Multimodal Analysis of Medical Data using Graphs and Deep Learning

Thesis submitted by
Anirban Dutta Choudhury

Doctor of Philosophy (Ph.D.)
in
Engineering

Department of Electronics &
Telecommunication Engineering,
Faculty Council of Engineering & Technology
Jadavpur University, Kolkata
India

2026

JADAVPUR UNIVERSITY
KOLKATA –700032, INDIA

INDEX NO.: 20/18/E

1. **Title of the Thesis:**

   Multimodal Analysis of Medical Data using Graphs and Deep Learning

2. **Name, Designation & Institution of the Supervisor:**

   Dr. Ananda Shankar Chowdhury
   Professor
   Department of Electronics & Telecommunication Engineering
   Jadavpur University,
   Kolkata 700032,
   India

3. **List of publication:**

   (Journals)

   1. **A. Dutta Choudhury**, A. S. Chowdhury: "Explainable hypergraphs for gait based Parkinson classification", *Pattern Recognition Letters* 186 (2024): 198-204.

   (Conferences)

   1. **A. Dutta Choudhury**, A. S. Chowdhury: "CHANGE: Cardiac Health Analysis Using Graph Eigenvalues", 40^{th} *Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (2018), pp. 4025-4029.
   2. **A. Dutta Choudhury**, A. S. Chowdhury: "CHAMPS: Cardiac health Hypergraph Analysis using Multimodal Physiological Signals", 41^{st} *An-*

nual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (2019), pp. 4640-4645.

3. **A. Dutta Choudhury**, D. Basak, R. Dutta and A. S. Chowdhury: "A Focal Weighted Binary Cross Entropy Loss for Explainable Sepsis Prediction", Fifth International Conference on Emerging Techniques in Computational Intelligence, ICETCI 2025, pp. 1-8. IEEE, 2025.

STATEMENT OF ORIGINALITY

I, ***Anirban Dutta Choudhury***, registered on ***08-June-2018*** do hereby declare that this thesis entitled "***Multimodal Analysis of Medical Data using Graphs and Deep Learning***" contains literature survey and original research work done by the undersigned candidate as part of Doctoral studies.

All the information in this thesis have been obtained and presented in accordance with existing academic rules and ethical conduct. I declare that, as required by these rules and conduct, I have fully cited and referred all materials and results that are not original to this work.

I also declare that I have checked this thesis as per the "Policy on Anti Plagiarism, Jadavpur University, 2019", and the level of similarity as checked by iThenticate software, after excluding my publications for this thesis, is 3%.

Anirban Dutta Choudhury

Signature of Candidate

Date: 13-April-2026

Certified by Supervisor(s):

Ananda Shankar Chowdhury

Signature of the Supervisor

Dr. Ananda Shankar Chowdhury

Professor, Dept. of ETCE

Jadavpur University, Kolkata – 700032

ANANDA SHANKAR CHOWDHURY
Professor
Electronics and Telecommunication Engineering
Jadavpur University, Kolkata-32

CERTIFICATE FROM THE SUPERVISOR/S

This is to certify that the thesis entitled "***Multimodal Analysis of Medical Data using Graphs and Deep Learning***" submitted by ***Mr. Anirban Dutta Choudhury***, who got his name registered on ***08/06/2018 [D-7/E/363/18]*** for the award of Ph.D (Engg.) degree of Jadavpur University is based upon his work under the supervision of ***Prof. Ananda Shankar Chowdhury*** and that neither his thesis nor any part of this thesis has been submitted for any degree/diploma or any other academic award anywhere before.

Signature of the Supervisor

Dr. Ananda Shankar Chowdhury

Professor, Dept. of ETCE

Jadavpur University, Kolkata – 700032

Date: 13-April-2026

ANANDA SHANKAR CHOWDHURY
Professor
Electronics and Telecommunication Engineering
Jadavpur University, Kolkata-32

Abstract

Physiological sensor data, such as those obtained from ECG, PCG, PPG, and VGRF modalities, medical records and metadata contain a rich collection of biomedical indicators that reflect a wide array of underlying medical conditions. Over the past decades, researchers have extensively employed machine learning (ML) and deep learning (DL) techniques to model the relationship between such sensor-derived data and disease-specific biomarkers.

Conventional shallow learning approaches typically involve the extraction of a limited number of handcrafted features, guided by domain expertise. These features may include time-domain statistics, frequency-domain coefficients, or combinations thereof. Although effective to a certain extent, such methods are often constrained by their reliance on feature engineering and limited representational capacity. In contrast, deep learning models automatically learn hierarchical feature representations from raw data. These models are capable of generating millions of parameters, enabling powerful non-linear mappings and classification decisions. However, these features often remain uninterpreted or disconnected, hindering their transparency and clinical interpretability.

To address this gap, the present thesis introduces a graph-based framework to capture the interrelationships between features in a structured and interpretable manner. Specifically, for the task of coronary artery disease (CAD) classification, we construct subject-specific graphs using shallow learning features. Unsupervised testing on photoplethysmogram (PPG) and metadata of 32 participants achieved 88% accuracy, demonstrating superior performance over a baseline and two state-of-the-art methods. Building upon this, we propose a hypergraph-based model that further enhances interpretability by representing individuals as hypergraphs—where hyperedges encapsulate higher-order interactions among multiple features. Validated using multimodal PPG, phonocardiogram (PCG) and metadata, the proposed method demonstrates 92% classification accuracy, driven by 98% sensitivity and 82% specificity.

A hybrid pipeline for Parkinson's Disease (PD) classification merges a novel explainable weighted hypergraph for feature extraction (Shallow Learning) with a ResNet architecture (Deep Learning) using multimodal Vertical Ground Reaction Force (VGRF) signals, gait speed and metadata. This dual-pronged strategy combines the interpretability of the graph-based models with the superior predictive performance of deep networks. The proposed pipeline achieves a superior 0.979 AUC on the PhysioNet Parkinson Dataset compared to isolated shallow (0.878) and deep learning (0.852) methods. The framework improves explainability through enhanced feature importance and demonstrates, via case studies, how the hybrid model rectifies individual component errors.

Furthermore, to improve model performance for highly imbalanced data, such as that found in sepsis prediction, we propose a novel loss function—Focal Weighted Binary Cross Entropy Loss. This approach not only addresses class imbalance but also assigns differential importance to more informative samples. Our sepsis prediction model leverages multimodal sensory data, medication records, and categorical metadata to ensure comprehensive and timely detection. The proposed method achieves a utility score of 0.4305 and 0.393 on two sepsis test sets which is at par with the best performance provided by the state of the art methods. Our SHAP analysis demonstrates that covering a wide range of medical specialties—a feature lacking in competitors—is our most impactful trait. This confirms that our loss function effectively addresses sepsis diagnosis by addressing multiple clinical domains.

Acknowledgment

"Make things as simple as possible,
but not simpler".
— Albert Einstein

Pursuing a Ph.D., in my earnest view, resembles entering into a profound partnership—akin to a miniature marriage—between the student and the advisor. In this intellectual and emotional journey, the advisor assumes the role of a guiding light, shaping the path and nurturing the student with wisdom, patience, and encouragement. I am deeply grateful to Professor Ananda S. Chowdhury for graciously accepting me as his doctoral student. That moment marked the beginning of an extraordinary voyage—one that has been as demanding as it has been enriching.

Beyond his pivotal academic guidance and critical insights that significantly contributed to the conceptualization and structure of this thesis, I have had the privilege of learning many invaluable life lessons from him. Among these, the art of supporting a student unwaveringly through both triumphs and setbacks stands out the most. Reflecting upon my journey, I often think that, had I been my own advisor, I might never have reached the end of this road. Professor Chowdhury played multiple roles throughout this journey—mentor, philosopher, and at times, a true friend—always guided by a philosophy rooted in the pursuit of scientific excellence and the cultivation of self-directed learning. I look forward to nurturing this meaningful relationship in the years to come, well beyond the formal conclusion of my doctoral journey.

I would be remiss if I did not acknowledge the unwavering support of my family. I express my heartfelt gratitude to my parents, Dr. Arunachal Dutta Choudhury and Dr. Sunanda Dutta Choudhury, who, despite modest beginnings, never wavered in their commitment to the value of education. Their encouragement laid the foundation for all that I have achieved. I am certain that they will be even more joyous than I am when this thesis reaches its final defense.

My brother, Arghadeep, has been a steadfast source of motivation, constantly

checking in on my progress and encouraging me to persevere. My deepest appreciation goes to my wife, Sohini, whose strength, patience, and unwavering support provided the emotional anchor I needed to complete this thesis. Her sacrifices and constant reassurance made this accomplishment possible. Lastly, to my beloved sons, Abhijnan and Ananyo, who endured many days and nights without their father's presence—I hope that in time, they will understand and take inspiration from this journey.

Jadavpur University
Kolkata – 700032

Anirban Dutta Choudhury
Data: 13-April-2026

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *Innovation distinguishes between a leader and a follower.*

Steve Jobs, The Innovation Secrets of Steve Jobs, 2001

This chapter provides an outline of the different components in time-series medical data analysis. In section 1.1, we discuss the motivation that prompted this research work. Section 1.2 presents an overview of diversified medical conditions touched upon in this thesis. Section 1.3 details an overview about the state of the art research on the medical conditions presented in section 1.2. We examine the research gaps in section 1.4, followed by the organization of this thesis in section 1.6. Finally, we highlight our contributions in section 1.5.

1.1 Prologue

The exponential growth of scientific and technological innovation, particularly in the domain of hardware-software co-design, has ushered in a transformative era for researchers and developers alike. Merely two to three decades ago—during the 1990s and early 2000s—researchers operated under severe constraints in terms of computational resources, data storage, and network connectivity. More critically, however, they lacked access to affordable, precise, and reliable sensor technologies. The landscape has drastically changed in the intervening years, with sensor technol-

ogy undergoing a remarkable evolution characterized by significant advancements in both functionality and affordability.

To illustrate this transformation, consider the example of inertial measurement, specifically the detection of acceleration. In the past, accelerometers—alongside gyroscopes and magnetometers—formed the foundational elements of motion sensing. The author of this thesis was involved in a project on pedestrian dead reckoning at NXP Semiconductors in Nijmegen, Netherlands, utilizing the Nokia N810 tablet, which was regarded as a state-of-the-art handheld device in 2008. Despite its premium pricing, the N810 featured only a single onboard motion sensor: the accelerometer. At the time, gyroscopes and magnetometers were either unavailable or not feasible to integrate due to cost, power, or application constraints.

Fast forward to the present day, and such limitations are almost inconceivable. Modern tablets and smartphones—regardless of price tier—routinely incorporate all three motion sensors. This shift is not merely a matter of hardware evolution but a reflection of broader socio-technical dynamics. Initially, accelerometers were integrated into handheld devices to enable automatic screen orientation switching (portrait to landscape and vice versa), a relatively simple task based on thresholding vertical acceleration. However, as the market for handheld and wearable devices expanded rapidly, manufacturers began equipping their products with an increasing number of sensors, often exceeding 15–20 in a single consumer-grade device within a decade of the N810's release.

This widespread sensor integration created a positive feedback loop. The proliferation of sensors across devices led to mass production, which in turn drove down costs while simultaneously improving sensor performance. Today, even budget smartwatches and smartphones—costing under 100 USD—can detect changes in acceleration as small as $1/1000^{\text{th}}$ of gravitational acceleration. While such high precision is excessive for simple orientation detection, it enabled a host of new applications. Fields such as healthcare, fitness, and interactive gaming began leveraging sensor data in novel and impactful ways. Experiments that once required costly laboratory setups could now be replicated with consumer-grade hardware. For instance,

accelerometers are widely used to count steps, infer walking gait, and estimate caloric expenditure during physical activity—all at virtually no marginal cost.

Concurrently, the plummeting costs of communication, computation, and data storage allowed for the collection and analysis of massive sensor-generated datasets. Traditional machine learning (ML) algorithms, though once sufficient, struggled to scale with this deluge of data. This challenge catalyzed the rise of deep learning (DL), which promised improved performance and accuracy across a wide array of tasks. However, DL models brought their own complications, most notably the lack of interpretability. As neural networks became more complex, understanding their internal reasoning became increasingly opaque—earning them the moniker "black-box" models. Researchers began to observe instances where DL models produced accurate predictions for entirely spurious, unexplainable, and non-repeatable reasons, casting doubt on their reliability in unseen scenarios.

These concerns have fueled the emergence of *explainable artificial intelligence* (XAI), a field that seeks to enhance transparency and trustworthiness by elucidating the decision-making processes of AI systems. It is within this context that the present thesis explores the utility of *Graph Signal Processing* (GSP) as a powerful tool for developing robust, interpretable, and high-performing diagnostic models, particularly in the domain of disease detection. GSP offers the unique advantage of bridging shallow and deep learning paradigms, enabling us to extract meaningful insights from structured data while preserving explainability. The forthcoming chapters will elaborate on how GSP-based methodologies can strike a critical balance between performance and interpretability, thus addressing one of the central challenges in modern AI research.

In the following sections of this chapter, we shall discuss the sensors (multiple modalities), different diseases, and the state-of-the-art research efforts in addressing the classification of these diseases. Furthermore, we discuss the research gaps that we aim to solve in this thesis.

1.2 Medical Conditions

The application of Artificial Intelligence (AI) in medical diagnostics has witnessed a dramatic surge over the past two decades, catalyzing transformative progress across multiple domains of healthcare. In this thesis, however, our focus is restricted to a curated set of medical conditions that are amenable to detection and classification using one-dimensional (1-D) time-series data, acquired through non-invasive physiological sensors. This section provides an overview of the sensors considered within the scope of this research and the corresponding medical conditions targeted for analysis and detection. The selected sensors were chosen based on three core criteria: clinical relevance, data accessibility via wearable or non-invasive devices, and their effectiveness in capturing physiological signals critical to disease prediction and monitoring.

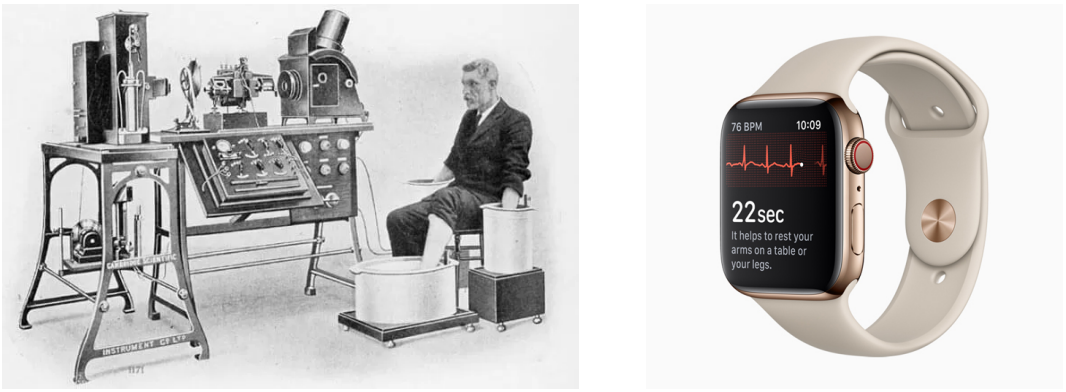
1.2.1 Sensors

The use of sensors to assess and monitor human health dates back centuries. One of the earliest examples of such instrumentation is the stethoscope, invented by René Laennec in 1816. Prior to this innovation, physicians relied on direct auscultation by placing their ear on a patient's chest to perceive cardiac activity. Laennec's wooden, monaural stethoscope revolutionized this practice by not only improving hygiene and ergonomics but also amplifying heart sounds via the physics of sound conduction through a hollow tube. More than two centuries later, the stethoscope remains a mainstay in clinical diagnostics—albeit in a modernized form. Contemporary stethoscopes now feature binaural earpieces, electronic amplification, digital recording capabilities, and wireless connectivity. The acoustic signal of the heart captured via a stethoscope is termed the phonocardiogram (PCG), a fundamental non-invasive tool in cardiovascular evaluation.

Another milestone in physiological sensing was the development of the electrocardiogram (ECG), which records the electrical activity of the heart via electrodes placed at specific sites on the body. In 1924, Willem Einthoven received the Nobel Prize for devising the first practical electrocardiographic system for clinical use.

As illustrated in Fig. 1.1, this once cumbersome apparatus has evolved into highly compact and accessible forms, such as smartwatches, enabling 24/7 heart monitoring for consumers. The ECG waveform comprises several diagnostically significant components:

1. **Peaks and Troughs:** Including the P, R, and T waves (peaks) and the Q and S waves (troughs).
2. **Segments and Intervals:** PR interval, QRS complex, ST segment, and QT interval—each of which serves as a biomarker for identifying deviations from normative cardiac function.
3. **Amplitude:** Variations in peak elevation, such as ST-segment elevation, are indicative of certain cardiac pathologies.



(a) First ECG device weighing 300 kg

(b) Apple Watch weighing 36 g

Figure 1.1: Evolution of ECG Devices

Plethysmography refers to the measurement of volumetric changes in an organ or body part, particularly as a function of blood flow. The term "photoelectric plethysmograph" was first introduced in 1936 by Alrick B. Hertzman and colleagues, laying the foundation for modern Photoplethysmography (PPG). Today, PPG is most commonly employed in pulse oximetry, where red and infrared light are transmitted through a fingertip to distinguish between oxygenated and deoxygenated blood. This technique, known as transmissive plethysmography, enables the estimation of

peripheral oxygen saturation (SpO_2). Reflective PPG (rPPG), an evolution of this technology, is now widely integrated into consumer-grade smartwatches and fitness trackers. It estimates SpO_2 by analyzing light reflected from the skin, typically at the wrist, offering a non-invasive, continuous monitoring option.

(a) Medical-grade oximeter

(b) PPG sensors in smartwatch

Figure 1.2: Evolution of PPG Devices

Summary: The inclusion of PCG, ECG, and PPG in this thesis is justified by their non-invasiveness, wide availability in clinical and consumer devices, and proven utility in detecting cardiovascular and neurological anomalies. These sensors generate rich 1-D time-series data streams, well-suited for analysis using signal processing and machine learning techniques.

1.2.2 Diseases

The diseases selected for study in this thesis share a common characteristic: they can be diagnosed or monitored using 1-D physiological signals, often in conjunction with auxiliary metadata (e.g., age, weight, height) and categorical indicators (e.g., gender, smoking status). Below, we detail three major disease categories considered in this work. The diseases are chosen as per the disability-adjusted life year (DALY) of all diseases in the world. The mortality rate is also weighed in selecting them. Moreover, the availability of 1-D signals was a factor, too. Upon visiting the above-mentioned points, we chose three distinct diseases, namely, Coronary Artery Disease, Parkinson's Disease, and Sepsis.

(A) Coronary Artery Disease (CAD)

Advancements in public health and medical interventions have drastically reduced mortality from communicable diseases (CDs) such as tuberculosis, measles, and even more recent threats like HIV, Zika, and Ebola. However, the global epidemiological burden has shifted toward non-communicable diseases (NCDs), which are chronic in nature and arise from a complex interplay of genetic, physiological, environmental, and behavioral factors. Among NCDs, cardiovascular diseases (CVDs) represent the leading cause of mortality, accounting for approximately 17.9 million deaths annually—about 31% of all global deaths [3–6].

Coronary Artery Disease (CAD), a major subset of CVDs, results from the accumulation of fatty deposits within the walls of coronary arteries, leading to restricted blood flow to cardiac muscle. The clinical gold standard for diagnosing CAD is angiography, an invasive imaging procedure that involves injecting contrast dye into the bloodstream. Due to its invasive nature and high cost (ranging from USD 300–500 in India to approximately USD 5000 in the United States), there is an urgent need for low-cost, non-invasive screening alternatives that can enable early diagnosis and preventive care.

(B) Parkinson's Disease (PD)

Neurological disorders are now recognized as the second leading cause of global mortality and a primary contributor to disability-adjusted life years (DALYs). Parkinson's Disease (PD), in particular, accounts for the majority of neurodegenerative disorder-related deaths, affecting millions worldwide [7]. PD is characterized by progressive motor dysfunction, including tremors, rigidity, and altered gait patterns. Early diagnosis of PD can significantly improve patient outcomes and quality of life.

Researchers have explored numerous modalities—such as speech, handwriting, movement patterns, and medical imaging—to differentiate PD patients from healthy controls [8–10]. Among these, gait analysis using pressure-sensitive insoles that capture Vertical Ground Reaction Forces (VGRF) has emerged as a cost-effective, non-invasive, and energy-efficient method. Unlike inertial sensors, VGRF-based systems

are simpler in design and capable of capturing high-resolution gait signatures critical for early-stage PD detection.

(C) Sepsis

Sepsis is a life-threatening condition triggered by an aberrant immune response to infection, resulting in multi-organ dysfunction. Despite advances in critical care, sepsis remains a leading cause of hospital mortality, implicated in nearly one-third of all in-hospital deaths in the United States alone [11, 12]. Each year, an estimated 31.5 million individuals globally develop sepsis, with 19.4 million progressing to severe stages and approximately 5.3 million fatalities reported [13].

The heterogeneity of clinical symptoms—ranging from fever and rapid heart rate to mental confusion and respiratory distress—makes early detection challenging. Timely diagnosis is crucial, as delayed intervention can lead to septic shock and death. Emerging AI-based technologies, such as early warning systems integrated with electronic health records (EHRs), offer promising avenues for improving sepsis recognition and treatment timelines. Nevertheless, challenges remain in modeling the complex pathophysiology and ensuring equitable healthcare access across populations.

Summary: The diseases targeted in this thesis—CAD, PD, and Sepsis—were selected based on three principal criteria: (1) their global health burden and clinical relevance, (2) the feasibility of detection through non-invasive, time-series sensor data, and (3) the potential impact of early, accessible AI-based diagnostics in improving patient outcomes.

1.3 Literature Review

The rapid advancement of artificial intelligence (AI) in healthcare has led to the development of a diverse array of diagnostic models and analytical pipelines for identifying and classifying medical conditions using physiological signals. This section provides an overview of the state-of-the-art methods relevant to the classification of Coronary Artery Disease (CAD), Parkinson disease (PD), and Sepsis prediction. We

categorize the existing body of research into two broad methodological paradigms: traditional machine learning approaches and modern deep learning frameworks.

1.3.1 Classical Machine Learning Methods

Classical machine learning (ML) approaches for medical signal analysis typically rely on explicit feature engineering, whereby informative features are extracted from physiological time series data such as photoplethysmogram (PPG), phonocardiogram (PCG), and electrocardiogram (ECG) signals [14]. Before feature extraction, raw signals undergo rigorous preprocessing, which may involve noise filtering, artifact removal, signal normalization, data imputation, and dimensional transformation [15]. These features are derived using a variety of signal processing techniques that span multiple domains.

In coronary artery disease (CAD) classification, handcrafted feature extraction from physiological signals such as ECG, PCG, and PPG remains a fundamental step for enhancing diagnostic performance. Time-domain features, including statistical measures such as mean, variance, skewness, kurtosis, and higher-order moments, capture the morphological characteristics and temporal variations of cardiac signals [16]. Frequency-domain features, typically obtained using Fourier Transform-based methods, provide insights into the spectral distribution of signal energy and have been effective in identifying abnormal cardiac rhythms and ischemic patterns [17]. To address the non-stationary nature of biomedical signals, time–frequency domain techniques such as Short-Time Fourier Transform (STFT) and Wavelet Transform (WT) are widely employed, enabling simultaneous localization of temporal and spectral information and improving the detection of transient cardiac events [16, 17]. In addition, entropy-based features, including Approximate Entropy (ApEn), Sample Entropy (SampEn), and Spectral Entropy, quantify the complexity and irregularity of cardiac signals, providing valuable biomarkers for distinguishing between healthy and CAD-affected subjects [16, 18]. The integration of these diverse feature representations with machine learning classifiers has been shown to significantly enhance the accuracy and robustness of CAD diagnosis systems.

In coronary artery disease (CAD) prediction, advanced signal processing techniques such as Empirical Mode Decomposition (EMD), Principal Component Analysis (PCA), and Independent Component Analysis (ICA) play a crucial role in extracting discriminative features from physiological signals. EMD has been widely applied for denoising and decomposition of non-linear and non-stationary signals such as ECG and PCG, enabling the separation of intrinsic oscillatory modes and removal of redundant noise components [19, 20]. PCA is extensively used as a dimensionality reduction technique to transform high-dimensional biomedical feature spaces into compact representations while preserving maximum variance, thereby improving classifier efficiency and reducing redundancy [21, 22]. Similarly, ICA facilitates the separation of statistically independent latent components from mixed cardiac signals, enhancing noise suppression and feature interpretability in CAD detection systems [21]. These techniques are often integrated with machine learning classifiers such as SVM, GMM, and neural networks, where studies have shown that ICA-based feature representations can achieve superior classification performance compared to PCA in CAD diagnosis tasks [21]. Overall, the combination of decomposition and dimensionality reduction techniques significantly improves the robustness and accuracy of CAD prediction models.

Banerjee *et al.* [23] demonstrated that incorporating multiple clinical metadata features—such as age, blood pressure, body mass index (BMI), and chest pain characteristics—can yield a sensitivity of 92% and a specificity of 90% for CAD detection. The PhysioNet Challenge 2016 [24] provided a large-scale heart sound dataset collected from leading hospitals worldwide, focusing on the classification of abnormal PCG signals arising from various cardiac conditions, including CAD and valvular disorders. The best-performing approach in this challenge, based on a modified AdaBoost algorithm, achieved an F1 score exceeding 86% [25]. These studies highlight the wide range of methodologies employed in CAD detection, predominantly involving machine learning classifiers such as support vector machines and random forests, along with diverse strategies including linear and non-linear modeling, and ensemble techniques like bootstrapping.

Classical machine learning (ML) techniques have been widely employed for the classification of coronary artery disease (CAD), particularly when using structured clinical data and non-invasive physiological signals such as electrocardiogram (ECG), photoplethysmogram (PPG), and phonocardiogram (PCG). These approaches typically follow a pipeline involving preprocessing, handcrafted feature extraction from time-domain, frequency-domain, and time–frequency representations, and subsequent classification using algorithms such as support vector machines (SVM), k-nearest neighbors (kNN), decision trees, and ensemble methods. Among these, SVM has consistently demonstrated strong performance due to its ability to handle high-dimensional feature spaces and nonlinear decision boundaries through kernel functions [26, 27]. Recent studies further show that optimized SVM variants, combined with feature selection techniques such as genetic algorithms or Bayesian optimization, can achieve classification accuracies exceeding 89%–97% in CAD detection tasks [27, 28].

Similarly, kNN-based classifiers have been effectively utilized due to their simplicity and strong discriminative capability in well-structured feature spaces. Recent comparative studies indicate that kNN can achieve very high classification performance, with reported accuracies exceeding 98% and AUC values close to 0.99 when applied to carefully selected clinical features [22]. However, kNN suffers from scalability issues and sensitivity to noise, particularly in high-dimensional biomedical datasets [26].

In addition to supervised learning, unsupervised techniques such as clustering (e.g., k-means, hierarchical clustering, and fuzzy C-means) have been explored to uncover latent structures in cardiac data. These methods are particularly useful for patient stratification, anomaly detection, and preprocessing prior to classification. Recent works integrating clustering with downstream classifiers have demonstrated improved feature representation and classification accuracy, highlighting the complementary role of unsupervised learning in CAD diagnosis [29, 30].

Despite the increasing adoption of deep learning, classical ML approaches remain highly relevant due to their interpretability, lower computational complexity,

and effectiveness on small-to-medium sized datasets. Comparative studies continue to show that models such as SVM and random forests often outperform or remain competitive with more complex approaches in structured clinical datasets, making them suitable for real-time clinical decision support systems [26, 31]. Overall, classical ML methods provide a robust and interpretable framework for CAD classification, particularly when combined with effective feature engineering and hybrid learning strategies.

Research on Parkinson's disease (PD) classification prior to 2015 predominantly focused on various supervised learning (SL) techniques, including Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Singular Value Decomposition (SVD), often employing different kernel configurations. The feature sets used in these studies comprised a range of time- and frequency-domain descriptors such as Coefficient of Variation (CV), Center of Pressure (CoP), peak force, as well as higher-order statistics like kurtosis and skewness derived from gait cycles [32–35]. Subsequent studies explored Local Pattern Transformation-based methods for PD classification [36, 37]. Ye *et al.* introduced an adaptive neuro-fuzzy inference system to capture the nonlinear dynamics of gait patterns [38]. Additionally, efforts have been made toward reconstructing Vertical Ground Reaction Force (VGRF) signals [39]. Khoury *et al.* provided a detailed comparative analysis of various methodologies applied to the PhysioPD2007DB dataset [40]. Nevertheless, SL-based classification methods generally exhibit limited performance, and even state-of-the-art approaches often overlook the underlying physical events, biological mechanisms, or statistical foundations associated with the extracted features.

The PhysioNet 2019 Challenge (PNC19) and its accompanying dataset (hereinafter referred to as PNC19DB) catalyzed significant advancements in the domain of early sepsis prediction using electronic health record (EHR) data. One of the most critical challenges highlighted through this competition was the pervasive issue of missing data, which accurately reflects the inconsistencies and sparsity typically encountered in real-world clinical settings [41]. To mitigate the effects of data incompleteness, a broad range of state-of-the-art (SOTA) methods integrated vari-

ous data imputation techniques. Notably, several models exploited the concept of *informative missingness*—where patterns in the absence of data themselves carry diagnostic significance. For example, the duration for which a specific variable remains unrecorded (e.g., hours a lab test is missing) was found to provide valuable predictive signals [42]. Leveraging such information, Singh *et al.* reported an area under the receiver operating characteristic (AUROC) of 0.8406 and a utility score of 0.404 using five-fold stratified cross-validation [43].

A particularly noteworthy submission utilized a signature-based regression framework [44]. This technique involved computing the path signature—a series of iterated integrals that summarize the geometric and temporal structure of time-series data. The path signature offers universal, invariant representations well-suited for modeling complex sequences. When this method was augmented with handcrafted features, the resulting model achieved a utility score of 0.43, ranking first among the 78 entries in the PNC19 Challenge.

In addition to imputation, many competitive approaches employed handcrafted features rooted in clinical knowledge. Examples include the Shock Index (defined as the ratio of heart rate to systolic blood pressure), the Bilirubin-to-Creatinine ratio, and partial Sequential Organ Failure Assessment (SOFA) scores, inspired by the clinical feature selection framework proposed in [45]. These features enriched model inputs with domain-specific insights that were difficult to learn directly from raw data, thereby boosting predictive performance.

A predominant modeling strategy involved the use of tree-based ensemble methods, particularly gradient boosting frameworks such as XGBoost [42–44]. Singh *et al.* constructed a robust XGBoost ensemble wherein 10% of the dataset was used for feature selection and hyperparameter tuning, while the remaining 90% was partitioned into five non-overlapping folds, each used to train a separate classifier. Predictions from these models were aggregated using the geometric mean of their outputs. This ensemble achieved an AUROC of 84.4% and a utility score of 0.4281, demonstrating a high level of discrimination and timeliness in sepsis onset prediction [43].

Another innovative methodology emerged from the Time-Phased Model [41],

which accounted for temporal progression during a patient's Intensive Care Unit (ICU) stay. The model divided the ICU timeline into three distinct phases—early (0–9 hours), middle (10–49 hours), and late (50+ hours)—and tailored predictive strategies for each. Gradient boosting models such as LightGBM were applied in the early and middle phases, while a Recurrent Neural Network (RNN) was employed in the late phase to better capture long-range dependencies. This hybrid approach achieved a utility score of 0.4149 through 10-fold cross-validation.

1.3.2 Deep Learning Methods

In contrast to classical machine learning methods, deep learning (DL) approaches have gained significant popularity due to their ability to automatically learn hierarchical features from raw or minimally preprocessed data, especially when operating on large-scale datasets. As data volumes increase, manual feature engineering becomes both computationally prohibitive and impractical, particularly when attempting to generalize across a wide array of medical conditions. Overall, while classical ML methods offer interpretability and domain-driven insight, DL approaches provide superior accuracy, scalability, and automation—making them a preferable choice in large-scale diagnostic systems for CAD. Deep neural networks, particularly those tailored for time series and sequential data, have demonstrated remarkable efficacy in cardiac signal classification. Common architectures include [15]:

- **Convolutional Neural Networks (CNNs):** These excel at capturing local patterns and morphological features in quasi-periodic signals such as ECG, PPG, and PCG.
- **Recurrent Neural Networks (RNNs):** RNNs and their variants, such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), are well-suited for modeling temporal dependencies and sequential structures.
- **Residual Networks (ResNets):** These architectures utilize skip connections to mitigate vanishing gradient issues and have shown superior performance in

classifying physiological signals, particularly when deeper networks are warranted.

CNN-based approaches have demonstrated strong performance in automatic feature extraction and classification tasks, achieving cardiologist-level accuracy in ECG arrhythmia detection [46, 47]. To capture temporal dependencies inherent in physiological signals, RNN-based models, especially Long Short-Term Memory (LSTM) networks, have been employed, often in combination with CNNs for improved performance [48]. Such hybrid architectures leverage both spatial and temporal features of the signals. In the context of heart sound analysis, ensemble and deep learning methods have also shown promising results, as demonstrated in the PhysioNet Challenge 2016 [24, 25]. Furthermore, deep residual networks (ResNet) enable the training of deeper architectures by mitigating vanishing gradient issues, thereby improving feature representation in complex biomedical signals [49]. More recently, deep learning techniques have also been successfully applied to PPG signals for non-invasive cardiac monitoring and arrhythmia detection [50].

