Therefore, Active Learning ensures better generalization and enhances the Deep Belief Network's ability to handle imbalanced data.

The technique is applied to both benchmark datasets and real-world industrial data to assess its performance. Five benchmark datasets from UCI Machine Learning Repository are used. The industrial dataset highlighted the effectiveness of the method in tackling multiclass imbalanced problems in a practical setting [21].

Oversampling Method Based on Labelling Budget

A pioneering active learning oversampling algorithm specifically designed for multi-class sparsely labeled data streams is introduced in [22] . This method effectively addresses stringent labeling budget constraints while dynamically managing class imbalance and concept drift. By simulating the behavior of a fully labeled and balanced data stream, the algorithm is both versatile and classifier-agnostic. Its implementation is structured around three core components: label acquisition, balancing strategies and instance generation.

- **Label Acquisition**: The method employs a random selection strategy based on an uniform distribution. If the generated random value is below the defined budget threshold, the classifier requests the instance's label. This strategy ensures that the estimated class proportions statistically reflect the overall data stream in the long run. To handle the non-stationary nature of data streams after a drift, old information about class proportions should not be considered for incoming instances.
- **Balancing Strategies**: The balancing module determines the number of synthetic instances to generate based on the current class imbalance and labeling budget. The function behind this process is inversely proportional to the labeling budget, dynamically increasing the number

of synthetic samples for minority classes as the budget becomes more constrained. This mechanism enables the method to simulate the behavior of a fully labeled stream while continuously adjusting to evolving class distributions and budget limitations.

- **Instance Generation**: Serves to be performed through two complementary techniques that create synthetic samples based on the calculated balancing ratio: Budget-Based Resampling and Budget-Based SMOTE. Budget-Based Resampling replicates existing instances multiple times, presenting them repeatedly to the classifier. In contrast, Budget-Based SMOTE extends the original SMOTE method to sparsely labeled streams. Using a sliding window of recent instances, the algorithm generates synthetic samples by identifying k-nearest neighbors from the same class and inserting new instances. Together, these techniques effectively mitigate the challenges of underrepresented classes in multi-class imbalanced streams, particularly under concept drift and limited labeling budgets.

The budget-based resampling algorithm addresses critical challenges associated with multi-class imbalanced data streams, including adapting effectively to non-stationary class distributions and handling concept drift. It ensures the efficient use of limited labeled data by functioning effectively within strict budget constraints while maintaining robust predictive performance. Conducted on 18 real-world datasets and 11 semi-synthetic streams, the experimental evaluation underscores the method's ability to significantly enhance classifier performance across a range of budgetary scenarios. The most substantial improvements were observed in cases with highly restricted budgets, demonstrating the method's adaptability and effectiveness. Furthermore, the statistical analyses the authors present highlights the robustness and reliability of the approach [22].

3.5. Previous Work on Imbalanced Learning in ECG-based Arrhythmia Detection

This section presents research related to the technical domain in which the nature of the selected dataset is included, i.e., ECG-based. It will focus on studies that handle the classification of cardiac arrhythmias, with particular emphasis on the detection of AFLAtrial Flutter, in contexts of imbalanced data. The approaches here discussed will provide important insights and contextual foundation for the subsequent chapters.

Automated Atrial Fibrillation Detection using a Hybrid CNN-LSTM Network on Imbalanced ECG Datasets

Although the study refers to atrial Fibrillation, it can be relevant and applicable to other conditions and because of that it consists of a hybrid deep learning architecture that combines a Convolutional Neural Network (CNN) with a Long Short-Term Memory(LSTM) network to detect and classify cardiac arrhythmias from ECG signals. The model incorporates Focal Loss, a modification of the traditional cross-entropy loss that mitigates class imbalance by emphasizing hard-to-classify minority samples [23].

The authors Petmezas et al. (2021) [23] use ECG signals from the MIT-BIH Atrial Fibrillation Database. The model is trained to classify several rhythm types, including Normal, Atrial Fibrillation, Atrial Flutter (AFL) and AV Junctional Rhythm. The CNN component extracts morphological features from denoised ECG beats, while the LSTM captures temporal dependencies in the signal. The use of Focal Loss further enhances sensitivity to underrepresented classes.

Despite the limited number of Atrial Flutter (AFL) samples, the model achieved high precision and recall, confirming its ability to detect minority rhythms effectively. This hybrid approach demonstrates how combining spatial feature extraction with temporal modeling can improve arrhythmia classification, offering a framework adaptable to future work on Atrial Flutter detection under class imbalance conditions.

Pretrained Foundation Model for Atrial Fibrillation and Atrial Flutter Detection

The proposed strategy investigates the use of deep learning to enhance the automated detection of atrial arrhythmias through ECG data [24]. The main objective is to fine-tune a pretrained ECG Foundation Model (ECG-FM) for the binary classification of Atrial Fibrillation and Atrial Flutter versus other cardiac rhythms using single-lead ECG recordings.

The authors Lee and Choi (2025) in [24] address several key challenges that commonly arise in automated ECG interpretation. First, it tackles the issue of limited and task-specific datasets, which often restrict the generalization ability of deep learning models. By leveraging transfer learning from a large, unlabeled set of ECG data, the pretrained ECG-FM is able to learn general morphological and temporal representations, improving robustness and data efficiency. Moreover, the study addresses class imbalance, a frequent problem in arrhythmia datasets where Atrial Fibrillation/Atrial Flutter events are underrepresented. To solve this problem, the authors employ a weighted loss function that assigns greater importance to minority-class samples. Finally, the study adapts the pretrained model for single-lead ECG analysis.

The model architecture is derived from the *Wav2Vec 2.0* framework, originally designed for self-supervised signal representation learning. It comprises

three main components: a convolutional feature extractor that captures low-level ECG morphology and frequency information; a transformer encoder which models long-range temporal dependencies and a classifier head that performs the final binary classification task.

From a clinical perspective, the study directly targets AFL detection, treating AFL and Atrial Fibrillation jointly as atrial arrhythmias of clinical relevance. Both conditions are characterized by rapid, irregular atrial activity, disrupted P-waves and irregular RR intervals, which pose challenges for automated analysis. The successful fine-tuning of the ECG-FM demonstrates that even single-lead ECG input can provide sufficient diagnostic information for accurate AFL and Atrial Fibrillation detection, underscoring the feasibility of implementing such models in real-world clinical and mobile health settings.

Overall, this research highlights its potential in overcoming the limitations of small and imbalanced ECG datasets. By combining large-scale pretrained representations with class-weighted fine-tuning, the approach achieves improved generalization and sensitivity to minority arrhythmias such as Atrial Flutter AFL.

Machine Learning Prediction of Atrial Flutter Mechanism from P-P wave Intervals

The present research focuses on developing a supervised learning model capable of discriminating the Atrial Flutter using 12-lead ECG signals [25]. The study is divided in two mechanisms of interest, the Focal Atrial Flutter (FAFL) and the Macroreentrant Atrial Flutter (MAFL). Clinically distinct, the FAFL originates from localized ectopic atrial activity, while the latter results from large, stable reentrant circuits.

What it is hypothesized in [25] is that the variability in the P-to-P interval series of an ECG can serve as a discriminative marker between these two mechanisms. While MAFL tends to produce highly stable atrial activations, FAFL displays greater temporal irregularity in the P-wave intervals. Well, using ECG recordings from 46 patients, where 41 are MAFL and only 5 are representative of FAFL, the study extracts descriptive statistical features from the P-to-P interval series, adopting a feature engineering process that is independent of signal delineation or waveform morphology.

Given the strong class imbalance - approximately a 8:1 ratio, the study implements several data augmentation methods, in particular SMOTE, modified-SMOTE and smoothed-bootstrap to mitigate model bias toward the majority class. Modified-SMOTE achieves the best performance in combination with Logistic Regression, yielding an accuracy of 77%, approximately, under a 400% augmentation of the minority class.

The Modified-SMOTE algorithm represents the paper's primary methodological innovation as it introduces a correction factor to the standard SMOTE formulation in order to preserve the variance and distributional characteristics of the original data. By re-expanding the variance of the synthetic samples, the algorithm minimizes artificial correlation and improves the fidelity of augmented data. Statistical validation confirms that Modified-SMOTE generates the most realistic synthetic data compared to other augmentation strategies.

Overall, this research provides an important contribution to imbalanced learning in ECG-based arrhythmia classification, showcasing how tailored data augmentation - particularly variance-preserving synthetic sampling - can improve diagnostic model performance in minority-class recognition tasks.

3.6. Conclusion

The previous analysis highlights the growing interest of the research community in addressing the scarcity of representative data for minority classes, including in medical practice in order to identify heart conditions like Atrial Flutter. The investigated strategies demonstrate significant progress but also emphasize persistent challenges, such as the careful selection of informative instances and the adaptation of traditional methods to highly imbalanced scenarios. Furthermore, in future phases of the research, exploring AL algorithms may reveal interesting outcomes for handling class imbalance in the particular case of AFL classification. Expanding the scope of investigation in this direction could contribute to the development of more robust and generalizable solutions, fostering advancements in the field and mitigating the limitations faced by current models.

As reported in the literature of section 3.5, most existing studies focus primarily on Atrial Fibrillation, whereas Atrial Flutter has received considerably less attention, yet justified by its lower prevalence. Nevertheless, Atrial Flutter remains an underexplored condition, with limited methodological advancements and few significant breakthroughs reported to date.

Consequently, this dissertation is both relevant and scientifically meaningful, as it addresses an existing research gap. The main objective is to investigate and evaluate the applicability of established techniques along with Active Learning when applied to a highly imbalanced dataset, aiming to reliably and accurately identify cases of Atrial Flutter. Furthermore, there is no evidence of combining Active Learning with the implementation of a Generative Adversarial Network GAN in a AFL context, and for that this dissertation is innovative.

[This page is intentionally left blank.]

CHAPTER 4

Methodology

At this stage of the dissertation, it is assumed that the reader is familiar with the technical/clinical concepts underlying the data. Therefore, this chapter outlines the methodological approach followed throughout the project and provides a comprehensive overview of the dataset. Besides, it details the preprocessing procedures implemented during the data understanding and data preparation phases which, naturally, cover data exploration, visualization, manipulation and evaluation processes.

This chapter, consisting of five sections, begins with a thorough description of the dataset and its storage structure, followed by a section concerning class distribution which includes a data assessment over demographic variables (i.e., age and sex). The Preprocessing Data section then addresses noise reduction in ECG signals, as well as, the feature extraction procedure and handling missing values during data preparation. Focused on AL strategies, the fourth section introduces the two selected algorithms and provides the reason for their selection in this context. The final section - Experiences - is divided into three sub-sections and outlines the components used in the classification model. This includes the data split configuration, the sampling methods, performance metrics and the integration of a GAN to support the experimental setup.

Thus, this chapter sets the foundation for the experimental work by presenting, preprocessing and applying the data to a demanding and sensitive classification model. The following sections guide the reader through each of these phases, ultimately framing the replicable experimental strategy.

4.1. Dataset

As mentioned in Chapter 2, the adopted dataset - entitled “A large scale 12-lead electrocardiogram database for arrhythmia study” [5] - is extracted from the *PhysioNet* open-access repository. The data was gathered under the auspices of Chapman University, Shaoxing People’s Hospital (Shaoxing Hospital Zhejiang University School of Medicine), and Ningbo First Hospital.

The dataset free availability serves the scientific community in conducting new studies on arrhythmia and other cardiovascular conditions [5]. The dataset is extracted from the repository in its entirety.

Originally, the dataset comprises 45 152 ECG recordings, representing different individuals, i.e., 1 record = 1 patient, identified by unique IDs. The patient's data is divided into two types of files, both identified by a unique patient ID. Each patient has his data represented in: a .mat format file, containing the 10 second long raw ECG signal information and a .hea file with complementary personal information and metadata. Furthermore, the package contains a file with the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) scheme, which provides a list of the standardized codes for medical conditions, enabling clear and consistent identification during diagnoses.

In order to simplify the browsing and data manipulation in future phases of the project, a preprocessing method is carried out to extract and integrate the ECG signals in a more accessible format. So, on one hand, the .mat files are accessed and the ECG signals are converted to .npy format - preserving the patient ID as the identifier of each file. On the other hand, the personal data in the .hea files is transferred to a CSV file, which includes demographic data such as the patient's age and sex, the SNOMED CT code identifying the diagnosed heart condition(s). Metadata is also included, with technical details concerning the ECG sampling frequency and the number of recorded samples - both variables with fixed values - 500 Hz and 5 000, respectively. Upon completing the data transfer to new formats, two records are retained as they cannot be extracted correctly from the original folders, therefore, the total amount of ECG recordings is updated to 45 150.

The file's extraction and transformation facilitates the subsequent phases of data understanding and preparation by making it easier to navigate the dataset, select and observe specific ECG signals.

4.2. Class Distribution

The data understanding phase comprehends both data assessment over the variables and ECG waveforms but also the class distribution of the dataset. This last matter is of the highest importance, since it indicates the degree of class imbalance present in the dataset - the premise of the study. In order to define the class distribution, a preliminary analysis over the personal data which holds information regarding the diagnosed conditions, mirrors several records with multiple heart conditions codes associated - indicated in the SNOMED CT column. This observation conveys the idea that when performing an ECG, it is a common outcome to get several pathologies at once, as the exam covers all heart's points of view and therefore, identifies different problems all around the organ.

However, as the project's purpose is to address the behavior in single-labeled Atrial Flutter cases, instances with plural heart conditions - including Atrial

Flutter - are excluded from the minority and join the majority class. The reason behind this decision is strategic; preserving pure Atrial Flutter characteristics and ECG morphologies, avoiding the integration of mixed condition AFL cases in the minority class. Adding all AFL-plus-other-conditions cases would set a precedent and compromise the classification process. Hence, restricting the minority class to exclusive Atrial Flutter diagnoses assures the model to be focused on representative patterns of the highlighted condition.

Thereby, the minority class records' filtering process selects the instances which only present at the SNOMED CT column the "164890007" code number - the official identifier of Atrial Flutter. As a result, the minority class consists of 1 483 AFL records while the remaining 43 667 samples are assigned to the majority class (Figure 4.1). In terms of percentage, Atrial Flutter records represent 3.3% of the dataset with the remaining 96.7% corresponding to the Non-Atrial Flutter class, i.e., the majority class.

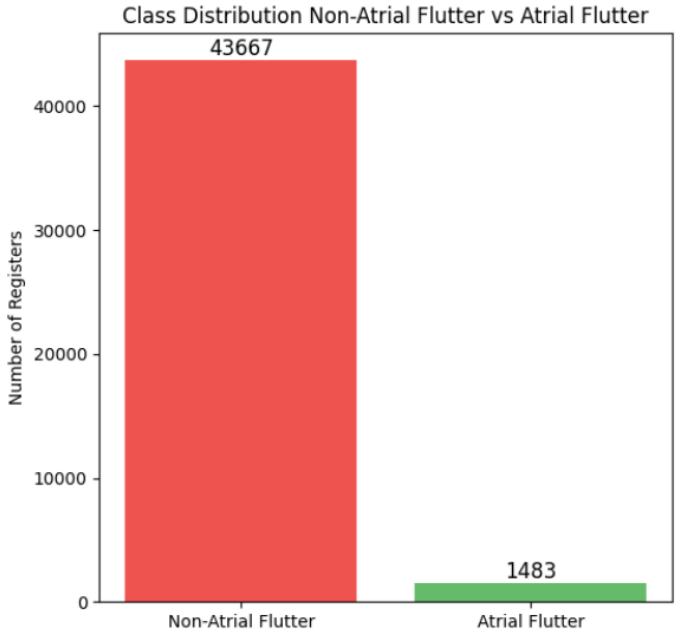


FIGURE 4.1. Class Distribution

Sex & Age Distribution over Classes

Besides the general class observations, assessing the patient's sex and age is of high value too since it reflects the Atrial Flutter's incidence demographically, as mentioned in Section 2.2.

