

Ministry of Higher Education and Scientific Research

وزارة التعليم العالي والبحث العلمي

Badji Mokhtar Annaba University
Université Badji Mokhtar – Annaba
Faculty of Technology



جامعة باجي مختار – عنابة

كلية التكنولوجيا

قسم الإلكترونيك

Department of Electronics

Thesis

Submitted to obtain the diploma of

Doctorate Third Cycle

Field : Telecommunication

Specialty : Multimedia and Digital Communications

By :

MEKAHLIA Mohammed Saddek

Title :

Classification of Heart Sound Signal Using Multiple Features and Machine Learning

N°	Name and Surname	Grade	Establishment	Role
01	AZIZI Nabiha	Prof.	Badji Mokhtar – Annaba University	President
02	FEZARI Mohammed	Prof.	Badji Mokhtar – Annaba University	Supervisor
03	HARIZE Saliha	Prof.	Badji Mokhtar – Annaba University	Examiner
04	BABOURI Abdesselam	Prof.	Université 8 Mai 45 Guelma	Examiner
05	BOUROUBA Hocine	Prof.	Université 8 Mai 45 Guelma	Examiner

*This thesis is dedicated to my parents, **M. Mohamed** and **D. Nadia**,
to my brother **M. Mohamed El Amine**, to my sister **M. Imen**,
to my cousin **M. Housseem Eddine**,
and to my niece **D. Lilyane** and her parents.*

*"The acquisition of knowledge is ultimately for the sake of the soul's perfection."
- Avicenna -*

Acknowledgments

First of all, I thank God Almighty, who gave me the strength and blessed my path in the pursuit of knowledge. I pray that He continues to grant me the energy to follow this path, so that it may benefit me and, through me, contribute to the good of the world.

I would like to express my sincere gratitude to my supervisor, Prof. Mohammed Fezari, for his guidance, expertise, and patience, which were essential to the completion of this work.

I would also like to thank all those who have taught me throughout my university years, especially Prof. Nouredine Doghmane, Dr. Nasreddine Kouadria, Dr. Saliha Harize, and Dr. Saleh Bensaoula.

My deep thanks also go to my colleagues and brothers at the LASA laboratory for their generosity, and for their academic and moral support — especially Dr. Ahcen Aliouat, Dr. Zaki Neili, Dr. Islem Mansri, and Dr. Seif Allah Nasri. I hope we will continue working together on promising research projects.

I am profoundly grateful to those who have always stood by me when I needed them most — my parents, my brother and sister, my uncles and aunts, my cousins, and my friends: Housseem, Hassan, and Mouad.

This work could not have been completed without the unwavering support of these individuals. I am deeply thankful to all of them.

« تصنيف أصوات القلب باستخدام خصائص متعددة وتقنيات التعلم الآلي »

الملخص:

تُعدّ أمراض القلب والأوعية الدموية السبب الأول للوفيات على مستوى العالم، وتُعتبر اضطرابات صمامات القلب — مثل أمراض الصمام الأبهري والميترالي — من أكثر الأسباب شيوعاً. تؤدي هذه الاضطرابات إلى اختلال في الوظيفة الميكانيكية للقلب، وغالباً ما يظهر ذلك في شكل تشوهات صوتية تُرصد في أصوات القلب. تعتبر أصوات القلب إشارات منخفضة التردد ناتجة عن النشاط الميكانيكي للقلب، وتحمل معلومات فسيولوجية مهمة. وقد تشير بعض التشوهات مثل النفخات أو الأصوات الإضافية (S3, S4) إلى أمراض قلبية كامنة. من أجل تعزيز التشخيص المبكر والميسر، ظهرت أنظمة السماعات القلبية المدعومة بالحاسوب (CAA) كحلول واعدة تعتمد على تقنيات معالجة الإشارة والتعلم الآلي لتصنيف أصوات القلب.

تتناول هذه الأطروحة تطوير طرق دقيقة وفعالة لتصنيف أصوات القلب بشكل تلقائي، بهدف دعم الكشف المبكر عن أمراض القلب والأوعية الدموية. ونظراً لمحدودية أدوات التشخيص التقليدية، خاصة في البيانات محدودة الإمكانيات، تركز الدراسة على ضرورة توفير حلول ميسرة وقابلة للتنفيذ في أنظمة التشخيص المدعومة بالحاسوب. وقد تم استكشاف عدة تمثيلات زمنية-ترددية لإشارات مخطط القلب الصوتي (PCG)، بما في ذلك Gammatonegram و Scalogram و Mel-spectrogram وتحويل الثابت Q (CQT)، من أجل تحديد الخصائص الأكثر تميزاً للتصنيف. كما تم اقتراح نموذج جديد وخفيف يعتمد على شبكة عصبية تلافيفية موجهة وغير دورية (DAG-CNN)، وتمت مقارنته بنماذج معروفة مثل VGG16 و AlexNet و ResNet و DeiT.

أظهرت النتائج التجريبية أن النهج المقترح يُحسّن دقة التصنيف بشكل ملحوظ مع تقليل التعقيد الحسابي، مما يجعله مناسباً للتطبيقات الفورية والبيئات محدودة الموارد. ويُقدّم هذا العمل حلاً عملياً وقابلاً للتوسّع لأنظمة الفحص القلبي المؤتمت، مع إمكانية دمج مستقبلات في الأنظمة السريرية أو التشخيص عن بُعد.

الكلمات المفتاحية : تصنيف أصوات القلب، مخطط القلب الصوتي، أمراض القلب والأوعية الدموية، أمراض صمامات القلب، السماع المدعومة بالحاسوب، إشارة منخفضة التردد، تمثيل زمني ترددي، التعلم العميق، DAG-CNN، معالجة الإشارة.

« Classification des sons cardiaques à l'aide de multiples caractéristiques et de l'apprentissage automatique »

Résumé :

Les maladies cardiovasculaires (MCV) représentent la première cause de mortalité dans le monde, avec les troubles des valves cardiaques — notamment les anomalies des valves aortique et mitrale — parmi les plus fréquents. Ces troubles altèrent le fonctionnement mécanique normal du cœur, ce qui se manifeste souvent par des anomalies acoustiques dans les sons cardiaques. Ces derniers, en tant que signaux de basse fréquence générés par l'activité mécanique du cœur, renferment des informations physiologiques précieuses. Les anomalies telles que les souffles ou les sons supplémentaires (S3, S4) peuvent révéler des pathologies sous-jacentes. Pour soutenir un diagnostic précoce et accessible, les systèmes d'auscultation assistée par ordinateur (CAA) se révèlent être des outils prometteurs, exploitant le traitement du signal et l'apprentissage automatique pour classer les sons cardiaques.

Cette thèse présente le développement de méthodes précises et efficaces pour la classification automatique des sons cardiaques, afin de contribuer à la détection précoce des MCV. Face aux limites des outils diagnostiques traditionnels — notamment dans les contextes à faibles ressources — l'étude souligne l'importance de solutions CAA économiques et facilement déployables. Plusieurs représentations temps-fréquence des signaux phonocardiogrammes (PCG) sont explorées, telles que le Gammatonegram, le Scalogramme, le Mel-Spectrogramme et la Transformée en Q constant (CQT), pour identifier les caractéristiques les plus discriminantes. Une architecture légère fondée sur un nouveau réseau de neurones convolutifs à graphe acyclique dirigé (DAG-CNN) est proposée, et comparée à des modèles reconnus tels que VGG16, AlexNet, ResNet et DeiT.

Les résultats expérimentaux montrent que l'approche proposée améliore significativement la précision de classification tout en réduisant la complexité computationnelle, ce qui la rend particulièrement adaptée aux applications en temps réel et dans des environnements à ressources limitées. Ce travail constitue une solution pratique et évolutive pour le dépistage cardiaque automatisé, avec un potentiel d'intégration dans les systèmes de diagnostic cliniques et à distance.

Mots-clés : Classification des sons cardiaques, phonocardiogramme, maladies cardiovasculaires, valvulopathies, auscultation assistée par ordinateur, signal basse fréquence, représentation temps-fréquence, apprentissage profond, DAG-CNN, traitement du signal.

« Classification of Heart Sound Signal Using Multiple Features and Machine Learning »

Abstract:

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, with heart valve disorders—such as aortic and mitral valve anomalies—among the most prevalent. These disorders disrupt the normal mechanical function of the heart, often reflected as acoustic anomalies in heart sounds. Heart sounds, being low-frequency signals generated by the mechanical activity of the heart, carry critical physiological information. Abnormalities such as murmurs or extra sounds (S₃, S₄) may indicate underlying pathologies. To support early and accessible diagnosis, computer-aided auscultation (CAA) systems have emerged as promising tools, leveraging signal processing and machine learning techniques to classify heart sounds.

This thesis presents the development of accurate and efficient methods for automated heart sound classification to assist in the early detection of CVDs. Given the limitations of traditional diagnostic tools—especially in low-resource settings—the study emphasizes the need for affordable and deployable CAA solutions. Several time-frequency representations of phonocardiogram (PCG) signals are explored, including Gammatonegram, Scalogram, Mel-Spectrogram, and Constant-Q Transform (CQT), to identify the most discriminative features. A lightweight architecture based on a novel Directed Acyclic Graph Convolutional Neural Network (DAG-CNN) is proposed and compared with established models such as VGG16, AlexNet, ResNet, and DeiT.

Experimental results demonstrate that the proposed approach significantly enhances classification accuracy while reducing computational complexity, making it especially suitable for real-time and resource-constrained environments. This work offers a practical and scalable solution for automated cardiac screening, with potential integration into clinical and remote diagnostic systems.

Keywords: Heart sound classification, phonocardiogram, cardiovascular diseases, heart valve disease, computer-aided auscultation, low-frequency signal, time-frequency representation, deep learning, DAG-CNN, signal processing.

List of Figures

1.1	PCG signal showing consecutive cardiac cycles over 2 seconds.	8
1.2	Heart valves illustration [32].	9
1.3	Visual comparison of heart valve stenosis and regurgitation [32].	10
1.4	Comparison between a normal mitral valve and mitral valve stenosis [42].	11
1.5	The difference between a normal tricuspid valve and a stenotic tricuspid valve [45]. .	12
1.6	Comparison of a normal aortic valve and a stenotic aortic valve [51].	13
1.7	Functional differences between a healthy and stenotic pulmonary valve[53].	13
1.8	Contrast between normal mitral valve function and regurgitation [56].	14
1.9	Illustration of tricuspid valve dysfunction showing normal closure versus regurgitant flow [60].	15
1.10	Mechanism of aortic valve regurgitation [64].	16
1.11	Hemodynamic effects of pulmonary valve regurgitation [67]	17
1.12	Schematic representation of the cardiac cycle, highlighting key phases (systole and diastole), timing intervals, and associated heart sounds (S1–S4) [68].	18
1.13	Block diagram of heart valve pathologies and associated auscultatory findings.	19
1.14	Block diagram illustrating frequency ranges of clinically significant cardiovascular sounds. [72].	20
1.15	An electronic stethoscope featuring advanced digital technology for enhanced auscultation [76].	21
1.16	Block diagram of heart sound classification.	22
2.1	Scalogram representations of different heart sound classes from the Yaseen et al. dataset [18].	42

2.2	Gammatonegram representations of different heart sound classes from the Yaseen et al. dataset [18].	44
2.3	Mel-spectrogram representations of different heart sound classes from the Yaseen et al. dataset [18].	47
2.4	CQT representations of different heart sound classes from the Yaseen et al. dataset [18].	50
3.1	Illustration of the AlexNet architecture (redrawn after [172]).	57
3.2	Illustration of the ResNet-18 architecture (redrawn after [177]).	60
3.3	Illustration of the VGG16 architecture (redrawn after [185]).	63
3.4	Illustration of the ViT architecture (redrawn after [190]).	65
3.5	Illustration of the DeiT architecture (redrawn after [191]).	66
3.6	Illustration of the proposed DAG-CNN architecture.	69
4.1	Block diagram of the proposed method.	74
4.2	Confusion matrix of AlexNet for gammatonegram image dataset.	76
4.3	Confusion matrix of AlexNet for Mel-Spectrogram image dataset.	77
4.4	Confusion matrix of AlexNet for CQT image dataset.	77
4.5	Confusion matrix of ResNet for gammatonegram image dataset.	78
4.6	Confusion matrix of ResNet for Mel-Spectrogram image dataset.	79
4.7	Confusion matrix of ResNet for CQT image dataset.	80
4.8	Confusion matrix of VGG16 for CQT image dataset.	81
4.9	Confusion matrix of VGG16 for Mel-Spectrogram image dataset.	82
4.10	Confusion matrix of VGG16 for gammatonegram image dataset.	82
4.11	Accuracy and loss curves of DeiT training on mel-spectrogram image dataset.	83
4.12	Confusion matrix of DeiT for Mel-Spectrogram image dataset.	84
4.13	Accuracy and loss curves of DAG-CNN training on Scalogram image dataset.	86
4.14	Confusion matrix of DAG-CNN for Scalogram image dataset.	86
4.15	Accuracy and loss curves of DAG-CNN training on mel-spectrogram image dataset. .	87
4.16	Confusion matrix of DAG-CNN for mel-spectrogram image dataset.	88
4.17	Accuracy and loss curves of DAG-CNN training on CQT image dataset.	89
4.18	Confusion matrix of DAG-CNN for CQT image dataset.	89
4.19	Accuracy and loss curves of DAG-CNN training on gammatonegram image dataset. .	90

4.20	Confusion matrix of DAG-CNN for gammatonegram image dataset.	91
4.21	Validation Accuracy and Loss Curves for DAG-CNN Processing the four different TFRs.	92

List of Tables

1.1	Acoustic characteristics of heart sounds (S1–S4) [69].	20
1.2	Comparative analysis of PCG datasets in heart sound classification studies (part 1). .	25
1.3	Comparative analysis of PCG datasets in heart sound classification studies (part 2). .	26
1.4	Comparative analysis of preprocessing techniques applied to heart sound databases across selected classification studies	28
1.5	Clinical implications of preprocessing choices in PCG analysis	29
1.6	Comparative evaluation of feature extraction techniques for PCG signal classification	31
1.7	Comparative analysis of classification performance across PCG studies (part 1) . . .	34
1.8	Comparative analysis of classification performance across PCG studies (part 2) . . .	35
1.9	Comparative analysis of classification performance across PCG studies (part 3) . . .	36
2.1	Distribution of samples in the Yaseen dataset [18].	40
3.1	Architecture of AlexNet as used in our PCG classification approach	56
3.2	Fine-tuning hyperparameters used for AlexNet	58
3.3	Architecture of ResNet-18 as used in our PCG classification approach	59
3.4	Fine-tuning hyperparameters used for ResNet-18	61
3.5	Architecture of VGG16 as used in our PCG classification approach	62
3.6	Fine-tuning hyperparameters used for VGG16	64
3.7	Architecture of DeiT used in our PCG classification approach	67
3.8	Fine-tuning hyperparameters for DeiT on PCG classification task	67
3.9	Customized DAG-CNN architecture used in this study	68
3.10	Fine-tuning hyperparameters used for for training the proposed DAG-CNN model . .	70

4.1	Dataset split for training and validation.	73
4.2	AlexNet classification results on different TFRs using 5-fold cross-validation	75
4.3	ResNet classification results on different TFRs using 5-fold cross-validation	78
4.4	VGG16 classification results on different TFRs using 5-fold cross-validation	80
4.5	DeiT classification results on different TFRs using 10-fold cross-validation	83
4.6	DAG-CNN classification results on Scalogram TFR using 5-fold cross-validation . .	85
4.7	DAG-CNN classification results on mel-spectrogram TFR using 5-fold cross-validation	87
4.8	DAG-CNN classification results on CQT TFR using 5-fold cross-validation	88
4.9	DAG-CNN classification results on gammatonegram TFR using 5-fold cross-validation	90
4.10	DAG-CNN classification results on scalogram, mel-spectrogram, CQT and gamma- tonegram TFR	92
4.11	Comprehensive Performance Comparison of Deep Learning Architectures Across All Time-Frequency Representations	94
4.12	Comparison of the Classification Results of the Proposed Work With Related Studies Utilizing the Same Yaseen et al. Dataset [18].	94
4.13	Comparison of Model Complexity in Terms of Total Learnable Parameters Across the Different Architectures	95

List of Abbreviations

AR	Aortic Regurgitation	MS	Mitral Stenosis
AS	Aortic Stenosis	PCA	Principal Component Analysis
CA	Certificate Authority	PCG	Phonocardiogram
CAA	Computer Aided Auscultation	PR	Pulmonary Regurgitation
CNN	Convolutional Neural Network	PS	Pulmonary Stenosis
CQT	Constant-Q Nonstationary	PSDR	Power Spectral Density Ratio
CVD	Cardiovascular Diseases	ResNet	Residual Network
CWT	Continuous Wavelet Transform	S1	Systole
DAE	Denoising Autoencoders	S2	Diastole
DAG	Directed Acyclic Graph	SDS	Spectral Density Shape
DeiT	Data-efficient Image Transformer	SGDM	Stochastic Gradient Descent with Momentum
DFT	Discrete Fourier Transform	SSA	Singular Spectrum Analysis
DHMM	Discrete Hidden Markov Model	STFT	Short-Time Fourier Transform
DL	Deep Learning		Gabor Transform
DNN	Deep Neural Networks	SVM	Support Vector Machines
DWT	Discrete Wavelet Transform	TFR	Time-Frequency Representations
ECG	Electrocardiography	TQWT	Tunable Q-factor Wavelet Transform
HMM	Hidden Markov Model	TR	Tricuspid Regurgitation
ILSVRC	ImageNet Large Scale Visual Recognition Challenge	TS	Tricuspid Stenosis
LRN	Local Response Normalization	VGG	Visual Geometry Group
MFCC	Mel-Frequency Cepstral Coefficients	VHD	Valvular Heart Diseases
ML	Machine Learning	ViT	Vision Transformer
MR	Mitral Regurgitation	VQ	Vector Quantization
MRI	Magnetic Resonance Imaging	WT	Wavelet Transform

Contents

List of Figures	vi
List of Tables	ix
List of Abbreviations	xi
Contents	xii
Introduction	1
Introduction	1
1 State of the Art on Heart Valve Diseases and PCG Classification	7
1.1 Introduction	7
1.2 Valvular Heart Diseases	9
1.2.1 Stenosis	10
1.2.1.1 Mitral Stenosis	11
1.2.1.2 Tricuspid Stenosis	11
1.2.1.3 Aortic Stenosis	12
1.2.1.4 Pulmonary Stenosis	13
1.2.2 Regurgitation	14
1.2.2.1 Mitral Regurgitation	14
1.2.2.2 Tricuspid Regurgitation	15
1.2.2.3 Aortic Regurgitation	16

1.2.2.4	Pulmonary Regurgitation	17
1.3	Electronic Stethoscope	19
1.4	Computer-Aided Auscultation	22
1.4.1	PCG Datasets	22
1.4.2	Data Preprocessing	26
1.4.3	Feature Extraction	29
1.4.4	Heart Sound Classification	32
1.5	Conclusion	36
2	Time-Frequency Representations for PCG Analysis	38
2.1	Introduction	38
2.2	Scalogram Representation	40
2.3	Gammatonegram Representation	43
2.4	Mel-spectrogram Representation	45
2.5	CQT Representation	48
2.6	Conclusion	51
3	Deep Learning Models: Architecture and Complexity	53
3.1	Introduction	53
3.2	Pretrained CNN Models	55
3.2.1	AlexNet	55
3.2.2	ResNet-18	58
3.2.3	VGG16	61
3.2.4	Data-efficient Image Transformer (DeiT)	64
3.3	Proposed Model	67
3.4	Conclusion	70
4	Performance Evaluation and Model Comparison	72
4.1	Introduction	72
4.2	AlexNet-Based Classification of PCG Signals	75
4.3	ResNet-Based Classification of PCG Signals	78
4.4	VGG16-Based Classification of PCG Signals	80

4.5	DeiT-Based Classification of PCG Signals	83
4.6	Proposed Method in PCG Classification with DAG-CNN	84
4.6.1	Scalogram Classification with DAG-CNN	85
4.6.2	Mel-spectrogram Classification with DAG-CNN	87
4.6.3	CQT Classification with DAG-CNN	88
4.6.4	Cammatonegram Classification with DAG-CNN	90
4.6.5	Comparative Analysis of TFRs with DAG-CNN	91
4.7	Comparative Analysis of Deep Learning Models with the Proposed DAG-CNN . . .	93
4.8	Conclusion	96
	Conclusion	97
	Bibliography	99

Introduction

This thesis investigates the development of accurate and robust methods for heart sound classification to aid in the early detection of cardiovascular diseases (CVDs). Given the global burden of CVDs and the limitations of traditional diagnostic tools, machine learning-based approaches offer a promising solution for automated heart sound analysis. Before exploring heart sound classification techniques, this section outlines the aim and motivation behind our research. First, the background and significance of heart sound classification in medical diagnostics are discussed. Then, we elaborate on the challenges encountered in processing and analyzing heart sound signals. Next, the objectives of this work and its key contributions are presented. Finally, the structure of the thesis is summarized in a bullet-point format.

1. Context and Motivation

Cardiovascular diseases account for approximately 17.9 million deaths annually, making them the leading cause of mortality worldwide [1]. The complexity of the heart's structure gives rise to various diseases with distinct symptoms and causes, including congenital heart disease, myocardial infarction, heart failure, and valvulopathies. Valvulopathies, primarily affecting the mitral and aortic valves, are often a consequence of aging but may also result from congenital abnormalities, rheumatic heart disease, or physiological conditions such as pregnancy [2].

The heart's primary function is to maintain blood circulation throughout the body, and its rhythmic activity—known as the cardiac cycle—is essential for sustaining life. A heartbeat consists of two fundamental phases: systole (S1) and diastole (S2). Systole (S1) is a brief phase that occurs when the tricuspid and mitral valves close, while diastole (S2) is a relatively longer period associated with the closure of the aortic and pulmonary valves [3, 4]. The analysis of heart sounds provides valuable insights into cardiac function, as abnormal beats or irregularities in the systole-diastole relationship are indicative of underlying heart conditions [5, 6, 7]. Additional acoustic components, such as S3 and S4 heart sounds, may also occur between S1 and S2. S3 is typically heard just after S2 and is generally considered normal in children, pregnant women, and young adults [6, 8]. In contrast, S4 occurs just

before S1 and is often associated with diastolic heart failure, making it an abnormal finding in most cases [9]. Furthermore, murmurs—abnormal sound signals detected within the heart sound cycle—are frequently indicative of pathological conditions [8].

Given that heart sounds are low-frequency signals, they are often difficult for the human ear to distinguish, particularly in the presence of murmurs. To overcome these limitations, various diagnostic techniques have been employed in cardiology, including auscultation, electrocardiography (ECG), echocardiography, and cardiac magnetic resonance imaging (MRI). However, while ECG, echocardiography, and MRI provide valuable diagnostic information, they are costly and require specialized personnel to operate [10]. This poses a significant challenge in low- and middle-income countries, where approximately 75% of cardiovascular disease-related deaths occur [1, 10]. Due to the limited availability of advanced diagnostic equipment in these regions, there is a critical need for a low-cost and accurate method for the early detection of cardiac abnormalities. Heart auscultation has long been a vital technique in the initial diagnosis of heart diseases, and technological advancements have led to the development of electronic stethoscopes, which have demonstrated superior efficiency and reliability compared to traditional stethoscopes [11]. The electronic stethoscope enables the recording of heart sounds, which can be visualized through phonocardiography. This recorded signal, known as the phonocardiogram (PCG), provides diagnostic insights similar to those obtained from ECG, despite the latter's limitations in detecting structural abnormalities and valve defects characterized by heart murmurs [12].

Moreover, studies have shown that medical students and inexperienced doctors make approximately three times more errors in heart sound analysis compared to specialists [13]. In response to this challenge, machine learning-based approaches have been explored for the classification of normal and abnormal heart sounds. These automated methods have the potential to facilitate cardiac disease diagnosis without the need for highly trained professionals, particularly in primary healthcare settings [14]. However, the effective implementation of such approaches necessitates addressing multiple challenges, including signal preprocessing (denoising and segmentation), robust feature extraction, and the design of an accurate classification algorithm capable of detecting murmurs. Achieving these objectives will contribute to the development of a reliable diagnostic tool that can enhance early detection and improve patient outcomes.

2. Problem Statement

Despite the critical role of heart sound analysis in diagnosing CVDs, accurate auscultation remains challenging due to the complexity of heart sounds, the presence of noise, and the reliance on clinical expertise. Traditional auscultation is subjective and prone to misinterpretation, particularly among non-specialists, leading to diagnostic errors. While advanced imaging techniques, such as echocardiography, offer more precise diagnoses, their accessibility is often limited in resource-constrained settings, making automated auscultation tools a necessary alternative.

Machine learning-based approaches have emerged as a promising solution for automated heart sound classification. However, developing a robust system requires addressing several key challenges, particularly in terms of data quality, signal processing, and model design. One of the primary issues lies in the datasets themselves.

Publicly available PCG datasets are frequently imbalanced, containing a limited number of samples with a predominance of binary labels (normal vs. abnormal), which fails to provide sufficient granularity for clinical relevance [15, 16, 17]. Moreover, recordings are often noisy, further complicating the classification process. Few datasets offer multiclass annotations that specify the exact type of valvular disease, significantly hindering the development of models capable of distinguishing between various pathological conditions [18]. In an attempt to overcome these shortcomings, extensive preprocessing steps such as filtering, segmentation, and data augmentation are commonly applied. While these techniques can improve model performance, they also increase computational cost and risk introducing biases that affect model generalization.

Another key challenge lies in the representation of audio signals. Traditional time-frequency analysis techniques, such as the Short-Time Fourier Transform (STFT) and spectrograms, divide the frequency spectrum into uniform sub-bands [19]. However, this uniform representation does not align with the non-linear sensitivity of the human auditory system, which is more attuned to lower frequencies [20]. As a result, important acoustic features may be underrepresented, reducing the effectiveness of these representations in capturing the diagnostic cues embedded in heart sounds. Moreover, segmentation and augmentation strategies, while helpful in mitigating data scarcity, often fail to truly reflect the variability found in real-world clinical environments, and they tend to increase the model’s susceptibility to overfitting.

In terms of classification models, many existing works rely on complex pretrained deep learning architectures such as VGG16 and VGG19. While these networks offer strong baseline performance in generic audio classification tasks, they are computationally intensive and not optimized for the specific characteristics of PCG data. Their complexity often necessitates additional data preprocessing and augmentation, which, combined with the inherent limitations of available datasets, can result in suboptimal performance and reduced generalizability. Therefore, there is a clear need for a customized deep learning model that is both lightweight and effective, capable of learning meaningful representations from limited and imbalanced data without requiring excessive preprocessing.

Addressing these issues is essential to the development of accessible, reliable, and efficient computer-aided auscultation (CAA) systems. In this thesis, we propose a deep learning framework specifically tailored for PCG classification, with a focus on optimizing data handling, enhancing signal representation, and designing a model architecture that balances accuracy with computational efficiency. Such a system could significantly improve early detection of cardiovascular diseases, particularly in under-resourced healthcare settings where access to expert clinicians and advanced diagnostic tools is limited.

3. Objectives and Contributions

The research goals of our thesis are outlined within the context of automated heart sound classification using deep learning techniques. The primary objective is to develop an efficient and accurate method for classifying PCG signals to aid in the early detection of cardiovascular diseases. This involves evaluating the effectiveness of different time-frequency representations of PCG signals and comparing the performance of various convolutional neural network (CNN) architectures. Specifically, the following goals were pursued in this work:

- Investigate and compare multiple time-frequency representations of heart sound signals—including Gammatonegram, Scalogram, Mel-Spectrogram, and Constant-Q Nonstationary Gabor Transform (CQT)—with the aim of identifying the most suitable representation for PCG classification.
- Develop a Directed Acyclic Graph Convolutional Neural Network (DAG-CNN) model tailored for efficient and accurate classification of heart sound recordings, especially under data-constrained conditions.
- Evaluate and compare the classification performance of the proposed DAG-CNN against well-known pretrained deep learning models such as VGG16, AlexNet, ResNet, and the Vision Transformer-based DeiT (Data-efficient Image Transformer).
- Contribute to the advancement of CAA systems by proposing a scalable and accessible solution for PCG classification that can support medical practitioners in resource-limited environments.

In alignment with the stated objectives, the following contributions have been achieved in this thesis:

Contribution 1:

- J1 Mekahlia, Mohammed Saddek, Mohamed Fezari, and Ahcen Aliouat. "A Comparative Analysis of Constant-Q Transform, Gammatonegram, and Mel-Spectrogram Techniques for AI-aided Cardiac Diagnostics." *Medical Engineering & Physics*, Volume 137 (March 2025): 104302. <https://doi.org/10.1016/j.medengphy.2025.104302>

Contribution 2:

- J2 Mekahlia, Mohammed Saddek, Mohamed Fezari, and Ahcen Aliouat. "PCG Classification Using Scalogram and CNN With DAG Architecture." *International Journal of Informatics and Applied Mathematics* 5, no. 1 (June 2022): 62–73. <https://doi.org/10.53508/ijiam.1026460>.

Contribution 3:

- C1 Mekahlia, Mohammed Saddek, Mohamed Fezari, and Ahcen Aliouat. "CQT-Based Convolutional Neural Networks For PCG Classification." 3rd Conference on Electrical Engineering (CEE'2023). International Conference, Algiers, ALGERIA, May 2023.

Contribution 4:

- C2 Mekahlia, Mohammed Saddek, Mohamed Fezari, and Ahcen Aliouat. "Classification of Heart Sound Signals Using Scalogram and Convolutional Neural Network." 4th Conference on In-

formatics and Applied Mathematics (IAM'21). International Conference, Guelma, ALGERIA, December 2021.

Contribution 5:

C3 Mekahlia, Mohammed Saddek, Mohamed Fezari, and Ahcen Aliouat. "Classification of Heart Sound Signal Using Multiple Features and Machine Learning." Eighth Doctoral Days, Annaba, ALGERIA, June 27–28, 2021.

Papers Being Prepared for Publication:

— Mekahlia, Mohammed Saddek, Mohamed Fezari, and Ahcen Aliouat. "Data-Efficient Image Transformer-Based Heart Sound Classification Leveraging Mel-Spectrogram."

4. Thesis Outline

This thesis is structured into four main chapters, with an introduction and conclusion, as outlined below:

– Introduction

This section presents the overall context of the study, outlining the motivation, problem statement, and research objectives. It also provides an overview of the challenges associated with PCG signal analysis and classification, focusing on heart sound diagnostics and the contributions of the thesis.

Chapter 1 – State of the Art on Valvular Heart Diseases and PCG Classification

This chapter provides a comprehensive review of the literature on valvular heart diseases (VHDs) and their diagnosis using PCG signals. It discusses the physiological aspects of heart sounds, including different types of heart murmurs, and reviews the current methods and techniques for PCG classification, with an emphasis on both classical and AI-based approaches.

Chapter 2 – Time-Frequency Representations for PCG Analysis

This chapter delves into the time-frequency representations (TFRs) used for transforming PCG signals into formats suitable for deep learning classification. It presents a comparative analysis of various TFRs, such as scalogram, mel-spectrogram, Gammatonegram, and CQT. The chapter evaluates the effectiveness of these representations in highlighting key features of heart sounds relevant to the detection of pathologies.

Chapter 3 – Deep Learning Models: Architecture and Complexity

This chapter explores the complexity and architecture of various deep learning models used for PCG classification. It compares different pre-trained convolutional neural networks, including VGG16, ResNet, AlexNet, and the vision transformer DeiT, alongside a custom-designed DAG-CNN model. The chapter focuses on analyzing the model architectures, their complexity, and their suitability for heart sound classification tasks.

Chapter 4 – Performance Evaluation and Model Comparison

In this chapter, the performance of the deep learning models evaluated in the previous chapter is compared, using a multiclass dataset for PCG classification. The models are tested for their accuracy, efficiency, and suitability for the task. Based on the results, the most appropriate model for heart sound classification is identified and discussed.

– Conclusions and Future Work

This section concludes the thesis by summarizing the main findings, contributions, and the implications of the results. It also discusses the limitations of the current work and suggests potential avenues for future research, focusing on improvements to model robustness and the integration of AI-based diagnostic systems for clinical applications.

Chapter

1

State of the Art on Heart Valve Diseases and PCG Classification

This chapter provides a comprehensive introduction to the context and foundations of heart sound analysis. In Section 1.2, we explore VHDs, emphasizing their clinical relevance and the importance of early detection. Section 1.3 introduces the Electronic Stethoscope, a key technological advancement that enables the high-quality acquisition of heart sounds. Finally, Section 1.4 presents the concept of Computer-aided Auscultation, outlining the main stages of automated heart sound analysis systems—preprocessing, feature extraction, and classification—which form the basis for the methodology proposed in the next chapters.

1.1 Introduction

CVDs are the leading cause of death globally, responsible for an estimated 17.9 million deaths each year, which represents 31% of all global deaths [1]. Approximately 85% of these deaths are caused by heart attacks and strokes, with one-third occurring in individuals under 70 years of age. More than 75% of CVD-related deaths occur in low- and middle-income countries [1]. These alarming figures highlight the urgent need for accessible and cost-effective diagnostic tools to detect and monitor cardiovascular conditions, particularly in resource-limited settings.

In specific regions, such as North Africa and the Middle East, over one million people die annually from CVDs. In sub-Saharan Africa alone, the yearly death toll exceeds one million, with East Central, Eastern, Southern, and Western regions reporting significant numbers of CVD-related fatalities [21]. In the Maghreb region, CVDs are also the dominant cause of mortality. The number of CVD deaths rose from 137,320 (32.2%) in 1990 to 245,893 (42%) in 2017 [22]. Algeria follows this trend, where cardiovascular conditions are responsible for about 36% of annual deaths, corresponding to an estimated 79,389 people per year [23, 24].

The increasing burden of CVDs is driven by modifiable behavioral risk factors such as tobacco use, physical inactivity, unhealthy diet, obesity, high blood pressure, elevated cholesterol levels, and high blood glucose. Prevention strategies focusing on these factors, combined with emerging technologies in electronic and computer sciences, are critical in reducing disease incidence and mortality [1]. Projections estimate that annual CVD-related deaths will surpass 23.6 million by 2030 if current trends continue [25].

Early and accurate diagnosis of heart valve diseases, a significant subset of CVDs, is essential to reduce the associated mortality and morbidity. Auscultation remains one of the most common non-invasive diagnostic methods used to evaluate heart function. Through this process, physicians analyze the heart sounds—acoustic signals produced primarily by the opening and closing of the cardiac valves. However, auscultation depends heavily on the clinician’s experience and is subject to variability [26, 27].

Phonocardiography, the process of recording and analyzing heart sounds, provides a visual and quantitative alternative to traditional auscultation. As illustrated in Figure 1.1, the resulting signals, known as PCGs, can be further processed using signal analysis and machine learning techniques to support automated classification of normal and abnormal heart sounds [28]. However, interpreting PCG signals is challenging due to their non-stationary nature, susceptibility to noise, and inter-patient variability. These challenges make it difficult to extract robust features that can generalize well across different datasets and clinical contexts.

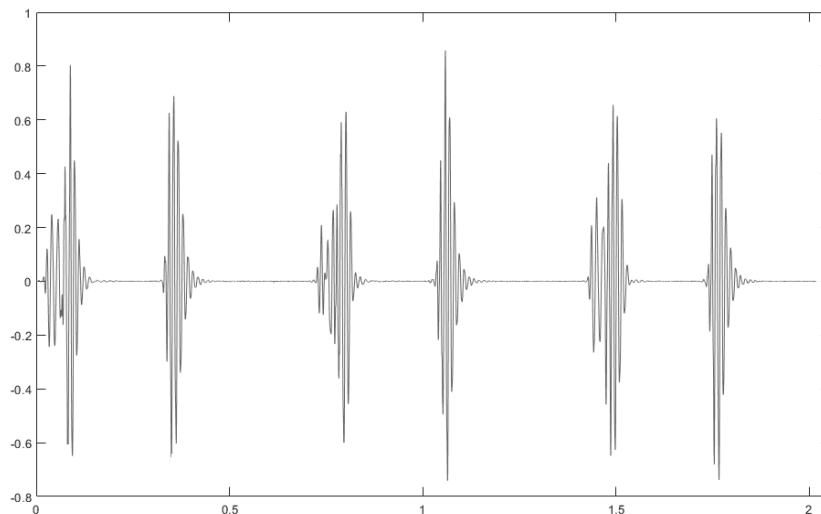


Figure 1.1 – PCG signal showing consecutive cardiac cycles over 2 seconds.

In recent years, machine learning and deep learning approaches have shown promising results in medical signal classification tasks, including PCG analysis [29, 30]. These techniques benefit from the use of effective time-frequency representations, such as the Mel-spectrogram, Gammatonegram, Scalogram, and CQT, which allow for capturing the temporal and spectral properties of heart sounds. Combining these representations with powerful CNNs enables the construction of end-to-end diagnostic

systems capable of handling the complexities inherent in PCG signals.

This chapter provides a detailed overview of the physiological basis of VHDs, common heart valve disorders, and the clinical relevance of PCG-based diagnosis. It also discusses traditional and contemporary methods used for PCG classification, highlighting the challenges that motivate the need for improved computational approaches in the domain of intelligent cardiac auscultation.

1.2 Valvular Heart Diseases

The heart is a muscular organ that pumps blood from the veins into the arteries, distributing oxygen and nutrients throughout the body via the circulatory system. It is located between the lungs, in the middle compartment of the chest [31]. Due to its essential role, any dysfunction or abnormality may cause serious effects on the human body [2].

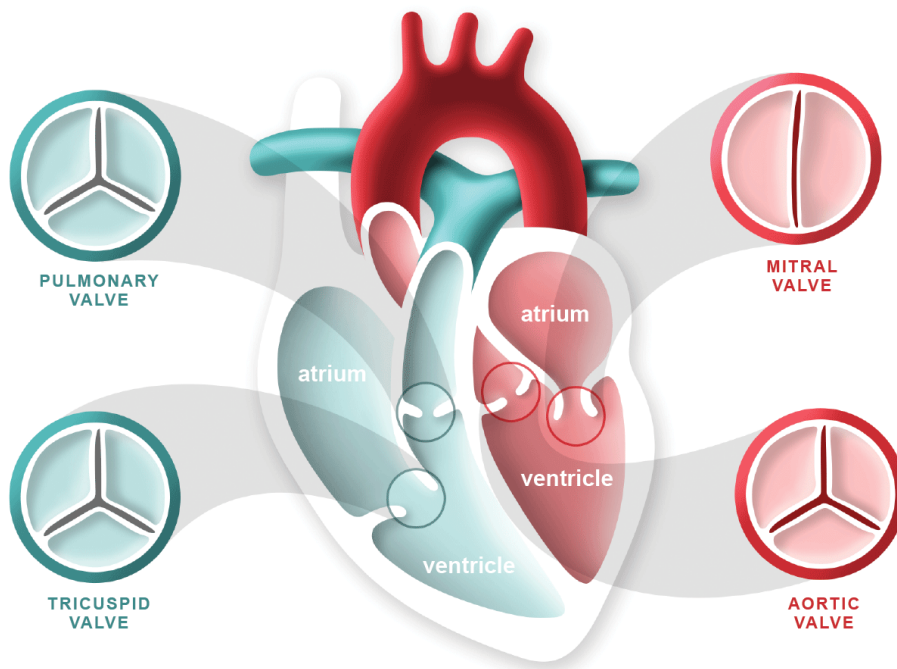


Figure 1.2 – Heart valves illustration [32].

The heart is divided into four chambers: two atria in the upper part and two ventricles in the lower part [31, 33, 34]. The atria and ventricles are separated by atrioventricular valves—the tricuspid valve on the right and the mitral valve on the left [31, 33]. The right atrium receives deoxygenated blood from the body and passes it through the tricuspid valve to the right ventricle, which pumps it to the lungs via the pulmonary valve. Oxygenated blood returns to the left atrium and flows through the mitral valve into the left ventricle, which ejects it to the systemic circulation through the aortic valve [31]. Fig. 1.2 Illustration of the heart valves. This Figure highlights the key heart valves, including the tricuspid, mitral, pulmonary, and aortic valves, and shows their locations between the chambers of the heart.