Recent advances (2022–2025) in deep learning have significantly improved coronary artery disease (CAD) detection by leveraging architectures such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Residual Networks (ResNet). CNN-based models have demonstrated superior capability in automatically extracting hierarchical spatial features from ECG and imaging data, achieving high diagnostic accuracy (up to 99%) in CAD prediction tasks [51, 52]. Comparative studies indicate that CNNs outperform traditional neural networks and LSTM models in discriminative feature learning due to their strong local pattern extraction ability [53]. However, LSTM networks excel in modeling long-term temporal dependencies in sequential cardiac signals, making them particularly effective for capturing subtle temporal variations associated with ischemic conditions, although often with slightly lower classification accuracy compared to CNNs [51]. ResNet architectures further enhance performance by introducing residual connections that enable deeper networks while mitigating vanishing gradient issues, leading to improved generalization and robustness; recent studies

report that ResNet-based models outperform standard CNNs in ECG classification tasks with accuracy exceeding 98% [54, 55]. Moreover, hybrid architectures such as CNN-LSTM have emerged as a powerful approach, combining spatial feature extraction with temporal modeling and achieving superior balance between sensitivity and specificity compared to standalone models [51, 56]. Overall, CNNs are preferred for feature extraction, LSTMs for temporal modeling, and ResNet for deep hierarchical learning, while hybrid models consistently deliver the best performance in modern CAD detection systems.

In recent years, state-of-the-art research in Parkinson Disease (PD) detection has progressively shifted from traditional SL methods toward deep learning (DL) approaches. A variety of DL architectures have been explored for PD classification, including Multi-Layer Perceptrons (MLP), deep 1D convolutional networks, Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) models [57–60]. Notably, these deep learning models have demonstrated remarkably high performance, often reporting sensitivity, specificity, and accuracy values of approximately 98%. Pham *et al.* even reported perfect classification performance, achieving 100% sensitivity and specificity [59]. Another emerging direction in DL involves the use of pre-trained architectures such as MobileNet and Xception, which are initially trained on large-scale image datasets [61]. Although swarm-based optimization techniques have been successfully applied to other types of medical data, their application in PD classification remains largely unexplored [62, 63]. Furthermore, most of these DL approaches lack interpretability and fail to provide meaningful explanations for their predictions.

Despite the success of deep learning (DL) in various biomedical domains, its adoption in sepsis prediction remains relatively limited. This is primarily due to the inherent challenges associated with clinical time-series data, which are often irregularly sampled, sparse, and noisy [64, 65]. Intensive Care Unit (ICU) datasets, such as *MIMIC-III* and the *eICU Collaborative Research Database*, contain heterogeneous patient records with substantial missing values, necessitating extensive preprocessing and imputation [66, 67]. Furthermore, sepsis labeling is complex and often retrospec-

tive, relying on clinical definitions such as Sepsis-3, which introduces ambiguity and noise in the ground truth [68]. The relatively small size of available datasets, coupled with significant class imbalance between septic and non-septic cases, further limits the effectiveness of data-hungry DL models [65]. Additionally, the need for interpretability in critical care settings discourages the use of black-box DL models, as clinicians require transparent and explainable predictions for decision-making. Real-time deployment constraints and the absence of substantial performance gains over simpler models further contribute to the limited prevalence of DL-based approaches in sepsis prediction.

A notable deep learning method in sepsis prediction presents the Time-Phased Model [69], which explicitly examined ICU length of stay (ICULOS) alongside the sepsis onset time (t_{sepsis}). This approach segmented the ICU timeline into three phases—early (0–9 hours), middle (10–49 hours), and late (50+ hours)—and applied phase-specific models. LightGBM was employed for the early and middle phases, while an RNN was utilized for the late phase, resulting in a utility score of 0.4149 under 10-fold cross-validation.

1.4 Research Gaps

This section captures the existing research gaps in the state of the art of CAD classification, PD classification, and sepsis prediction.

1.4.1 Coronary Artery Disease Classification

Coronary Artery Disease (CAD) detection is a well-established and extensively studied problem in the literature, with researchers proposing a wide range of innovative methodologies over the past few decades. Despite this progress, several notable research gaps remain.

Firstly, classical machine learning techniques—often referred to as shallow learning (SL) methods—are computationally efficient but have reached a stage of performance stagnation. While newer feature extraction techniques have been introduced, they typically yield only incremental improvements with limited overall gains. More

importantly, the fundamental processing pipeline in SL-based approaches has largely remained unchanged, resulting in a predominantly trial-and-error–driven research paradigm.

Secondly, the inherent interrelationships among features have not been adequately explored. In most existing approaches, features are treated as independent numerical entities, and SL methods are applied directly to these representations. This simplification overlooks the underlying structural dependencies between features, leading to a potential loss of valuable information that could otherwise be leveraged for improved classification performance.

Finally, there has been very limited adoption of graph signal processing techniques in the state of the art for CAD classification. Given the potential of such methods to model and exploit complex feature interdependencies, their underutilization represents a significant gap and an opportunity for further research.

1.4.2 Parkinson Disease Classification

At the risk of repetition, the limitations identified in the previous subsections—namely, the predominance of a trial-and-error–driven research paradigm, the lack of systematic exploitation of interdependencies among state-of-the-art features, and the absence of graph signal processing–based approaches—also persist as major research gaps in Parkinson's disease (PD) classification.

Furthermore, the vertical ground reaction force (VGRF) signals used in PD analysis comprise sixteen parallel channels, resulting in a highly complex and interdependent feature space. This complexity is significantly greater than that encountered in CAD classification, thereby necessitating more sophisticated modeling frameworks capable of capturing these intricate relationships.

Finally, the current state of the art lacks robust and interpretable classification techniques for PD. Existing approaches primarily emphasize predictive performance, with limited focus on explainability, thereby restricting their clinical applicability. Addressing this gap by developing explainable PD classification methods constitutes another key objective of this thesis.

1.4.3 Sepsis Prediction

Sepsis prediction is a relatively recent research problem compared to CAD and PD classification, and it presents a distinct set of challenges. One of the primary difficulties lies in the handling of missing data, which has been extensively addressed in the state of the art through various imputation and modeling strategies. However, an equally critical issue arises from the inherent class imbalance (data skewness) present in sepsis datasets. Despite this, most existing studies predominantly rely on conventional, textbook loss functions without adequately accounting for this imbalance. In our view, this represents a significant gap in the current literature. Accordingly, this thesis aims to address this limitation by exploring and developing more suitable loss functions tailored to the skewed nature of sepsis prediction data.

Furthermore, there is a notable lack of emphasis on the explainability of sepsis detection models. While predictive performance has been the primary focus in existing studies, limited attention has been given to interpreting model decisions and ensuring clinical transparency. This gap is particularly critical in healthcare applications, where trust, interpretability, and actionable insights are essential.

1.5 Contributions of the Thesis

This thesis presents five major contributions toward disease classification using multimodal medical data:

1. A novel framework is developed to exploit the interrelationships among state-of-the-art features through graph and hypergraph representations, enabling a structured understanding of feature dependencies.

2. New discriminative features are derived from the constructed graph and hypergraph structures, enhancing the representational capacity beyond conventional feature engineering approaches.

3. An explainable hypergraph-based model is proposed, wherein the interpretability of the learned representations is systematically validated using feature importance analysis and relevant clinical use cases.

4. A hybrid classification pipeline is designed by integrating shallow learning and deep learning techniques, thereby preserving the interpretability of shallow models while leveraging the superior predictive performance of deep architectures.

5. We introduce a novel loss function with multiple tunable parameters designed to address class imbalance. The proposed formulation emphasizes hard-to-classify samples while also contributing to improved explainability.

1.6 Organization of the Thesis

This thesis is structured to progressively build on the use of graph-based and deep learning methodologies for classification and prediction tasks across a range of medical conditions, utilizing both unimodal and multimodal physiological signals. A chapter-wise summary is outlined below to guide the reader through the logical flow and contributions of the work:

- **Chapter 2** presents the theoretical foundations reused in the following chapters of this thesis. Here, we present the graph theory basics and the concept of explainability.

- **Chapter 3** explores unimodal CAD classification using photoplethysmogram (PPG). We propose a graphical construct for each subject using the features extracted from the single modality and metadata. Then, we derive spectral features from it to classify CAD.

- **Chapter 4** builds upon the work in Chapter 3 by introducing a hypergraph-based model that captures complex relationships in multimodal physiological data (photoplethysmogram (PPG) and phonocardiogram (PCG)). This chapter extends the graph construction paradigm to handle heterogeneous signal sources for improved CAD classification accuracy and robustness. Here, we propose a novel hypergraph for each subject, using features extracted from multiple physiological signals and metadata.

- **Chapter 5** explores multimodal Parkinson Disease (PD) classification, using VGRF (Vertical Ground Reaction Force), metadata, and gait speed. Here, we propose a novel hybrid framework that combines the interpretability of hypergraph-based representations with the high learning capacity of deep neural networks. This integrated approach aims to improve classification performance while maintaining clinical transparency.

- **Chapter 6** addresses the early prediction of sepsis using a combination of multimodal time-series data and categorical clinical variables. We design a prediction framework that leverages engineered explainable features and optimizes the training process using a focal-weighted Binary Cross Entropy Loss function to better handle class imbalance and enhance prediction reliability in critical care environments.

- **Chapter 7** concludes the thesis by summarizing the major findings and methodological contributions across all chapters. We also outline potential directions for future research, particularly emphasizing the potential roles of graph neural network (GNN), higher-dimension medical data, Large Language Model (LLM) in medical signal analysis.

Chapter 2

Theoretical Foundations

> *If you don't get this elementary, but mildly unnatural, mathematics of elementary probability into your repertoire, then you go through a long life like a one-legged man in an ass-kicking contest.*

Charlie Munger, Speech at USC Business School, 1994

This chapter explores basic graph theory and the concept of explainability foundational to this thesis. Key concepts discussed include the construction of graphs and hypergraphs, and state-of-the-art explainability approaches such as LIME and SHAP.

In this chapter, we present the fundamentals of a few topics we use extensively in the following chapters. The reader should consider the following sections as a prerequisite before proceeding towards the rest of the thesis. We cover multiple topics such as basics of Graph Theory, Explainability in Artificial Intelligence etc.

2.1 Graph Theory Basics

In this subsection, we place the definitions of the technical terms [70] used in the rest of the chapter.

Definition 1 (Graph)**.** *A Graph $G = (V, E)$ is a set of vertices $V = v_1, v_2, \ldots v_m$; connected by a set of edges $E = e_w, e_2, \ldots e_n$. A weighted graph has a real numeric value w associated with each edge E, also known as the weight of the edge E. Hence,*

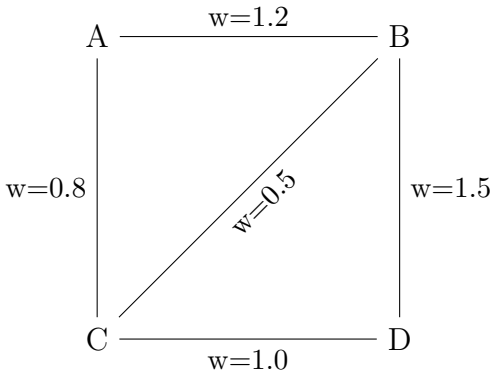


Figure 2.1: An example of a graph

a weighted graph is denoted as $G = (V, E, W)$, where W is the set of weights. Please note that all the graphs used in the following chapters are weighted graphs. In figure 2.1, we present a graph with four nodes and five edges. Each edge has edge weight assigned with it.

Definition 2 (Degree Matrix)**.** *Given a graph $G = (V, E)$ the degree matrix D is defined as a $|V| \times |V|$ matrix with the following entries:*

$$d_{i,j} = \begin{cases} degree(v_i) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

where $degree(v_i)$ denotes the number of edges connected to the vertex v_i.

Definition 3 (Incidence Matrix)**.** *For an undirected graph $G = (V, E)$, the Incidence matrix is defined as:*

$$B_{ij} = \begin{cases} 1 & \text{if vertex } v_i \text{ is incident to edge } e_j \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

Definition 4 (Adjacency Matrix)**.** *Adjacency Matrix for a graph $G = (V, E)$ is defined as a $|V| \times |V|$ matrix A with the following entries:*

$$a_{i,j} = \begin{cases} 1 & \text{if } \exists \text{ an edge connecting } v_i \text{ and } v_j \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

Definition 5 (Laplacian Matrix)**.** *Laplacian Matrix for a graph $G = (V, E)$, where*

$|V| = n$ is defined as the following:

$$L = D - A \tag{2.4}$$

where D is the degree matrix and A is the adjacent matrix.

Furthermore, we introduce some more definitions concerning the structure of a graph.

Definition 6 (Induced Subgraph)**.** *Let $G = (V, E)$ be an undirected graph. For a subset of vertices $V' \subseteq V$, the* induced subgraph *$G[V']$ is the graph whose vertex set is V' and whose edge set is*

$$E' = \{(u, v) \in E \mid u \in V',\ v \in V'\}. \tag{2.5}$$

That is, $G[V']$ contains all edges of G that have both endpoints in V'.

Definition 7 (Clique)**.** *A subset of vertices $C \subseteq V$ in a graph $G = (V, E)$ is called a* clique *if the induced subgraph $G[C]$ is a complete graph. In other words, every pair of distinct vertices in C is connected by an edge:*

$$\forall u, v \in C,\ u \neq v \Rightarrow (u, v) \in E. \tag{2.6}$$

Definition 8 (Hyperedge)**.** *A hyperedge e is defined as an edge which connects a set of vertices $V = v_1, v_2, \ldots v_p$, where $p \geq 3$.*

In figure 2.2, we present a hypergraph with four nodes, five edges and two hyperedges. Each edge and hyperedge has edge weight assigned with it.

Definition 9 (Hypergraph)**.** *A hypergraph G is defined as an edge $G = (V, E)$ where V is a set of vertices and E is a non-empty set of hyperedges. A weighted hypergraph has a real number w associated with each hyperedge E, also known as the weight of the hyperedge E. Hence, a weighted hypergraph is denoted as $G = (V, E, W)$, where W is the set of weights. Please note that all the hypergraphs used in the presented approach are weighted hypergraphs.*

Definition 10 (Incidence Matrix of Hypergraph)**.** *An incidence matrix of a hypergraph $G = (V, E)$ is defined by a $|V| \times |E|$ matrix H with the following entries:*

$$h_{v,e} = \begin{cases} 1 & \text{if } v \in e \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

where v and e denote a vertex and hyperedge respectively.

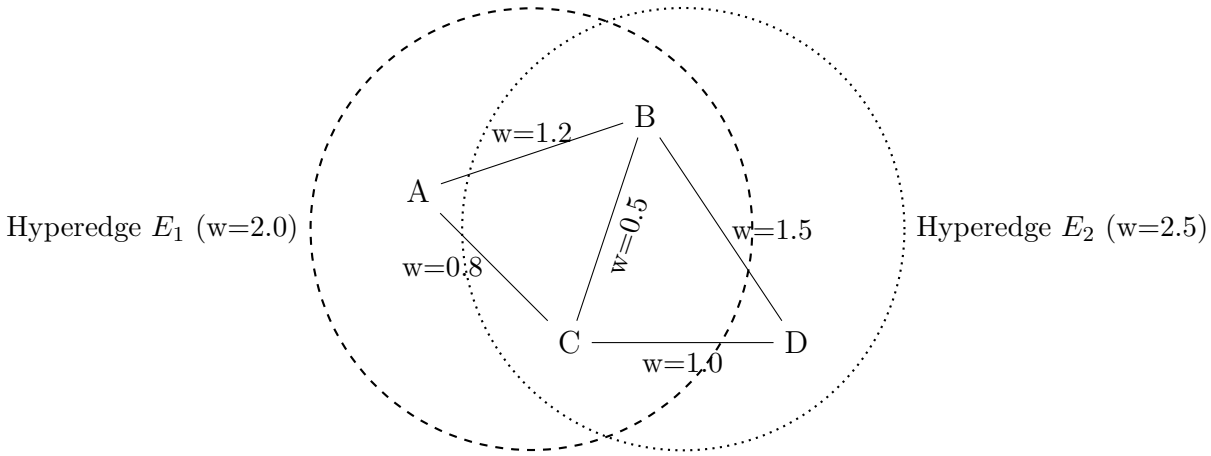


Figure 2.2: An example of a hypergraph

Definition 11 (Adjacency Matrix of Hypergraph)**.** *Let W denote the diagonal matrix containing the weights of hyperedges. Adjacency Matrix of a hypergraph G is defined as:*

$$A = HWH^T - D_v \tag{2.8}$$

where D_v denotes the diagonal matrix containing the vertex degrees.

Definition 12 (Laplacian Matrix of Hypergraph)**.** *Laplacian of a hypergraph G is defined as:*

$$\Delta = \frac{1}{2}(I - D_v^{-1/2} A D_v^{-1/2}) \tag{2.9}$$

where I denotes the identity matrix.

2.2 Explainability

With the advent of high performance Deep Learning (DL), the machine learning community has become more successful and hence, deep learning is presently more

popular compared to shallow learning techniques. However, this new approach introduced the world to a hugely complex chain of logic, learning millions or even billions of parameters. It is practically impossible to retrace the decisions taken by the deep learning architectures. In parallel, the legal implications of the decisions taken by DL were unclear, but important, especially for AI used in medical sciences. European GDPR (General Data Protection Regulation) restricts to use of black-box approaches affecting human life [71].

Explainable AI (xAI) was coined in 2019 by DARPA [72], and since then caught the eye of the AI researchers. The term is self-explanatory in the true sense of the word. It makes AI explainable and tries to reveal and/or introspect why the AI behaved in a certain manner. Needless to say, this form of AI is instrumental in enhancing the trust factor needed by the domain experts, i.e., the physicians and nurses in the field of medical science.

The researchers have explored numerous avenues to reach xAI. Holzinger et. al reviewed 17 methods: LIME, Anchors, GraphLIME, LRP, DTD, PDA, TCAV, XGNN, SHAP, ASV, Break-Down, Shapley Flow, Textual Explanations of Visual Models, Integrated Gradients, Causal Models, Meaningful Perturbations, and X-NeSyL [73]. However, detailing each of them is out of scope for this thesis. Out of these, LIME (Local Interpretable Model Agnostic Explanations) and SHAP (SHapley Additive exPlanations) stand out as the most popular methods, owing to their effectiveness and ease of deployment. In this chapter, we shall limit our discussions on these two techniques.

2.3 LIME (Local Interpretable Model-agnostic Explanations)

LIME (Local Interpretable Model-agnostic Explanations) is an explainability technique designed to make complex, black-box machine learning models more understandable [74]. It provides human-interpretable explanations for individual predictions by approximating the model's behavior in the vicinity of a specific data instance. The method is "model-agnostic," meaning it can be applied to any machine

learning model, regardless of architecture.

Many advanced models, such as deep neural networks, gradient-boosting machines, and ensemble methods, achieve high accuracy but are often difficult to interpret. This lack of transparency makes it challenging to trust the model's decisions, especially in critical fields, such as, healthcare, finance, and law. LIME addresses this problem by generating simple, interpretable explanations that approximate how the model arrives at a specific decision.

LIME creates local approximations of a complex model's decision-making process. It operates in the following steps:

1. Select an Instance to Explain: The first step is choosing the data point (instance) for which an explanation is required.

2. Generate Perturbations: LIME creates a set of new, synthetic data points by perturbing (slightly altering) the original instance. These perturbed samples are labeled using the complex model.

3. Assign Weights: LIME assigns higher weights to the perturbed samples that are closer to the original instance and lower weights to those farther away.

4. Train a Simple Interpretable Model: A simple, human-understandable model (such as a linear regression or decision tree) is trained on the perturbed data. This model approximates the complex model's decision boundary for the chosen instance.

5. Generate an Explanation: The interpretable model provides feature importance scores that explain why the complex model made a specific prediction for the given instance.

LIME explains decisions for specific instances rather than providing a global model explanation. Model-Agnostic Nature: It can be applied to any type of machine learning model, including deep learning, random forests, and support vector machines. LIME can be used for various data types, including tabular, image, and

text data. Limitations of LIME include approximation errors. LIME uses a simplified model that may not fully capture the original model's complexity. Furthermore, the quality of explanations depends on how perturbations are generated. Most importantly, the explanations churned out by LIME are only accurate for the local instance, not necessarily for the whole model.

LIME is a powerful tool for interpreting black-box machine learning models, making AI more transparent and trustworthy. However, it should be used with caution, considering its limitations and ensuring it aligns with the specific interpretability needs of an application.

2.4 SHAP (SHapley Additive exPlanations)

SHAP (SHapley Additive exPlanations) is a model-agnostic technique used to explain the predictions of machine learning models [75]. It is based on Shapley values, a concept from cooperative game theory, which fairly distributes a model's prediction among its input features. Unlike other explanation methods, SHAP provides consistent and theoretically sound feature importance scores, ensuring that the most influential features receive the highest attribution.

Shapley values originate from cooperative game theory, where players contribute differently to the total game outcome. In machine learning:

- The "players" are the input features.
- The "game" is the model's prediction.
- The goal is to fairly distribute the prediction among the features, considering their individual contributions.

Mathematically, a feature's Shapley value is computed by averaging its marginal contributions across all possible feature subsets. This ensures that each feature's contribution is fairly assigned, making SHAP one of the most rigorous explainability techniques.

SHAP explains a model's prediction by comparing it to a baseline prediction, typically the average model output. The process follows these steps:

1. Select an Instance: Choose the data point for which we want an explanation.
2. Generate Feature Combinations: Consider different subsets of input features.
3. Compute Model Predictions: Measure how adding each feature changes the prediction.
4. Average the Contributions: Calculate each feature's impact across all possible subsets to determine its Shapley value.
5. Provide an Explanation: The SHAP values show how much each feature contributes to pushing the prediction away from the baseline.

SHAP ensures that more important features always get higher scores. It can explain both individual predictions and overall model behavior. It is based on game theory, making it mathematically rigorous. Unlike LIME, SHAP considers feature interactions and handles correlated features better.

SHAP provides different implementations depending on the complexity of the model. Some of these rae listed below:

- Kernel SHAP: A model-agnostic version that approximates Shapley values using sampling.
- Tree SHAP: Optimized for tree-based models (e.g., XGBoost, LightGBM).
- Deep SHAP: Designed for deep learning models.
- Linear SHAP: Works efficiently with linear models.

Calculating exact Shapley values is expensive, especially for large datasets. It requires more computation than LIME. Unlike LIME, which trains a simple model, SHAP computes contributions for all feature subsets, making it slower. Some implementations (like Kernel SHAP) approximate the values, which may introduce minor inaccuracies.

SHAP is one of the most powerful and theoretically sound methods for interpreting machine learning models. By fairly distributing prediction contributions

among input features, it ensures transparency and trustworthiness in AI decision-making. While computationally expensive, its advantages in accuracy and fairness makes it a preferred choice in critical fields such as healthcare, finance, and legal AI applications.

Chapter 3

Coronary Artery Disease Classification

> *The problem with heart disease is that the first symptom is often fatal.*

Michael Phelps, World Heart Day, 2023

The development of machine learning techniques paved the way for automated disease classification, enabling timely interventions. These techniques require non-invasive Photoplethysmogram (PPG) collected from the fingertip, paving the way for unobtrusive 24×7 *in-home monitoring to be possible.*

In this chapter, we target Coronary Artery Disease classification (CAD) using graphs. Here we consider CAD from a single modality of physiological sensors, namely Photoplethysmogram (PPG). We build a graph for each subject and derived more distinctive features, nothcing up the classfication performance.

3.1 Introduction

Coronary Artery Disease (CAD) is a medical condition of developing plaques inside coronary artery. CAD is also known as Ischemic Heart Disease (IHD) and accounts for almost 16% of global death, thereby making it the most common cause of human mortality [76]. Though, in the developed countries, prevalence of CAD has dipped in the last 30 years [77], it stands as a major concern for the developing world. As true

for most of the cardiac problems, CAD results from gradual and systemic deposition of foreign body material over a long period of time. Hence, early detection of CAD is imperative for preventive healthcare. This chapter presents a new methodology of feature representation technique for classifying CAD.

The clinical gold standard for CAD estimation is angiography, which uses a special dye and camera to take pictures of blood flow. This is an invasive procedure and requires continuous supervision of experienced clinicians. In order to provide an alternative, researchers analyzed several non-invasive physiological signals including (but not limited to) Photoplethysmogram (PPG) and Phonocardiogram (PCG). There are proven fusion techniques which used multiple sensor signals in conjunction and performed feature or decision level fusion. However, multiple sensors offer reduced usability, and hence may not always suit the unobtrusive 24×7 use case. In this chapter, we have modified the way features are represented traditionally and decided to limit this work to only one sensor - PPG. PPG is a very popular non-invasive signal which measures volumetric blood flow in capillaries. Low-cost pulse oximeters are available off-the-shelf which are capable of recording, storing and/ or streaming continuous PPG. Banerjee et al. [1] demonstrated the efficacy of wide variation of features extracted from the PPG signal in classifying CAD. The features included time and frequency domain parameters extracted from PPG. For example, Heart Rate Variability (HRV) related features are known to have significant medical significance - moderately high HRV is a favorable indication of normalcy [78], whereas, extremely low HRV is an indicator to multiple medical conditions including cardiac issues and high stress. Angius et al. introduced more features from PPG, such as, relative crest time to identify cardiovascular disorder in hospital dataset [79].

Graph signal processing (GSP) has been applied in recent times to solve problems in different domains. Sandryhaila *et al.* demonstrated the effectiveness of frequency component derived from graph signals in semi-supervised classification in sensor networks as well as image and text datasets [80]. Researchers also used graph theory to analyze complex biological system [81], e.g., applying cluster analysis on various networks.

In this chapter, we extend the knowledge of GSP to introspect the dependencies of PPG features and metadata features, such as, age, weight, height and gender with respect to CAD. We have two major contributions in this work, which are enlisted below:

- The first contribution lies in the representation of cardiac health as a biologically interpretable graph with available PPG features and metadata features.
- The second contribution is the formulation of a spectral feature vector using the top three eigenvalues for classifying the above set of graphs into CAD and non-CAD. To the best of our knowledge, this is the first approach to classify CAD using GSP till date.

3.2 Cardiac Health as Graph

The novelty of the present work lies in representing the features in a more biologically meaningful manner. In particular, we represent cardiac health/patient data in form of graphs (CHG). In order to achieve that, we have followed a set of strategies detailed in the following subsections.

3.2.1 Biomedical Relation amongst Physiological Features

In this step, we form a graph with the subset of features extracted from a physiological signal which in our case is the PPG signal. The features are chosen in such a way that each subset of features satisfies the following properties: (1) they individually proved to be essential and found to have significant correlation with CAD, and (2) they originate from a *single biological phenomenon.*

In the proposed approach, these features are connected as a *complete weighted graph* $G = (V, E)$, where V and E represent the vertices and edges (Fig. 3.1). Each node/vertex represents a feature. An edge between any two vertices indicates that the features mutually influence each other. The weight $w(e_{ij})$ of such an edge e_{ij} is deemed as the product of the strengths $(\phi(i), \phi(j))$ of the nodes (i, j) it connects.

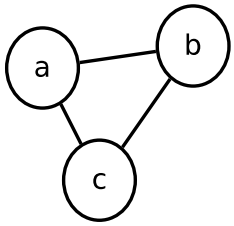


Figure 3.1: A complete graph where nodes are features calculated from a single biological phenomenon

Please note that $\phi(i)$ is the node weight of i, which in turn is the normalized value of the i^{th} base feature. So, we write:

$$w(e_{ij}) = \phi(i) * \phi(j) \tag{3.1}$$

3.2.2 Graph with PPG Features

In the previous section, we constitute a complete graph which connects features originated from a single biological property of PPG signal. However, for a typical physiological signal, such as PPG, there will be multiple such biological properties, each of which can culminate into various features. Hence, we need to combine multiple such isolated graphs (each representing a disjoint group of features) in a single graph (say CHG_1). Therefore, CHG_1 essentially captures the state of the art features as well as their mutual connections, effectively representing the cardiac health of a person in the cohort.

Let us assume there are multiple such isolated graphs. Let the i^{th} graph be denoted by $G_i = (V_i, E_i)$, where V_i and E_i are the vertices and edges of G_i respectively. Each G_i represents an unique phenomenon related to cardiac health and V_i represents the features associated to that phenomenon. To keep the graph CHG_1 connected, we introduce a non-existent node (say, a *ghost node*, g). To keep the effect of ghost node minimal, following constructions are undertaken: (1) the ghost node g is connected to all the nodes of all the isolated graphs and (2) the new edges resulting form these connections are assigned insignificantly small weights. So, $CHG_1 = (\mathcal{V}_1, \mathcal{E}_1)$ such that $\mathcal{V}_1 = \bigcup_i V_i \cup \{g\}$, and $\mathcal{E}_1 = \bigcup_i E_i \cup \bigcup_j e_{gj}$, where, e_{gj}

is the edge connecting the ghost node g with the node j. In other words, each G_i becomes an *induced subgrah* of CHG_1. Moreover, since each G_i is also a complete graph, every G_i is a *clique* of CHG_1. Fig. 3.2 shows a *representative* CHG_1 where a ghost node g connects to all the nodes of three subgraphs, namely, G_p, G_q and G_r.

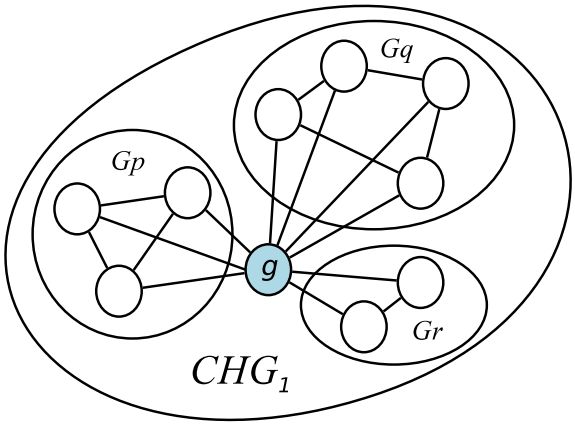
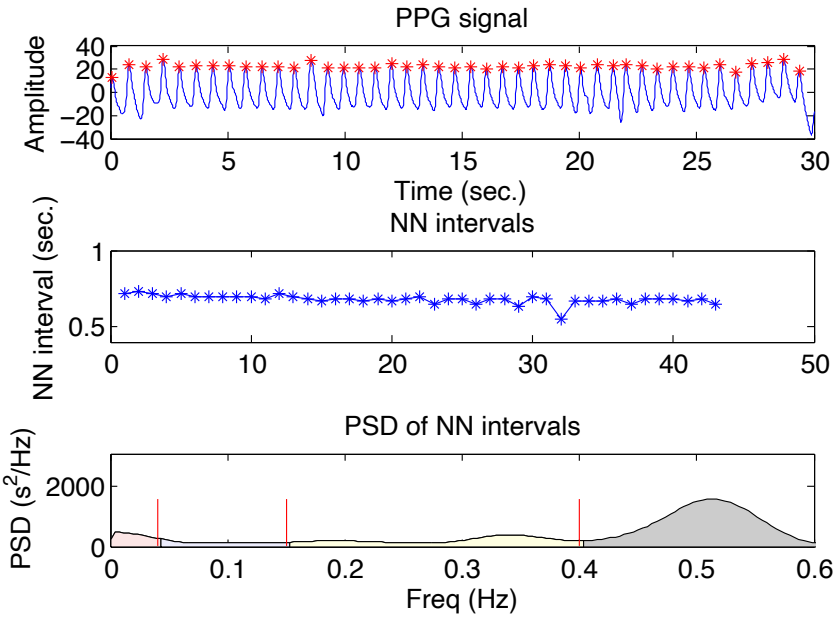


Figure 3.2: A *representative* CHG_1 where a ghost node is connected to three subgraphs

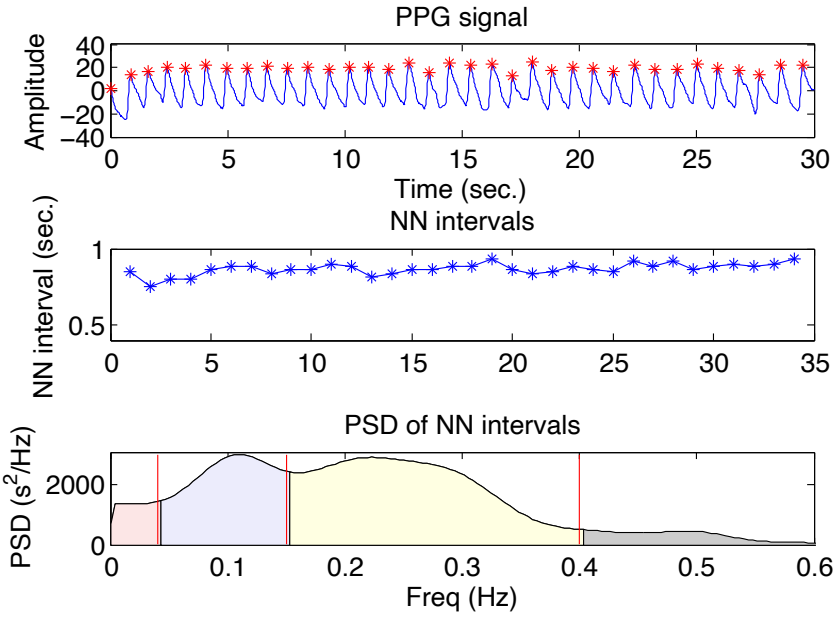
We consider the features presented in [1] and [79] for forming CHG_1. For example, the first subgraph G_1 is constructed using the features which are related to the powers in different frequency ranges of NN intervals in a PPG signal (Fig. 3.3). An NN interval is measured by the time interval between two consecutive normal heartbeats. The nodes of G_1 are normalized spectral power in the following ranges: 0 to 0.04 Hz (f_1), 0.04 Hz to 0.15 Hz (f_2) and 0.15 to 0.50 Hz (f_3). Hence $G_1 = (V_1, E_1)$ where $V_1 = \{f_1, f_2, f_3\}$ and $E_1 = \{e_{12}, e_{13}, e_{23}\}$. The node weights are calculated according to Eqn. 3.1.

Similarly, the subsequent subgraphs are constructed in the following manner - mean and standard deviation of cardiac cycle (G_2: f_4, f_5), mean and standard deviation of systole duration normalized to cardiac cycle (G_3: f_6, f_7), mean and standard deviation of diastole duration normalized to cardiac cycle (G_4: f_8, f_9), mean and standard deviations of ratio of systolic time to diastolic time (G_5: f_{10}, f_{11}), mean and standard deviation of peak to peak distances (G_6: f_{12}, f_{13}), mean and standard deviation of systole and diastole durations (G_7: f_{14}, f_{15}) and mean and standard deviation of systole and diastole durations (G_8: f_{16}, f_{17}).