Regarding the sex variable, the column presents three possible values - Male, Female and Unknown (see Table 4.1). Both classes reveal a predominance of Male individuals over Female - in total, 25 452 Male patients versus 19 676 Female. Additionally, 22 people have labeled their sex as Unknown, all of whom belong to the majority class.

Concerning the sex distribution on Atrial Flutter records, as Figure 4.2 indicates, 62.2% of patients correspond to Male and 37.8% to Female.

TABLE 4.1. Class Distribution by Sex

	Non-Atrial Flutter	Atrial Flutter	Total
Male	24 530	922	25 452
Female	19 115	561	19 676
Unknown	22	0	22
Total	43 667	1 483	45 150

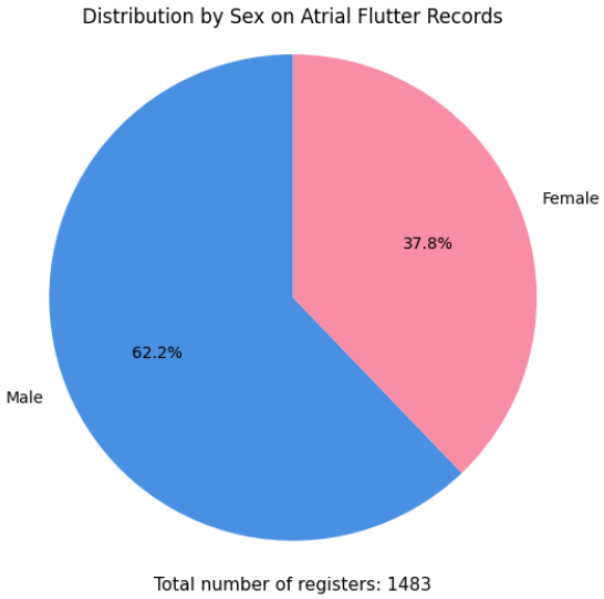


FIGURE 4.2. Sex Distribution on Atrial Flutter

On the other hand, the age parameter reveals a greater Atrial Flutter incidence over older people (50+ years old) compared to younger individuals. Table 4.2 shows the effect alongside with the patient's sex. The difference between ages is blatant in both classes with emphasis in the minority class that presents 855 Male patients and 521 Female patients with 50+ years old, compared to 62 and 40 cases, respectively, which are below the 50 years old age. There are some patients which are not contemplated in this count: 55 patients are encountered with no attributed age, in which five of them belong to the minority class - and two others have legible age but Unknown sex label. These specific patients combine in total 57, which subtracted to the initial 45 150 patients result in the 45 093 presented in Table 4.2 .

To conclude, the pie chart in Figure 4.3 outlines the age distribution within the minority class. About 93.1% of patients are older than 50 years old, while 6.9% represent younger individuals, enhancing the tendency invoked previously in Section 2.2 regarding the condition's incidence.

TABLE 4.2. Class Distribution by Sex and Age

Gender	Age	Non-Atrial Flutter	Atrial Flutter	Total
Male	< 50	6 372	62	6 434
	50+	18 138	855	18 993
Female	< 50	5881	40	5 921
	50+	13 224	521	13 745
	Total	43 615	1 478	45 093

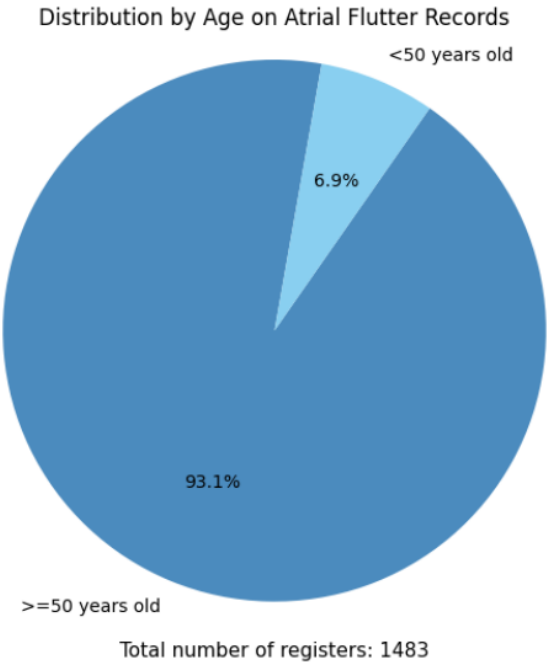


FIGURE 4.3. Age Distribution on Atrial Flutter

4.3. Preprocessing Data

4.3.1. ECG Signal Noise Cleaning

As suggested in literature of the area, ECG signal denoising is a recommended step during data understanding. In fact, the Python *Neurokit2* library is commonly used for that purpose since it is designed for signal visualization and waveform inspection. For that reason, *Neurokit2* was initially adopted to identify potential noise associated to the ECG signals and remove it. Nonetheless, the resulting outcomes did not reach satisfactory levels in terms of effectiveness, as the clean function did not remove the noise entirely.

Alternatively, the Butterworth filter comes as a solution for the noise cleaning process, as a suggestion by the supervisors. According to literature [26], the Butterworth filter is usually employed in ECG analysis to minimize noise and enhance signal clarity. The method improves both human and machine data interpretation and presents three possible filter configurations: low-pass, high-pass and band-pass. To choose the most adequate configuration, context

and data characteristics must be taken into consideration, since each variation removes high-frequency noise, low-frequency noise and isolate a specific frequency range, respectively.

In order to determine the most suitable filter option, a series of tests are conducted, adjusting the frequency parameters on the Butterworth. Figures 4.4, 4.5 and 4.6 represent the low-pass, high-pass and band-pass implementation, respectively, over the same ECG signal, which corresponds to an Atrial Flutter example. The criteria to choose the right configuration goes by the noise reduction along with the signal preservation. The low-pass filter unravels to present the best outcome, with a cutoff frequency of 25 Hz, which removes high-frequency components that could obscure the ECG morphology and maintaining clinically relevant features such as the QRS complex and P wave amplitude, for instance (see Figure 4.4).

Consequently, the Butterworth low-pass filter is applied to the entire dataset, ensuring all ECG signals are equally denoised.

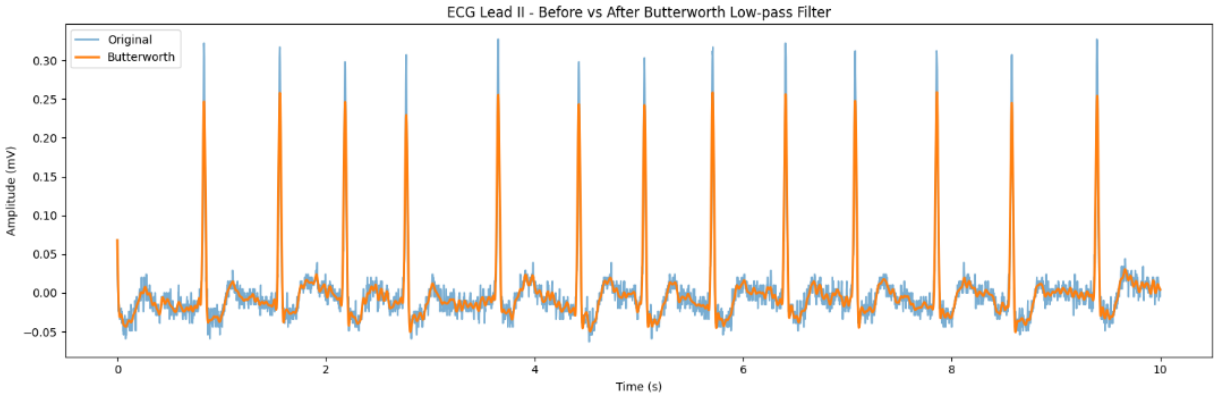


FIGURE 4.4. Low-pass Filter on ECG

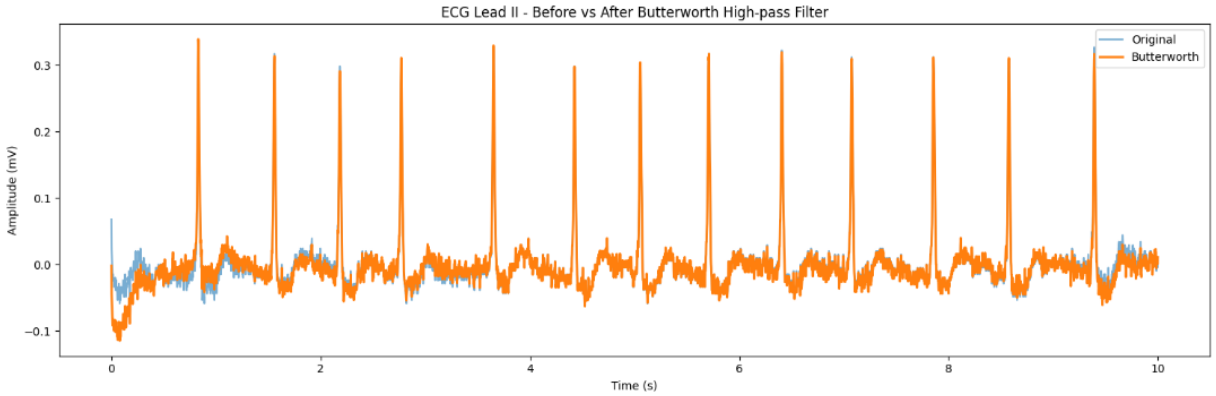


FIGURE 4.5. High-pass Filter on ECG

4.3.2. Feature Extraction

ECG signals are inherently time-series data represented as continuous waveforms. However, ML algorithms require input data to be in a structured and

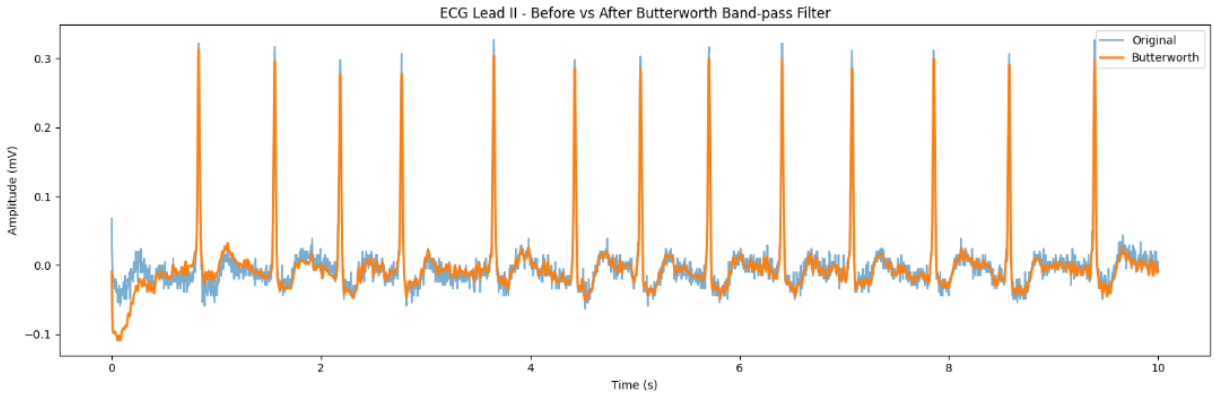


FIGURE 4.6. Band-pass Filter on ECG

tabular format composed of numerical parameters. The transformation from ECG traces to measurable features is essential because models cannot directly interpret raw signal waveforms - instead, they must rely on meaningful, quantifiable characteristics extracted from the signals. Hence, feature extraction is performed after the noise cleaning signals, making sure the variables are taken from the properly cleaned ECG signals and in shape for classification matters.

Dr. Rosário was consulted during this phase to help defining the list of features to be extracted from the records. The selection criteria was simple: to integrate both general ECG reading parameters and particular features specific to Atrial Flutter. Features like P-wave amplitude, PQ interval duration and Heart Rate Variability (HRV) metrics are included, even though only three HRV metrics are extracted, i.e., MeanNN, SDNN and RMSSD, due to relevance in Atrial Flutter cases. QRS complex duration and RR wave ratio are included as well. For the extraction process, the *delineate* function from *Neurokit2* is employed with emphasis on the wave peaks.

The extraction enables to represent each recording as a set of numerical features that capture the relevant information from the respective ECGs' waveforms. This structured representation ensures efficient training, comparison and interpretation of predictive models, particularly for tasks like classification.

Along with other typical features, the saw-tooth waves, characteristic of Atrial Flutter, are most easily observed in the inferior leads (i.e., II, III and aVF) and in lead V1, where their shape and regularity are most distinct (see Figure 4.7).

However, lead II usually presents the clearest saw-tooth morphology, as its orientation closely aligns with the main direction of atrial depolarization. Thus, lead II is particularly suitable for detecting the presence and frequency of flutter waves. Moreover, given the size of the dataset, only lead II is considered for the

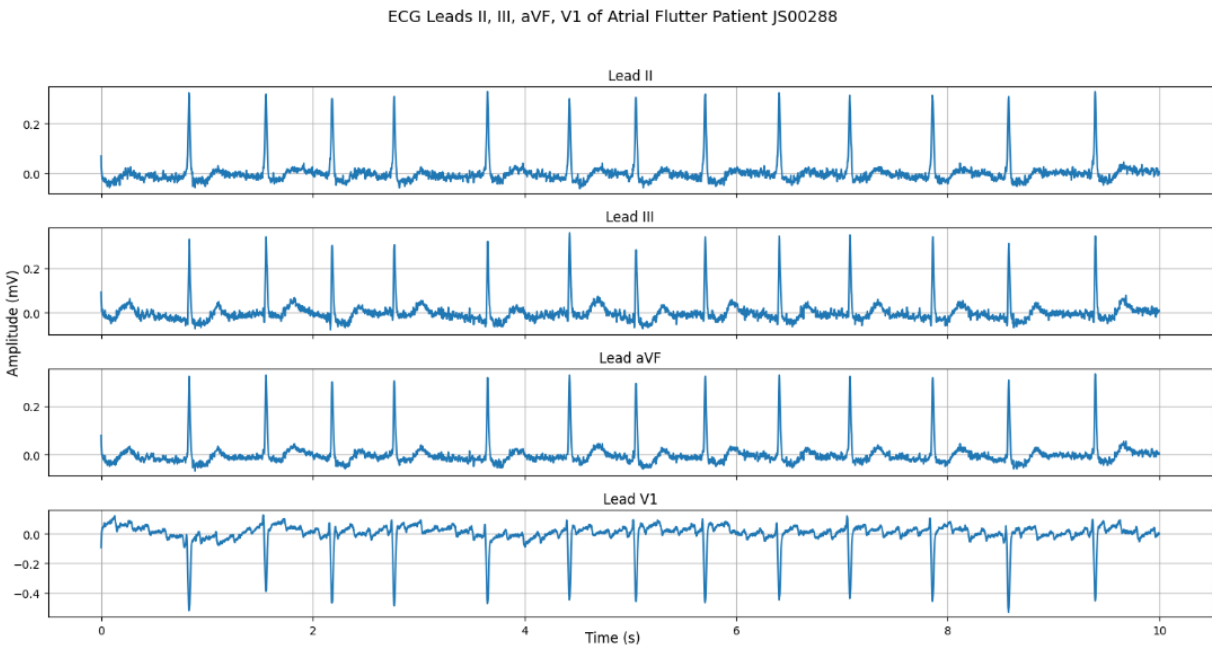


FIGURE 4.7. Leads II, III, aVF and V1 of Atrial Flutter Record

feature extraction process. Such approach provides a more robust and computationally efficient analysis; including features from multiple leads might emphasize redundancy with extensive information related to the same record and result in potential overfitting. Furthermore, initial experiments proved that extracting multiple leads confused the classifier and skewed the results having multiple variables from different yet related and similar leads.