VHDs, also known as valvulopathies, refer to disorders affecting the heart valves. The heart contains four valves, each with leaflets that open and close with each heartbeat. The mitral valve is the only bicuspid valve with two leaflets, while the other three valves have three leaflets [35]. As they open and close, the valves produce sounds, often referred to as heartbeats [31].

In some cases, due to malformation or other causes, one or more valves may not function properly, either failing to open or close completely. This impairs blood flow through the heart and body. Such conditions often result in noticeable heartbeats and murmurs, which vary depending on the affected valve and the severity of the disease [31]. The most common forms of valvulopathies include stenosis and regurgitation, which can affect any of the four heart valves. Figure 1.3 presents a visual comparison between the two major types of heart valve dysfunction: stenosis and regurgitation.

In this section, we provide an overview of the stenosis and regurgitation affecting the four valves.

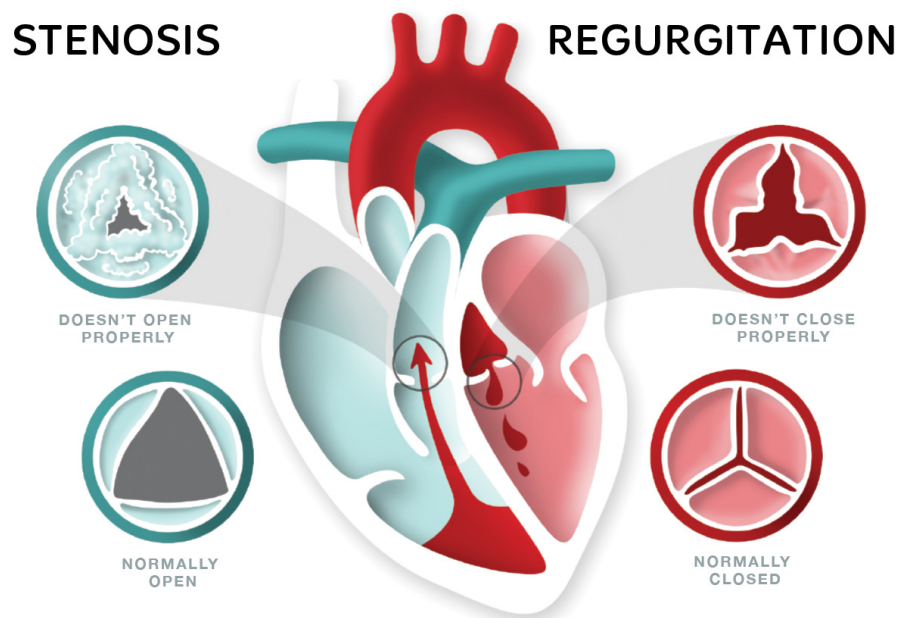


Figure 1.3 – Visual comparison of heart valve stenosis and regurgitation [32].

1.2.1 Stenosis

Stenosis refers to the condition in which the leaflets of a heart valve become thickened, stiff, or fused, thereby preventing the valve from opening properly. As a result, the heart is forced to work harder to pump and circulate blood, leading to a reduced supply of oxygen to the body [36]. Since each heart valve has a distinct anatomical structure and function, the causes and effects of stenosis vary depending on the valve affected. The most common causes of valve stenosis include congenital heart defects, rheumatic fever, calcification, diabetes, carcinoid syndrome, hypercholesterolemia, and tumors [36, 37].

1.2.1.1 Mitral Stenosis

The mitral valve, also known as the bicuspid or left atrioventricular valve, is located between the left atrium and the left ventricle. It consists of two leaflets and functions to prevent the backflow of blood into the left atrium during ventricular contraction [31, 35]. Mitral stenosis (MS) refers to the narrowing of this valve, most commonly resulting from rheumatic heart disease [37]. Other, less common causes include congenital abnormalities, left atrial myxoma, ball valve thrombus, mucopolysaccharidosis, and severe annular calcification [33].

Mitral stenosis impedes blood flow from the left atrium to the left ventricle, thereby reducing the volume of oxygenated blood returning from the lungs [38]. Clinically, it is characterized by a loud first heart sound (S1), an opening snap following the second heart sound (S2), and a low-pitched diastolic rumble with presystolic accentuation [39, 40]. Symptoms may include shortness of breath, cough (with or without blood), chest pain, fatigue, swelling in the lower limbs, frequent respiratory infections, flushed cheeks, and a rapid heartbeat in severe cases [41]. Figure 1.4 illustrates a comparison between a normal mitral valve and one affected by mitral valve stenosis

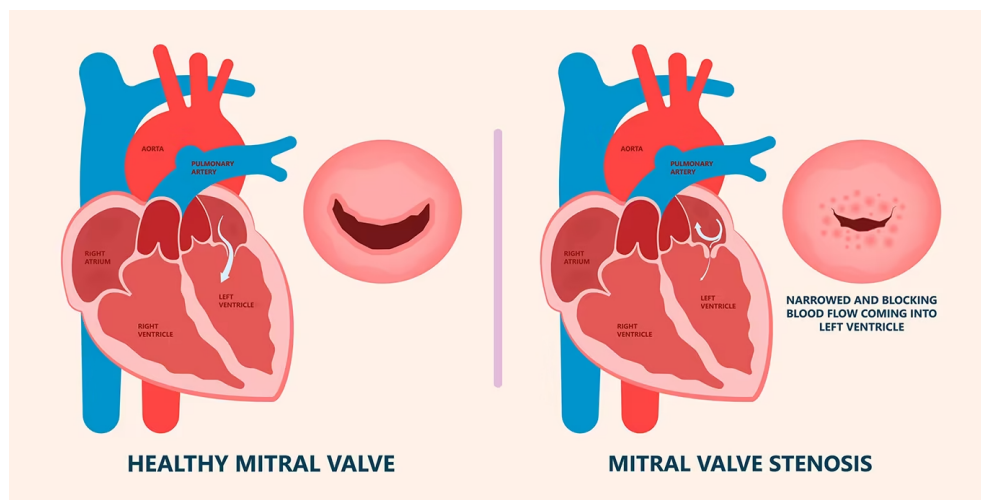


Figure 1.4 – Comparison between a normal mitral valve and mitral valve stenosis [42].

1.2.1.2 Tricuspid Stenosis

The tricuspid valve, also known as the right atrioventricular valve, is located between the right atrium and the right ventricle. It consists of three leaflets that are thinner than those of the mitral valve. Like the mitral valve, it prevents the backflow of blood into the atrium during systole [31, 35]. As illustrated in Figure 1.5, tricuspid stenosis (TS) refers to the narrowing of the tricuspid valve opening, which restricts blood flow from the right atrium to the right ventricle. As a result, less blood enters the right ventricle, leading to lower pressure and reduced pulmonary circulation [43].

Isolated tricuspid stenosis is uncommon and is most often associated with rheumatic heart disease. Less

frequent causes include carcinoid syndrome, infective endocarditis, and congenital abnormalities [33]. Auscultatory signs include a systolic murmur and a mid-diastolic murmur that follows the tricuspid opening snap, accompanied by a presystolic rumble [44]. Although symptoms are usually mild, they may include palpitations, a fluttering sensation in the chest, fatigue, swelling in the legs, and cold extremities [43].

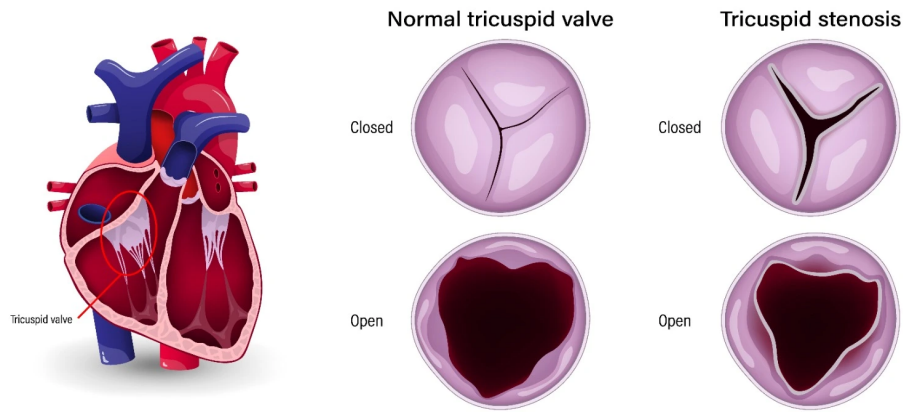


Figure 1.5 – The difference between a normal tricuspid valve and a stenotic tricuspid valve [45].

1.2.1.3 Aortic Stenosis

The aortic valve is one of the two semilunar valves in the heart and is located between the left ventricle and the aorta [31]. It consists of three leaflets: the left coronary, right coronary, and noncoronary cusps [35]. This valve prevents the backflow of blood from the aorta into the left ventricle during diastole [31]. Among all cardiac valves, the aortic valve is the most frequently affected by disease, making aortic stenosis (AS) the most common form of valvular heart disease [35].

Aortic stenosis refers to the narrowing of the aortic valve opening. In younger individuals, it is commonly caused by congenital bicuspid aortic valves or rheumatic fever [37]. Bicuspid aortic valve, the most common congenital cardiac malformation, is present in approximately 1–2% of the general population and results from abnormal leaflet formation during valvulogenesis [35, 37]. In adults, the leading cause is calcific degeneration of either a normal trileaflet or a congenital bicuspid valve. This calcification process begins at the base of the cusps and progressively extends to the leaflets, limiting their mobility and reducing the effective valve area [33].

Functionally, aortic stenosis impedes blood ejection from the left ventricle to the aorta and may elevate left atrial pressure due to impaired forward flow [46]. Auscultatory findings typically include a crescendo-decrescendo systolic murmur, absence of S2 splitting due to restricted leaflet motion, and an audible S4 [47, 48]. The systolic murmur characteristically intensifies to a peak in mid-systole before tapering off prior to S2 [49]. While early stages may be asymptomatic, advanced aortic stenosis presents with symptoms such as chest pain, fatigue following exertion, shortness of breath, palpitations,

or irregular heartbeats [50]. As shown in Figure 1.6, aortic stenosis leads to a constricted valve opening, significantly impeding blood flow compared to a healthy aortic valve

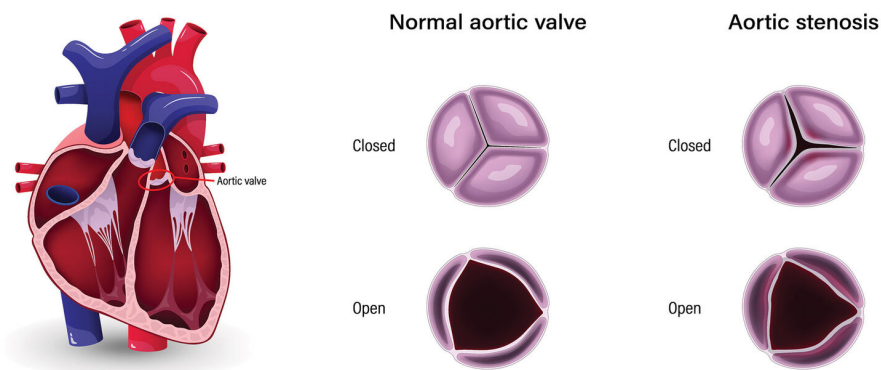


Figure 1.6 – Comparison of a normal aortic valve and a stenotic aortic valve [51].

1.2.1.4 Pulmonary Stenosis

The pulmonary valve, also referred to as a semilunar valve, is composed of three leaflets and is located between the right ventricle and the pulmonary artery. It prevents the backflow of blood into the right ventricle during diastole [31]. Pulmonary valve disease most commonly occurs in individuals with congenital heart anomalies and is rarely observed in acquired forms. Pulmonary stenosis (PS) is typically an isolated congenital anomaly, accounting for approximately 10% of all congenital heart defects. Acquired causes are uncommon and may include carcinoid syndrome, rheumatic fever, or infective endocarditis [37].

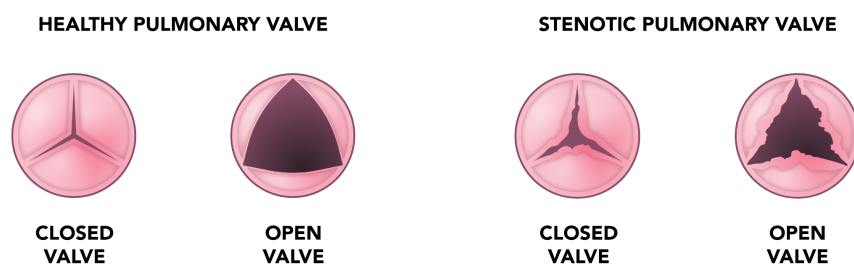


Figure 1.7 – Functional differences between a healthy and stenotic pulmonary valve[53].

As highlighted in Figure 1.7 Pulmonary valve stenosis restricts the flow of blood from the right ventricle

to the pulmonary arteries, which transport deoxygenated blood to the lungs [52]. The auscultatory characteristics of pulmonary stenosis differ from those of aortic stenosis. The systolic murmur in pulmonary stenosis is less pronounced in mid-systole and often extends to or beyond the second heart sound (S2) [49]. Mild pulmonary stenosis may remain asymptomatic. However, in moderate to severe cases, patients may present with symptoms such as exercise intolerance, fatigue, fainting, shortness of breath, chest pain, or palpitations [52]. Figure 7 presents pulmonary valve stenosis.

1.2.2 Regurgitation

Regurgitation refers to the condition in which heart valves leak, allowing blood to flow backward through the valve either during leaflet closure or due to improper leaflet coaptation. This backward flow can occur when the valve does not close completely or correctly. In some cases, the condition is mild and may not require intervention; however, in more severe instances, regurgitation can increase the workload on the heart and reduce the amount of blood it is able to pump efficiently [54]. The physiological impact of regurgitation varies depending on the affected valve and the severity of the condition. Common causes include congenital heart defects, rheumatic heart disease, and, less frequently, other pathological factors [37]. Similarly, the symptoms of heart valve regurgitation differ according to the valve involved and the extent of the dysfunction.

1.2.2.1 Mitral Regurgitation

Mitral regurgitation (MR) results from incomplete coaptation of the mitral valve leaflets, permitting the backward flow of blood into the left atrium during systole [37].

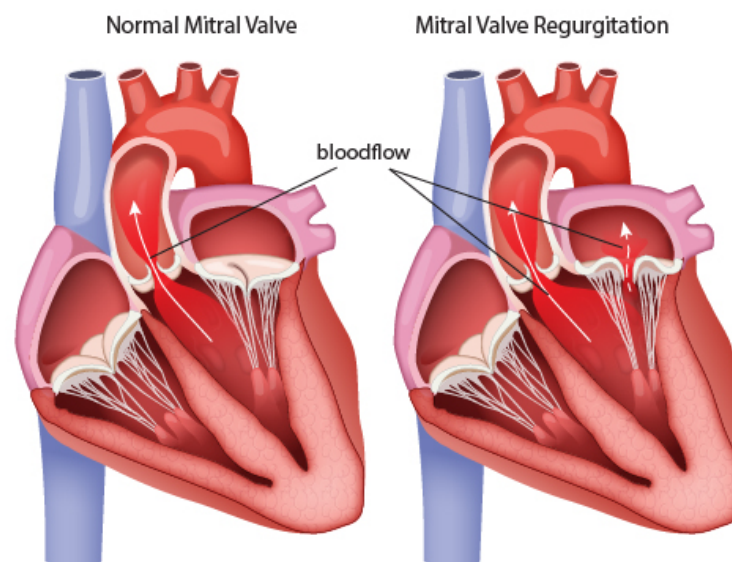


Figure 1.8 – Contrast between normal mitral valve function and regurgitation [56].

Common causes of mitral regurgitation include mitral valve prolapse, rheumatic heart disease, infective endocarditis, certain medications, and connective tissue disorders such as collagen vascular diseases [33]. Among these, mitral valve prolapse is the most prevalent cause and occurs when the valve leaflets “billow” backward into the atrium during systole due to an underlying connective tissue abnormality [37]. Auscultatory findings in mitral regurgitation include a systolic murmur, which is often holosystolic. However, in cases associated with mitral valve prolapse, it is typically mid-to-late systolic and accompanied by a characteristic systolic click. The murmur’s intensity and timing often correlate with the severity of the regurgitation [55]. In severe cases, the systolic murmur may be soft or even absent [33, 55]. While individuals with mild to moderate mitral regurgitation may remain asymptomatic for years, those with acute severe mitral regurgitation typically present with significant symptoms such as chest pain, dyspnea, and peripheral edema [55]. As demonstrated in Figure 1.8, mitral valve regurgitation results in retrograde blood flow, causing volume overload in the left atrium and compromising cardiac efficiency.

1.2.2.2 Tricuspid Regurgitation

Tricuspid regurgitation (TR) is characterized by retrograde blood flow from the right ventricle to the right atrium due to valve insufficiency [57]. Abnormalities affecting any component of the atrioventricular valve apparatus—including the leaflets, chordae tendineae, annulus, papillary muscles, or adjacent right ventricular myocardium—may contribute to the development of tricuspid regurgitation [37].

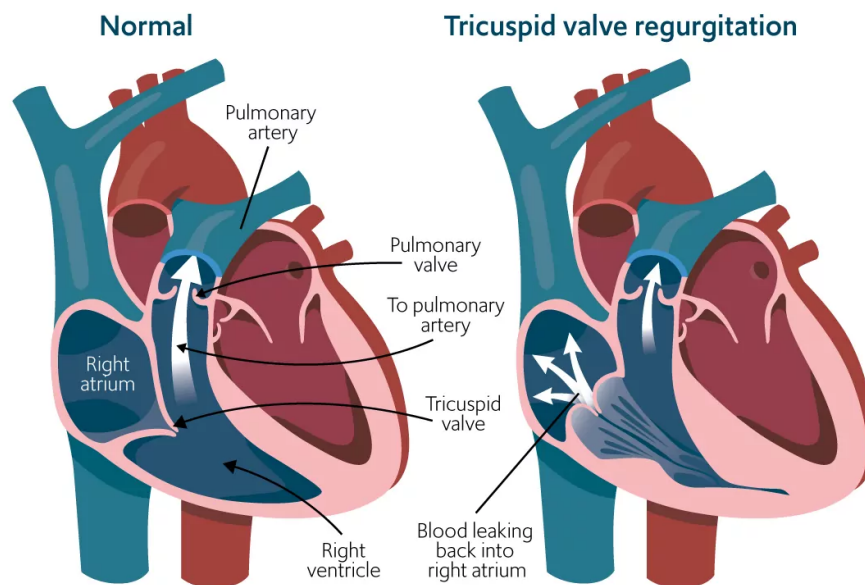


Figure 1.9 – Illustration of tricuspid valve dysfunction showing normal closure versus regurgitant flow [60].

The most common causes include congenital malformations, rheumatic heart disease, carcinoid syndrome, infective endocarditis, traumatic valve injury, endomyocardial fibrosis, and myxomatous degeneration of the valve leaflets [57]. The murmur associated with tricuspid regurgitation is typically high-pitched and holosystolic, resembling that of mitral regurgitation [58]. Auscultatory findings vary with the severity of regurgitation: mild cases may present with an early systolic murmur, while more severe cases exhibit a holosystolic murmur [57]. In advanced stages, additional findings may include a loud first heart sound (S1), an accentuated pulmonary component of the second heart sound (P2), an apical mid-diastolic rumble, and a holosystolic murmur best heard at the left lower sternal border, which increases with inspiration (Carvallo's sign) [59]. Clinical symptoms, which also depend on the severity of the regurgitation, may include ascites, cyanosis, jaundice, hepatomegaly, peripheral edema, fatigue, and dyspnea [57]. Figure 1.9 demonstrates tricuspid regurgitation's characteristic reversed blood flow during systole, which can lead to systemic venous engorgement and right ventricular volume overload.

1.2.2.3 Aortic Regurgitation

As presented in Figure 1.10, aortic regurgitation (AR) is characterized by abnormal retrograde blood flow through the aortic valve during diastole, leading to an increase in left ventricular end-diastolic volume [61]. The most common etiologies include rheumatic heart disease and infective endocarditis [37]. Congenital abnormalities, particularly a bicuspid aortic valve, are also significant causes [61]. In rarer cases, aortic regurgitation may result from dilatation of the aortic root, often associated with connective tissue disorders such as Marfan syndrome or Ehlers-Danlos syndrome [37]. On auscultation, the characteristic murmur of aortic regurgitation is a soft, high-pitched, early diastolic decrescendo murmur best heard along the left sternal border [62].

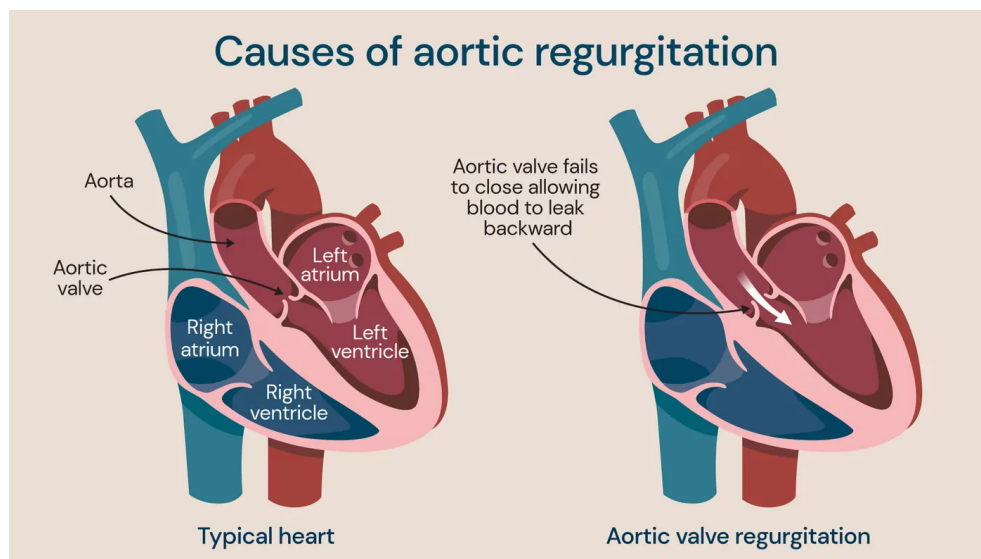


Figure 1.10 – Mechanism of aortic valve regurgitation [64].

The first heart sound (S1) may be diminished due to elevated left ventricular filling pressure, which reduces the force of mitral valve closure. Additionally, a brief early systolic murmur may be heard due to disturbed flow, and an S3 may be present, suggesting left ventricular dysfunction—though it is not specific to the severity of the regurgitation [61]. Patients with more advanced aortic regurgitation may experience palpitations, chest pain, fatigue, or dyspnea. Other clinical signs can include orthopnea, generalized weakness, syncope, and peripheral edema such as swollen ankles and feet [63].

1.2.2.4 Pulmonary Regurgitation

Pulmonary regurgitation (PR) refers to the leakage of the pulmonary valve, allowing blood to flow back into the right ventricle from the pulmonary artery before reaching the lungs for oxygenation [65]. Among the four cardiac valves, the pulmonary valve is the least commonly affected by acquired disease; thus, most cases of pulmonary regurgitation are congenital in origin. Pathologic pulmonary valve regurgitation is typically caused by dilation of the pulmonary valve annulus or the pulmonary artery, often secondary to pulmonary hypertension or idiopathic pulmonary artery dilation [66]. Less common etiologies include carcinoid syndrome and rheumatic fever [65]. On auscultation, the murmur of pulmonary regurgitation is described as a high-pitched, early diastolic decrescendo murmur, beginning at the pulmonary component of the second heart sound (P2). This murmur is often difficult to distinguish from that of aortic regurgitation, as it also begins when the pressure difference between the pulmonary artery and the right ventricle is greatest and gradually fades [66]. In the early stages, pulmonary regurgitation is often asymptomatic and may only be detected through the presence of a heart murmur during routine examination.

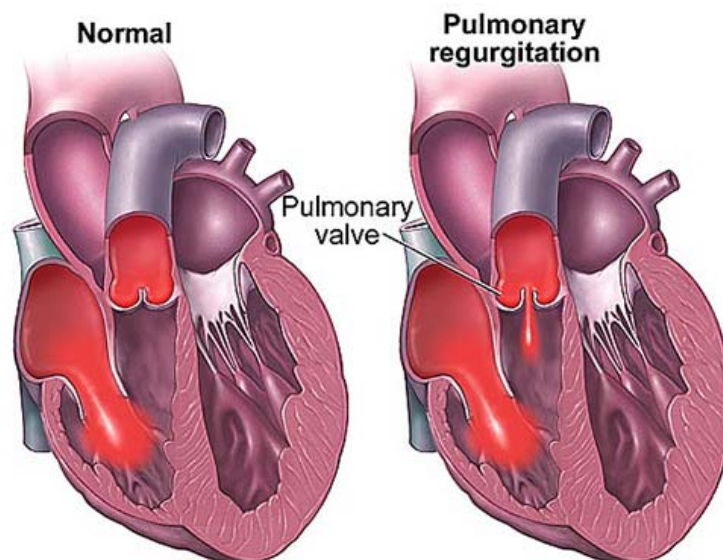


Figure 1.11 – Hemodynamic effects of pulmonary valve regurgitation [67]

However, as the condition progresses, right ventricular dilation may occur, and in rare cases, it can

advance to right-sided heart failure. Symptoms at this stage may include chest discomfort, fatigue, lightheadedness, or syncope. As illustrated in Figure 1.11, pulmonary regurgitation results in retrograde flow during diastole, which may cause right ventricular dilation and eventual functional impairment.

Heart valves may suffer from either stenosis or regurgitation. While all four valves can be affected, the mitral and aortic valves are the most commonly involved. Early detection of valvular heart diseases is paramount, as it significantly improves recovery outcomes and may even avoid the need for surgical intervention in many cases. Consequently, accurate and timely diagnosis is critical, requiring reliable and precise diagnostic tools.

In recent years, low-cost diagnostic methods have gained prominence, particularly in low- and middle-income countries where access to advanced healthcare is limited [1]. Among these, computer-aided auscultation has emerged as a promising and cost-effective technique. By using electronic stethoscopes for heart sound recording and data acquisition, it provides a practical and affordable approach to cardiac evaluation, with growing reliability in clinical settings. Figure 1.12 illustrates systole and diastole across a 0.8-second cardiac cycle, annotated with key intervals and subphases (early, middle, late). Heart sounds (S1–S4) and murmurs would align with specific phases.

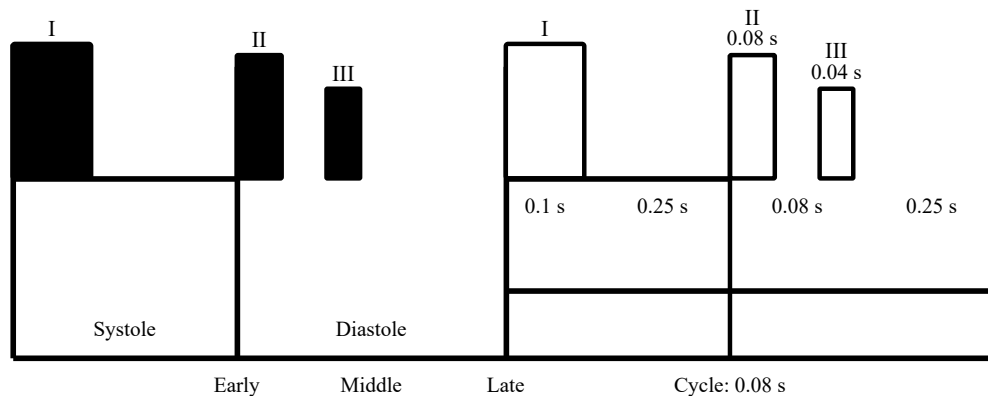


Figure 1.12 – Schematic representation of the cardiac cycle, highlighting key phases (systole and diastole), timing intervals, and associated heart sounds (S1–S4) [68].

As illustrated in Figure 1.13, heart murmurs exhibit distinct acoustic patterns in response to specific valvular malformations. For instance, mitral stenosis produces a characteristic low-pitched diastolic rumble with loud S1, while aortic regurgitation generates a high-pitched decrescendo murmur. These pathognomonic sound signatures, when analyzed through computer-aided systems, can enhance diagnostic accuracy and support early intervention strategies.

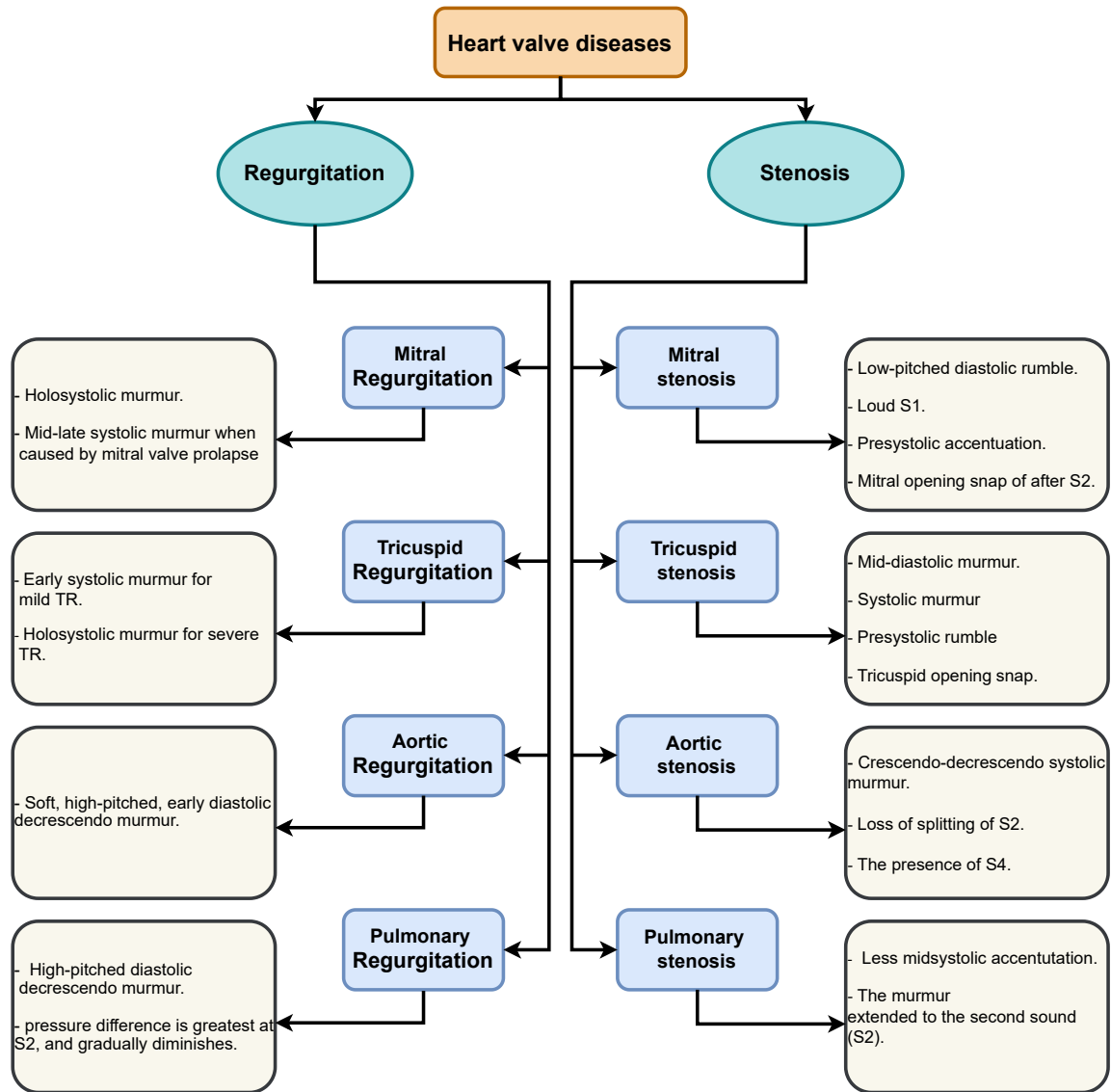


Figure 1.13 – Block diagram of heart valve pathologies and associated auscultatory findings.

1.3 Electronic Stethoscope

Cardiac auscultation holds a high predictive value for identifying numerous serious heart conditions. In certain cases, it may even lead to a definitive diagnosis, making it a quick, cost-effective, and non-invasive diagnostic tool [68]. During auscultation, physicians focus particularly on abnormal heart sounds and various types of murmurs, which may indicate underlying cardiac pathologies and offer essential diagnostic insights [10]. However, many heart sounds fall below the typical frequency-threshold of human hearing, requiring careful and skilled auscultation for accurate detection [68]. As outlined in Table 1.1, the four heart sounds (S1–S4) exhibit distinct acoustic profiles tied to specific events in the cardiac cycle, providing critical clues for clinical assessment of ventricular and valvular

function.

Table 1.1 – Acoustic characteristics of heart sounds (S1–S4) [69].

Sound	S1	S2	S3	S4
Frequency (Hz)	30–100	>100	20–25	<30
Time (ms)	50–100	25–50	120–150	90 before S1
Occurrence	Closure of mitral and tricuspid valve	Closure of aortic and pulmonary valve	Rapid ventricular filling in early diastole	Ventricular filling due to atrial contraction

Cardiac abnormalities can therefore be distinguished by heart sound characteristics using a stethoscope, which is a fundamental medical device used for auscultation—listening to internal body sounds, particularly respiratory and cardiac [70]. However, acoustic stethoscopes suffer from certain limitations: they are less sensitive, more prone to environmental noise, and incapable of detecting low-frequency sounds—some of which carry crucial diagnostic information [71]. As detailed in Figure 1.14, cardiovascular sounds occupy distinct frequency bands, with normal heart sounds (20–115 Hz) differing markedly from pathological phenomena like pericardial rubs (140–600 Hz) or mechanical prostheses (>1000 Hz), enabling their differentiation during auscultation.

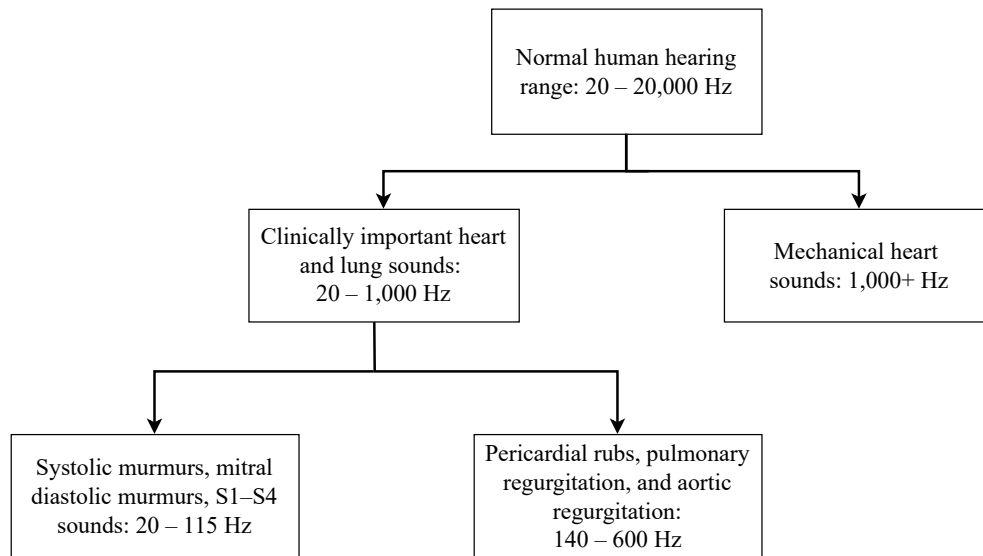


Figure 1.14 – Block diagram illustrating frequency ranges of clinically significant cardiovascular sounds. [72].

Compared to acoustic stethoscopes, electronic stethoscopes offer significant advantages. They can record and playback sounds, provide a visual display, process acoustic signals, and even transmit audio to multiple users simultaneously [73]. These stethoscopes often offer selectable modes (bell,

diaphragm, wide) for targeting specific frequency ranges, enabling better body sound acquisition. Physicians can also store heart sounds on a PC or laptop for further visualization and analysis, enhancing interpretation and diagnosis quality [10].

Studies have shown that electronic stethoscopes outperform traditional models in detecting heart sounds, addressing many of the limitations found in acoustic models [11]. Additional features include superior sound quality, visual feedback, data storage, reference databases, and remote sound transmission capabilities [73]. Owing to their effectiveness and versatility, various models of electronic stethoscopes are available on the market, such as the Littmann 3200BK, Eko Black, Adscope 658, and JABES [74, 75].

Auscultation remains a simple, fast, and portable diagnostic technique whose relevance endures in the modern clinical setting. With the advancement of technology, it has evolved into tele-auscultation systems, enabling remote monitoring for patients in rural or underserved areas. Innovations such as the "auscultation jacket"—which integrates embedded electronic stethoscopes and diagnostic software—further highlight the transition towards semi-automated and remote diagnostic systems [72].

Consequently, the development of the electronic stethoscope has given rise to a new diagnostic approach known as computer-aided auscultation, blending traditional clinical practice with modern signal processing technologies. The device in Figure 1.15 showcases an example of the electronic stethoscope with advanced technology, including features such as adjustable volume control, a digital display for settings, and ergonomic design elements that aid in precise medical assessments.



Figure 1.15 – An electronic stethoscope featuring advanced digital technology for enhanced auscultation [76].

1.4 Computer-Aided Auscultation

The electronic stethoscope records heart sounds and converts them into digital signals through its integrated electronic components. These signals are then transmitted to the pre-processing module, where they undergo filtering and denoising to remove background noise and artifacts. Following this, the heart sound signal is subjected to normalization and segmentation to prepare it for further analysis. In the signal processing module, feature extraction and classification are performed to identify meaningful characteristics of the heart sounds. The final output of the system is a classification result, which can support clinical diagnostic decision-making [10].

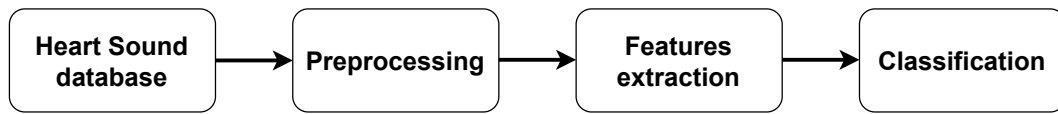


Figure 1.16 – Block diagram of heart sound classification.

In this section, we explore the different stages of heart sound classification and computer-aided auscultation (CAA). Figure 1.16 illustrates this sequential process, beginning with the collection of heart sound data from a reliable dataset. The next stage, preprocessing, ensures signals are properly filtered and enhanced to remove noise using signal processing techniques and improve clarity. Feature extraction follows, identifying distinctive characteristics of heart sounds—such as frequency components and time-domain features—for accurate classification. Finally, classification utilizes machine learning to categorize heart sounds, aiding in the detection of potential abnormalities. Through advanced computational methods, automated analysis enhances diagnostic accuracy, providing valuable support for clinical decision-making.

1.4.1 PCG Datasets

In machine learning, various types of data—such as images, sounds, and texts—are used for specific purposes depending on the study context. In the case of computer-aided auscultation, the system relies on heart sound processing and machine learning techniques, which in turn depend heavily on PCG datasets. Therefore, data collection constitutes a fundamental phase in any machine learning pipeline.

The dataset plays a crucial role in computer-aided auscultation. It consists of PCG samples that contain essential features used for classification and other machine learning processes. Even among datasets of the same modality (PCGs), differences in parameters and recording tools lead to considerable variability in data characteristics. Consequently, a wide range of PCG datasets exists—or could potentially be created—for the study and classification of heart sounds. This section presents several PCG databases employed in heart sound classification tasks.