In this way, we form the graph CHG_1, with a total of 17 nodes (f_1-f_{17}; each node



(a) CAD subject



(b) non-CAD subject

Figure 3.3: Sample PSD of nn intervals; obtained from PPG, for a CAD and non-CAD subject [1]

representing a PPG feature) distributed in 8 subgraphs (G_1 to G_8; each subgraph representing different biological phenomenon). All the 17 feature nodes are also

connected to one ghost node.

3.2.3 Graph with PPG Features and Metadata Features

We now show how a more informative graph CHG_2 can be built with the inclusion of metadata features such as age, weight and height. The metadata features have the following properties: (1) they individually do not necessarily show a very strong relation with CAD, and (2) in conjunction with the PPG features they may prove to emerge as significant differentiating factor. For example, only higher body weight may not necessarily correlate with CAD as a tall person tend to weigh more. However, both high weight and lesser energy in HRV tend to correlate with CAD. Consideration of such conditions is expected to inherently strengthen the CAD vs. non-CAD classification problem [82].

We first create a *complete graph* $G_m = (V_m, E_m)$ from scratch, having nodes as metadata features. We now discuss the construction of CHG_2 where the ghost node g in CHG_1 is replaced by G_m, where each node of G_m is connected to rest of the nodes of CHG_1. Hence, G_m becomes the *metadata subgraph* of CHG_2. So, $CHG_2 = (\mathcal{V}_2, \mathcal{E}_2)$ such that $\mathcal{V}_2 = \bigcup_i V_i \cup V_m$ and $\mathcal{E}_2 = \bigcup_i E_i \bigcup_{j \in V_i, k \in V_m} e_{jk} \cup E_m$ where e_{jk} is the edge connecting a node j in V_i with a node k in V_m. All the edge weights are recalculated according to Eqn. 3.1.

A *representative* CHG_2 having subgraphs G_m, G_1 and G_2 are shown in Fig. 3.4. Please note that the CHG_2 used in the proposed solution contains 17 PPG features (in G_1 to G_8) and 3 metadata features (in G_m). Further it is worth noting that G_m and each G_i is a clique in CHG_2.

3.2.4 Spectral Features for Graph Classification

Our goal is to classify the cohort into two groups, namely, CAD and non-CAD. In order to achieve this task, spectral measures are employed in the present work. Spectral measures has proven to be effective cardiac screening research area [83]. After representing each patient data using the final graph (CHG_2), we have computed the graph Laplacian and obtained its eigenvalues. Let, $\mathcal{L}(CHG_2)$, $\mathcal{D}(CHG_2)$ and

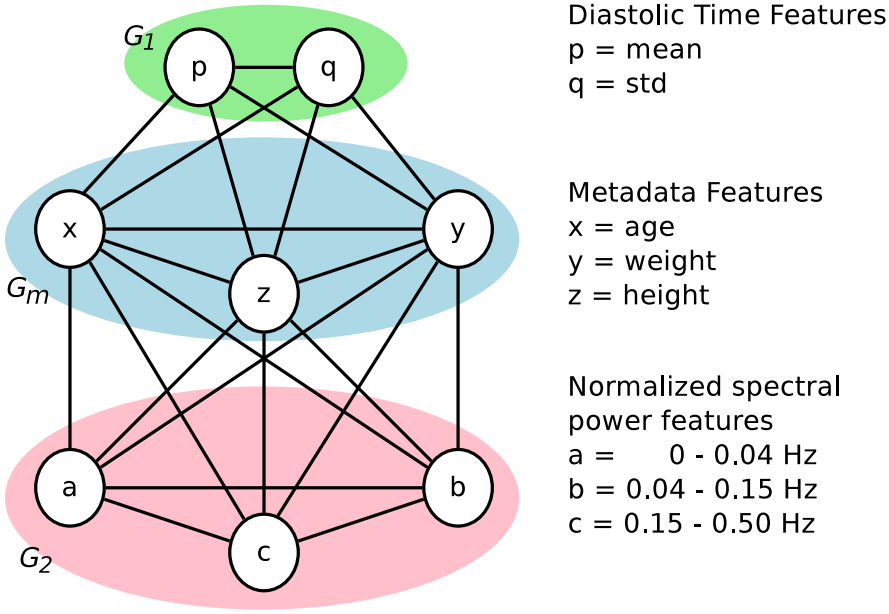
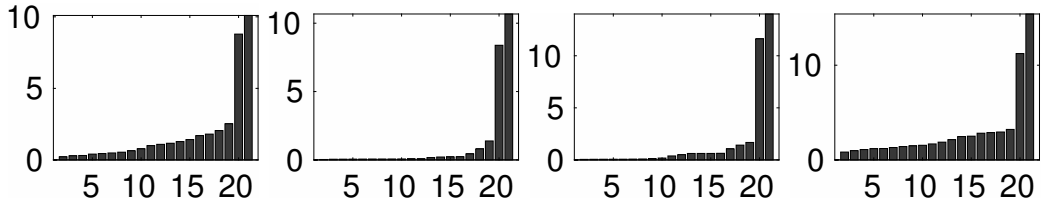


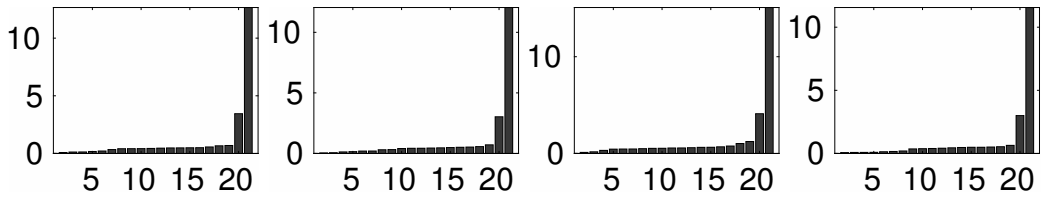
Figure 3.4: *Representative* CHG_2 where metadata features are connected to all nodes of two subgraphs

$\mathcal{A}(CHG_2)$ denote the Laplacian, diagonal degree and weighted adjacency matrices of CHG_2 respectively. The Laplacian matrix of a graph is defined as the difference between its diagonal degree matrix and adjacency matrix [84]. So, we can write:

$$\mathcal{L}(CHG_2) = \mathcal{D}(CHG_2) - \mathcal{A}(CHG_2) \tag{3.2}$$

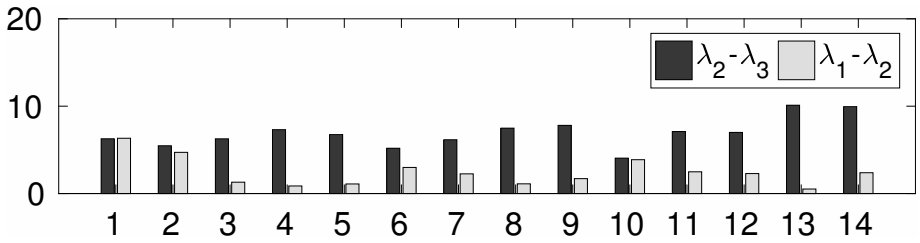


(a) CAD subjects

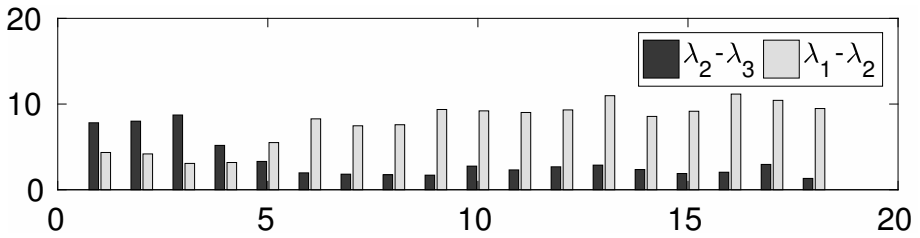


(b) non-CAD subjects

Figure 3.5: Sample Eigenvalues (sorted by increasing amplitude); obtained from CHG_2, for four CAD and four non-CAD subjects



(a) CAD participants



(b) Non-CAD participants

Figure 3.6: Eigenvalue Analysis: graph spectral feature vector $\psi(CHG_2)$; obtained using Eqn. 3.3

Studies suggest that maximum energy tends to get stored in few largest eigenvalues [85]. As shown in Fig. 3.5(a) and Fig. 3.5(b), for each CHG_2, CAD or otherwise, more than 50% of the energy is stored in the top three eigenvalues. It is worth noting that the difference between the first and second eigenvalues as well as that of the second and third eigenvalues exhibit very distinct characteristics for CAD and non-CAD participants. Therefore, we propose a derived feature vector $\psi(CHG_2)$ to capture the above mentioned characteristics of the three eigenvalues $(\lambda_i, i = 1, \cdots 3)$, which in turn encapsulates the distinctive spectral properties of the graph CHG_2.

$$\psi(CHG_2) = [\lambda_1 - \lambda_2, \lambda_2 - \lambda_3] \tag{3.3}$$

Variations of the elements of $\psi(CHG_2)$ for CAD and non-CAD classes are respectively demonstrated in Fig. 3.6(a) and Fig. 3.6(b). Finally, k-means algorithm is applied on the above spectral feature vectors for classification.

3.3 Experiments and Results

3.3.1 Dataset

All our experiments were conducted on PPG data and meta data collected from 32 consenting participants. Please note that we could not use MIMIC dataset for evaluation as used in [1], as metadata features were absent in MIMIC. The data collection protocol has been reviewed by the respective IRB (Institution Review Board) of Tata Consultancy Services and Fortis Hospital, Kolkata, India. FDA approved medical grade oximeter (CMS 50D+) was used for data collection. The ground truth of CAD in each participant was obtained through angiogram, which is the current gold standard for blockage detection. Metadata such as age, weight, height were also measured. Final dataset of 32 participants (12 female and 20 male) had age variation of 47 ± 21 years, weight variation of 68 ± 10 kilograms and height variation of 164 ± 9 cm.

3.3.2 Protocol

In this work, we have taken features extracted from PPG and explored their internal dependencies using graphs. A novel way of applying graph signal processing technique has been deployed where mutual relations amongst the features were retained, leading to richer (more discriminatory) features.

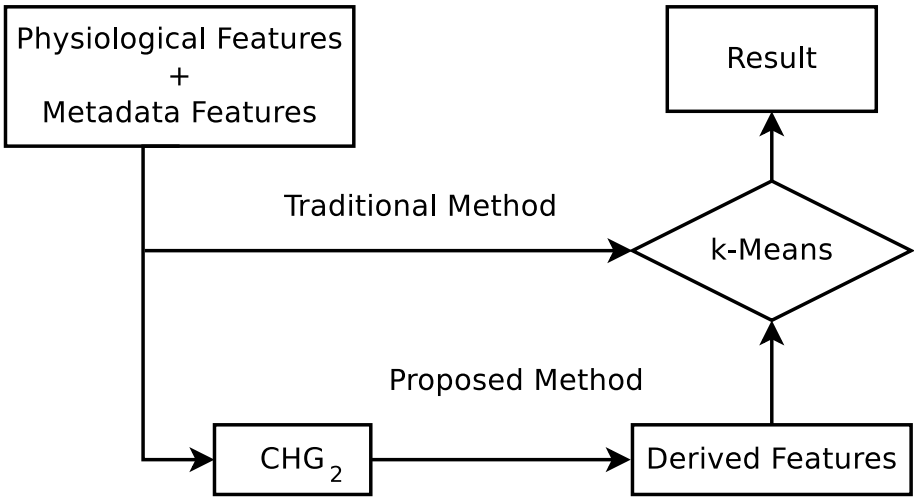


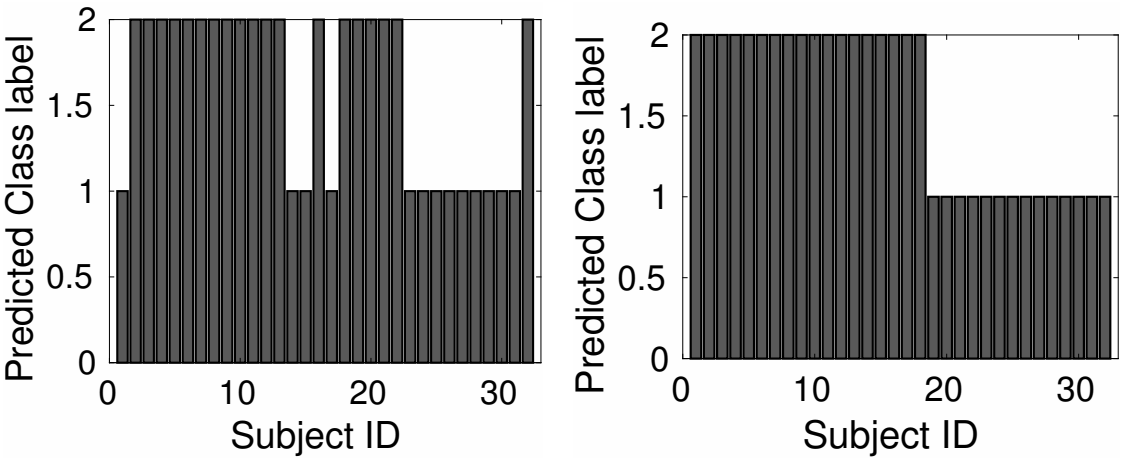
Figure 3.7: Validation Technique of CAD classification using PPG

Hence, a straight forward validation of the proposed method would be to compare the performance of an unsupervised machine learning technique on the proposed features $\psi(CHG_2)$ against that of the traditional features $(f_1, \cdots f_{17})$. Since, this is a control and case study for the condition of CAD, we essentially deal with two classes. So, in this work, k-means clustering technique is used with k=2. In [86], Halkidi et al. proposed external, internal and relative criteria to measure cluster validity. Since the ground truth classes are known, those were used as external criteria in this work. The detailed block diagram of the validation methodology is presented in Fig. 3.7.

3.3.3 Performance Comparisons

As shown in Fig. 3.8, CHG_2 resulted in 88% overall accuracy, as compared to 72% when the features (collection of the PPG features and metadata features) were fed into k-means as-is (simple aggregation), which we deem as the *baseline method*. This is because simple aggregation fails to exploit the dependencies within the PPG features, within the metadata features and between the PPG and the metadata features. In sharp contrast, our construction of CHG_2 (Fig. 3.4) incorporates all the above dependencies successfully. Moreover, the proposed spectral features $\psi(CHG_2)$ properly capture the difference in behavior of the CAD and the non-CAD classes (Fig. 3.6).

In our application, sensitivity (S_e) measures the algorithm's ability to correctly identify a subject with CAD as positive, whereas specificity (S_p) reflects the ability to determine the subject who does not have CAD as negative. Hence, in this context, higher the S_e, lower the probability of a CAD patient being wrongly diagnosed. In other words, lower S_e would lead to more number of sick people missing the net, and hence can prove to be fatal. On the other hand, higher the S_p, lower the probability that a non-CAD participant will unnecessarily go through the invasive test. However, the cost implication for lower S_p is purely procedural and financial, and not fatal. Hence, in the context of preventive healthcare, higher sensitivity is a preferable condition compared to higher specificity. As shown in Tables 3.1 and



(a) Traditional features (b) Proposed features

Figure 3.8: k-means clustering performance: Prediction on 32 subjects sorted according to ground truth class label; *Subjects 1-14 belong to class 2 and subjects 17-32 belong to class 1.*

3.2, both True Positive and True Negative have improved in the proposed approach over the baseline method, resulting in better sensitivity (1.00) as well as specificity (0.78). It is to be noted that for any screening system, sensitivity closer to unity is required before mass scale deployment.

Table 3.1: Confusion Matrix of baseline method

Traditional Feature		*Predicted*	
		CAD	non CAD
Actual	CAD	12	2
	non-CAD	7	11

Table 3.2: Confusion Matrix of CHG_2

CHG_2		*Predicted*	
		CAD	non CAD
Actual	CAD	14	0
	non-CAD	5	13

In addition to the above baseline approach, we have compared our method with two state-of-the-art techniques [79] and [1]. Note that both the above methods are supervised in nature employing leave-one-out cross validation (LOOCV). In spite of

having the advantage of supervised classification, those two methods yielded 60% and 80% accuracy respectively. In Table 3.3, we show the sensitivity, specificity and accuracy of the above supervised methods, our CHG_2 and the baseline. The results indicate that CHG_2 supersedes [79], [1] and the baseline method by yielding best accuracy and sensitivity values. In terms of specificity values, our method is the second best.

Table 3.3: Comparison of Sensitivity (S_e), Specificity (S_p) and Accuracy (A_{cc}) Values

Method	***Classification***	***Se***	***Sp***	***Acc***
[5]	Supervised LOOCV	0.53	0.67	0.60
[3]	Supervised LOOCV	0.73	0.87	0.80
baseline	Unsupervised	0.86	0.61	0.72
CHG_2	Unsupervised	1.00	0.78	0.88

3.4 Discussion

In this chapter, we have addressed non-invasive cardiac health analysis using PPG features and metadata features. We investigated the feasibility of graph based feature representation leveraging their internal dependencies. We have further shown that spectral features derived from the graph eigenvalues result in improved classification accuracy of CAD and non-CAD. In the future, we plan to incorporate additional metadata related to sedentary lifestyle in our graph to further improve the results. Extension of the proposed representation to problems in other domains, such as mental health prediction can be explored as well.

Chapter 4

Multimodal Coronary Artery Disease Classification

> *The important graphs are the ones where some things are not connected to some other things. When the unenlightened ones try to be profound, they draw endless verbal comparisons . . . until their graph is fully connected and also totally useless.*

Eliezer S. Yudkowsky, The Virtue of Narrowness, 2007

Multimodal physiological signal captures the fingerprint of Coronary Artery Disease. Using multiple signals make the classification process more robust and accurate. Formation of novel hypergraphs using features extracted from various signals and metadata help us explore intricate connections among them, which were not analyzed before in the state-of-the-art research.

4.1 Introduction

Over the last century, the landscape of global health has undergone a dramatic transformation. The advancement of medical science has led to a substantial decline in mortality rates associated with communicable diseases (CDs), such as tuberculosis and measles. Even more recent infectious threats—including Human Immunodeficiency Virus (HIV), Zika virus, and Ebola—are now managed more effectively

with regard to their global mortality impact. In contrast, the widespread adoption of sedentary lifestyles has contributed to a surge in non-communicable diseases (NCDs), which now represent the leading cause of death worldwide [3].

Unlike CDs, which are often acute and preventable, NCDs are chronic conditions influenced by a complex interplay of genetic, physiological, environmental, and behavioral factors. Among these, cardiovascular diseases (CVDs) have emerged as particularly lethal. As reported by both the World Health Organization (WHO) [4] and the American Heart Association (AHA) [5], CVDs account for approximately 17.9 million deaths annually—representing 31% of total global mortality and nearly half of all deaths attributed to NCDs, which collectively comprise 71% of worldwide fatalities [6]. In figure 4.1, we can observe that the prevalence of cardiac disease is more than 5%-6% for almost all the geographies globally.

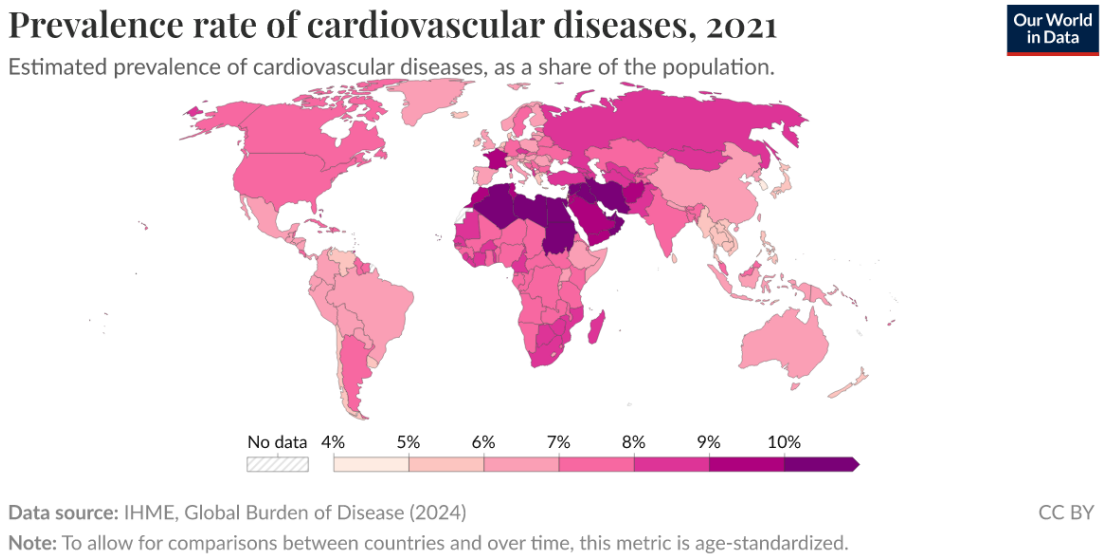


Figure 4.1: Prevalence of Cardiovascular Diseases[1]

Coronary Artery Disease (CAD), a prevalent subset of cardiovascular disease (CVD), involves the narrowing or blockage of coronary arteries—the blood vessels responsible for supplying oxygen-rich blood to the heart muscle. The primary pathological mechanism underlying CAD is the accumulation of fatty deposits, or atherosclerotic plaques, on the inner walls of the coronary arteries. Over time, this

[1] https://ourworldindata.org/grapher/prevalence-rate-of-cardiovascular-disease?time=2021

condition may extend to other blood vessels throughout the body, contributing to systemic vascular dysfunction [87].

The current clinical gold standard for CAD diagnosis is coronary angiography, an invasive imaging technique that involves the injection of contrast dye into the bloodstream to visualize arterial blockages. However, the invasive nature of this procedure, coupled with its high cost—ranging from approximately USD 300–500 in India to nearly USD 5,000 in the United States—highlights the pressing need for a low-cost, non-invasive, and scalable screening alternative.

In response to this need, researchers have explored the use of non-invasive physiological signals—such as photoplethysmograms (PPG), phonocardiograms (PCG), and electrocardiograms (ECG)—to detect CAD with encouraging levels of accuracy. For instance, Banerjee *et al.* [23] demonstrated that incorporating demographic and clinical metadata (e.g., age, blood pressure, body mass index, chest pain characteristics) can yield a CAD detection sensitivity and specificity of 92% and 90%, respectively. Furthermore, the PhysioNet Computing in Cardiology Challenge 2016 [24] curated a comprehensive heart sound dataset from major hospitals worldwide to facilitate the classification of abnormal PCGs—a category encompassing various cardiac pathologies including CAD and valvular disorders. The top-performing model, which employed a modified AdaBoost algorithm, achieved an F1 score exceeding 86% [25].

A wide range of machine learning methodologies have since been applied to this task, including support vector machines, random forests, and ensemble strategies utilizing linear, non-linear, and bootstrapped models [23,88,89]. More recently, deep learning approaches have shown promise in extracting informative representations from cardiac time-series data.

In our prior work [90], we showed how Graph Signal Processing (GSP) can enhance CAD detection using features derived from PPG signals. By representing these features as nodes in a graph, we were able to model their interrelationships and derive new graph-based descriptors. However, the proposed method had two primary limitations: (1) while it effectively captured biologically driven feature de-

pendencies, it did not scale well to include other relevant interrelationships, such as, statistical similarity across features, and (2) the graph construction was restricted to a single physiological signal, thereby limiting the potential benefits of multimodal integration.

In the present study, we address these limitations by extending our GSP framework to incorporate multiple physiological signals. Signal processing literature often emphasizes that corroborating evidence from multiple modalities enhances the accuracy and robustness of inference. Accordingly, we analyze both the biological and statistical interdependencies among features extracted from the PCG and the PPG signals and introduce a novel hypergraph-based representation that captures these complex relationships. To the best of our knowledge, this is the first work of its kind which employs a hypergraph framework for modeling features derived from heterogeneous time-series cardiac signals.

The concept of hypergraphs—where edges are capable of connecting more than two nodes—has been established for over three decades [91]. However, significant developments in hypergraph theory and its associated tools and operators have emerged primarily within the past decade. For instance, the introduction of the hypergraph Laplacian by Zhou et al. [70] marked a notable advancement. Traditional approaches employing hypergraphs typically involve modeling the entire population of records within a single hypergraph structure, followed by operations, such as clustering, cuts, partitioning, and embedding.

In contrast, our work adopts a fundamentally different problem formulation. Specifically, we leverage multiple modalities for each subject to construct corresponding hypergraphs. This results in a collection of homogeneous hypergraphs [92] —one per subject—with consistent structural configurations but varying edge weights (*i.e.*, connection strengths among nodes). Subsequently, we aim to classify subjects into two categories: those with coronary artery disease (CAD) and those without.

The key contributions of this chapter are as follows:

1. We design hyperedges that connect nodes based on diverse properties, such as biological origin and statistical relationships, culminating in a novel hy-

pergraph architecture that effectively captures the cardiac health profile of individuals.

2. We propose a richer set of derived features, specifically hypergraph Laplacians, extending beyond conventional feature sets extracted from physiological signals.

4.2 Multimodal Signal as Hypergraph

Features extracted from multimodal physiological signals can encompass a diverse range of representations, including time-domain characteristics, frequency-domain properties, wavelet transforms, and morphological descriptors. Typically, subsets of these features are derived from specific biological phenomena; for example, multiple representations may be obtained from beat-to-beat cardiac cycles. Furthermore, when features are extracted across a sequence of temporal windows—such as features computed for each cardiac cycle during a session of multiomodal physiological signals such as photoplethysmogram (PPG) or phonocardiogram (PCG)—various statistical aggregation techniques are employed to condense the temporal feature set into a singular representative value. Common aggregators include measures, such as, mean, median, kurtosis, and centroid, among others.

The primary novelty of the present work lies in proposing a more biologically and statistically meaningful framework for representing features derived from PCG and PPG signals. Specifically, we extend the concept of Cardiac Health as a Graph, previously introduced in [90], by introducing a new paradigm: Cardiac Health as a Hypergraph. This enhanced representation leverages the rich relational structure of hypergraphs to better capture the complex interdependencies between physiological features, ultimately facilitating a more nuanced assessment of cardiac health.

4.2.1 Biomedical, Statistical Representations of Features

In this step, we form different statistical representations of features extracted from multimodal physiological signals (PPG and PCG signals).

PPG

We analyze and consider the following features [23] extracted from a session of PPG: normalized spectral power in the following ranges: 0 to 0.04 Hz (F_{ppg_1}), 0.04 Hz to 0.15 Hz (F_{ppg_2}) and 0.15 to 0.50 Hz (F_{ppg_3}); mean and standard deviation of cardiac cycle (F_{ppg_4}, F_{ppg_5}); mean of systole and diastole duration normalized to cardiac cycle (F_{ppg_6}, F_{ppg_7}) and root mean square of cardiac cycle (F_{ppg_8}).

$$F_{ppg} = \begin{bmatrix} F_{ppg_1} & F_{ppg_2} & \cdots & F_{ppg_8} \end{bmatrix}$$

PCG

For a comprehensive analysis of phonocardiogram (PCG) signals, it is essential to accurately extract the fundamental heart sounds. To achieve this, we employed the heart sound segmentation algorithm proposed by Liang *et al.* [87], which facilitates the reliable detection of the first (S1) and second (S2) heart sounds. Once the cardiac cycles were segmented, a set of features was calculated for each individual cardiac cycle within a session of PCG recording. These features include: the ratio of spectral features (f_{spec}); the spectral centroid ($f_{centroid}$), roll-off ($f_{rolloff}$), and spectral flux (f_{flux}); the kurtosis (f_{kurt}) of the time-domain signal; and entropy-based measures such as Natural Entropy (f_{nat}) and Tsallis Entropy ($f_{tsallis}$), as introduced in [23].

Each session typically contains multiple cardiac cycles, with the exact number varying according to the individual's heart rate and the duration of the PCG recording. Consequently, a mechanism is required to summarize the set of feature values obtained across all cardiac cycles into a single representative feature per session. To this end, we apply statistical aggregation techniques, specifically calculating the mean, median, and variance of the feature arrays (e.g., f_{spec}), where the size of each array corresponds to the number of cardiac cycles within the session. The aggregated statistical features are then compiled to form a new feature set, denoted as F_{spec}, which is an array of length three representing the mean, median, and variance of f_{spec}.

Similarly, for the features extracted per cardiac cycle ($f_{pcg_{1-7}}$), the mean, me-

dian, and variance were computed to generate intermediate aggregated features. As a result, the features extracted from PCG signals were systematically extended, enriching the representation of cardiac dynamics at the session level. Let F_{spec} denote an array of length three containing these aggregated statistics, with analogous formulations applied to the other features extracted from the cardiac cycles.

Mean, *median* and *variance* of $f_{pcg_{1-7}}$ were calculated as intermediate features. Hence, the features extracted from PCG got extended to the following. Let F_{spec} be an array of length 3,

$$F_{spec} = \begin{bmatrix} F_{spec_1} \\ F_{spec_2} \\ F_{spec_3} \end{bmatrix} = \begin{bmatrix} mean(f_{spec}) \\ median(f_{spec}) \\ variance(f_{spec}) \end{bmatrix} \tag{4.1}$$

and consequently, F_{spec}, $F_{centroid}$, $F_{rolloff}$, F_{flux}, F_{kurt}, F_{nat} and $F_{tsallis}$ are calculated in a similar manner from f_{spec}, $f_{centroid}$, $f_{rolloff}$, f_{flux}, f_{kurt}, f_{nat} and $f_{tsallis}$ respectively. Finally, this leads us to the the entire PCG feature set

$$F_{pcg} = \begin{bmatrix} F_{spec_1} & F_{spec_2} & F_{spec_3} \\ F_{centroid_1} & F_{centroid_2} & F_{centroid_3} \\ F_{rolloff_1} & F_{rolloff_2} & F_{rolloff_3} \\ F_{flux_1} & F_{flux_2} & F_{flux_3} \\ F_{kurt_1} & F_{kurt_2} & F_{kurt_3} \\ F_{nat_1} & F_{nat_2} & F_{nat_3} \\ F_{tsallis_1} & F_{tsallis_2} & F_{tsallis_3} \end{bmatrix} \tag{4.2}$$

4.2.2 Hyperedges connecting PPG and PCG Features

In our previous work [90], we introduced a *complete weighted graph* denoted as $G = (V, E)$, where V represents the set of vertices and E denotes the set of edges. Each vertex in this graph corresponded to a distinct physiological feature, and an edge was established between two vertices only if the corresponding features were derived from the same underlying biological phenomenon. This biologically motivated connectivity ensured that the graph structure reflected inherent physiological

relationships among features.

Building upon this foundational concept, the present work proposes an extension, wherein, features originating from a common biological source are not merely connected pairwise but grouped collectively through a single connection, giving rise to a *hyperedge*. This transformation leads to the construction of a hypergraph, wherein, each hyperedge can connect more than two vertices simultaneously, thereby capturing higher-order relationships among groups of related features.

As illustrated in Figure 4.2, an example of such a hyperedge-based representation can be observed in the case of photoplethysmography (PPG) features. Specifically, the spectral power features extracted from different frequency bands—namely F_{ppg_1}, F_{ppg_2}, and F_{ppg_3} (described in detail in Section 4.2.1)—are jointly connected via hyperedge e_1, as they all pertain to spectral characteristics of the PPG signal. Similarly, the remaining PPG features, F_{ppg_4} through F_{ppg_8}, are derived from time-domain characteristics across various segments of the cardiac cycle. These features are grouped together via hyperedge e_2 in Figure 4.2, reflecting their shared temporal origin.

This transition from pairwise graph-based modeling to hypergraph-based representation enables a more expressive and biologically faithful characterization of the inter-feature relationships within and across physiological modalities.

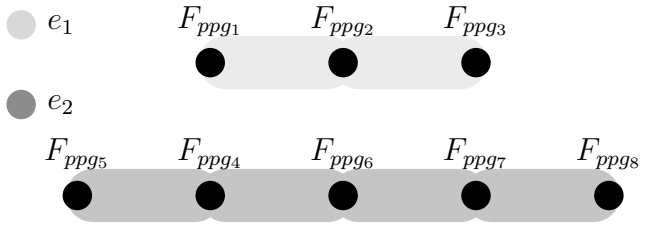


Figure 4.2: Two sample hyperedges from PPG features: e_1 connecting spectral powers of different ranges from PPG and e_2 connecting all the features related to cardiac cycle

This work further extends the aforementioned hypergraph-based modeling paradigm by incorporating connections not only among features derived from similar biological phenomena, but also among those representing different statistical aggregators of the same underlying signal property. This enables the representation of

intra-feature relationships arising from statistical summarization over time.

For instance, as illustrated in Figure 4.3, we construct a hyperedge, denoted as e_1, that connects the three elements of the array F_{spec}—each representing a distinct statistical aggregate (mean, median, and variance) of the spectral features extracted from a session of phonocardiogram (PCG), as described in Section 4.2.1. Similarly, all features related to the spectral centroid, computed through different statistical aggregation techniques and collectively forming the array $F_{centroid}$, are grouped via the hyperedge e_2.

This design is systematically applied across additional sets of features. Specifically, we construct five additional hyperedges to capture the aggregated forms of the following feature arrays: $F_{rolloff}$, F_{flux}, F_{kurt}, F_{nat}, and $F_{tsallis}$—each of which encapsulates a distinct aspect of the PCG signal, computed over multiple cardiac cycles and summarized using statistical measures.

By constructing hyperedges in this manner, the resulting hypergraph architecture effectively models both biological origins and statistical relationships among features, leading to a more holistic and structured representation of cardiac acoustic data.