The extraction procedure unfolds around the P, Q, R and S wave peaks in order to identify QRS complexes and RR intervals and wave amplitudes. PQ intervals are calculated separately from the other variables, due to the parameter's sensitivity and variability and the P-wave amplitude is recalculated to assure consistency. However, the amplitude values proved to be worse, as it provides one more missing value than the first attempt in the extraction function. Nevertheless, the PQ interval results are acceptable. Once the feature's computation is complete, the variables are merged with the respective patient's personal data through the record ID.

The resulting CSV file comprises both medical information relating to the patient's ECG exam, but also his age, sex and diagnose indicating wether he has Atrial Flutter or not.

4.3.3. Data Preparation

The data preparation process includes both analysis and the missing values treatment, with special attention to the minority class instances. Initially, the data is inspected to determine the total number of records containing missing entries. The majority class presents 154 cases with missing values, whereas

the minority class only has five cases, as Table 4.3 illustrates - all of which exhibit missing data for both the P-wave amplitude and PQ interval duration. In addition, records with missing age values (Nan) are assigned the label "*Unknown*" to distinguish between genuinely missing entries and those where age is not originally attributed. This distinction is considered less critical than missing ECG related parameters.

Despite the attempt to recalculate the P-wave amplitude variable, the morphology of the ECG signals - characterized by sharp peaks and frequent waveforms, as illustrated in Figure 4.8 - prevented *Neurokit2* from detecting P or Q-wave peaks for potential imputation. Given the small number of affected samples (5 out of the 1 483), it was concluded that applying an imputation method would not significantly impact the overall dataset. Furthermore, neither assigning zeros nor generating synthetic values using the kNN algorithm seemed appropriate, as such strategies might mislead the model during classification and introduce overfitting, particularly due to the natural absence of detectable waveforms in these cases.

TABLE 4.3. Class Distribution With Missing Values

	Non-Atrial Flutter	Atrial Flutter	Total
Missing Values	154	5	159
Correct Values	43 301	1478	44 779

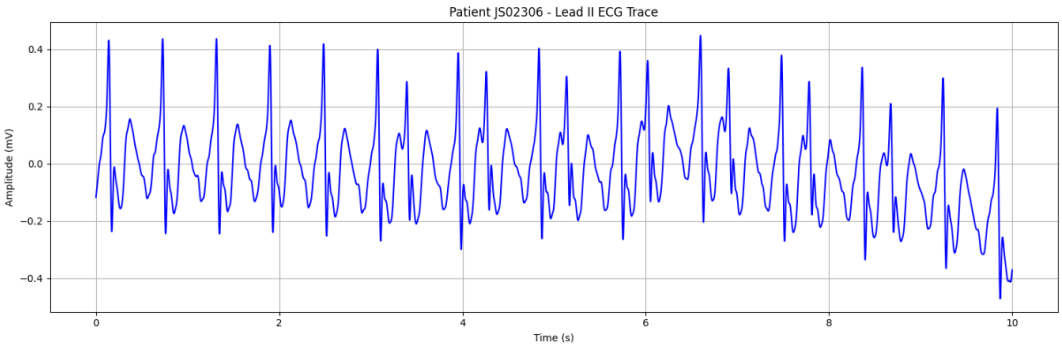


FIGURE 4.8. Atrial Flutter Record's ECG Lead II

As a result, the five minority class instances and the 154 records from the majority class were excluded from the dataset, updating the final total to the 44 779 patients presented in Table 4.3. This removal does not compromise the representativeness of the data, even though the minority class contains fewer instances - 1 478 records. The exclusion of 154 samples from the majority class only represents 0.35% of its total, which is negligible given that over 43 000 majority class instances remain.

Ultimately, the sex column which originally has d Female, Male and Unknown labels, is converted into binary values, i.e., 0, 1 and -1, respectively, to ensure consistency across the dataset.

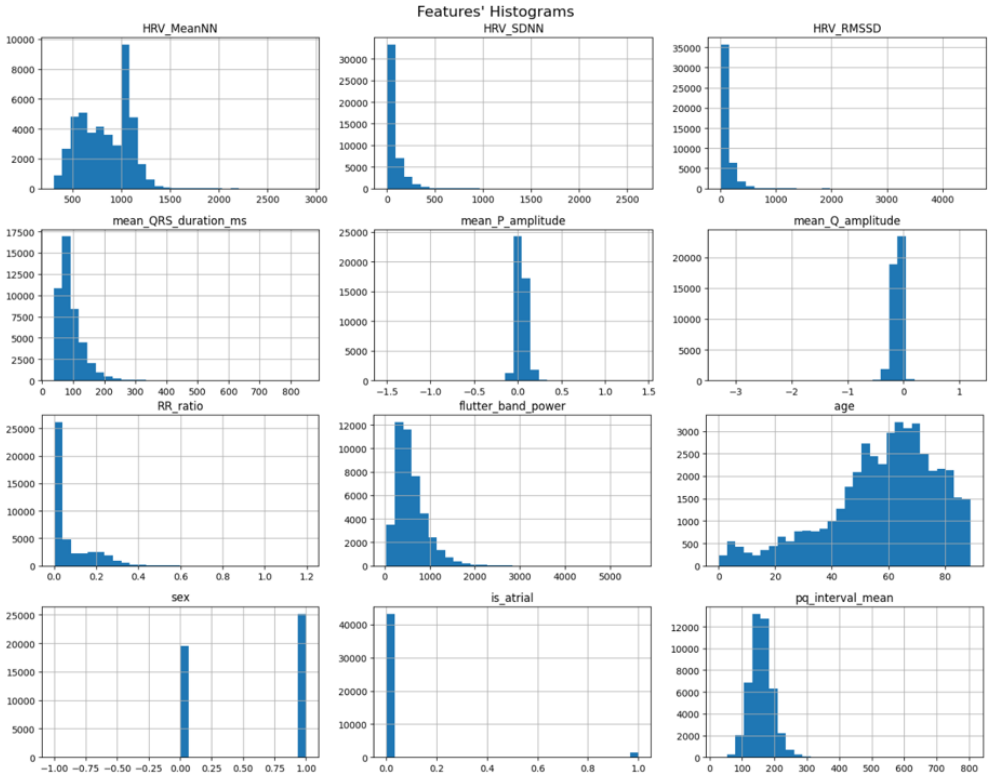


FIGURE 4.9. Histograms

The exploratory data analysis concludes with an overview of the features' distribution, where histograms are generated to observe the values range (see Figure 4.9). Additionally, the plots in Figure 4.10 identify numerical patterns between classes and facilitate a comparison of feature ranges between non-Atrial Flutter and Atrial Flutter records. These plots will have a major role when implementing the Generative Adversarial Network in sub-section 4.4.3.

4.4. Experiences

This section presents the experimental setup used throughout the study. Configuration settings and model specifications are included, as well as the algorithms applied, including the selected Active Learning strategies and evaluation metrics to assess the classifier's performance. Finally, the section concludes with an overview of GAN, which was incorporated as part of the experimental procedure.

4.4.1. Dataset split & Classifier

The experimental process starts by diving the dataset into two subsets: the training and test sets. In order to ensure consistency in the evaluation procedure, the data split is applied at a *testsize=0.2*, being the remaining 80% allocated to the training set. Therefore, as shown in Table 4.4, the dataset is composed by a total of 44 779 records. On one hand, the training set contains 35 823 samples while the test set holds 8 956 instances. In total, there are 1

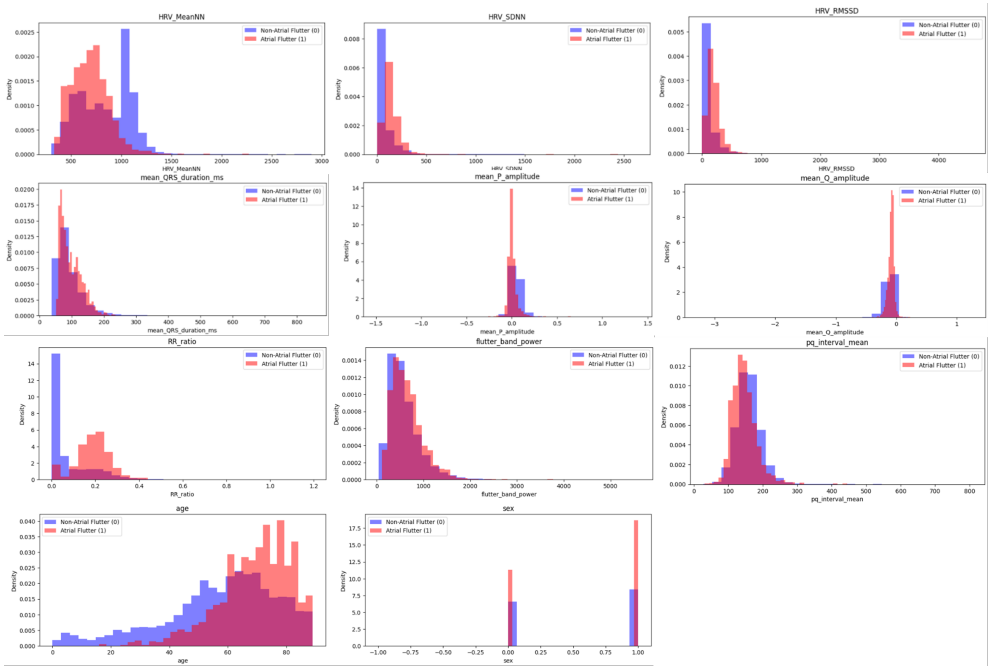


FIGURE 4.10. Histograms per Class

478 Atrial Flutter cases, where 1 182 are assigned to the training test and 296 to the test set. After the data is properly split, the classifier is trained and the standard sampling methods and AL algorithms are employed.

TABLE 4.4. Data split summary - Original Data

	Atrial Flutter	Atrial Flutter	Total
Training Set	34 641	1 182	35 823
Test Set	8 660	296	8 956
Total	43 301	1 478	44 779

The XGBoost classifier is selected due to its proven performance and adaptability in imbalanced learning scenarios. It combines a high predictive accuracy, handling efficiently skewed data distributions. Besides, its robustness to deal with data augmentation and Active Learning strategies smoothly make it particularly suitable for this study's iterative learning process.

In addition, to evaluate the classifier's performance under different data distribution conditions, various sampling and learning strategies are applied. These methods aim to balance the dataset to different extents, allowing a comparative analysis of their respective advantages and disadvantages in the context of highly imbalanced classification tasks.

4.4.2. Techniques & Performance Metrics

During experimental tests, the methods Undersampling, Oversampling, SMOTE-Tomek, Isolation Forest and the two Active Learning algorithms previously introduced: CBAL and FIMPS are implemented. Each of these methods, except

for the Active Learning algorithms - which select informative instances without explicit resampling - are tested on four different class imbalance ratios (i.e., 1:1, 2:1, 3:1 and 4:1) to observe the classifier's behavior in different imbalance situations.

After training, a confusion matrix is generated per ratio and method to assess the classifier's predictive ability, with particular attention in identifying the minority class (Atrial Flutter). The outcomes are then stored in structured dataframes for easy access for plotting and comparative analysis across different ratios and methods.

The performance metrics selection is based on relevance in imbalance classification problems. According to literature, accuracy, precision, recall, f1-score and ROC-AUC prove to be the most adequate in similar contexts. Nevertheless, precision and recall are the most considered due to their significance: precision indicates how many positive predictions are correct and recall reflects the model's capacity to detect all true positive cases.

Even though the results obtained from the baseline were not catastrophic, it can be interesting, from a research perspective, to implement a GAN and broaden knowledge and conclusions, assessing its impact in a situation of severe imbalance.

4.4.3. Generative Adversarial Network

A GAN is a deep generative model designed to generate realistic data by learning patterns from existing training datasets and frequently used in cases of imbalance class representation. The model is composed by two networks: the generator and the discriminator. While the generator works to produce similar samples to real data, the discriminator, on the other hand, tries to distinguish the synthetic data from the legitimate. During the training process, the discriminator learns how to improve his classification ability, by refining its parameters to detect artificial samples more accurately. Whereas the generator's main objective is to produce the most realistic instances possible so that the discriminator has trouble identifying the synthetic ones as fake and therefore labels them as real.

The dynamic generator-discriminator has proven to be effective and fruitful as the generated instances improve restoring the imbalance issue and their resemblance to the real data becomes closer. Moreover, considering the context in which the data is inserted, several studies suggest that generating synthetic ECG signals can be a valuable tool, helping with the identification of heart diseases and rebalancing the class distribution [27].

Therefore, in the present study, a GAN is implemented with the purpose of increasing the Atrial Flutter representation and even the existing underrepresentation between classes. Besides, the incentive to employ such measure

is moved by the bad results obtained by the classifier when training with the baseline data. The GAN came up as a well-proven solution to mitigate the previous outcomes with no synthetic sample method, as it is designed to produce with high reliability synthetic samples of a certain class.

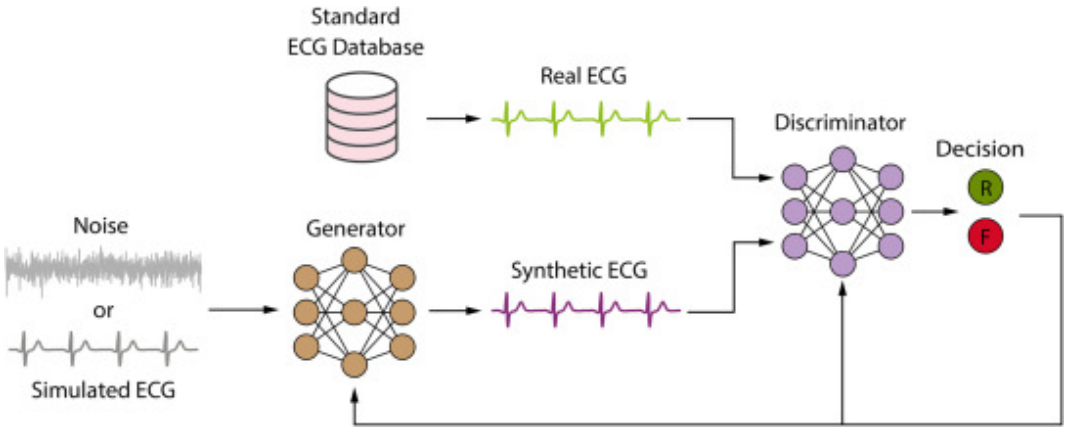


FIGURE 4.11. Functional Diagram of a GAN model

Apart from the mentioned advantages, the highest value of using a GAN is the strong attempt to safeguard the original data and its characteristics intact, preserving its inherit nature [28]. This is crucial to assure the data integrity and reliability when comparing synthetic data to the real one.

Concerning the GAN's implementation, it has taken place after consulting the features histograms to assess which variables make a bigger difference in terms of value range between Atrial and Non-Atrial ECGs records. The distributions reveal the features belonging to lead II, the PQ segment measure, the P-wave amplitude along with the age and gender of the patient have the strongest class-separating potential. Hence, these selected features are considered for the GAN implementation and training. Considering the ratio between majority and minority representation, the GAN generates in total 2 956 synthetic samples (the double of the original amount of minority class records) during the training process which takes 2000 epochs to complete.

