UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CAMPUS DE SÃO JOÃO DA BOA VISTA

ISADORA RAPHAELA CAZELATO SILVA

Aplicação de Redes Neurais Convolucionais na Detecção de Sopros Cardíacos em Sinais de Fonocardiograma

São João da Boa Vista
2026

Isadora Raphaela Cazelato Silva

Aplicação de Redes Neurais Convolucionais na Detecção de Sopros Cardíacos em Sinais de Fonocardiograma

Trabalho de Graduação apresentado ao Conselho de Curso de Graduação em Engenharia Aeronáutica do Campus de São João da Boa Vista, Universidade Estatual Paulista, como parte dos requisitos para obtenção do diploma de Graduação em Engenharia Aeronáutica.

Orientador: Profª Drª. Leandra Isabel de Abreu

São João da Boa Vista
2026

S586a

Silva, Isadora Raphaela Cazelato
Aplicação de Redes Neurais Convolucionais na Detecção de Sopros Cardíacos em Sinais de Fonocardiograma / Isadora Raphaela Cazelato Silva. -- São João da Boa Vista, 2026
38 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Engenharia Aeronáutica) - Universidade Estadual Paulista (UNESP), Faculdade de Engenharia, São João da Boa Vista
Orientadora: Leandra Isabel de Abreu

1. Hemodinâmica. 2. Aprendizagem profunda. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"

FACULDADE DE ENGENHARIA - CÂMPUS DE SÃO JOÃO DA BOA VISTA

GRADUAÇÃO EM ENGENHARIA AERONÁUTICA

TRABALHO DE CONCLUSÃO DE CURSO

APLICAÇÃO DE REDES NEURAIS CONVOLUCIONAIS NA DETECÇÃO DE SOPROS CARDÍACOS EM SINAIS DE FONOCARDIOGRAMA

Aluno: Isadora Raphaela Cazelato Silva

Orientador: Prof.ª Dr.ª Leandra Isabel de Abreu

Banca Examinadora:

- Leandra Isabel de Abreu (Orientadora)
- Ivan Aritz Aldaya Garde (Examinador)
- Wolney Dalla Pria Neto (Examinador)

Os formulários de avaliação e a ata da defesa, na qual consta a aprovação do trabalho, devidamente assinados pela banca encontram-se no prontuário eletrônico do aluno.

São João da Boa Vista, 16 de outubro de 2026.

RESUMO

Este trabalho investiga a aplicação de técnicas de aprendizado profundo na detecção e classificação automática de anomalias em sinais de fonocardiograma (FCG). As doenças cardiovasculares exigem triagens eficientes, mas a auscultação clínica tradicional apresenta limitações relacionadas à subjetividade e variação entre observadores. Para contornar essa questão, utilizou-se a base de dados pediátrica CirCor DigiScope para treinar e avaliar um modelo computacional capaz de identificar sopros cardíacos. A metodologia consistiu no pré-processamento dos sinais de áudio mediante segmentação temporal em janelas uniformes de 5 segundos, seguida da conversão para representações visuais bidimensionais no domínio tempo-frequência (espectrogramas) utilizando a Transformada de Fourier de Tempo Curto (TFTC). As imagens em escala de cinza alimentaram a arquitetura convolucional ResNet-50, adaptada por meio de transferência de aprendizado (*transfer learning*). No conjunto de teste independente, o classificador alcançou uma acurácia global de 80,0% e uma pontuação F_1 ponderada de 0,82. A análise de desempenho, quando comparada proporcionalmente à triagem diagnóstica médica da mesma base, demonstrou que a rede neural atua como uma ferramenta altamente específica. Enquanto o diagnóstico humano apresentou uma Taxa de Falsos Positivos de 38,0%, o modelo computacional reduziu essa métrica para 18,4%, mitigando significativamente os alarmes falsos. Embora a triagem clínica tenha mantido uma maior sensibilidade (83,2% contra 73,0% da rede), conclui-se que o modelo proposto oferece um equilíbrio diagnóstico valioso, atuando como um filtro inicial promissor para reduzir encaminhamentos médicos e exames ecocardiográficos desnecessários sem comprometer o rastreio patológico.

Palavras-chave: fonocardiograma; aprendizado profundo; ResNet-50; espectrograma; CirCor DigiScope.