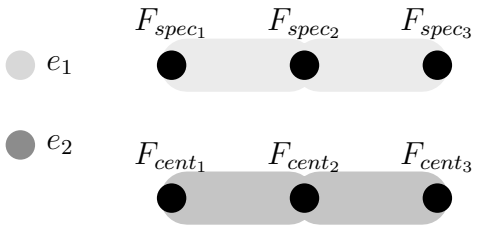


Figure 4.3: Two sample hyperedges from PCG features: e_1 connecting the statistical aggregators of spectral power ratios and e_2 connecting the statistical aggregators of spectral centroids

Hence, we have described nine hyperedges in this subsection: two hyperedges connecting the PPG features and seven hyperedges connecting the PCG features.

4.2.3 Hypergraph Formation: Connecting PPG and PCG Features to Metadata

In this subsection, we present the final hypergraph structure designed to represent each individual subject in the dataset. The primary objective is to construct a *hy-*

pergraph architecture capable of accommodating the various types of hyperedges described in Section 4.2.2, thereby providing a comprehensive and interpretable model of subject-specific data. Following relations are considered:

1. Several key relational structures are incorporated into the proposed hypergraph. Notably, metadata features, which often exhibit a strong correlation with cardiac health risks, are explicitly modeled. For instance, an individual's body weight, when considered in isolation, can serve as a crude indicator of cardiac risk. However, a more meaningful measure emerges when weight is normalized by height, resulting in the widely accepted Body Mass Index (BMI),

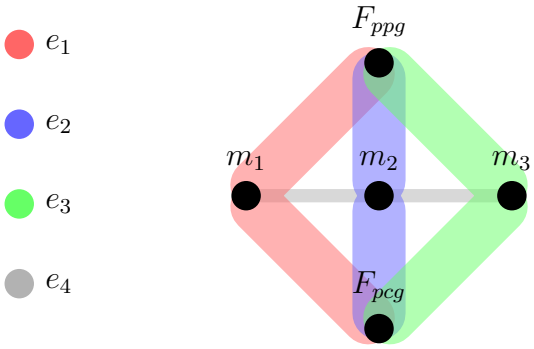


Figure 4.4: Four hyperedges which connect metadata to PPG and PCG features: one hyperedge ($e4$) connecting all the metadata and the three hyperedges ($e1$, $e2$ and $e3$) each consisting of one metadata, PPG and PCG features (F_{ppg} and F_{pcg})

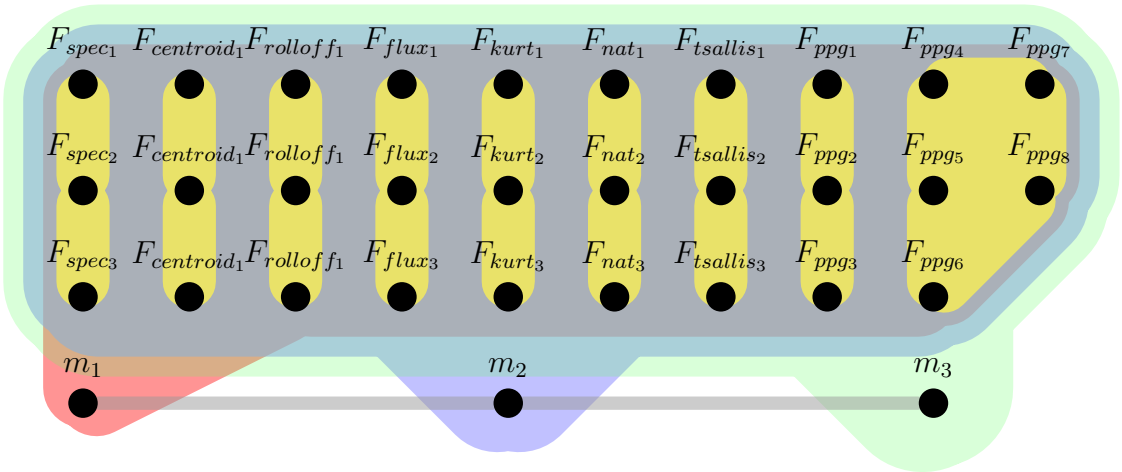


Figure 4.5: Fully populated hypergraph; each node is a feature; each yellow hyperedge connects features sharing a singular physiological signal; red, blue and green hyperedges connect one metadata and all physiological features and the grey hyperedge connects all metadata features

defined as:

$$\text{BMI} = \frac{\text{Weight (kg)}}{(\text{Height (m)})^2}$$

Clinically, a BMI greater than thirty is categorized as obesity, a condition known to significantly elevate the risk of developing cardiac problems. Similarly, age serves as a critical factor in risk stratification; while conditions such as obesity are deleterious across all age groups, their impact becomes markedly more severe in older individuals compared to younger counterparts.

Accordingly, in the proposed hypergraph structure, a set of metadata features, denoted as F_{meta}, is constructed as follows:

$$F_{meta} = \begin{bmatrix} weight \\ height \\ age \end{bmatrix} \tag{4.3}$$

Each element—weight (m_1), height (m_2), and age (m_3)—is represented as a distinct vertex in the hypergraph. Vertex weights are assigned based on the normalized values of the corresponding features to ensure comparability across different scales. As depicted in Figure 4.4, these three vertices are interconnected via a hyperedge, labeled e_4, capturing their joint influence on cardiac risk profiles.

This integration of critical metadata into the hypergraph structure enhances its capacity to model both physiological signals and associated risk factors, providing a richer and more holistic representation of each subject's cardiac health status.

2. Beyond their individual contributions, metadata features can play a critical role when considered in conjunction with features derived from physiological signals. For example, an elevated resting heart rate, when observed alongside low heart rate variability (HRV) and excessive body weight, is a well-

established compound risk indicator for cardiovascular disease. Such combinatorial relationships underscore the importance of modeling interactions between metadata and signal-derived features.

To capture these interactions within the hypergraph framework, we introduce hyperedge e_1, which connects the metadata feature corresponding to body weight (m_1) to the full set of features extracted from photoplethysmography (PPG) and phonocardiogram (PCG) signals. These physiological features, denoted as F_{ppg} and F_{pcg} respectively (as defined in Section 4.2.1), represent dynamic aspects of cardiac function that may be modulated or contextualized by metadata such as weight.

3. In a similar fashion, and as depicted in Figure 4.4, additional hyperedges—namely e_2 and e_3—are constructed to link the other metadata features to the signal-derived feature sets. Specifically, hyperedge e_2 connects height (m_2) to the features in F_{ppg}, while hyperedge e_3 connects age (m_3) to the features in F_{pcg}. These connections enable the hypergraph to represent higher-order interactions between individual-specific physiological measurements and demographic characteristics, thereby facilitating more nuanced analysis of cardiovascular health risk.

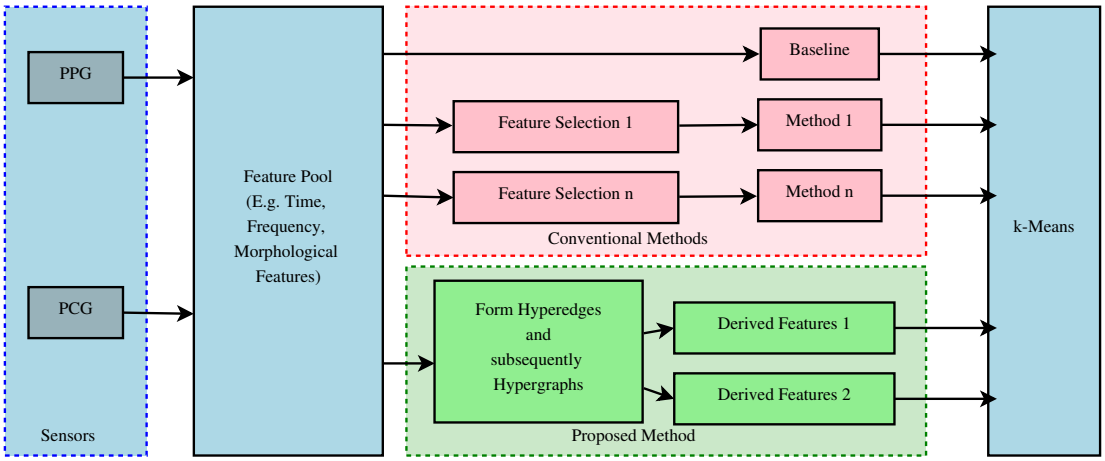


Figure 4.6: Flow diagram for Multimodal CAD Classification Performance Validation Protocol

Each subject within the dataset is represented by a distinct hypergraph structure (refer to Fig. 4.5). This hypergraph comprises thirty-two vertices and thirteen hyperedges. These thirteen hyperedges can be categorized as follows: (a) nine hyperedges, each linking vertices derived from physiological signals (such as PPG or PCG), (b) three hyperedges, where each one connects all physiological features with a corresponding metadata, and (c) one hyperedge that interlinks all metadata.

It is important to note that all hypergraphs in the dataset share an identical structural configuration, including the number of vertices, the number of hyperedges, the cardinality of each hyperedge, and the vertices contained within each hyperedge. The only variation across the hypergraphs lies in the edge weights. Each hypergraph represents data from a single subject.

The weight of an edge, denoted as e_{ij} connecting nodes i and j with respective node weights n_i and n_j respectively, is computed using the following expression:

$$w(e_{ij}) = e^{(-1)\times|n_i - n_j|} \tag{4.4}$$

Furthermore, the weight of a hyperedge, $w(h_e)$, is defined as the sum of the weights of all edges that connect the member nodes of that particular hyperedge. If the hyperedge h_e connects the set of nodes $\{N\}$, its weight is calculated as:

$$w(h_e) = \frac{1}{2} \times \sum_{i,j \in \{N\}, i \neq j} w(e_{ij}) \tag{4.5}$$

The factor $(1/2)$ is introduced to account for the potential double counting of edge weights.

4.2.4 Derived Features

We sim to derive new features from the hypergraph we formed with state of the art features. We extract the following spectral features from the hypergraph: (a) hypergraph Laplacians and (b) hypergraph eigenvalues. In addition, as described in our previous work [90], we also compute the hypergraph Laplacians (Eqn. 2.8) and their corresponding eigenvalues for comparative analysis. To assess the effectiveness of the

hypergraph spectral features, we further include a classification approach that relies solely on the hypergraph weights. All proposed methodologies, in conjunction with state-of-the-art techniques, are evaluated using an unsupervised validation framework, as illustrated in Fig. 4.6.

4.3 Experiments and Results

4.3.1 Dataset

The experiments were conducted using PPG and PCG data collected from 86 consenting participants. The data collection protocol was reviewed and approved by the respective Institutional Review Boards (IRB) of Tata Consultancy Services (TCS) and Fortis Hospital, Kolkata, India. To capture the data, cost-effective and non-intrusive devices were employed, specifically the CMS 50D+ pulse oximeter (priced under 50 USD) and the TCS in-house digital stethoscope (priced under 10 USD). The ground truth for coronary artery disease (CAD) was established through angiography, which is recognized as the gold standard procedure in this field. The final dataset, comprising 86 participants, displayed the following demographic characteristics: an average age of 47 ± 18 years, an average weight of 64 ± 13 kilograms, and an average height of 163 ± 9 cm.

4.3.2 Performance Measurement Protocol

We extract a total of twenty-one features from the PCG signal F_{pcg}) and eight features from the PPG signal (F_{ppg}), as detailed in Sections 4.2.1 and 4.2.1, respectively. Additionally, three metadata features (F_{meta})are gathered during the data collection process, as described in Section 4.2.3. Collectively, the classification based on these thirty-two features forms the foundational classifier, which is referred to as the *baseline method* throughout the remainder of this chapter.

In the study presented in this chapter, we adopt a straightforward unsupervised validation approach, as illustrated in Fig. 4.6. All methods, including state-of-the-art techniques, the baseline method, interim/intermediate methods, and the

proposed method, begin by utilizing the aforementioned thirty-two features. It is important to note that while state-of-the-art methods directly feed these features into the classification process, the hypergraph-based methods first derive novel features from F_{ppg}, F_{pcg} and F_{meta}, and subsequently proceed with classification. The output from each method is then processed using the k-Means algorithm, where the value of k is set to two, corresponding to the two classes: CAD and non-CAD.

4.3.3 Performance Analysis

There are several metrics available for evaluating the performance of clustering algorithms. In the current work, we utilize the Silhouette Value (S), which is defined for each data point in the clusters as follows:

$$S_i = \frac{q_i - p_i}{max(p_i, q_i)}$$

Here, p_i represents the mean distance from the i^{th} point to all other points within the same cluster to which it belongs, while q_i denotes the minimum mean distance from the i^{th} point to any other points in a different cluster, minimized over all possible clusters. The Silhouette Value (S) quantifies how well each point fits within its own cluster in comparison to its proximity to points in other clusters. A higher S_i value indicates that the point is well-matched to its own cluster and poorly matched to other clusters.

Since the Silhouette Value produces a set of values, we report the mean and standard deviation of these values for each method. A higher mean value, coupled with a lower standard deviation, indicates better performance of the clustering method [93].

In the following, we present the performance of the graph spectral processing (GSP)-based methods alongside the baseline method, as summarized in Table 4.1. The baseline method achieves an accuracy of 78%, demonstrating the strong discriminative nature of the features. However, it exhibits a notable shortcoming in terms of specificity, which drops to 53%, essentially making it a random chance prediction. It is important to highlight that this result diverges significantly from those reported in the related literature, where similar features have been used. The po-

tential reasons for this discrepancy include: (1) the absence of supervised training, which prevents the model from learning the proper weighting of features, and (2) the lack of feature selection, as all features were applied without any refinement.

Upon transitioning to the graph-based spectral approaches, we observe an overall improvement in performance, with the exception of the *Graph Eigenvalues* approach. Specifically, both hypergraph-based spectral methods—namely, the Laplacian and Eigenvalues approaches—yield superior results, achieving accuracy levels of 88% and 92%, respectively, outperforming their corresponding graph-based spectral counterparts, which show accuracy rates of 74% and 90% for Laplacian and Eigenvalues methods, respectively. Additionally, the approach relying solely on hypergraph weights performs comparatively worse, with an accuracy of 80%, falling behind the hypergraph-based spectral methods.

Overall, the *Hypergraph Laplacian* method emerges as the best performer among all strategies tested. Furthermore, as indicated in Table 4.1, this method also demonstrates the highest mean and lowest standard deviation in the *Silhouette Value*, reflecting the superior clustering performance of the *Hypergraph Laplacian* approach relative to all other methods.

Table 4.1: Comparison of Graph and Hypergraph based performance against Baseline method (Sensitivity, Specificity, Accuracy, Mean and Standard deviation of the Silhouette Value)

Method	*Se*	*Sp*	*Acc*	*S*
Baseline (Raw Features)	94	53	78	39 ± 64
Graph Eigenvalues	69	82	74	32 ± 63
HyperGraph Weights	74	85	80	45 ± 60
HyperGraph Eigenvalues	94	79	88	66 ± 56
Graph Laplacian	94	82	90	70 ± 55
HyperGraph Laplacian	98	82	92	76 ± 51

Finally, we compare the *Hypergraph Laplacian* approach to several state-of-the-art methods. In a seminal study by Schmidt *et al.* [89], the method achieved only 61% accuracy. This relatively low performance can be attributed to the fact that their approach relied solely on the PCG signal, thereby omitting the PPG information, which may have limited the model's overall performance. In contrast, other

contemporary methods demonstrated significantly higher accuracy rates. Notably, the approach by Banerjee *et al.* [23], which utilized a combination of PPG, PCG, and various metadata such as age, weight, blood pressure, and chest pain, achieved impressive results. While the *Hypergraph Laplacian* approach offers marginally better accuracy than the method proposed by Banerjee *et al.*, the sensitivity of the proposed method is notably closer to unity, reaching 98%, compared to the 92% sensitivity achieved by Banerjee *et al.*

Table 4.2: Performance of proposed method against state-of-the-art methods (Sensitivity, Specificity and Accuracy)

Method	**Supervised**	**Sensors**	**Se**	**Sp**	**Acc**
Schmidt *et al.* [89]	Yes	PCG	47	82	61
Choudhury *et al.* [88]	Yes	PPG & PCG	82	83	82
Banerjee *et al.* [23]	Yes	PPG & PCG & Metadata	92	90	91
HyperGraph Laplacian	No	PPG & PCG & Metadata	98	82	92

4.4 Conclusions

In the field of non-invasive coronary artery disease (CAD) screening, most existing research focuses primarily on novel feature extraction techniques, followed by various permutations and combinations of artificial intelligence methods. However, a significant gap in these studies is the failure to leverage the interconnections between the features themselves. In this work, we introduce the concept of hyperedges, which facilitates the creation of a hypergraph representation capable of accommodating multimodal physiological data. This novel approach allows for a more comprehensive understanding of the relationships between the features.

The proposed unsupervised classifier was evaluated using PPG and PCG data collected from the subjects in a hospital setting. The results demonstrate that even being unsupervised, the proposed method outperforms all existing state-of-the-art supervised techniques. Notably, the method also achieves near-perfect sensitivity, which is critical for the effectiveness of a screening method.

Looking ahead, we plan to automate the construction process of the hypergraph to streamline the implementation of this approach. Additionally, we aim to explore

the use of kernel methods that would incorporate the proposed hypergraph structure in a supervised learning framework, potentially further enhancing the method's performance and applicability.

Chapter 5

Multimodal Parkinson Disease Classification

> *It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong.*
>
> Richard P. Feynman, The Character of Physical Law, 1965

Unobtrusive and accurate detection of Parkinson's Disease (PD) remains a critical objective in clinical and biomedical research. While numerous studies have investigated the use of Vertical Ground Reaction Force (VGRF) signals for the robust classification of PD, the majority of these approaches focus predominantly on classification accuracy, often at the expense of clinical interpretability. That is, they rarely provide medically acceptable explanations that justify the model's predictions in a manner comprehensible to healthcare professionals. The methodology proposed in this chapter addresses this gap by integrating deep learning with a hypergraph-based representation within a hybrid learning framework. This approach not only improves classification performance but also facilitates explainability through the use of hypergraphs, which inherently model complex relationships among multimodal features. Consequently, the resulting system offers both high diagnostic reliability and interpretability, thus aligning more closely with the practical needs of medical diagnosis and decision-making.

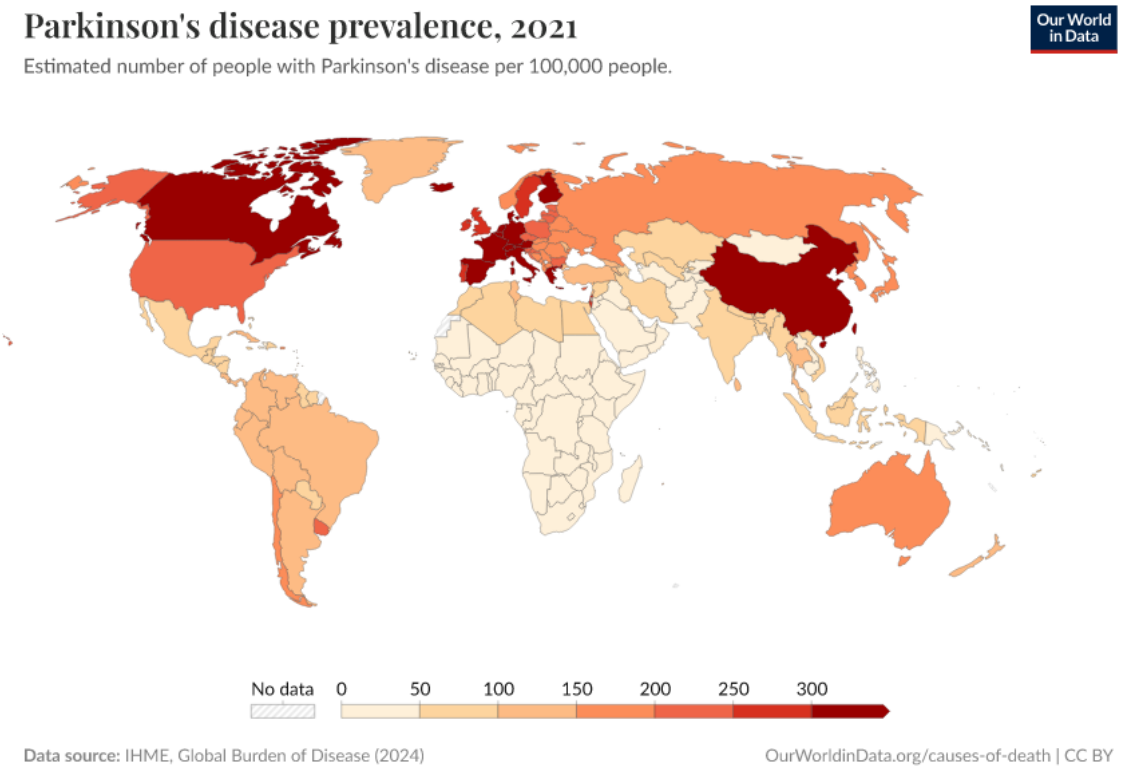


Figure 5.1: Global prevalence of Parkinson Disease[1]

5.1 Introduction

Neurological disorders represent one of the foremost global health challenges, ranking as a leading contributor to disability-adjusted life years (DALYs)—a comprehensive measure that combines years of life lost (YLL) due to premature mortality and years lived with disability (YLD). Moreover, neurological conditions are the second leading cause of mortality worldwide, accounting for approximately nine million deaths annually. Notably, Parkinson's Disease (PD) alone is estimated to be responsible for nearly two-thirds of these fatalities, making it the most lethal neuro-degenerative condition globally and the most prevalent among neuro-degenerative disorders [7]. As captured in Fig. 5.1, prevalence of Parkinson disease data is not available for a large number of the developing nations. However, if we inspect the data from the rest of geography, the prevalence is extremely high.

Patients afflicted with PD frequently exhibit distinct abnormalities in gait, which

[1] https://ourworldindata.org/grapher/parkinsons-disease-prevalence-ihme?time=2021

serve as important diagnostic markers. Early detection of such motor impairments is critical, as timely intervention can substantially enhance both quality of life and life expectancy. In recent years, there has been growing research interest in distinguishing PD patients from healthy controls through the extraction of discriminative features derived from various sensor modalities. These include voice recordings, handwriting dynamics, motion analysis, and neuroimaging techniques [8–10].

Among the available modalities, pressure sensors embedded in footwear have emerged as a promising low-cost and minimally invasive alternative. These sensors are capable of capturing high-resolution, multi-point Vertical Ground Reaction Force (VGRF) signals, which reflect subtle gait dynamics with high fidelity. VGRF offers a distinct advantage over other motion sensing technologies by reducing system complexity, energy consumption, and overall deployment costs.

In this study, we concentrate on the classification of Parkinson's Disease using VGRF signals. Our work leverages the most comprehensive and publicly accessible dataset annotated for PD—the Physionet Parkinson Dataset, originally introduced by Hausdorff et al. in 2007 [94]. For clarity, we refer to this dataset as PhysioPD2007DB throughout the remainder of this chapter.

Although visual sensing technologies have demonstrated considerable efficacy in the detection and monitoring of Parkinson's Disease (PD), their practical deployment is often hindered by significant ethical and operational challenges. Chief among these are concerns related to privacy and data security, particularly in scenarios involving continuous video monitoring or personal recordings. Moreover, visual data processing typically incurs high computational overhead, rendering real-time analysis resource-intensive and potentially impractical in low-power or embedded systems [95].

In an alternative approach, Mera et al. [96] developed a home-based PD monitoring system that utilized motion sensing through tri-axial accelerometers and gyroscopes. While effective in capturing relevant movement patterns, the system's design included a finger-worn motion sensor paired with a wrist-worn control module. Such a configuration introduces a degree of physical intrusiveness, potentially

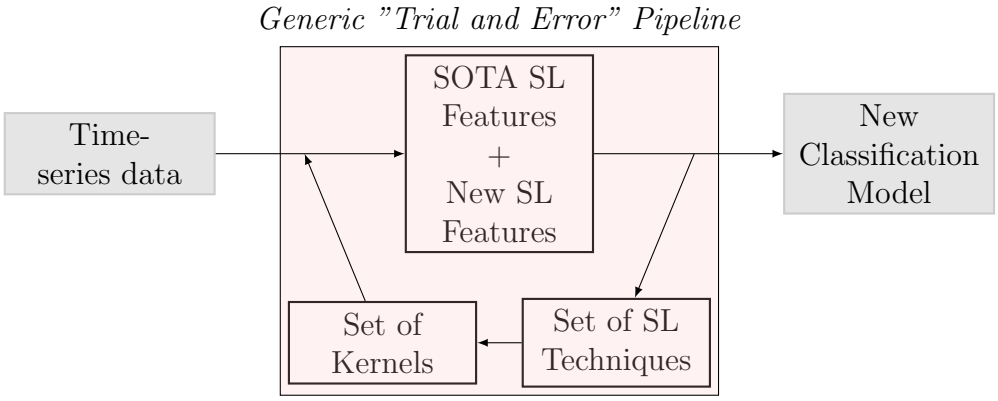


Figure 5.2: SOTA SL features are usually represented using feature vectors, ignoring the physical events, biological phenomena or statistical operations

impacting user comfort and long-term adherence, thereby limiting its applicability as a truly non-invasive monitoring solution.

The contemporary state-of-the-art (SOTA) methodologies for classifying Parkinson's Disease (PD) using Vertical Ground Reaction Force (VGRF) time-series data can be broadly classified into two principal paradigms: (i) Classical Machine Learning, herein referred to as Shallow Learning (SL), and (ii) Deep Learning (DL). SL approaches are characterized by the extraction of hand-crafted features from VGRF signals, which are subsequently utilized as input to traditional classifiers such as support vector machines, decision trees, and k-nearest neighbors [34, 40], as illustrated in Figure 5.2. In contrast, DL methodologies employ multi-layered neural networks to automatically learn discriminative feature hierarchies directly from raw or minimally processed input data, often demonstrating improved classification accuracy [57, 58, 97].

While DL techniques have shown superior performance in classification tasks, they are often criticized for their limited interpretability, opaque decision-making process, and high computational overhead. Conversely, SL models benefit from greater transparency, explainability, and computational efficiency. However, the SL paradigm commonly follows a repetitive and largely empirical pipeline. As demonstrated in Figure 5.2, SL-based studies frequently adhere to a standardized framework wherein researchers augment existing SOTA feature sets by appending newly derived features. These augmented feature sets are then subjected to an array of

feature selection techniques and classifier configurations (e.g., tuning of kernel functions, regularization parameters, or neighborhood sizes). Ultimately, the optimal configuration is typically chosen based on performance metrics derived from empirical evaluation.

Recognizing the limitations and complementary strengths of both paradigms, this chapter proposes a novel hybrid framework that integrates the interpretability of SL with the better predictive capabilities of DL. The goal is to formulate a classification approach that not only achieves high accuracy but also provides clinically relevant explanations, essential for medical decision-making. The key contributions of the work described in this chapter are enumerated as follows:

1. We systematically explore the physical principles, biological underpinnings, and statistical interdependencies that govern the relationships among existing state-of-the-art Shallow Learning (SL) features. These interconnections are formalized through the construction of an interpretable hypergraph structure, as detailed in Section 5.3.1. This hypergraph not only provides meaningful insights into feature associations but also serves as the foundation for deriving a new set of highly discriminative features, which are subsequently utilized to develop the SL component of our hybrid classification framework (Section 5.3.2).

2. In parallel, we design and implement a Deep Learning (DL) module using a modified ResNet architecture, specifically tailored to accommodate the multichannel and temporal characteristics of the VGRF time-series signals (Section 5.3.4). To effectively leverage the complementary strengths of both learning paradigms, we propose an adaptive fusion strategy that combines the classification probabilities from the SL and DL components. The weights for this fusion are learned from the training data, resulting in a cohesive and robust hybrid classification pipeline (Section 5.3.5).

3. The interpretability of the SL component is rigorously validated through a dual-pronged approach: (i) enhanced permutation feature importance analysis, and (ii) real-world clinical case studies that illustrate both successful and

failed classifications (Section 5.4.7). This approach ensures that the SL arm delivers transparent, medically interpretable justifications for its predictions, thereby aligning with the needs of clinical decision-making and diagnostic support systems.

4. Lastly, we evaluate the classification efficacy of the hypergraph-derived features and the overall performance of the proposed hybrid SL-DL framework on the binary classification task distinguishing Parkinson's Disease (PD) patients from healthy individuals. The hybrid framework achieves an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.979, substantially outperforming the individual SL and DL models, which yield AUC values of 0.878 and 0.852, respectively (Section 5.4.5). This underscores the effectiveness of the hybrid methodology in capturing both performance and interpretability.

The remainder of this chapter is structured as follows: Section 5.2 presents a comprehensive overview of the significance of explainability in medical artificial intelligence, with a particular focus on current state-of-the-art (SOTA) methodologies for Parkinson's Disease (PD) classification. Section 5.3 introduces the proposed explainable hypergraph framework and elaborates on the architectural details of both the Shallow Learning (SL) and Deep Learning (DL) components that constitute the hybrid classification pipeline. Section 6.3 reports the experimental outcomes, including performance evaluations of various baseline models, and benchmarks the proposed hybrid framework against existing SOTA results. Additionally, this section discusses the practical implications and limitations of the proposed methodology. Finally, Section 5.5 summarizes the key findings and contributions of the study, while also outlining promising avenues for future research.

5.2 Related Works

This section is organized into two key subsections. The first subsection highlights the critical importance of explainability in the development and deployment of medical artificial intelligence (AI) systems. The second subsection provides a comprehen-

sive overview of Parkinson's Disease (PD) classification methodologies that leverage either Shallow Learning (SL) or Deep Learning (DL) approaches.

5.2.1 Shallow Learning (SL) Approaches for PD Classification

Prior to the proliferation of deep learning methods, most research on PD classification relied on shallow learning techniques. These studies employed classical algorithms such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Singular Value Decomposition (SVD), often with custom kernel configurations or distance metrics. The features used in these approaches were primarily handcrafted and derived from either the time or frequency domains. Common examples include the Coefficient of Variation (CV), Center of Pressure (CoP), peak force, as well as the kurtosis and skewness of gait cycles [32–35].

In addition to these traditional feature sets, researchers have also explored the use of local pattern transformations for PD classification [36,37]. Ye *et al.* [38] proposed an adaptive neuro-fuzzy inference system designed to capture the nonlinear dynamics of gait, while others have investigated methods for reconstructing VGRF signals to enhance data quality [39]. A thorough comparative analysis of shallow classifiers was conducted by Khoury *et al.*, based on the widely used PhysioPD2007DB dataset [40].

Despite these efforts, SL-based approaches generally exhibit suboptimal performance in terms of classification accuracy and robustness. Furthermore, current state-of-the-art SL methods often overlook the fundamental physical, biological, and statistical principles underpinning the engineered features. This lack of contextual grounding limits their interpretability and clinical applicability.

5.2.2 Explainability in Medical AI

Explainability plays a pivotal role in medical AI, as it fosters transparency, builds clinical trust, and enables effective human oversight. Importantly, achieving explainability does not necessarily require constructing an entirely symbolic or rule-based model. Rather, it is sufficient if the segments of the model that are instrumental in

decision-making can be interpreted in a clinically meaningful manner [98,99]. Tjoa *et al.* observe that the opaque and complex nature of deep learning models—commonly referred to as the "black-box" problem—remains largely unresolved and presents a major barrier to the widespread clinical adoption of DL-based systems [100].

A majority of the explainable artificial intelligence (xAI) techniques currently employed in healthcare focus on two-dimensional (2D) imaging data, such as X-rays, magnetic resonance imaging (MRI), or computed tomography (CT) scans. In their seminal work, Adadi and Berrada [99] catalog various xAI strategies, with feature visualization emerging as one of the most prominent and widely used methods.

However, the adoption of xAI for one-dimensional (1D) time-series data, such as electrocardiograms (ECG), electroencephalograms (EEG), or vertical ground reaction force (VGRF) signals, remains significantly limited. This shortfall is particularly problematic in clinical domains where time-series analysis is crucial. Zvuloni *et al.* [101], for instance, emphasize the lack of interpretability in deep feature representations used for ECG classification, underscoring the urgent need for more transparent approaches.

In the specific context of Parkinson's Disease classification using VGRF time-series data, there is a notable absence of dedicated xAI methodologies. Existing frameworks either lack interpretability or fail to establish clinically grounded reasoning for their predictions. This research aims to address that gap by developing an explainable hybrid learning framework that integrates interpretable SL features with the predictive strength of DL architectures.

5.2.3 Deep Learning (DL) Approaches for PD Classification

In recent years, there has been a significant paradigm shift toward the adoption of deep learning models for PD classification tasks. A diverse array of DL architectures has been proposed, including Multi-Layer Perceptrons (MLPs), 1D convolutional networks, Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks [57–60,97,102]. These models have demonstrated impressive performance metrics, frequently achieving sensitivities, specificities, and overall accura-

cies exceeding 98%. Remarkably, Pham *et al.* [59] reported a perfect classification rate with 100% sensitivity and specificity on their dataset.

Another emerging trend involves the use of pre-trained convolutional architectures such as MobileNet and Xception, originally trained on large-scale image datasets [61]. Although these transfer learning strategies have shown promise, their applicability to 1D biomedical signals remains an area of active research. Additionally, swarm intelligence algorithms—despite proving effective in other medical domains—have not yet been explored for PD classification based on VGRF data [62,63].

A major drawback of these deep learning techniques lies in their lack of interpretability. Their internal decision processes are typically opaque, limiting their adoption in clinical workflows that demand transparency and traceability.

5.2.4 Motivation for a Hybrid Approach

The inherent trade-offs between SL and DL approaches—namely, the interpretability of the former and the predictive strength of the latter—motivate the development of a hybrid learning paradigm. In this work, we propose a framework that builds upon derived SL features, based on the tenets of the physical, biological, and statistical foundations of the existing handcrafted features. These features are then integrated with a DL component to form a hybrid architecture. This approach aims to retain the clinical explainability of the manual features while leveraging the superior classification capabilities of the deep learning models.

5.3 Proposed Method

This section delineates the components and design principles underlying the proposed hybrid framework for Parkinson's Disease (PD) classification. An overview of the complete architecture is illustrated in Fig. 5.3. The method integrates both shallow and deep learning methodologies into a unified classification pipeline to enhance both performance and interpretability.