The amount of synthetic instances + the real minority class registers represent 9.27% of the total dataset. That being said, 2 956 new synthetic samples seem an adequate number to impute in the minority class. Firstly, the imbalance remains and it is considerable; secondly, given the disparity, duplicate the minority instances is enough to provide a more robust basis for classifier training.

The data split implemented in the baseline experiences - Scenario 1 - is the same as in this one - Scenario 2. Table 4.5 indicates the data distribution over the training and test sets. In total, there are 4 434 samples from the minority class. Therefore, for the Scenario 2 experiments, the total amount increases to 47 735.

Table 4.5. Data split summary - Data with GAN samples

	Non-Atrial Flutter	Atrial Flutter	Total
Training Set	34 641	3 547	38 188
Test Set	8 660	887	9 547
Total	43 301	4 434	47 735

CHAPTER 5

Results

This chapter presents the outcomes obtained during the experiments phase, including both dataset setups: Scenarios 1 and 2. On one hand, Scenario 1 refers to experiments using the original data, whereas Scenario 2 employs the data plus synthetic samples generated by the GAN. The primary objective is to evaluate the performance of the proposed methods under both contexts and determine whether the inclusion of GAN-generated data positively influences the classifier's behavior and overall performance.

The chapter is structured into 6 main sections. Each section corresponds to one of the multiple algorithms tested during the modeling phase.Therefore, each section is named after the three standard sampling methods - Undersampling, Oversampling and SMOTE-Tomek, the Isolation Forest, and the two AL methods, CBAL and FIMPS.

5.1. Undersampling

Scenario 1

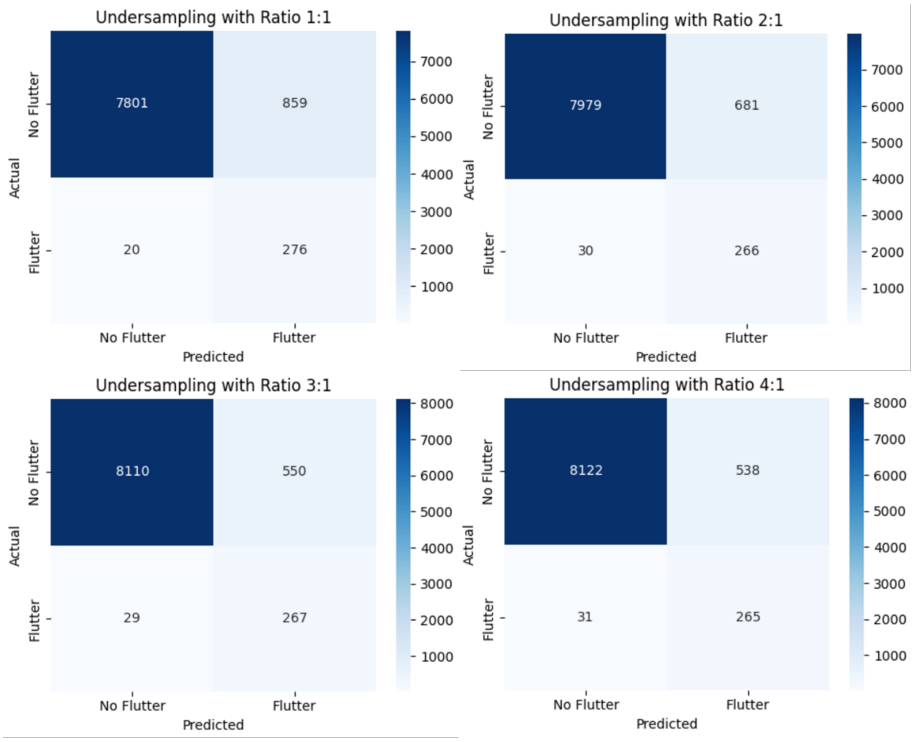


FIGURE 5.1. Scenario 1 - Undersampling Confusion Matrix

Figure 5.1 starts to illustrate how undersampling influences the model's ability to correctly classify Atrial Flutter instances, but at the time struggles to control false positives. For instance, at 1:1 ratio, the model achieves a high recall of 0.9324, successfully identifying almost all AFL cases - 276 true positives out of 296 minority instances. However, this comes at the cost of a relatively large number of false positives - 859, which lowers precision to 0.2432.

In ratio 2:1 configuration, the false positives slightly increase to 30 instances, but the false negatives have a considerable reduction to 681 when comparing to the 1:1 ratio. In this case, the recall presents a score of 0.8986 and both precision and F1-Sore increase to 0.2809 and 0.4280, respectively. These scores suggest that the configuration 2:1 provides a more balanced trade-off between sensitivity and specificity, as shown by the improvement in F1-Score. Moreover, the 3:1 ratio presents a F1-score of 0.4798 and precision at 0.3268.

Lastly, the 4:1 ratio presents the best metrics of all configurations, except for recall (0.8953). Although the 4:4 ratio misclassifies 31 samples as false positives, it is the setup presenting the most balanced results (Table 5.1).

TABLE 5.1. Scenario 1 - Undersampling Performance Metrics per Ratio

Ratio / Metric	Accuracy	Precision	Recall	F1-Score	ROC-AUC
1:1	0.9019	0.2432	**0.9324**	0.3857	0.9166
2:1	0.9206	0.2809	0.8986	0.4280	0.9100
3:1	0.9353	0.3268	0.9020	0.4798	**0.9193**
4:1	**0.9365**	**0.3300**	0.8953	**0.4823**	0.9166

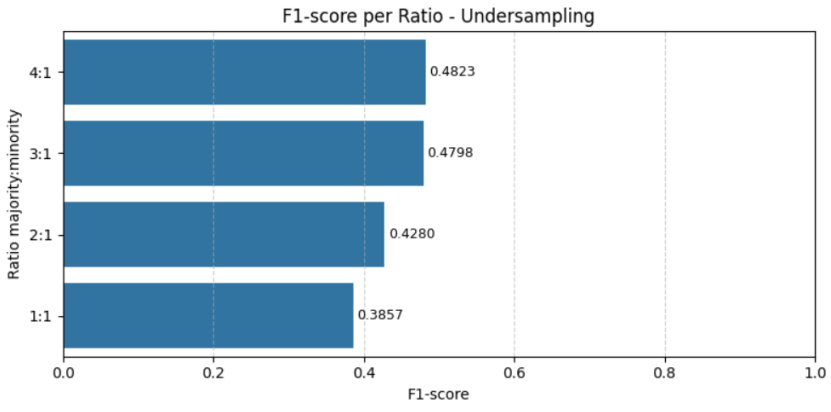


FIGURE 5.2. Scenario 1 - F1-Score in Undersampling

Overall, it is fair to say, according to Figure 5.2, that the best ratio setup considering the best trade-off between sensitivity and specificity is 4:1, with the highest F1-score of 0.4823. The model performs better with the highest imbalance tested.

Scenario 2

Looking at the Scenario 2, with the GAN-produced synthetic samples, the model demonstrates a strong improvement in recall with a score of 0.9628, compared to the same experiment employed in Scenario 1, which was 0.9324. Besides, this configuration correctly identifies 854 out of 887 Atrial Flutter cases, even though the number of false positives remains considerable - 523 records. On the other hand, precision and F1-score are substantially higher than the homonymous experiment in Scenario 1, now presenting 0.6202 and 0.7544 scores, respectively, instead of 0.2432 and 0.3857 (Figure 5.3).

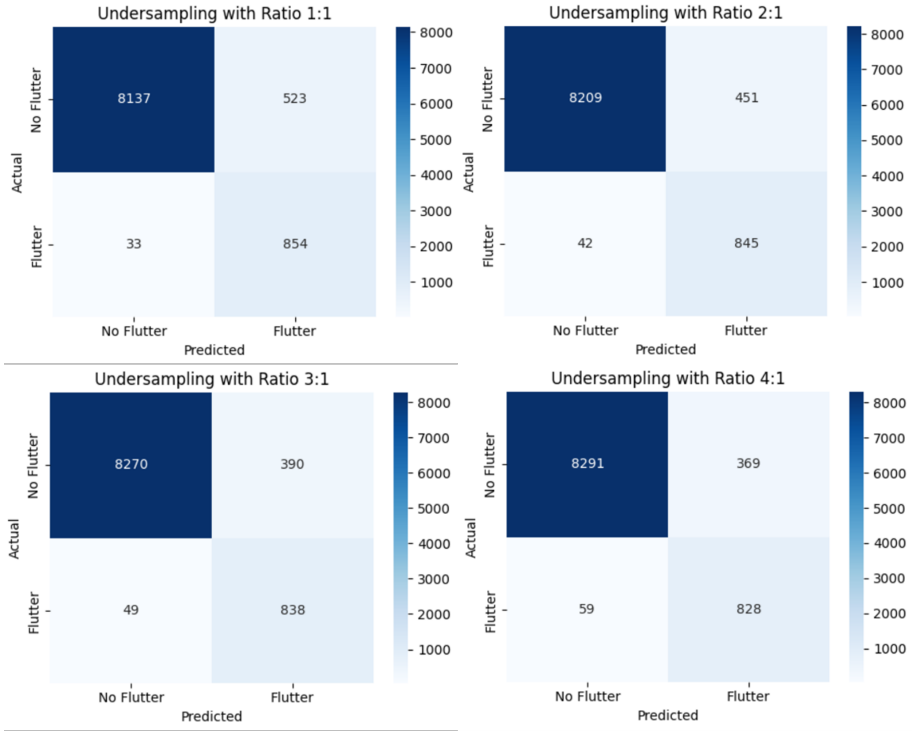


FIGURE 5.3. Scenario 2 - Confusion Matrix of Undersampling

As for the 2:1 ratio configuration, the classifier maintains a very high recall of 0.9526 while reducing false positives to 451, leading to a higher precision (0.6520) and F1-score (0.7742). The outcome suggests that the model benefits from the additional synthetic data, as it can recognize minority cases more confidently without overly compromising specificity (Table 5.2).

TABLE 5.2. Scenario 2 - Undersampling Performance Metrics per Ratio

Ratio / Metric	Accuracy	Precision	Recall	F1-Score	ROC-AUC
1:1	0.9418	0.6202	**0.9628**	0.7544	**0.9512**
2:1	0.9484	0.6520	0.9526	0.7742	0.9503
3:1	0.9540	0.6824	0.9448	0.7924	0.9499
4:1	**0.9552**	**0.6917**	0.9335	**0.7946**	0.9454

The trend continues at 3:1 ratio, with a steady decrease in false positives - 390 instances - and a minor reduction in recall, 0.9448. While both precision (0.6824) and F1-score (0.7924) reach their highest values so far, accuracy (0.9540) remains consistently strong. This configuration therefore provides the most balanced performance, achieving an optimal trade-off between identifying *Flutter* cases and limiting false alarms.

Finally, the 4:1 ratio slightly increases the number of false negatives to 59, while further reducing false positives to 369. Although recall decreases marginally to 0.9335, precision improves to 0.697 and F1-score presents a score of 0.7946, remaining comparable to the 3:1 ratio. The metric ROC-AUC, 0.9554, confirms a stable separation between classes, across all configurations, indicating robust model behavior even under different sampling intensities (Figure 5.4).

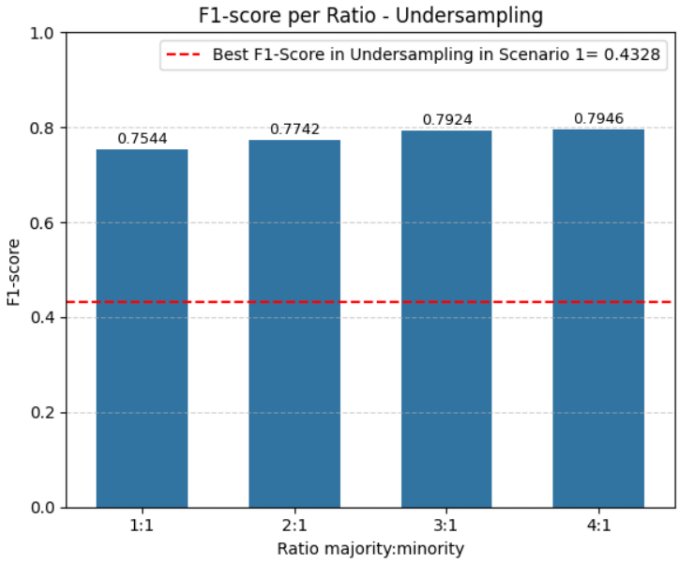


FIGURE 5.4. Scenario 2 - F1-Score in Undersampling

Overall, based only on this first experience, it can be said, the use of GAN-generated synthetic data is already noticeable compared with experiments which only use the original data. As Figure 5.4 outlines, all F1-Score values in this experience surpass the best result obtained in the homonymous experiment in Scenario 1 (0.4328). This means, the model deals better when given more information to the training process.

5.2. Oversampling

Scenario 1

Figure 5.5 presents the results from oversampling techniques over the four ratios and the numbers outline an improvement from the previous method (see Figure 5.6).

Starting with ratio 1:1, the model achieves the highest recall of 0.7061 by successfully identifying most Atrial Flutter instances - 209 true positives out of 296 minority samples. Even though the 272 false positives reduce precision to 0.4345, resulting in a moderate F1-Score of 0.5380, the ROC-AUC metric indicates reasonable class separability.

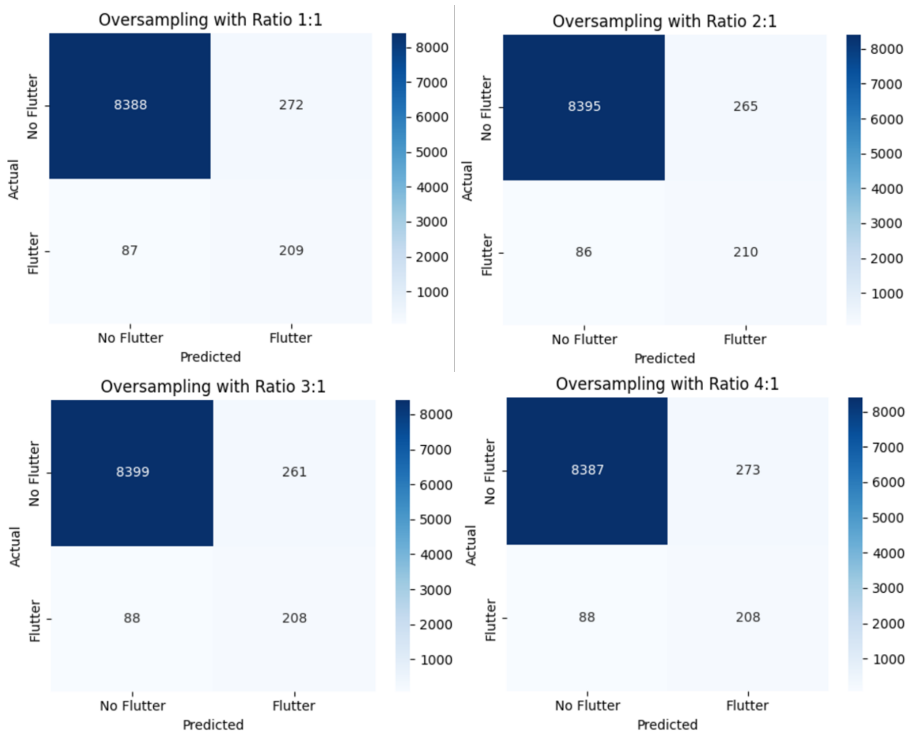


FIGURE 5.5. Scenario 1 - Oversampling Confusion Matrix

Regarding the 2:1 ratio, the performance slightly improves both in precision (0.4421) and F1-Score (0.5447) while maintaining recall at 0.7095. The reduction in false positives to 265 instances suggests that this particular configuration enhances the classifier's discriminative ability without compromising its sensitivity to minority cases (is better than 1:1 ratio) (Table 5.3).