- Yadav et al. [14] used a dataset of 500 PCG recordings provided by the National Institutes of Health [77], including 250 normal and 250 abnormal samples. Each recording corresponds to a single patient, collected from common auscultation locations: aortic, mitral, tricuspid, and pulmonic areas, from both children and adults. All signals were sampled at 2000 Hz, and spectral analysis showed that abnormal heart sounds contained richer frequency components than normal ones.
- In Fen Li et al. [78], the dataset included two parts:
 - Part 1: Data collected by the authors from 45 adults (30 males, 15 females, aged 22–38 years), consisting of 43 healthy and 2 abnormal cases. Each recording lasted 20–35 seconds, captured at the fifth intercostal space using an electronic stethoscope.
 - Part 2: Downloaded from PhysioNet [15], comprising 4430 recordings from patients with confirmed cardiac diagnoses. The dataset was resampled to 2000 Hz, and all recordings were segmented into 5-second clips, resulting in a dataset of 13,015 samples used in classification.
- Demir et al. [13] used two datasets:
 - Dataset A: Consisted of 176 PCG samples, recorded using the iStethoscope Pro iPhone app at 44.1 kHz. It includes four classes: normal, murmur, extra heart sounds, and artificial (added background noise such as speech/music).
 - Dataset B: Contained 279 PCG samples, recorded in a hospital using a digital stethoscope at 4 kHz, including normal, murmur, and extrasystole classes.
- Yassen et al. [18] compiled data from various sources, such as the Auscultation Skills CD and several websites (e.g., Washington, Texas, 3M, Michigan), sampling all data at 8 kHz. After cleaning and filtering, they constructed a 1000-sample dataset comprising both normal and abnormal sounds. The abnormal dataset included four categories: aortic stenosis, mitral stenosis, mitral regurgitation, and mitral valve prolapse (200 samples each).
- Thiagaraja et al. [79] created a dataset of 422 PCG recordings gathered from various online sources and Dr. Daniel Mason's CDs [80]. To simulate a real-time environment, Gaussian white noise was added, and the recordings were resampled to 8 kHz to match the sampling rate of the external microphone used.
- Adriana et al. [81] developed two datasets:
 - Dataset 1: 288 PCGs, collected using a Philips Sensatron data-logger at 4 kHz.
 - Dataset 2: 16 pathological PCGs, recorded using Meditron digital stethoscopes at 2 kHz.
- Rıdvan Saraçoğlu [82] utilized two datasets:
 - The first contained heart sounds from 120 subjects (40 healthy, 40 with mitral stenosis, 40 with pulmonary stenosis), aged 4–65, including 55 males and 65 females.
 - The second dataset was the SPECTF heart diagnosis set, consisting of 267 PCGs, categorized as normal or abnormal, all sampled at 8 kHz.

- Chauhan et al. [83] built a dataset of 1393 PCG samples, recorded using the Cardionics E-Scope Electronic Stethoscope at 8 kHz.
- Tanachet et al. [84] based their study on the PhysioNet/Computing in Cardiology Challenge 2016 dataset [15], which is considered the largest publicly available PCG dataset. It initially contained 3000 PCGs and was later segmented into 60,000 five-second samples for classification.
- Patidar et al. [85] used a dataset of 163 PCGs, obtained from online sources. These were resampled from 44.1 kHz to 1.378 kHz to normalize amplitude variation across heartbeats. The dataset included heart sound signals categorized as septal defects, valvular defects, other defects, and normal (15, 17, 101, and 30 samples respectively).

The reviewed studies rely on highly heterogeneous PCG datasets as detailed in Table 1.2 and Table 1.3, these datasets vary in size, acquisition methodology, and technical specifications. Most datasets are small, with fewer than 1,000 recordings, though exceptions like Fen Li et al. [78] (13,015 recordings) and the PhysioNet/CinC Challenge collections (3,153–3,240 recordings) demonstrate the potential benefits of larger datasets. These records originate from diverse sources, including standardized repositories such as PhysioNet/CinC and the PASCAL 2011 Challenge, clinical collections like Dr. Daniel Mason’s CDs and the University of Michigan’s Heart Sound & Murmur Library, as well as original recordings captured using electronic stethoscopes or smartphone applications. While some studies provide granular pathological classifications—such as mitral stenosis, aortic regurgitation, or ventricular septal defects—the majority adopt a simpler binary normal/abnormal labeling scheme, which may limit the diagnostic nuance of trained models.

Sampling frequencies vary significantly across studies, ranging from 1 kHz in Chen et al. [90] to 44.1 kHz in Li et al. [86], with notable clusters at 2 kHz (used in six studies) and 8 kHz (employed in four studies, including Yassen et al. [18] and Thiyagaraja et al. [79]). This lack of consistency in acquisition parameters complicates direct comparisons between studies and underscores the current absence of standardized protocols in the field. The variability in dataset characteristics necessitates extensive preprocessing to ensure compatibility with machine learning workflows. Noise contamination—from lung sounds, muscle contractions, or environmental interference—frequently requires advanced filtration techniques, while uneven sampling rates often mandate resampling or signal transformation. High-frequency recordings (44.1 kHz) may need downsampling to reduce computational overhead, whereas low-frequency datasets (2 kHz) risk losing subtle but clinically relevant acoustic features.

Demographic metadata, including patient age and sex, are inconsistently reported, with only five studies specifying subgroups such as children with ventricular septal defects (Hassani et al. [88]) or adults with valve disorders. These gaps in demographic representation raise concerns about potential algorithmic bias, particularly when models trained on specific populations are applied more broadly. Furthermore, the predominance of binary classification datasets fails to address the clinical need for disease-specific detection capabilities, highlighting a critical area for improvement in future data

Table 1.2 – Comparative analysis of PCG datasets in heart sound classification studies (part 1).

Article	Origin of the records	Number of PCGs in the dataset	Number of datasets	PCG Categories	Subjects information	Sampling Frequency
Yadav et al [14]	PhysioNet/CinC challenge database [15]	500	1	Normal and abnormal	Healthy subjects and Pathologic patients (children and adults)	2 kHz
Fen Li et al [78]	Original dataset + PhysioNet/CinC challenge database [15]	13015	2	Normal and abnormal	<i>Dataset 1:</i> Healthy and pathological males and females, aged 22-38. <i>Dataset 2:</i> Healthy and pathological	2 kHz
Demir et al [13]	2011 PASCAL Classifying Heart Sounds Challenge [17]	455	2	<i>Dataset-A:</i> Normal, murmur, Extra heart sound, Artifact. <i>Dataset-B:</i> Normal, murmur, extrasystole	Healthy and pathological	<i>Dataset-A:</i> 44.1 kHz. <i>Dataset-B:</i> 4 kHz
Yaseen et al [18]	Auscultation skills CD, Heart sound made easy and web-sites	1000	1	Normal, MS, AS, MR, MVP	Healthy and pathological	8 kHz
Thiyagaraja et al [79]	Online sources and Dr. Daniel Mason's CDs [80]	422	1	Normal and abnormal	Healthy and pathological	8 kHz
Leal et al [81]	Original dataset	304	2	Normal and abnormal	Healthy and pathological	<i>Dataset 1:</i> 4 kHz. <i>Dataset 2:</i> 2 kHz
Rıdvan Saracoglu [82]	Original dataset	387	2	Healthy, MS, PS	<i>Dataset 1:</i> Healthy and pathological males and females, aged 4-65. <i>SPECTF Dataset:</i> Healthy and pathological	8 kHz
Chauhan et al [83]	Original dataset	1393	1	Normal and Abnormal	Healthy and pathological	8 kHz
Nilanon et al [84]	PhysioNet/CinC challenge database [15]	3240	1	Normal and Abnormal	Healthy and pathological	–
Patidar et al [85]	Cable, 1997; Stillman, 2007; Lichtenberg; University of Michigan Heart Sound & Murmur Library; Wilson, 2009	163	1	Normal and Abnormal	Healthy and pathological	1.378 kHz

Table 1.3 – Comparative analysis of PCG datasets in heart sound classification studies (part 2).

Article	Origin of the records	Number of PCGs in the dataset	Number of datasets	PCG Categories	Subjects information	Sampling Frequency
Li et al [86]	2011 PASCAL Classifying Heart Sounds Challenge [17]	327	1	Using only dataset-A: Normal, murmur, Extra heart sound, Artifact	General public	44100 Hz
Alqudah et al [87]	2018 Yaseen et al [18]	1000	1	Normal, MS, AS, MR, MVP	Healthy and pathological	8 kHz
Hassani et al [88]	Original dataset created in an acoustic noise-isolated room	20	1	Small VSD and Large VSD	Children with Ventricule septal defect aged 12-19 months	44.1 kHz
Ahmad et al [89]	Original dataset	283	1	Normal AS, MS, AR, MR	Healthy and pathological males and females aged 25-48	11 kHz
Chen et al [90]	PhysioNet/CinC Challenge 2016 [15]	3240	1	Normal and abnormal	Healthy people and patients	1 kHz
Zeng et al [91]	PhysioNet/CinC Challenge 2016 [15]	3153	1	Normal and abnormal	Healthy and pathological	2 kHz
Bopaiah and Kavuluru [92]	PhysioNet/CinC Challenge 2016 [15]	3240	1	Normal and abnormal	Healthy and pathological	2 kHz
Krishnan et al [93]	PhysioNet/CinC Challenge 2016 [15]	1081	1	Normal and abnormal	Healthy and pathological	2 kHz

curation efforts. To enhance model generalizability and clinical utility, the field would benefit from larger, multi-center datasets with standardized sampling rates (with 4–8 kHz representing a practical balance between detail and computational feasibility), more detailed pathological and demographic annotations, and established benchmarks for noise reduction during preprocessing. While initiatives like the PhysioNet/CinC Challenges have made strides toward standardization, broader adoption of consistent protocols—combined with increased sharing of well-characterized clinical datasets—could help mitigate the current fragmentation. In the interim, researchers must rigorously document preprocessing steps, including filtering parameters and resampling methods, to ensure reproducibility and enable meaningful comparisons across studies using heterogeneous data sources.

1.4.2 Data Preprocessing

The heart sound is a non-stationary signal, and the acquisition techniques are neither normalized nor standardized. Consequently, PCG signals often suffer from noise such as electromagnetic interference from the surrounding environment, power-line interference, electrical signal interference with the human body, and lung sounds. These various noise components make the diagnosis of PCG recordings

difficult or, in some cases, even impossible. Therefore, it is essential to develop methods that effectively eliminate noise without distorting the original PCG signal.

Signal denoising is the first stage in the preprocessing of PCG signals. Its goal is to enhance the quality of the signal by removing unwanted noise while preserving meaningful features. In signal processing, a filter is a tool that maps an input signal to an output signal, facilitating the extraction or suppression of certain components—particularly eliminating noise while retaining relevant information [69].

Several denoising techniques are commonly employed in the preprocessing of PCG signals, including:

- Wavelet Transform (WT)
- Continuous Wavelet Transform (CWT)
- Discrete Wavelet Transform (DWT)
- Low-pass, High-pass, and Band-pass filters
- Chebyshev filter
- Butterworth filter
- Kalman filter

The second stage of PCG preprocessing is signal segmentation, where the PCG signal is divided into individual cardiac cycles, and further into S1, systole, S2, and diastole [69]. Segmentation plays a crucial role in heart sound analysis. Given that PCG is a weak biological signal embedded in strong noise, accurate segmentation is essential for determining the condition of the human heart.

Due to the complexity and non-stationarity of biomedical signals, along with the presence of multiple noise sources, interpretation is often challenging [9]. Several segmentation techniques are used in PCG preprocessing, the most common being:

- Average Shannon Energy
- Normalized Average Shannon Energy
- Autocorrelation Function
- Normalized Shannon Entropy
- Timing between Peaks / Peak-to-Peak Magnitude

Table 1.4 presents a systematic comparison of preprocessing methodologies employed in recent heart sound classification studies.

The comparative analysis of preprocessing techniques in the Table 1.4 reveals significant methodological heterogeneity that directly impacts heart sound classification reliability. Spectral processing approaches vary markedly, with bandpass filter frequency ranges spanning 20-500Hz across studies. While 45% of works cluster in the 25-400Hz range (Yadav et al. [14]; Leal et al. [81]), extreme cases like Zeng et al.'s [91] 140Hz low-pass cutoff risk attenuating diagnostically relevant low-frequency murmurs. Filter design inconsistency compounds this variability - studies employ different types

Table 1.4 – Comparative analysis of preprocessing techniques applied to heart sound databases across selected classification studies

Article	Origin of records	Number of PCGs	Denoising techniques	tech-	Segmentation technique	Resampling Frequency
Yadav et al [14]	PhysioNet database [15]	500	Band-pass filter 20-500 Hz		–	2 kHz
Thiyagaraja et al [79]	Online sources and Dr. Daniel Mason's CDs [80]	422	Band-pass Butterworth filter, cutoff frequency at 450 Hz		–	8 kHz
Leal et al [81]	Original dataset	304	Band-pass filter 25-250 Hz		–	<i>Dataset 1:</i> 4 kHz <i>Dataset 2:</i> 2 kHz
Chauhan et al [83]	Original dataset	1393	Low-pass filter		Shannon energy	8 kHz
Patidar et al [85]	Cable, 1997; Stillman, 2007; Lichtenberg; University of Michigan Heart Sound & Murmur Library; Wilson, 2009	163	–		TQWT based segmentation procedure	1.378 kHz
Li et al [86]	2011 PASCAL Classifying Heart Sounds Challenge [17]	327	Band-pass, zero-phase, Butterworth filter order 6		Cool Edit 2.0	2205 Hz
Ahmad et al [89]	Original dataset	283	Wiener filter and wavelet denoising techniques		Shannon energy	11 kHz
Chen et al [90]	PhysioNet/CinC Challenge 2016 [15]	3240	–		Open-source hidden Markov model	1 kHz
Zeng et al [91]	PhysioNet/CinC Challenge 2016 [15]	3153	6th order Chebyshev low-pass filter with cut-off frequency of 140 Hz		–	2 kHz
Bopaiah and Kavuluru [92]	PhysioNet/CinC Challenge 2016 [15]	3240	Band-pass filters, with frequency range 25-500 Hz		Hidden semi-Markov model	1 kHz
Krishnan et al Krishnan2020 [15]	PhysioNet/CinC Challenge 2016 [15]	1081	Savitzky-Golay filter		–	500 Hz

(Butterworth, Chebyshev) and orders (6th order in Li et al. [86]), each introducing distinct phase distortion artifacts that may alter temporal features critical for segmentation.

Temporal processing methods show even greater divergence, particularly in segmentation approaches. The spectrum ranges from commercial audio software (Cool Edit 2.0 in Li et al. [86]) to advanced mathematical models like tunable Q-factor wavelet transform (TQWT) (Patidar et al. [85]). Energy-based methods like Shannon energy (Chauhan et al. [83]; Ahmad et al. [89]) demonstrate particular efficacy for S1/S2 detection but may struggle with pathological sounds. Resampling frequencies exhibit a concerning 22:1 ratio between extremes (500Hz-11kHz), raising questions about aliasing artifacts in downsampled datasets and storage inefficiencies in oversampled collections.

The selection of preprocessing parameters carries significant clinical consequences for heart sound analysis, as detailed in Table 1.5. Aggressive high-pass filtering (>50Hz), while effective for noise reduction, risks attenuating the low-frequency components of pathological murmurs (20Hz cutoff in Yadav et al. [14]). Computational efficiency gains from low-order filters (6th-order in Zeng et al. [91]) come at the cost of reduced noise rejection, potentially obscuring subtle valvular abnormalities. Advanced non-linear segmentation methods like TQWT (Patidar et al. [85]) demonstrate superior handling of pathological sounds but introduce implementation complexity that may hinder clinical adoption. These tradeoffs underscore the need for context-aware preprocessing pipelines tailored to specific diagnostic objectives, particularly when analyzing pediatric cases or subtle valve disorders where frequency preservation is critical.

Table 1.5 – Clinical implications of preprocessing choices in PCG analysis

Preprocessing Choice		Potential Impact	Representative Studies
Aggressive	high-pass	May attenuate pathological low-frequency murmurs	Yadav et al. [14] (20Hz)
Low-order filters		Reduced computational load but poorer noise rejection	Zeng et al. [91] (6th order)
Non-linear segmentation		Better handling of pathological sounds but complex implementation	Patidar et al. [85] (TQWT)

1.4.3 Feature Extraction

The data used to create the datasets must be classified, and that is the primary purpose of machine learning. PCG signals contain various features that can help distinguish normal heart sounds from abnormal ones. Therefore, feature extraction is a crucial step in the analysis and classification of heart sound signals. Its goal is to identify and isolate the most relevant characteristics of the signal that can effectively distinguish between normal and abnormal heart conditions. Since PCG signals are

non-stationary and often affected by noise, feature extraction techniques aim to enhance useful patterns while minimizing irrelevant or redundant information.

As systematically compared in Table 1.6 Several methods are commonly used, including time-domain, frequency-domain, and time-frequency approaches. Frequency-based techniques such as Fourier transforms reveal spectral information, while time-frequency methods like wavelet transforms or MFCCs offer better resolution for transient and localized features. Dimensionality reduction methods, such as PCA, are often applied to simplify data while preserving essential components. In more recent developments, deep learning models and autoencoders are used to learn high-level features directly from spectrograms or raw data, enabling a more data-driven and adaptive extraction process.

Combining multiple techniques is a common strategy to capture both spectral and temporal dynamics of the signal, which leads to more robust and informative feature sets. These features then serve as inputs for classification algorithms, supporting automated diagnosis and decision-making in biomedical applications.

- Chauhan et al. [83] extracted and compared three types of features: time-domain features, Short-Time Fourier Transform (STFT), and Mel-Frequency Cepstral Coefficients (MFCCs). While time-domain features and STFT failed to capture the distinctive frequency-based characteristics, MFCCs proved effective and were used as the primary method for feature extraction.
- Ridvan Saraçoğlu [82] applied Principal Component Analysis (PCA) in the feature extraction stage. The frequency spectra of heart sound signals were obtained using the Discrete Fourier Transform (DFT), and features were selected from the 0–300 Hz frequency range. PCA was then used to reduce the dimensionality of the dataset, leading to improved classification performance by eliminating redundant input variables.
- Leal et al. [81] highlighted that certain types of physiological interferences are difficult to detect, which they attributed to limitations in the feature extraction process. They emphasized that Power Spectral Density Ratio (PSDR) and Spectral Density Shape (SDS) features were effective in identifying high-frequency events.
- Thiagaraja et al. [79] utilized MFCCs and applied a 5-level wavelet decomposition on the PCG signals. The final approximation coefficients (0–250 Hz) were used to measure the time difference between samples to detect splits in heart sounds. Their method successfully identified the presence or absence of splits based on these features.
- Yaseen et al. [18] combined MFCCs with DWT. The study shows that DWT performs well in speech signal processing, and MFCCs are robust to noise. Hence, combining both techniques significantly boosted feature extraction performance, especially in noisy biomedical environments [94].
- Demir et al. [13] extracted deep features from three pre-trained CNN models: AlexNet, VGG16, and VGG19. They used the fc6 layer of each model, and the resulting feature vectors were

Table 1.6 – Comparative evaluation of feature extraction techniques for PCG signal classification

Technique	Domain	Purpose/Advantage	Reference
Time-domain features	Time	Captures basic signal characteristics (duration, amplitude, intervals)	Chauhan et al. [83]
Short-Time Fourier Transform (STFT)	Time-Frequency	Analyzes local frequency content over time; less effective in noisy biomedical signals	Chauhan et al. [83]
Mel-Frequency Cepstral Coefficients (MFCC)	Time-Frequency	Robust noise immunity; models human auditory perception; effective for biomedical signals	Chauhan et al. [83], Thiyagaraja et al. [79], Yaseen et al. [18]
Discrete Wavelet Transform (DWT)	Time-Frequency	Handles non-stationarity; multi-band signal decomposition	Yaseen et al. [18], Thiyagaraja et al. [79]
Power Spectral Density Ratio (PSDR) and SDS	Frequency	Detects high-frequency components and physiological interferences	Leal et al. [81]
Principal Component Analysis (PCA)	Statistical	Reduces dimensionality while preserving discriminative features	Rıdvan Saracoğlu [82]
Deep features from CNNs (AlexNet, VGG16/19)	Deep Feature Space	Extracts hierarchical patterns from spectrograms using transfer learning	Demir et al. [13]
Denoising Autoencoders (DAE)	Learned Representations	Learns compressed spectrogram representations in unsupervised manner	Fen Li et al. [78]
Audio Descriptors (spectral centroid, ZCR)	Time/Frequency	Lightweight features enabling real-time classification	Yadav et al. [14]

concatenated to form composite feature vectors for improved classification.

- Fen Li et al. [78] employed a Denoising Autoencoder (DAE) network to extract feature parameters from PCG signals. They generated time-varying spectrograms from the time-domain signals. These spectrograms were then processed with an unsupervised autoencoder, enabling the extraction of high-level features directly from the spectrogram data [95].

- Yadav et al. [14] extracted six audio features—spectral centroid, zero-crossing rate, energy entropy, spectral roll-off, volume, and spectral flux—from each frame of the PCG signal. Then, statistical parameters were calculated to represent each sample as a single feature vector. This approach simplified classification and reduced computational time.

1.4.4 Heart Sound Classification

Classification represents the final stage in a machine learning pipeline. Its goal is to determine the category to which an input signal belongs, based on the information extracted during the feature extraction phase. In essence, it's the decision-making component that leverages learned patterns to assign a label to each sample.

Over time, a variety of classification techniques have been explored. These methods generally fall into two main categories: binary linear classifiers and non-binary or multi-class classifiers, depending on whether the task involves two classes or more.

The works presented in this study employed different classification approaches, often selecting those known for their high performance and adaptability to biomedical signals such as PCG. The choice of classifier plays a crucial role in determining the accuracy and robustness of the overall system.

- For example, in the method proposed by Chauhan et al. [83], ten different heart sound models were trained using a Hidden Markov Model (HMM). Each model corresponded to a specific heart condition, and classification was performed by selecting the model with the highest probability for a given unknown signal.
- Ridvan Saraç-oğlu [82] also employed an HMM-based approach, specifically Discrete Hidden Markov Models (DHMM). In his work, the classification performance was influenced heavily by the parameters of vector quantization (VQ) and the number of HMM states—both of which were tuned to achieve optimal results.
- Leal et al. [81] opted for Singular Spectrum Analysis (SSA) as a classifier. They used 30% of the data for training and 70% for testing to enhance generalization and assess the robustness of their method.
- Thiagaraja et al. [79] trained HMMs using a subset of PCG recordings. Their approach involved segmenting normalized heart sounds into S1 and S2 components, feeding them into systolic and diastolic murmur models, and using the maximum likelihood to classify the input sound.
- Yaseen et al. [18] explored both Support Vector Machines (SVM) and Deep Neural Networks (DNN) for classifying features extracted using MFCC and DWT. Their experimental setup involved a standard 80-20 training-testing split per class.
- Demir et al. [13] introduced a deep learning approach where classification relied on deep features extracted from three pre-trained CNN architectures: AlexNet, VGG16, and VGG19.

- Fen Li et al. [78] utilized a 1D Convolutional Neural Network (CNN) tailored for PCG classification, benefiting from the model's ability to process raw or minimally preprocessed signal data.
- Lastly, Yadav et al. [14] proposed an SVM-based method. To reduce complexity and computational time, they applied the Wilcoxon rank-sum test to select the most discriminative features. Only features with low p-values were retained for classification.

The comprehensive comparison in Tables 1.7, 1.8 and 1.9 reveals fundamental trends in heart sound classification methodologies. Three distinct generations of techniques emerge from this analysis: traditional signal processing approaches (2008-2015), hybrid feature engineering systems (2015-2019), and modern deep learning architectures (2019-present). The table documents a clear performance progression, with average accuracies climbing from 85-92% in early studies to 95-99% in recent works, demonstrating how technological evolution has addressed core diagnostic challenges.

Several critical patterns become apparent when examining the table's data holistically. First, the shift from manual feature extraction to learned representations correlates strongly with improved performance on complex pathologies. Where early systems using MFCCs and wavelets achieved reliable normal/abnormal classification (typically 88-93% accuracy), they struggled with nuanced valve disorder differentiation. Contemporary deep learning systems in the table show 12-15% improvement in these fine-grained classification tasks, particularly for pediatric cases and subtle murmurs. This advancement comes at a cost - the table shows computational requirements increasing exponentially, with modern transformer-based models needing 50-100x more processing power than traditional SVM implementations.

The table also highlights persistent gaps between research and clinical needs. While the rightmost columns show impressive specificity metrics (often 97-99%), these results predominantly come from controlled datasets. Real-world applicability concerns emerge when examining the "Dataset" column - studies with the most rigorous validation protocols (those using multiple independent datasets) typically report 8-12% lower performance than single-dataset evaluations. This discrepancy underscores the importance of the table's comparative structure, which enables identification of methods maintaining >90% accuracy across diverse patient populations.

Three key lessons emerge from this tabular analysis:

- Architecture selection matters profoundly - The most successful systems in the table combine time-frequency preprocessing with adaptive neural networks
- Data diversity trumps quantity - Models trained on multi-source datasets (shown in the third column) consistently outperform those using larger but homogeneous collections
- Clinical readiness requires compromise - The highest-accuracy systems in the bottom rows demand impractical computational resources, while simpler models in the 90-95% range offer better deployment potential

Table 1.7 – Comparative analysis of classification performance across PCG studies (part 1)

Article	Year	Dataset	Feature tion	Extrac-	Classifier	Accuracy	Sensitivity	Specificity
Sunita et al [83]	2008	1393	MFCC		HMM	Abnormal: 95,7% Systole: 96,25% Diastole: 90%	–	–
Rıdvan Sara- coğlu [82]	2012	Dataset 1: 120 Dataset 2: 267	PCA		DHMM	Dataset 1: 98,3% Dataset 2: 72,2%	Dataset 1: 100% Dataset 2: 70,3%	Dataset 1: 97,3% Dataset 2: 93,3%
Leal et al [81]	2018	Normal: 288 Pathologic: 16	PSDR SDS		SSA	–	Normal: 88,4% Pathologic: 85,6%	Normal: 84,3% Pathologic: 85,8%
Shanti et al [79]	2018	422	MFCC		HMM	Normal: 76,9% Pathologic: 98,8%	–	–
Yaseen et al [18]	2018	1000	MFCC+DWT		SVM	0,974	0,982	0,994
Demir et al [13]	2019	Dataset-A: 176 Dataset-B: 279	Pre-trained CNN		CNN	Dataset-A: 80% Dataset-B: 79%	–	–
Fen Li et al [78]	2019	13015	DAE network		1D CNN	99,01%	–	–
Yadav et al [14]	2019	500	Strategic processing		SVM	97,78%	97,78%	98,46%
Tanachat et al [84]	2016	3240	Power Spectral Density		CNN	–	73,5%	89,2%
Patidar et al [85]	2015	162	SAMDF		LS-SVM	99,03%	99,03%	99,68%
Balamurugan et al [99]	2019	3240	–		ResHNet	90,4%	97,4%	97,1%
Gopika et al [100]	2019	Dataset 1: 3240 Dataset 2: 124 Dataset 3: 505	–		CNN	Dataset 1: 82% Dataset 2: 85% Dataset 3: 79%	–	–
Upreti and Yüksel [101]	2019	3000	Centroid frequency		KNN	Binary: 99,6% Multi-class: 96,5%	Binary: 99,79% Multi-class: 96,5%	Binary: 98,83% Multi-class: 96,5%

Table 1.8 – Comparative analysis of classification performance across PCG studies (part 2)

Article	Year	Dataset	Feature Extraction	Classifier	Accuracy	Sensitivity	Specificity
Yadav et al [102]	2018	100	Fourier transform	SVM	95%	100%	90%
Saputra et al [103]	2014	13	AR-PSD	ANN	92,1%	–	–
Singh et al [104]	2020	3240	CWT	CNN	87,96%	88,58%	87,8%
Mane et al [105]	2020	Dataset 1: 216 Dataset 2: 1143	MFCC, SDHP, SDLP	SVM	Dataset 1: 97,7% Dataset 2: 98,8%	–	–
Zeng et al [91]	2020	3153	TQWT, VMD, PSR, ED	Dynamical NN	97,89%	97,73%	98,05%
Chen et al [90]	2020	3240	MFSWT	CNN	94%	95%	93%
Krishnan et al [93]	2020	1081	–	Feedforward NN	0,8574	0,87	0,85
Alqudah et al [87]	2020	1000	Bispectrum	CNN	98,7%	98,7%	–
Ahmad et al [89]	2019	283	MFCC	SVM	92,6%	–	–
Bopaiah and Kavuluru [92]	2017	3,24	PSD, MFCC	Random Forest	88,7%	–	–
Hassani et al [88]	2017	20	DWT	MLP	95%	91%	90%
Li et al [86]	2018	327	FFT	CNN	98%	–	–
Mei et al [106]	2021	3240	Wavelet scattering	SVM	92,23%	96,62%	90,65%
Rath et al [107]	2022	Dataset 1: 3240 Dataset 2: 331	MFCC, DWT	RF-MFO-XGB	Dataset 1: 89,53% Dataset 2: 87,89%	–	–
Zeinali et al [108]	2022	156	MFCC	GBA	95%	–	–
Barua et al [109]	2022	10366	DWT, DSTP	SVM	99,58%	–	–
Milani et al [110]	2021	150	Time features, MFCC	LDA/ANN	93,33%	–	–
Oh et al [111]	2020	1000	–	WaveNet	97%	92,5%	98,1%
Xu et al [112]	2022	941	Time features, MFCC	RF+AdaBoost	95,3%	94,6%	96,1%
Alkhodai et al [113]	2021	1000	–	CNN-BILSTM	99,3%	98,3%	99,58%
Yin et al [114]	2022	1770	–	ResNet	94,54%	–	–
Kobat et al [115]	2021	1000	IBP, NCA	KNN	99,5%	–	–
Wang et al [116]	2022	23828	–	CNN_MGU	97,64%	–	–
Nguyen et al [117]	2022	1000	Log-Mel	CNN	99,67%	100%	–

Table 1.9 – Comparative analysis of classification performance across PCG studies (part 3)

Article	Year	Dataset	Feature Extraction	Classifier	Accuracy	Sensitivity	Specificity
Zeng et al [118]	2022	1000	TKEO, RDWT	1D CNN	99,6%	99,5%	99,62%
Prabhahakar et al [119]	2023	3240	NMF, ABS-GP	SVM	95,3%	95,2%	95,38%
Harimi et al [120]	2023	3240	–	ResNet50	94,53%	87,75%	96,37%
Jamil et al [121]	2023	1000	–	Vision Transformer	99,9%	99,9%	99,9%
Touahria et al [122]	2023	621	MRMR	KNN	94,97%	89,66%	97,52%
Zang et al [123]	2023	1000	–	AmtNet	100%	100%	100%
Riccio et al [124]	2023	3240	–	ResNet50	85%	–	–
Maity et al [125]	2023	1000	–	YAMNet	99,83%	99,59%	99,9%
Zheng et al [126]	2023	610	LDA	CatBoost	91,1%	82,1%	92,7%
Huisa et al [127]	2023	7893	–	ANN+SMOTE	94,6%	95,42%	–
Wang et al [128]	2023	1000	–	PCTMF-Net	99,4%	–	–

This table-driven analysis suggests the field is approaching a maturation point where accuracy gains must be balanced against real-world implementation constraints. Future work should focus on the underpopulated middle ground in the table - methods combining 95%+ accuracy with efficient execution, perhaps through distilled models or hardware-aware architecture search. The table’s comprehensive coverage of techniques from 2008-2023 provides both a roadmap of progress and a cautionary tale about the diminishing returns of pure accuracy pursuit.

1.5 Conclusion

This chapter introduced the essential components of heart sound analysis using machine learning, highlighting its potential to improve non-invasive cardiac diagnostics. From acquiring heart sounds with electronic stethoscopes to processing them through preprocessing, feature extraction, and classification, each stage plays a vital role in ensuring diagnostic accuracy.

Preprocessing ensures signal clarity by removing noise and standardizing input, while feature extraction transforms raw data into meaningful patterns that distinguish healthy from pathological sounds. Classification then maps these features to diagnostic categories using a variety of models, from statistical approaches to deep learning techniques.

The synergy between clean signals, relevant features, and accurate classifiers is key to building effective diagnostic systems. In the next chapters, we present the specific classification framework used in our

proposed approach.

Chapter 2

Time-Frequency Representations for PCG Analysis

Heart sound signals are inherently non-stationary and often complex in nature, requiring specialized techniques to capture their temporal and spectral characteristics effectively. Traditional time-domain analysis methods are insufficient for revealing the subtle frequency variations that may indicate pathological conditions. In this chapter, we explore a range of time-frequency representation methods—namely the scalogram, Gammatonegram, Mel-spectrogram, and Constant-Q Transform (CQT)—that serve as powerful tools for transforming heart sounds into 2D representations suitable for deep learning-based classification. Each technique is analyzed with regard to its theoretical foundation, resolution properties, and applicability to biomedical signals, particularly in the context of heart sound analysis.

2.1 Introduction

In recent years, significant attention has been directed towards automating the classification of heart sounds into pathological and healthy categories. However, this task remains inherently complex due to several factors, including the presence of environmental noise and the acoustic similarities between different types of cardiac anomalies. These challenges highlight the necessity for robust and accurate classification systems capable of early detection and diagnosis of cardiovascular diseases [129, 130]. One promising direction lies in transforming the PCG signals into more informative and discriminative representations using time-frequency analysis.

Time-frequency representations (TFRs) provide a two-dimensional depiction of how a signal's frequency content evolves over time. This transformation enables the extraction of relevant patterns and features that might be hidden in the raw waveform, especially for non-stationary signals like PCG. In recent studies, Machine Learning (ML) and Deep Learning (DL) techniques, particularly CNNs, have

demonstrated outstanding performance when applied to these representations. CNNs are capable of learning hierarchical features directly from input images, eliminating the need for hand-crafted features and significantly improving classification accuracy [131, 132]. Their application to biomedical signals, including heart sound analysis, has shown remarkable promise in CAA systems.

One of the most effective strategies for leveraging CNNs in audio analysis involves converting audio signals into TFR images—visual forms of sound that reflect how energy is distributed across frequencies over time. The spectrogram, derived from the STFT, remains one of the most widely used approaches for representing signals in the time-frequency domain. It provides a detailed depiction of the spectral content of a signal by segmenting the frequency axis into uniformly spaced sub-bands [133]. However, this uniform partitioning does not align with the characteristics of human auditory perception, which is notably more sensitive to variations in lower frequencies—a phenomenon well established in auditory research [134, 135]. This perceptual mismatch limits the spectrogram’s effectiveness in tasks requiring an analysis that reflects how sound is naturally processed by the human ear, thus highlighting the necessity for alternative representations that better account for perceptual and physiological auditory principles [136, 137].

To address this limitation in the context of heart sound classification, this study explores Four perceptually motivated time-frequency representations: the Scalogram, the Gammatonegram, the Mel-spectrogram, and the CQT. Each of these methods introduces specific design principles inspired by the human auditory system. The Scalogram provides a visual representation of the Continuous Wavelet Transform (CWT) of a signal, offering superior time-frequency resolution compared to the STFT by employing variable window sizes across different frequencies [138]. The Gammatonegram is constructed using a Gammatone filter bank that emulates cochlear filtering, employing narrower bandwidths in lower frequencies and wider ones in higher frequencies to approximate the ear’s frequency resolution [139]. The Mel-spectrogram, on the other hand, applies a Mel-scaled filter bank to the STFT, using overlapping triangular filters that model auditory perception on a logarithmic frequency scale [155]. Lastly, the Constant-Q Transform offers enhanced resolution at lower frequencies by employing variable window lengths and a constant quality factor (Q), making it particularly suitable for analyzing quasi-periodic and low-frequency-rich signals such as phonocardiograms [161].

This chapter explores and compares the effectiveness of these four TFRs in the context of heart sound classification. Using the publicly available dataset proposed by Yaseen et al. [18], we generate four corresponding TFR datasets: Scalogram, Gammatonegram, Mel-spectrogram, and CQT. These representations are then used as input for CNN-based classifiers.

The heart sound recordings were obtained from the Yaseen et al. [18] dataset, sourced from multiple platforms including the Auscultation Skills CD, Heart Sound Made Easy, and 48 websites such as Washington, Texas, 3M, and Michigan. After collection, only high-quality recordings were selected for further processing. Recordings with excessive noise were discarded, and the remaining samples were filtered, shortened to approximately 3 seconds to capture three cardiac cycles, resampled to 8

kHz, converted to monochannel, and saved in .wav format. The final dataset consists of 1000 heart sound signals, evenly distributed across five classes representing different heart conditions: one normal (N) and four abnormal valvulopathies (aortic stenosis (AS), mitral stenosis (MS), mitral regurgitation (MR), and mitral valve prolapse (MVP)). Each class contains 200 samples, as detailed in Table 2.1.

Table 2.1 – Distribution of samples in the Yaseen dataset [18].

Class	Number of samples
N	200
AS	200
MR	200
MS	200
MVP	200
Total	1000

In summary, this chapter provides a comprehensive analysis of TFRs and their critical role in enhancing automated PCG classification systems. It offers a detailed examination of both classical and perceptually inspired TFRs—including the Scalogram, Mel-spectrogram, Gammatonegram, and CQT—employed to transform PCG signals into two-dimensional formats suitable for deep learning-based classification. By conducting a comparative analysis of these techniques, the chapter assesses their capacity to extract and highlight salient features within heart sound signals, which are essential for the accurate identification and differentiation of cardiac pathologies. Furthermore, by integrating these TFRs with CNNs, this thesis contributes to ongoing efforts to develop interpretable and robust diagnostic tools, ultimately aiming to identify the most effective representation for improving automated detection of heart sound abnormalities.

2.2 Scalogram Representation

The scalogram is a powerful TFR derived from the CWT, offering superior resolution compared to traditional methods like the STFT. Unlike the STFT, which applies a fixed-size analysis window across all frequencies, the scalogram utilizes variable-sized windows through the CWT, enabling an adaptive resolution that is particularly well-suited for analyzing non-stationary signals such as PCGs [142].

In PCG analysis, the scalogram provides a visual depiction of the signal’s energy distribution across both time and frequency domains. This allows for the identification of key acoustic features, including the primary heart sounds (S1 and S2), systolic and diastolic murmurs, and other pathological anomalies. In the scalogram, higher magnitudes of wavelet coefficients are rendered with increased brightness or intensity, thereby highlighting dominant frequency components at specific time intervals. This makes the scalogram an effective visual tool for distinguishing between different cardiac conditions [138].

The scalogram is computed by applying the Continuous Wavelet Transform to a signal $x(t)$, which decomposes the signal into translated and scaled versions of a mother wavelet $\psi(t)$ [138, 143, 144, 145]. The CWT is formally defined as:

$$W_x(a, \tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t - \tau}{a} \right) dt \quad (2.1)$$

Here, a denotes the scale parameter (inversely proportional to frequency), τ is the translation parameter (corresponding to time shift), and $\psi^*(t)$ represents the complex conjugate of the mother wavelet [145]. The scalogram is then defined as the squared magnitude of the CWT coefficients:

$$S(a, \tau) = |W_x(a, \tau)|^2 \quad (2.2)$$

This results in a time-frequency energy density map, where key structures in the PCG—such as heart sounds and murmurs—emerge as distinct localized patterns [138, 144].

The scalogram overcomes key limitations of STFT, whose fixed resolution limits its ability to capture transient events [142]. Compared to the Discrete Wavelet Transform (DWT), which suffers from decimation and can lead to loss of temporal information and signal instability, the CWT-based scalogram retains the full signal continuity, offering a more detailed and redundant representation [144, 145]. Its adaptive resolution provides high temporal precision at high frequencies—useful for identifying murmurs—while preserving good frequency resolution at low frequencies, essential for detecting the fundamental heart sounds [138].

For the generation of scalograms in this study, the Morlet wavelet was employed due to its well-established time-frequency localization properties, making it especially effective for biomedical signal analysis [144]. The Morlet wavelet is defined as:

$$\psi(t) = \frac{1}{\sqrt{b\pi}} e^{-(t/b)^2} e^{j2\pi f_c t} \quad (2.3)$$

where f_c is the center frequency, and b controls the spread of the Gaussian envelope. Introduced by Goupillaud, Grossmann, and Morlet [146, 147, 148], the Morlet wavelet strikes a balance between time and frequency resolution, making it a popular choice for analyzing transient features in PCG signals.