We begin by revisiting the set of established features commonly used in high-performing state-of-the-art (SOTA) shallow learning (SL) approaches. These fea-

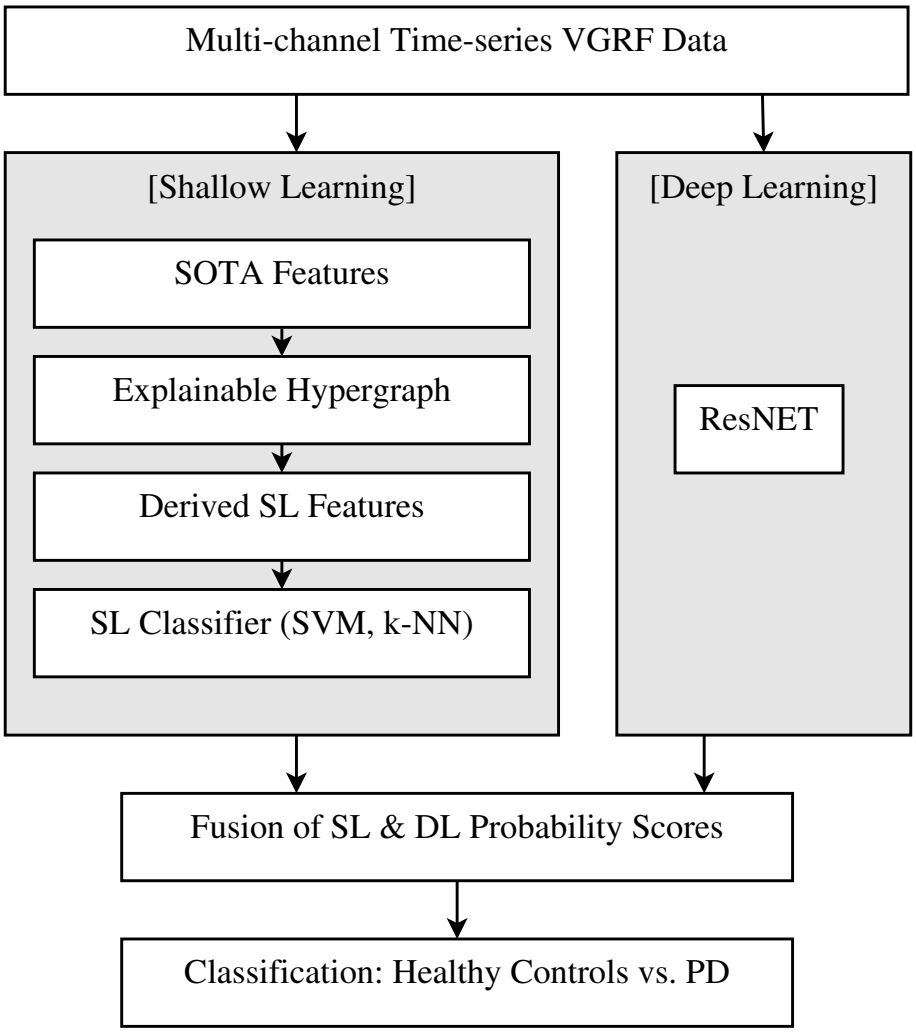


Figure 5.3: Block diagram of the proposed hybrid pipeline consisting of SL and DL arms

tures are then employed as input to construct a novel, interpretable hypergraph model, which enables the derivation of more informative and distinctive representations. These derived representations are subsequently processed using conventional machine learning classifiers, thereby constituting the *SL arm* of the proposed framework.

Parallel to this, a *DL arm* is developed to process the raw multichannel vertical ground reaction force (VGRF) time-series data directly. This deep branch is responsible for learning abstract feature representations from the input data through a deep neural network architecture. The outputs of both the SL and DL arms are finally fused using a decision-level integration strategy, as shown in Fig. 5.3, to yield

the final classification output.

5.3.1 Multimodal SL Features

This subsection provides a concise overview of the key handcrafted features employed in SOTA SL-based PD classification frameworks, particularly as outlined in the work by Alam *et al.* [34]. As listed in Table 5.1, these multimodal features serve as the foundational input to our explainable hypergraph model. Features 1-10 in Table 5.1 are derived from pressure sensors, commonly known as VGRF (Vertical Ground Reaction Force) sensors, features 11-14 are metadata, and feature 15 is a different modality (speed sensor).

1. **Coefficient of Variation (CV) of Swing and Stride Time:** The CV measures the extent of variability relative to the mean swing and stride times (Eq. 5.1). A higher CV reflects increased temporal irregularity in gait, which is more frequently observed in individuals with PD.

$$CV(\%) = \frac{\sigma}{|\mu|} \times 100 \tag{5.1}$$

   where CV is the coefficient of variation, σ is the standard deviation, and μ is the mean.

2. **Center of Pressure (CoP) Features:** In healthy individuals, the CoP typically shifts from the heel to the toe during the stance phase of walking. In contrast, PD patients often exhibit abnormal CoP trajectories due to disruptions in motor control, making these features highly discriminative.

3. **Maximum Peak Force during Heel Strike:** During the initial contact phase of the gait cycle, the peak force exerted by the heel is generally reduced in PD patients compared to healthy controls. This feature captures that biomechanical deficiency.

4. **Kurtosis and Skewness of the Gait Cycle:** Healthy gait is characterized by rhythmic, symmetric cycles. In contrast, PD-related gait is often arrhythmic and asymmetric, which is reflected in deviations in the kurtosis (tailedness) and skewness (asymmetry) of gait cycles.

5. **Metadata Features:** The PhysioPD2007DB dataset [94] provides five demographic and physiological attributes: age (ranging from 36 to 86 years), body weight (47 to 105 kilograms), height (1.48 to 1.85 meters), gender (male/female), and average gait speed (ranging from 0.36 m/s to 1.5 m/s). These are represented as Features 11 through 15 in Table 5.1. Due to the noticeable skewness and missing values in some of these metadata entries, we employ the median as the imputation strategy to ensure robust feature preprocessing.

It is important to note that the objective of this work is not to expand the existing set of shallow features by introducing new ones. Instead, our focus lies in transforming this well-established feature set into a richer and more discriminative feature space using an explainable hypergraph framework. This approach is intended to enhance model performance while preserving interpretability.

5.3.2 Explainable Hypergraph Construction

In this subsection, we provide a formal introduction to the concept of hypergraphs, which serves as the mathematical foundation for constructing the explainable feature relationships used in our proposed classification framework. The hypergraph

Table 5.1: SOTA SL Features used in PD classification using VGRF

Feature Count	**Feature Name**	**Feature Count**	**Feature Name**
1	CV_{swing}	9	$kurtosis_{gait}$
2	CV_{stance}	10	$skew_{gait}$
3	$mean(CoP_x)$	11	age
4	$std(CoP_x)$	12	weight
5	$mean(CoP_y)$	13	height
6	$std(CoP_y)$	14	gender
7	$mean(peak_{force})$	15	average speed
8	$std(peak_{force})$		

formalism enables modeling of complex, higher-order associations among input features, which go beyond pairwise interactions and facilitate interpretability rooted in physical, biological, or statistical relevance [103].

Through this construction, the hypergraph encapsulates the underlying explainable structure among the SOTA SL features, allowing the derivation of new feature representations that preserve interpretability while enhancing discriminative power. The related definitions of hypergraph related mathematical constructs can be found in Section 2.1.

We elaborate on the construction of an interpretable hypergraph structure built upon the fifteen state-of-the-art (SOTA) shallow learning features listed in Table 5.1. The core objective of this hypergraph is to capture explainable relationships among the features—specifically those grounded in shared physical events, biological phenomena, or common statistical derivations. By organizing the features based on these explainable associations, we aim to derive a structured, higher-order representation that preserves both interpretability and analytical richness.

To illustrate this process, consider the four Center of Pressure (CoP) features: mean(CoP_x), std(CoP_x), mean(CoP_y), and std(CoP_y). Although these features differ statistically, they are all computed from the same underlying physical measurement—namely, the 2D time-synchronized Cartesian coordinates of the CoP. As such, they are connected via a hyperedge denoted by e_4 (see Fig. 5.4(a)). This connection exemplifies a shared physical origin.

In a different instance, edge e_1 is formed by connecting two features that both represent the coefficient of variation (CV), applied respectively to swing time and stride time. Despite representing asynchronous events in the gait cycle, these features share a common statistical operation and are therefore logically linked under this category of explainability.

This methodology is systematically extended to other feature combinations as detailed in Table 5.2. For example:

- **Physical origin** connections include e_2, e_3, and e_5, which group various statistical representations of the same fundamental signals: CoP_x, CoP_y, and peak

ground reaction force.

- **Statistical operation**-based groupings are reflected in hyperedges e_6 and e_7, which connect features computed using similar statistical aggregators, such as mean and standard deviation.

- **Biological similarity**-based groupings include hyperedge e_9, which connects all five metadata features (age, height, weight, gender, and gait speed), treating them as biologically related variables.

- **Cross-domain** biological relationships are captured through hyperedges e_{10} to e_{14}, each connecting a specific metadata feature to the full set of non-metadata features. These edges are grounded in explainable biological relevance, except for e_{14}, which explicitly connects gait speed (a measurable physical event) to all non-metadata features.

In total, the explainable hypergraph comprises five binary edges (e_1–e_3, e_5, and e_8) and nine multi-node hyperedges (e_4, e_6, e_7, and e_9–e_{14}), interlinking all fifteen SOTA features. A detailed summary of these connections and their interpretative justifications is provided in Table 5.2.

The weight of the edge e_{ij} connecting vertices v_i and v_j having respective weights n_i and n_j is calculated in the following manner:

$$w(e_{ij}) = e^{(-1)\times|n_i - n_j|} \tag{5.2}$$

The weight of a hyperedge is computed as the sum of all individual edge weights among its constituent vertices, adjusted to avoid double-counting due to symmetry in undirected graphs. Formally, the hyperedge weight is given by:

$$w(h_e) = \frac{1}{2} \sum_{i,j \in v_1,\ldots,v_p \;\; i \neq j} w(e_{ij}) \tag{5.3}$$

The factor of $\frac{1}{2}$ compensates for the bidirectional nature of edge contributions in an undirected hypergraph, as is the case in our framework.

The primary output of this structure is the hypergraph weight matrix, which encodes the strength of relationships among the features. The weights of binary edges are computed according to Equation 5.2, while those of hyperedges follow the formulation in Equation 5.3. These weights reflect both the magnitude of similarity between features and the nature of their interpretability.

Due to the inherent density and complexity of the final hypergraph, a complete visual rendering is impractical. As such, a simplified schematic—highlighting the core topology and relationships—is provided in Fig. 5.4(b) to illustrate the conceptual framework.

To ensure numerical robustness, especially in the presence of outliers or skewed distributions among feature values, we apply a robust rescaling procedure to each weight prior to downstream processing. Specifically, each value is centered around the median and scaled according to the interquartile range (IQR), i.e., the span between the 25^{th} and 75^{th} percentiles [104]. This normalization is independently applied to each element of the weight matrix, thereby improving consistency across features of differing scales and units.

5.3.3 Shallow Learning (SL) Framework

Numerous state-of-the-art (SOTA) studies in Parkinson's Disease (PD) classification have employed a wide variety of shallow learning (SL) techniques, ranging from linear classifiers to non-linear ensemble models. While a comprehensive review of all such SL methods is beyond the scope of this chapter, our objective is not to introduce a novel SL classifier, but rather to enhance the feature representation fed into well-established and empirically validated classification models. By doing so, we seek to assess whether enriched, explainable features derived from our hypergraph framework can improve predictive performance in the context of existing SL algorithms.

To identify the most suitable and historically effective SL methods for this task, we draw upon the extensive comparative analysis presented by Mei *et al.* [105], which surveys the application of machine learning techniques to the PhysioPD2007DB

dataset over the past fifteen years. Their findings identify Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN) as two of the most consistently successful algorithms for PD classification on this dataset. Accordingly, we adopt these two models as the backbone of our SL classification pipeline.

i) **Support Vector Machine (SVM):** SVM is a supervised learning algorithm that constructs optimal separating hyperplanes to distinguish between classes. In this work, we employ both linear and non-linear variants of SVM. Notably, the Radial Basis Function (RBF) kernel variant operates in a non-parametric regime by leveraging pairwise similarity measures (i.e., kernel values) rather than assuming a fixed number of parameters. The RBF kernel computes a Gaussian similarity between all training data points, enabling it to capture complex, non-linear class boundaries.

ii) **k-Nearest Neighbors (k-NN):** The k-NN algorithm is an instance-based, non-parametric classifier that predicts the class of a data point based on the majority class among its k nearest neighbors in the training set. These neighbors are identified using a pre-defined distance metric—typically Euclidean distance. In alignment with standard best practices, we select the value of k according to the square root heuristic, i.e., $k = \sqrt{N}$, where N is the total number of training samples. This choice balances bias and variance in the model's predictions.

By integrating these classifiers into our hybrid pipeline and providing them with derived features from the explainable hypergraph model, we aim to evaluate the additive value of interpretability-focused feature engineering within a shallow learning framework. Experimental results in Section 6.3 will further demonstrate the effectiveness of this strategy.

5.3.4 Deep Learning (DL) Architecture

In parallel with the shallow learning (SL) framework, we construct a deep learning (DL) pipeline to directly extract discriminative representations from the raw Vertical

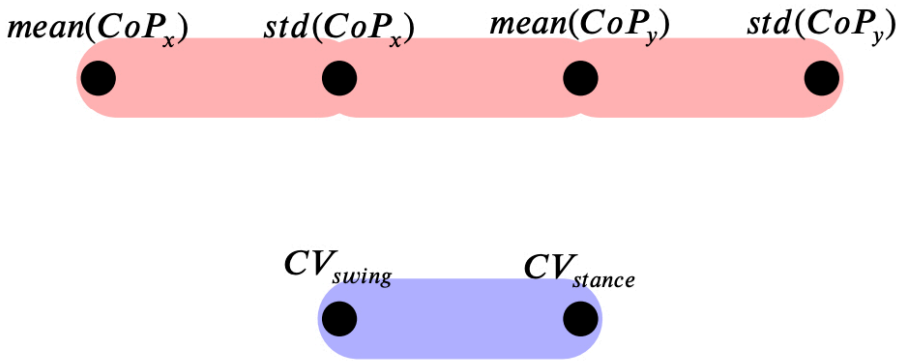
Table 5.2: Explainability in formation of the (hyper)edges

(hyper) edge	Vertices connected by the (hyper)edge	Explanation of Connection
e_1	CV_{swing}, CV_{stance}	Statistical
e_2	mean(CoP_x), std(CoP_x)	Physical
e_3	mean(CoP_y), std(CoP_y)	Physical
e_4	mean(CoP_x), std(CoP_x), mean(CoP_y), std(CoP_y)	Physical
e_5	mean($peak_{force}$), std($peak_{force}$)	Physical
e_6	mean(CoP_x), mean(CoP_y), mean($peak_{force}$)	Statistical
e_7	std(CoP_x), std(CoP_y), std($peak_{force}$)	Statistical
e_8	$kurtosis_{gait}$, $skew_{gait}$	Statistical
e_9	age, weight, height, gender, speed	Biological
e_{10}	age, all non-metadata features	Biological
e_{11}	weight, all non-metadata features	Biological
e_{12}	height, all non-metadata features	Biological
e_{13}	gender, all non-metadata features	Biological
e_{14}	speed, all non-metadata features	Physical

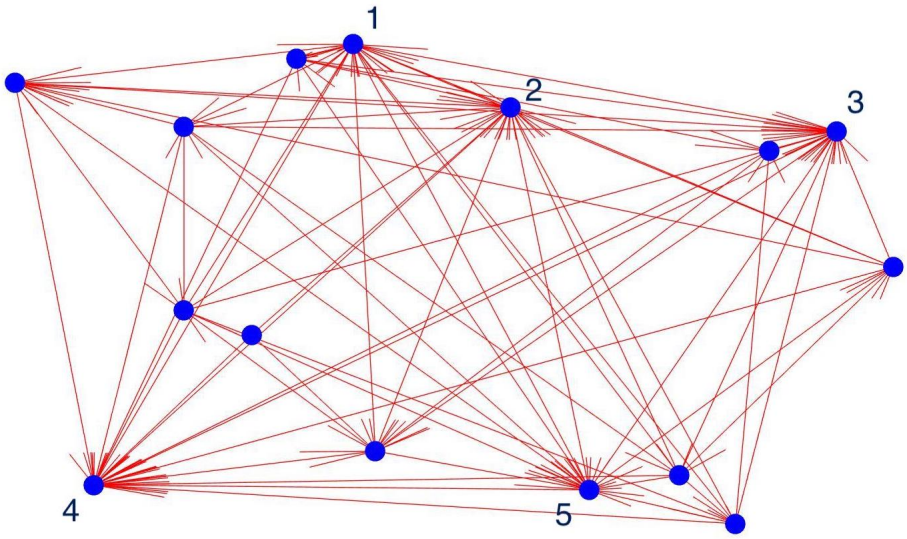
Ground Reaction Force (VGRF) time-series data. To ensure uniformity across all input samples, a standardized preprocessing procedure is applied. Specifically, we fix the input sequence length to 100 seconds, which approximates the median duration of VGRF recordings in the PhysioPD2007DB dataset (ranging from 40 to 122 seconds per subject). For instances where the recording exceeds 100 seconds, the data is truncated. Conversely, recordings shorter than 100 seconds are cyclically repeated to reach the target duration. This ensures consistent input dimensionality across all subjects.

To enhance robustness to outliers and inter-subject variability, we employ a robust rescaling technique analogous to the preprocessing of SL features (see Section 5.3.2). This normalization strategy removes the median and scales values relative to the interquartile range (i.e., between the 25th and 75th percentiles) for each individual input channel.

The backbone of our DL model is inspired by the deep convolutional ResNET architecture proposed by Hannun et al. [46], originally developed for raw ECG signal



(a) Example of (hyper)edges



(b) Explainable hypergraph

Figure 5.4: Construction of Explainable Hyperedge; (a) hyperedge e_4 (red) is connecting the statistical aggregators of CoP and hyperedge e_1 (blue) is an edge connecting the CV of swing and stance; (b) The complete hypergraph with 15 nodes and 14 (hyper)edges; edges and hyperedges are shown as complete and incomplete red edges respectively; Five meta-data nodes, namely age, weight, height, gender, and average speed (marked as 1 - 5) are the most connected vertices of the hypergraph

classification. Hannun's model comprises 33 convolutional layers followed by a fully connected layer and a softmax output. To tailor this architecture to our task and dataset characteristics, the following key modifications are introduced:

i) **Input dimensionality adaptation:** The input to the modified ResNET model is defined as a two-dimensional tensor of shape 18×10000, where 10,000 represents the number of time points corresponding to 100 seconds of VGRF

data sampled at 100 Hz. The 18 channels represent eight VGRF signals from each foot, in addition to one total force channel per foot, consistent with the channel structure of PhysioPD2007DB [94].

ii) **Dropout regularization:** The original ResNET architecture utilized a dropout rate of 0.2, which was appropriate for large-scale datasets comprising over 100,000 samples. However, given the relatively limited size of PhysioPD2007DB (306 subjects), we increase the dropout rate to 0.5 to mitigate the risk of overfitting and improve generalization performance.

iii) **Kernel subsampling adjustments:** Due to the shorter length of VGRF recordings in comparison to Hannun's ECG data, we reduce the subsampling intervals in the convolutional layers to [1, 2, 1, 2]. This enables finer-grained feature extraction from shorter time-series segments without excessive downsampling.

Through these architectural and hyperparameter modifications, we adapt Hannun's ResNET to effectively learn hierarchical representations from multichannel biomechanical gait data for PD classification.

5.3.5 Fusion of SL and DL Probabilities

The hybrid framework culminates in the fusion of prediction scores derived from the shallow and deep learning branches. Each of these arms independently produces a probabilistic estimate for PD classification—denoted as $P_{shallow}$ from the SL pipeline and P_{deep} from the DL pipeline. To integrate these two sources of information, we employ a weighted linear fusion strategy, defined as:

$$P_{\text{final}} = W_{\text{deep}} \cdot P_{\text{deep}} + W_{\text{shallow}} \cdot P_{\text{shallow}} \tag{5.4}$$

Here, W_{deep} and W_{shallow} are scalar weights that control the contribution of the respective branches to the final prediction. As a baseline, we assign equal weights ($W_{\text{deep}} = W_{\text{shallow}} = 0.5$), thereby computing the simple arithmetic mean of the

two probability scores. However, this uniform weighting may not yield optimal classification performance in practice.

To address this, we introduce a data-driven alternative, where the fusion weights are learned from training data via linear regression. This approach enables adaptive weighting based on the predictive strengths of each branch, potentially enhancing overall classification accuracy.

It is important to note that while both SL and DL pipelines are applied to the same underlying time-series data, they operate in fundamentally distinct representational spaces. The SL branch relies on a compact set of hypergraph-derived, interpretable features, whereas the DL branch processes the full high-dimensional multichannel time series. Consequently, the predictive outputs $P_{shallow}$ and P_{deep} are only moderately correlated. In our experiments, the Pearson correlation coefficient between these two probability scores is observed to be approximately 0.34, suggesting a non-trivial degree of independence. This statistical complementarity justifies their fusion and increases the likelihood of performance gains through ensemble learning.

5.4 Experiments and Results

In this section, we present the performance of the proposed hybrid pipeline. We compare the SL and DL baselines separately, followed by the hybrid baselines. Moreover, we compare the performance of the best baseline with SOTA baseline with the state of the art research. Finally, we present the explainability aspect of the presented solution through importance of the derived features and multiple case studies.

5.4.1 Dataset and Experimental Protocol

All experiments presented in this study are conducted using the complete PhysioPD2007DB dataset, comprising a total of 166 subjects—93 diagnosed with Parkinson's disease (PD) and 73 age-matched healthy control subjects. PhysioPD2007DB is a consolidated dataset, constructed by aggregating three independent studies led by different research groups, as originally described in [94]. It is considered a benchmark dataset for PD gait analysis using wearable sensors.

Each subject's metadata includes demographic and physiological parameters such as age, gender, weight, height, and average gait speed. Gait data was collected using wearable pressure-sensitive insoles, each embedded with eight Vertical Ground Reaction Force (VGRF) sensors strategically positioned around the periphery of the foot to capture pressure distribution across various anatomical regions. A portable data acquisition unit was secured at the waist during recording to wirelessly transmit sensor data.

Participants were instructed to walk at their natural pace on level ground for a duration of up to two minutes. This naturalistic walking protocol was chosen to reflect real-world gait patterns, thereby enhancing the ecological validity of the dataset [106]. The VGRF data was recorded at a frequency of 100 Hz, capturing high-resolution time series from 16 sensor channels (eight per foot). In total, 306 distinct walking trials were collected, approximately 30

To evaluate model performance, the dataset was divided into disjoint training and testing sets. Stratified sampling based on class labels was used to preserve the class distribution in both subsets. Additionally, strict subject-level separation was enforced to ensure that no individual appeared in both the training and testing sets. This split yielded 244 walking trials in the training set (80

Performance Evaluation Criteria. While prior studies have predominantly reported classification accuracy as the primary performance metric, we contend that accuracy alone is not a robust measure, particularly in the presence of class imbalance. Accuracy, as well as precision and F-score, when calculated on skewed dataset, become sensitive to the underlying class distribution, leading to biased conclusion. [107].

In contrast, the Receiver Operating Characteristic (ROC) curve offers a class-distribution-invariant evaluation framework. Each point on the ROC curve corresponds to a specific decision threshold and represents a pair of true positive rate (TPR) and false positive rate (FPR) values. Both TPR and FPR are columnar ratios derived from the confusion matrix, rendering the ROC curve agnostic to skewed class proportions. Therefore, the Area Under the ROC Curve (AUC) is adopted

as the primary metric for evaluating classifier performance in this study, providing a comprehensive view of model discrimination across the entire range of possible thresholds.

For completeness and to facilitate direct comparison with prior state-of-the-art (SOTA) methods applied to PhysioPD2007DB, we additionally report traditional evaluation metrics—sensitivity (recall), specificity, and classification accuracy—at an empirically determined optimal operating point on the ROC curve.

Model Output and Thresholding. It is also important to distinguish between the types of classifiers considered. Certain algorithms, such as decision trees or rule-based systems, produce deterministic class labels—yielding a single confusion matrix and a single set of evaluation scores. In contrast, probabilistic classifiers output confidence scores for class membership. These scores can be converted into binary decisions using varying threshold values, enabling the generation of multiple confusion matrices and a corresponding ROC curve [107].

In this chapter, all models are configured to output class probability scores. These scores are evaluated using ROC analysis, with AUC serving as the aggregate performance measure. Sensitivity, specificity, and accuracy are additionally reported at the optimal ROC threshold, allowing a balanced performance comparison with existing literature.

5.4.2 Shallow Learning (SL) Baselines

To establish the reliability and validity of our proposed shallow learning (SL) architecture, we begin by evaluating the statistical separation between the feature representations of Parkinson's disease (PD) and healthy control subjects. We compute the Pearson correlation coefficient between the mean feature matrix of the PD class and that of the healthy control group. This coefficient is found to be -0.44, indicating a moderate negative correlation between the average feature profiles of the two classes—supporting their distinctiveness in the feature space.

To further assess intra-class consistency and inter-class variability, we calculate the Coefficient of Variation (CoV)of the mean feature matrix across three groups:

all subjects combined, PD subjects, and healthy controls. As shown in Table 5.3, the CoV of all subjects is exceptionally high (350), suggesting a wide distribution and high variability in the overall dataset. In contrast, the CoV values within each class are substantially lower, 1.67 for PD subjects and 0.89 for healthy controls. This reduction in intra-class variability demonstrates that the derived features are more homogeneous within each group, which in turn substantiates the robustness of the constructed feature space and validates the reliability of our SL feature engineering pipeline.

The primary derived feature for the SL classification models is the weight matrix of an explainable hypergraph, constructed from 15 state-of-the-art (SOTA) biomechanical and metadata-based features, as previously outlined in Section 5.3.2 and detailed in Table 5.1. This weight matrix is of dimension 15×15, yielding a maximum of 225 possible connections between the features. However, due to the principled and interpretable hypergraph construction, not all features are connected—resulting in a sparse matrix with only 84 non-zero entries. These 84 non-zero weights represent meaningful feature interactions and serve as the final derived features used for SL classification.

As described in Section 5.3.3, we evaluate three SL classifiers based on these features: k-Nearest Neighbors (k-NN), linear Support Vector Machine (SVM), and cubic-kernel SVM. For the k-NN classifier, we set the number of neighbors k=14, which approximates $\sqrt{N}$, where N=244 is the number of training samples. This selection aligns with standard heuristic guidelines for k-NN optimization.

The SL baselines therefore provide a strong, interpretable benchmark grounded in handcrafted, biologically informed feature engineering. The combination of hypergraph-derived features and well-established classification algorithms ensures a fair and principled comparison with deep learning methods discussed in subsequent sections.

Table 5.3: Variability among features, derived from the explainable hypergraph, built using SOTA SL features

Class	Absolute Coefficient of Variation of Mean Feature Matrix
All Subjects	350
PD	1.67
Healthy Control	0.89

5.4.3 Deep Learning (DL) Baselines

To establish strong deep learning (DL) baselines, we implement a modified version of Hannun's ResNet architecture, customized to accommodate the unique characteristics of the PhysioPD2007DB dataset, as previously described in Section 5.3.4. We evaluate multiple training configurations to understand the impact of learning rate schedules and normalization techniques on the model's performance.

Learning Rate Configurations

We experiment with two different training regimes:

- **Fixed Learning Rate:** A range of learning rates is empirically tested, and the optimal fixed learning rate is determined to be 10^{-5}, which offers a reasonable trade-off between training speed and stability.
- **Adaptive Learning Rate (ReduceLROnPlateau):** To further enhance convergence and mitigate overfitting, we adopt an adaptive learning rate schedule. Specifically, the learning rate is reduced by a factor of 10% whenever the validation AUC fails to improve over a window of five consecutive epochs. This widely-used approach, known as *ReduceLROnPlateau* [2], enables the model to adjust its learning pace dynamically based on observed performance.

These configurations yield two DL baselines—one trained with a static learning rate and the other with an adaptive, performance-aware learning rate.

[2] https://keras.io/api/callbacks/reduce_lr_on_plateau/

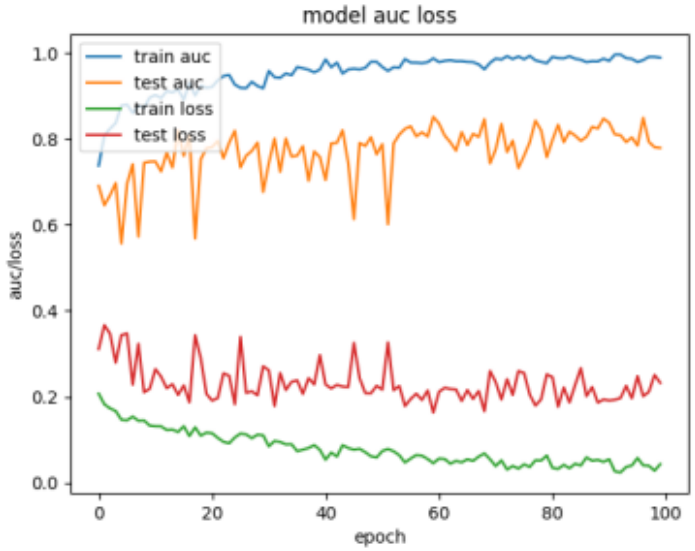


Figure 5.5: AUC loss without Normalization

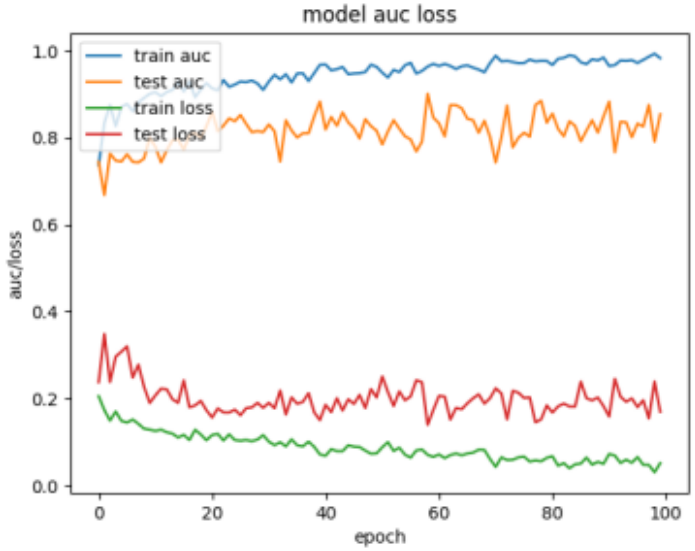


Figure 5.6: AUC loss with Normalization

Impact of Robust Normalization

We also investigate the role of robust feature scaling (see Section 5.3.4) in improving training stability. Training the deep network on unnormalized data leads to high oscillations in both validation AUC and accuracy across epochs (see Fig. 5.5). These fluctuations make it difficult to identify convergence trends or detect overfitting. However, once robust normalization is applied, these oscillations are markedly

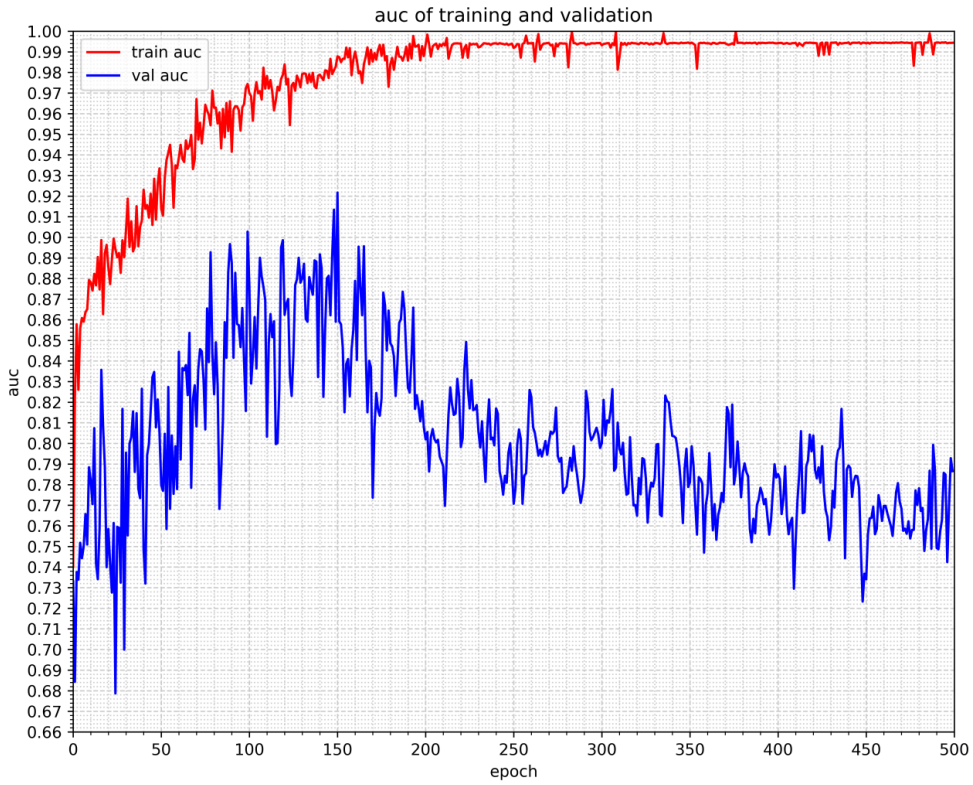


Figure 5.7: Reduce Learning Rate On Plateau for PD Classification

reduced, and training proceeds with greater consistency (see Fig. 5.6). Despite this improvement, the performance curves—particularly the AUC over epochs—still exhibit some noise and lack the expected smooth trajectory typically seen in well-optimized models. This behavior is primarily attributed to the use of a constant learning rate.