ABSTRACT

This work investigates the application of deep learning techniques for the automated detection and classification of anomalies in phonocardiogram (PCG) signals. Cardiovascular diseases require efficient screening, yet traditional clinical auscultation presents limitations related to subjectivity and inter-observer variability. To overcome this issue, the pediatric CirCor DigiScope dataset was utilized to train and evaluate a computational model capable of identifying heart murmurs. The methodology consisted of pre-processing the audio signals through temporal segmentation into uniform 5-second windows, followed by their conversion into two-dimensional visual representations in the time-frequency domain (spectrograms) using the Short-Time Fourier Transform (STFT). The grayscale images were fed into a ResNet-50 convolutional architecture, adapted via transfer learning. On the independent test set, the classifier achieved an overall accuracy of 80.0% and a weighted F_1-score of 0.82. Performance analysis, when proportionally compared to the medical diagnostic screening of the same dataset, demonstrated that the neural network acts as a highly specific tool. While human diagnosis presented a False Positive Rate of 38.0%, the computational model reduced this metric to 18.4%, significantly mitigating false alarms. Although clinical screening maintained higher sensitivity (83.2% versus the network's 73.0%), it is concluded that the proposed model offers a valuable diagnostic balance, acting as a promising initial filter to reduce unnecessary medical referrals and echocardiographic exams without compromising pathological screening.

Keywords: phonocardiogram; deep learning; ResNet-50; spectrogram; CirCor DigiScope.

LISTA DE ILUSTRAÇÕES

Figura 1 – Representação tempo-frequência ilustrando a capacidade da STFT de localizar eventos no tempo: o Sinal A exibe componentes contínuas e constantes, enquanto o Sinal B mapeia transições espectrais em intervalos específicos. 17
Figura 2 – Representação esquemática da arquitetura de uma Rede Neural Convolucional. . . 20
Figura 3 – Exemplo de um Espectrograma de potência log-frequência gerado 29
Figura 4 – Distribuição da duração dos áudios cardíacos na base de dados, justificando a necessidade de padronização. 30
Figura 5 – Sinal acústico no domínio do tempo evidenciando artefatos e transientes de alta amplitude (Amplitude vs. Tempo em segundos). 31
Figura 6 – Espectrograma de potência log-frequência de um sinal segmentado (Frequência em Hz vs. Tempo em segundos). 31
Figura 7 – Amostra de lote de espectrogramas convertidos para escala de cinza, classificados como normais (Saudável) e anormais (Não Saudável), submetidos à rede neural. . 32
Figura 8 – Evolução da função de perda (*loss*) de treinamento e de validação por época. . . . 32
Figura 9 – Matriz de confusão resultante da avaliação do modelo no conjunto de teste. 33
Figura 10 – Matriz de confusão refletindo o desempenho do diagnóstico clínico (médico) na base CirCor. 33

LISTA DE TABELAS

Tabela 1 – Métricas estatísticas de desempenho obtidas pela ResNet-50 no conjunto de teste. 35

LISTA DE SÍMBOLOS

F_1	Pontuação que representa a média harmônica entre precisão e sensibilidade (*F1-score*)
S_1	Primeira Bulha Cardíaca
S_2	Segunda Bulha Cardíaca
Re	Número de Reynolds
ρ	Densidade do sangue
v	Velocidade média do escoamento
D	Diâmetro da câmara ou vaso condutor
μ	Viscosidade dinâmica do sangue
$x[n]$	Sequência discreta / Sinal discretizado
$X[m, k]$	Transformada de Fourier de Tempo Curto (STFT) de uma sequência discreta
m	Índice da janela ao longo do tempo
R	Deslocamento da janela (*hop size*)
k	Índice discreto de frequência
$w[n]$	Função de janelamento simétrica
$S[m, k]$	Espectrograma (magnitude da resposta da STFT)
$S_{dB}[m, k]$	Espectrograma em escala logarítmica (decibéis)
$\mathcal{F}(x)$	Transformação computada pelas convoluções no bloco residual
x	Vetor de entrada do estágio convolucional
VP	Verdadeiro Positivo
FN	Falso Negativo
FP	Falso Positivo
VN	Verdadeiro Negativo
Acc	Acurácia
t	Tempo (em segundos)

SNR_{treino}	Relação sinal-ruído de treino
SNR_{teste}	Relação sinal-ruído de teste