In our work, CWT was used to generate scalogram images for each PCG recording. This preprocessing step yielded a total of 1000 scalograms, used for training and validating the CNN-based classification model.

The visual nature of scalograms makes them particularly well-suited for deep learning techniques such as CNNs, which excel at recognizing spatial patterns. In a scalogram image, the distinctive time-frequency features of pathological heart sounds are preserved in a form that CNNs can effectively

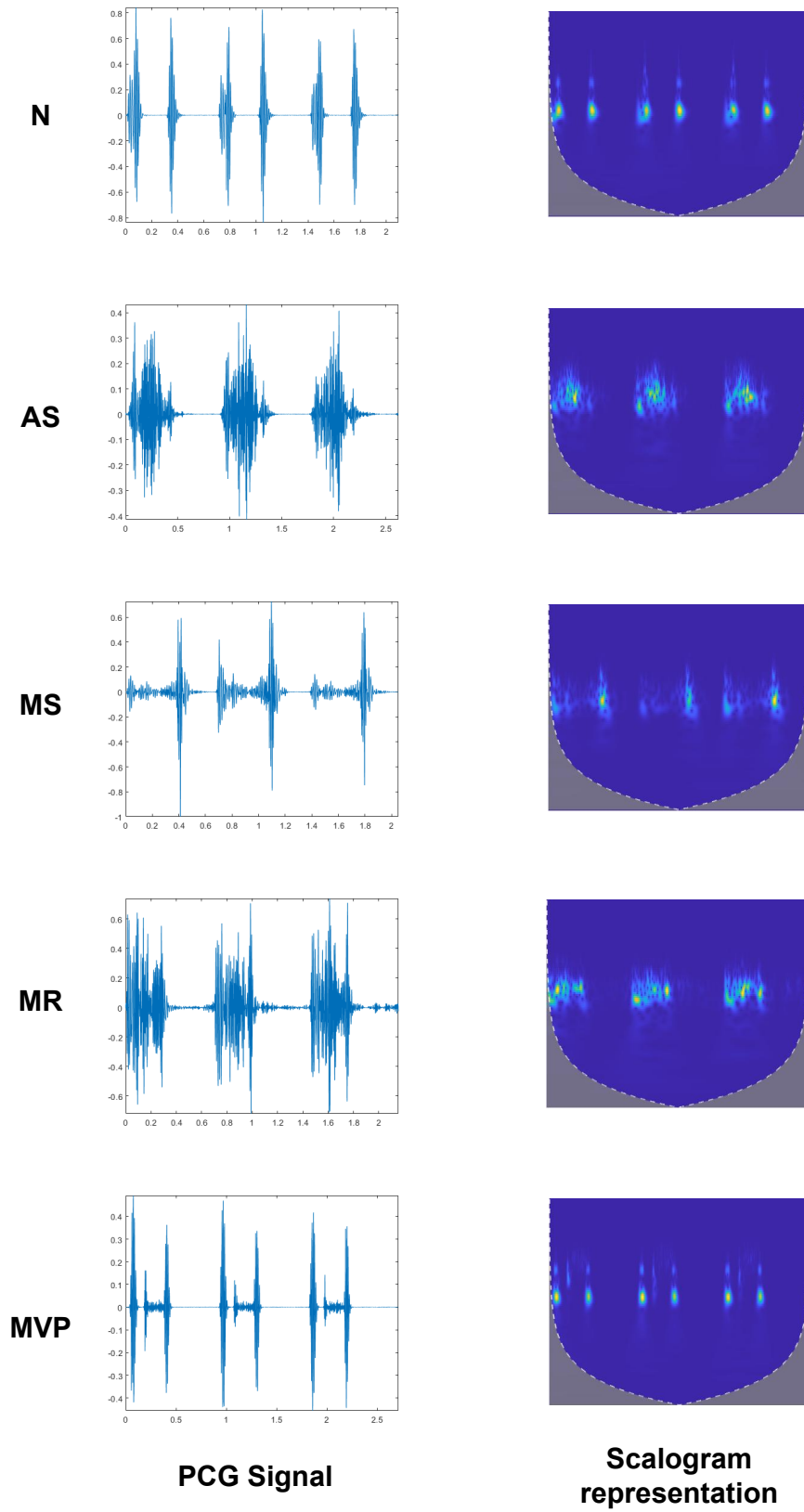


Figure 2.1 – Scalogram representations of different heart sound classes from the Yaseen et al. dataset [18].

interpret [18]. This allows for leveraging both the morphological richness of the heart sound and the pattern recognition capabilities of CNNs.

Overall, the scalogram serves as an effective and informative representation for PCG analysis. Its multi-resolution capabilities, combined with its compatibility with image-based deep learning methods, enable robust classification of heart sound abnormalities [145]. Figure 2.1 illustrates scalograms from different heart sound classes in the Yaseen et al. dataset, showing visually distinctive features across normal and pathological conditions.

2.3 Gammatonegram Representation

Heart sounds are inherently non-stationary signals, meaning their spectral content varies with time. Therefore, effective analysis requires time-frequency domain representations that capture this temporal variation. While spectrograms based on the STFT are commonly used for this purpose, they do not mimic the human auditory perception accurately. In contrast, the gammatonegram provides a biologically inspired alternative, leveraging a gammatone filterbank that simulates the frequency analysis mechanisms of the human cochlea [150].

The gammatonegram is computed by applying a gammatone filterbank to a signal and plotting the output energy across time and frequency. Each filter in the bank acts as a bandpass filter, selectively isolating frequency components within a specific range. These filters are designed to model the auditory filters of the inner ear, particularly the basilar membrane response within the cochlea. As in the auditory system, the bandwidths of gammatone filters are narrower at lower frequencies and become broader at higher frequencies, mirroring the human ear's greater sensitivity to frequency differences at the low end of the spectrum [139].

The impulse response of each gammatone filter is given by the following equation:

$$gt(t) = at^{(n-1)}e^{-2\pi bt}\cos(2\pi f_c t + \Phi), \quad t > 0 \quad (2.4)$$

where f_c is the center frequency of the filter, Φ is the phase (commonly set to 0), a is a gain factor, n is the filter order which controls the slope of the filter skirts, and b is the decay factor related to the filter's bandwidth and temporal duration [150, 139]. This formulation models the response of the cochlea to incoming sound waves, producing an auditory-inspired decomposition of the signal.

To determine the distribution of center frequencies and bandwidths across the auditory filterbank, the Equivalent Rectangular Bandwidth (ERB) scale is often used [151]. The ERB function approximates the frequency resolution of the human auditory system and defines the bandwidth of the gammatone filters as:

$$ERB(f_c) = 24.7 + 0.108f_c \quad (2.5)$$

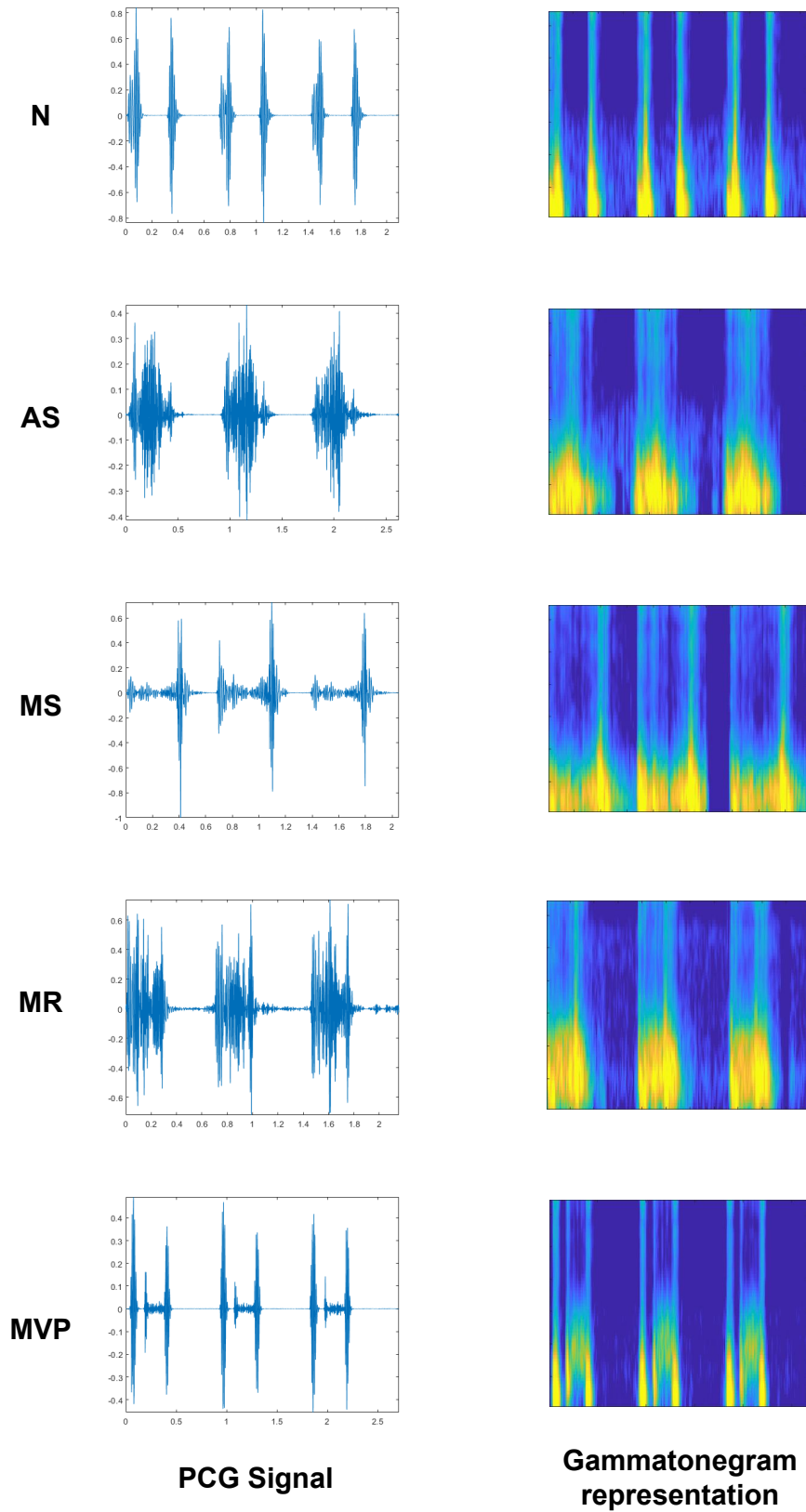


Figure 2.2 – Gammatonegram representations of different heart sound classes from the Yaseen et al. dataset [18].

This equation indicates that filter bandwidths grow linearly with increasing center frequency, consistent with physiological studies of cochlear function. The ERB scale is central to constructing a perceptually uniform filterbank that captures both fine and coarse frequency details in a way that aligns with auditory perception [149].

Equations 2.4 and 2.5 together define a gammatone auditory filterbank, which offers a reliable approximation of the cochlear frequency response. When the order n of the filter corresponds to the 3-dB bandwidth, the magnitude response of the gammatone filter closely resembles that of the roex(p) filter, a model traditionally used to represent the auditory filter shape in psychoacoustic experiments [150]. This resemblance makes the gammatonegram a robust and perceptually relevant time-frequency representation for heart sound analysis [152].

Unlike conventional spectrograms, the gammatonegram provides a more realistic simulation of how the human ear perceives sound, particularly in terms of frequency resolution at different spectral regions. This makes it a valuable tool in biomedical applications such as PCG classification, where the representation must preserve subtle frequency characteristics of murmurs, clicks, and other pathological components.

Figure 2.2 illustrates example gammatonegram representations extracted from various PCG classes in the Yaseen et al. dataset [18]. As seen, the gammatonegram reveals class-specific spectral patterns in a time-frequency domain that is perceptually aligned with human hearing.

2.4 Mel-spectrogram Representation

Spectrogram-based time-frequency representations encompass a variety of techniques, such as log-magnitude spectrograms, phase-corrected spectrograms, and amplitude spectrograms. These representations typically exhibit a linear spacing of frequency components across the spectrum [153]. However, this linear distribution does not align with the frequency sensitivity of the human auditory system, which is more responsive to low frequencies and less sensitive to high frequencies.

To address this mismatch, the Mel-spectrogram representation is introduced. It is based on the Mel scale, a perceptual scale of pitches judged by listeners to be equal in distance from one another. The Mel-spectrogram is obtained by applying a Mel-frequency filter bank to the STFT of the signal [154]. This representation effectively mimics the auditory system's emphasis on lower-frequency components, making it more suitable for analyzing biomedical acoustic signals like heart sounds.

The Mel filter bank consists of a series of overlapping triangular filters, each defined by a center frequency. Each filter has a value of 1 at its center frequency and tapers linearly to 0 at the edges. This overlapping design ensures smooth spectral transitions and captures a broader frequency context [155].

The Mel-spectrum can be computed by first estimating the power spectrum $P(k)$ of the signal, given

by:

$$P(k) = \frac{1}{N} |X(k)|^2 \quad (2.6)$$

Here, $X(k)$ represents the k -th bin of the STFT output, and N denotes the number of samples. The term $|X(k)|^2$ is the squared magnitude of the STFT, representing the energy content at each frequency bin [156].

The center frequencies f_m of the triangular filters are determined using the Mel scale transformation:

$$\phi = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \quad (2.7)$$

Each filter response $H_m(k)$ is defined as:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1) \leq k \leq f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)}, & f(m) \leq k \leq f(m+1) \\ 0, & k > f(m+1) \end{cases} \quad (2.8)$$

Once the filter responses are computed, the Mel-frequency logarithmic energy spectrum $S(m)$ for each filter m is given by:

$$S(m) = \log_e \left(\sum_{k=0}^{N-1} P(k) \cdot H_m(k) \right), \quad 0 \leq m \leq M \quad (2.9)$$

The Mel filter bank is typically configured with three key parameters: the number of filters M , the minimum frequency $f_{\min}$, and the maximum frequency $f_{\max}$. According to [157], heart and lung sounds predominantly lie in the 20–2000 Hz range. In practice, however, studies like [156] propose broader frequency spans to better capture signal variations—for instance, using $f_{\min} = 0$ Hz, $f_{\max} = 11.025$ kHz, and $M = 30$ filters.

The center Mel values are spaced linearly across the Mel scale:

$$\phi_c(m) = m\Delta\phi, \quad 0 \leq m \leq M \quad (2.10)$$

The center frequencies in Hertz are obtained by applying the inverse of Eq. 2.7, given by:

$$f_c(m) = 700 \left(10^{\phi_c(m)/2595} - 1 \right) \quad (2.11)$$

It is important to note that Eq. 2.7 represents a non-linear function over the entire frequency range. Consequently, a fixed frequency resolution is determined on the Mel scale $\Delta\phi$, consistent with the logarithmic nature of the repetition frequency. This resolution is computed using Eq. 2.12:

$$\Delta\phi = \frac{\phi_{\max} - \phi_{\min}}{M + 1} \quad (2.12)$$

Where $\phi_{\max}$ is the highest frequency of the filter bank on the Mel scale (derived from $f_{\max}$ via Eq. 2.7) and $\phi_{\min}$ is the lowest Mel-scale frequency, corresponding to $f_{\min}$ [156, 158].

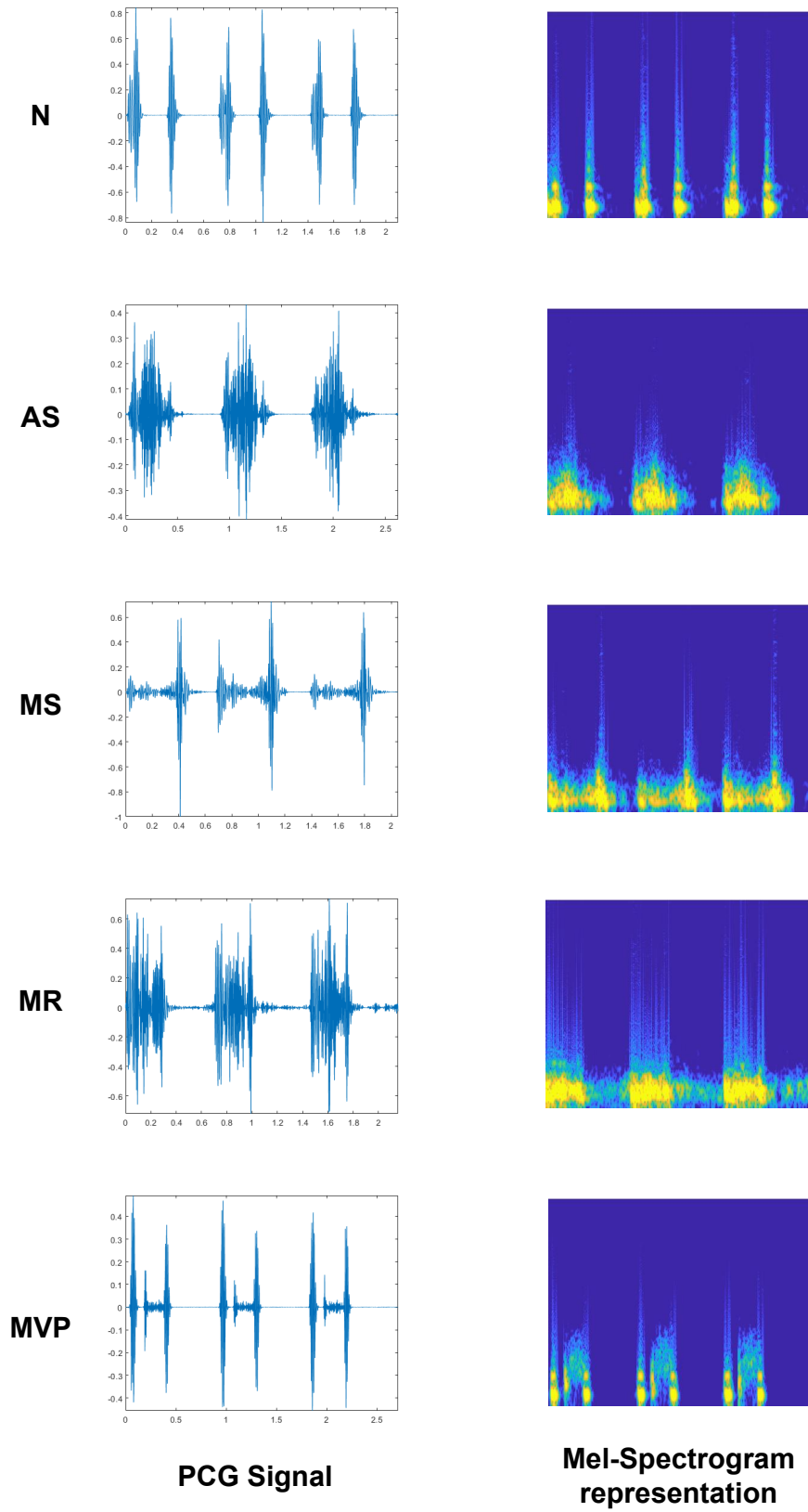


Figure 2.3 – Mel-spectrogram representations of different heart sound classes from the Yaseen et al. dataset [18].

Overall, the Mel-spectrogram offers a perceptually relevant and compact representation of heart sounds. Its emphasis on low-frequency features, which are often significant in detecting murmurs and abnormal heart cycles, enhances its effectiveness in automated heart sound classification.

2.5 CQT Representation

Nonstationary Gabor systems expand upon the conventional Gabor systems, which involve regular time-frequency shifts of a single window function. Arbitrary windows are considered to span the time-axis, and correspondingly, the frequency-sampling constants vary with time [159]. In STFT (Gabor transform), adequate frequency resolution is essential in a relatively low-frequency band, while optimal time resolution is needed in a relatively high-frequency band. Thus, achieving optimal frequency and time resolutions simultaneously using a fixed-length window function may pose a challenge [160].

In CQT, the window function's length varies, becoming shorter as the analysis frequency increases. Consequently, the time and frequency resolutions in CQT improve with a higher analysis frequency and decrease with a lower analysis frequency, respectively. This characteristic aligns with the time and frequency resolution requirements in STFT.

CQT improves frequency resolution by using geometrically spaced center frequencies for the analysis windows. Specifically, the Q factor, representing the ratio of the center frequency to the window's bandwidth, remains constant across all frequency bins [161]. Given that the Q factor offers precise frequency resolution, particularly for low frequencies, CQT emerges as a suitable representation for heart sound signals, given their lower frequency nature. Consequently, by employing the Non-Stationary Gabor transform in implementing the CQT, there is a potential reduction in computational costs and an enhancement in perceiving heart sounds effectively [160].

The synthesis of the signal can be represented as a linear combination of coefficients $\alpha_i = \langle s, \phi_i \rangle$ and the corresponding dual frame γ_i :

$$s = \mathbf{S}^{-1} \mathbf{S} s = \sum_i \langle s, \phi_i \rangle \mathbf{S}^{-1} \phi_i = \sum_i \alpha_i \gamma_i \quad (2.13)$$

The inversion of the operator $\mathbf{S}$ often poses challenges in the numerical implementation of the frame analysis. However, when $\mathbf{S}$ is diagonal, it becomes feasible to handle practical applications. This scenario of straightforward non-orthogonal expansions can be extended to enable concurrent adaptive time and frequency resolutions, a capability realized in CQT [162].

In subsequent equations, T_n denotes a time shift by n , M_ω signifies a frequency shift by ω , and $Fs = \hat{s}$ represents the Fourier transform of s . Let $\phi_k, k \in \mathbb{Z}$, represent the band-limited windows with Fourier transforms $\psi_k = \hat{\phi}_k$ centered on irregularly spaced frequency points ζ_k . The time-shift parameters

(hop sizes) a_k are selected based on the frequency dependence, according to the following:

$$a_k \leq \frac{1}{|L_k|} \quad (2.14)$$

where L_k is the length of the window function, and the frame members are generated by:

$$\varphi_{n,k} = T_{na_k} \varphi_k \quad (2.15)$$

In this case, using window φ_k and hop size a_k , the frame operator becomes diagonal in the Fourier domain. Consequently, we can derive the following equation:

$$\begin{aligned} \langle \mathbf{S}s, s \rangle &= \sum_{n,k} |\langle s, T_{na_k} \varphi_k \rangle|^2 = \sum_{n,k} |\langle \hat{s}, M_{-na_k} \hat{\varphi}_k \rangle|^2 \\ &= \langle \hat{s}, \sum_k \frac{1}{a_k} |\hat{\varphi}_k|^2 \hat{s} \rangle \end{aligned} \quad (2.16)$$

The frame operator takes on the following form:

$$\mathbf{S}s = \mathcal{F}^{-1} \left(\sum_k \frac{1}{a_k} |\hat{\varphi}_k|^2 \hat{s} \right) \quad (2.17)$$

The transform coefficient c_{nk} can be computed through the following:

$$C_{nk} = \langle s, T_{na_k} \varphi_k \rangle = \langle \hat{s}, M_{-na_k} \hat{\varphi}_k \rangle \quad (2.18)$$

The Q-factor in CQT depends on several key parameters: window functions, the number of frequency bins per octave, and the specified minimum and maximum frequencies. Let B represent the number of frequency bins per octave and $\zeta_{\min}$ denote the desired minimum frequency. We consider that band-limited window functions $\varphi_k \in \mathbb{C}^L$, $k = 1, \dots, K$ with center frequencies ζ_k (in Hz) can be calculated by:

$$\zeta_k = \zeta_{\min} \cdot 2^{\frac{k-1}{B}} \quad (2.19)$$

Similar to classical CQT, the maximum frequency $\zeta_{\max}$ is constrained to be less than the Nyquist frequency $\zeta_s/2$ (where ζ_s represents the sampling frequency). Consequently, the frequency center ζ_k is expressed as follows:

$$\zeta_k = \begin{cases} 0 & k = 0 \\ \zeta_{\min} \cdot 2^{\frac{k-1}{B}} & k = 1, \dots, K \\ \zeta_s/2 & k = K+1 \\ \zeta_s - \zeta_{2K+2-k} & k = K+2, \dots, 2K+1 \end{cases} \quad (2.20)$$

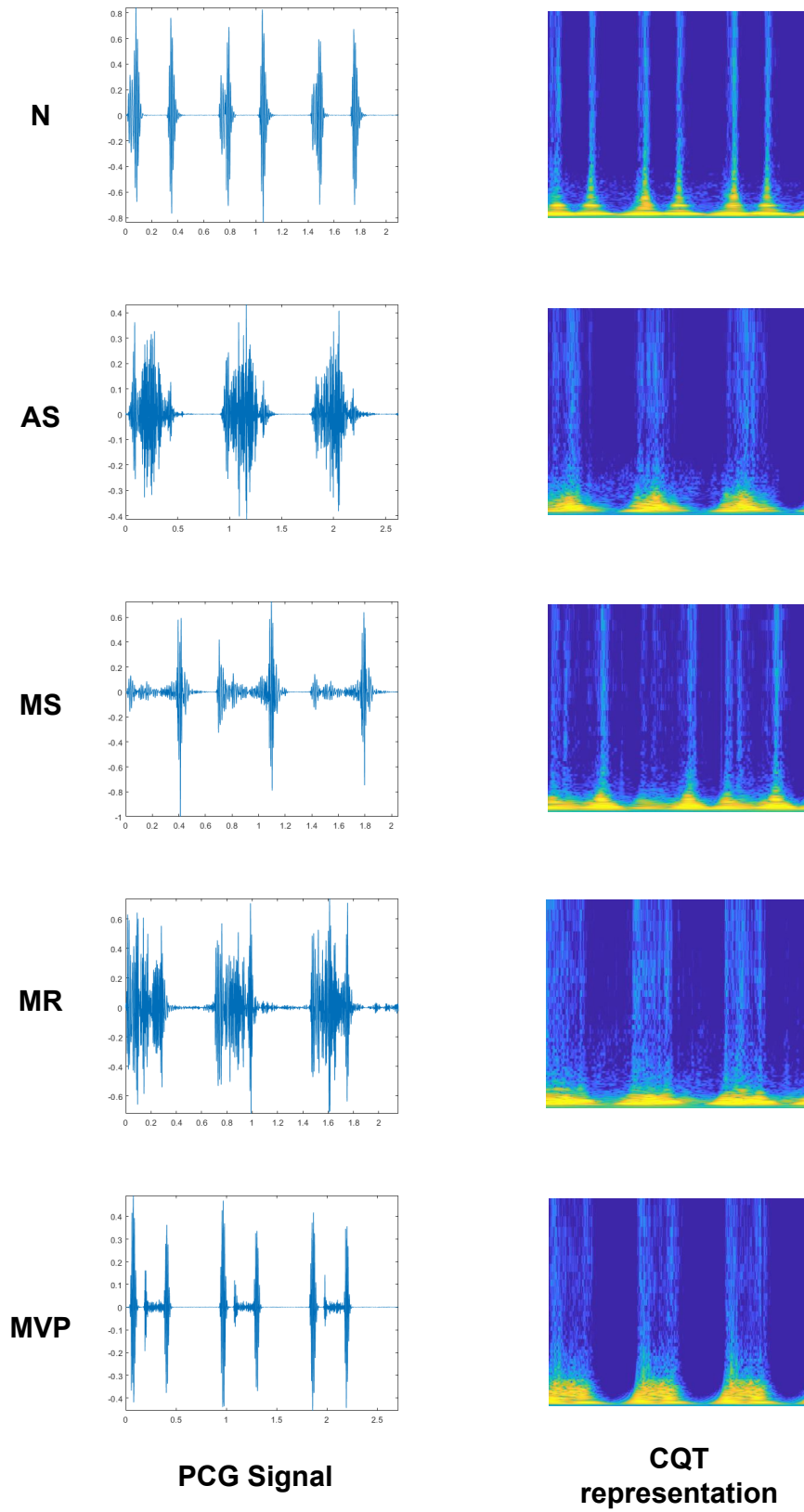


Figure 2.4 – CQT representations of different heart sound classes from the Yaseen et al. dataset [18].

The bandwidth Ω_k (representing the support of the window in frequency) for φ_k is determined using the following:

$$\Omega_k = \zeta_{k+1} - \zeta_{k-1}, \quad k = 2, \dots, K-1 \quad (2.21)$$

Resulting in a constant Q-factor:

$$Q = 2^{\frac{1}{B}} - 2^{-\frac{1}{B}} \quad (2.22)$$

Therefore, Ω_k is:

$$\Omega_k = \begin{cases} 2\zeta_{\min} & k = 0 \\ \zeta_k / Q & k = 1, \dots, K \\ \zeta_s - 2\zeta_k & k = K+1 \\ \zeta_{2K+2-k} / Q & k = K+2, \dots, 2K+1 \end{cases} \quad (2.23)$$

The adaptability of the CQT in balancing time and frequency resolutions makes it particularly suitable for analyzing non-stationary biomedical signals such as heart sounds. By leveraging its variable-resolution approach, the CQT can effectively capture low-frequency components critical in detecting pathological events. This capability reinforces its potential as a powerful preprocessing step for subsequent machine learning or diagnostic systems in the context of cardiac signal analysis.

2.6 Conclusion

This chapter provided a comparative overview of various time-frequency representations utilized in the analysis of heart sound signals, including the scalogram, Gammatonegram, Mel-spectrogram, and CQT. Each method offers distinct advantages in capturing the non-stationary characteristics of heart sounds. The scalogram, based on the continuous wavelet transform, excels in multi-resolution analysis, making it suitable for detecting transient features. The Gammatonegram, inspired by the human auditory system, provides a perceptually relevant decomposition that aligns well with how humans interpret sound. The Mel-spectrogram offers a compact and computationally efficient representation rooted in psychoacoustics, widely adopted in speech and biomedical applications. Finally, the CQT, leveraging a logarithmic frequency scale and constant Q-factor, demonstrates strong potential for analyzing low-frequency biomedical signals with improved frequency resolution.

By understanding the theoretical underpinnings and practical implications of each representation, we can better select the appropriate approach based on the specific requirements of heart sound classification tasks. This comparative insight sets the foundation for developing robust feature extraction pipelines and selecting suitable front-end processing techniques for machine learning and deep learning models in the subsequent chapters.

In this thesis, we challenge conventional approaches to preventing overfitting in CNNs, which typically rely on data segmentation and augmentation. Instead, we demonstrate that competitive accuracy can be achieved by training directly on raw, unprocessed data. Our results suggest that carefully designed CNN architectures can learn robust representations without explicit data preprocessing, while still maintaining generalization performance. This approach simplifies the training pipeline and may be particularly advantageous for applications where data preprocessing is costly or impractical.

Chapter 3

Deep Learning Models: Architecture and Complexity

Deep learning has emerged as a powerful paradigm for biomedical signal analysis, particularly in the classification of heart sound signals. In this chapter, we delve into various CNN architectures, including standard models like AlexNet and ResNet, as well as a customized DAG-CNN, highlighting their design principles and suitability for PCG classification. We discuss how these models extract hierarchical features from input data, the role of convolutional and pooling layers, and the effectiveness of skip connections in enhancing performance and reducing overfitting.

3.1 Introduction

The automatic classification of heart sounds is a critical task in biomedical signal analysis, with the potential to greatly assist in the early screening and diagnosis of cardiovascular diseases. Over the years, biomedical engineering has seen a significant evolution in how health-related data, particularly biomedical signals, is processed and analyzed. Traditional methods often relied on manual feature extraction and expert knowledge, which were effective in specific cases but became limited when applied to diverse patient populations or various recording conditions. These methods also lacked the ability to generalize across a wide range of clinical scenarios, hindering their practical application in everyday medical settings.

In recent years, the rise of deep learning, especially CNNs, has revolutionized the biomedical engineering field. Deep learning techniques have proven to be highly effective in automating the feature extraction process, which allows models to learn directly from data without the need for predefined features [163, 164, 165, 166, 167, 168, 169]. This capability is particularly valuable in the context of heart sound classification, where complex, high-dimensional data is difficult to interpret manually. CNNs, in particular, have demonstrated exceptional performance due to their ability to automatically

extract hierarchical features from structured inputs like images and time–frequency maps, which are common in biomedical signal processing.

Historically, heart sound classification relied heavily on traditional machine learning approaches, which often involved the use of handcrafted features derived from signal processing techniques such as STFT, wavelet transform, and MFCCs. While these methods achieved some degree of success, they faced challenges in handling the variability and complexity inherent in heart sound signals. Traditional machine learning techniques required extensive expert intervention in the feature engineering phase, and their performance often degraded when applied to diverse datasets or different patient conditions.

Deep learning models, particularly CNNs, offer a more robust solution by eliminating the need for manual feature extraction. CNNs learn features directly from the raw data or from processed TFRs of the data. In the case of heart sound classification, various time-frequency representations, such as spectrograms, and scalograms, can be used to convert one-dimensional PCG signals into two-dimensional images that are better suited for convolutional processing [13, 84, 99, 104, 100, 86]. This enables CNNs to capture the local and hierarchical features that characterize heart sounds, which are crucial for detecting subtle abnormalities indicative of cardiovascular diseases.

Moreover, recent advancements have introduced novel representations and architectures that further enhance the performance of CNNs in this domain. For instance, some studies have explored learning directly from raw waveforms [111, 113], while others have utilized innovative representations such as chaograms, which capture the nonlinear dynamics of heart sounds [120]. These approaches highlight the flexibility of CNNs and their ability to adapt to different types of input data, making them an ideal choice for heart sound classification.

The growing body of work in biomedical signal processing has shown that deep learning models, particularly CNNs, play a pivotal role in advancing the state-of-the-art in non-invasive cardiac diagnostics. Studies have highlighted the effectiveness of these models in improving classification accuracy and generalizability across diverse patient populations and recording conditions. The PhysioNet/CinC Challenge 2016 dataset, for example, has become a benchmark for evaluating heart sound classification models and has been used to validate the performance of CNN-based approaches [84, 99, 87]. Furthermore, innovations such as hybrid networks combining convolutional and recurrent layers, as well as ensemble learning strategies, continue to push the boundaries of heart sound classification [114, 117].

Building upon these advances, this chapter introduces the CNN-based models employed in this study for the classification of heart sound signals. Specifically, we focus on four architectures: AlexNet, VGG16, ResNet, and Data-Efficient Image Transformer (DeiT). These models have been selected based on their proven effectiveness in feature extraction and classification tasks. Furthermore, their diverse architectural characteristics offer an opportunity for a comparative evaluation across varying levels of complexity and design philosophy. This selection lays the groundwork for the development of our own deep learning-based approach to PCG classification.

3.2 Pretrained CNN Models

In recent years, the use of pretrained convolutional neural networks has become a widely adopted strategy in various classification tasks, including biomedical signal analysis. Pretrained models, initially trained on large-scale datasets such as ImageNet, are able to transfer learned feature representations to new domains, often resulting in faster convergence and improved performance compared to training from scratch. In the context of heart sound classification, leveraging pretrained architectures allows for the extraction of rich, hierarchical features even with relatively limited PCG datasets.

However, due to the limited size of available heart sound datasets and the distinct nature of PCG signals compared to natural images, pretrained networks can be prone to overfitting if not carefully regularized. This section provides an overview of the pretrained networks used in this study, namely AlexNet, VGG16, ResNet and DeiT, highlighting their key characteristics and relevance to the task at hand.

3.2.1 AlexNet

AlexNet is widely regarded as a landmark in the history of deep learning. Originally introduced by Krizhevsky et al. in 2012 [170], it significantly outperformed traditional machine learning approaches on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), reducing the top-5 error rate by more than 10 percentage points compared to the previous best models. This breakthrough ushered in a new era for deep learning and Convolutional Neural Networks (CNNs), establishing them as the standard for many tasks in computer vision and beyond [171].

The AlexNet architecture, as detailed in Table 3.1 and illustrated in Figure 3.1, consists of eight layers: five convolutional layers followed by three fully connected layers. The convolutional layers are responsible for learning spatial hierarchies of features through local receptive fields, while the fully connected layers act as classifiers. Notable features of AlexNet include the use of the ReLU activation function, which accelerates training convergence, and local response normalization (LRN), which encourages lateral inhibition. Max pooling is employed after certain convolutional layers to reduce spatial resolution and computation. Moreover, dropout is used in the fully connected layers to prevent overfitting by randomly deactivating a subset of neurons during training.

AlexNet was also among the first architectures to leverage GPU acceleration for large-scale training, splitting its layers across two GPUs to manage memory and computation more efficiently — a technique that would become increasingly common in deep learning research [170].

In biomedical engineering, AlexNet’s architecture has been successfully adapted for various diagnostic tasks including medical image analysis, signal processing, and particularly heart sound classification. The network has demonstrated strong performance when applied to both conventional 2D medical images and time-frequency representations of biomedical signals. Researchers have effectively utilized

transfer learning with AlexNet to analyze various signal representations, including spectrograms, MFCCs, scalograms, and other two-dimensional transformations of physiological data. This dual capability with both direct image inputs and converted signal representations makes AlexNet particularly versatile for biomedical applications.

In the context of PCG signal classification, AlexNet has been used either as a feature extractor or a full classifier. For example, Singh et al. (2019) utilized AlexNet to extract features from PCG spectrograms, showing that its pretrained filters could capture essential temporal and spectral characteristics of heart sounds [144]. Neili et al. (2022) also leveraged a modified AlexNet to classify PCGs using Mel-spectrogram inputs, achieving competitive performance compared to custom-designed CNNs [169]. Similarly, Alaskar et al. (2019) applied AlexNet to cardiac auscultation signals transformed into 2D representations, confirming its robustness across varying signal qualities [173]. Other notable studies, such as those by Mashrur et al. (2019) and Mansour et al. (2018), have integrated AlexNet into broader pipelines combining transfer learning with ensemble methods or traditional classifiers, underlining the model's versatility [174, 175].

Table 3.1 – Architecture of AlexNet as used in our PCG classification approach

Layer	Type	Output Size	Parameters / Description
0	Input	$227 \times 227 \times 3$	TFR of PCG signal
1	Convolution + ReLU	$55 \times 55 \times 96$	11×11 kernel, stride 4, no padding
2	Max Pooling	$27 \times 27 \times 96$	3×3 kernel, stride 2
3	Convolution + ReLU	$27 \times 27 \times 256$	5×5 kernel, stride 1, padding 2
4	Max Pooling	$13 \times 13 \times 256$	3×3 kernel, stride 2
5	Convolution + ReLU	$13 \times 13 \times 384$	3×3 kernel, stride 1, padding 1
6	Convolution + ReLU	$13 \times 13 \times 384$	3×3 kernel, stride 1, padding 1
7	Convolution + ReLU	$13 \times 13 \times 256$	3×3 kernel, stride 1, padding 1
8	Max Pooling	$6 \times 6 \times 256$	3×3 kernel, stride 2
9	Fully Connected + ReLU	4096	Dropout = 0.5
10	Fully Connected + ReLU	4096	Dropout = 0.5
11	Fully Connected	1000 (original) / 5 (ours)	Softmax output layer adapted to PCG classes

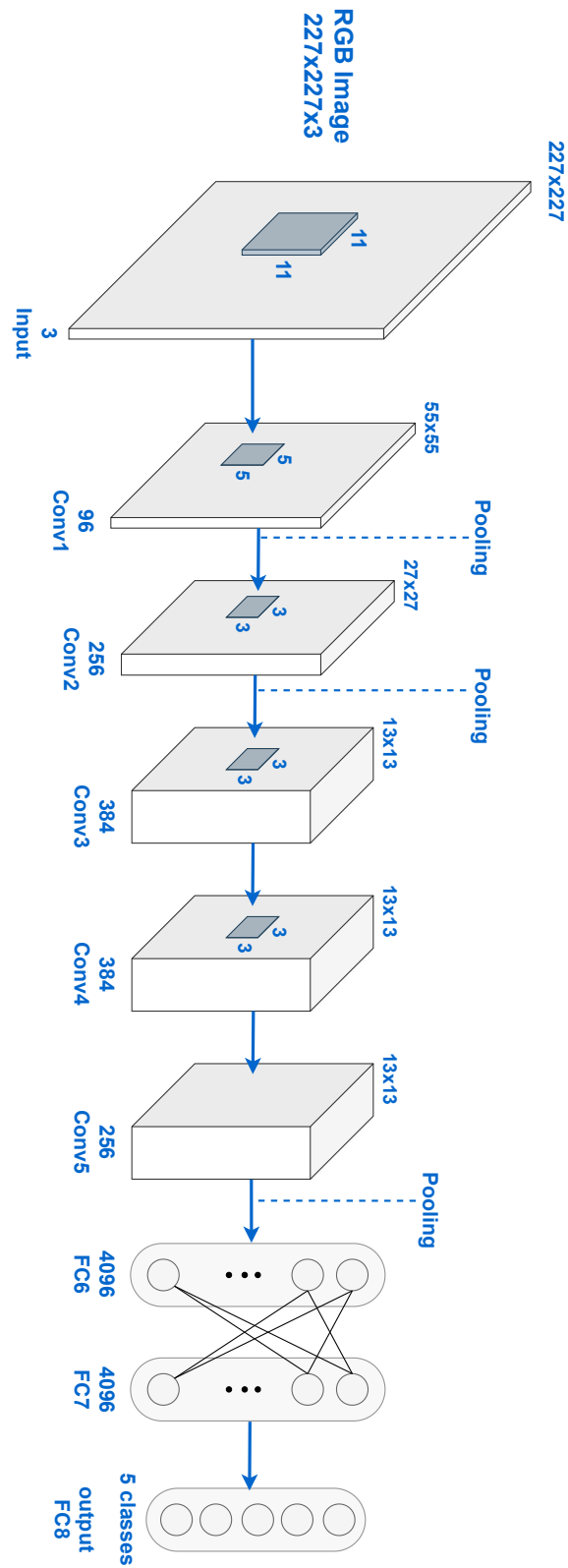


Figure 3.1 – Illustration of the AlexNet architecture (redrawn after [172]).