Transition to Adaptive Learning Rate

To overcome the limitations of a fixed learning rate, we initially experiment with exponential decay, which gradually reduces the learning rate over time. This results in a modest improvement in validation AUC, from 88.2% to 89.0%. However, the test accuracy stagnates at 94.6%, indicating that the decay schedule may be too aggressive, causing the optimizer to settle into a local minimum and exacerbating overfitting. To address this, we transition to the *ReduceLROnPlateau* policy.

This adaptive strategy reduces the learning rate only when meaningful progress in validation performance has stalled, making it both responsive and efficient.

Training Dynamics and Optimal Epoch

The learning curves under the adaptive learning rate regime provide insights into effective and healthy model training, as illustrated in Fig. 5.7. Key observations include:

(i) From epoch 1 to approximately epoch 150, both validation and test AUC exhibit a steady upward trend, indicating that the model is consistently learning and generalizing well during this phase.

(ii) The performance curves are not flat or oscillatory but instead reflect a clear, positive gradient. This suggests that the model is making measurable improvements in every batch of epochs.

(iii) When smoothed using a moving average (window size of 10 epochs), both the validation and test AUC curves reveal a near-monotonic increase, further confirming the model's stable learning dynamics.

(iv) The highest validation AUC is observed around epoch 150. After this point, a gradual decline is seen, with the validation AUC asymptotically approaching a value near 0.75, while the training AUC continues to increase and surpasses 0.99 by epoch 220.

(v) This divergence between training and validation performance beyond epoch 150 indicates the onset of overfitting. Consequently, we designate epoch 150 as the optimal stopping point for training.

These analyses validate the effectiveness of adaptive learning strategies in stabilizing training, mitigating overfitting, and improving generalization. Together, they serve as a robust foundation for subsequent model comparisons and fusion strategies.

5.4.4 Fusion Baselines

To capitalize on the complementary strengths of the shallow learning (SL) and deep learning (DL) approaches, we implement fusion strategies that combine their respective classification outputs. Specifically, we propose two distinct fusion baselines that integrate the predicted probability scores generated by the best-performing SL and DL models.

Naive Fusion (Equal Weights)

The first baseline employs a straightforward linear fusion strategy, wherein equal importance is assigned to both learning modalities. That is, the final classification probability P_{final} is computed as the average of the individual SL and DL probabilities, $P_{shallow}$ and P_{deep}, respectively:

$$P_{final} = 0.5 \cdot P_{shallow} + 0.5 \cdot P_{deep} \tag{5.5}$$

This naive fusion method assumes equal predictive power from both learning pipelines and serves as a baseline for evaluating the potential benefits of more sophisticated fusion strategies.

Adaptive Fusion (Learned Weights)

Recognizing that SL and DL models may not contribute equally to final prediction performance, we further explore a data-driven approach that learns optimal fusion weights directly from the training set. To this end, we apply linear regression to fit a weighted combination of $P_{shallow}$ and P_{deep} in a manner that maximizes the area under the ROC curve (AUC) on the training data. This regression yields adaptive weights of $W_{shallow} : W_{deep} = 2 : 1$, indicating a stronger predictive influence of the SL classifier in the fused model.

The resulting fusion probability for each test instance is then computed using

these learned weights:

$$P_{final} = \frac{2}{3} \cdot P_{shallow} + \frac{1}{3} \cdot P_{deep} \tag{5.6}$$

This adaptive fusion strategy enables a more flexible and data-sensitive integration of the two learning pipelines, potentially enhancing generalization performance.

In summary, we evaluate two fusion baselines:

- **Equal-weight fusion:** Assumes equal reliability of SL and DL predictions.
- **Adaptive-weight fusion:** Learns optimal weighting from training data to maximize AUC.

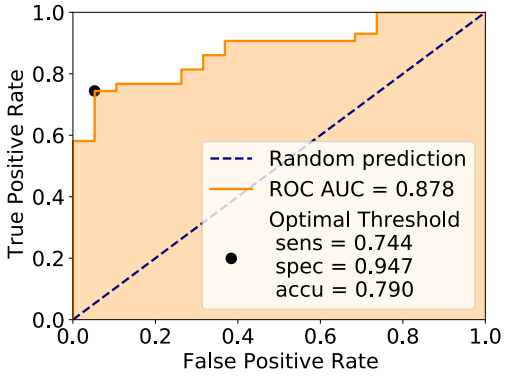
These fusion models serve as critical benchmarks for assessing the added value of combining distinct learning paradigms for Parkinson's Disease classification based on VGRF signals.

5.4.5 Comparison of Baselines

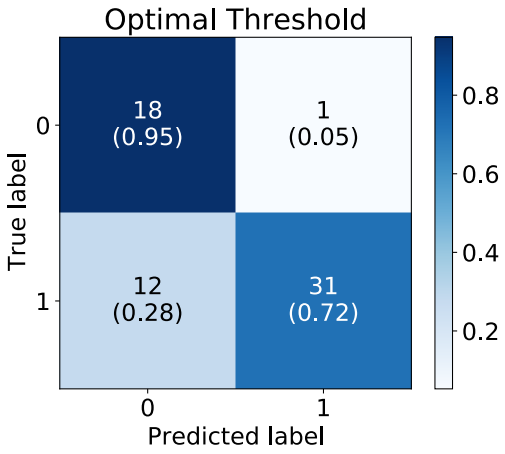
In Sections 5.4.2, 5.4.3, and 5.4.4, we introduced six distinct baselines encompassing shallow learning (SL), deep learning (DL), and fusion strategies. Table 5.4 summarizes the classification performance metrics—including AUC, sensitivity, specificity, and accuracy—of all the baselines.

Among the shallow learning models, the Support Vector Machine (SVM) with a linear kernel demonstrates reasonable discriminative ability, achieving an AUC of 82.5%. However, this model exhibits significant imbalance between sensitivity (69.8%) and specificity (89.5%), implying an asymmetry in class-level predictive performance. When transitioning to the more expressive SVM with a cubic kernel, there is a slight decline in AUC to 81.8%, but the sensitivity-specificity gap narrows, indicating improved stability in binary classification.

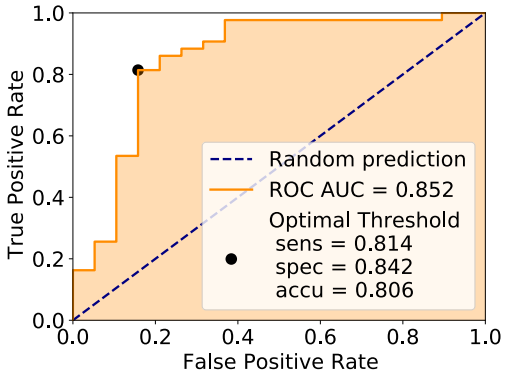
Interestingly, the k-Nearest Neighbors (k-NN) classifier, although simple and parameter-free, shows superior performance among SL models with an AUC of 87.8%. However, it reintroduces imbalance between class sensitivities, pushing specificity to 94.7% at the cost of reducing sensitivity to 74.4%.



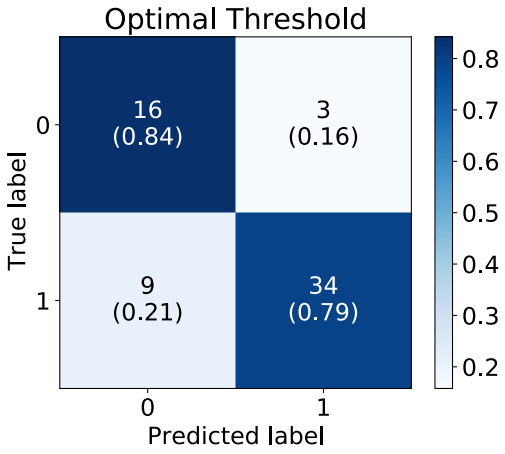
(a) ROC AUC - Shallow Learning



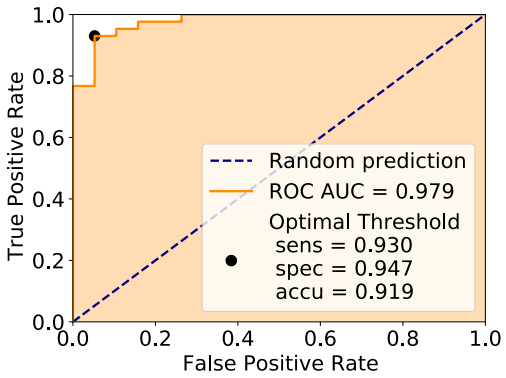
(b) CM - Shallow Learning



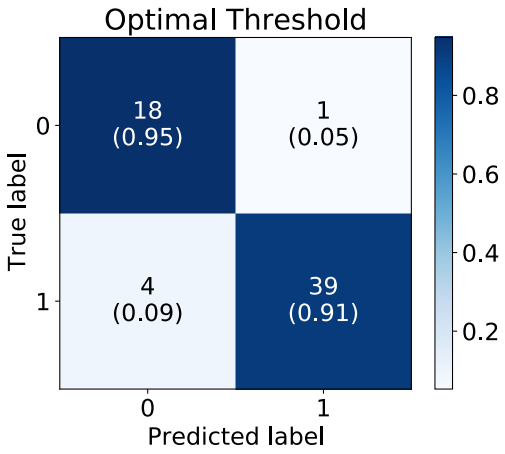
(c) ROC AUC - Deep Learning



(d) CM - Deep Learning



(e) ROC AUC - Hybrid Learning



(f) CM - Hybrid Learning

Figure 5.8: Comparison of the best of SL, DL and Hybrid Baselines (Table 5.4); Optimal thresholds are marked in black circles in 5.8(a), 5.8(c) and 5.8(e); 5.8(b), 5.8(d) and 5.8(f) are the respective Confusion Matrices (CM) corresponding to those optimal points

Moving to deep learning baselines, the initial ResNet-based model with a static learning rate achieves a modest AUC of 81.4%, suggesting room for optimization. Upon introducing an adaptive learning rate strategy—specifically, the `ReduceLROnPlateau` policy—the DL model yields a more balanced classification, with sensitivity and specificity improving to 81.4% and 84.2% respectively, alongside an increased AUC of 85.2%. This outcome underscores the importance of dynamically adjusting the learning rate when model performance plateaus, as discussed in Section 5.4.3.

Nevertheless, despite these improvements, the best-performing DL model does not surpass the highest AUC achieved by the SL counterpart (k-NN, AUC = 87.8%). This highlights the discriminative strength of the handcrafted features derived from the explainable hypergraph formulation (cf. Table 5.2). It is worth noting that the k-NN model operates without any trainable parameters, in contrast to the ResNet architecture, which comprises approximately 141,000 trainable parameters.

Fusion baselines demonstrate the synergistic benefits of combining SL and DL pipelines. The naive fusion approach—where equal weights are assigned to the outputs of SL and DL models—yields a notable increase in performance, achieving an AUC of 95.6%. However, the sensitivity and accuracy values under this configuration (86.0% and 87.1%, respectively) suggest that further gains are possible.

To that end, the adaptive fusion strategy, which employs training data to learn optimal weighting coefficients, significantly enhances model performance. Assigning greater importance to the SL output ($W_{shallow} : W_{deep} = 2 : 1$), this method improves sensitivity to 93.0%, accuracy to 94.7%, and elevates the AUC to 97.9%, thereby establishing itself as the most effective configuration.

To identify the most favorable threshold for each model, we select the optimal operating point on the ROC curve—defined as the point closest to the ideal performance of true positive rate (TPR) = 1 and false positive rate (FPR) = 0. The optimal thresholds are indicated in the ROC plots corresponding to SL (Fig. 5.8(a)), DL (Fig. 5.8(c)), and fusion (Fig. 5.8(e)) methods.

Notably, the hybrid fusion approach demonstrates a more desirable operating

point relative to both SL and DL baselines. This is further supported by the confusion matrices corresponding to the optimal thresholds (Figs. 5.8(b), 5.8(d), and 5.8(f)). The adaptive fusion method correctly classifies 57 out of 62 test instances—surpassing the performance of the best SL and DL methods, which yield 49 and 50 correct classifications, respectively.

In conclusion, the adaptive-weight fusion baseline—leveraging both handcrafted and deep representations—achieves the highest overall performance and constitutes our proposed method. It effectively combines the interpretability and precision of shallow learning with the capacity and flexibility of deep learning, thereby delivering a robust and generalizable model for Parkinson's disease detection from VGRF data.

Table 5.4: Classification Performance of baselines: SL, DL and hybrid methods; performance metrics are in %; the gray highlighted rows present the best performance in respective category

Method	Baselines	Sens	Spec	Acc	AUC
SL	SVM Linear	69.8	89.5	74.2	82.5
	SVM Cubic	79.1	84.2	79.0	81.8
	k-NN	74.4	94.7	79.0	87.8
DL	Fixed Learning	83.2	77.1	79.2	81.4
	Adaptive Learning	81.4	84.2	80.6	85.2
Hybrid (Fusion of SL & DL)	Equal Weights	86.0	94.7	87.1	95.6
	Adaptive Weights	93.0	94.7	91.9	97.9

5.4.6 Comparison with State-of-the-Art Methods

As discussed in Section 5.4.1, the Area Under the Receiver Operating Characteristic Curve (AUC) has been selected as the principal performance metric owing to its insensitivity to class imbalance and its comprehensive measure of a classifier's ability to distinguish between classes. However, as evident from Table 5.5, several existing state-of-the-art (SOTA) approaches applied to this dataset do not report AUC, making direct comparisons challenging.

Additionally, these studies exhibit considerable heterogeneity in terms of validation methodology, employing strategies such as leave-one-out (LOO) cross-validation, 10-fold cross-validation (CV), or fixed train-test splits. This variability

Table 5.5: Comparison of the best Baseline with SOTA methods

SOTA	Method, Year	Sens (%)	Spec (%)	Acc (%)	AUC (%)	Validation Method
[33]	MLP, '10	90.3	77.6	86.4	92.0	10-Fold CV
[34]	SVM Cubic, '17	94.4	96.6	95.7	98.0	LOO
[97]	MLP, '18	100	83.0	91.2	92.0	1 random set
[102]	MLP, 16	88.9	82.2	90.3	-	10-Fold CV
[57]	1D-Convnet, '20	98.1	100	98.7	-	10-Fold CV
[58]	RNN LSTM, '18	-	-	98.6	-	10-Fold CV
[32]	LC-KSVD, '13	-	-	83.4	-	10-Fold CV
[35]	k-NN, '15	-	-	83	-	1 random set
[38]	Neuro-fuzzy, '18	86.7	93.7	90.3	-	LOO
[59]	Tensor, '17	100	100	100	-	LOO, CV
[36]	Local pattern, '21	99	95.7	97.6	-	CV
[60]	1D-CNN, '23	95.2	94.1	95.2	-	10-Fold CV
[37]	S-ELBP, '23	99.0	95.7	97.6	-	-
	Hybrid Pipeline	93.0	94.7	91.9	97.9	1 random set

in experimental design further complicates performance comparisons.

Despite these differences, the proposed method demonstrates competitive and, in some cases, superior performance. For instance, the AUC score achieved by our approach significantly exceeds those reported in [33] and [97], and is on par with that reported in [34]. However, it is important to note that the work by Alam et al. [34] employs the LOO strategy, which allows training on nearly the entire dataset (i.e., $N-1$ samples) while testing on a single observation. In contrast, our approach follows an 80-20 train-validation split, rendering our performance evaluation more conservative and reflective of generalization capability.

Further, works such as [57] and [59] report perfect performance (100%) on certain metrics. While impressive, such results should be interpreted with caution as they may stem from overfitting or overly optimistic validation protocols.

Other DL-based approaches [58, 60] and SL-based frameworks [32, 35–38, 102] have also demonstrated competitive results on this problem. However, a significant limitation of these methods is the lack of attention to model interpretability or explainability.

In summary, although our method may not universally outperform all existing approaches across all metrics, it offers comparable classification performance to the best-reported methods while introducing a novel explainable shallow learning (SL) module based on a handcrafted hypergraph. This additional interpretability constitutes a key advantage of our framework.

5.4.7 Discussion on Explainability

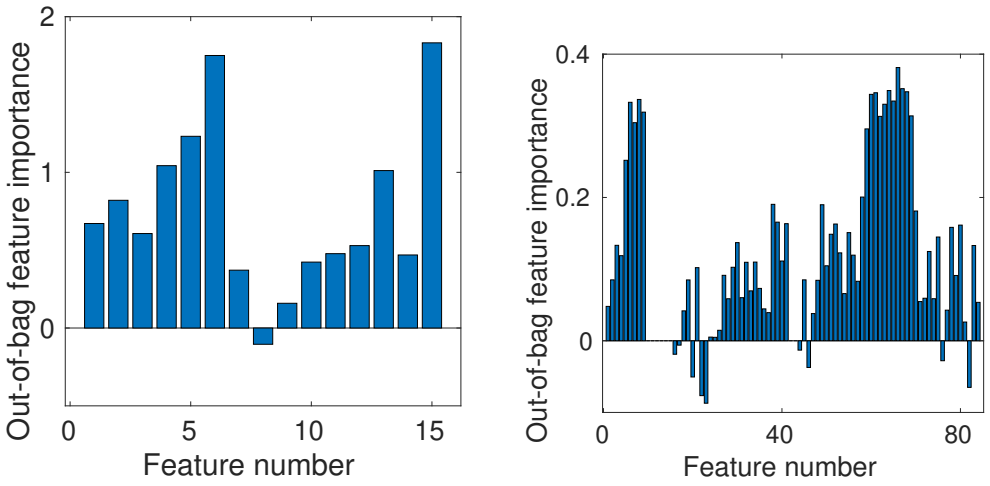
Interpretability is a critical component in the deployment of machine learning models, more so in healthcare domains, where model transparency and accountability are of paramount importance. Explainable Artificial Intelligence (xAI) aims to address these concerns. A variety of approaches, broadly categorized as model-agnostic or model-specific have been developed for this purpose. In this study, we adopt model-agnostic techniques due to their flexibility and applicability across different model architectures [99].

Among model-agnostic methods, post-hoc explainability techniques are most commonly used. These techniques can be further classified into four principal categories: (i) Visualization-based methods, (ii) Knowledge extraction, (iii) Influence-based methods, and (iv) Example-based explanations [99]. For this work, we focus on two such categories: (a) Influence-based methods, specifically feature importance, and (b) Example-based explanations.

Feature Importance

To quantify feature relevance, we employ Permutation Feature Importance (PFI), a widely accepted method introduced by Breiman [108]. PFI measures the decline in model performance when a specific feature's values are randomly permuted. For ensemble models, this process is repeated across all constituent models (e.g., trees in a random forest), and the results are averaged and normalized by the corresponding standard deviation.

A feature set with a higher proportion of features exhibiting substantial PFI values is generally considered more informative. To operationalize this, we define the



(a) SOTA Features (b) Hypergraph Features

Figure 5.9: Feature Importance plots of SOTA features and the features derived from the explainable hypergraph

top quartile of features (in terms of PFI score) as "high-importance" features. As illustrated in Figure 5.9, the baseline feature set—consisting of 15 features—contains only 2 high-importance features (∼13%), as shown in Figure 5.9(a). In contrast, our derived feature set, consisting of 84 features constructed via the proposed hypergraph methodology, comprises 16 high-importance features (∼19%), as shown in Figure 5.9(b).

This improvement in feature relevance substantiates the efficacy of our handcrafted explainable feature engineering pipeline and its ability to extract more discriminative and meaningful representations from the raw data.

Example-Based Explanation

To further illustrate the efficacy and interpretability of our proposed explainable hybrid learning pipeline, we present two contrasting case studies based on individual test subjects. These examples underscore the complementary strengths of the shallow and deep learning components and demonstrate how explainability can be integrated into model predictions.

Case Study 1: Subject ID *SiPt08* Subject *SiPt08* is a 63-year-old male, weighing 100 kilograms and standing 1.75 meters tall, yielding a Body Mass Index (BMI) of 32.7, which falls in the obese category. This subject has been clinically annotated as having Parkinson's Disease (PD), with a Hoehn and Yahr (H&Y) scale rating of 2, indicating an early stage of the disease.

The deep learning (DL) component of our model, which processes raw vertical ground reaction force (VGRF) signals, fails to assign a sufficiently high PD probability to this subject, resulting in a misclassification. A plausible explanation for this outcome is that early-stage PD may not manifest prominently in spatio-temporal gait patterns, which are typically exploited by deep networks for diagnosis.

In contrast, the shallow learning (SL) component, which leverages a handcrafted hypergraph-based feature set, correctly classifies the subject as PD. Specifically, the hypergraph incorporates informative hyperedges that connect anthropometric attributes (e.g., height and weight) with conventional state-of-the-art (SOTA) gait features. These combinations effectively capture the subject's obesity-related profile, as indicated by features such as those listed in rows 9, 11, and 12 of Table 5.2. This enriched feature representation enables the SL arm to produce a high PD probability.

The success of the SL module in this case is not only empirically validated but also aligned with findings from existing literature, which suggest a strong correlation between obesity and increased risk of neurodegenerative disorders, including PD [109]. Thus, this case exemplifies how the SL module provides an interpretable rationale for overriding the incorrect decision of the DL module, thereby showcasing the strength of the hybrid approach.

Case Study 2: Subject ID *GaPt08* Subject *GaPt08* is a 68-year-old female, weighing 57 kilograms and measuring 1.63 meters in height, resulting in a BMI of 21.5—within the healthy range. Like *SiPt08*, she is also diagnosed with early-stage PD (H&Y scale rating of 2).

In this case, the SL module fails to yield a sufficiently high PD probability. This is attributed to the lack of strong discriminative signals within the SOTA features and anthropometric data. The subject's healthy BMI does not present a risk indicator

similar to the previous case, and thus the hypergraph-based feature set lacks the necessary discriminatory power for accurate classification.

However, the DL module succeeds in classifying *GaPt08* correctly as a PD patient. This result suggests that, despite the absence of explicit feature-level indicators, the deep network is capable of extracting subtle, complex spatio-temporal patterns from the raw VGRF data that are indicative of PD. These latent features may not be immediately interpretable but contribute meaningfully to the final decision.

This example highlights the complementary nature of the two modules: while the SL arm prioritizes interpretability and domain-informed feature representations, the DL arm excels in uncovering abstract data-driven patterns. The fusion of both allows the hybrid pipeline to combine interpretability and predictive strength, leading to robust and explainable decision-making.

5.4.8 Discussion

The proposed hybrid learning framework offers a significant advantage, particularly in clinical and biomedical contexts, by combining predictive performance with partial interpretability. One of the key strengths of our approach lies in its ability to retain the high classification accuracy typically associated with deep learning (DL) models, while simultaneously enhancing interpretability through the shallow learning (SL) component that utilizes an explainable hypergraph-based feature set.

This dual capability is especially beneficial in clinical decision-making environments, where, domain experts often prefer interpretable outputs to aid diagnosis and treatment planning. The SL arm, grounded in handcrafted features and hypergraph theory, provides transparency and traceability in the reasoning process behind classification. As a result, medical professionals can better understand and trust the model's recommendations, especially when supported by medically relevant attributes such as age, body mass index, or gait parameters.

Despite these strengths, the current hybrid framework has notable limitations. Foremost among them is the partial nature of its explainability. While the SL compo-

nent is explicitly designed to be interpretable, the DL module remains a "black box." This limits the full transparency of the overall decision pipeline, as the contribution of the DL arm cannot be directly scrutinized or explained using domain-specific knowledge or interpretable features.

Another important limitation lies in the adaptive fusion strategy used to combine outputs from the SL and DL arms. The adaptive weights are learned from the available training data, making the fusion process inherently data-dependent. While this improves performance on the specific dataset used in our study, it may compromise the model's generalizability across different populations or clinical settings. A more robust approach would involve learning these fusion weights from larger and more diverse datasets, or alternatively, deducing them from established clinical guidelines or domain knowledge, thereby anchoring the model more firmly in medical reasoning.

In summary, the proposed hybrid model provides a promising step toward explainable and accurate automated diagnosis. However, future work is required to enhance the interpretability of the DL component and to establish a more generalized and clinically grounded fusion mechanism.

5.5 Conclusion

In this study, we proposed a novel hybrid classification framework for the detection of Parkinson's Disease (PD) using vertical ground reaction force (VGRF) signals. The proposed pipeline integrates both shallow learning (SL) and deep learning (DL) paradigms to achieve a balanced trade-off between interpretability and predictive performance.

The SL component is designed around an explainable hypergraph-based model, where state-of-the-art (SOTA) gait features are connected based on their physical, biological, and statistical relevance. This structured representation enhances transparency by allowing medical practitioners to trace the rationale behind classification decisions. In parallel, the DL arm is constructed using a customized deep Residual Network (ResNet) architecture, which is trained to autonomously learn

spatio-temporal features from raw VGRF signals. The outputs of the SL and DL arms are subsequently combined using adaptively learned fusion weights to produce a unified classification decision.

Extensive experimental evaluations confirm the efficacy of the hybrid approach, outperforming standalone SL and DL baselines in terms of AUC, sensitivity, and accuracy. Furthermore, we bolster the claim of explainability by employing Permutation Feature Importance (PFI) to identify the most influential features and providing detailed case-based examples that illustrate how the hybrid model handles contrasting clinical scenarios. These qualitative and quantitative insights demonstrate the practical utility of our framework in real-world diagnostic contexts.

Looking forward, we aim to enhance the system's interactivity and clinical applicability by incorporating a human-AI collaboration layer, enabling healthcare professionals to engage with the model through intuitive interfaces and receive medically grounded explanations, in line with the vision of explainable and trustworthy AI systems [98].

Chapter 6

Multimodal Explainable Sepsis Prediction

> *Prediction is very difficult, especially about the future.*
>
> Niels Bohr, Critical Care Medicine, August 2011, Volume 39, Issue 8

We present an explainable artificial intelligence (xAI) framework for sepsis prediction, addressing the challenges posed by its dynamic nature and high rate of missing clinical data. Our approach enhances XGBoost with a novel focal weighted binary cross-entropy loss, incorporating three tunable parameters—α, β, and γ—to balance prediction probabilities, penalize misclassifications, and emphasize hard examples, respectively. We derive the gradient and Hessian of the proposed loss to ensure its integration into XGBoost's optimization. Experimental evaluations on a public dataset demonstrate superior performance over state-of-the-art baselines, while SHAP-based interpretability reveals the clinical relevance of diverse medical features captured by our method.

6.1 Introduction

Sepsis is a critical medical condition arising from a dysregulated host immune response to infection, often leading to life-threatening organ dysfunction. It continues

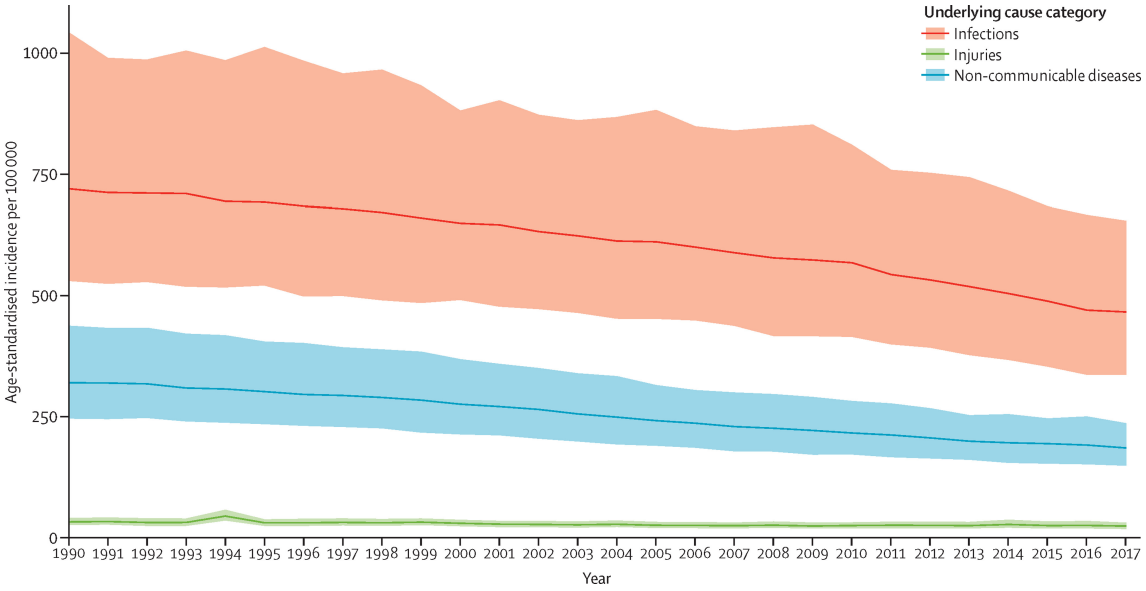


Figure 6.1: Sepsis affects more people compared to injuries and non-communicable diseases [2]

to impose a significant burden on global healthcare systems and remains one of the foremost causes of morbidity and mortality worldwide. The etiological sources of sepsis are varied, including but not limited to infections in the respiratory tract, urinary system, gastrointestinal organs, or dermal lesions [11, 12]. According to the Centers for Disease Control and Prevention (CDC), sepsis is implicated in nearly one-third of all hospital-related deaths, highlighting its devastating clinical impact. Globally, it is estimated that approximately 31.5 million individuals develop sepsis annually, with 19.4 million progressing to severe stages, resulting in an estimated 5.3 million deaths per year [12]. These staggering statistics underscore the complexity and urgency of early detection and effective clinical management of sepsis. As depicted in Fig. 6.1, Sepsis has higher prevalence than the diseases we already covered in this thesis [2]. Sepsis itself is not contagious and cannot be spread from person to person. However, the infections that can lead to sepsis, such as bacterial, viral, or fungal infections, can be contagious. For example, someone with a contagious respiratory infection like COVID-19 or the flu could potentially spread that infection to others, and if that infection leads to sepsis, the original infection is what was spread, not the sepsis itself.

The clinical presentation of sepsis often evolves rapidly, necessitating prompt

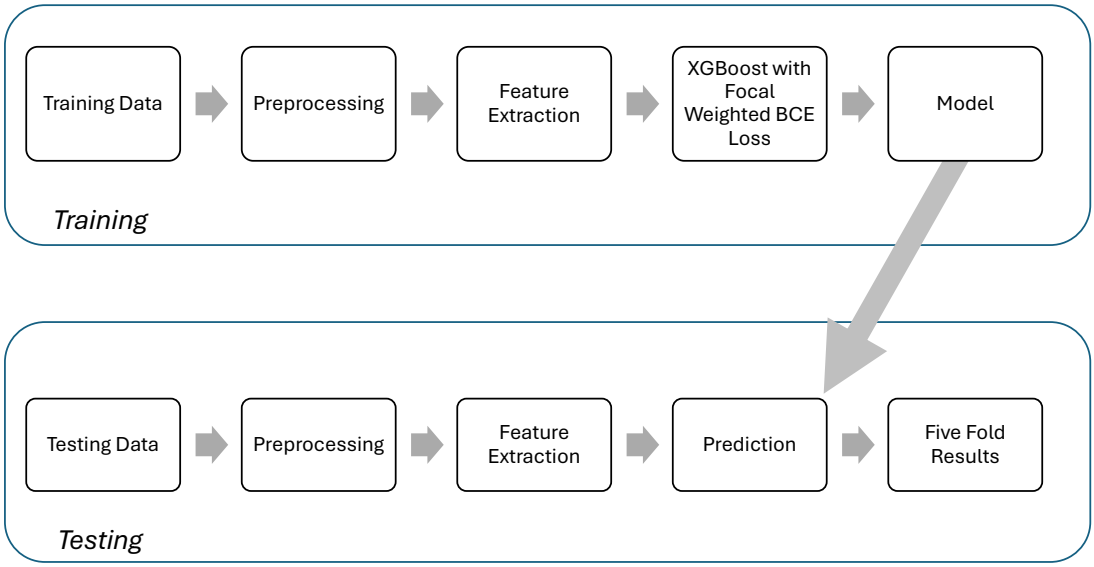


Figure 6.2: System Diagram

recognition and intervention. Typical symptoms may include fever, altered mental status, tachycardia, and respiratory distress—symptoms that can escalate into septic shock if left untreated. While the adoption of bundled care strategies and advancements in critical care medicine have led to improvements in patient outcomes, the global burden of sepsis remains alarmingly high. In recent years, technological innovations such as artificial intelligence (AI), electronic health records (EHRs), and automated early warning systems have emerged as promising tools for enhancing the early detection and management of sepsis. These technologies, along with improvements in public health education and healthcare access, hold potential for reducing the incidence and fatality associated with sepsis. Nevertheless, the inherent pathophysiological complexity of the condition and disparities in healthcare infrastructure continue to present considerable challenges [13].

In this context, the PhysioNet Computing in Cardiology Challenge 2019 (PNC19), and its associated dataset (PNC19DB), provided a standardized platform to evaluate early sepsis prediction models [41]. A notable characteristic of this dataset is the prevalence of missing data, which realistically reflects clinical scenarios in hospital settings. Many state-of-the-art (SOTA) methods employed various

imputation techniques to handle missingness. Additionally, the concept of "informative missingness"—which treats missing data patterns as informative signals rather than mere gaps—was leveraged to improve prediction accuracy. For instance, Hammoud et al. [42] demonstrated that the duration of missing values for a given feature can serve as a valuable predictor. One leading submission in the challenge achieved an AUROC of 0.8406 and a utility score of 0.404 through stratified 5-fold cross-validation [43].

Feature engineering played a significant role in several submissions. Manually derived features such as the Shock Index (heart rate divided by systolic blood pressure), Bilirubin-to-Creatinine ratio, and partial Sequential Organ Failure Assessment (SOFA) scores were integrated to capture domain-specific insights [45]. Gradient boosting models, particularly XGBoost, emerged as popular choices due to their interpretability and performance [42–44]. For example, Singh et al. [43] trained an ensemble of five XGBoost models using a data-splitting strategy where 10% of the data was allocated for feature selection and hyperparameter tuning, and the remaining 90% was split into five disjoint subsets. Each subset trained a separate model, and final predictions were obtained via geometric mean aggregation, resulting in an AUROC of 0.844 and a utility score of 0.4281.