SUMÁRIO

1 INTRODUÇÃO 11

2 OBJETIVOS 13

3 FUNDAMENTAÇÃO TEÓRICA 14

3.1 FISIOLOGIA CARDÍACA E SINAIS DE FONOCARDIOGRAMA 14

3.1.1 Gênese das Bulhas Cardíacas Normais (S_1 e S_2) 14

3.1.2 Sopros Cardíacos e Mecanismos de Turbulência 14

3.1.3 Auscultação Digital e o Sinal de Fonocardiograma 15

3.2 PROCESSAMENTO DIGITAL DE SINAIS E REPRESENTAÇÕES TEMPO-FREQUÊNCIA 16

3.2.1 Limitações da Transformada de Fourier Clássica 16

3.2.2 Transformada de Fourier de Tempo Curto 16

3.2.3 Representação por Espectrogramas como Entrada para Redes Neurais 17

3.3 APRENDIZADO PROFUNDO E REDES NEURAIS ARTIFICIAIS 18

3.3.1 O Problema da Degradação e o Aprendizado Residual 20

3.3.2 A Arquitetura ResNet-50 e o Bloco *Bottleneck* 21

3.3.3 Transferência de Aprendizado (*Transfer Learning*) e Ajuste Fino 22

3.4 MÉTRICAS DE DESEMPENHO E AVALIAÇÃO DE MODELOS CLÍNICOS . 23

3.4.1 Matriz de Confusão 23

3.4.2 Métricas Estatísticas Derivadas 24

3.4.2.1 Acurácia (Acc): 24

3.4.2.2 Sensibilidade ou Revocação (*Sensitivity* / *Recall*): 24

3.4.2.3 Especificidade (*Specificity*): 24

3.4.2.4 Precisão (*Precision*): 24

3.4.2.5 Pontuação F_1 (F_1-score): 25

4 METODOLOGIA 26

4.1 BASE DE DADOS E CARACTERIZAÇÃO AMOSTRAL 26

4.1.1 Origem e Protocolo de Coleta dos Dados 26

4.1.2 Anotações Clínicas e Definição das Classes de Estudo 26

4.2 AMBIENTE COMPUTACIONAL E FERRAMENTAS DE DESENVOLVIMENTO 27

4.3 PIPELINE DE PRÉ-PROCESSAMENTO DE SINAIS ACÚSTICOS 27

4.3.1 Segmentação e Padronização Temporal 28

4.3.2 Geração dos Espectrogramas via STFT 28

4.4 CRITÉRIOS DE AVALIAÇÃO E VALIDAÇÃO EXPERIMENTAL 29

4.4.1 Matriz de Confusão 29

4.4.2 Métricas Estatísticas de Desempenho 29

5 RESULTADOS E DISCUSSÕES 30
5.1 ANÁLISE DO PRÉ-PROCESSAMENTO DOS DADOS 30
5.2 CONVERGÊNCIA E DINÂMICA DE TREINAMENTO 32
5.3 MATRIZ DE CONFUSÃO NO CONJUNTO DE TESTE 33
5.4 MÉTRICAS CONSOLIDADAS DE DESEMPENHO 34

6 CONCLUSÃO 36

1 INTRODUÇÃO

A detecção automática de anomalias em sinais biomédicos tem se tornado um importante campo de pesquisa dentro da inteligência artificial, impulsionado principalmente pelos avanços recentes em técnicas de aprendizado de máquina e aprendizado profundo (**??**). Esses métodos apresentam elevada capacidade de identificar padrões complexos e relações não lineares em grandes volumes de dados, permitindo a análise eficiente de sinais fisiológicos que, muitas vezes, possuem alta variabilidade e características difíceis de serem interpretadas apenas por métodos tradicionais. Nesse contexto, redes neurais profundas têm demonstrado grande eficácia em tarefas de classificação, reconhecimento de padrões e apoio ao diagnóstico médico assistido por computador.