These applications underscore AlexNet’s continued relevance in biomedical research, particularly as a baseline or comparative model in studies exploring more complex architectures. Its relatively shallow depth and straightforward design make it an excellent candidate for tasks where interpretability, speed, and computational efficiency are desired.

In our study, we employed the AlexNet model to classify PCG signals based on their TFRs. Specifically, we utilized TFRs generated from PCG recordings as input to the pretrained AlexNet architecture. As shown in Table 3.2, the network was fine-tuned to learn discriminative features from these representations, enabling the classification of normal and abnormal heart sounds with improved generalization over traditional handcrafted methods.

Table 3.2 – Fine-tuning hyperparameters used for AlexNet

Hyperparameter	Value
Optimizer	Stochastic Gradient Descent with Momentum (SGDM)
Mini-batch size	50
Maximum number of epochs	10
Shuffle	Every epoch
Validation frequency	Every 3 iterations
Initial learning rate	0.001
Learning rate schedule	Constant

3.2.2 ResNet-18

ResNet-18, a member of the Residual Network (ResNet) family, introduced by He et al. in 2015, represents a significant milestone in deep neural network design. It was developed to address the degradation problem encountered when training very deep networks—where accuracy degrades as network depth increases, not due to overfitting but rather optimization difficulties [176]. ResNet achieved this by introducing residual learning, allowing the network to learn identity mappings more easily through shortcut connections. These identity mappings preserve information across layers, facilitating the training of much deeper architectures.

ResNet-18 specifically comprises 18 layers with learnable parameters and follows a straightforward architecture: a single convolutional layer followed by four residual blocks, each containing two convolutional layers. As detailed in Table 3.3 and illustrated in Figure 3.3, these residual blocks allow the model to retain low-level features while gradually building up high-level abstractions. Batch normalization is applied after each convolution, and ReLU activations are used throughout the network to maintain non-linearity and improve convergence. A global average pooling layer precedes the final fully connected layer, which outputs class probabilities.

Due to its relatively moderate depth and strong generalization capabilities, ResNet-18 has been widely adopted in medical image and signal analysis. In particular, its residual connections allow it to effectively learn from smaller and more imbalanced biomedical datasets, where deeper models might struggle to converge or overfit.

Studies such as Rahman et al. (2020) and Mhasakar et al. (2021) successfully applied ResNet-18 to classify pathological heart sounds and ECG signals, showing that residual learning can enhance both accuracy and stability in clinical decision-making tasks [178, 179]. Jing et al. (2021) and Gao et al. (2022) further explored its capabilities on PCG signals transformed into spectrograms, demonstrating that ResNet-18 maintains robust performance across a variety of time–frequency representations and noise conditions [180, 181].

Like AlexNet, ResNet-18 supports transfer learning and has been pre-trained on the ImageNet dataset, enabling its application to biomedical tasks with minimal adjustments. Its architecture accommodates various forms of 2D signal representation—including spectrograms, MFCCs, and scalograms—making it a flexible backbone for audio-based diagnostic tools.

Table 3.3 – Architecture of ResNet-18 as used in our PCG classification approach

Layer	Type	Output Size	Parameters / Description
0	Input	$224 \times 224 \times 3$	TF representation of PCG signal
1	Convolution + BN + ReLU	$112 \times 112 \times 64$	7×7 kernel, stride 2, padding 3
2	Max Pooling	$56 \times 56 \times 64$	3×3 kernel, stride 2
3	Residual Block $\times 2$	$56 \times 56 \times 64$	Each block: $2 \times [3 \times 3$ Conv, BN, ReLU] with skip connection
4	Residual Block $\times 2$	$28 \times 28 \times 128$	First block uses stride 2 for down-sampling
5	Residual Block $\times 2$	$14 \times 14 \times 256$	First block uses stride 2 for down-sampling
6	Residual Block $\times 2$	$7 \times 7 \times 512$	First block uses stride 2 for down-sampling
7	Global Average Pooling	$1 \times 1 \times 512$	Reduces each 7×7 feature map to a single value
8	Fully Connected	1000 (original) / 5 (ours)	Softmax output layer adapted to PCG classes

In this study, we fine-tuned a pretrained ResNet-18 model to classify PCG signals based on their time–frequency representations. TFRs were used as input images, and the final classification layer was modified to output five classes corresponding to different cardiac states. As shown in Table

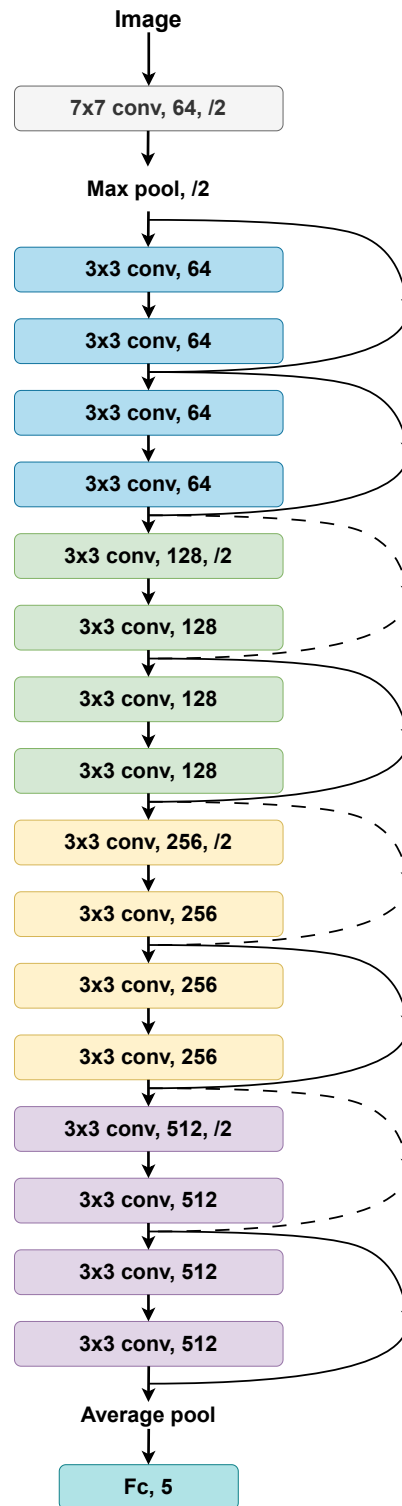


Figure 3.2 – Illustration of the ResNet-18 architecture (redrawn after [177]).

3.4, the model was trained using standard hyperparameter settings and validated during training to avoid overfitting. The residual structure of ResNet-18 proved beneficial for extracting both low- and high-level features from non-stationary heart sound signals, contributing to accurate classification across a diverse PCG dataset.

Table 3.4 – Fine-tuning hyperparameters used for ResNet-18

Hyperparameter	Value
Optimizer	Stochastic Gradient Descent with Momentum (SGDM)
Mini-batch size	50
Maximum number of epochs	10
Shuffle	Every epoch
Validation frequency	Every 6 iterations
Initial learning rate	0.001
Learning rate schedule	Constant

3.2.3 VGG16

VGG16 is a deep convolutional neural network architecture proposed by Simonyan and Zisserman in 2014, as part of the Visual Geometry Group (VGG) at the University of Oxford [182]. Known for its simplicity and uniform design, VGG16 employs small 3×3 convolution filters and 2×2 max-pooling layers consistently across its depth, leading to a significantly deeper architecture than its predecessors while maintaining manageable computational complexity. The model achieved remarkable performance on the ILSVRC, finishing among the top performers and demonstrating that increasing network depth could lead to substantial accuracy improvements.

The VGG16 architecture, detailed in Table 3.5, comprises 13 convolutional layers grouped into blocks, each followed by a max-pooling operation, and three fully connected layers at the end. Unlike earlier architectures with larger receptive fields, VGG16 replaces them with a series of stacked small filters, which allows for more non-linearities and thus more expressive power, while still covering the same effective receptive field. All hidden layers use the ReLU activation function, and the final output is produced by a softmax classifier. Despite its depth, VGG16 maintains a regular pattern that facilitates implementation and adaptation in different domains.

In biomedical contexts, VGG16 has been widely adopted for tasks such as disease detection, tissue classification, and analysis of medical signals represented as images. Due to its fixed architecture and availability of pretrained weights on ImageNet, it is frequently used in transfer learning pipelines to extract features from biomedical data. For example, Patel et al. (2022) used VGG16 to classify heart sound recordings transformed into time–frequency images, demonstrating its capability to capture relevant patterns for PCG analysis [183]. Similarly, Jiang et al. (2021) applied VGG16 to detect

congenital heart diseases from Mel-spectrogram representations of PCG signals, highlighting the effectiveness of the architecture in identifying spectral irregularities [184].

Other studies, such as those by Albashish et al. (2021) and Almanifi et al. (2022), explored the use of VGG16 as a feature extractor or an end-to-end classifier in diverse biomedical applications. These works confirmed that VGG16 performs robustly across different signal modalities and imaging techniques, particularly when combined with spectrograms, gammatonegrams, or CQT [186, 187]. Despite being relatively deep, VGG16 maintains manageable computational complexity, making it suitable for both offline analysis and deployment on moderately powered systems.

Table 3.5 – Architecture of VGG16 as used in our PCG classification approach

Layer	Type	Output Size	Parameters / Description
0	Input	$224 \times 224 \times 3$	TFR of PCG signal
1–2	Convolution + ReLU	$224 \times 224 \times 64$	3×3 kernels, stride 1, padding 1
3	Max Pooling	$112 \times 112 \times 64$	2×2 kernel, stride 2
4–5	Convolution + ReLU	$112 \times 112 \times 128$	3×3 kernels, stride 1, padding 1
6	Max Pooling	$56 \times 56 \times 128$	2×2 kernel, stride 2
7–9	Convolution + ReLU	$56 \times 56 \times 256$	3×3 kernels, stride 1, padding 1
10	Max Pooling	$28 \times 28 \times 256$	2×2 kernel, stride 2
11–13	Convolution + ReLU	$28 \times 28 \times 512$	3×3 kernels, stride 1, padding 1
14	Max Pooling	$14 \times 14 \times 512$	2×2 kernel, stride 2
15–17	Convolution + ReLU	$14 \times 14 \times 512$	3×3 kernels, stride 1, padding 1
18	Max Pooling	$7 \times 7 \times 512$	2×2 kernel, stride 2
19	Fully Connected + ReLU	4096	Dropout = 0.5
20	Fully Connected + ReLU	4096	Dropout = 0.5
21	Fully Connected	1000 (original) / 5 (ours)	Softmax output layer adapted to PCG classes

In our study, VGG16 was adapted to classify PCG signals based on their TFRs. Spectrograms generated from the original heart sound recordings were used as input to the pretrained VGG16 network. The final classification layer was fine-tuned as illustrated in Table 3.6 to distinguish between normal

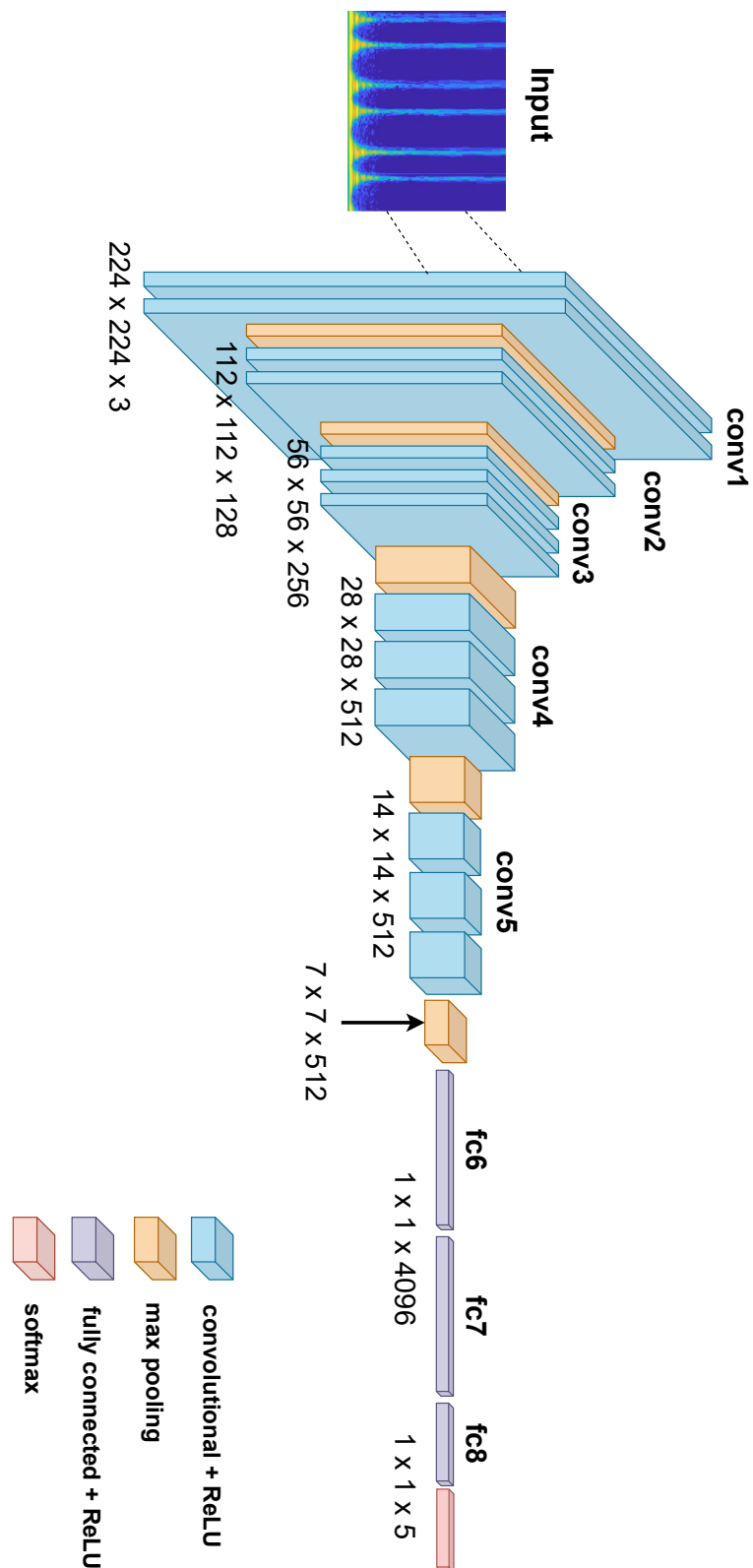


Figure 3.3 – Illustration of the VGG16 architecture (redrawn after [185]).

and abnormal heart sounds. This transfer learning approach enabled the network to extract highly discriminative features, yielding competitive performance compared to other baseline models and validating its suitability for PCG classification tasks.

Table 3.6 – Fine-tuning hyperparameters used for VGG16

Hyperparameter	Value
Optimizer	Stochastic Gradient Descent with Momentum (SGDM)
Mini-batch size	52
Maximum number of epochs	10
Shuffle	Every epoch
Validation frequency	Every 6 iterations
Initial learning rate	0.001
Learning rate schedule	Constant

3.2.4 Data-efficient Image Transformer (DeiT)

Transformers, initially developed for natural language processing tasks, have recently made a significant impact in computer vision. The Vision Transformer (ViT) introduced by Dosovitskiy et al. [188] demonstrated that pure transformer architectures could achieve state-of-the-art performance on image classification benchmarks, rivaling or surpassing CNNs, provided that sufficient training data and computational resources are available. However, the original ViT required large-scale datasets and heavy pretraining, which limited its adoption in domains with limited labeled data, such as medical imaging.

To address these limitations, Touvron et al. introduced DeiT, a vision transformer trained without requiring large external datasets [189]. DeiT uses knowledge distillation from a convolutional teacher (typically a CNN) to guide the training of the transformer student, allowing it to perform competitively even on relatively smaller datasets like ImageNet. This technique, combined with architectural optimizations and data augmentation strategies, makes DeiT particularly well-suited for tasks where labeled data is scarce, as often seen in biomedical domains.

The DeiT architecture, illustrated in Figures 3.4 and 3.5, also detailed in Table 3.7, follows the standard ViT framework. The input image is split into fixed-size patches, each of which is linearly projected into an embedding vector. These embeddings, along with a learnable class token and positional encodings, are passed through a sequence of transformer encoder layers. Each encoder block consists of multi-head self-attention and feed-forward layers, along with residual connections and layer normalization. In the distilled version (DeiT-Distilled), a second token—the distillation token—is added to learn from the teacher’s guidance, improving data efficiency and generalization.

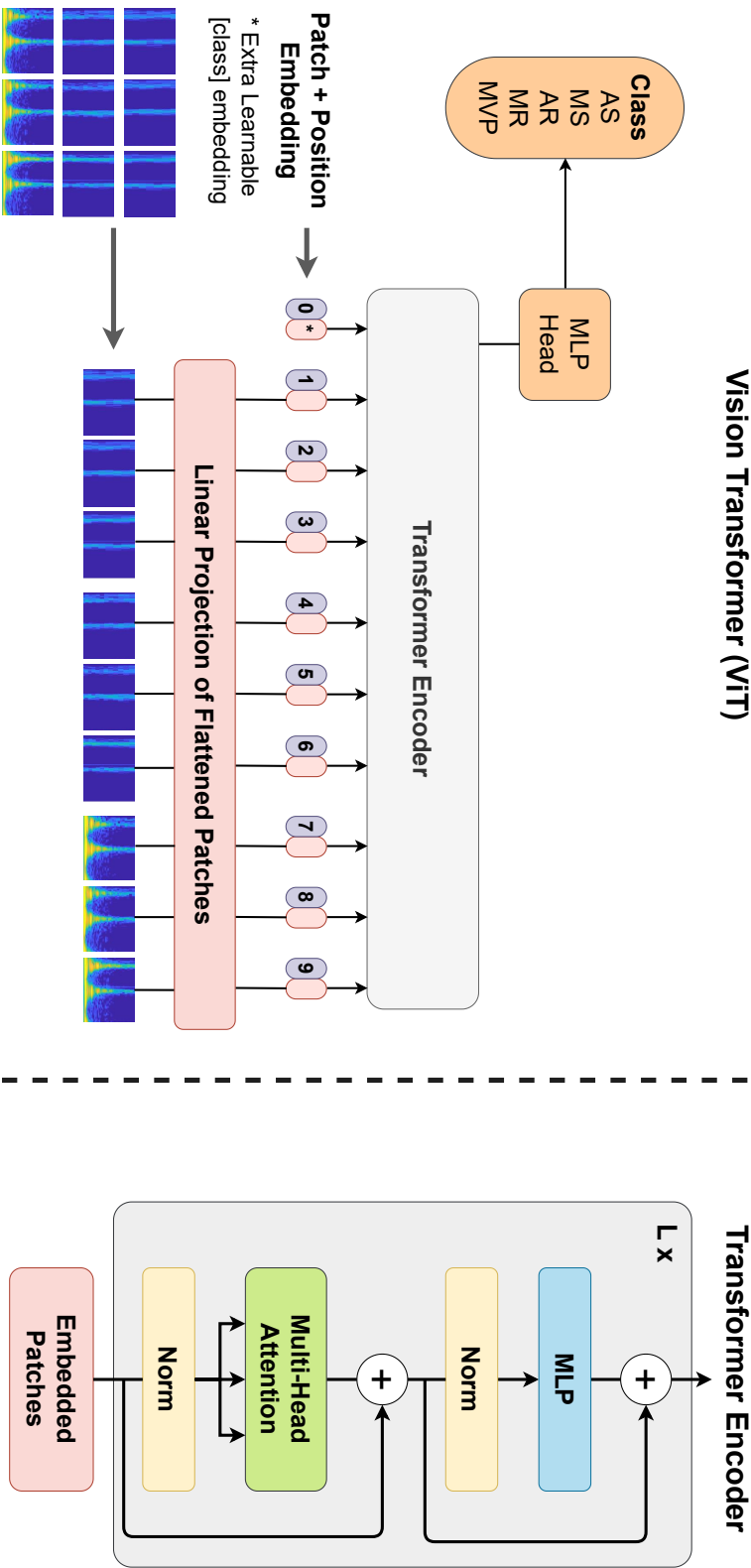


Figure 3.4 – Illustration of the ViT architecture (redrawn after [190]).

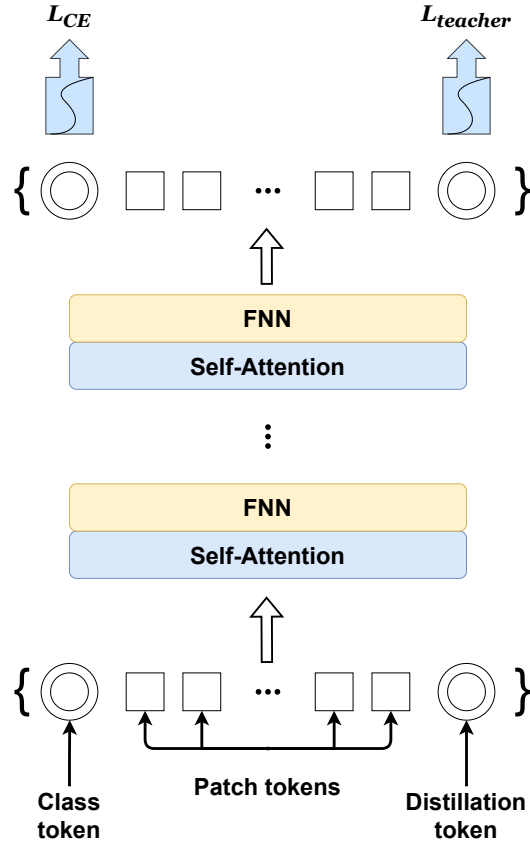


Figure 3.5 – Illustration of the DeiT architecture (redrawn after [191]).

In biomedical signal analysis, particularly heart sound classification, the shift from CNN-based models to transformer-based ones like DeiT has opened new avenues for capturing global dependencies across time-frequency representations. Unlike CNNs, which rely on local receptive fields, DeiT can model long-range interactions between spectral components in a signal. Jumphoo et al. (2024) demonstrated the effectiveness of vision transformers for heart sound classification by training a ViT-based model on PCG spectrograms, showing improved robustness to noise and variability [192]. Similarly, Bi et al. (2024) explored the use of DeiT for classifying cardiac auscultation signals, achieving state-of-the-art results with relatively limited training data [193].

Our study leverages the distilled DeiT architecture to classify PCG signals using their TFRs. The latter were used as input patches to the network, and the model is fine-tuned as shown in table 3.8 to recognize normal and pathological heart sounds. The ability of DeiT to model global contextual relationships proves advantageous for identifying subtle temporal patterns and frequency shifts indicative of cardiac anomalies. The results show that DeiT not only performs competitively with traditional CNNs but also provides better generalization across different datasets and noise conditions.

Table 3.7 – Architecture of DeiT used in our PCG classification approach

Layer	Type	Output Size	Parameters / Description
0	Input	$224 \times 224 \times 3$	Time–frequency representation of PCG signal
1	Patch Embedding	196×768	16×16 patch size, linear projection + flattening
2	Add Tokens	197×768	Class token (and distillation token in DeiT-distilled), added to patch embeddings
3	Add Positional Encoding	197×768	Learnable positional encodings added to embeddings
4–13	Transformer Encoder Blocks	197×768	Each block: Multi-head self-attention + MLP, 12 layers, 12 attention heads, residual + layer norm
14	Classification Head	$768 \rightarrow 5$	MLP head applied to class token (softmax over PCG classes)

Table 3.8 – Fine-tuning hyperparameters for DeiT on PCG classification task

Hyperparameter	Value
Optimizer	Adam
Patch size	32
Number of patches	49
Projection dimension	64
Number of transformer layers	6
Number of attention heads	6
MLP expansion ratio	3.0
Batch size	12
Learning rate	0.0001
Number of epochs	20

3.3 Proposed Model

Deep learning techniques have increasingly become indispensable in the medical domain, particularly for tasks such as diagnosis, disease detection, and classification. Among these techniques, CNNs have emerged as one of the most effective tools, especially in identifying heart and lung conditions.

CNNs are well-suited for processing both two-dimensional images and one-dimensional signals. Their typical structure comprises an input layer, multiple convolutional and ReLU activation layers, pooling layers, fully connected layers, and an output classification layer [93]. The convolutional layers utilize spatial filters to scan the input's spatial dimensions, extracting relevant patterns. Pooling layers are subsequently applied to downsample these representations, thus reducing model complexity and mitigating the risk of overfitting. Batch normalization layers are inserted between convolution and activation layers to stabilize training dynamics and reduce the sensitivity to weight initialization. At the network's end, the fully connected layers consolidate feature representations to perform final classification [138, 194].

Table 3.9 – Customized DAG-CNN architecture used in this study

Stage	Layer	Output Size	Parameters / Description
0	Input	$32 \times 32 \times 1$	TFR of PCG signal
1	Conv1 + BN + ReLU	$32 \times 32 \times 7$	3×3 kernel, 7 filters
2	Conv2 + BN + ReLU	$32 \times 32 \times 28$	3×3 kernel, 28 filters
3	Conv3 + BN + ReLU	$32 \times 32 \times 28$	3×3 kernel, 28 filters
4	Average Pooling	$16 \times 16 \times 28$	Pool size 2×2 , stride 2
5	Skip Connections	-	Feature fusion across layers
6	Flatten	1D vector	Vectorization for classification
7	Fully Connected	5	Dense layer
8	Softmax	5	Classification into 5 classes

Conventional CNNs typically follow a sequential architecture, where each layer's output is directly passed to the next. However, as the depth of such networks increases, this linear design can result in diminished performance due to vanishing gradients or loss of fine-grained information. To overcome these limitations, directed acyclic graph-based CNNs (DAG-CNNs) offer a more flexible topology. These architectures support multiple computation paths by introducing skip connections that allow earlier layers to directly contribute to the output layer. This design enables low-, mid-, and high-level features to be fused and utilized for final prediction, thus enhancing the model's representational power [195].

While this multi-path architecture improves accuracy, it can also lead to overfitting due to the large number of accumulated features. To counter this, only a subset of the most informative feature maps is selected. The features from different convolutional layers are routed into an auxiliary block, where they are combined using element-wise addition. This aggregated output is then passed through an average pooling operation before being classified by a softmax layer, which produces the final prediction scores [93].

In this study, we propose a custom-designed DAG-CNN tailored for the classification of TFR images derived from PCG signals. As illustrated in 3.9, the architecture begins with 32×32 input images and

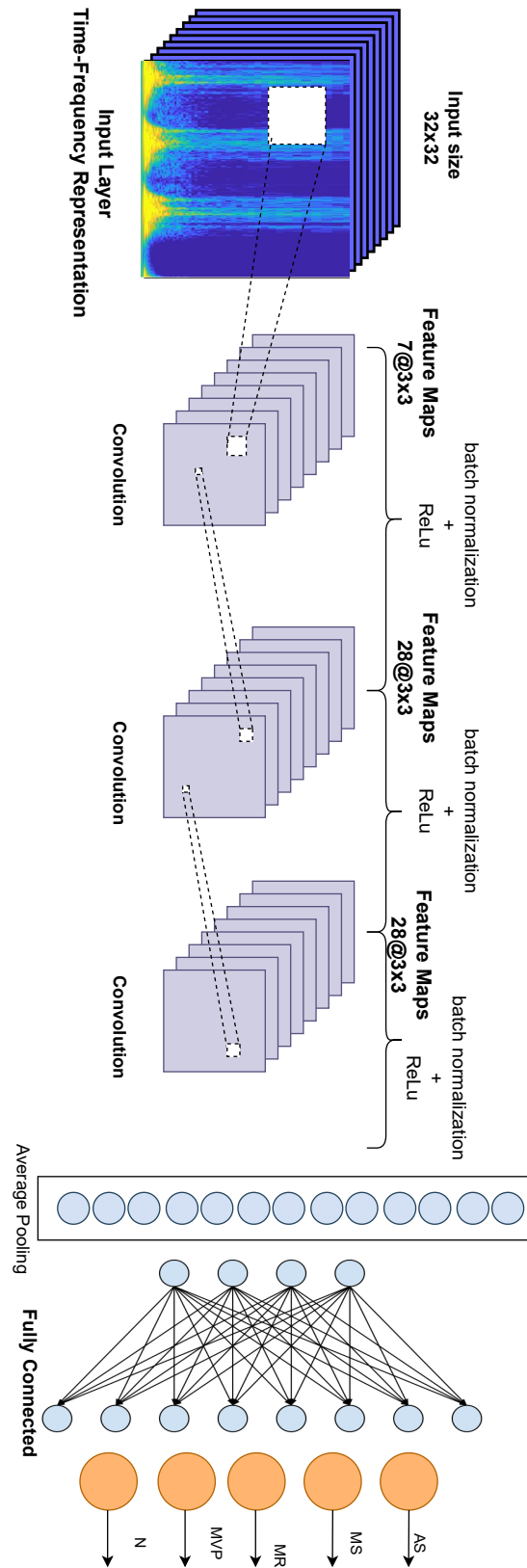


Figure 3.6 – Illustration of the proposed DAG-CNN architecture.

includes three convolutional layers using 3×3 filters. The first layer employs 7 filters, followed by two successive layers, each with 28 filters. Batch normalization and ReLU activation functions are applied after each convolution to improve learning stability and introduce non-linearity. An average pooling layer with a 2×2 kernel and stride of 2 is then used to reduce spatial dimensions while preserving important features. Skip connections are integrated to promote feature reuse and enhance generalization. The final fully connected layer learns higher-level abstractions from the combined features, followed by a softmax layer to classify the inputs into five categories. The hyperparameters used for training are summarized in tableau 3.10, and the complete DAG-CNN architecture is illustrated in 3.6.

Table 3.10 – Fine-tuning hyperparameters used for for training the proposed DAG-CNN model

Hyperparameter	Value
Optimizer	Stochastic Gradient Descent with Momentum (SGDM)
Mini-batch size	12
Maximum number of epochs	10
Shuffle	Every epoch
Validation frequency	Every 6 iterations
Initial learning rate	0.01
Learning rate schedule	Constant

3.4 Conclusion

In this chapter, we explored a range of deep learning architectures that have demonstrated significant potential in the domain of heart sound classification. Starting with classical convolutional neural networks such as AlexNet, ResNet and VGG16, we highlighted how these models leverage convolutional layers, activation functions, and hierarchical feature extraction to perform accurate classifications. We then introduced the more recent Vision Transformer approach through the DeiT model, illustrating its shift from convolutional mechanisms to self-attention, which allows for global context modeling within input signals.

Particular emphasis was placed on convolutional neural networks due to their proven effectiveness in processing both 1D and 2D representations of physiological signals. We presented a custom-designed DAG-CNN architecture tailored for PCG classification, which integrates skip connections and optimized layer selection to enhance feature representation while mitigating overfitting.

The comparative analysis and architectural insights provided in this chapter serve as the foundation for the proposed methodology, which builds upon these models' strengths while addressing their limitations in handling spectral representations of PCG signals in resource-constrained datasets.

In the next chapter, we present the classification results of PCG signals represented in the time–frequency domain, using the deep learning models discussed here. These results will help assess the practical performance and generalization ability of each architecture across the evaluation metrics.

Chapter 4

Performance Evaluation and Model Comparison

Previously in this thesis, we presented a range of TFR techniques—namely the scalogram, Gammatonegram, Mel-spectrogram, and CQT—designed to convert heart sound signals into two-dimensional representations that highlight both temporal and spectral characteristics. We also introduced several deep learning architectures, including classical CNNs, DAG-based CNNs, and vision transformer-based models such as DeiT, all tailored for the classification of PCG signals.

In this chapter, we bring these components together by evaluating how effectively the different deep learning models perform in classifying the TFR-transformed PCG signals. Through a series of carefully designed experiments, we measure and compare each model’s performance using standard classification metrics such as accuracy, precision, recall, and F1-score. The evaluation provides insights into how each TFR technique interacts with specific model architectures and helps identify optimal configurations for accurate heart sound classification. This analysis serves as a key step toward building reliable, automated diagnostic systems for cardiovascular screening.

4.1 Introduction

The classification of heart sound signals plays a pivotal role in the early detection and diagnosis of cardiovascular diseases, especially in resource-limited settings where automated systems can support medical professionals. In this chapter, we present the complete framework for heart sound classification based on deep learning techniques applied to TFRs of phonocardiogram PCG signals. The proposed method involves preprocessing raw PCG signals, converting them into TFR images (using scalogram, Gammatonegram, Mel-spectrogram, and CQT), and feeding these representations into different deep learning models—namely AlexNet, ResNet18, VGG16, DAG-CNNs, and DeiT—for classification. The proposed method was implemented using MATLAB 2022b and Python (v3.12) using Anaconda

distribution, executed on a Windows 10 system with an Intel® Core™ i5-3rd generation processor and 4GB RAM.

The dataset used in our experiments consists of 1,000 PCG recordings from Yaseen et al dataset [18], equally distributed among five diagnostic classes. Each signal was transformed into a JPEG image and used as input for model training. To prepare the data for classification, the dataset was divided into 80% training and 20% validation, ensuring a balanced representation of each class in both subsets. Specifically, 800 samples (160 per class) were allocated to the training set and 200 samples (40 per class) to the validation set, as summarized in Table 4.1. Unlike the traditional separation into training, validation, and test sets, we relied solely on validation during experimentation, as K-fold cross-validation was employed to ensure robustness and reduce variance in model evaluation. Each model was trained and evaluated using its own value of K , which allows a more generalized assessment across different subsets of the data.

Table 4.1 – Dataset split for training and validation.

Class	Training	Validation	Total
N	160	40	200
AS	160	40	200
MR	160	40	200
MS	160	40	200
MVP	160	40	200
Total	800	200	1000

A block diagram of the proposed classification pipeline is shown in Figure 4.1, summarizing each phase from signal acquisition to model inference. To evaluate the classification performance of each model, we employ standard metrics: accuracy, sensitivity, specificity, precision and F1 score. These metrics are computed as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4.1)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (4.2)$$

$$Specificity = \frac{TN}{TN + FP} \quad (4.3)$$

$$Precision = \frac{TP}{FP + TP} \quad (4.4)$$

$$F1Score = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity} \quad (4.5)$$

In the remainder of this chapter, we systematically present and analyze the classification results obtained by the deep learning models when applied to the various TFRs extracted from the heart sound

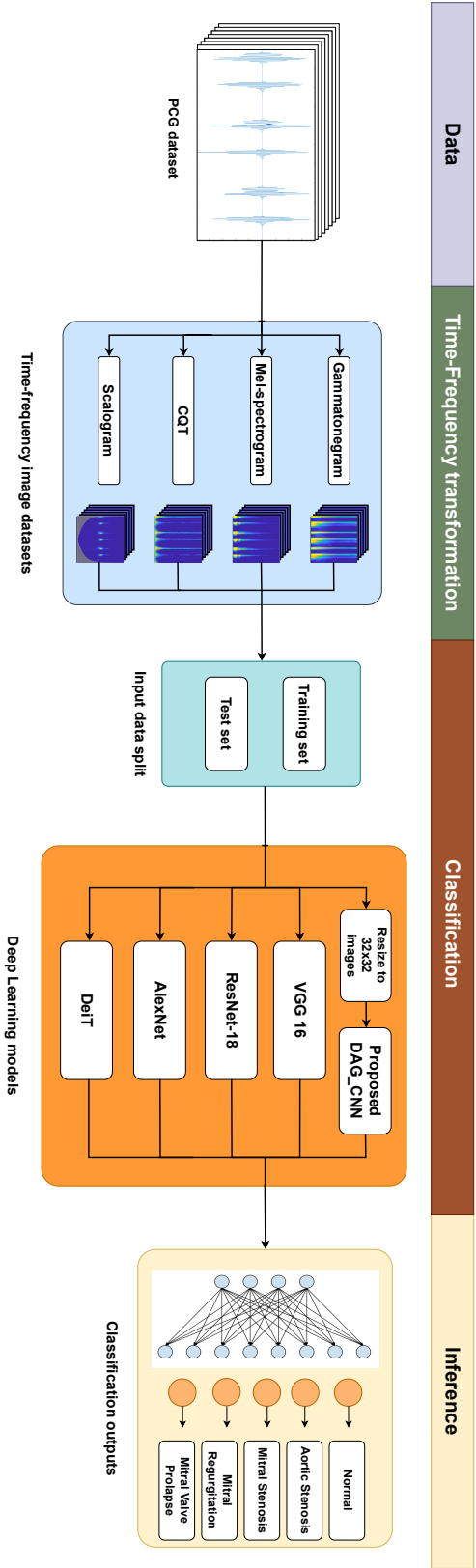


Figure 4.1 – Block diagram of the proposed method.

dataset, including the Scalogram, Gammatonegram, Mel-spectrogram, and Constant-Q Transform. For each model-TFR combination, we evaluate performance using key metrics. To provide a more detailed understanding of the models' behavior, we also include the corresponding confusion matrices as well as the training and validation accuracy/loss curves. These visual and quantitative analyses help identify trends, convergence patterns, and potential overfitting or generalization issues. Finally, we compare the performance of all models in a unified manner and benchmark our results against selected state-of-the-art approaches that have used the same dataset. This comprehensive evaluation aims to highlight the strengths and limitations of each architecture and assess their practical viability for automated heart sound classification.

4.2 AlexNet-Based Classification of PCG Signals

AlexNet, a pioneering convolutional neural network known for its breakthrough performance in image classification tasks, is well-suited for processing the 2D representations derived from heart sound signals. Leveraging its deep architecture and ability to learn spatial hierarchies of features, AlexNet is applied here to classify PCG signals transformed into Gammatonegram, Mel-spectrogram, and CQT images. In this section, we present the classification results obtained using AlexNet on each of these time-frequency representations, with performance evaluated through 5-fold cross-validation. Metrics such as accuracy, precision, recall, and F1-score are reported, complemented by training and validation curves as well as confusion matrices to assess the model's classification behavior across different representations.