Comparatively, at the 3:1 ratio, the number of false positives decreases even more to 261 cases, which maintains recall stable at 0.7027. Although the performance gains are modest, both accuracy (0.9610) and precision (0.4435) reach their highest values, indicating the best overall balance between detecting positive cases and limiting misclassifications of the majority class.

TABLE 5.3. Scenario 1 - Oversampling Performance Metrics per Ratio

Ratio / Metric	Accuracy	Precision	Recall	F1-Score	ROC-AUC
1:1	0.9599	0.4345	0.7061	0.5380	0.8373
2:1	0.9608	0.4421	**0.7095**	**0.5447**	**0.8394**
3:1	**0.9610**	**0.4435**	0.7027	0.5438	0.8363
4:1	0.9597	0.4324	0.7027	0.5354	0.8356

Ratio 4:1 shows a slight deterioration in performance across all metrics. For instance, precision drops to 0.4324 and F1-score decreases to 0.5354, suggesting that oversampling beyond the 3:1 proportion introduces redundancy without actual improvements.

This time, the 3:1 ratio presents the most balanced and consistent results, achieving the best combination of accuracy, precision and recall, as well as a stable ROC-AUC score value. However, it is important not to forget that in contexts of imbalanced data, due to the unevenness in representation, accuracy tends to be higher but it does not necessarily mean the model is learning better. It mirrors the consequence of a higher representativeness of one class over the other, i.e., majority vs minority class sizes.

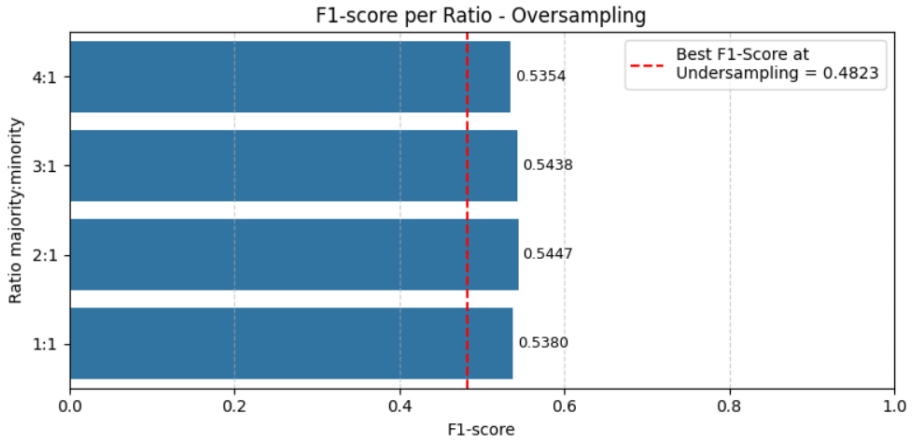


FIGURE 5.6. Scenario 1 - F1-Score in Oversampling

Moreover, when compared with the undersampling approach, it can be said the oversampling strategy performs better. As Figure 5.6 shows, the best F1-Score encountered in undersampling in Scenario 1 experiments (0.4823) is lower than all F1 values when employing the oversampling, meaning the model adapts better to data augmentation rather than cutting it.

Scenario 2

In ratio 1:1, the model achieves the highest recall (0.8985) among the four configurations, correctly identifying 797 true positives out of 887 Atrial Flutter instances - it improves comparing to Scenario 1 where there are 209 true positives from a total of 296 AFL cases. However, despite this strong sensitivity, the relatively higher number of false positives (289) reduces precision to 0.7339, resulting in an F1-score of 0.8079. The ROC-AUC of 0.93 indicates solid class separability, though the trade-off between sensitivity and specificity remains evident (Figure 5.7).

In the 2:1 ratio configuration, both precision (0.7406) and F1-score (0.8092) show slight improvements while recall (0.8918) remains high. The reduction

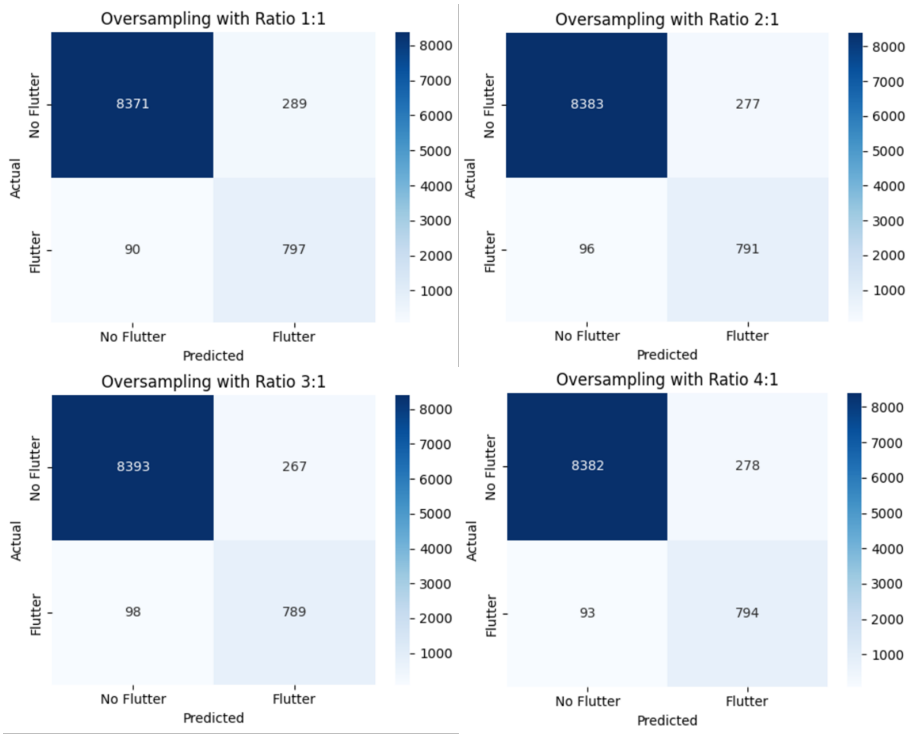


FIGURE 5.7. Scenario 2 - Confusion Matrix of Oversampling

in false positives to 277 suggests that this oversampling level helps the classifier better discriminate between classes without substantially compromising its ability to detect minority cases.

Moving to the 3:1 ratio, the model achieves its best overall performance, with accuracy rising to 0.9618, precision increasing to 0.7472, and a F1-score of 0.8121. On the other hand, recall slightly decreases to 0.8895. The continued decrease in false positives (267) indicates a more stable and balanced detection capability between the positive and negative classes (Table 5.4).

TABLE 5.4. Scenario 2 - Oversampling Performance Metrics per Ratio

Ratio / Metric	Accuracy	Precision	Recall	F1-Score	ROC-AUC
1:1	0.9603	0.7339	**0.8985**	0.8079	**0.9326**
2:1	0.9609	0.7406	0.8918	0.8092	0.9299
3:1	**0.9618**	**0.7472**	0.8895	**0.8121**	0.9293
4:1	0.9611	0.7407	0.8952	0.8106	0.9315

In contrast, the 4:1 ratio configuration shows a minor decline in most metrics. While recall (0.8952) remains comparable to earlier configurations, precision (0.7407) and F1-score (0.8106) exhibit no further gains and ROC-AUC (0.9315) remains steady. This suggests that oversampling beyond 3:1 introduces redundancy without offering meaningful performance improvements.

Overall, the 3:1 ratio delivers the most balanced and consistent performance, achieving the best trade-off between precision, recall, and accuracy,

while maintaining a stable ROC-AUC value (nearly 0.93). Nonetheless, accuracy should still be interpreted with caution, as its relatively high value partly reflects the model's success in correctly classifying numerous majority-class samples, potentially masking the impact of remaining false negatives.

Once more, the overall performance of the model with GAN-generated synthetic data surpasses the results obtained in the same experience but in Scenario 1, specially in Recall, with an increase of 0.19 and F1-score improving more than 0.26. As Figure 5.8 illustrates, all F1-Score of Oversampling under these conditions (Scenario 2) surpass the best result F1-Score encountered in Scenario 1.

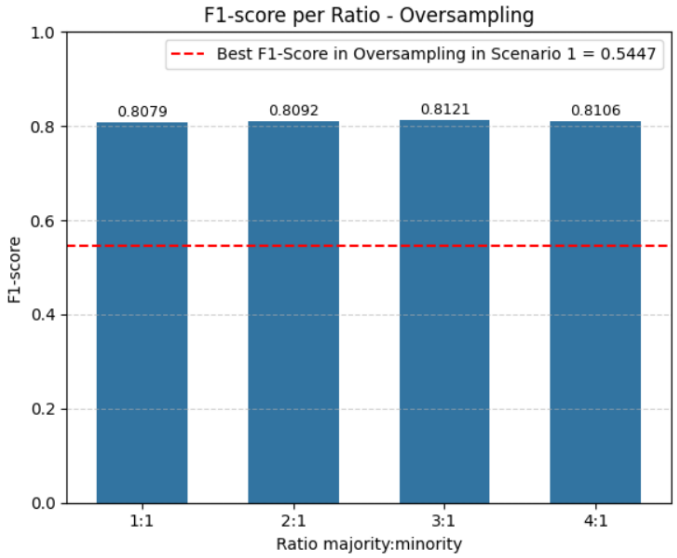


FIGURE 5.8. Scenario 2 - F1-Score in Oversampling

5.3. SMOTE-Tomek

Scenario 1

Instead of employing the traditional commonly used SMOTE method, the hybrid approach of SMOTE-Tomek is implemented. This technique combines oversampling associated to original SMOTE with the removal of Tomek links to eliminate borderline or ambiguous samples. It aims to balance the dataset while simultaneously improving the quality of the decision boundary between classes.

As Figure 5.9 illustrates , across all ratios, the classifier exhibits a strong and consistent performance, with small fluctuations in the evaluation metrics (Table 5.5). In ratio 1:1 ratio, for instance, the model achieves a recall of 0.6993, successfully identifying most Atrial Flutter cases, though this sensitivity came with a lower precision of 0.4148 due to an increased number of 292 false positives. The resulting F1-Score equals to 0.5208 and ROC-AUC of

0.8328 indicate a well-balanced and effective classification performance (Table 5.5).

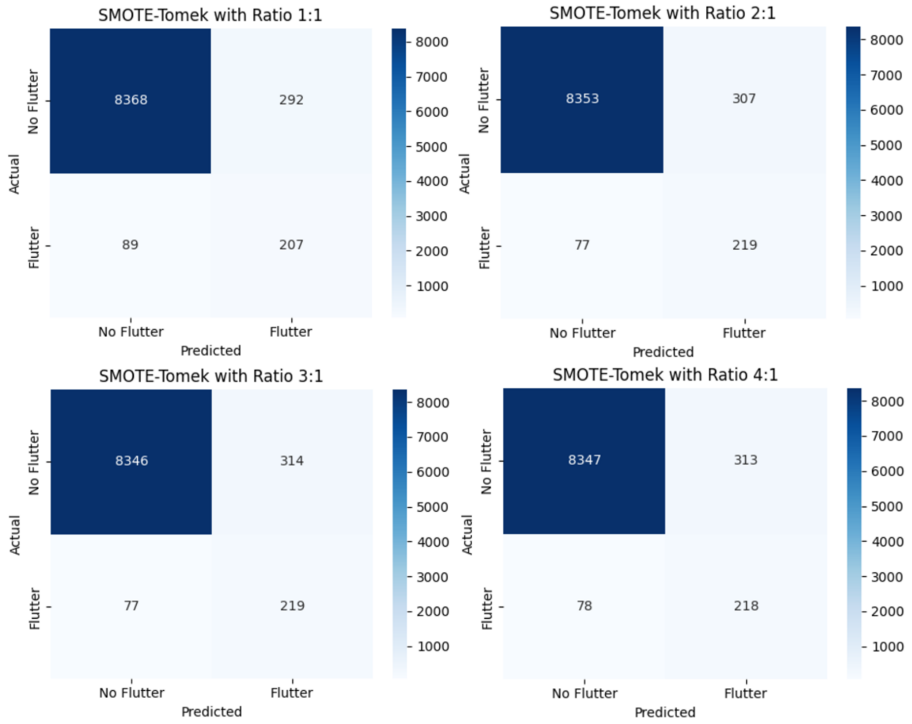


FIGURE 5.9. Scenario 1 - SMOTE-Tomek Confusion Matrix

Looking at the 2:1 configuration, it revealed a noticeable improvement in both recall (0.7399) and F1-score (0.5328) while maintaining a similar level of precision with a score of 0.4164. This suggests that a modest increase in majority-class representation helps to stabilize the model's decision boundary without compromising its ability to detect minority instances.

At 3:1 ratio, performance remains consistent, with recall and F1-score analogous to the previous configuration, 0.7399 and 0.5283, respectively. Although precision decreases a little bit, accuracy shows a value of 0.9563 and a high ROC-AUC as well, 0.8518, which may indicate a reliable global separability between the two classes. The 4:1 ratio maintains similar behavior, confirming that further increases in sampling proportion yield negligible improvements in classification quality.

TABLE 5.5. Scenario 1 - SMOTE-Tomek Performance Metrics per Ratio

Ratio / Metric	Accuracy	Precision	Recall	F1-Score	ROC-AUC
1:1	**0.9575**	0.4148	0.6993	0.5208	0.8328
2:1	0.9571	**0.4164**	**0.7399**	**0.5328**	**0.8522**
3:1	0.9563	0.4109	**0.7399**	0.5283	0.8518
4:1	0.9563	0.4105	0.7365	0.5272	0.8502

In general, the results demonstrate that the SMOTE-Tomek approach effectively mitigates class imbalance and produces stable and reliable outcomes

across all configurations. However, the F1-Scores compared to the best of Oversampling Scenario 1 fall short, as all of them fall behind the 0.5447 mark (Figure 5.10). Nevertheless, separately speaking, the SMOTE-Tomek 3:1 ratio reveals to have the most consisting results (highest F1-Score, 0.5328) among all other configurations.

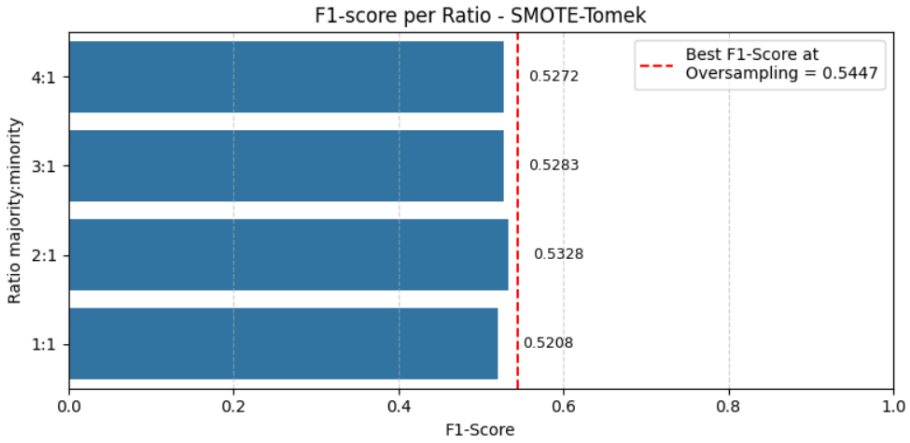


FIGURE 5.10. Scenario 1 - F1-Score in SMOTE-Tomek

Scenario 2

The application of the SMOTE-Tomek technique resulted in consistently strong classification performance across all oversampling ratios, maintaining a solid balance between precision and recall. As shown in Figure 5.11 and summarized in Table 5.6, this hybrid approach continues to enhance minority representation while effectively minimizing overlap between classes.