Entre as diversas aplicações dessas técnicas na área da saúde, destaca-se a análise automática de sinais cardíacos. As doenças cardiovasculares representam uma das principais causas de morbidade e mortalidade em todo o mundo (**??**), o que torna fundamental o desenvolvimento de métodos capazes de auxiliar na detecção precoce de anomalias cardíacas. A auscultação cardíaca, realizada com o auxílio de um estetoscópio, é uma prática clínica amplamente utilizada como ferramenta inicial de triagem, permitindo a identificação de sons característicos produzidos pelo funcionamento das válvulas cardíacas e pelo fluxo sanguíneo. Alterações nesses sons, como a presença de sopros cardíacos, podem indicar a existência de diferentes tipos de patologias.

Apesar de sua importância clínica, a auscultação cardíaca apresenta algumas limitações. A interpretação dos sons cardíacos depende fortemente da experiência e da habilidade do profissional de saúde, podendo apresentar variações significativas entre diferentes observadores (**??**). Além disso, em ambientes com poucos recursos ou com acesso limitado a especialistas, a avaliação adequada desses sinais pode se tornar ainda mais desafiadora. Nesse cenário, o desenvolvimento de ferramentas computacionais capazes de analisar automaticamente os sons cardíacos surge como uma alternativa promissora para auxiliar profissionais da saúde na identificação de possíveis anomalias.

Com o avanço dos estetoscópios eletrônicos, tornou-se possível registrar digitalmente os sons cardíacos na forma de fonocardiogramas (PCG), permitindo que esses sinais sejam armazenados, processados e analisados por métodos computacionais. A disponibilidade desses dados em formato digital possibilita a aplicação de técnicas de processamento de sinais e aprendizado de máquina para a identificação automática de padrões associados a condições normais ou patológicas. Entre as diferentes abordagens existentes, destaca-se o uso de redes neurais convolucionais (CNNs), que têm se mostrado particularmente eficazes em tarefas de classificação de imagens e reconhecimento de padrões complexos (**??**).

Uma estratégia amplamente utilizada para a análise de sinais de áudio com CNNs consiste em transformar esses sinais em representações visuais, como espectrogramas ou outras formas de representação tempo-frequência (**??**). Essa abordagem permite tratar o problema como uma tarefa de classificação de imagens, explorando a capacidade das CNNs de extrair automaticamente características relevantes a partir dos dados. Dessa forma, padrões associados a sons cardíacos normais e anômalos podem ser identificados diretamente a partir dessas representações, reduzindo a necessidade de extração manual

de características.

Importa ressaltar que, embora a aplicação empírica deste estudo esteja concentrada no diagnóstico cardiovascular, o arcabouço metodológico fundamenta-se em competências de engenharia de caráter eminentemente transversal. O processamento digital de sinais acústicos e a decomposição tempo-frequência por meio de transformadas de Fourier constituem pilares fundamentais tanto na análise biológica quanto na engenharia aeronáutica. Sob essa ótica, a estratégia de mapear séries temporais não estacionárias em representações visuais bidimensionais para classificação via redes neurais convolucionais encontra direta correlação prática com a indústria aeroespacial, a exemplo de rotinas de monitoramento de integridade estrutural (*Structural Health Monitoring* – SHM), detecção precoce de trincas por emissão acústica e inspeção de vibrações anômalas em fuselagens e componentes rotativos. Contudo, apesar do amplo potencial de transferência dessas técnicas entre diferentes campos da engenharia, sua transposição eficaz para o setor da saúde ainda esbarra em desafios específicos de modelagem e disponibilidade de dados.

Apesar dos avanços obtidos nessa área, muitos dos sistemas desenvolvidos até o momento concentram-se principalmente na distinção binária entre sinais cardíacos normais e anormais, sem fornecer informações detalhadas sobre a natureza específica das anomalias detectadas. Essa limitação está frequentemente relacionada à escassez de bases de dados públicas contendo anotações clínicas detalhadas, o que dificulta o treinamento e a avaliação de modelos mais sofisticados.

Nesse contexto, destaca-se a base de dados pediátrica CirCor, que reúne um amplo conjunto de gravações de fonocardiogramas coletadas por meio de estetoscópios eletrônicos. Essa base é composta por 5.282 gravações provenientes de 1.568 pacientes e contém mais de 215 mil batimentos cardíacos anotados manualmente por médicos (**??**). Um de seus principais diferenciais é o alto nível de detalhamento das anotações associadas aos sopros cardíacos, incluindo informações relacionadas ao momento de ocorrência no ciclo cardíaco, à forma do sopro, à tonalidade, à intensidade e à localização auscultatória. Essas características tornam a base CirCor um recurso valioso e promissor para o avanço dos estudos voltados ao reconhecimento automatizado de padrões em fonocardiogramas.