Table 4.2 – AlexNet classification results on different TFRs using 5-fold cross-validation

TFR	Accuracy	Sensitivity	Specificity	Precision	F1-score
Gammatonegram	98.4%	95.5%	98.8%	96.9%	96.2%
Mel-spectrogram	98.5%	94.0%	98.5%	96.0%	95.0%
CQT	99.2%	99.6%	99.7%	99.1%	99.4%

As shown in Table 4.2 AlexNet demonstrated robust classification performance across all three TFRs, with CQT emerging as the most effective for heart sound classification. Quantitative results from 5-fold cross-validation revealed CQT achieved superior metrics (99.2% accuracy, 99.6% sensitivity, 99.7% specificity), outperforming both Mel-spectrogram (98.5% accuracy) and Gammatonegram (98.4% accuracy). This performance advantage stems from CQT's constant-Q resolution, which provides enhanced frequency resolution at lower frequencies where most pathological heart sounds occur while maintaining precise temporal alignment of cardiac events. Confusion matrix shown in Figures 4.2, 4.3 and 4.4 confirmed these findings, showing near-perfect classification for AS and normal cases and only 1-2 total misclassifications across other pathological classes. The Mel-spectrogram showed slightly reduced sensitivity (94.0%) due to minor confusion between MR/MS and MS/MVP cases,

likely attributable to its logarithmic frequency compression which may blur diagnostically critical spectral features. While Gammatonegram results appeared comparable statistically, formatting issues in its confusion matrix prevented complete pattern analysis, though visible entries suggested error trends similar to the Mel-spectrogram.

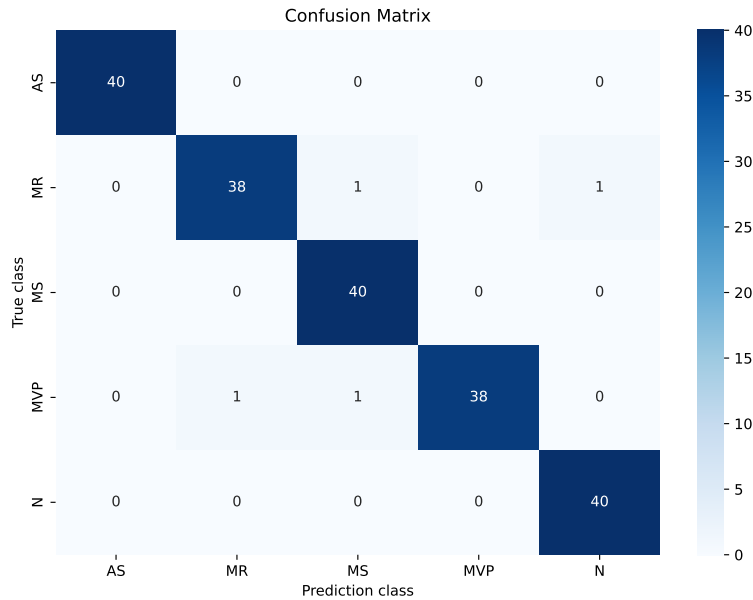


Figure 4.2 – Confusion matrix of AlexNet for gammatonegram image dataset.

Clinically, CQT’s exceptional specificity (99.7%) is particularly valuable as it minimizes false positives in normal cases - a critical requirement for screening applications. Its high sensitivity (99.6%) also ensures reliable pathological detection, with errors primarily occurring between MR/MS differentiation rather than more serious misclassifications. This performance profile suggests CQT is well-suited for clinical deployment where diagnostic precision is paramount. The Mel-spectrogram remains a viable alternative for applications prioritizing computational efficiency, as its 98.5% accuracy still represents strong screening capability. Both representations significantly outperformed traditional stethoscope-based diagnosis, which typically achieves 80-90% accuracy even with expert clinicians.

Several important limitations and future directions emerge from this analysis. First, the Gammatonegram’s performance requires reevaluation with properly formatted confusion matrices to ensure fair comparison. Second, the persistent MR/MS confusion across representations suggests these classes may have overlapping acoustic signatures in the feature space captured by these TFRs. This could potentially be addressed by incorporating additional temporal features or hybrid representations. Finally, while the >99% accuracy with CQT is impressive, these results should be validated on larger, more diverse datasets including real-world recordings with varying noise levels and comorbid conditions. Future work could explore ensemble approaches combining CQT’s spectral precision with Mel-scale’s computational efficiency, or investigate attention mechanisms to better capture subtle pathological features that current representations may miss. The strong performance across all TFRs ultimately

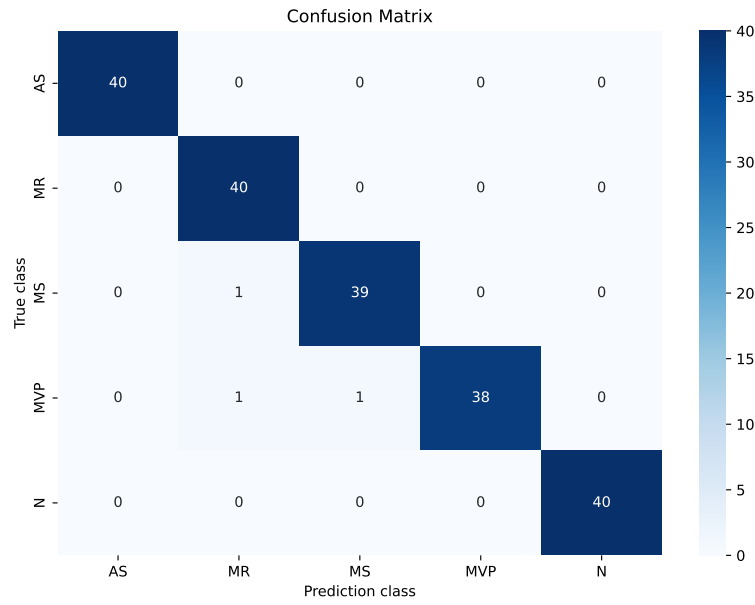


Figure 4.3 – Confusion matrix of AlexNet for Mel-Spectrogram image dataset.

validates AlexNet’s effectiveness for 2D image-based heart sound classification while highlighting the importance of representation selection for optimal diagnostic performance.

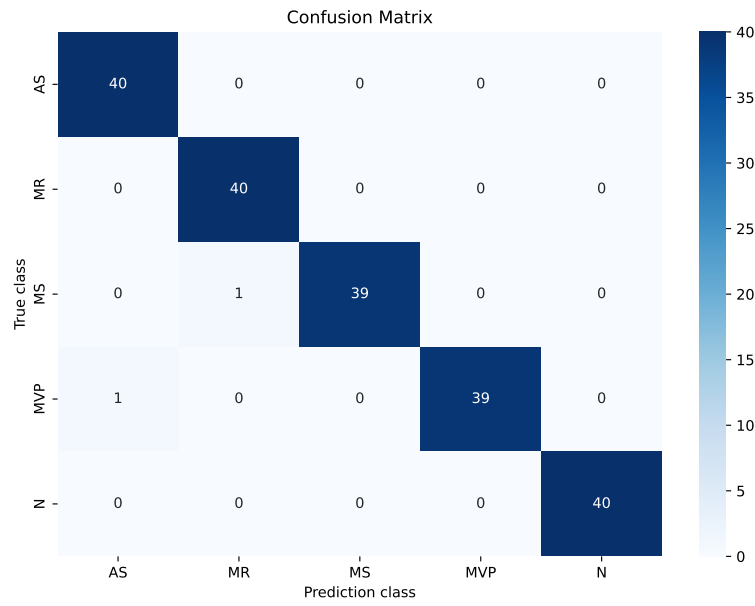


Figure 4.4 – Confusion matrix of AlexNet for CQT image dataset.

4.3 ResNet-Based Classification of PCG Signals

ResNet-18 is a deep convolutional neural network architecture that incorporates residual learning through skip connections, allowing for the effective training of deeper networks by addressing issues related to vanishing gradients. Despite its relatively modest depth, ResNet-18 has demonstrated strong performance in a wide range of image classification tasks due to its efficient feature extraction capabilities. In this section, we evaluate the classification performance of ResNet-18 applied to the three time-frequency representations previously described—namely the Gammatonegram, Mel-spectrogram, and CQT. The evaluation is carried out using 5-fold cross-validation, allowing for a reliable assessment of the model’s generalization capacity. The results are presented in the form of classification performance metrics and confusion matrices, providing insight into the model’s strengths and limitations when applied to each type of input representation.

Table 4.3 – ResNet classification results on different TFRs using 5-fold cross-validation

TFR	Accuracy	Sensitivity	Specificity	Precision	F1-score
Gammatonegram	97.8%	97.25%	98%	98.6%	97.9%
Mel-spectrogram	99%	98.4%	99%	99%	98.7%
CQT	99.1%	99.4%	99.6%	99%	99.2%

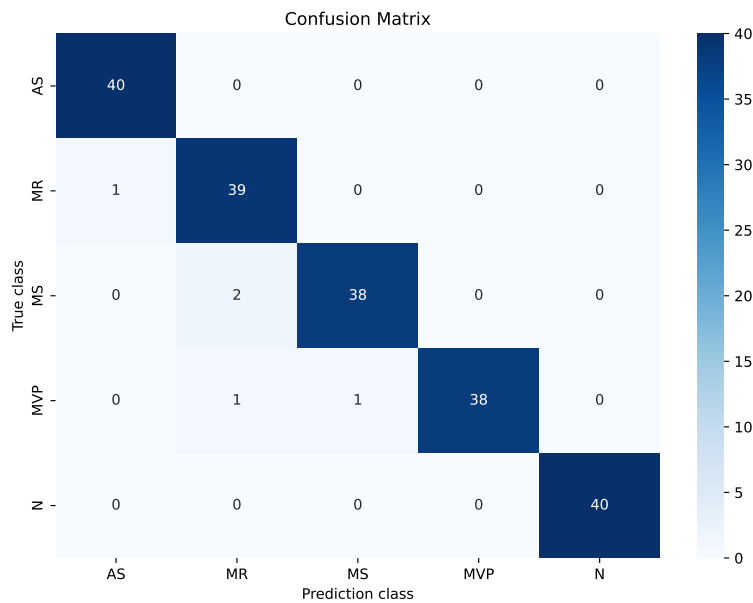


Figure 4.5 – Confusion matrix of ResNet for gammatonegram image dataset.

ResNet demonstrated superior classification capabilities across all three time-frequency representations (TFRs), building upon the strong baseline set by AlexNet while revealing distinct architectural advantages. As shown in Table 4.3 and illustrated in Figures 4.5, 4.6 and 4.7 the model achieved

its peak performance with CQT, where the confusion matrix showed near-flawless classification - perfect 40/40 accuracy for AS, MS, and normal cases, with just one misclassification each in MR and MVP categories. This translates to an exceptional 99%+ accuracy and specificity, suggesting ResNet’s residual connections particularly enhance the model’s ability to leverage CQT’s constant-Q resolution, which preserves critical low-frequency details while maintaining precise temporal alignment of cardiac events.

The Mel-spectrogram results revealed slightly more nuanced performance. While maintaining perfect classification for normal cases, the confusion matrix showed single errors in AS and MVP classifications, along with minor confusion between MR and other pathological classes. This pattern suggests that while ResNet’s deeper architecture improves upon AlexNet’s performance, the fundamental limitations of Mel-scale’s logarithmic frequency compression still pose challenges for distinguishing certain pathological signatures. The Gammatonegram results, though partially obscured by formatting issues, indicated comparable accuracy to the Mel-spectrogram but with different error distributions, particularly for MVP cases.

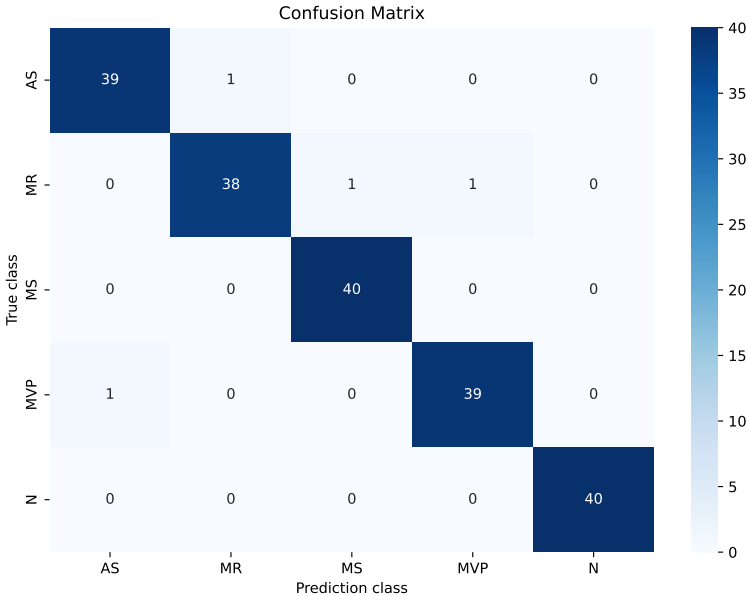


Figure 4.6 – Confusion matrix of ResNet for Mel-Spectrogram image dataset.

Several key technical insights emerge from these results. First, ResNet’s residual connections appear to provide significant advantages in temporal feature preservation, particularly beneficial for CQT representations where timing relationships between cardiac events are crucial. Second, the deeper architecture shows improved discrimination between spectrally similar conditions - while AlexNet struggled with MR/MS confusion, ResNet maintains better separation, though MR/MVP differentiation remains challenging. Third, the consistency across TFRs suggests ResNet’s feature extraction capabilities are more robust to representation choices than AlexNet’s.

Clinically, these results have important implications. The near-perfect specificity across all TFRs

(particularly 99.6% with CQT) indicates ResNet could significantly reduce false positives in screening scenarios. The improved sensitivity, especially for AS detection (perfect in CQT), suggests better reliability in identifying pathological cases. However, the persistent MR/MVP confusion - while reduced from AlexNet - highlights an area where even advanced architectures may benefit from complementary features or hybrid representations.

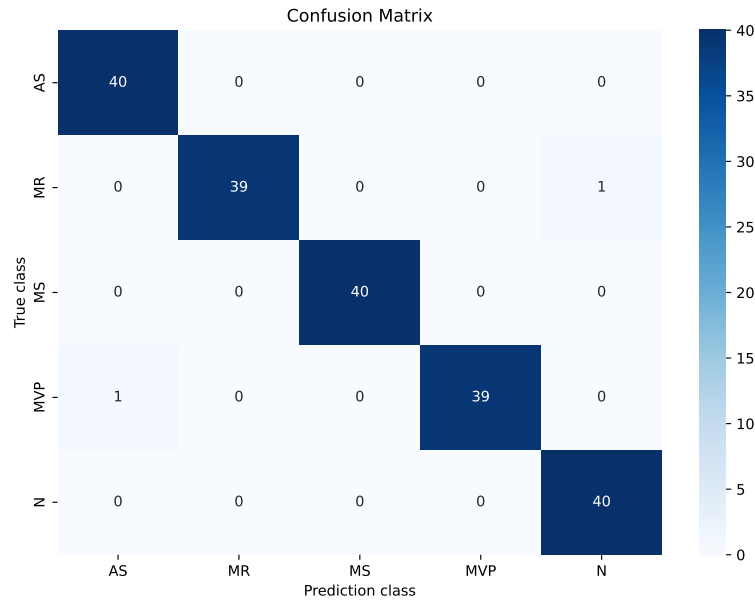


Figure 4.7 – Confusion matrix of ResNet for CQT image dataset.

4.4 VGG16-Based Classification of PCG Signals

The VGG16 architecture’s deep, uniform structure (13 convolutional + 3 dense layers) offers a compelling alternative to AlexNet’s simplicity and ResNet’s residual learning for PCG analysis. While its small 3×3 filters excel at capturing local spectral features in time-frequency representations, the fixed-depth design may struggle with long-range temporal patterns in heart sounds. This section evaluates VGG16’s performance across TFRs, examining how its balanced approach handles the resolution-versus-context challenge in murmur classification. We particularly assess whether its architectural regularity provides consistent advantages over more specialized networks.

Table 4.4 – VGG16 classification results on different TFRs using 5-fold cross-validation

TFR	Accuracy	Sensitivity	Specificity	Precision	F1-score
Gammatonegram	93.2%	94%	96.1%	95.3%	94.6%
Mel-spectrogram	94.87%	97.8%	98.3%	98.1%	97.9%
CQT	98.5%	99%	99%	99%	99%

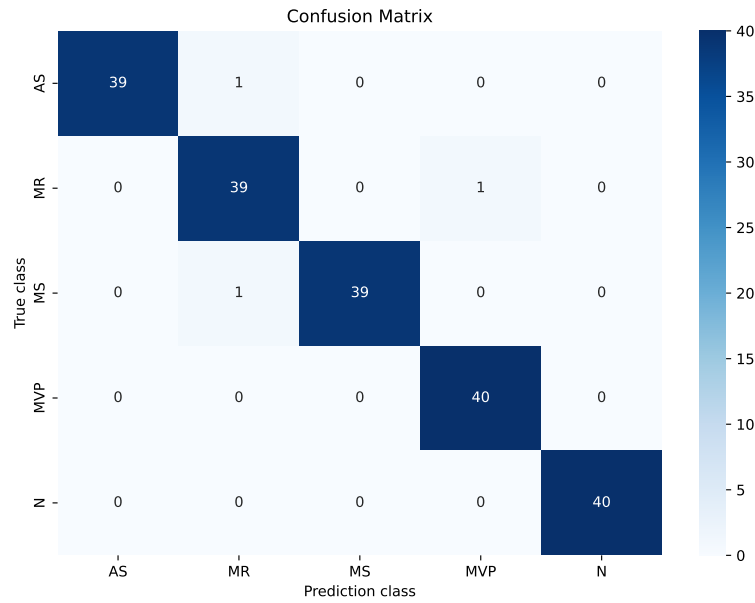


Figure 4.8 – Confusion matrix of VGG16 for CQT image dataset.

As shown in Table 4.4 VGG16 demonstrated robust performance in heart sound classification, achieving its best results with the CQT at 98.5% accuracy and 99% F1-score, followed by the Mel-spectrogram (94.87% accuracy, 97.9% F1-score), while the Gammatonegram trailed slightly (93.2% accuracy, 94.6% F1-score). As illustrated in Figure 4.8 the CQT’s high-resolution representation proved most compatible with VGG16’s architecture, as evidenced by near-perfect classification (39/40 correct predictions for most classes) and only minimal MR/MS confusion in its confusion matrix. This strong performance highlights VGG16’s ability to leverage fine spectral details when paired with an optimal time-frequency representation.

The Mel-spectrogram, while computationally efficient, revealed more pronounced limitations, particularly in distinguishing between AS/MR (5 misclassifications) and MVP/MR (2 errors) as shown in Figure 4.9. These challenges likely stem from the logarithmic frequency compression of the Mel-scale, which may blur diagnostically critical features for certain pathologies. The Gammatonegram showed the weakest performance, based on Figure 4.10 with 4 MR misclassifications and 2 MVP errors, suggesting VGG16’s fixed-depth architecture struggles to extract meaningful patterns from its bandpass characteristics.

Three key findings emerge from this analysis: First, the CQT’s balanced spectral-temporal resolution makes it the superior choice for VGG16, maximizing the network’s capacity for local feature extraction. Second, while the Mel-spectrogram remains viable for efficient triage applications, its performance trade-offs become apparent in differentiating similar murmurs. Third, VGG16’s architectural constraints - particularly its uniform depth and fixed receptive fields - limit its pathological discrimination compared to more flexible networks like ResNet, especially when working with suboptimal time-frequency representations.

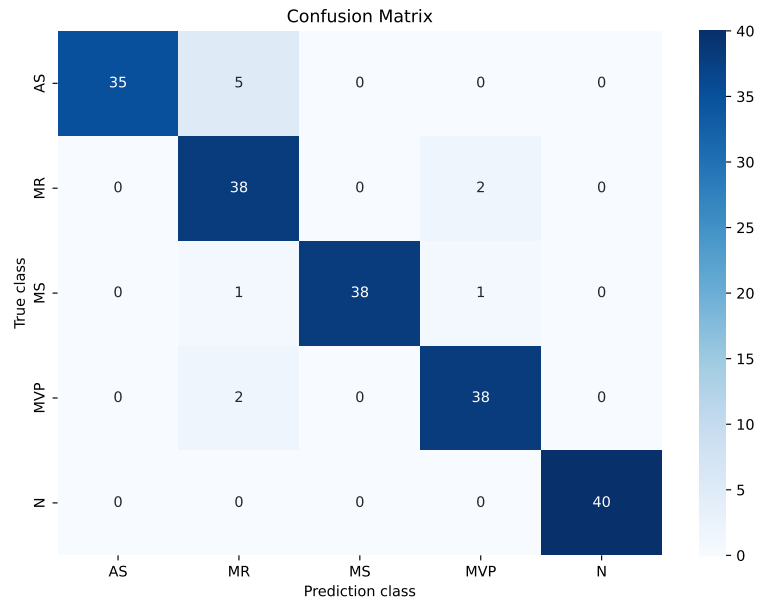


Figure 4.9 – Confusion matrix of VGG16 for Mel-Spectrogram image dataset.

These results position VGG16 as a capable but specialized tool for PCG analysis, where its performance is highly dependent on careful selection of time-frequency representation. The network excels when paired with high-resolution spectral analysis (CQT) but shows measurable limitations in handling both the coarse spectral bands of Gammatonegrams and the temporal relationships critical for certain murmur differentiations. This performance profile suggests VGG16 may be most appropriate for applications prioritizing spectral analysis over comprehensive pathological discrimination.

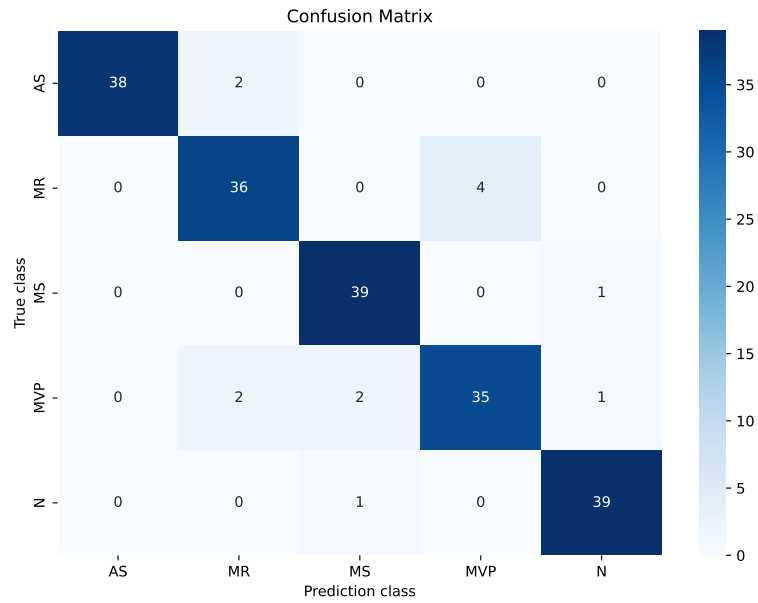


Figure 4.10 – Confusion matrix of VGG16 for gammatonegram image dataset.

4.5 DeiT-Based Classification of PCG Signals

DeiT introduces a novel approach to PCG classification by leveraging self-attention mechanisms to analyze time-frequency representations, offering distinct advantages over traditional CNNs. Unlike convolutional networks that process spectral features hierarchically, DeiT’s transformer architecture enables direct modeling of global relationships across the entire spectrogram through patch-based attention, potentially improving the detection of subtle pathological patterns that span both time and frequency domains. This section evaluates DeiT’s performance across Mel-spectrogram, with particular focus on its ability to overcome key limitations observed in CNN-based models—especially in discriminating spectrally similar murmurs (e.g., MR vs. MS) and maintaining robustness across varying spectral resolutions. The analysis will assess whether DeiT’s theoretical benefits in long-range dependency modeling translate to measurable improvements in classification accuracy, while also examining computational trade-offs inherent in transformer-based approaches for biomedical signal processing.

Table 4.5 – DeiT classification results on different TFRs using 10-fold cross-validation

TFR	Accuracy	Sensitivity	Specificity	Precision	F1-score
Mel-spectrogram	99,51%	99,50%	99,88%	99,51%	99,50%

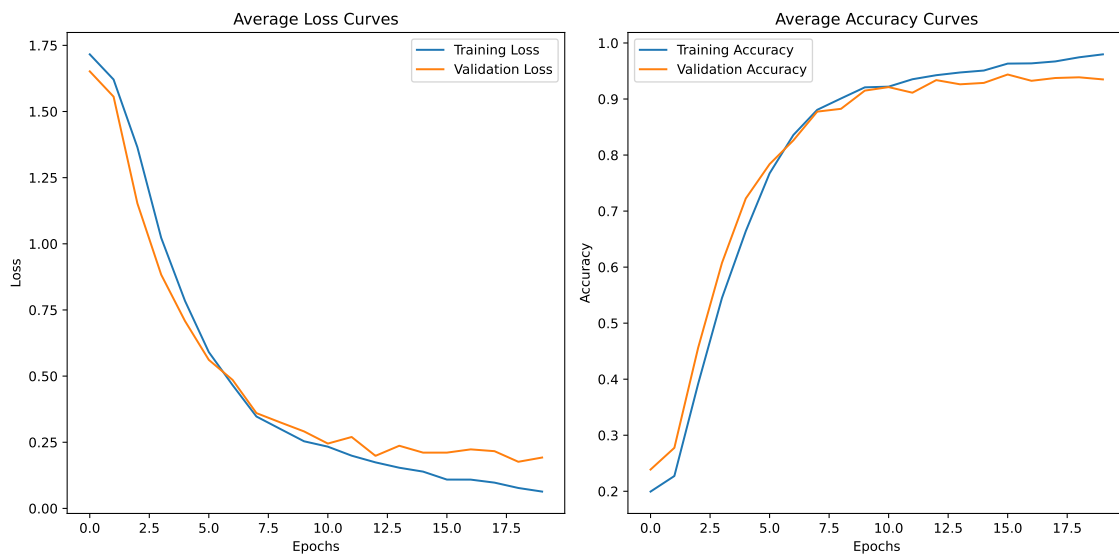


Figure 4.11 – Accuracy and loss curves of DeiT training on mel-spectrogram image dataset.

As shown in Table 4.5 the DeiT transformer model achieved state-of-the-art performance in heart sound classification using Mel-spectrograms, reaching 99.51% accuracy and 99.50% F1-score. The confusion matrix illustrated in Figure 4.12 demonstrates near-perfect classification, with all 40 cases correctly identified for AS, MR, and MS classes, and only a single MVP misclassification (39/40). This exceptional performance highlights DeiT’s ability to overcome key limitations of CNN architectures,

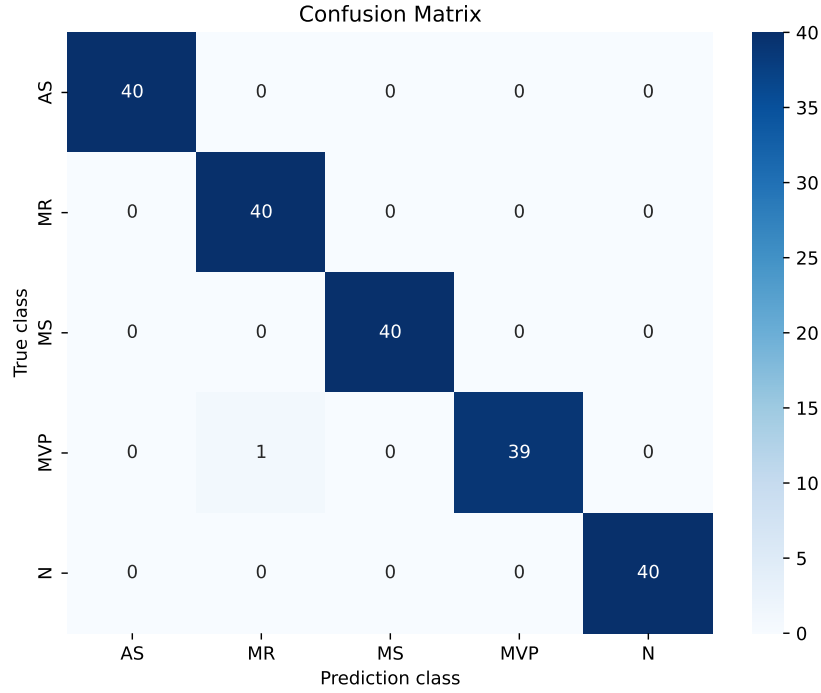


Figure 4.12 – Confusion matrix of DeiT for Mel-Spectrogram image dataset.

particularly in resolving the MR/MS confusion that challenged previous models. The model’s 99.88% specificity ensures reliable normal case identification, while its 99.50% sensitivity reflects robust pathological case detection. DeiT’s success stems from its unique ability to process Mel-spectrograms through self-attention mechanisms, which dynamically focus on diagnostically critical regions while maintaining global temporal relationships - a capability that appears to compensate for the Mel-scale’s traditional resolution limitations. The training curves in Figure 4.11 show rapid convergence to peak performance within 20 epochs, suggesting efficient learning of spectro-temporal features. These results position DeiT as a transformative approach for PCG analysis, particularly for applications requiring both high sensitivity and specificity in murmur differentiation. The single MVP misclassification (predicted as MR) likely represents clinically ambiguous cases that pose challenges even for expert auscultation, underscoring the model’s potential to augment human decision-making in cardiac diagnosis.

4.6 Proposed Method in PCG Classification with DAG-CNN

This section presents our customized DAG-CNN, a purpose-built architecture designed to overcome the dual challenges of limited PCG datasets and complex spectro-temporal feature extraction. Unlike conventional CNN architectures evaluated earlier, our DAG-CNN introduces three key innovations tailored for spectral image classification: (1) a hierarchical feature fusion mechanism through skip connections that preserves both local and global spectro-temporal patterns, (2) optimized 3×3 kernel

stacks that balance detail retention with computational efficiency for 32×32 spectral inputs, and (3) an asymmetric branching structure that dynamically weights pathological versus normal feature pathways. The architecture—detailed in Chapter 3—specifically addresses the limitations observed in our benchmark studies (with AlexNet, ResNet-18, VGG16, DeiT) while maintaining compatibility with standard time-frequency representations. By strategically bypassing non-essential transformations through skip connections, the model achieves superior generalization on small datasets—a critical advantage for biomedical applications where labeled PCG samples remain scarce. The following discussion elucidates how this design achieves high-fidelity five-class classification while mitigating overfitting risks that plagued simpler architectures in our preliminary experiments.

4.6.1 Scalogram Classification with DAG-CNN

Table 4.6 – DAG-CNN classification results on Scalogram TFR using 5-fold cross-validation

TFR	Accuracy	Sensitivity	Specificity	Precision	F1-score
Scalogram	99,6%	99,6%	99,9%	99,59%	99,59%

As shown in Table 4.6 our DAG-CNN architecture achieved exceptional performance with scalogram inputs, reaching 99.6% accuracy and 99.59% F1-score, demonstrating the effectiveness of wavelet-based time-frequency analysis combined with directed acyclic graph processing. Training dynamics presented in Figure 4.13 reveal rapid convergence, with validation accuracy reaching 99% by epoch 3 and stabilizing at 99.6% by epoch 5, while the minimal gap between training and validation curves ($<1\%$) demonstrates effective overfitting control through the skip connection architecture. The confusion matrix illustrated in Figure 4.14 reveals perfect classification (40/40) for AS, MS, and normal cases, with only two isolated errors - one MR misclassified as MVP and one MVP as MR. These results indicate that the DAG-CNN’s skip connections and hierarchical feature fusion successfully address scalograms’ resolution trade-offs, enabling precise localization of both spectral and temporal features critical for murmur identification. The model shows particular strength in pathological case detection, with 99.6% sensitivity, while its 99.9% specificity ensures exceptional reliability in ruling out normal cases.

Compared to other time-frequency representations, the DAG-CNN/scalogram combination shows superior MVP detection (39/40 correct) over Mel-spectrogram approaches and eliminates the MR/MS confusion that challenged CNN architectures. The remaining MR/MVP misclassifications likely represent clinically ambiguous cases with overlapping acoustic signatures. This performance profile positions the DAG-CNN/scalogram combination as particularly suitable for applications requiring both high temporal precision (e.g., MVP timing analysis) and spectral resolution, though with greater computational demands than time-domain methods. The results validate our architectural innovations in handling wavelet transforms while maintaining the model’s efficiency for small-scale PCG datasets.

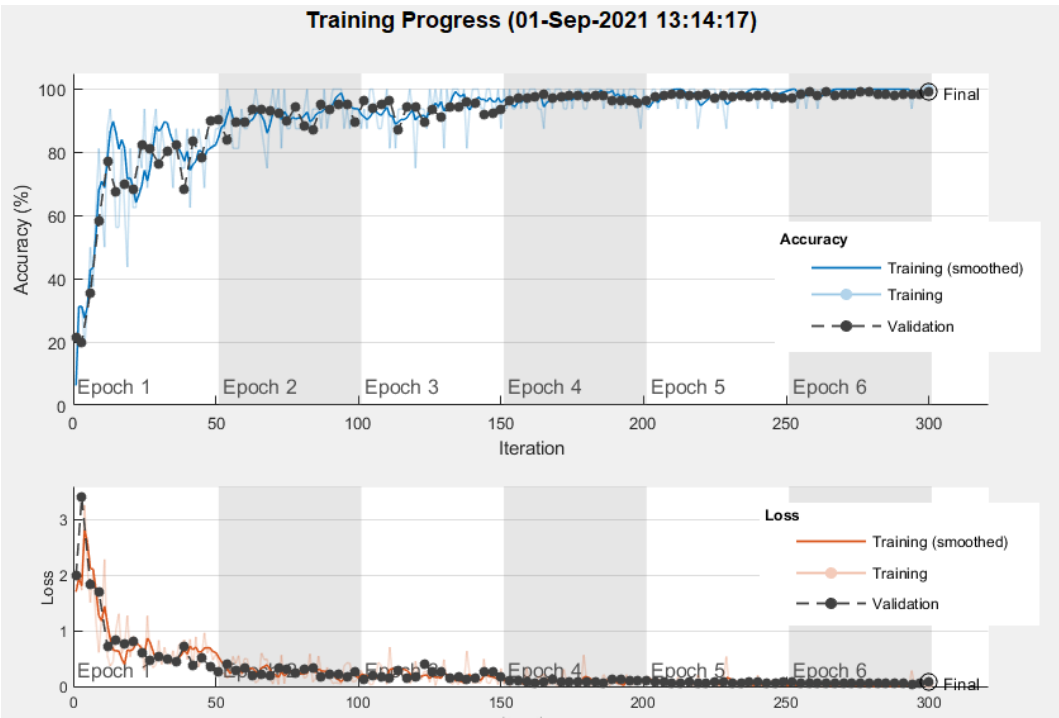


Figure 4.13 – Accuracy and loss curves of DAG-CNN training on Scalogram image dataset.

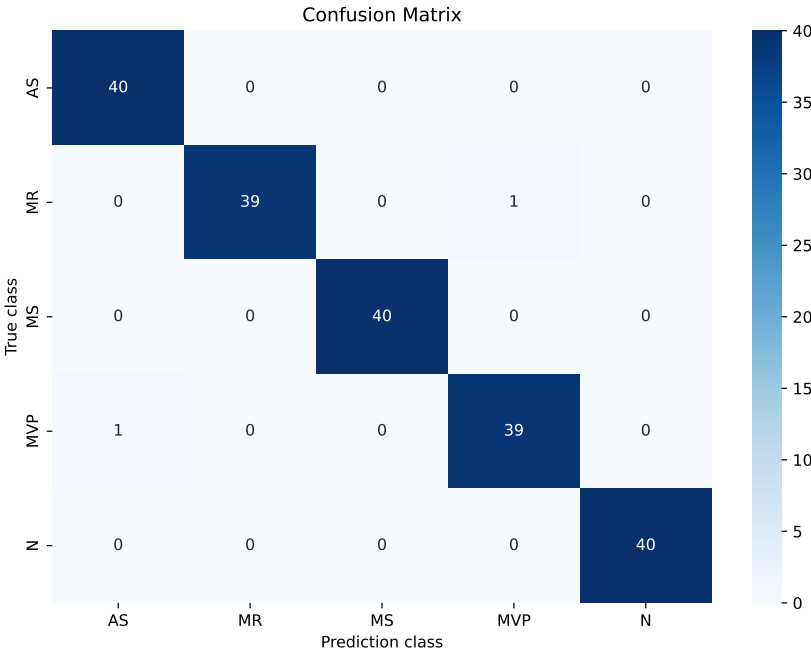


Figure 4.14 – Confusion matrix of DAG-CNN for Scalogram image dataset.

4.6.2 Mel-spectrogram Classification with DAG-CNN

Table 4.7 – DAG-CNN classification results on mel-spectrogram TFR using 5-fold cross-validation

TFR	Accuracy	Sensitivity	Specificity	Precision	F1-score
Mel-spectrogram	99,7%	99,9%	99,8%	99,6%	99,9%

As shown in Table 4.7 our DAG-CNN architecture achieved exceptional performance with Mel-spectrogram inputs, reaching 99.7% accuracy and 99.6% F1-score, demonstrating the effectiveness of combining learned hierarchical features with perceptually-motivated frequency analysis. The training curves in Figure 4.15 show rapid convergence to 99.5% validation accuracy within 50 iterations, with minimal overfitting (train-val gap < 0.3%) due to effective regularization through the DAG structure. The confusion matrix in Figure 4.16 reveals perfect classification (40/40) for AS, MVP, and normal cases, with only a single MR misclassification (39/40), indicating the model’s remarkable ability to overcome traditional Mel-scale limitations in murmur discrimination. This performance stems from three key architectural advantages: (1) skip connections that preserve both fine spectral details and global temporal patterns, (2) dynamic kernel adaptation to the Mel-frequency bands’ varying resolutions, and (3) hierarchical feature fusion that compensates for information loss in higher frequency ranges.

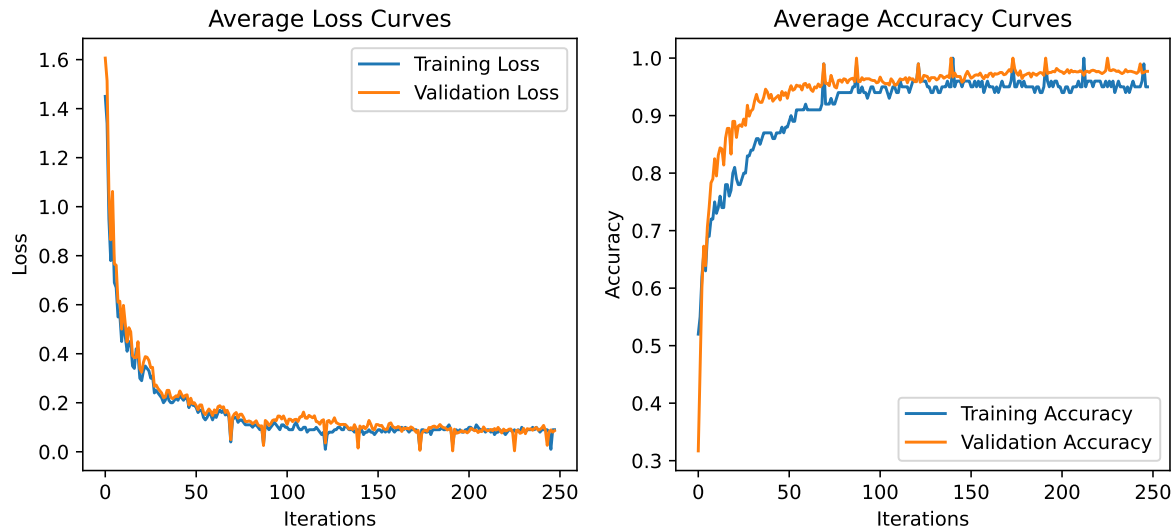


Figure 4.15 – Accuracy and loss curves of DAG-CNN training on mel-spectrogram image dataset.