Innovative modeling strategies also emerged, such as the Time-Phased Model, which stratified the ICU length of stay (ICULOS) into early (0–9 hours), middle (10–49 hours), and late (50+ hours) phases, employing LightGBM for early and middle stages, and a recurrent neural network (RNN) for the late stage. This strategy achieved a utility score of 0.4149 using 10-fold cross-validation. Another noteworthy approach utilized a signature-based regression model [44], which extracted path signatures—collections of iterated integrals from multivariate time series—to capture temporal dynamics in the data. This approach, when combined with engineered features, achieved a utility score of 0.43 and ranked first among 78 participating teams in the challenge.

Subsequent research compared classical machine learning models (e.g., Logistic Regression, Random Forest, Gradient Boosting) and deep learning architectures

(e.g., LSTM, CNN) for time-series analysis [110]. A recent study [111] trained an XGBoost model on 80% of the dataset, using 107 features (both original and derived), and evaluated its performance on the remaining 20% and on an entirely unseen prospective dataset. Using the normalized utility score prescribed by the PhysioNet Sepsis Challenge, the model achieved a score of 0.494. However, due to the dynamic and heterogeneous nature of sepsis, such clinical criteria may not always be consistently present, thereby reducing model reliability [112]. Moreover, the presence of co-morbid conditions in critically ill ICU patients often leads to confounding symptoms that complicate sepsis diagnosis [113].

This study utilizes the clinical data provided by the PNC19 dataset to develop a novel explainable artificial intelligence (xAI) framework for early sepsis detection. The two principal contributions of this chapter are as follows:

1. **Novel Loss Function Design:** We introduce a focal weighted binary cross-entropy loss function that incorporates three tunable parameters: α (to modulate the influence of prediction probabilities), β (to differentially penalize false positives and false negatives), and γ (to emphasize harder-to-classify examples, inspired by focal loss). This formulation allows the model to achieve improved utility scores with the same feature set.

2. **Explainability through SHAP Analysis:** We perform a comprehensive SHAP-based feature attribution analysis to elucidate model decisions. The results reveal that the most influential features span across multiple medical domains—ranging from respiratory and renal indicators to metabolic and hematological parameters—thereby aligning with the multi-systemic clinical nature of sepsis. This holistic perspective is absent in other competing models and underscores the practical interpretability of our approach.

The remainder of the chapter is organized as follows: Section 6.2 elaborates on the proposed focal weighted binary cross-entropy loss, as well as data pre-processing, feature engineering, and model training. Section 6.3 compares our model's performance with existing SOTA methods using the PNC19 utility metric and explores

explainability aspects. Finally, Section 6.4 presents concluding remarks and outlines directions for future research.

6.2 Methods

This section outlines the end-to-end pipeline of the proposed methodology for early sepsis detection. As illustrated in Figure 6.2, the data processing flow is divided into training and testing phases. The central innovation of our approach is the focal weighted loss function, which is incorporated within the XGBoost model for classification. Below, we describe the key steps of the pipeline, focusing on pre-processing, feature engineering, and the details of the loss function integration.

6.2.1 Multimodal Feature Engineering

In this study, the dataset contains clinical variables from multiple modalities, such as vital sign variables, laboratory variables, and demographic variables. Altogether, the features contain forty different modalities.

The dataset used in this study, PNC19DB, contains a significant proportion of missing data, particularly within the lab test and vital sign features. Specifically, the lab test features have an average missingness of 32.4%, while the vital sign features exhibit an even higher missingness rate of 94.9%. In contrast, patient demographic features are complete with no missing data [114]. To address the missing data, we apply multiple strategies depending on the type of feature.

For the variables in PNC19DB that exhibit long-tailed distributions, a logarithmic transformation is performed to reduce the skewness of the data. Subsequently, z-score normalization is applied to standardize the data, ensuring that all features have a mean of zero and a standard deviation of one. Missing values in the dataset are handled using zero-forward-filling imputation, where each missing value is replaced by the last known value at the preceding time step [115].

To further enhance the feature set and account for missingness patterns, we augment the data with a binary mask vector of length 40. Each element in the mask vector indicates whether the corresponding feature value is present or missing.

Specifically, the mask vector is set to 1 if the original feature is present, and to 0 if the feature is missing and subsequently imputed. This provides the model with additional information on the reliability of each feature value.

Furthermore, we introduce a delta vector of length 40, which represents the differences between consecutive feature values at discrete time steps. This delta vector is used to capture temporal dynamics within the data. The final feature vector is constructed by appending the delta vector to the original feature vector, forming a vector of length 120. This vector is then augmented by including the feature vectors from the preceding 1 to 6 hours, thereby incorporating temporal dependencies in the model's input. If there is insufficient data for a given hour (i.e., fewer than H hours' worth of data), we pad the feature vector with zeros to align the length of the input.

The resulting input to the classifier is a matrix of size $L \times H = 120 \times 6$, where L represents the length of the feature vector (120) and H represents the number of previous hours (6). This augmented feature matrix is then fed into the classifier to make predictions on sepsis detection.

6.2.2 Focal weighted BCE loss

Let us start with the equation of binary cross-entropy (BCE) loss, which we intend to expand going forward. BCE is defined as follows.

$$\begin{aligned} &L(y,p) = \\ &-\frac{1}{T}\sum_{t=1}^{T}\Big((1-p_t)y_t \times log p_t \\ &+ p_t \times (1-y_t) log(1-p_t)\Big) \end{aligned} \tag{6.1}$$

where T is the number of time steps recorded for a patient, y_t is the sepsis label output at time step t (i.e., $y_t = 0$ or $y_t = 1$), and p_t is the classifier's output at time step t (i.e., $0 \leq p_t \leq 1$).

In the PNC19DB dataset, there is a significant class imbalance, with 2,932 subjects diagnosed with sepsis and 37,404 ICU patients who do not exhibit any signs

of sepsis [114]. This results in a heavily skewed distribution, where the minority class (sepsis) constitutes only 7.2% of the data, while the majority class (non-sepsis) makes up a dominant 92.8%. It is important to note that even among the subjects who are diagnosed with sepsis, there are extended periods during their ICU stay where they are not classified as having sepsis. In contrast, subjects in the non-sepsis group consistently remain without sepsis annotations throughout their entire ICU stay.

This class imbalance issue becomes even more pronounced when considering the hourly data labeling, as each hour of every subject's ICU stay must be marked as either sepsis or non-sepsis. As a result, the hourly class imbalance is even more severe, exacerbating the challenges in training a robust model for sepsis detection.

To address these issues and ensure that the model effectively learns from both the minority and majority classes, we introduce several modulating factors in the loss function. Specifically, we incorporate a factor β into the binary cross-entropy loss function, as shown in Equation 6.2. The factor β serves to adjust the penalty applied to false positive and false negative predictions. When $\beta = 1$, equal penalties are applied to false positives and false negatives, ensuring a balanced treatment of both types of errors. However, if a higher penalty is desired for false positives, β is set to a value greater than 1. Conversely, if the objective is to assign greater importance to false negatives, β is assigned a value less than 1. This flexibility allows the model to better handle the class imbalance and optimize performance based on the specific requirements of the sepsis detection task.

$$
\begin{aligned}
&L_1(y, p) = \\
&-\frac{1}{T}\sum_{t=1}^{T}\Big((1-p_t)y_t \times \log p_t \\
&+ p_t \times \beta(1-y_t)\log(1-p_t)\Big)
\end{aligned}
\tag{6.2}
$$

In the next stage, we consider the introduction of the focal loss function. The focal loss, as originally proposed by [116], is expressed in the following form:

$$
FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{6.3}
$$

where p_t represents the probability of the t-th instance, $\gamma \geq 0$ is a tunable focusing parameter, and $\alpha \in [0, 1]$ is a weighting factor that can be either based on class frequency or optimized as a hyperparameter. The primary advantage of the focal loss is its ability to assign more weight to hard-to-classify examples during the training process, ensuring that they contribute more significantly to the model's learning. Notably, when $\gamma = 0$, the focal loss reduces to the traditional binary cross-entropy loss, with no emphasis on difficult examples.

Building upon the original definition of the focal loss in Equation 6.3, we incorporate it into the weighted binary cross-entropy loss function described in Equation 6.2. This results in the formulation of the Focal Weighted Binary Cross-Entropy (FWBCE) loss, which is defined as follows:

$$\begin{aligned} &L_2(y, p) = \\ &-\frac{1}{T}\sum_{t=1}^{T}\Big(\alpha(1-p_t)^{\gamma} y_t \times \log p_t \\ &+ (1-\alpha)p_t^{\gamma} \times \beta(1-y_t)\log(1-p_t)\Big) \end{aligned} \tag{6.4}$$

To optimize the FWBCE loss function, we compute the gradient and Hessian of $L_2(y, p)$ by taking the first-order and second-order partial derivatives with respect to time t. These derivations are crucial for updating the model parameters during training. The final expressions for the gradient and Hessian are provided below, and the corresponding derivations are outlined in the Appendix section of this chapter.

$$\begin{aligned} &\mathcal{G}\mathit{rad}\Big(L_2(y, p)\Big) = \\ &-\phi\Big(-\gamma p \log(p)(1-p)^{\gamma} + (1-p)^{\gamma+1}\Big) \\ &-\psi\Big(\gamma p(1-p)\log(1-p) - p^{\gamma+1}\Big) \end{aligned} \tag{6.5}$$

$$
\begin{aligned}
&\mathcal{H}ess\Big(L_2(y,p)\Big) = \\
&-\phi\Big(-\gamma^2 p^2(1-p)\log(p) - \gamma p(1-p)^{\gamma+1}\log(p) \\
&-\gamma p(1-p)^{\gamma+1} - (\gamma+1)p(1-p)^{\gamma+1}\Big) \\
&-\psi\Big(\gamma p(1-p)^2\log(1-p) - \gamma p^2(1-p) \\
&-\gamma p^2(1-p)\log(1-p) - (\gamma+1)p^{\gamma+1}(1-p)\Big)
\end{aligned}
\tag{6.6}
$$

where $\phi = \alpha y$ and $\psi = (1-\alpha)\beta(1-y)$.

The parameters α, β, and γ play critical roles in mitigating the effects of class imbalance in the training dataset. Each of these parameters approaches the problem in a distinct manner. The parameter α provides linear control over the loss contribution from both p and $(1-p)$. The parameter β, on the other hand, serves as an additional linear coefficient to penalize false positives and false negatives. Meanwhile, γ is introduced in the gradient and Hessian expressions as a non-linear function, emphasizing harder-to-classify examples during training.

When training the model, the parameters α and β should be optimized first, as they directly influence the weighting of the loss function. Once these parameters have been adjusted, γ can be tuned to control the focus on difficult examples, further improving the model's ability to handle imbalanced classes and optimize performance in challenging scenarios.

6.2.3 Classifier

The classification model employed in this study is based on gradient boosting, specifically utilizing the XGBoost library [115, 117]. This approach constructs up to 100 decision trees as part of the boosting process. The model is evaluated using a stratified 5-fold cross-validation strategy, ensuring that each fold maintains a balanced representation of both the positive and the negative classes. During training, the model is subjected to 40 epochs, with a fixed learning rate of 0.2 to ensure optimal convergence. To further ensure robust performance evaluation, the train and test

Medical Specialties	Parameters captured in PNC19DB
Cardiovascular	HR, SBP, MAP, DBP, Troponin I
Respiratory	O_2Sat, Resp, $EtCO_2$, FiO_2, SaO_2, $PaCO_2$
Coagulation	Hgb, PTT, Fibrionogen
Blood corpuscle	WBC, Platelets, Hct
Liver	Bilirubin total, Bilirubin direct, BUN, AST
Renal	Creatinine
Minerals	Magnesium, Phosphate, Potassium, Calcium, Chloride
Biochemical Parameters	Lactate, Glucose, pH BaseExcess, HCO3, Alkalinephos
Misc.	Temp
Demography	Age, Gender, Unit 1, Unit 2, HospAdmTIme, ICULOS

Table 6.1: Parameters captured in PNC19 dataset, categorized into separate medical specialties

datasets in each fold are designed such that there is no overlap of subjects, effectively preventing data leakage between training and testing phases. This methodological setup allows for an unbiased and reliable assessment of the model's performance.

6.3 Experiments and Results

In this section, we discuss the dataset, experimental protocol, and results.

6.3.1 Dataset

The dataset comprises over 40,000 time series records collected from two hospital systems, with each patient represented in an individual .psv file. Each file contains 40 columns spanning vital signs, laboratory values, and demographics, with the 41st column indicating the target variable, sepsisLabel. The clinical variables include 40 features: 8 vital signs,, 26 laboratory values and 6 demographic features.Missing and erroneous data were retained to reflect real-world ICU conditions. For patients who developed sepsis, the label was '0' up to 6 hours before sepsis and '1' otherwise.

For all experiments of this chapter, the complete public dataset of PNC19, i.e, PNC19DB, is used. The dataset consists of two different datasets collected from two different hospitals, named as set A and set B. Sets A and B have 20,336 and 20,000 unique patients, respectively. For each subject, the dataset has a placeholder

for 40 clinical variables: eight vital sign variables, 26 laboratory variables, and six demographic variables (Table 6.1). However, as one can expect, a large percentage of laboratory variables were not collected, leading to a huge percentage of missing values in them. Here, we present a different view of these variables categorized into different groups of medical specialties such as cardiovascular, respiratory, coagulation, blood corpuscle, liver function, minerals, other biochemical parameters, demography, and miscellaneous [41]. The reason for creating these categories is to show that sepsis is clinically derived from an amalgamation of multiple medical specialties [11]. Later in this chapter, we intend to demonstrate how these categories can play a role in building an xAI sepsis prediction.

6.3.2 Protocol

We use stratified 5-fold cross-validation with zero subject overlap between train and test. For each fold, the training set contained data from four-fifth of the patients, resulting in approximately 1 million time slices, each representing a data window. The resulting model was evaluated on the remaining one-fifth of the patients.

Performance is measured using a utility score defined by the Physionet 2019 challenge organizers as follows [41]. For patients who eventually had sepsis, according to at least one SepsisLabel entry with value 1, the function rewards classifiers that correctly predict sepsis within 12 hours before the earliest and 3 hours after the latest onset timestamp of sepsis, t-sepsis. The optimal incentive for accurate predictions within the time frame is set as a parameter, valued at 1.0. In contrast, penalties are incurred for predictions made after this time frame. Predictions made more than 12 hours before t-sepsis are considered very early detections and incur a penalty, with the penalty upper limit being (-0.05). Similarly, after the 3-hour threshold post-sepsis, late predictions are also penalized with a maximum penalty of (-2.0).

6.3.3 Hyperparameter tuning

As discussed in section 6.2.2, we tuned the α, β and γ variables as hyperparameters on the PNC19DB. Heuristically, using a five-fold validation, α is set to 0.5. α proved

SOTA	Method (year)	Utility	
		Set A	Set B
[118]	Signature based regression (2020)	0.430	
[115]	XGBoost (2019)	0.416	0.376
[119]	XGBoost (2019)	0.4281	
[69]	XGBoost (2019)	0.4149	
[43]	XGBoost (2019)	0.404	
[120]	XGBoost (2019)	0.411	
[121]	Weighted squared error and squared log error (2022)	0.47 (1 random test)	
[122]	Random Forest (2022)	0.427 (1 random test)	
[114]	Reinforcement learning (2023)	0.415	0.450
[110]	XGBoost (2024)	0.472 (3-fold test)	
Ours	XGBoost with Focal Weighted Loss	0.4305	0.393

Table 6.2: Comparison of focal weighted loss based solution ($\alpha = 0.5$, $\beta = 0.025, \& \ \gamma = 0.5$) with SOTA methods on Set A and B

to be too sensitive for Sepsis detection. Conversely, values of β was not drawn from experimentation. The utility function of PNC19, counts false negatives as -2 and false positives as -0.05. β is set to 0.025 as the utility function of PNC19 imposes a penalty (0.05:2=1:40) [41]. Then, we tune the focal loss γ (Eqn. 6.5 and Eqn. 6.6). As expected, the performance advantage of γ eventually gets nullified once γ gets close to 1. We select $\gamma = 0.5$ heuristically, after running 5-fold cross-validation for the entire dataset.

6.3.4 Comparison with SOTA methods

Here we compare our results with the state of the art results on sepsis prediction.

6.3.5 Explainability Analysis

We restrict the SOTA comparison within traditional machine learning methods, i.e., non-deep learning methods. We do not consider SOTA deep learning methods employed for sepsis prediction, as they fail to provide explainability. All SOTA methods do not report individual results on datasets A and B. Moreover, the validation methodology is not uniform either. Some report on the hidden test dataset of PNC19, or 3/5/10 on sets A and B. Only a few of them analyze the aspect of explainability. We compare our 5-fold result with the as-is reported cross-validation

results of ten SOTA methods in Table 6.2. Our utility score is at par with the best results. The performance of the proposed methods is the best among the methods that use 0.5 as the cut-off of probabilities for prediction. Two methods that beat us optimized the cutoff threshold (as opposed to the standard cutoff of 0.5) on the regressed values to maximize the utility score on the training set ([110, 121]).

The best XGBoost method stood 2^{nd} in the official PNC19 [115]. They used only the β parameter. After α and γ parameters are introduced in the proposed focal weighted loss, we see a 3.4% and 4.5% improvement in sets A and B, respectively, using the same set of features they used [115]. Signature-based regression using lightGBM got the 1^{st} rank in official PNC19 owing to the novel features [118]. However, they do not offer explainability as a part of their solution. On the other hand, the unofficial 1^{st} rank submission in PNC19 provides a good utility score of explainability with SHAP [123]. However, the top-most impactful features they presented do not cover all medical specialties in Table 6.1. For example, no parameter from the "Coagulation" is listed in their top 20 features. Conversely, the proposed solution covers all specialties in Table 6.1.

There are several ways to achieve explainability in an xAI solution [99]. In this chapter, we select *Influence methods*. We further choose Feature Importance as our choice of Influence method. We use the SHAP tool to showcase the impact of individual features.

In Fig. 6.3, we list the top twenty features that impacted the final prediction. Interestingly, "ICULOS hour 6" (ICU length of stay at $t = T - 6$ hours) has the highest impact in sepsis, i.e., class 1 prediction. Conversely, the "HospAdmTime hour 1" (Time between hospital and ICU admission at $t = T - 1$ hours) has the highest impact in non-sepsis, *i.e.*, class 0 prediction. This is intuitively explainable as once sepsis is clinically determined, the subject will most likely be shifted to the ICU, if not done already. The second most impactful features for sepsis and non-sepsis are "Temp hour 1" and "Unit1 hour 1" (Administrative identifier for medical ICU)", respectively. This is also perfectly explainable as fever, a severe response to an infection, is a common symptom of sepsis. On the other hand, subjects in the

medical ICU, as opposed to the surgical ICU, are less likely to have sepsis. Similarly, we can medically explain other features of this list.

However, the more important part of the explainability of the proposed solution in sepsis prediction is as follows. Sepsis patients can have different symptoms at different times and may appear differently in different people. This is the reason we introduced Table 6.1 to present different medical faculties in a grouped format. It is unlikely that only one faculty of medical science will have an impact on sepsis pre-

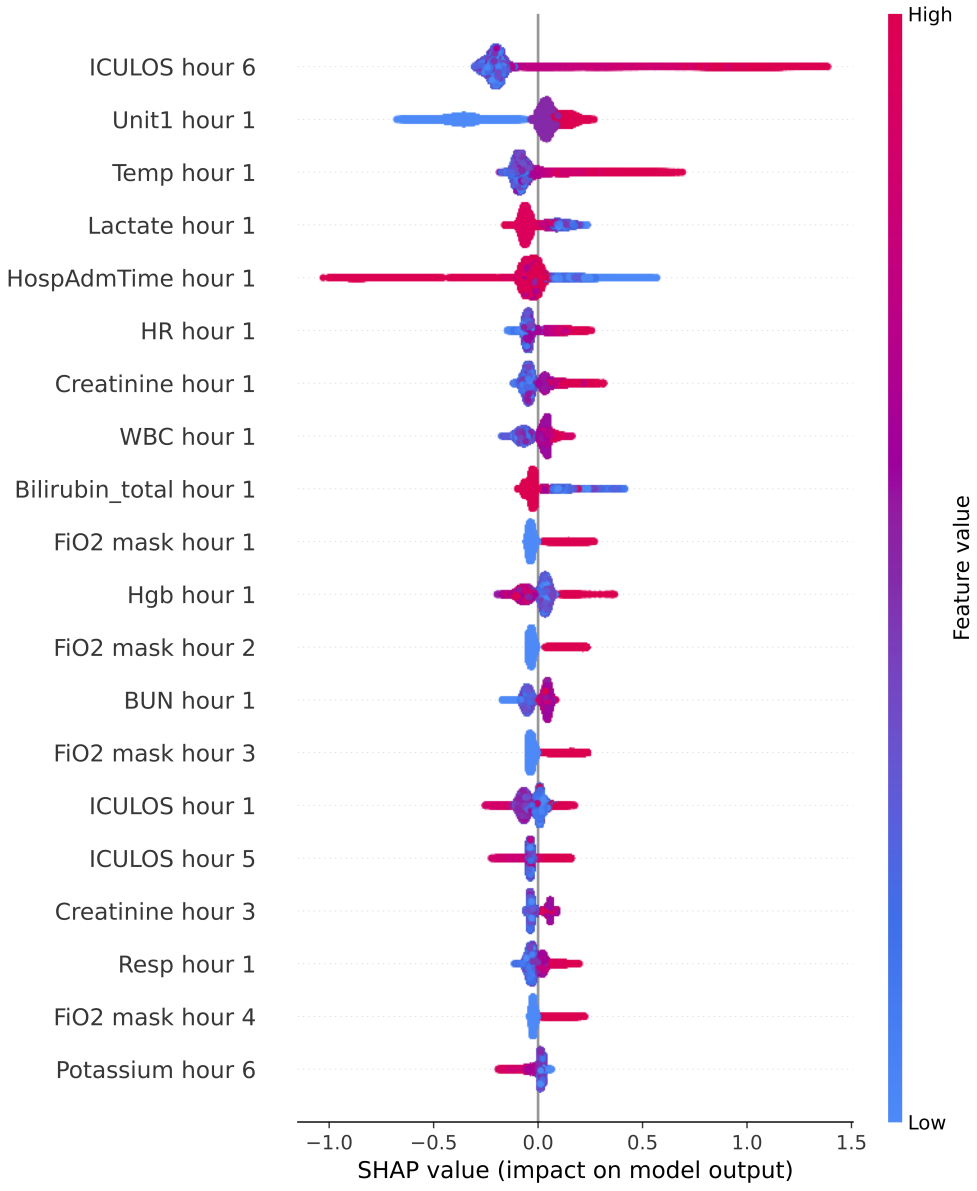


Figure 6.3: Impact on prediction —a single dot is an explanation on each feature row

diction. If we compare Table 6.1 with Fig. 6.3, we can see that the proposed pipeline has taken features from each of these faculties into consideration. This proves that

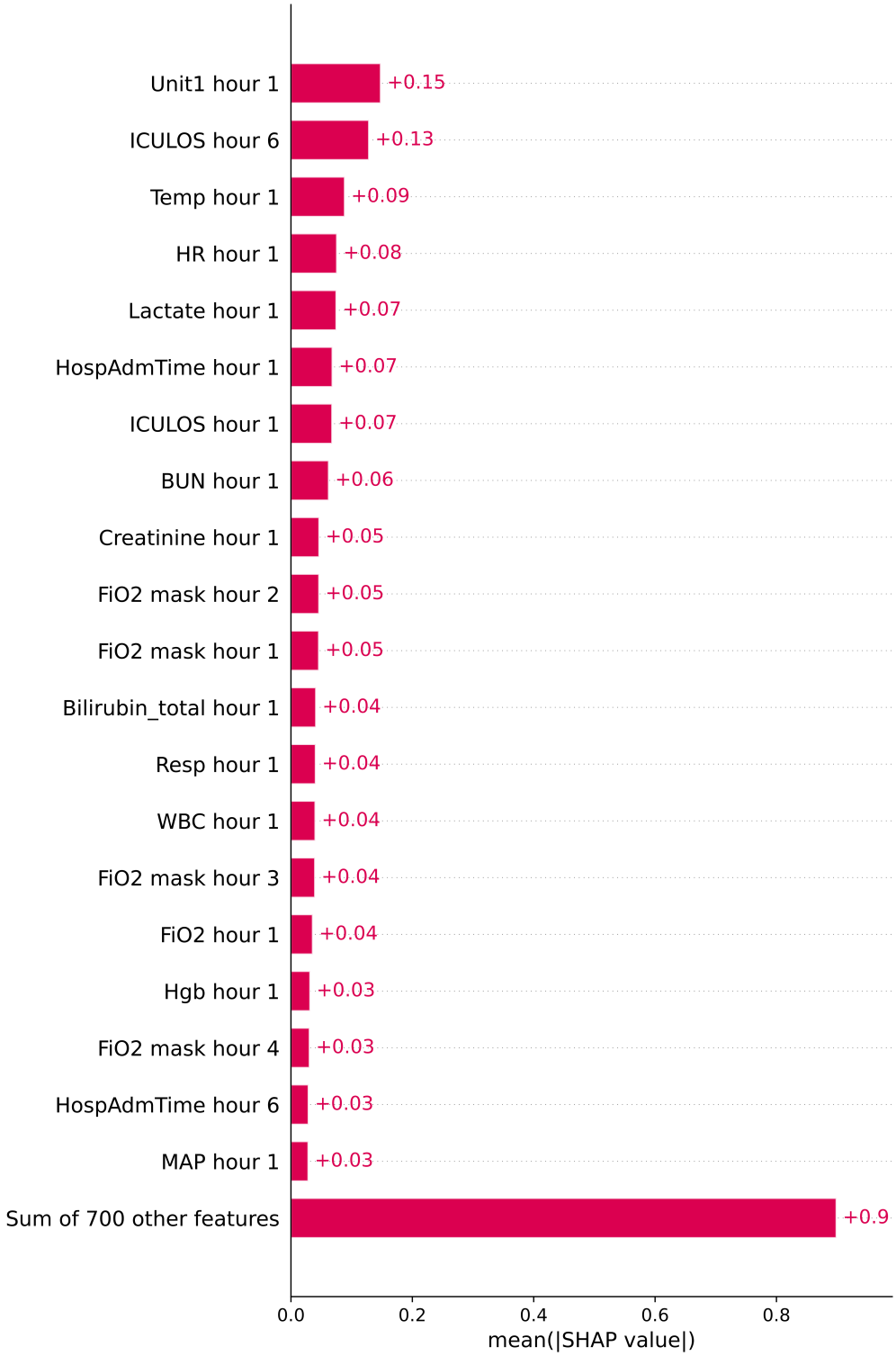


Figure 6.4: Global feature importance —Mean SHAP values of top twenty features

the proposed pipeline achieves a holistic explainability in sepsis prediction. The only SOTA method that discussed explainability failed to achieve this [123].

Furthermore, we dive into the mean absolute SHAP values of individual features, generated during prediction (Fig. 6.4). It is to be noted that the top 20 features' contribution exceeds the rest 700 features' contribution. Furthermore, we observe only two six-hour-old features in the top 20 list, namely "ICULOS hour 6" and "HospAdmTime hour 6". This is understandable as per the aforementioned discussion. However, one particular feature, namely FiO_2 stands out as the missingness of this feature for the past few hours (hours 1, 2, 3, and 4) got listed in the top 20 features (Fig. 6.4). The fact that FiO_2 is missing means that the subject is not currently under scrutiny and hence, not likely to be a sepsis patient.

6.4 Conclusion

In this chapter, we addressed the problem of binary sepsis prediction with an xAI solution. The two major highlights of our solution are a focal weighted binary cross-entropy loss and a SHAP-based explainability. The proposed method enhances sepsis prediction by improving performance and interpretability. Through integration of tunable hyperparameters, we achieve a better utility score than the state-of-the-art XGBoost-based approaches. Experimental results validate the efficacy of our solution in capturing diverse medical indicators, reinforcing its potential for real-time clinical deployment and further optimization. On the other hand, SHAP analysis confirms that our model considers a comprehensive range of medical specialties, ensuring a holistic and interpretable prediction framework. Since these features are inherently interconnected, a possible future work is to apply graph neural network to derive more disctintive classification feature engineering.

6.5 Appendix

6.5.1 Grad and Hess Derivation for Focal Wighted BCE Loss

In this appendix, we derive the expressions for the gradient and the hessian of the focal weighted binary cross-entropy loss. We start with the differentiation of Eqn. 6.4.

$$
\begin{aligned}
&\mathcal{G}\mathit{rad}(L_2) \\
&= (-1) \times \Big(\phi_t(\gamma(1-p_t)^{\gamma-1} \times (-1) \times log(p_t)) \\
&+ (1-p_t)^{\gamma} \times \frac{1}{p_t}) \\
&+ \psi_t(\gamma \times p_t^{\gamma-1} \times log(1-p_t) \\
&+ p_t^{\gamma} \times \frac{1}{1-p_t} \times (-1)\Big) \times p_t(1-p_t)) \\
&= -\phi_t\Big(-\gamma(1-p_t)^{\gamma} \times p_t \times log(p_t) \\
&+ (1-p_t)^{\gamma+1}\Big) \\
&- \psi_t\Big(\gamma \times p_t \times log(1-p_t) \times (1-p_t) \\
&- p_t^{\gamma+1}\Big)
\end{aligned}
\tag{6.7}
$$

where ϕ and ψ are defined in Eqn. 6.8.

$$
\begin{aligned}
\phi &= \alpha \times y \\
\psi &= (1-\alpha) \times \beta \times (1-y)
\end{aligned}
\tag{6.8}
$$

Please note that the Eqn. 6.7 presents the gradient for the t^{th} term. The generic representation is given in Eq. 6.5. We need to differentiate Eqn. 6.7 once more to get the hessian.

$$
\begin{aligned}
&\mathcal{H}ess(L_2) \\
&= (-1) \times \Bigg(\phi_t \Big(-\gamma^2 \times (1-p_t)^{\gamma-1} \times (-1) \\
&- \gamma \times (1-p_t)^{\gamma} \times log(p_t) - \gamma \times (1-p_t)^{\gamma} \\
&+ (\gamma+1) \times (1-p_t)^{\gamma} \times (-1) \Big) \\
&+ \psi_t \Big(\gamma \times (1-p_t) \times log(1-p_t) \\
&+ \gamma \times p_t(-1) + \gamma \times p_t \times log(1-p_t)(-1) \\
&- (\gamma+1) \times p_t^{\gamma} \Big) \Bigg) \times p_t \times (1-p_t) \\
&= -\phi_t \Big(\gamma^2 \times p_t^2 \times (1-p_t) \times log(p_t) \\
&- \gamma \times p_t \times (1-p_t)^{\gamma+1} \times log(p_t) \\
&- \gamma \times p_t \times (1-p_t)^{\gamma+1} \\
&- (\gamma+1) \times p_t \times (1-p_t)^{\gamma+1} \Big) \\
&- \psi_t \Big(\gamma \times p_t \times (1-p_t)^2 \times log(1-p_t) \\
&- \gamma p_t^2 \times (1-p_t) \\
&- \gamma \times p_t^2 \times (1-p_t) \times log(1-p_t) \\
&- (\gamma+1) \times p_t^{\gamma+1} \times (1-p_t) \Big)
\end{aligned}
\tag{6.9}
$$

Similarly, Eq. 6.9 represents the hessian for the t^{th} term. The general form of the equation is given in Eqn. 6.6.

6.5.2 Grad and Hess without α and γ

Here, we calculate the form of the Grad and Hess expression, if decide to we remove the effect of α and γ. This reduces the the equations of the previous subsections of the state of the art weighted BCE function.

Let us also derive the simpler version of Eqn. 6.7 and Eqn. 6.9. If we assign $\alpha = \frac{1}{2}$, ϕ_t and ψ_t gets modified as shown in Eqn. 6.10.

$$\begin{aligned}\phi_t &= \frac{y_t}{2} \\ \psi_t &= \frac{\beta}{2} \times (1 - y_t)\end{aligned} \tag{6.10}$$

Furthermore, if we assign $\gamma = 0$, $\mathcal{G}\mathcal{r}\mathcal{a}\mathcal{d}$ changes to the following expression in Eqn. 6.11.

$$\begin{aligned}&\mathcal{G}\mathcal{r}\mathcal{a}\mathcal{d}(L_2) \\ &= \frac{y_t}{2} \times (1 - p_t) + \frac{\beta}{2} \times (1 - p_t) \times (-p_t) \\ &= \frac{y_t}{2} - \frac{p_t \times y_t}{2} - \frac{\beta \times p_t}{2} + \frac{\beta \times y_t \times p_t}{2} \\ &= -\frac{p_t}{2} \times (\beta - \beta \times y_t + y_t) + y_t\end{aligned} \tag{6.11}$$

The generic form of Eqn. 6.11 is presented in Eqn. 6.12

$$\begin{aligned}&\mathcal{G}\mathcal{r}\mathcal{a}\mathcal{d}(L_2) \\ &= -\frac{p}{2} \times (\beta - \beta y + y) + y\end{aligned} \tag{6.12}$$

Similarly, if we assign $\alpha = \frac{1}{2}$ and $\gamma = 0$ in Eqn. 6.9, $\mathcal{H}\mathcal{e}\mathcal{s}\mathcal{s}$ changes to the following expression in Eqn. 6.13.

$$\begin{aligned}&\mathcal{H}\mathcal{e}\mathcal{s}\mathcal{s}(L_2) \\ &= \frac{y_t}{2} \times p_t(1 - p_t) + \frac{\beta}{2}(1 - y_t)p_t(1 - p_t) \\ &= \frac{p_t}{2} \times (\beta - \beta y_t + y_t)(1 - p_t)\end{aligned} \tag{6.13}$$

The generic form of Eqn. 6.13 is presented in Eqn. 6.14

$$\begin{aligned}&\mathcal{H}\mathcal{e}\mathcal{s}\mathcal{s}(L_2) \\ &= \frac{p}{2} \times (\beta - \beta y + y)(1 - p)\end{aligned} \tag{6.14}$$

Chapter 7

Conclusion

> *We shall not cease from exploration*
> *And the end of all our exploring*
> *Will be to arrive where we started*
> *And know the place for the first time*

T.S. Eliot, Little Gidding, September 1942

This thesis has addressed several contemporary challenges in the diagnosis of multiple diseases through the integration of graph signal processing techniques and deep learning frameworks. Across the preceding chapters, we have presented comprehensive descriptions of our proposed methodologies, supported by rigorous experimental results—both qualitative and quantitative—that substantiate their effectiveness. In this concluding chapter, we summarize the major contributions of the thesis and outline promising directions for future research.