Dessa forma, a combinação entre bases de dados clínicas ricas em anotações e técnicas avançadas de aprendizado profundo estabelece um campo fértil para superar as limitações diagnósticas tradicionais, impulsionando o desenvolvimento de sistemas mais robustos, precisos e acessíveis para a triagem cardiovascular assistida por computador.

2 OBJETIVOS

O presente trabalho tem como objetivo geral investigar e implementar técnicas de aprendizado de máquina e aprendizado profundo aplicadas à detecção e classificação automática de anomalias em sinais de FCG, empregando representações visuais tempo-frequência a partir dos registros da base de dados CirCor DigiScope.

Para atingir este propósito, são estabelecidos os seguintes objetivos específicos:

1. Analisar e caracterizar a base de dados pediátrica CirCor, compreendendo a distribuição dos dados e o nível de detalhamento das anotações clínicas associadas aos sopros cardíacos;
2. Realizar o pré-processamento dos sinais de áudio de FCG, aplicando técnicas adequadas de filtragem e atenuação de ruídos inerentes à auscultação eletrônica;
3. Transformar os sinais acústicos unidimensionais em representações visuais bidimensionais no domínio tempo-frequência (como espectrogramas), permitindo o tratamento da classificação sob a ótica da visão computacional;
4. Implementar e treinar modelos baseados em CNNs capazes de extrair características intrínsecas e identificar padrões patológicos nos registros;
5. Avaliar o desempenho dos modelos por meio de métricas quantitativas de classificação (como acurácia, sensibilidade, especificidade e *F1-score*), discutindo sua aplicabilidade no contexto do diagnóstico médico assistido por computador.

3 FUNDAMENTAÇÃO TEÓRICA

3.1 FISIOLOGIA CARDÍACA E SINAIS DE FONOCARDIOGRAMA

O coração humano atua como uma bomba eletromecânica composta por quatro câmaras: dois átrios e dois ventrículos. O funcionamento cíclico dessas estruturas é regido pelo ciclo cardíaco, dividido primariamente em duas fases hemodinâmicas: a sístole (contração e ejeção ventricular) e a diástole (relaxamento e enchimento ventricular). Durante as transições mecânicas entre essas fases, ocorrem variações abruptas de pressão e fluxo sanguíneo que resultam na movimentação e no fechamento das válvulas cardíacas atrioventriculares (mitral e tricúspide) e semilunares (aórtica e pulmonar) (**??**).

A desaceleração súbita da coluna sanguínea combinada à vibração mecânica das cúspides valvares e das paredes miocárdicas adjacentes produz fenômenos acústicos audíveis na superfície torácica, conhecidos clinicamente como bulhas cardíacas ou sons cardíacos fundamentais.

3.1.1 Gênese das Bulhas Cardíacas Normais (S_1 e S_2)

Em um ciclo hemodinâmico saudável, o padrão acústico básico é caracterizado por dois eventos sonoros dominantes:

- **Primeira Bulha Cardíaca (S_1):** Marca o início da sístole ventricular e o término da diástole. Sua gênese decorre principalmente do fechamento das válvulas atrioventriculares (mitral e tricúspide), em resposta ao aumento abrupto da pressão intraventricular que supera a pressão atrial. É predominantemente associado ao som tradicionalmente onomatopeico “tum”.
- **Segunda Bulha Cardíaca (S_2):** Delimita o término da sístole ventricular e o início da diástole. É gerada pela desaceleração rápida do fluxo retrógrado e consequente fechamento mecânico das válvulas semilunares (aórtica e pulmonar). Representa o som correspondente a “tá”.

3.1.2 Sopros Cardíacos e Mecanismos de Turbulência

Enquanto os sons cardíacos normais se originam de impactos estruturais e desacelerações do fluxo laminar, os sons anômalos, genericamente classificados como sopros cardíacos (*heart murmurs*), são ruídos acústicos gerados pela transição de um regime de fluxo laminar para um regime de fluxo turbulento no interior do sistema cardiovascular.