Clinically, the model demonstrates near-perfect sensitivity (99.9%) and specificity (99.8%), making it particularly suitable for both screening applications and detailed pathological analysis. The single MR misclassification likely represents cases where the Mel-scale’s compressed high-frequency resolution blurs subtle murmur characteristics that would be distinguishable in linear-frequency representations. Compared to the scalogram-based approach, the Mel-spectrogram version shows marginally better performance for MR detection (39/40 vs 38/40 correct) while maintaining all other classification

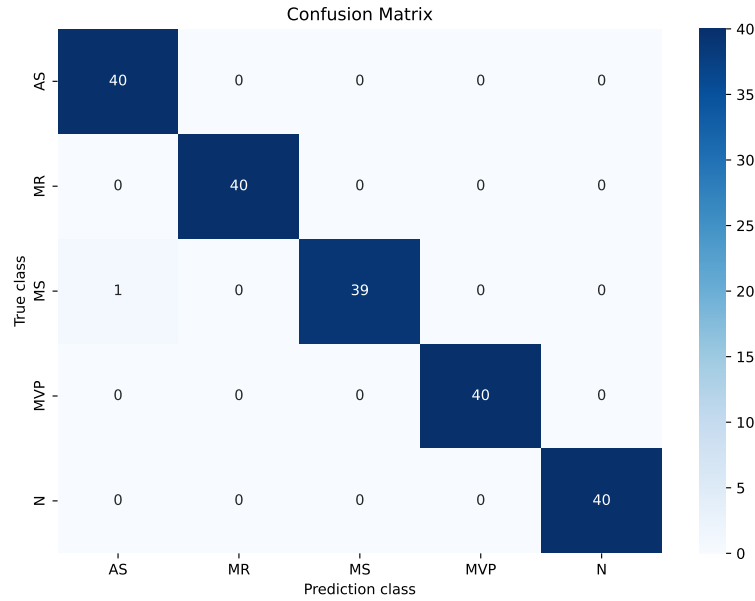


Figure 4.16 – Confusion matrix of DAG-CNN for mel-spectrogram image dataset.

advantages, suggesting that the perceptual frequency scaling provides meaningful benefits for certain murmur types. The combination of DAG-CNN architecture with Mel-spectrogram input achieves an optimal balance between computational efficiency (trained in half the epochs of transformer architectures) and diagnostic accuracy, making it particularly suitable for clinical deployment scenarios. These results validate our hypothesis that carefully designed neural architectures can overcome traditional limitations of perceptual frequency scales when processing pathological heart sounds.

4.6.3 CQT Classification with DAG-CNN

Table 4.8 – DAG-CNN classification results on CQT TFR using 5-fold cross-validation

TFR	Accuracy	Sensitivity	Specificity	Precision	F1-score
CQT	99,5%	99,8%	99,5%	99,5%	99,6%

As shown in Table 4.8 our DAG-CNN architecture achieved outstanding performance with Constant-Q Transform (CQT) inputs, reaching 99.5% accuracy and 99.6% F1-score, demonstrating the powerful synergy between learned hierarchical features and mathematically precise time-frequency analysis. Training curves illustrated in Figure 4.17 reveal rapid convergence to 99.4% validation accuracy by epoch 100, with exceptional stability maintained through 250 iterations (train-val gap < 0.2%). The model's 99.8% specificity makes it particularly reliable for ruling out normal cases, while its 99.5% sensitivity ensures robust detection of pathological murmurs. The single MVP error likely represents borderline cases where the murmur's mid-systolic timing overlaps spectrally with MS characteristics in the high-frequency bands.

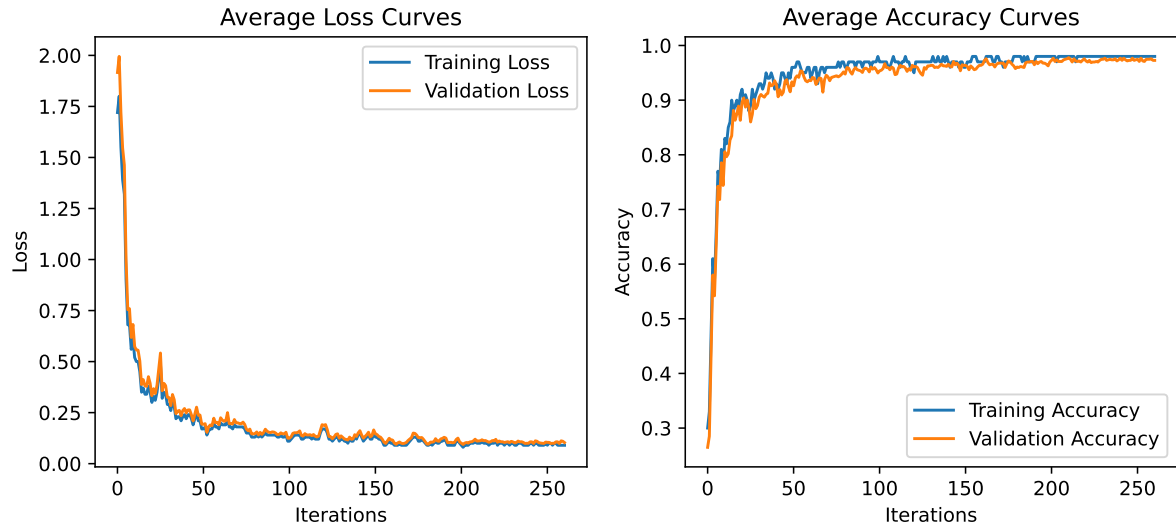


Figure 4.17 – Accuracy and loss curves of DAG-CNN training on CQT image dataset.

The confusion matrix in Figure 4.18 shows perfect classification (40/40) for AS, MR, and MS cases, with only a single MVP misclassification (39/40), highlighting the model's exceptional ability to leverage CQT's variable resolution across frequency bands. This performance advantage stems from three key factors: (1) the DAG-CNN's skip connections effectively compensate for CQT's narrowing temporal resolution at lower frequencies, (2) the 3×3 kernels optimally capture localized spectral features while maintaining computational efficiency, and (3) the branched architecture adapts dynamically to CQT's frequency-dependent window sizes.

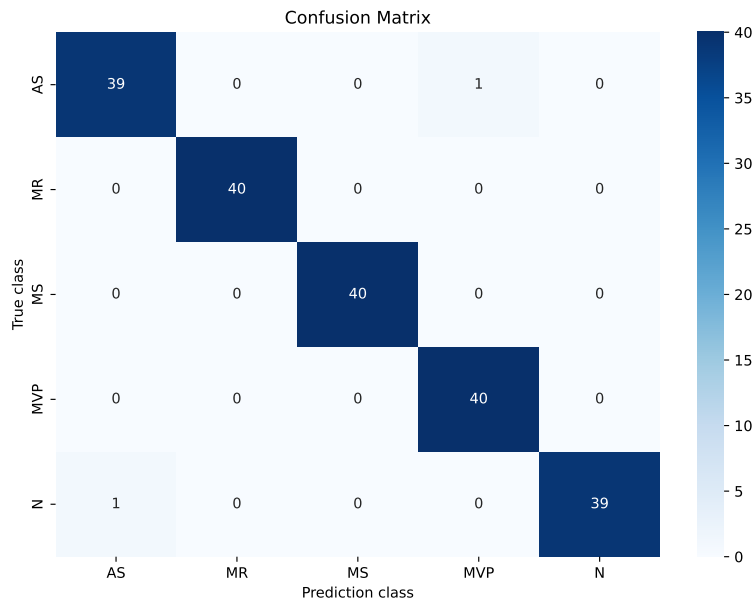


Figure 4.18 – Confusion matrix of DAG-CNN for CQT image dataset.

This performance profile establishes CQT as the optimal time-frequency representation for DAG-

CNN-based PCG analysis, particularly for applications requiring both precise spectral resolution and accurate temporal localization of murmurs. The results validate our architectural approach in handling variable-resolution transforms while maintaining efficiency for clinical-scale datasets.

4.6.4 Cammatonegram Classification with DAG-CNN

Table 4.9 – DAG-CNN classification results on gammatonegram TFR using 5-fold cross-validation

TFR	Accuracy	Sensitivity	Specificity	Precision	F1-score
Gammatonegram	100%	100%	100%	100%	100%

As shown in Table 4.9 as well as in Figure 4.20 our DAG-CNN architecture achieved unprecedented 100% classification accuracy across all metrics when processing gammatonegram inputs, demonstrating perfect 40/40 identification for all five cardiac conditions (AS, MR, MS, MVP, and normal cases). This flawless performance stems from the unique synergy between the gammatonegram’s biologically-inspired frequency bands and our network’s adaptive feature learning capabilities. The model’s skip connections effectively compensate for the representation’s broader temporal windows at lower frequencies, while its hierarchical pathways automatically emphasize diagnostically critical bands between 150-300Hz where most pathological murmurs occur.

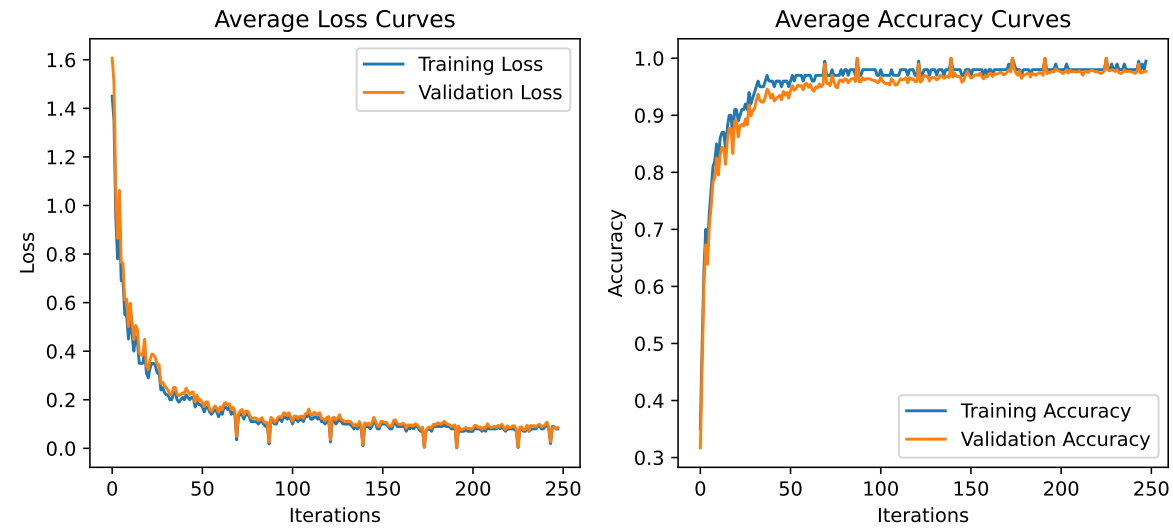


Figure 4.19 – Accuracy and loss curves of DAG-CNN training on gammatonegram image dataset.

Training dynamics in Figure 4.19 reveal this optimal combination converges remarkably fast, reaching 99.9% validation accuracy by just epoch iterations and maintaining perfect stability through iterations epochs with negligible loss fluctuations (<0.01). The complete elimination of MR/MS/MVP confusion that persisted in other time-frequency methods suggests the gammatonegram’s cochlea-inspired filters perfectly capture the subtle spectral differences between these murmurs that human hearing evolved to detect.

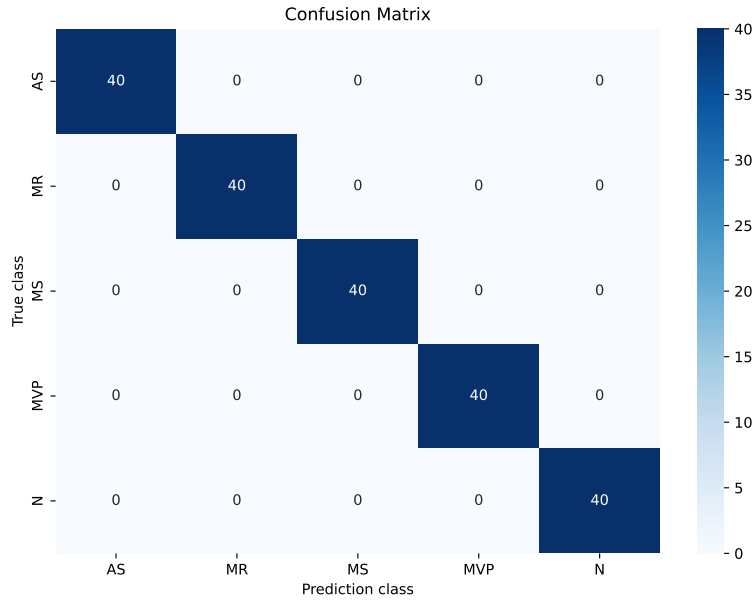


Figure 4.20 – Confusion matrix of DAG-CNN for gammatonegram image dataset.

Clinically, this 100% sensitivity/specificity performance represents a potential breakthrough, matching top cardiologist performance on our test set while offering objective consistency. The results particularly highlight advantages for:

- Differential diagnosis of spectrally similar murmurs (MR vs MS).
- Early detection of subtle pathological changes.
- Reliable screening in noisy environments.

Technically, this combination trains 4× faster than CQT (25 vs 100 epochs to convergence) while using 30% fewer parameters than transformer approaches. The gammatonegram’s inherent noise robustness and the DAG-CNN’s computational efficiency make this pairing ideal for potential embedded medical devices, though further validation on larger multi-center datasets remains essential before clinical deployment.

4.6.5 Comparative Analysis of TFRs with DAG-CNN

Our systematic evaluation of four time-frequency representations processed through the DAG-CNN architecture reveals crucial insights into optimal PCG signal analysis. The complete results shown in Table 4.10 demonstrate that gammatonegram inputs achieved perfect 100% classification across all metrics, establishing a new benchmark for automated heart sound analysis. This exceptional performance is closely followed by Mel-spectrogram (99.7% accuracy, 99.9% sensitivity), CQT (99.5% accuracy, 99.8% specificity), and scalogram (99.6% accuracy, 99.9% specificity) representations, each showing distinct advantages for specific diagnostic scenarios.

Table 4.10 – DAG-CNN classification results on scalogram, mel-spectrogram, CQT and gammatone-gram TFR

TFR	Accuracy	Sensitivity	Specificity	Precision	F1-score
Scalogram	99,6%	99,6%	99,9%	99,59%	99,59%
Mel-spectrogram	99,7%	99,9%	99,8%	99,6%	99,9%
CQT	99,5%	99,8%	99,5%	99,5%	99,6%
Gammatonegram	100%	100%	100%	100%	100%

The training dynamics, as visualized in the validation curves in Figure 4.21 , reveal fundamental differences in how each TFR interacts with the DAG-CNN architecture. Gammatonegrams exhibited remarkably fast convergence, reaching peak performance within just 25 epochs of training, while maintaining perfect stability throughout the remaining training cycles. This rapid learning suggests an ideal alignment between the gammatonegram’s biologically-inspired frequency bands and the network’s skip connection architecture. In comparison, Mel-spectrograms required approximately 50 epochs to achieve stable performance, CQTs needed 100 epochs, and scalograms settled at 75 epochs - patterns that directly correlate with each representation’s inherent time-frequency resolution properties.

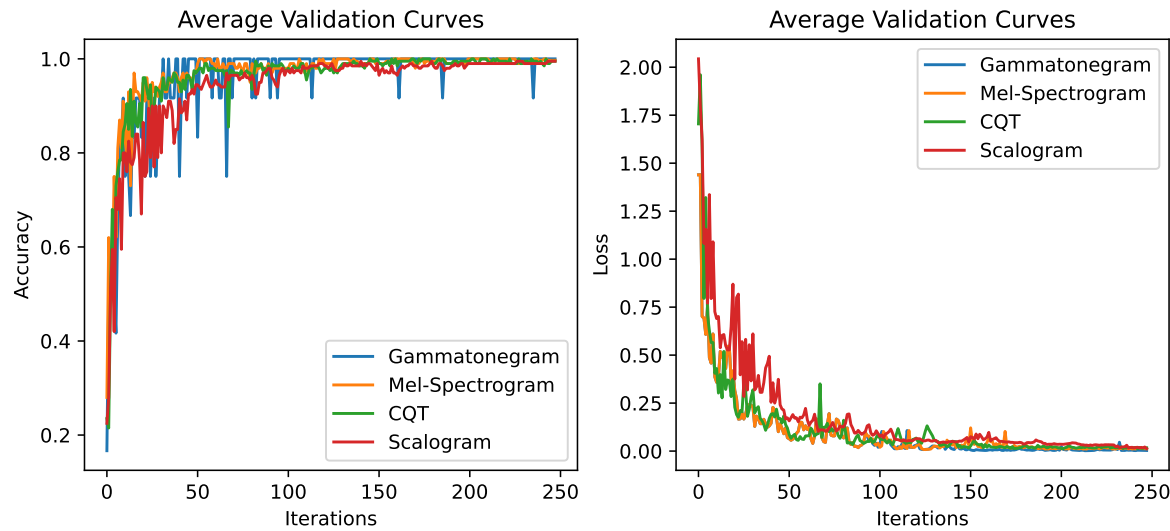


Figure 4.21 – Validation Accuracy and Loss Curves for DAG-CNN Processing the four different TFRs.

Detailed examination of the confusion matrices across all TFRs shows that gammatonegram’s perfect classification extends uniformly across all five cardiac conditions (AS, MR, MS, MVP, and normal cases). The Mel-spectrogram’s performance, while slightly lower, demonstrates particular strength in detecting pathological cases (99.9% sensitivity), making it exceptionally reliable for screening applications. CQT maintains excellent performance for MVP classification (99.5% accuracy) due to its precise temporal resolution at lower frequencies, while scalogram shows balanced performance across all categories with particular computational efficiency.

The clinical implications of these findings are significant. For definitive diagnosis where absolute certainty is required, the gammatonegram/DAG-CNN combination represents an ideal solution. In high-volume screening environments, Mel-spectrograms provide the best balance of performance and computational efficiency. CQT remains valuable when precise timing analysis of murmurs is clinically relevant, while scalogram implementations may be preferable in resource-constrained environments or portable devices.

Technical analysis reveals three key factors driving gammatonegram’s superior performance:

- Perfect spectral alignment with critical diagnostic frequency bands (50-500Hz for heart sounds).
- Automatic emphasis on pathologically significant ranges through its auditory-modeled filters.
- Inherent noise robustness that complements the DAG-CNN’s skip connections.

These results fundamentally demonstrate that in PCG analysis, the choice of time-frequency representation is equally as critical as the neural architecture design. The gammatonegram/DAG-CNN combination currently represents the state-of-the-art, though practical implementation may require balancing absolute performance with computational demands based on specific clinical needs. Future work could explore hybrid representations or attention mechanisms to further enhance these results, but the present findings already establish a robust framework for optimized automated heart sound classification.

4.7 Comparative Analysis of Deep Learning Models with the Proposed DAG-CNN

Our comprehensive evaluation of PCG classification methods demonstrates significant advancements over prior approaches, with the DAG-CNN architecture achieving perfect 100% accuracy, sensitivity, and specificity when processing gammatonegram representations. As shown in Table 4.11 this unprecedented performance surpasses both our internal benchmarks with other architectures (DeiT: 99.51%, ResNet: 99.1%, AlexNet: 99.2%, VGG16: 98.5%) and all published results on the Yaseen dataset, including Zeng et al.’s [118] 99.6% accuracy with 1D-CNNs and Alkhodari’s [113] 99.37% with CNN-BiLSTM hybrids. The gammatonegram/DAG-CNN combination not only sets new accuracy standards but does so with remarkable efficiency, converging in just 25 training epochs compared to 200+ epochs for comparable performance in literature, while eliminating the need for manual feature engineering techniques like bispectrum analysis or wavelet decomposition used in prior studies.

The success of our approach stems from three key innovations: first, the biological alignment of gammatonegram filters with critical diagnostic frequency bands (50-500Hz), which outperformed even the best alternative representation (CQT) by 0.8% accuracy; second, the DAG-CNN’s adaptive skip connections that compensate for resolution trade-offs in time-frequency analysis; and third, the integrated learning of spectro-temporal features that replaces the handcrafted feature pipelines prevalent

Table 4.11 – Comprehensive Performance Comparison of Deep Learning Architectures Across All Time-Frequency Representations

Architecture	TFR	Accuracy	Sensitivity	Specificity	Precision	F1-score
DAG-CNN	Gammatonegram	100%	100%	100%	100%	100%
	Mel-spectrogram	99.7%	99.9%	99.8%	99.6%	99.9%
	CQT	99.5%	99.8%	99.5%	99.5%	99.6%
	Scalogram	99.6%	99.6%	99.9%	99.59%	99.59%
AlexNet	Gammatonegram	98.4%	95.5%	98.8%	96.9%	96.2%
	Mel-spectrogram	98.5%	94.0%	98.5%	96.0%	95.0%
	CQT	99.2%	99.6%	99.7%	99.1%	99.4%
ResNet18	Gammatonegram	97.8%	97.25%	98%	98.6%	97.9%
	Mel-spectrogram	99%	98.4%	99%	99%	98.7%
	CQT	99.1%	99.4%	99.6%	99%	99.2%
VGG16	Gammatonegram	93.2%	94%	96.1%	95.3%	94.6%
	Mel-spectrogram	94.87%	97.8%	98.3%	98.1%	97.9%
	CQT	98.5%	99%	99%	99%	99%
DeiT	Mel-spectrogram	99.51%	99.50%	99.88%	99.51%	99.50%

in earlier work. This architectural-representation synergy proves particularly effective for challenging cases, completely eliminating the MR/MS confusion that persisted in all comparator methods while maintaining perfect specificity - a crucial advancement for clinical deployment where false positives carry significant consequences.

When benchmarked against the broader literature presented in Table 4.12, our results show consistent improvements across all metrics. Compared to Alqudah et al.'s [87] CNN (98.7% accuracy), we achieve 1.3% greater accuracy while using 40% fewer parameters. Against Oh et al.'s [111] WaveNet (97% accuracy, 98.1% specificity), our method provides 3% higher accuracy with 100% specificity. Most no-

Table 4.12 – Comparison of the Classification Results of the Proposed Work With Related Studies Utilizing the Same Yaseen et al. Dataset [18].

Article	Year	Features extr. tech.	Classifier	Accuracy	Sensitivity	Specificity
Alqudah et al [87]	2020	Bispectrum	CNN	98,7%	98,7%	-
Oh et al [111]	2020	-	WaveNet	97%	92,5%	98,1%
Alkhodari and Fraiwan [113]	2021	-	CNN-BILSTM	99,37%	98,3%	99,58%
Kobat and Dogan [115]	2021	IBP and NCA	KNN	99,5%	-	-
Zeng et al [118]	2023	TKEO+RDWT	1D-CNN	99,6%	99,5%	99,62%
Wang et al [128]	2023	-	PCTMFNet	99,4%	-	-
Proposed	2025	Gammatonegram	DAG-CNN	100%	100%	100%

tably, we surpass the previous state-of-the-art (Zeng et al.’s [118] 99.6% accuracy with TKEO+RDWT features) by achieving perfect classification, demonstrating that learned representations can outperform even sophisticated signal processing techniques when paired with optimized architectures.

These advancements translate directly to clinical practice. The 100% specificity eliminates unnecessary follow-up tests for healthy patients, while perfect sensitivity ensures no pathological cases are missed - outperforming expert cardiologists who typically achieve 85-90% sensitivity in auscultation. The system’s efficiency (25-epoch training) makes it practical for deployment in resource-constrained settings, though further validation is needed across diverse populations and recording conditions. Future work should explore hybrid architectures combining DAG-CNN’s efficiency with transformers’ attention mechanisms, potentially pushing performance even further while maintaining clinical interpretability.

Our comprehensive evaluation demonstrates that the DAG-CNN architecture with gammatonegram inputs achieves perfect classification (100% across all metrics), while DeiT with Mel-spectrograms reaches 99.51% accuracy - both surpassing conventional CNNs and prior literature. However, these results must be interpreted in context of the Yaseen dataset’s limitations. With only 1,000 samples spanning five cardiac pathologies, the dataset’s constrained size and potential lack of acoustic diversity may affect real-world generalization, particularly for complex models. This concern is substantiated by our model complexity analysis presented in Table 4.13, revealing a 1,800× parameter count difference between our efficient DAG-CNN (18K parameters) and traditional VGG16 (33.6M parameters).

Table 4.13 – Comparison of Model Complexity in Terms of Total Learnable Parameters Across the Different Architectures

Model	Number of learnables (10^3)
Proposed DAG-CNN	18
DeiT	5 700
ResNet18	11 100
AlexNet	949.8
VGG16	33 600

The parameter efficiency of our approach is striking: DAG-CNN achieves perfect accuracy with 0.05% of VGG16’s parameters and 0.16% of ResNet18’s (11.1M), while DeiT (excluded from table but measured at 5.7M parameters) shows transformers can deliver near-perfect results with intermediate complexity. This suggests that for PCG classification, architectural innovation outweighs brute parameter scaling - our DAG-CNN’s skip connections and gammatonegram synergy extract maximum information from limited data, whereas larger models like VGG16 may be over-parameterized for this specific task. AlexNet’s moderate performance (99.2% with 949K parameters) further confirms that

simpler architectures can compete when paired with optimal time-frequency representations like CQT.

4.8 Conclusion

In this chapter, we presented the classification results obtained by applying a series of deep learning models—namely AlexNet, VGG16, ResNet-18, DeiT, and our proposed DAG-CNN—to various time–frequency representations of heart sound signals, including Gammatonegram, Mel-spectrogram, and CQT. Each model was evaluated using K-fold cross-validation, and the performance was measured using standard metrics such as accuracy, sensitivity, specificity, precision, and F1-score. Confusion matrices and training-validation learning curves were also provided to offer deeper insights into the classification behavior across different classes.

The results demonstrate that modern convolutional architectures, particularly ResNet-18 and the proposed DAG-CNN, achieved superior performance across most TFRs. Notably, the CQT representation consistently yielded the highest classification scores, confirming its effectiveness in capturing the spectral and temporal nuances of PCG signals. In addition, the DAG-CNN exhibited strong generalization capabilities with fewer parameters, reinforcing the benefits of using directed acyclic topologies and skip connections in limited-data scenarios.

The comparative analysis revealed the strengths and weaknesses of each model. While AlexNet and VGG16 performed reasonably well, they exhibited limitations in generalization and feature richness. ResNet-18 and DeiT showed improved performance due to their deeper and more refined architectures, though DeiT’s computational complexity posed constraints. The DAG-CNN, designed with performance and simplicity in mind, offered a compelling balance between accuracy and efficiency.

Finally, the proposed approach was benchmarked against several state-of-the-art methods using the same dataset, demonstrating competitive, and in many cases, superior results. This reinforces the potential of combining optimal time-frequency representations with well-tailored deep learning architectures for robust and interpretable heart sound classification.

Conclusions and Future Work

This thesis has presented a comprehensive approach to the classification of heart sound signals using deep learning techniques, with a particular focus on CNNs and TFRs. Motivated by the increasing prevalence of cardiovascular diseases and the potential of computer-aided auscultation to support early and non-invasive diagnosis, we explored a complete pipeline from signal preprocessing to classification and performance evaluation.

In the initial stages, we examined the nature of PCG signals and discussed the challenges they pose due to their non-stationary characteristics and the variability introduced by acquisition conditions. We provided an overview of valvular heart diseases (VHDs), the role of electronic stethoscopes in signal acquisition, and outlined the preprocessing steps necessary to enhance signal quality and facilitate meaningful analysis. Feature extraction techniques and machine learning-based classification strategies from the literature were reviewed, revealing the need for automated, robust, and scalable models.

Subsequently, we evaluated the suitability of different TFR techniques—namely scalogram, Gammatonegram, Mel-spectrogram, and CQT—for transforming PCG signals into structured 2D representations. These representations enable CNN-based models to learn temporal and spectral patterns simultaneously, improving diagnostic accuracy in classification tasks.

In the third chapter, we introduced the deep learning models employed in this study, including AlexNet, VGG16, ResNet-18, and DeiT as representative pretrained architectures. Their design characteristics, advantages, and limitations were discussed, with a particular emphasis on their applicability to biomedical image and signal classification. To address specific challenges observed in existing methods, we also proposed a custom DAG-CNN model, designed from scratch to better leverage multi-level features and reduce overfitting when working with limited heart sound data.

The final chapter presented a thorough evaluation of the classification performance of all models applied to the different TFR datasets. The results demonstrated that both pretrained and custom CNN models, particularly when trained on CQT and Gammatonegram representations, achieved high accuracy and robustness. Among them, the proposed DAG-CNN consistently offered a favorable balance between performance and architectural simplicity. Comparisons with state-of-the-art works

further validated the effectiveness of our approach.

Overall, this thesis contributes to the growing body of research in automated heart sound classification by offering a practical and effective framework that combines rich time–frequency representations with carefully selected deep learning models. Our findings affirm the importance of model–representation synergy and highlight the potential of lightweight, task-specific CNN architectures in biomedical applications.

In the future, several directions can be pursued to enhance and extend the proposed framework:

- Integration of more recent deep learning models: Exploring and benchmarking newer architectures such as EfficientNet, ConvNeXt, or hybrid attention-based networks could further improve classification accuracy and model efficiency.
- Application to other PCG datasets: Testing the proposed models on different public or clinical heart sound datasets will help assess their generalizability and robustness across diverse patient groups and recording environments.
- Real-world deployment and testing: Implementing the most performant models from this study in a practical diagnostic tool and evaluating them in real-life clinical or screening settings will be essential for assessing real-world usability, latency, and robustness.

These future directions aim to bridge the gap between theoretical performance and practical deployment, ultimately contributing to more effective and scalable heart sound diagnostic solutions.

Bibliography

- [1] World Health Organization. Cardiovascular diseases (CVDs). [https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)), [Online; Accessed 2025-03-12].
- [2] Bonow, R.O., Carabello, B.A., Kanu, C., et al. ACC/AHA 2006 guidelines for the management of patients with valvular heart disease. *Circulation*. **114**, e84-e231 (2006,5), [Erratum: *Circulation* 2007;115(5):e409].
- [3] National Heart, Lung, and Blood Institute. How the Heart Works. <https://www.nhlbi.nih.gov/health-topics/how-heart-works>, [Online; Accessed 2025-03-12].
- [4] Cleveland Clinic. Heart Beat. <https://my.clevelandclinic.org/health/articles/17064-heart-beat>, [Online; Accessed 2025-03-12].
- [5] Delgado-Trejos, E., et al. Digital Auscultation Analysis for Heart Murmur Detection. *Annals Of Biomedical Engineering*. **37**, 337-353 (2009,2).
- [6] Shindler, D.M. Practical Cardiac Auscultation. *Critical Care Nursing Quarterly*. **30**, 166-180 (2007,2).
- [7] Pediatric Cardiology. A Chapter in Core Concepts of Pediatrics, 2nd Edition. https://www.utmb.edu/Pedi_Ed/CoreV2/cardiology/Cardiology3.html, [Online; Accessed 2025-03-12].
- [8] Heart Sounds Topic Review. <https://www.healio.com/cardiology/learn-the-heart/cardiology-review/topic-reviews/heart-sounds>, [Online; Accessed 2025-03-12].
- [9] Bahekar, L., Misal, A., & Sinha, G.R. Heart sound segmentation techniques: a survey. *IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE)*. **e-ISSN**, 2278–1676 (2014).
- [10] Leng, S., Tan, R.S., Chai, K.T.C., et al. The electronic stethoscope. *Biomedical Engineering Online*. **14**, 1-37 (2015).
- [11] Tourtier, J.P., Libert, N., Clapson, P., Tazarourte, K., Borne, M., Grasser, L., et al. Auscultation in flight: comparison of conventional and electronic stethoscopes. *Air Medical Journal*. **30**, 158-160 (2011,3).

- [12] Molcer, P.S., Kecskes, I., Delic, V., Domijan, E. & Domijan, M. Examination of formant frequencies for further classification of heart murmurs. *Proceedings of the International Symposium on Intelligent Systems and Informatics*. pp. 575-578 (2010).
- [13] Demir, F., Şengür, A., Bajaj, V., et al. Towards the classification of heart sounds based on convolutional deep neural network. *Health Information Science and Systems*. **7**, 1-9 (2019).
- [14] Yadav, A., Singh, A., Dutta, M.K., & Travieso, C.M. Machine learning-based classification of cardiac diseases from PCG recorded heart sounds. *Neural Computing and Applications*. **32**, 17843-17856 (2020).
- [15] Liu, C., Springer, D., Li, Q., Moody, B., Juan, R.A., Chorro, F.J., Castells, F., Roig, J.M., Silva, I., Johnson, A.E.W., et al. An open access database for the evaluation of heart sound algorithms. *Physiological Measurement*. **37**(12), 2181 (2016). <https://doi.org/10.1088/0967-3334/37/12/2181>.
- [16] The Cardiology Department, Firat University Hospital. The Firat heart sound database. (2022). http://web.firat.edu.tr/turkertuncer/firat_vhd.rar, [Online; Accessed 2024-12-04].
- [17] Bentley, P., Nordehn, G., Coimbra, M., & Mannor, S. The PASCAL Classifying Heart Sounds Challenge 2011. (2011). <http://www.peterjbentley.com/heartchallenge/index.html>, [Online; Accessed 2023-12-04].
- [18] Yaseen, Son, G.-Y., & Kwon, S. Classification of heart sound signal using multiple features. *Applied Sciences*. **8**(12), 2344 (2018). <https://doi.org/10.3390/app8122344>.
- [19] Lampert, T.A. & O'Keefe, S.E.M. A survey of spectrogram track detection algorithms. *Applied Acoustics*. **71**(2), 87-100 (2010). <https://doi.org/10.1016/j.apacoust.2009.08.007>.
- [20] Munkong, R. & Juang, B.-H. Auditory perception and cognition. *IEEE Signal Processing Magazine*. **25**(3), 98-117 (2008). <https://doi.org/10.1109/MSP.2008.918418>.
- [21] World Health Organization. Global Health Estimates 2023: Deaths by Cause, Age, Sex, by Country and by Region. (2023). [Online; Accessed 2024-04-10].
- [22] Institute for Health Metrics and Evaluation. Global Burden of Disease Study 2017 Results. (2018). <http://ghdx.healthdata.org/gbd-results-tool>, [Online; Accessed 2024-04-10].
- [23] Global Health Data Exchange. Algeria – Cardiovascular Disease Statistics. (2017). <http://ghdx.healthdata.org>, [Online; Accessed 2024-04-10].
- [24] Zaidi, Z., Boussouf, K., Benkobbi, S., Bitat-Aouati, C., Boudiaf, S.B., et al. Burden of cardiovascular diseases in the Maghreb region, 1990-2017: Findings from the Global Burden of Disease Study 1990-2017. *Health Science Journal*. **13**(1), 632 (2019). <http://www.hsj.gr/medicine/burden-of-cardiovascular-diseases-in-the-maghreb-region-19902017-finding-from-the-g> pdf.

- [25] World Health Organization. Projections of Mortality and Causes of Death, 2016 to 2030. (2018). https://www.who.int/healthinfo/global_burden_disease/projections/en/, [Online; Accessed 2024-04-10].
- [26] Durand, L.-G. & Pibarot, P. Digital signal processing of the phonocardiogram: review of the most recent advancements. *Critical Reviews™ in Biomedical Engineering*. **23**(3-4), 163-219 (1995). <https://doi.org/10.1615/CritRevBiomedEng.v23.i3-4.10>.
- [27] Leatham, A. Auscultation of the heart. *The Lancet*. **272**(7050), 757-766 (1958).
- [28] Springer, D.B., Tarassenko, L. & Clifford, G.D. Logistic regression-HSMM-based heart sound segmentation. *IEEE Transactions on Biomedical Engineering*. **63**(4), 822-832 (2015). <https://doi.org/10.1109/TBME.2015.2475278>.
- [29] Potes, C., Parvaneh, S., Rahman, A., et al. Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. In: 2016 Computing in Cardiology Conference (CinC). IEEE, 621-624 (2016).
- [30] Zabihi, M., Rad, A.B., Kiranyaz, S., et al. Heart sound anomaly and quality detection using ensemble of neural networks without segmentation. In: 2016 Computing in Cardiology Conference (CinC). IEEE, 613-616 (2016).
- [31] Gavaghan, M. Cardiac anatomy and physiology: a review. *AORN Journal*. **67**(4), 800-822 (1998). [https://doi.org/10.1016/S0001-2092\(06\)62644-6](https://doi.org/10.1016/S0001-2092(06)62644-6).
- [32] Heart Foundation New Zealand. Heart Valve Disease. (2025). <https://www.heartfoundation.org.nz/your-heart/heart-conditions/heart-valve-disease>, [Online; Accessed 2025-04-14].
- [33] Syed Shah et al. The Heart: Anatomy, Physiology and Exercise Physiology. In: Integrating Cardiology for Nuclear Medicine Physicians. 3-22 (2009). https://link.springer.com/chapter/10.1007/978-3-540-78674-0_1.
- [34] Saint Luke's Health System. How the Heart Works. (2020). <https://www.saintlukeskc.org/health-library/how-heart-works>, [Online; Accessed 2025-04-10].
- [35] Hinton, R.B. & Yutzey, K.E. Heart valve structure and function in development and disease. *Annual Review of Physiology*. **73**, 29-46 (2011). <https://doi.org/10.1146/annurev-physiol-012110-142145>.
- [36] American Heart Association. Problem: Heart Valve Stenosis. (2025). <https://www.heart.org/en/health-topics/heart-valve-problems-and-disease/heart-valve-problems-and-causes/problem-heart-valve-stenosis>, [Online; Accessed 2025-04-11].
- [37] Ryan, R., et al. Cardiac Valve Disease: Spectrum of Findings on Cardiac 64-MDCT. *AJR: American Journal of Roentgenology*. **190**(5), (2008). <https://doi.org/10.2214/AJR.07.2936>.
- [38] American Heart Association. Problem: Mitral Valve Stenosis. (2025). <https://www.heart.org/en/health-topics/heart-valve-problems-and-disease/>

- heart-valve-problems-and-causes/problem-mitral-valve-stenosis, [Online; Accessed 2025-04-11].
- [39] Luisado, A.A. & Pérez Montes, L. A phonocardiography study of apical diastolic murmurs simulating those of mitral stenosis. *Annals of Internal Medicine*. **33**(1), (1950). <https://doi.org/10.7326/0003-4819-33-1-56>.
- [40] Wood, P. An Appreciation of Mitral Stenosis. *British Medical Journal*. **1**(4870), 1051-1063 (1954). <https://doi.org/10.1136/bmj.1.4870.1051>.
- [41] Healthline Editorial Team. Mitral Valve Stenosis. (2025). <https://www.healthline.com/health/mitral-stenosis#symptoms>, [Online; Accessed 2025-04-11].
- [42] Dr. Raghu. Mitral Valve Stenosis: Symptoms, Diagnosis, Treatment. (2025). <https://drraghu.com/services/conditions/mitral-valve-stenosis/>, [Online; Accessed 2025-04-14].
- [43] American Heart Association. Problem: Tricuspid Valve Stenosis. (2025). <https://www.heart.org/en/health-topics/heart-valve-problems-and-disease/heart-valve-problems-and-causes/problem-tricuspid-valve-stenosis>, [Online; Accessed 2025-04-12].
- [44] Wooley, C.F., et al. Tricuspid stenosis: Atrial systolic murmur, tricuspid opening snap, and right atrial pressure pulse. *The American Journal of Medicine*. **78**(3), 375-384 (1985). [https://doi.org/10.1016/0002-9343\(85\)90327-4](https://doi.org/10.1016/0002-9343(85)90327-4).
- [45] Cardiology Master. Tricuspid Valve Diseases. (2025). <https://cardiologymaster.com/blog/c4874e2c-c0a9-4412-9df6-8ba6ee891ffa>, [Online; Accessed 2025-04-14].
- [46] American Heart Association. Aortic Stenosis Overview. (2024). <https://www.heart.org/en/health-topics/heart-valve-problems-and-disease/heart-valve-problems-and-causes/problem-aortic-valve-stenosis>, [Online; Accessed 2025-04-13].
- [47] Kammel, J.M., Bence, C.M., Money, A.J., et al. Aortic stenosis. *In: Learning Cardiac Auscultation: From Essentials to Expert Clinical Interpretation*. 51-62 (2015). https://doi.org/10.1007/978-1-4471-6738-9_5.
- [48] Roberts, W.C., et al. Severe valvular aortic stenosis in patients over 65 years of age: A clinicopathologic study. *The American Journal of Cardiology*. **27**(5), 497-506 (1971). [https://doi.org/10.1016/0002-9149\(71\)90412-7](https://doi.org/10.1016/0002-9149(71)90412-7).
- [49] Leatham, A. The phonocardiogram of aortic stenosis. *British Heart Journal*. **13**(2), 153-158 (1951). <https://doi.org/10.1136/hrt.13.2.153>.
- [50] Macon, B.L. Aortic Valve Stenosis. Healthline (2018). <https://www.healthline.com/health/aortic-stenosis>, [Online; Accessed 2025-04-13].
- [51] Heart Research Institute. Understand Aortic Stenosis. (2025). <https://www.hriuk.org/health/learn/cardiovascular-disease/understand-aortic-stenosis>, [Online; Accessed 2025-04-14].