At the 1:1 ratio, the classifier achieves one of its best performances, with a recall of 0.8918 and a precision of 0.7635. The model correctly identifies 791 true Atrial Flutter cases, while producing 245 false positives. The resulting F1-score (0.8227) and ROC-AUC (0.9317) indicate a strong and balanced trade-off between sensitivity and specificity, confirming that this configuration captures the majority of positive instances with limited noise.

In the 2:1 ratio, recall further increases to 0.9087, the highest among all configurations, while maintaining robust precision (0.7449). This setup also achieves the highest ROC-AUC (0.9384), reflecting superior class separability and strong generalization. Although false positives increase slightly (276), this configuration effectively reduces false negatives (81), yielding the best sensitivity overall and demonstrating the model's ability to identify the greatest number of Atrial Flutter cases.

For the 3:1 ratio, performance remains stable, though both precision (0.7298) and recall (0.9042) experience a slight decrease.; in Scenario 1 the value stands at 0.7399. The F1-score (0.8077) suggests that while the classifier continues

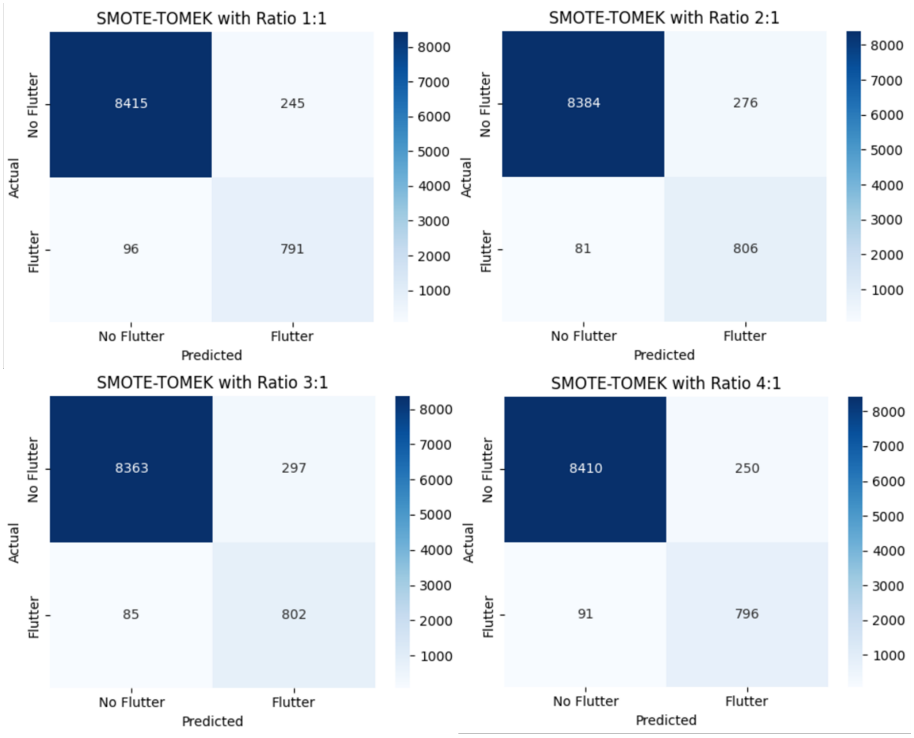


FIGURE 5.11. Scenario 2 - Confusion Matrix of SMOTE-Tomek

to perform well, additional oversampling begins to yield diminishing returns, with little to no improvement in class discrimination despite higher computational effort. The moderate increase in false positives (297) further supports this observation.

TABLE 5.6. Scenario 2 - SMOTE-Tomek Performance Metrics per Ratio

Ratio / Metric	Accuracy	Precision	Recall	F1-Score	ROC-AUC
1:1	**0.9643**	**0.7635**	0.8918	0.8227	0.9317
2:1	0.9626	0.7449	**0.9087**	0.8187	**0.9384**
3:1	0.9600	0.7298	0.9042	0.8077	0.9350
4:1	**0.9643**	0.7610	0.8974	**0.8236**	0.9343

Finally, the 4:1 ratio restores performance to levels comparable with the 1:1 configuration, achieving precision of 0.7610, recall of 0.8974 and the highest F1-score (0.8236) across all ratios. The relatively low number of false positives (250) reinforces the stability and consistency of this configuration. However, it is important to note that while both the 1:1 and 4:1 ratios deliver top-level metrics, the 1:1 setup represents an artificial class balance not reflective of the real data distribution, whereas the 4:1 configuration, though more imbalanced, offers a more realistic representation of the dataset's underlying structure.

Overall, the 2:1 ratio emerges as a well-balanced compromise between false positives and false negatives. Despite a modest increase in false positives, it achieves the lowest number of false negatives and the highest recall, marking

it as the most sensitive configuration for detecting Atrial Flutter cases without a substantial loss in precision.

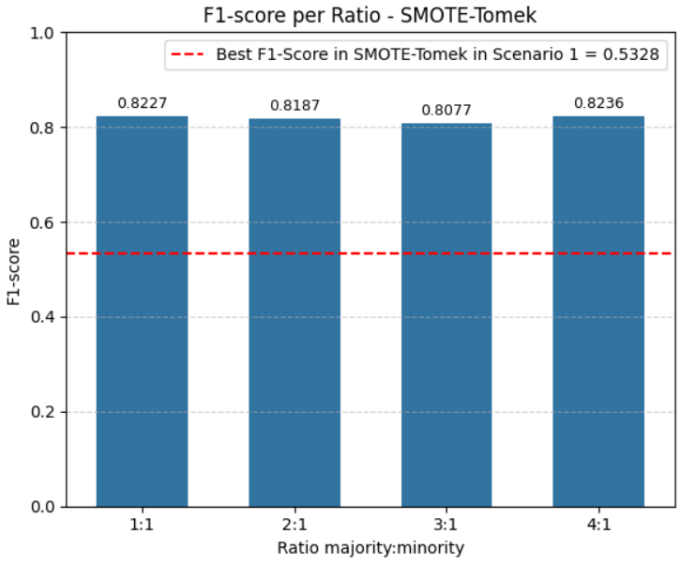


FIGURE 5.12. Scenario 2 - F1-Score in SMOTE-Tomek

The tendency repeats itself, as all F1-Score values are higher than the biggest F1-Score from SMOTE-Tomek Experiment in Scenario 1. It is unanimous, concerning the traditional sampling methods, Scenario 2 setup proves to perform better than Scenario 1.

5.4. Isolation Forest

Scenario 1

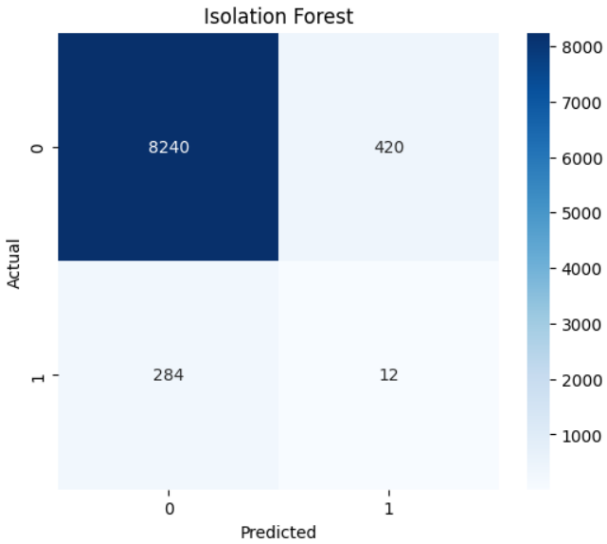


FIGURE 5.13. Scenario 1 - Isolation Forest Confusion Matrix

The Isolation Forest algorithm is employed at the suggestion of the supervisors and it consists of an unsupervised anomaly detection approach; in this

case, to identify Atrial Flutter cases by modeling them as statistical outliers. The method isolates observations by recursively dividing the data through randomly selected features and split values. Instances that can be isolated in fewer random partitions are considered more anomalous. This makes the algorithm particularly effective when anomalies are rare and clearly distinct from the majority of the data [29].

During the experiments, the model correctly identified 12 true Atrial Flutter cases and 8 240 true negatives, while misclassifying 284 false negatives and 420 false positives (Figure 5.13).

Although the overall accuracy appears high, this value is misleading due to the strong class imbalance. The very low precision (0.0278) and recall of 0.0405, indicate that the model was largely unable to distinguish Atrial Flutter instances from non-Flutter cases. The ROC-AUC close to 0.5 further confirms that the algorithm's discriminative ability is equivalent to random guessing (Figure 5.14).

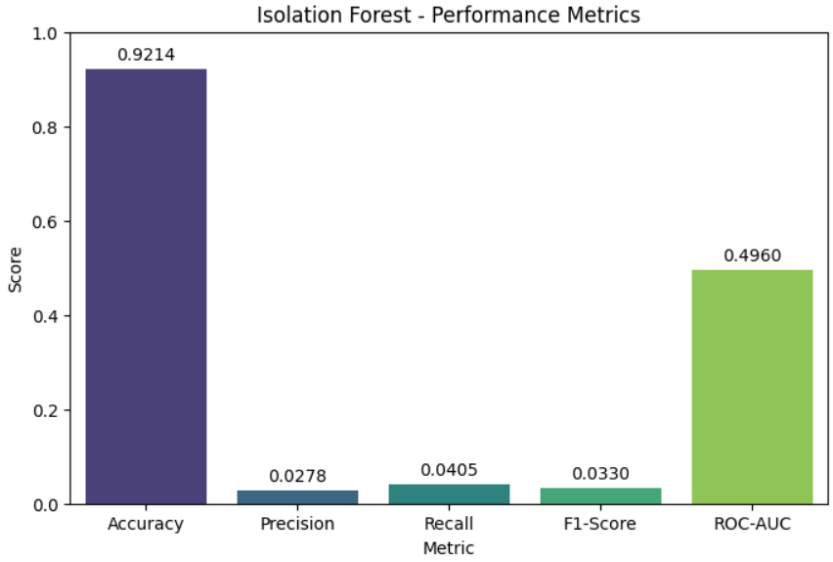


FIGURE 5.14. Scenario 1 - Isolation Forest Performance Metrics

The weak performance can be explained by the nature of the data and the composition of the majority class. Since all samples correspond to ECG signals, the physiological characteristics of the heart activity are inherently similar across patients. Moreover, the majority class aggregates a wide range of cardiac conditions—such as Atrial Fibrillation and other arrhythmias—while the minority class only includes AFL cases. This heterogeneity within the majority class introduces overlapping feature distributions, preventing the Isolation Forest from identifying Atrial Flutter as a distinct anomaly.

In summary, the Isolation Forest failed to capture clear separation between the classes because Atrial Flutter is not an extreme outlier, but rather a minority condition embedded within a spectrum of other pathological patterns.

These findings suggest that anomaly detection techniques are not well suited for this type of medical data, where both classes represent abnormal cardiac behaviors with subtle physiological differences. Supervised or semi-supervised approaches specifically designed for imbalanced learning are therefore more appropriate in this context.

Scenario 2

As Figure 5.15 illustrates, the model correctly identifies 227 true positives, but at the same time misclassifies 660 instances as false negatives, indicating that a considerable number of Atrial Flutter cases remain undetected. Comparing to Scenario 1 which presents a significant percentage of 95.94% false negatives, in Scenario 2 the indicator decreases to 74.4%, which means that with the increase of AFL cases, the classifier improves its capability of distinguish between positive cases.

This limitation is reflected in the low recall with a score of 0.2285, meaning the model captures only a small portion of the actual positive cases (227 out of 887). On the other hand, the precision value, 0.3690, shows that over 1/3 of the instances predicted as positive are indeed Atrial Flutter, suggesting a moderate level of confidence in its positive classifications (Figure 5.16). Besides that, the ROC-AUC metric presents a rather modest score of 0.5942.

Despite the restricted recall, the overall accuracy (0.8920) demonstrates that the Isolation Forest performs well in distinguishing between the two classes from a global perspective. However, its focus on identifying extreme outliers rather than capturing broader minority patterns likely causes it to overlook many valid positive instances, leading to a lower F1-score, presenting a score of 0.2822 (see Figure 5.16).

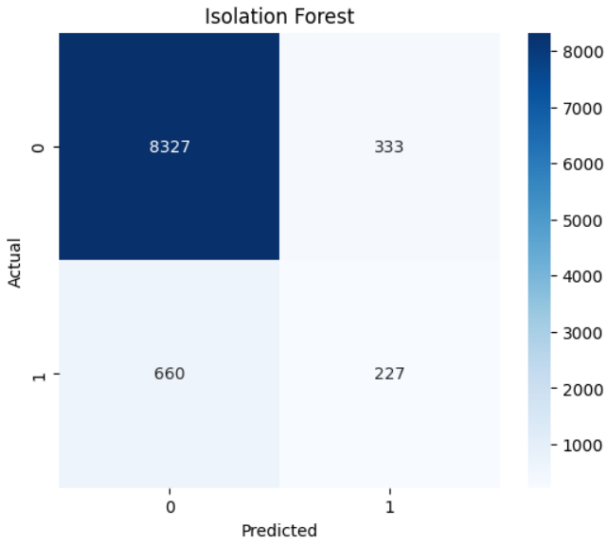


FIGURE 5.15. Scenario 2 - Confusion Matrix of Isolation Forest

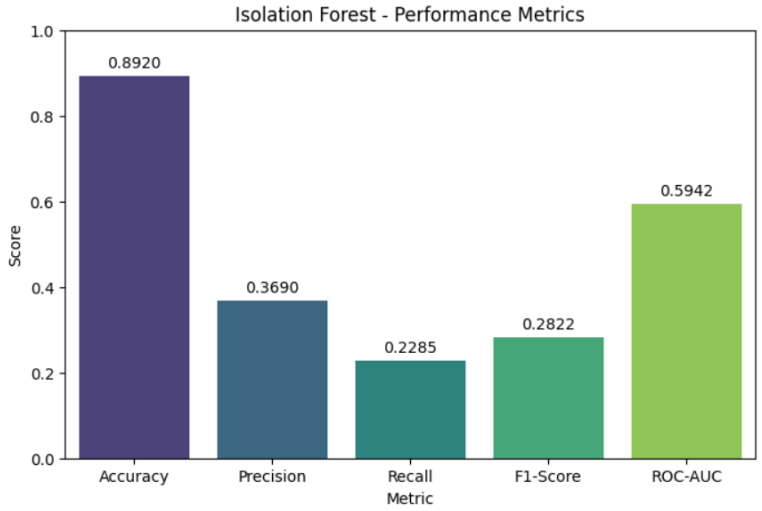


FIGURE 5.16. Scenario 2 - Performance Metrics of Isolation Forest

In summary, the Isolation Forest displays a weak overall separability with limited sensitivity to minority-class samples. While effective at flagging highly distinctive anomalies, it struggles to detect more subtle manifestations of Atrial Flutter. These results highlight the inherent trade-off of anomaly detection models in imbalanced classification tasks, as they offer robustness and simplicity but may underperform when the minority class exhibits complex or overlapping distributions, which is the case.

5.5. CBAL

Scenario 1

As said previously, this algorithm selects instances for labeling based on the model's confidence, prioritizing samples that are located near the decision boundary where classification uncertainty is higher.