7.1 Concluding Remarks

This thesis is structured into seven chapters, comprising an introduction (Chapter 1), a theoretical foundations chapter (Chapter 2), and the current concluding chapter. The four intermediate chapters (Chapters 3 through 6) are each devoted to addressing specific research questions, each centered on novel approaches to medical diagnosis using graph-based learning and hybrid architectures.

We initiated our investigation with a focus on coronary artery disease (CAD),

where the primary objective was to differentiate between diseased and control populations. To that end, we proposed a graph-based approach wherein a graphical structure was superimposed on traditional feature spaces to preserve more nuanced interrelationships among features. Specifically, edges in the graph were derived using a combination of biological, physical, and statistical associations—connections that are typically neglected in conventional machine learning pipelines. Experimental evaluation clearly demonstrated that the graph-derived features enhanced discriminative capacity and preserved richer information compared to traditional methods.

Subsequently, we extended our approach to handle multimodal physiological signals. In this context, we introduced a hypergraph-based representation to capture complex interactions among features extracted from photoplethysmography (PPG), phonocardiography (PCG), and electrocardiography (ECG) signals. The use of hyperedges enabled modeling of mutual interdependencies across modalities, resulting in a novel hypergraph structure capable of expressing relationships hitherto unexplored in the literature. The empirical results confirmed the superior representational power and classification performance of our proposed hypergraph framework.

Building upon these foundations, we developed a hybrid architecture combining shallow learning techniques with deep learning methodologies. This was motivated by the increasing emphasis on interpretability in AI-driven medical diagnostics. We highlighted the critical need for a human-in-the-loop paradigm in medical explainable AI (xAI) and introduced an explainable hypergraph framework to bridge this gap. Parallel to this, we also explored the efficacy of deep learning models, and ultimately demonstrated that a judicious combination of shallow and deep learning components could yield a high-performing yet interpretable solution—a major step toward trustworthy AI in clinical settings.

In the penultimate part of the thesis, we turned our attention to optimizing loss functions in the context of sepsis prediction. Here, we proposed a novel loss function termed focal weighted binary cross-entropy, specifically designed to improve explainability without sacrificing predictive accuracy. This formulation offered enhanced interpretability from the perspective of clinical significance, aligning better with

real-world diagnostic priorities. Furthermore, we demonstrated that our approach improved feature attribution when evaluated using the SHAP (SHapley Additive exPlanations) method, thereby reinforcing the utility of our loss function from both a performance and interpretability standpoint.

7.2 Future Directions

While the contributions made in this thesis offer substantive progress toward the diagnosis and classification of complex medical conditions using graph-based learning, several promising avenues remain open for future research.

A logical progression of our work would involve leveraging recent advances in *Graph Neural Networks* (GNNs). GNNs inherently utilize the connectivity information in graph structures, similar to our approaches, but with the added advantage of end-to-end learning. GNNs can be employed for a variety of problem formulations including node classification, graph classification, and link prediction. Their primary strength lies in the ability to learn rich and meaningful embeddings by incorporating both node-level features and topological information. These embeddings can capture subtle dependencies and relationships that conventional methods may overlook, potentially enhancing both model performance and explainability.

Another significant extension would involve the application of our methodologies to higher-dimensional medical data. While the present research has focused on 1D time-series signals—such as ECG, PPG, PCG, and pressure sensor data—future investigations could incorporate 2D and 3D imaging data, such as X-rays, MRI, and CT scans, which are commonly used in cancer diagnostics and other complex disease assessments. We posit that the graph- and hypergraph-based techniques described in this thesis are scalable and adaptable to such higher-dimensional modalities. In doing so, they could facilitate the construction of more intricate graphical models that not only improve performance but also enhance interpretability, which remains a crucial unmet need in multi-dimensional medical data analysis.

Moreover, the rapid emergence of *Large Language Models* (LLMs) in recent years has opened up a new frontier in explainable AI. LLMs—and their special-

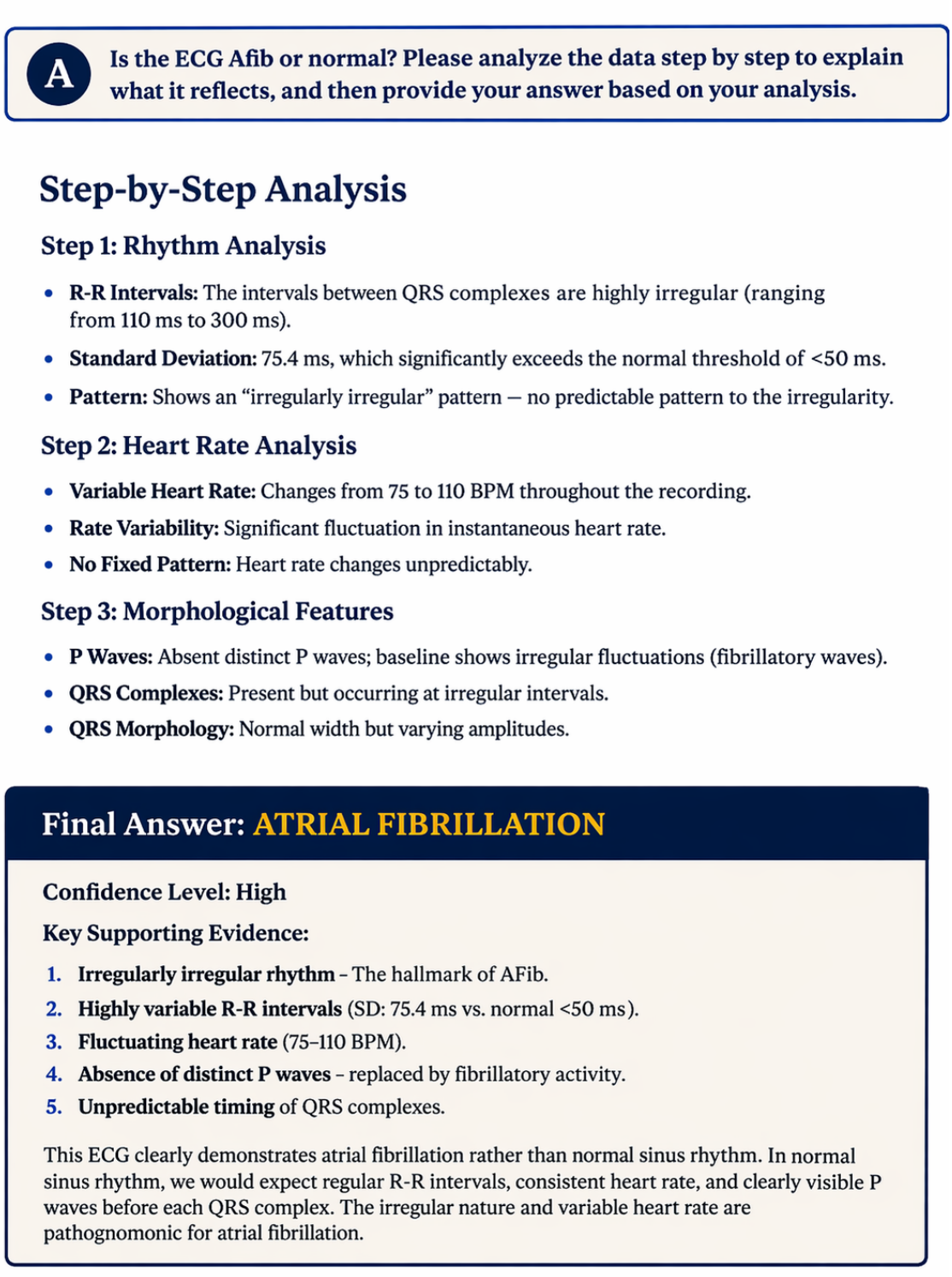


Figure 7.1: Prompt and response on 1-D ECG data when uploaded to Claude.ai 4.0

ized variants such as Vision-Language Models (VLMs) and Signal-Language Models (SLMs)—offer inherently interpretable outputs, often presented in natural lan-

guage. As of 2025, emerging research demonstrates that LLMs are increasingly being adapted to handle not only textual data but also time-series signals such as ECG. In a recent effort to validate this capability, we examined the work of Tuo An et al., who reported that Claude[1] exhibited exceptional performance in interpreting ECG signals [124]. As shown in Fig. 7.1, upon uploading a sample ECG from a patient with atrial fibrillation (AF), the model returned a detailed and medically relevant explanation, underscoring the potential of LLMs in real-time, interpretable diagnostics.

These developments highlight an exciting research opportunity: integrating LLMs into graph-based or hybrid diagnostic pipelines to enhance interpretability, generalizability, and user trust. Given their natural language generation capabilities and increasing modality versatility, LLMs represent a compelling next step for advancing explainable AI in healthcare.

By pursuing these avenues, future research can further improve diagnostic accuracy, increase interpretability, and enhance clinical trust—ultimately contributing to more robust and transparent AI systems for healthcare applications.

[1] https://claude.ai

Bibliography

[1] R. Banerjee, R. Vempada, K. Mandana, A. D. Choudhury, and A. Pal, "Identifying coronary artery disease from photoplethysmogram," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 2016, pp. 1084–1088.

[2] K. E. Rudd, S. C. Johnson, K. M. Agesa, K. A. Shackelford, D. Tsoi, D. R. Kievlan, D. V. Colombara, K. S. Ikuta, N. Kissoon, S. Finfer *et al.*, "Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study," *The Lancet*, vol. 395, no. 10219, pp. 200–211, 2020.

[3] M. H. Forouzanfar, A. Afshin, L. T. Alexander, H. R. Anderson, Z. A. Bhutta, S. Biryukov, M. Brauer, R. Burnett, K. Cercy, F. J. Charlson *et al.*, "Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the global burden of disease study 2015," *The Lancet*, vol. 388, no. 10053, pp. 1659–1724, 2016.

[4] W. H. Organization *et al.*, "Hearts: technical package for cardiovascular disease management in primary health care," 2016.

[5] D. Mozaffarian, E. J. Benjamin, A. S. Go, D. K. Arnett, M. J. Blaha, M. Cushman, S. R. Das, S. de Ferranti, J.-P. Després, H. J. Fullerton *et al.*, "Heart disease and stroke statistics—2016 update: a report from the american heart association," *Circulation*, pp. CIR–0 000 000 000 000 350, 2015.

[6] S. Leeder, S. Raymond, H. Greenberg, H. Liu, K. Esson *et al.*, "A race against time: the challenge of cardiovascular disease in developing economies," *New York: Columbia University*, 2004.

[7] V. L. Feigin, E. Nichols, T. Alam, M. S. Bannick, E. Beghi, N. Blake, W. J. Culpepper, E. R. Dorsey, A. Elbaz, R. G. Ellenbogen *et al.*, "Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the global burden of disease study 2016," *The Lancet Neurology*, vol. 18, no. 5, pp. 459–480, 2019.

[8] L. S. Bernardo, A. Quezada, R. Munoz, F. M. Maia, C. R. Pereira, W. Wu, and V. H. C. de Albuquerque, "Handwritten pattern recognition for early parkinson's disease diagnosis," *Pattern recognition letters*, vol. 125, pp. 78–84, 2019.

[9] A. Kurmi, S. Biswas, S. Sen, A. Sinitca, D. Kaplun, and R. Sarkar, "An ensemble of cnn models for parkinson's disease detection using datscan images," *Diagnostics*, vol. 12, no. 5, p. 1173, 2022.

[10] R. Khaskhoussy and Y. B. Ayed, "Improving parkinson's disease recognition through voice analysis using deep learning," *Pattern Recognition Letters*, vol. 168, pp. 64–70, 2023.

[11] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith *et al.*, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *Jama*, vol. 315, no. 8, pp. 801–810, 2016.

[12] J. Hajj, N. Blaine, J. Salavaci, and D. Jacoby, "The "centrality of sepsis": a review on incidence, mortality, and cost of care," in *Healthcare*, vol. 6, no. 3. MDPI, 2018, p. 90.

[13] C. Fleischmann, A. Scherag, N. K. Adhikari, C. S. Hartog, T. Tsaganos, P. Schlattmann, D. C. Angus, and K. Reinhart, "Assessment of global incidence and mortality of hospital-treated sepsis. current estimates and limi-

tations," *American journal of respiratory and critical care medicine*, vol. 193, no. 3, pp. 259–272, 2016.

[14] A. Ameen, I. E. Fattoh, T. Abd El-Hafeez, and K. Ahmed, "Advances in ecg and pcg-based cardiovascular disease classification: a review of deep learning and machine learning methods," *Journal of Big Data*, vol. 11, no. 1, p. 159, 2024.

[15] Y. N. Fuadah and K. M. Lim, "Advances in cardiovascular signal analysis with future directions: a review of machine learning and deep learning models for cardiovascular disease classification based on ecg, pcg, and ppg signals," *Biomedical Engineering Letters*, vol. 15, no. 4, pp. 619–660, 2025.

[16] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, A. Gertych, and R. S. Tan, "A deep convolutional neural network model to classify heartbeats," *Computers in Biology and Medicine*, vol. 89, pp. 389–396, 2017.

[17] P. S. Addison, *Wavelet Transforms and the ECG: A Review.* IOP Publishing, 2005, vol. 26, no. 5.

[18] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, pp. H2039–H2049, 2000.

[19] U. R. Acharya *et al.*, "Automated detection of coronary artery disease using different durations of ecg segments with convolutional neural network," *Knowledge-Based Systems*, vol. 132, pp. 62–71, 2017.

[20] K. Iqtidar, U. Qamar, S. Aziz, and M. U. Khan, "Phonocardiogram signal analysis for classification of coronary artery diseases using mfcc and 1d adaptive local ternary patterns," *Computers in Biology and Medicine*, vol. 138, p. 104926, 2021.

[21] M. Kumar, R. Pachori, and U. Acharya, "Automated diagnosis of coronary artery disease affected patients using lda, pca, ica and discrete wavelet transform," *Knowledge-Based Systems*, vol. 43, pp. 25–36, 2013.

[22] A. Çakmak, G. Akyilmaz, A. G. Köse, G. Keskin, and L. Uğur, "Machine learning-based prediction of coronary artery disease using clinical and behavioral data: A comparative study," *Diagnostics*, vol. 16, no. 2, p. 318, 2026.

[23] R. Banerjee, S. Bhattacharya, S. Bandyopadhyay, A. Pal, and K. M. Mandana, "Non-invasive detection of coronary artery disease based on clinical information and cardiovascular signals: A two-stage classification approach," in *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, June 2018, pp. 205–210.

[24] G. D. Clifford, C. Liu, B. Moody, D. Springer, I. Silva, Q. Li, and R. G. Mark, "Classification of normal/abnormal heart sound recordings: The physionet/-computing in cardiology challenge 2016," in *2016 Computing in Cardiology Conference (CinC)*, Sep. 2016, pp. 609–612.

[25] C. Potes, S. Parvaneh, A. Rahman, and B. Conroy, "Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds," in *2016 Computing in Cardiology Conference (CinC)*, Sep. 2016, pp. 621–624.

[26] M. Yuvali, B. Yaman, and O. Tosun, "Classification comparison of machine learning algorithms using two independent cad datasets," *Mathematics*, vol. 10, no. 3, p. 311, 2022.

[27] J. Hassannataj Joloudari *et al.*, "Hybrid svm model with feature selection for coronary artery disease diagnosis," *Biomedical Signal Processing and Control*, 2022.

[28] F. Khozeimeh *et al.*, "Active learning and ensemble-based cad detection using svm," *Expert Systems with Applications*, 2023.

[29] J. H. Joloudari *et al.*, "Fcm-dnn: diagnosing coronary artery disease by fuzzy c-means clustering," *Expert Systems with Applications*, 2022.

[30] S. Upadhyay, A. K. Sagar, and N. R. Roy, "Benchmarking ensemble learning approaches for coronary artery disease classification," *International Journal of Computational Intelligence Systems*, 2026.

[31] A. Garavand, C. Salehnasab, A. Behmanesh, N. Aslani, A. H. Zadeh, and M. Ghaderzadeh, "Efficient model for coronary artery disease diagnosis: a comparative study of several machine learning algorithms," *Journal of Healthcare Engineering*, vol. 2022, no. 1, p. 5359540, 2022.

[32] Y. Zhang, P. O. Ogunbona, W. Li, B. Munro, and G. G. Wallace, "Pathological gait detection of parkinson's disease using sparse representation," in *2013 Int. Conf. on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2013, pp. 1–8.

[33] M.-J. Yang, H.-R. Zheng, H.-Y. Wang, S. Mcclean, and N. Harris, "Combining feature ranking with pca: An application to gait analysis," in *2010 Int. Conf. on Machine Learning and Cybernetics*, vol. 1. IEEE, 2010, pp. 494–499.

[34] M. N. Alam, A. Garg, T. T. K. Munia, R. Fazel-Rezai, and K. Tavakolian, "Vertical ground reaction force marker for parkinson's disease," *PLoS One*, vol. 12, no. 5, p. e0175951, 2017.

[35] R. Alkhatib, M. Diab, B. Moslem, C. Corbier, and M. E. Badaoui, "Gait-ground reaction force sensors selection based on roc curve evaluation," *Journal of Computer and Communications*, vol. 3, no. 03, p. 13, 2015.

[36] S. J. Priya, A. J. Rani, M. Subathra, M. A. Mohammed, R. Damaševičius, and N. Ubendran, "Local pattern transformation based feature extraction for recognition of parkinson's disease based on gait signals," *Diagnostics*, vol. 11, no. 8, p. 1395, 2021.

[37] P. Klinton Amaladass, M. Subathra, S. Jeba Priya, and M. Sivakumar, "Enhanced local pattern transformation based feature extraction for identification of parkinson's disease using gait signals," *SN Computer Science*, vol. 4, no. 2, p. 200, 2023.

[38] Q. Ye, Y. Xia, and Z. Yao, "Classification of gait patterns in patients with neurodegenerative disease using adaptive neuro-fuzzy inference system," *Computational and mathematical methods in medicine*, vol. 2018, 2018.

[39] A. R. Alhaidar, M. Y. Sikkandar, and A. A. Alkathiry, "Reconstruction of dual tasking gait pattern in parkinson's disease subjects using radial basis function based artificial intelligence," *Journal of Intelligent & Fuzzy Systems*, vol. 39, no. 4, pp. 5437–5448, 2020.

[40] N. Khoury, F. Attal, Y. Amirat, L. Oukhellouand, and S. Mohammed, "Data-driven based approach to aid parkinson's disease diagnosis," *Sensors*, vol. 19, no. 2, p. 242, 2019.

[41] M. A. Reyna, C. S. Josef, R. Jeter, S. P. Shashikumar, M. B. Westover, S. Nemati, G. D. Clifford, and A. Sharma, "Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019," *Critical care medicine*, vol. 48, no. 2, pp. 210–217, 2020.

[42] I. Hammoud, I. Ramakrishnan, and M. Henry, "Early prediction of sepsis using gradient boosting decision trees with optimal sample weighting," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[43] J. Singh, K. Oshiro, R. Krishnan, M. Sato, T. Ohkuma, and N. Kato, "Utilizing informative missingness for early prediction of sepsis," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. 1–4.

[44] J. H. Morrill, A. Kormilitzin, A. J. Nevado-Holgado, S. Swaminathan, S. D. Howison, and T. J. Lyons, "Utilization of the signature method to identify the early onset of sepsis from multivariate physiological time series in critical care monitoring," *Critical Care Medicine*, vol. 48, no. 10, pp. e976–e981, 2020.

[45] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, "A targeted real-time early warning score (trewscore) for septic shock," *Science translational medicine*, vol. 7, no. 299, pp. 299ra122–299ra122, 2015.

[46] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection with convolutional neural networks," *Nature Medicine*, vol. 25, no. 1, pp. 65–69, 2019.

[47] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ecg classification by 1-d convolutional neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664–675, 2015.

[48] A. Peimankar and S. Puthusserypady, "Dens-ecg: A deep learning approach for ecg signal delineation," *IEEE Access*, vol. 8, pp. 123 507–123 516, 2020.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[50] Y. Bulut *et al.*, "Deep learning-based arrhythmia detection using photoplethysmography signals," *Biomedical Signal Processing and Control*, 2024.

[51] S. T. Vasudeva, S. S. Rao *et al.*, "Development of a convolutional neural network model to predict coronary artery disease based on ecg signals," *Applied Sciences*, vol. 12, no. 15, p. 7711, 2022.

[52] A. R. W. Sait and A. K. Dutta, "Deep-learning-based coronary artery disease detection using ct images," *Diagnostics*, vol. 13, no. 7, p. 1312, 2023.

[53] N. Azdaki *et al.*, "Deep learning-based cad diagnosis using cnns," *Intelligent Systems with Applications*, p. 200507, 2025.

[54] F. Khan, X. Yu, Z. Yuan, and A. Rehman, "Ecg classification using 1-d convolutional deep residual neural network," *PLOS ONE*, vol. 18, no. 4, p. e0284791, 2023.

[55] N. K. Prajapati, H. K. Patel *et al.*, "Coronary artery disease detection using pre-trained resnet and vgg-16," *SSRG International Journal of Electronics and Communication Engineering*, vol. 11, no. 8, pp. 160–171, 2024.

[56] A. Ameen, I. Eldesouky *et al.*, "Advances in ecg and pcg-based cardiovascular disease classification: A review of deep learning methods," *Journal of Big Data*, vol. 11, p. 159, 2024.

[57] I. El Maachi, G.-A. Bilodeau, and W. Bouachir, "Deep 1d-convnet for accurate parkinson disease detection and severity prediction from gait," *Expert Systems with Applications*, vol. 143, p. 113075, 2020.

[58] A. Zhao, L. Qi, J. Li, J. Dong, and H. Yu, "A hybrid spatio-temporal model for detection and severity rating of parkinson's disease from gait data," *Neurocomputing*, vol. 315, pp. 1–8, 2018.

[59] T. D. Pham and H. Yan, "Tensor decomposition of gait dynamics in parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 8, pp. 1820–1827, 2017.

[60] N. P. Sharma, I. Junaid, and S. Ari, "Early diagnosis of parkinson's disease and severity assessment based on gait using 1d-cnn," in *2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*. IEEE, 2023, pp. 1–6.

[61] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[62] Y. B. Özçelik and A. Altan, "Overcoming nonlinear dynamics in diabetic retinopathy classification: a robust ai-based model with chaotic swarm intelligence optimization and recurrent long short-term memory," *Fractal and Fractional*, vol. 7, no. 8, p. 598, 2023.

[63] A. Altan and S. Karasu, "Recognition of covid-19 disease from x-ray images by hybrid model consisting of 2d curvelet transform, chaotic salp swarm algorithm and deep learning technique," *Chaos, solitons & fractals*, vol. 140, p. 110071, 2020.

[64] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell, "Learning to diagnose with lstm recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2016.

[65] J. Futoma, S. Hariharan, and K. Heller, "Learning to detect sepsis with a multitask gaussian process rnn classifier," *arXiv preprint arXiv:1703.07771*, 2017.

[66] A. E. Johnson, T. J. Pollard, L. Shen, H. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 2016.

[67] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eicu collaborative research database," *Scientific Data*, vol. 5, p. 180178, 2018.

[68] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer *et al.*, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, 2016.

[69] X. Li, Y. Kang, X. Jia, J. Wang, and G. Xie, "Tasp: a time-phased model for sepsis prediction," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[70] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," in *Advances in neural information processing systems*, 2007, pp. 1601–1608.

[71] D. Schneeberger, K. Stöger, and A. Holzinger, "The european legal framework for medical ai," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 2020, pp. 209–226.

[72] G. D. A. DW, "Darpa's explainable artificial intelligence program," *AI Mag*, vol. 40, no. 2, p. 44, 2019.

[73] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek, "Explainable ai methods-a brief overview," in *International workshop on extending explainable AI beyond deep models and classifiers*. Springer, 2020, pp. 13–38.

[74] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[75] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[76] I. Abubakar, T. Tillmann, and A. Banerjee, "Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the global burden of disease study 2013," *Lancet*, vol. 385, no. 9963, pp. 117–171, 2015.

[77] A. E. Moran, M. H. Forouzanfar, G. Roth, G. A. Mensah, M. Ezzati, A. Flaxman, C. J. Murray, and M. Naghavi, "The global burden of ischemic heart disease in 1990 and 2010: the global burden of disease 2010 study," *Circulation*, pp. CIRCULATIONAHA–113, 2014.

[78] S. Dua, X. Du, S. V. Sree, and T. A. VI, "Novel classification of coronary artery disease using heart rate variability analysis," *Journal of Mechanics in Medicine and Biology*, vol. 12, no. 04, p. 1240017, 2012.

[79] G. Angius, D. Barcellona, E. Cauli, L. Meloni, and L. Raffo, "Myocardial infarction and antiphospholipid syndrome: a first study on finger ppg waveforms effects," in *Computing in Cardiology (CinC), 2012*. IEEE, 2012, pp. 517–520.

[80] A. Sandryhaila and J. M. Moura, "Discrete signal processing on graphs: Frequency analysis." *IEEE Trans. Signal Processing*, vol. 62, no. 12, pp. 3042–3054, 2014.

[81] G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, J. Aerts, R. Schneider, and P. G. Bagos, "Using graph theory to analyze biological networks," *BioData mining*, vol. 4, no. 1, p. 10, 2011.

[82] D. F. Dietrich, U. Ackermann-Liebrich, C. Schindler, J.-C. Barthélémy, O. Brändli, D. R. Gold, B. Knöpfli, N. M. Probst-Hensch, F. Roche, J.-M.

Tschopp *et al.*, "Effect of physical activity on heart rate variability in normal weight, overweight and obese subjects: results from the sapaldia study," *European journal of applied physiology*, vol. 104, no. 3, pp. 557–565, 2008.

[83] B. Mika and D. Komorowski, "Higher-order spectral analysis combined with a convolution neural network for atrial fibrillation detection-preliminary study," *Sensors*, vol. 24, no. 13, p. 4171, 2024.

[84] B. Bollobás, *Modern graph theory*. Springer Science & Business Media, 2013, vol. 184.

[85] V. N. Gangapure, S. Nanda, and A. S. Chowdhury, "Superpixel based causal multisensor video fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017. [Online]. Available: https://doi.org/10.1109/TCSVT.2017.2662743

[86] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of intelligent information systems*, vol. 17, no. 2, pp. 107–145, 2001.

[87] S. Mendis, P. Puska, B. Norrving, W. H. Organization *et al.*, *Global atlas on cardiovascular disease prevention and control*. Geneva: World Health Organization, 2011.

[88] A. D. Choudhury, R. Banerjee, A. Pal, and K. Mandana, "A fusion approach for non-invasive detection of coronary artery disease," in *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*. ACM, 2017, pp. 217–220.

[89] S. E. Schmidt, J. Hansen, H. Zimmermann, D. Hammershøi, E. Toft, and J. J. Struijk, "Coronary artery disease and low frequency heart sound signatures," in *Computing in Cardiology, 2011*. IEEE, 2011, pp. 481–484.

[90] A. D. Choudhury and A. S. Chowdhury, "Change: Cardiac health analysis using graph eigenvalues," in *2018 40th Annual International Conference of*

the IEEE Engineering in Medicine and Biology Society (EMBC), July 2018, pp. 4025–4029.

[91] C. Berge, *Hypergraphs: combinatorics of finite sets*. Elsevier, 1984, vol. 45.

[92] G. Lee, S. Y. Lee, and K. Shin, "Villain: Self-supervised learning on homogeneous hypergraphs without features via virtual label propagation," in *Proceedings of the ACM Web Conference 2024*, 2024, pp. 594–605.

[93] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009, vol. 344.

[94] J. M. Hausdorff, J. Lowenthal, T. Herman, L. Gruendlinger, C. Peretz, and N. Giladi, "Rhythmic auditory stimulation modulates gait variability in parkinson's disease," *European Journal of Neuroscience*, vol. 26, no. 8, pp. 2369–2375, 2007.

[95] N. De Vries *et al.*, "Exploring the parkinson patients' perspective on home-based video recording for movement analysis: a qualitative study," *BMC neurology*, vol. 19, no. 1, pp. 1–6, 2019.

[96] T. O. Mera *et al.*, "Feasibility of home-based automated parkinson's disease motor assessment," *Journal of neuroscience methods*, vol. 203, no. 1, pp. 152–156, 2012.

[97] H. Alaskar and A. Hussain, "Prediction of parkinson disease using gait signals," in *2018 11th Int. Conf. on Developments in eSystems Engineering (DeSE)*. IEEE, 2018, pp. 23–26.

[98] A. T. Holzinger and H. Muller, "Toward human–ai interfaces to support explainability and causability in medical ai," *Computer*, vol. 54, no. 10, pp. 78–86, 2021.

[99] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.

[100] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[101] E. Zvuloni, J. Read, A. H. Ribeiro, A. L. P. Ribeiro, and J. A. Behar, "On merging feature engineering and deep learning for diagnosis, risk prediction and age estimation based on the 12-lead ecg," *IEEE Transactions on Biomedical Engineering*, 2023.

[102] Ö. F. Ertuğrul, Y. Kaya, R. Tekin, and M. N. Almalı, "Detection of parkinson's disease by shifted one dimensional local binary patterns from gait," *Expert Systems with Applications*, vol. 56, pp. 156–163, 2016.

[103] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," *Advances in neural information processing systems*, vol. 19, 2006.

[104] V. N. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, "Study the influence of normalization/transformation process on the accuracy of supervised classification," in *2020 Third Int. Conf. on Smart Systems and Inventive Technology (ICSSIT)*, 2020, pp. 729–735.

[105] J. Mei, C. Desrosiers, and J. Frasnelli, "Machine learning for the diagnosis of parkinson's disease: A review of literature," *Frontiers in aging neuroscience*, vol. 13, p. 184, 2021.

[106] Physionet.org, "Gait in Parkinson's Disease 1.0.0 Data," https://physionet.org/files/gaitpdb/1.0.0/, 2008, [Online; accessed 9-Sep-2022].

[107] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, rOC Analysis in Pattern Recognition. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016786550500303X

[108] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[109] J. N. Mazon, A. H. de Mello, G. K. Ferreira, and G. T. Rezin, "The impact of obesity on neurodegenerative diseases," *Life sciences*, vol. 182, pp. 22–28, 2017.

[110] B. Mutlu, M. E. Yeşilyurt, N. Shahbazi, M. S. Güzel, and E. A. Sezer, "Early prediction of sepsis: A comparative assessment on patients' covariates," *Biomedical Signal Processing and Control*, vol. 95, p. 106400, 2024.

[111] Q. Wu, F. Ye, Q. Gu, F. Shao, X. Long, Z. Zhan, J. Zhang, J. He, Y. Zhang, and Q. Xiao, "A customised down-sampling machine learning approach for sepsis prediction," *International Journal of Medical Informatics*, vol. 184, p. 105365, 2024.

[112] Z. Zhang, N. J. Smischney, H. Zhang, S. Van Poucke, P. Tsirigotis, J. Rello, P. M. Honore, W. S. Kuan, J. J. Ray, J. Zhou *et al.*, "Ame evidence series 001—the society for translational medicine: clinical practice guidelines for diagnosis and early identification of sepsis in the hospital," *Journal of Thoracic Disease*, vol. 8, no. 9, p. 2654, 2016.

[113] J.-L. Vincent, "The clinical challenge of sepsis identification and monitoring," *PLoS medicine*, vol. 13, no. 5, p. e1002022, 2016.

[114] G. Tsang and X. Xie, "Deep learning based sepsis intervention: the modelling and prediction of severe sepsis onset," in *2020 25th international conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 8671–8678.

[115] J. A. Du, N. Sadr, and P. de Chazal, "Automated prediction of sepsis onset using gradient boosted decision trees," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[116] T.-Y. Ross and G. Dollár, "Focal loss for dense object detection," in *proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2980–2988.

[117] H. B. Mann, "Nonparametric tests against trend," *Econometrica: Journal of the econometric society*, pp. 245–259, 1945.

[118] J. Morrill, A. Kormilitzin, A. Nevado-Holgado, S. Swaminathan, S. Howison, and T. Lyons, "The signature-based model for early detection of sepsis from electronic health records in the intensive care unit," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[119] M. Zabihi, S. Kiranyaz, and M. Gabbouj, "Sepsis prediction in intensive care unit using ensemble of xgboost models," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[120] M. Yang, X. Wang, H. Gao, Y. Li, X. Liu, J. Li, and C. Liu, "Early prediction of sepsis using multi-feature fusion based xgboost learning and bayesian optimization," in *The IEEE conference on computing in cardiology (CinC)*, vol. 46, 2019, pp. 1–4.

[121] S. Liu, B. Fu, W. Wang, M. Liu, and X. Sun, "Dynamic sepsis prediction for intensive care unit patients using xgboost-based model with novel time-dependent features," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4258–4269, 2022.

[122] M. Chen and A. Hernández, "Towards an explainable model for sepsis detection based on sensitivity analysis," *IRBM*, vol. 43, no. 1, pp. 75–86, 2022.

[123] M. Yang, C. Liu, X. Wang, Y. Li, H. Gao, X. Liu, and J. Li, "An explainable artificial intelligence predictor for early detection of sepsis," *Critical care medicine*, vol. 48, no. 11, pp. e1091–e1096, 2020.

[124] T. An, Y. Zhou, H. Zou, and J. Yang, "Iot-llm: Enhancing real-world iot task reasoning with large language models," *arXiv preprint arXiv:2410.02429*, 2024.