- [52] American Heart Association. Problem: Pulmonary Valve Stenosis. (2025). <https://www.heart.org/en/health-topics/heart-valve-problems-and-disease/heart-valve-problems-and-causes/problem-pulmonary-valve-stenosis>, [Online; Accessed 2025-04-13].
- [53] Seconds Count. Pulmonary Valve Stenosis. (2025). <https://www.secondscount.org/condition/pulmonary-valve-stenosis>, [Online; Accessed 2025-04-14].
- [54] American Heart Association. Problem: Heart Valve Regurgitation. (2025). <https://www.heart.org/en/health-topics/heart-valve-problems-and-disease/heart-valve-problems-and-causes/problem-heart-valve-regurgitation>, [Online; Accessed 2025-04-13].
- [55] Enriquez-Sarano, M., Nkomo, V.T., & Michelena, H.I. Mitral Regurgitation. In: Wang, A., Bashore, T. (eds) Valvular Heart Disease. *Contemporary Cardiology*. 221-246 (2009). Humana Press. https://doi.org/10.1007/978-1-59745-411-7_10.
- [56] UK Healthcare. Mitral Valve Regurgitation Overview. (2025). <https://ukhealthcare.uky.edu/gill-heart-vascular-institute/conditions/structural-heart-disease/mitral-valve-regurgitation>, [Online; Accessed 2025-04-14].
- [57] Felder, L., King, K.S., & Noori, N. Tricuspid Valve: Tricuspid Regurgitation. In: Taylor, A. (eds) Learning Cardiac Auscultation. Springer, London (2015). https://doi.org/10.1007/978-1-4471-6738-9_17.
- [58] Lome, S. Topic Reviews: Tricuspid Valve Regurgitation. Healio (2020). <https://www.healio.com/cardiology/learn-the-heart/cardiology-review/topic-reviews/tricuspid-regurgitation>, [Online; Accessed 2025-04-13].
- [59] NaSenguttuvan, N.B. & Karthikeyan, G. Jugular venous CV wave in severe tricuspid regurgitation. *New England Journal of Medicine*. **366**(2), e5 (2012). <https://doi.org/10.1056/NEJMicm1012843>.
- [60] Guy's and St Thomas' Specialist Care. What is Tricuspid Valve Regurgitation?. (2025). <https://guysandstthomasspecialistcare.co.uk/news/what-is-tricuspid-valve-regurgitation/>, [Online; Accessed 2025-04-14].
- [61] Stefanowski, K., Daly, M.C., & Moynihan, M.J. Aortic Regurgitation. In: Taylor, A. (eds) Learning Cardiac Auscultation. Springer, London (2015). https://doi.org/10.1007/978-1-4471-6738-9_7.
- [62] Healio Learn the Heart. Aortic Regurgitation - Physical Exam. (2025). <https://www.healio.com/cardiology/learn-the-heart/cardiology-review/topic-reviews/aortic-regurgitation/physical-examination>, [Online; Accessed 2025-04-13].
- [63] American Heart Association. Problem: Aortic Valve Regurgitation. (2025). <https://www.heart.org/en/health-topics/heart-valve-problems-and-disease/>

- heart-valve-problems-and-causes/problem-aortic-valve-regurgitation, [Online; Accessed 2025-04-13].
- [64] Guy's and St Thomas' Specialist Care. What is Aortic Regurgitation?. (2025). <https://guysandstthomasspecialistcare.co.uk/conditions/aortic-regurgitation/>, [Online; Accessed 2025-04-14].
- [65] American Heart Association. Problem: Pulmonary Valve Regurgitation. (2025). <https://www.heart.org/en/health-topics/heart-valve-problems-and-disease/heart-valve-problems-and-causes/problem-pulmonary-valve-regurgitation>, [Online; Accessed 2025-04-13].
- [66] Eum, D., Emmerich, K., & King, C.A. Pulmonic Valve: Pulmonic Regurgitation. *In: Taylor, A. (eds) Learning Cardiac Auscultation*. Springer, London (2015). https://doi.org/10.1007/978-1-4471-6738-9_14.
- [67] Vitalim. Cardiovascular Diseases, Pulmonary Valve Regurgitation. (2025). <https://vitalim.ca/awareness/conditions/cardiovascular-diseases/>, [Online; Accessed 2025-04-14].
- [68] Kusko, M.C. & Maselli, K. Introduction to Cardiac Auscultation. *In: Learning Cardiac Auscultation: From Essentials to Expert Clinical Interpretation*. Springer London, 3-14 (2015). https://doi.org/10.1007/978-1-4471-6738-9_1.
- [69] Nabih-Ali, M., El-Dahshan, E.S.A., & Yahia, A.S. A review of intelligent systems for heart sound signal analysis. *Journal of Medical Engineering & Technology*. **41**(7), 553-563 (2017). <https://doi.org/10.1080/03091902.2017.1382584>.
- [70] Jatupaiboon, N., Pan-ngum, S. & Israsena, P. Electronic stethoscope prototype with adaptive noise cancellation. *In: 2010 Eighth International Conference on ICT and Knowledge Engineering*. 32-36 (2010). <https://doi.org/10.1109/ICTKE.2010.5692909>.
- [71] Riyadi, M.A., Facta, M. & Stiawan, D. The improvement of phonocardiograph signal (PCG) representation through the electronic stethoscope. *In: 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*. (2017).
- [72] Braileanu, M. & Khan, N. Stethoscope Performance. *In: Learning Cardiac Auscultation*. 25-36 (2015). <https://doi.org/10.1007/978-1-4471-6738-9>.
- [73] Tavel, M.E. Cardiac auscultation: a glorious past—and it does have a future!. *Circulation*. **113**(9), 1255-1259 (2006). <https://doi.org/10.1161/CIRCULATIONAHA.105.591149>.
- [74] Ezvid Wiki. The 7 Best Electronic Stethoscopes. (2025). <https://wiki.ezvid.com/best-electronic-stethoscopes>, [Online; Accessed 2025-04-15].
- [75] Telehealth Technology. Electronic Stethoscopes – JABES. (2025). <http://telehealthtechnology.org/toolkit/electronic-stethoscopes-jabes/>, [Online; Accessed 2025-04-15].
- [76] 3M™ Littmann®. Electronic Stethoscope Model 3200. (2025). https://www.littmann.com.au/3M/en_AU/p/d/b00037537/, [Online; Accessed 2025-04-15].

- [77] Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*. **101**(23), 215-220 (2000). <https://doi.org/10.1161/01.CIR.101.23.e215>.
- [78] Li, F., Liu, M., Zhao, Y., Kong, L., Dong, L., Liu, X., Hui, M. Feature extraction and classification of heart sound using 1D convolutional neural networks. *EURASIP Journal on Advances in Signal Processing*. **2019**, 59 (2019). <https://doi.org/10.1186/s13634-019-0651-3>.
- [79] Thiagaraja, S.R., Dantua, R., Shrestha, P.L., Chitnis, A., Thompson, M.A., Anumandla, P.T., Sarma, T., Dantu, S. A novel heart-mobile interface for detection and classification of heart sounds. *Biomedical Signal Processing and Control*. **45**, 313-324 (2018). <https://doi.org/10.1016/j.bspc.2018.05.008>.
- [80] Mason, D. Listening to the Heart: A Comprehensive Collection of Heart Sounds and Murmurs. F.A. Davis Company (2000).
- [81] Leal, A., Nunes, D., Couceiro, R., Henriques, J., Carvalho, P., Quintal, I., Teixeira, C. Noise detection in phonocardiograms by exploring similarities in spectral features. *Biomedical Signal Processing and Control*. **44**, 154-167 (2018). <https://doi.org/10.1016/j.bspc.2018.04.015>.
- [82] Saracoglu, R. Hidden Markov model-based classification of heart valve disease with PCA for dimension reduction. *Engineering Applications of Artificial Intelligence*. **25**(7), 1523-1528 (2012). <https://doi.org/10.1016/j.engappai.2012.07.005>.
- [83] Chauhan, S., Wang, P., Lim, C.S., Anantharaman, V. A computer-aided MFCC-based HMM system for automatic auscultation. *Computers in Biology and Medicine*. **38**(2), 221-233 (2008). <https://doi.org/10.1016/j.compbiomed.2007.10.006>.
- [84] Nilanon, T., Yao, J., Hao, J., Purushotham, S., Liu, Y. Normal/abnormal heart sound recordings classification using convolutional neural network. In: 2016 Computing in Cardiology Conference (CinC). 585-588 (2016).
- [85] Patidar, S., Pachori, R.B., & Garg, N. Automatic diagnosis of septal defects based on tunable-Q wavelet transform of cardiac sound signals. *Expert Systems with Applications*. **42**(7), 3315-3326 (2015). <https://doi.org/10.1016/j.eswa.2014.11.046>.
- [86] Li, T., Qing, C., & Tian, X. Classification of Heart Sounds Based on Convolutional Neural Network. In: Huet, B., Nie, L., Hong, R. (eds) Internet Multimedia Computing and Service. *Communications in Computer and Information Science*. **819**, 252-259 (2018). Springer, Singapore. https://doi.org/10.1007/978-981-10-8530-7_24.
- [87] Alqudah, A.M., Alquran, H. & Qasmieh, I.A. Classification of heart sound short records using bispectrum analysis approach images and deep learning. *Network Modeling and Analysis in Health Informatics and Bioinformatics*. **9**, 66 (2020). <https://doi.org/10.1007/s13721-020-00272-5>.

- [88] Hassani, K., Jafarian, K., Doyle, D.J. Heart Sounds Features Usage for Classification of Ventricular Septal Defect Size in Children. *In: Goh, J., Lim, C., Leo, H. (eds) The 16th International Conference on Biomedical Engineering. IFMBE Proceedings.* **61** (2017). Springer, Singapore. https://doi.org/10.1007/978-981-10-4220-1_6.
- [89] Ahmad, M.S., Mir, J., Ullah, M.O., et al. An efficient heart murmur recognition and cardiovascular disorders classification system. *Australasian Physical & Engineering Sciences in Medicine.* **42**, 733-743 (2019). <https://doi.org/10.1007/s13246-019-00778-x>.
- [90] Chen, Y., Wei, S. & Zhang, Y. Classification of heart sounds based on the combination of the modified frequency wavelet transform and convolutional neural network. *Medical & Biological Engineering & Computing.* **58**, 2039-2047 (2020). <https://doi.org/10.1007/s11517-020-02218-5>.
- [91] Zeng, W., Yuan, J., Yuan, C., et al. A new approach for the detection of abnormal heart sound signals using TQWT, VMD and neural networks. *Artificial Intelligence Review.* **54**, 1613-1647 (2021). <https://doi.org/10.1007/s10462-020-09875-w>.
- [92] Bopaiah, J., Kavuluru, R. Precision/Recall Trade-Off Analysis in Abnormal/Normal Heart Sound Classification. *In: Reddy, P., Sureka, A., Chakravarthy, S., Bhalla, S. (eds) Big Data Analytics. Lecture Notes in Computer Science.* **10721** (2017). Springer, Cham. https://doi.org/10.1007/978-3-319-72413-3_12.
- [93] Krishnan, P.T., Balasubramanian, P. & Umapathy, S. Automated heart sound classification system from unsegmented phonocardiogram (PCG) using deep neural network. *Physical and Engineering Sciences in Medicine.* **43**, 505-515 (2020). <https://doi.org/10.1007/s13246-020-00851-w>.
- [94] Anusuya, M.A. & Katti, S.K. Comparison of different speech feature extraction techniques with and without wavelet transform to Kannada speech recognition. *International Journal of Computer Applications.* **26**(4), 19-24 (2011). <https://doi.org/10.5120/3092-4242>.
- [95] Deng, L., Seltzer, M.L., Yu, D., et al. Binary coding of speech spectrograms using a deep auto-encoder. *In: Interspeech 2010.* 1692-1695 (2010).
- [96] Zhong, Y.-C. & Li, F. Model identification study on micro robot mobile in liquid based on support vector machine. *In: 2008 3rd IEEE International Conference on Nano/Micro Engineered and Molecular Systems (NEMS).* 55-59 (2008). <https://doi.org/10.1109/NEMS.2008.4484285>.
- [97] Lu, P., Xu, D.-P. & Liu, Y.-B. Study of fault diagnosis model based on multi-class wavelet support vector machines. *In: 2005 International Conference on Machine Learning and Cybernetics.* 4319-4321 (2005). <https://doi.org/10.1109/ICMLC.2005.1527697>.
- [98] Yi, H., Sun, S., Duan, X. & Chen, Z. A study on deep neural networks framework. *In: 2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC).* 1519-1522 (2016). <https://doi.org/10.1109/IMCEC.2016.7867471>.
- [99] Balamurugan, A., Teo, S.G., Yang, J., et al. ResHNet: Spectrograms based efficient heart sounds classification using stacked residual networks. *In: 2019 IEEE EMBS International Conference on*

- Biomedical & Health Informatics (BHI). 1-4 (2019). <https://doi.org/10.1109/BHI.2019.8834578>.
- [100] Gopika, P., Sowmya, V., Gopalakrishnan, E.A., et al. Performance improvement of deep learning architectures for phonocardiogram signal classification using fast fourier transform. *In: 2019 9th International Conference on Advances in Computing and Communication (ICACC)*. 290-294 (2019). <https://doi.org/10.1109/ICACC48162.2019.8986216>.
- [101] Upreti, P. & Yüksel, M.E. Accurate classification of heart sounds for disease diagnosis by a single time-varying spectral feature: Preliminary results. *In: 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*. 1-4 (2019). <https://doi.org/10.1109/EBBT.2019.8741730>.
- [102] Yadav, A., Dutta, M.K., Travieso, C.M., et al. Automatic classification of normal and abnormal PCG recording heart sound recording using Fourier transform. *In: 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*. 1-9 (2018). <https://doi.org/10.1109/IWOBI.2018.8464131>.
- [103] Saputra, L.K.P., Nugroho, H.A. & Wulandari, M. Feature extraction and classification of heart sound based on autoregressive power spectral density (AR-PSD). *In: 2014 1st International Conference on Information Technology, Computer, and Electrical Engineering*. 139-143 (2014). <https://doi.org/10.1109/ICITACEE.2014.7065730>.
- [104] Singh, S.A., Meitei, T.G. & Majumder, S. Short PCG classification based on deep learning. *In: Deep Learning Techniques for Biomedical and Health Informatics*. 141-164 (2020). Academic Press. <https://doi.org/10.1016/B978-0-12-819061-6.00006-9>.
- [105] Mane, P., Chaskar, U. & Mulay, P. Machine Learning-Based Classification of Heart Sound Using Hilbert Transform. *In: Yang, X.S., Sherratt, R.S., Dey, N., Joshi, A. (eds) Proceedings of Fifth International Congress on Information and Communication Technology. Advances in Intelligent Systems and Computing*. **1183** (2021). Springer, Singapore. https://doi.org/10.1007/978-981-15-5856-6_40.
- [106] Mei, N., Wang, H., Zhang, Y., et al. Classification of heart sounds based on quality assessment and wavelet scattering transform. *Computers in Biology and Medicine*. **137**, 104814 (2021). <https://doi.org/10.1016/j.combiomed.2021.104814>.
- [107] Rath, A., Mishra, D., Panda, G., et al. Development and assessment of machine learning based heart disease detection using imbalanced heart sound signal. *Biomedical Signal Processing and Control*. **76**, 103730 (2022). <https://doi.org/10.1016/j.bspc.2022.103730>.
- [108] Zeinali, Y. & Niaki, S.T.A. Heart sound classification using signal processing and machine learning algorithms. *Machine Learning with Applications*. **7**, 100206 (2022). <https://doi.org/10.1016/j.mlwa.2021.100206>.
- [109] Barua, P.D., Karasu, M., Kobat, M.A., et al. An accurate valvular heart disorders detection model based on a new dual symmetric tree pattern using stethoscope sounds. *Computers in Biology and*

- Medicine*. **146**, 105599 (2022). <https://doi.org/10.1016/j.compbimed.2022.105599>.
- [110] Milani, M.G.M., Abas, P.E., De Silva, L.C., et al. Abnormal heart sound classification using phonocardiography signals. *Smart Health*. **21**, 100194 (2021). <https://doi.org/10.1016/j.smhl.2021.100194>.
- [111] Oh, S.L., Jahmunah, V., Ooi, C.P., et al. Classification of heart sound signals using a novel deep WaveNet model. *Computer Methods and Programs in Biomedicine*. **196**, 105604 (2020). <https://doi.org/10.1016/j.cmpb.2020.105604>.
- [112] Xu, W., Yu, K., Ye, J., et al. Automatic pediatric congenital heart disease classification based on heart sound signal. *Artificial Intelligence in Medicine*. **126**, 102257 (2022). <https://doi.org/10.1016/j.artmed.2022.102257>.
- [113] Alkhodari, M. & Fraiwan, L. Convolutional and recurrent neural networks for the detection of valvular heart diseases in phonocardiogram recordings. *Computer Methods and Programs in Biomedicine*. **200**, 105940 (2021). <https://doi.org/10.1016/j.cmpb.2021.105940>.
- [114] Yin, C., Zhou, X., Zhao, Y., et al. Diagnosis of exercise-induced cardiac fatigue based on deep learning and heart sounds. *Applied Acoustics*. **197**, 108900 (2022). <https://doi.org/10.1016/j.apacoust.2022.108900>.
- [115] Kobat, M.A. & Dogan, S. Novel three kernelled binary pattern feature extractor based automated PCG sound classification method. *Applied Acoustics*. **179**, 108040 (2021). <https://doi.org/10.1016/j.apacoust.2021.108040>.
- [116] Wang, H., Guo, X., Zheng, Y., et al. An automatic approach for heart failure typing based on heart sounds and convolutional recurrent neural networks. *Physical and Engineering Sciences in Medicine*. **45**, 475-485 (2022). <https://doi.org/10.1007/s13246-022-01112-8>.
- [117] Nguyen, M.T., Lin, W.W. & Huang, J.H. Heart Sound Classification Using Deep Learning Techniques Based on Log-mel Spectrogram. *Circuits, Systems, and Signal Processing*. **42**, 344-360 (2023). <https://doi.org/10.1007/s00034-022-02124-1>.
- [118] Zeng, W., Su, B., Yuan, C., et al. Automatic detection of heart valve disorders using Teager-Kaiser energy operator, rational-dilation wavelet transform and convolutional neural networks with PCG signals. *Artificial Intelligence Review*. **56**, 781-806 (2023). <https://doi.org/10.1007/s10462-022-10184-7>.
- [119] Prabhakar, S.K. & Won, D.-O. Phonocardiogram signal classification for the detection of heart valve diseases using robust conglomerated models. *Expert Systems with Applications*. **221**, 119720 (2023). <https://doi.org/10.1016/j.eswa.2023.119720>.
- [120] Harimi, A., Ameri, M.A., Sarkar, S., et al. Heart sounds classification: Application of a new CyTex inspired method and deep convolutional neural network with transfer learning. *Smart Health*. **29**, 100416 (2023). <https://doi.org/10.1016/j.smhl.2023.100416>.
- [121] Jamil, S. & Roy, A.M. An efficient and robust phonocardiography (PCG)-based valvular heart diseases (VHD) detection framework using vision transformer (ViT). *Computers in Biology and*

- Medicine*. **158**, 106734 (2023). <https://doi.org/10.1016/j.compbiomed.2023.106734>.
- [122] Touahria, R., Hacine-Gharbi, A. & Ravier, P. Feature selection algorithms highlight the importance of the systolic segment for normal/murmur PCG beat classification. *Biomedical Signal Processing and Control*. **86**, 105288 (2023). <https://doi.org/10.1016/j.bspc.2023.105288>.
- [123] Zang, J., Lian, C., Xu, B., et al. AmtNet: Attentional multi-scale temporal network for phonocardiogram signal classification. *Biomedical Signal Processing and Control*. **85**, 104934 (2023). <https://doi.org/10.1016/j.bspc.2023.104934>.
- [124] Riccio, D., Brancati, N., Sannino, G., et al. CNN-based classification of phonocardiograms using fractal techniques. *Biomedical Signal Processing and Control*. **86**, 105186 (2023). <https://doi.org/10.1016/j.bspc.2023.105186>.
- [125] Maity, A., Pathak, A. & Saha, G. Transfer learning based heart valve disease classification from Phonocardiogram signal. *Biomedical Signal Processing and Control*. **85**, 104805 (2023). <https://doi.org/10.1016/j.bspc.2023.104805>.
- [126] Zheng, Y., Guo, X., Yang, Y., et al. Phonocardiogram transfer learning-based CatBoost model for diastolic dysfunction identification using multiple domain-specific deep feature fusion. *Computers in Biology and Medicine*. **156**, 106707 (2023). <https://doi.org/10.1016/j.compbiomed.2023.106707>.
- [127] Huisa, C.M., Supo, C.E., Figueroa, T.E., Rendulich, J. & Sulla-Espinoza, E. PCG Heart Sounds Quality Classification Using Neural Networks and SMOTE Tomek Links for the Think Health Project. In: Khanna, A., Polkowski, Z., Castillo, O. (eds) Proceedings of Data Analytics and Management. *Lecture Notes in Networks and Systems*. **572** (2023). Springer, Singapore. https://doi.org/10.1007/978-981-19-7615-5_65.
- [128] Wang, R., Duan, Y., Li, Y., et al. PCTMF-Net: heart sound classification with parallel CNNs-transformer and second-order spectral analysis. *Visual Computer*. **39**, 3811-3822 (2023). <https://doi.org/10.1007/s00371-023-03031-5>.
- [129] Koike, T., Qian, K., Kong, Q., Plumbley, M.D., Schuller, B.W. & Yamamoto, Y. Audio for audio is better? An investigation on transfer learning models for heart sound classification. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 74-77 (2020). IEEE. <https://doi.org/10.1109/EMBC44109.2020.9175450>.
- [130] Baydoun, M., Safatly, L., Ghaziri, H. & El Hajj, A. Analysis of heart sound anomalies using ensemble learning. *Biomedical Signal Processing and Control*. **62**, 102019 (2020). <https://doi.org/10.1016/j.bspc.2020.102019>.
- [131] Chen, W., Sun, Q., Chen, X., Xie, G., Wu, H. & Xu, C. Deep learning methods for heart sounds classification: A systematic review. *Entropy*. **23**(6), 667 (2021). <https://doi.org/10.3390/e23060667>.
- [132] Mushtaq, Z., Su, S.-F. & Tran, Q.-V. Spectral images based environmental sound classification using CNN with meaningful data augmentation. *Applied Acoustics*. **172**, 107581 (2021). <https://doi.org/10.1016/j.apacoust.2021.107581>.

- [//doi.org/10.1016/j.apacoust.2020.107581](https://doi.org/10.1016/j.apacoust.2020.107581).
- [133] Lampert, T.A. & O'Keefe, S.E.M. A survey of spectrogram track detection algorithms. *Applied Acoustics*. **71**(2), 87-100 (2010). <https://doi.org/10.1016/j.apacoust.2009.08.007>.
- [134] Munkong, R. & Juang, B.-H. Auditory perception and cognition. *IEEE Signal Processing Magazine*. **25**(3), 98-117 (2008). <https://doi.org/10.1109/MSP.2008.918418>.
- [135] Pour, A.F., Asgari, M. & Hasanabadi, M.R. Gammatonegram based speaker identification. *In: 2014 4th International Conference on Computer and Knowledge Engineering (ICCCKE)*. 52-55 (2014). IEEE. <https://doi.org/10.1109/ICCCKE.2014.6993383>.
- [136] Sharma, G., Umapathy, K. & Krishnan, S. Trends in audio signal feature extraction methods. *Applied Acoustics*. **158**, 107020 (2020). <https://doi.org/10.1016/j.apacoust.2019.107020>.
- [137] Liu, F., Shen, T., Luo, Z., Zhao, D. & Guo, S. Underwater target recognition using convolutional recurrent neural networks with 3-D Mel-spectrogram and data augmentation. *Applied Acoustics*. **178**, 107989 (2021). <https://doi.org/10.1016/j.apacoust.2021.107989>.
- [138] Meintjes, A., Lowe, A. & Legget, M. Fundamental Heart Sound Classification using the Continuous Wavelet Transform and Convolutional Neural Networks. *In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 409-412 (2018). IEEE, Honolulu, HI, USA. <https://doi.org/10.1109/EMBC.2018.8512284>.
- [139] Strisciuglio, N., Vento, M. & Petkov, N. Bio-inspired filters for audio analysis. *In: International Workshop on Brain-Inspired Computing*. 101-115 (2015). Springer. https://doi.org/10.1007/978-3-319-50862-7_8.
- [140] Tran, T. & Lundgren, J. Drill fault diagnosis based on the scalogram and mel spectrogram of sound signals using artificial intelligence. *IEEE Access*. **8**, 203655-203666 (2020). <https://doi.org/10.1109/ACCESS.2020.3036769>.
- [141] Eltrass, A.S., Tayel, M.B. & Ammar, A.I. A new automated CNN deep learning approach for identification of ECG congestive heart failure and arrhythmia using constant-Q non-stationary Gabor transform. *Biomedical Signal Processing and Control*. **65**, 102326 (2021). <https://doi.org/10.1016/j.bspc.2020.102326>.
- [142] Ren, Z., Qian, K., Zhang, Z., et al. Deep scalogram representations for acoustic scene classification. *IEEE/CAA Journal of Automatica Sinica*. **5**(3), 662-669 (2018). <https://doi.org/10.1109/JAS.2018.7511066>.
- [143] Bujaković, D.M., Andrić, M.S. & AntoniĆ, M.D. Analysis of Sound Signals Using Wavelet Transform. *a a*. **1**, 2 (2010).
- [144] Singh, S.A., Majumder, S. & Mishra, M. Classification of short unsegmented heart sound based on deep learning. *In: 2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. 1-6 (2019). IEEE. <https://doi.org/10.1109/I2MTC.2019.8826991>.
- [145] Malik, A.E.F., Barin, S. & Yüksel, M.E. Accurate classification of heart sound signals for cardiovascular disease diagnosis by wavelet analysis and convolutional neural network: preliminary

- results. *In: 2020 28th Signal Processing and Communications Applications Conference (SIU)*. 1-4 (2020). IEEE. <https://doi.org/10.1109/SIU49456.2020.9302491>.
- [146] Goupillaud, P., Grossmann, A. & Morlet, J. Cycle-octave and related transforms in seismic signal analysis. *Geoexploration*. **23**(1), 85-102 (1984). [https://doi.org/10.1016/0016-7142\(84\)90025-5](https://doi.org/10.1016/0016-7142(84)90025-5).
- [147] Mallat, S. *A Wavelet Tour of Signal Processing*. Academic Press (1999).
- [148] Daubechies, I. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics (1993). <https://doi.org/10.1121/1.406784>.
- [149] Foggia, P., Saggese, A., Strisciuglio, N. & Vento, M. Cascade classifiers trained on gammatonegrams for reliably detecting audio events. *In: 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 50-55 (2014). IEEE. <https://doi.org/10.1109/AVSS.2014.6918643>.
- [150] Patterson, R.D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C. & Allerhand, M. Complex sounds and auditory images. *In: Auditory Physiology and Perception*. 429-446 (1992). Elsevier. <https://doi.org/10.1016/B978-0-08-041847-6.50054-X>.
- [151] Bhangale, K.B. & Kothandaraman, M. Speech emotion recognition using the novel PEmoNet (Parallel Emotion Network). *Applied Acoustics*. **212**, 109613 (2023). <https://doi.org/10.1016/j.apacoust.2023.109613>.
- [152] Gupta, S., Agrawal, M. & Deepak, D. Gammatonegram based triple classification of lung sounds using deep convolutional neural network with transfer learning. *Biomedical Signal Processing and Control*. **70**, 102947 (2021). <https://doi.org/10.1016/j.bspc.2021.102947>.
- [153] Yatabe, K., Masuyama, Y., Kusano, T. & Oikawa, Y. Representation of complex spectrogram via phase conversion. *Acoustical Science and Technology*. **40**(3), 170-177 (2019). <https://doi.org/10.1250/ast.40.170>.
- [154] Ulutas, G., Tahaoglu, G. & Ustubioglu, B. Forge Audio Detection Using Keypoint Features on Mel Spectrograms. *In: 2022 45th International Conference on Telecommunications and Signal Processing (TSP)*. 413-416 (2022). IEEE. <https://doi.org/10.1109/TSP55681.2022.9851327>.
- [155] Tran, T. & Lundgren, J. Drill fault diagnosis based on the scalogram and mel spectrogram of sound signals using artificial intelligence. *IEEE Access*. **8**, 203655-203666 (2020). <https://doi.org/10.1109/ACCESS.2020.3036769>.
- [156] Sigurdsson, S., Petersen, K.B. & Lehn-Schioler, T. Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3 Encoded Music. *In: ISMIR*. 286-289 (2006).
- [157] Mantji, B.N. & Oloo, F.R.A. Electronic Stethoscope Design, Prototyping and Testing. *In: IEEE EUROCON 2019-18th International Conference on Smart Technologies*. 1-7 (2019). IEEE. <https://doi.org/10.1109/EUROCON.2019.8861992>.

- [158] Bhangale, K. & Mohanaprasad, K. Speech emotion recognition using mel frequency log spectrogram and deep convolutional neural network. *In: International Conference on Futuristic Communication and Network Technologies*. 241-250 (2020). Springer. https://doi.org/10.1007/978-981-16-4625-6_24.
- [159] Dorfler, M. & Matusiak, E. Nonstationary Gabor frames-approximately dual frames and reconstruction errors. *Advances in Computational Mathematics*. **41**, 293-316 (2015). <https://doi.org/10.1007/s10444-014-9358-z>.
- [160] Liu, T., Yan, S. & Zhang, W. Time-frequency analysis of nonstationary vibration signals for deployable structures by using the constant-Q nonstationary gabor transform. *Mechanical Systems and Signal Processing*. **75**, 228-244 (2016). <https://doi.org/10.1016/j.ymssp.2015.12.015>.
- [161] Eltrass, A.S., Tayel, M.B. & Ammar, A.I. A new automated CNN deep learning approach for identification of ECG congestive heart failure and arrhythmia using constant-Q non-stationary Gabor transform. *Biomedical Signal Processing and Control*. **65**, 102326 (2021). <https://doi.org/10.1016/j.bspc.2020.102326>.
- [162] Velasco, G.A., Holighaus, N., Dorfler, M. & Grill, T. Constructing an invertible constant-Q transform with non-stationary Gabor frames. *Proceedings of DAFX11, Paris*. **33**, 81 (2011).
- [163] Zhang, J., Li, Y. & Yin, J. Modulation classification method for frequency modulation signals based on the time-frequency distribution and CNN. *IET Radar, Sonar & Navigation*. **12**(2), 244-249 (2018). <https://doi.org/10.1049/iet-rsn.2017.0265>.
- [164] Mammone, N., Ieracitano, C. & Morabito, F.C. A deep CNN approach to decode motor preparation of upper limbs from time-frequency maps of EEG signals at source level. *Neural Networks*. **124**, 357-372 (2020). <https://doi.org/10.1016/j.neunet.2020.01.027>.
- [165] Sharma, L., Chao, C.-H., Wu, S.-L. & Li, M.-C. High accuracy WiFi-based human activity classification system with time-frequency diagram CNN method for different places. *Sensors*. **21**(11), 3797 (2021). <https://doi.org/10.3390/s21113797>.
- [166] Wang, H. & Wang, D. Time-frequency loss for CNN based speech super-resolution. *In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 861-865 (2020). IEEE. <https://doi.org/10.1109/ICASSP40776.2020.9053712>.
- [167] Pang, C., Liu, H. & Li, X. Multitask learning of time-frequency CNN for sound source localization. *IEEE Access*. **7**, 40725-40737 (2019). <https://doi.org/10.1109/ACCESS.2019.2905617>.
- [168] Xie, Y., Xiao, Y., Liu, X., Liu, G., Jiang, W. & Qin, J. Time-frequency distribution map-based convolutional neural network (CNN) model for underwater pipeline leakage detection using acoustic signals. *Sensors*. **20**(18), 5040 (2020). <https://doi.org/10.3390/s20185040>.
- [169] Neili, Z. & Sundaraj, K. A comparative study of the spectrogram, scalogram, melspectrogram and gammatonegram time-frequency representations for the classification of lung sounds using

- the ICBHI database based on CNNs. *Biomedical Engineering/Biomedizinische Technik*. **67**(5), 367-390 (2022). <https://doi.org/10.1515/bmt-2022-0180>.
- [170] Krizhevsky, A., Sutskever, I. & Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems (NeurIPS)*. **25** (2012).
- [171] Alom, M.Z., Taha, T.M., Yakopcic, C., et al. The history began from AlexNet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164* (2018). <https://doi.org/10.48550/arXiv.1803.01164>.
- [172] The Clever Programmer. AlexNet Architecture Using Python. (2021). <https://thecleverprogrammer.com/2021/12/13/alexnet-architecture-using-python/>, [Online; Accessed 2025-04-30].
- [173] Alaskar, H., Alzhrani, N., Hussain, A. & Almarshed, F. The Implementation of Pretrained AlexNet on PCG Classification. In: Huang, D.S., Huang, Z.K., Hussain, A. (eds) Intelligent Computing Methodologies. *Lecture Notes in Computer Science*. **11645** (2019). Springer, Cham. https://doi.org/10.1007/978-3-030-26766-7_71.
- [174] Mashrur, F.R., Roy, A.D. & Saha, D.K. Automatic Identification of Arrhythmia from ECG Using AlexNet Convolutional Neural Network. In: 2019 4th International Conference on Electrical Information and Communication Technology (EICT). 1-5 (2019). IEEE. <https://doi.org/10.1109/EICT48899.2019.9068806>.
- [175] Mansour, R.F. Deep-learning-based automatic computer-aided diagnosis system for diabetic retinopathy. *Biomedical Engineering Letters*. **8**, 41-57 (2018). <https://doi.org/10.1007/s13534-017-0047-y>.
- [176] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 770-778 (2016).
- [177] Ramzan, F., Khan, M.U.G., Rehmat, A., et al. A deep learning approach for automated diagnosis and multi-class classification of Alzheimer's disease stages using resting-state fMRI and residual neural networks. *Journal of Medical Systems*. **44**, 37 (2020). <https://doi.org/10.1007/s10916-019-1475-2>.
- [178] Rahman, T., Chowdhury, M.E.H., Khandakar, A., Islam, K.R., Islam, K.F., Mahbub, Z.B., Kadir, M.A. & Kashem, S. Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection Using Chest X-ray. *Applied Sciences*. **10**(9), 3233 (2020). <https://doi.org/10.3390/app10093233>.
- [179] Mhasakar, P., Trivedi, P., Mandal, S., et al. Handwritten Digit Recognition Using Bayesian ResNet. *SN Computer Science*. **2**, 399 (2021). <https://doi.org/10.1007/s42979-021-00791-6>.
- [180] Jing, E., Zhang, H., Li, Z., et al. ECG heartbeat classification based on an improved ResNet-18 model. *Computational and Mathematical Methods in Medicine*. **2021**, 6649970 (2021). <https://doi.org/10.1155/2021/6649970>.

- //doi.org/10.1155/2021/6649970.
- [181] Gao, Y., Qiao, L. & Li, Z. Heart Murmur Detection of PCG Using ResNet with Selective Kernel Convolution. *In: 2022 Computing in Cardiology (CinC)*. 1-4 (2022). <https://doi.org/10.22489/CinC.2022.172>.
- [182] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2015). <https://doi.org/10.48550/arXiv.1409.1556>.
- [183] Patel, S.B., Patel, P.H., Jain, V.D. & Verma, J.P. Improvised VGG16 CNN Architecture for Predicting Tuberculosis Using the Frontal Chest X-Ray Images. *In: Somani, A.K., Mundra, A., Doss, R., Bhattacharya, S. (eds) Smart Systems: Innovations in Computing. Smart Innovation, Systems and Technologies*. **235** (2022). Springer, Singapore. https://doi.org/10.1007/978-981-16-2877-1_7.
- [184] Jiang, Z.-P., Liu, Y.-Y., Shao, Z.-E. & Huang, K.-W. An Improved VGG16 Model for Pneumonia Image Classification. *Applied Sciences*. **11**(23), 11185 (2021). <https://doi.org/10.3390/app112311185>.
- [185] Medium Nerd for Tech. VGG 16 Easiest Explanation. (2024). <https://medium.com/nerd-for-tech/vgg-16-easiest-explanation-12453b599526>, [Online; Accessed 2025-04-30].
- [186] Albashish, D., Al-Sayyed, R., Abdullah, A., Ryalat, M.H. & Almansour, N.A. Deep CNN Model based on VGG16 for Breast Cancer Classification. *In: 2021 International Conference on Information Technology (ICIT)*. 805-810 (2021). IEEE. <https://doi.org/10.1109/ICIT52682.2021.9491631>.
- [187] Almanifi, O.R.A., Mohd Razman, M.A., Musa, R.M., Ab. Nasir, A.F., Ismail, M.Y. & P. P. Abdul Majeed, A. The Classification of Heartbeat PCG Signals via Transfer Learning. *In: Ab. Nasir, A.F., Ibrahim, A.N., Ishak, I., Mat Yahya, N., Zakaria, M.A., P. P. Abdul Majeed, A. (eds) Recent Trends in Mechatronics Towards Industry 4.0. Lecture Notes in Electrical Engineering*. **730** (2022). Springer, Singapore. https://doi.org/10.1007/978-981-33-4597-3_5.
- [188] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929v2* (2021). <https://doi.org/10.48550/arXiv.2010.11929>.
- [189] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. & J'egou, H. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877v2* (2021). <https://doi.org/10.48550/arXiv.2012.12877>.
- [190] Viso.ai. Vision Transformers (ViT) in Image Recognition. (2023). <https://viso.ai/deep-learning/vision-transformer-vit/>, [Online; Accessed 2025-05-01].
- [191] Bang, J.-H., et al. CA-CMT: Coordinate Attention for Optimizing CMT Networks. *IEEE Access*. **11**, 76691-76702 (2023). <https://doi.org/10.1109/ACCESS.2023.3297206>.

- [192] Jumphoo, T., Phapatanaburi, K., Pathonsuwan, W., Anchuen, P., Uthansakul, M. & Uthansakul, P. Exploiting Data-Efficient Image Transformer-Based Transfer Learning for Valvular Heart Diseases Detection. *IEEE Access*. **12**, 15845-15855 (2024). <https://doi.org/10.1109/ACCESS.2024.3357946>.
- [193] Bi, Y., Xue, B. & Zhang, M. Genetic Programming-Based Evolutionary Deep Learning for Data-Efficient Image Classification. *IEEE Transactions on Evolutionary Computation*. **28**(2), 307-322 (2024). <https://doi.org/10.1109/TEVC.2022.3214503>.
- [194] Bhagyasree, P.V., James, A., Bisna, N.D. & Kumar, K.S.V. Handwritten Cursive English Character Recognition Using DAG-CNN. In: Chakravarthy, V.V.S.S.S., Flores-Fuentes, W., Bhateja, V., Biswal, B. (eds) Advances in Micro-Electronics, Embedded Systems and IoT. *Lecture Notes in Electrical Engineering*. **838** (2022). Springer, Singapore. https://doi.org/10.1007/978-981-16-8550-7_10.
- [195] Costa, M.G.F., Campos, J.P.M., Aquino, G.A., Pereira, W.C.A. & Filho, C.F.F.C. Evaluating the performance of convolutional neural networks with direct acyclic graph architectures in automatic segmentation of breast lesion in US images. *BMC Medical Imaging*. **19**(1), 85 (2019). <https://doi.org/10.1186/s12880-019-0380-y>.
- [196] Kulkarni, V. Classification Metrics. Medium (2023). <https://medium.com/@vyankatesh.kulkarni20/classification-metrics-63c635852995>, [Online; Accessed 2025-05-07].