As the confusion matrix shows in Figure 5.17, the model correctly identifies 196 true positive cases of Atrial Flutter and misclassifies 100 false negatives. Well, this ratio of 100/296 minority cases wrongly labeled corresponds to 33.78% of the total amount of AFL instances in the test set, a considerable number. Moreover, the classifier misclassifies 1 118 as false positives.

This pattern results in a recall of 0.6622, indicating that the model successfully recognized approximately 2/3 of the minority-class samples. However, the large number of false positives significantly impacts precision with a score of 0.1492, suggesting that many of the instances predicted as positive do not actually correspond to Atrial Flutter.

The overall accuracy 0.8640 remains relatively high due to the prominence of the majority class, while the F1-score (0.2435) reflects the imbalance between precision and recall. The ROC-AUC value (0.8968) indicates strong class

separability from a probabilistic standpoint, showing that the model can still distinguish between positive and negative instances even if its final threshold leads to misclassifications, which indeed occurs (Figure 5.18).

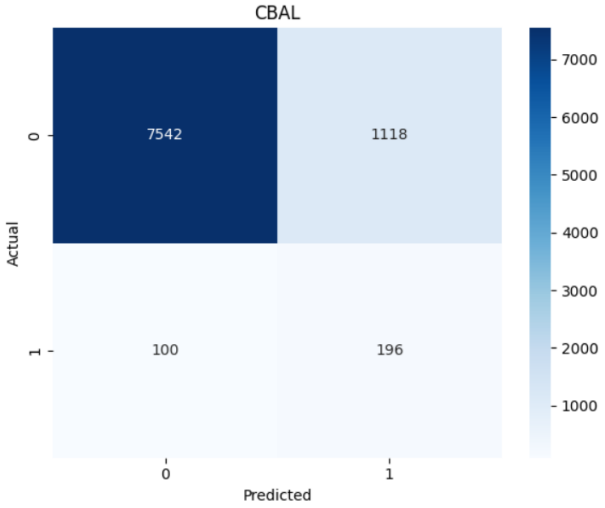


FIGURE 5.17. Scenario 1 - CBAL Confusion Matrix

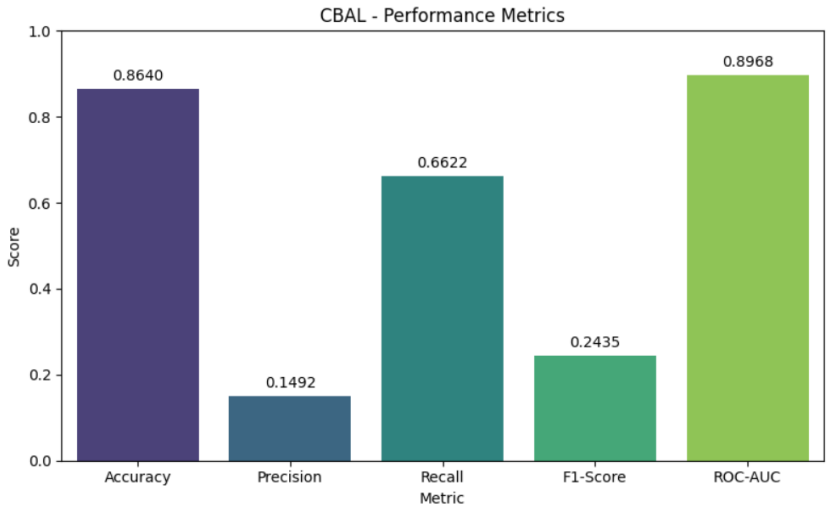


FIGURE 5.18. Scenario 1 - CBAL Performance Metrics

In summary, CBAL demonstrates good sensitivity but limited precision when trained on the original data. Although the algorithm effectively detects minority instances, the imbalance of the dataset prevents its ability to maintain reliable positive predictions. These results highlight the typical trade-off in active learning under severe class imbalance - improved recall achieved at the expense of a higher false-positive rate. Nonetheless, the overall evaluation metrics outline a more realist and reliable set of values than the traditional sampling methods.

Scenario 2

The implementation of the Active Learning certainty-based algorithm revealed better results in general in Scenario 2 than in the first experiment. As the confusion matrix show, in Figure 5.19, the classifier is able to detect 697 out of 887 Atrial Flutter cases, while misclassifying 190 instances - the false negatives, achieving an optimistic percentage of 21.4% false positives. In Scenario 1, the percentage stands at 33.8%.

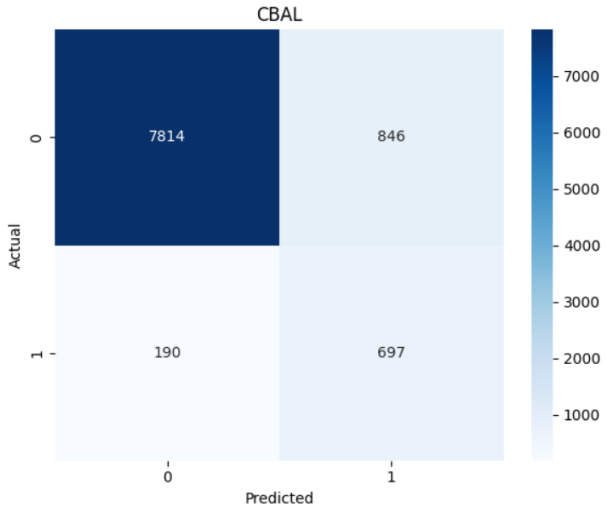


FIGURE 5.19. Scenario 2 - Confusion Matrix of CBAL

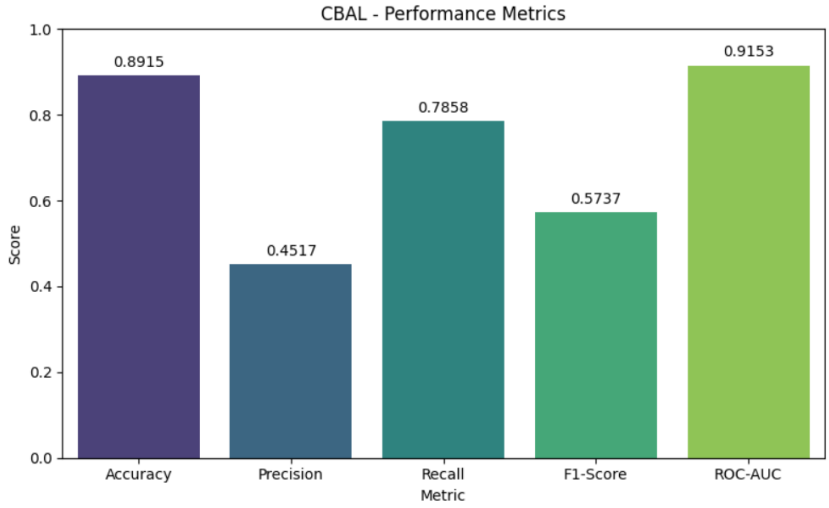


FIGURE 5.20. Scenario 2 - Performance Metrics of CBAL

Regarding the performance metrics, recall achieves an optimistic score of 0.7858 which indicates that the model captures the majority of Atrial Flutter records. However, the high number of false positives - 846 cases - reflects a tendency to overestimate the positive class, which impacts precision (decreases a bit comparing to other methods). The F1-score, 0.5737, highlights a moderate balance between precision and recall, meanwhile, the ROC-AUC, with

0.9153 score outlines a strong class separability and confirms that the CBAL strategy successfully enhances the model's discriminative capacity compared to baseline configurations (Figure 5.20).

In general, results demonstrate that CBAL effectively mitigates class imbalance by prioritizing uncertain and potentially informative instances during training. The algorithm achieves high recall - improves from 0.6622 in Scenario 1 to 0.7858 and strong overall separability, though at the cost of increased false positives, which remains a problem, as well. This reflects a common trade-off in active learning for imbalanced datasets, as improved minority class cases detection often comes with reduced precision, highlighting the need for further fine-tuning or hybridization with balancing techniques to achieve optimal performance. Nonetheless, the progress from Scenario 1 to Scenario 2 in terms of CBAL effectiveness is noted, proving the GAN use was not in vain.

5.6. FIMPS

Scenario 1

At last, the FIM-Based Pairwise Selection algorithm is applied to assess its effectiveness in improving minority-class recognition through a pairwise sampling strategy.

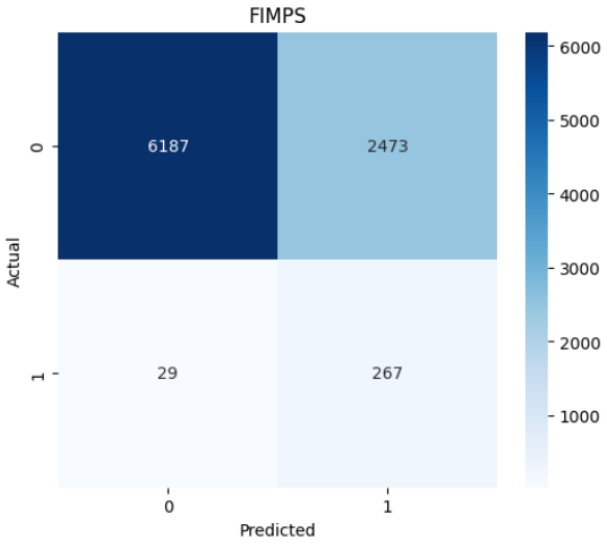


FIGURE 5.21. Scenario 1 - FIMPS Confusion Matrix

As shown in Figure 5.21 , the model correctly identifies 267 Atrial Flutter cases out of 296, presenting the second lowest number of false negatives (29), after the undersampling 2:2 ratio. Effectively, this result is reflected in the evaluation metrics, presented in Figure 5.22, where recall achieves a high score of 0.9020. On the other hand, due to the considerable number of false positives, i.e., 2473 instances, the precision is inevitably low (0.0974), which is also verified through the modest F1-Score of 0.1759.

Despite the large number of false positives, the ROC-AUC value (0.8462) suggests that the model still retains a reasonable degree of discriminative labeling. However, the accuracy (0.7206) confirms that the overestimation of positive cases negatively impacts global performance.

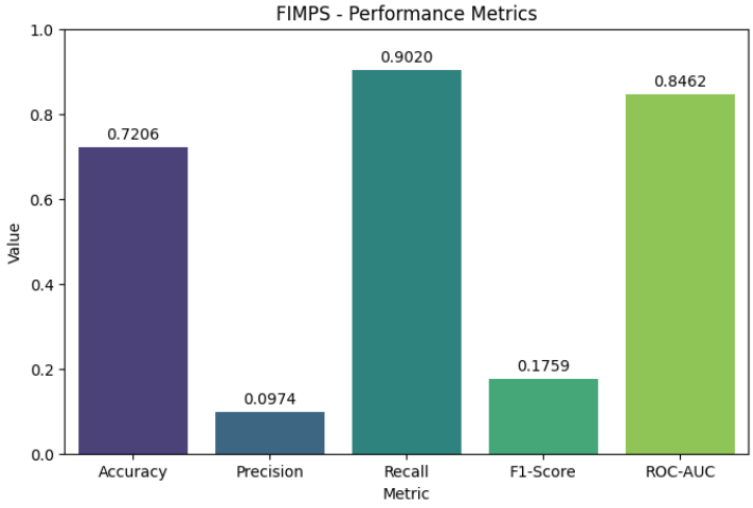


FIGURE 5.22. Scenario 1 - FIMPS Performance Metrics

In summary, the results reveal that FIMPS prioritizes sensitivity over precision, successfully detecting almost all Atrial Flutter instances but producing many false alarms. This behavior aligns with the algorithm's objective of emphasizing minority detection through uncertainty-based selection, though it also exposes its tendency to overgeneralize the positive class in highly imbalanced datasets. While effective for capturing rare events, further tuning or integration with balancing techniques would be necessary to enhance its precision and overall stability.

Scenario 2

The performance metrics in Figure 5.24 reveals that the model achieves a very high recall (0.9154) and slightly higher than in Scenario 1 (0.9020), correctly identifying almost all Atrial Flutter cases (812 true positives out of 887).

It comes, nonetheless, at the cost of a substantial number of false positives - 2 410, which sharply reduces precision, with the lowest score encountered from all experiments: 0.2520. The difference between sensitivity and specificity is reflected in the F1-score, suggesting that although the model is highly sensitive to the minority class, its ability to correctly distinguish positive from negative cases is limited (Figure 5.24). This can be explained due to the similarity between some records which may be confused as AFL, since there are some which belong to the majority class that have AFL and other associated heart conditions.

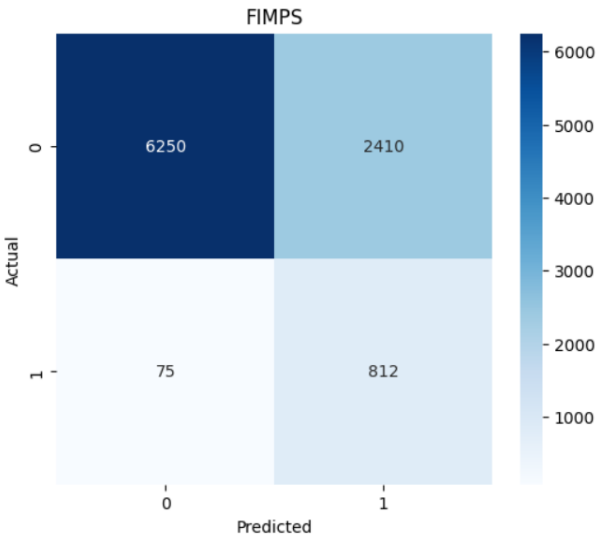


FIGURE 5.23. Scenario 2 - Confusion Matrix of FIMPS

The score 0.7397 in accuracy is considerably lower than in previous models, primarily due to the large number of false alarms, even though the ROC-AUC presents a high score, 0.8887. These results imply that while FIMPS fulfills its objective of maximizing detection of minority instances, it tends to overgeneralize the positive class, leading to reduced overall classification precision. Nevertheless, the low accuracy reflects the expected reality present in class imbalance contexts, because until now all experiments have been unrealistically presenting values that are too high, due to the corrected classified majority class instances.

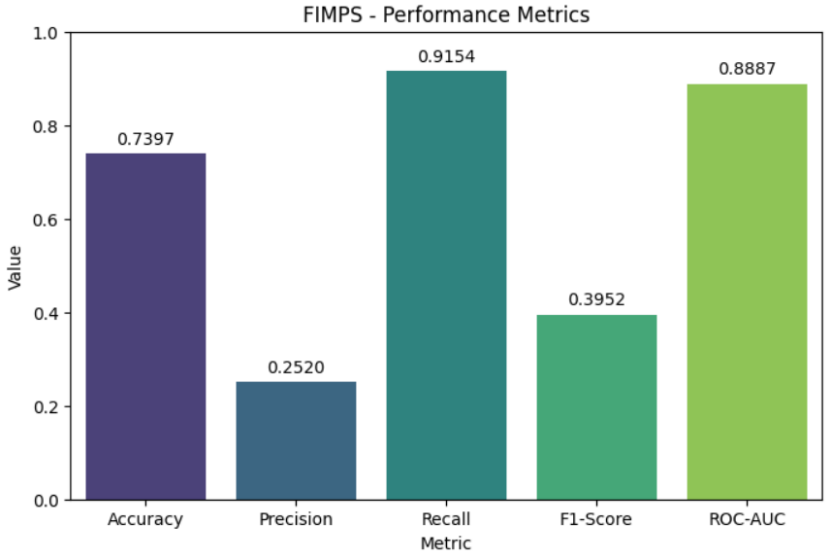


FIGURE 5.24. Scenario 2 - Performance Metrics of FIMPS

CHAPTER 6

Conclusion

This chapter serves to answer the research questions initially posed in Chapter 1 and to address the future work that may follow this study.

As a reminder to the reader, the drawn research questions will be presented one by one followed by their response and respective reflections.

RQ1: "How effective are the CBAL and FIMPS Active Learning algorithms in improving classifier performance on severely imbalanced ECG datasets, compared to traditional sampling?"

In both experimental scenarios, the Active Learning algorithms CBAL and FIMPS did not outperform traditional resampling techniques such as undersampling, oversampling, or SMOTE-Tomek. Despite their conceptual advantages, the obtained results reveal that Active Learning struggled to achieve competitive performance in this highly imbalanced and heterogeneous ECG dataset.

Between the two algorithms,FIMPS generally achieved higher recall, showing a stronger ability to detect minority-class instances of Atrial Flutter. However, this came at the cost of a large number of false positives, which reduced overall precision and accuracy. The CBAL algorithm, in contrast, produced fewer false positives and thus higher precision, but it failed to identify a sufficient number of true positives, leading to lower recall and F1-score. These results illustrate the distinct learning dynamics of both algorithms: whileFIMPS emphasizes the exploitation of informative samples, CBAL relies on the existence of a clearly defined decision boundary - a condition that was not fully met in this dataset.

We believe the limited effectiveness of both Active Learning methods may be attributed to the intrinsic complexity and overlap between classes. The majority class is not homogeneous, containing ECG recordings of various cardiac conditions, some even associated with Atrial Flutter, which blurs the separation between classes. Consequently, the classifiers are unable to establish a stable or meaningful decision boundary, leading the Active Learning process to select samples that are not truly informative, introducing redundancy.

RQ2: "How does the application of a GAN in the minority class impact on the performance of classification models and how does it influence Active Learning methods?"

Integrating synthetic data generated through a GAN had a significant positive impact on model performance, particularly in addressing the severe class imbalance inherent to the ECG-related dataset. The network produces 2 956 minority class samples to enrich the representation of Atrial Flutter and the results demonstrate a consistent improvement across all evaluation metrics when compared with the experiments that only use the original data.

TheCBAL algorithm, previously struggling with low precision and unstable decision boundaries in Scenario 1, achieved a considerable increase in both precision and F1-score, during Scenario 2 experiments. The result outlines a stronger ability to correctly identify Atrial Flutter instances while maintaining a lower rate of false positives. Similarly, theFIMPS technique showed enhanced recall and ROC-AUC, confirming that the GAN-generated data provided to the AL algorithms more informative samples with a more coherent representation of the minority class.

These improvements suggest that the GAN not only increased the amount of available data but also expanded the diversity and structural consistency within the minority class, enabling the models to learn more robust decision boundaries. As a result, AL methods could operate more effectively, identifying genuinely informative instances rather than noisy or redundant ones, which was a limitation in the original data scenario (Scenario 1).

While the risk of overfitting remains a potential concern in data augmentation processes, the observed performance results - especially in recall, F1-score and ROC-AUC - indicate that the synthetic data contributed meaningfully to model generalization rather than simply increasing accuracy. Overall, the integration of GAN-generated samples proved to be a valuable upgrade, enhancing the detection of rare Atrial Flutter cases and increasing the effectiveness of Active Learning algorithms under conditions of severe imbalance.

RQ3: "Does the integration of ECG signal analysis (data understanding) lead to a measurable improvement in Atrial Flutter detection in the context of class imbalance?"

The results obtained throughout the experimental process clearly demonstrate that data understanding and feature selection play a decisive role in the identification of Atrial Flutter under class imbalance conditions. Although the prior tests are not covered in this document, the truth is that numerous tests

were carried out until the optimal and most informative variables to include in the final dataset were found.

Additionally, in the earliest classification experiments, the model's performance was unexpectedly poor, inconsistent even and the reasons to that phenomenon were unaware at first. Subsequent analysis revealed that the source of the cause to the unfortunate outcomes was largely due to the lack of a structured feature selection process. The initial dataset included all extracted variables without a filtering criteria, which negatively affected the classifier's ability to distinguish both Atrial Flutter and majority-class instances. Moreover, the downgrading in performance resulted from the excessive and redundant information provided to the model.

Originally, the extracted features were related to four ECG leads (out of the 12), from which more than ten distinct metrics were added - including allHRV metrics, PQ and ST intervals, P, Q, R, S waves, and other parameters - resulting in over 40 variables per signal. Such range of features not only increased computational complexity given the dataset's size (over 45,000 records) but also introduced confusion in the learning process, i.e., many of these features represented similar physiological information. The fact that the same variables were being extracted from four different leads (II, III, aVF and V1) which presented identical values to the same features, led to the dissipation of the discriminative perception of the input space. As a consequence, the classifier struggled to recognize subtle inter-class differences and tended to overfit on irrelevant variations.

To address this issue, an iterative feature selection process was implemented to identify the most informative variables while discarding redundant ones. This theme will be further dissected in the next research question.

This methodological step back was essential for the classification model to fully function without any obstacles from the ECG signal representation point of view. Moreover, it highlights a key conclusion that a thorough understanding of the data and careful selection of relevant features substantially improves model performance in this case of Atrial Flutter detection and perhaps in similar cases which use morphological data as well.

RQ4: "What are the distinguishing ECG signal characteristics to identify Atrial Flutter, and how can these features be exploited to enhance model performance in highly imbalanced classification scenarios?"

During feature extraction, several morphological and temporal parameters were initially computed from the ECG signals. Feature refinement enabled to retain only the features that captured the essential physiological aspects of the ECG waveforms. Hence, the number of leads has been reduced from four

to one, lead II, for being the most complete due to its clinical reliability and representativeness in rhythm analysis. However, following the feature distribution and class separability analysis, particularly through histogram inspection, specific variables emerged as the most distinctive indicators between the Atrial Flutter and non-Atrial Flutter cases. These histograms were specifically plotted in order to identify the most interesting variables to include in the construction of the GAN.

The histograms presented in Figure 4.10 revealed that the features with the strongest separation capability are MeanNN and SDNN (HRV metrics), the QRS complex duration, RR ration, the mean amplitude of P-wave and the PQ interval. These parameters capture fundamental rhythm irregularities and conduction abnormalities - both Atrial Flutter characteristics - and are integrated in the list of selected features to be used in the GAN training. Additionally, age and sex are also covered in the network design, as contextual demographic variables, providing complementary information for performance interpretation and potential clinical relevance.

Therefore, the refined feature selection mentioned above was employed with the purpose of prioritizing precisely these parameters, as they exhibit the most distinct value ranges and distributions between classes. The generated synthetic data contributed to better class representation focused on features with measurable discriminative capacity, improving model learning dynamics under imbalance.

This methodological focus directly supports the notion that effective feature selection - rooted in physiological understanding of the ECG signal - enhances the model's ability to generalize across unevenly distributed data. The exploration and visualization of these variables not only clarified their contribution to class separation but also helped explain how each feature shapes the overall dataset structure. Consequently, this compact yet informative set of features constitutes the most relevant basis for reliable Atrial Flutter classification.

Ultimately, minimizing the noise and redundancy in the list of variables makes the classifier better at generalizing across imbalanced classes, leading to more reliable and clinically meaningful results.

Scientific Contributions and Limitations

This dissertation presents a set of methodological and empirical contributions to the study of atrial flutter in highly imbalanced learning scenarios. The results highlight the critical importance of feature engineering and feature integration, showing that these steps had a substantial impact on the overall performance of the classification models. In particular, clinically informed

ECG feature selection proved to be one of the most sensitive components of the proposed pipeline.

From an experimental perspective, the findings indicate that Active Learning did not achieve the expected performance improvements over traditional sampling-based approaches in the presence of severe class imbalance and class overlap. Conversely, experiments involving synthetic data generation through Generative Adversarial Networks yielded more consistent performance gains, particularly in terms of sensitivity and precision for Atrial Flutter detection. Finally, the definition of classes and the presence of mixed or overlapping cardiac conditions in the majority class (Atrial Flutter + other heart conditions) were identified as limiting factors, significantly influencing the obtained results and underscoring the importance of careful class construction in clinical machine learning studies. Moreover, the usage of a unique dataset - the one from *PhysioNet* - compromises validation as the model adapts to the data. It would be a good practice to assess the model's performance by introducing an external dataset.

Future Work

While this study has demonstrated the potential of combining synthetic data generation and Active Learning to address class imbalance in ECG-based Atrial Flutter detection, the findings also reveal important limitations that can guide for future research. The experiments made showed that Active Learning algorithms such as CBAL and FIMPS struggle to outperform traditional sampling methods when the dataset presents a high degree of class overlap and heterogeneity. Therefore, future studies should focus on refining both class definition and feature extraction to establish clearer and more separable decision boundaries. The solution may reside in adding more insightful morphological indicators to the dataset in order to distinguish the subtle differences that may exist between Atrial Flutter and other arrhythmias, improving model discrimination capacity.

Another interesting direction involves enhancing the GAN training process through more advanced architectures, such as conditional GANs or diffusion-based generative models, to produce physiologically consistent synthetic signals that better represent minority-class variability. Integrating these high-quality synthetic data with adaptive Active Learning strategies may further improve sample selection efficiency, particularly when combined with mechanisms to evaluate local uncertainty more robustly.

Evidently, there is still plenty room for improvement related to this subject. However, it is essential to emphasize that the quality and interpretability of the data remain the most decisive factors in achieving reliable results. The

clarity and coherence of the data, both in terms of feature representation and diagnostic consistency, are fundamental to the model's ability to distinguish between classes and to ensure that any technical enhancement translates into genuine clinical insight.

Bibliography

[1] Haibo He and E. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, pp. 1263–1284, Sept. 2009.

[2] B. Settles, "Active Learning Literature Survey," 2009.

[3] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," *Artificial Intelligence Review*, vol. 57, p. 137, May 2024.

[4] J. R. Venable, J. Pries-Heje, and R. L. Baskerville, "Choosing a design science research methodology," in *Proceedings of the DESRIST 2017 Conference*, 2017.

[5] J. Zheng, H. Guo, and H. Chu, "A large-scale 12-lead electrocardiogram database for arrhythmia study." PhysioNet, 2022. Dataset, version 1.0.0. RRID: SCR_007345.

[6] Serviço Nacional de Saúde, "Eletrocardiograma." `https://www.sns24.gov.pt/pt/tema/exames-e-tratamentos/eletrocardiograma/`, 2025. Accessed: 2025-09-03.

[7] CUF Saúde, "Eletrocardiograma (ecg)." `https://www.cuf.pt/saude-a-z/eletrocardiograma-ecg`, 2025. Accessed: 2025-09-03.

[8] Geeky Medics, "Understanding an ecg." `https://geekymedics.com/understanding-an-ecg/`, 2025. Accessed: 2025-09-03.

[9] Life in the FastLane, "Atrial flutter." `https://litfl.com/atrial-flutter-ecg-library/`, 2024. Accessed: 2025-09-07.

[10] J. Granada, W. Uribe, P.-H. Chyou, K. Maassen, R. Vierkant, P. N. Smith, J. Hayes, E. Eaker, and H. Vidaillet, "Incidence and predictors of atrial flutter in the general population," *Journal of the American College of Cardiology*, vol. 36, pp. 2242–2246, Dec. 2000.

[11] National Center for Biotechnology Information, "Arrhythmias." `https://www.ncbi.nlm.nih.gov/books/NBK540985/`, 2020. Accessed: 2025-09-03.

[12] J. Fu and S. Lee, "Certainty-based active learning for sampling imbalanced datasets," *Neurocomputing*, vol. 119, pp. 350–358, Nov. 2013.

[13] L. Chen, X. Tian, and L. Cai, "FIM-Based Pairwise Selection for Active Learning on Imbalanced Datasets," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, (Kowloon Tong, Hong Kong), pp. 1876–1881, IEEE, Oct. 2015.

[14] Y. Xiao and M. Watson, "Guidance on Conducting a Systematic Literature Review," *Journal of Planning Education and Research*, vol. 39, pp. 93–112, Mar. 2019.

[15] R. L. S. Silva and Frâncila Weidt Neiva, "Systematic Literature Review in Computer Science - A Practical Guide," 2016. Publisher: 1174.

[16] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, p. n71, Mar. 2021.

[17] J. Attenberg and Ertekin, "Class Imbalance and Active Learning," in *Imbalanced Learning* (H. He and Y. Ma, eds.), pp. 101–149, Wiley, 1 ed., June 2013.

[18] T. M. Hospedales, Shaogang Gong, and Tao Xiang, "Finding Rare Classes: Active Learning with Generative and Discriminative Models," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, pp. 374–386, Feb. 2013.

[19] Y. Zhang and D. Wang, "A Cost-Sensitive Ensemble Method for Class-Imbalanced Datasets," *Abstract and Applied Analysis*, vol. 2013, pp. 1–6, 2013.

[20] Jia Guo, Xin Wan, Hao Lin, Peng Li, Guannan Liu, and Yueying He, "An active learning method based on mistake sampling for large scale imbalanced classification," in *2017 International Conference on Service Systems and Service Management*, (Dalian, China), pp. 1–6, IEEE, June 2017.

[21] J. Deng, J. Sun, W. Peng, D. Zhang, and V. Vyatkin, "Imbalanced multiclass classification with active learning in strip rolling process," *Knowledge-Based Systems*, vol. 255, p. 109754, Nov. 2022.

[22] G. Aguiar and A. Cano, "An active learning budget-based oversampling approach for partially labeled multi-class imbalanced data streams," in *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, (Tallinn Estonia), pp. 382–389, ACM, Mar. 2023.

[23] G. Petmezas, K. Haris, L. Stefanopoulos, V. Kilintzis, A. Tzavelis, J. A. Rogers, A. K. Katsaggelos, and N. Maglaveras, "Automated Atrial Fibrillation Detection using a Hybrid CNN-LSTM Network on Imbalanced ECG Datasets," *Biomedical Signal Processing and Control*, vol. 63, p. 102194, Jan. 2021.

[24] M. S. Lee and Y.-S. Choi, "Transfer Learning with a Pretrained Foundation Model for Atrial Fibrillation and Flutter Detection on Single-Lead ECG,"

in *2025 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC)*, (Seoul, Korea, Republic of), pp. 1–5, IEEE, July 2025.

[25] M. U. Gul, M. H. Kamarul Azman, K. A. Kadir, J. A. Shah, and S. Hussen, "Supervised Machine Learning Based Noninvasive Prediction of Atrial Flutter Mechanism from P-to-P Interval Variability under Imbalanced Dataset Conditions," *Computational Intelligence and Neuroscience*, vol. 2023, p. 8162325, Jan. 2023.

[26] B. Salsekar and D. Wadhwani, "Filtering of ecg signal using butterworth filter and its feature extraction," *International Journal of Engineering Science and Technology*, vol. 4, 04 2012.

[27] B. Zanchi, G. Monachino, L. Fiorillo, G. Conte, A. Auricchio, A. Tzovara, and F. D. Faraci, "Synthetic ECG signals generation: A scoping review," *Computers in Biology and Medicine*, vol. 184, p. 109453, Jan. 2025.

[28] L. Berger, M. Haberbusch, and F. Moscato, "Generative adversarial networks in electrocardiogram synthesis: Recent developments and challenges," *Artificial Intelligence in Medicine*, vol. 143, p. 102632, Sept. 2023.

[29] O'Sullivan, C., "Isolation forest guide: Explanation and python implementation." `https://www.datacamp.com/tutorial/isolation-forest`, 2025. Accessed: 2025-09-28.