Sob condições ideais, o fluxo sanguíneo é laminar e silencioso. De acordo com a dinâmica de fluidos hemodinâmica, a transição para a turbulência é governada pelo número de Reynolds (Re) (**??**), expresso por:

$$Re = \frac{\rho \cdot v \cdot D}{\mu} \tag{1}$$

onde ρ é a densidade do sangue, v é a velocidade média do escoamento, D é o diâmetro da câmara ou vaso condutor e μ representa a viscosidade dinâmica do sangue. Quando patologias cardíacas alteram

esses parâmetros geométricos ou dinâmicos, ultrapassa-se o limite crítico de Reynolds, originando vórtices turbulentos que induzem vibrações mecânicas de maior duração na parede do vaso e no miocárdio.

Considerando o sangue em regime macroscópico nas cavidades cardíacas e grandes vasos como um fluido incompressível e de comportamento aproximadamente newtoniano em altas taxas de cisalhamento, a densidade ρ e a viscosidade μ mantêm-se essencialmente constantes. Dessa forma, o aumento expressivo de Re decorre primariamente de variações na geometria do escoamento (D) e na velocidade linear do fluido (v). Quando patologias cardíacas alteram esses parâmetros dinâmicos, ultrapassa-se o limite crítico de Reynolds, originando vórtices turbulentos que induzem vibrações mecânicas de maior duração na parede do vaso e no miocárdio.

Ao contrário das bulhas normais (S_1 e S_2), que são eventos transitórios concentrados em frequências predominantemente baixas (em geral abaixo de 150 Hz), os sopros ocupam porções mais extensas dos intervalos do ciclo acústico (sístole, diástole ou ambas) e concentram energia em componentes espectrais significativamente mais elevados, frequentemente estendendo-se de 200 Hz a mais de 600 Hz.

3.1.3 Auscultação Digital e o Sinal de Fonocardiograma

A avaliação tradicional desses sons ocorre via auscultação clínica analógica com estetoscópio acústico, prática limitada pela baixa sensibilidade do ouvido humano a frequências inferiores a 50 Hz, pela atenuação tecidual torácica e pela subjetividade inerente à interpretação clínica individual.

A introdução de estetoscópios eletrônicos digitais viabilizou a captação e transdução direta dessas vibrações mecânicas em sinais elétricos discretizados, produzindo o registro gráfico temporal denominado **Fonocardiograma**.

O sinal de FCG constitui uma série temporal unidimensional contínua das variações acústicas cardíacas. Do ponto de vista técnico e de processamento de sinais, o FCG apresenta particularidades determinantes para projetos computacionais:

- **Faixa Espectral Relevante:** A grande maioria da energia acústica dos eventos cardíacos fisiológicos e patológicos encontra-se concentrada na faixa subaudível e audível baixa, estendendo-se essencialmente de 10 Hz até cerca de 800 Hz;

- **Não Estacionariedade:** Por se originar de processos biológicos dinâmicos com modulações de ritmo, amplitude e pressão intrínsecas a cada batimento, o FCG é um sinal não estacionário, tornando a aplicação direta de transformadas puramente temporais ou puramente frequenciais ineficaz para capturar a evolução temporal de suas componentes espectrais;

- **Susceptibilidade a Ruídos Inerentes:** Sendo coletado por transdutores piezoelétricos ou microfones capacitivos acoplados à pele, o sinal bruto contém interferências externas (ruídos ambientes de consultório), ruídos fisiológicos intrínsecos (sons respiratórios e atrito tecidual/muscular) e artefatos mecânicos causados pela movimentação do sensor sobre a superfície cutânea.

Portanto, a caracterização adequada e o tratamento prévio dessas particularidades do sinal de PCG são etapas indispensáveis para que modelos computacionais consigam extrair com robustez os padrões sonoros distintivos associados às patologias cardíacas.

3.2 PROCESSAMENTO DIGITAL DE SINAIS E REPRESENTAÇÕES TEMPO-FREQUÊNCIA

O processamento digital de sinais biomédicos compreende o conjunto de técnicas matemáticas e computacionais empregadas para transformar, filtrar e extrair informações de séries temporais discretizadas. No contexto do sinal de fonocardiograma (PCG), o tratamento digital prévio é mandatório para atenuar ruídos espúrios e mapear os componentes acústicos fundamentais em representações que facilitem a discriminação automática por algoritmos de aprendizado de máquina.

3.2.1 Limitações da Transformada de Fourier Clássica

Uma vez obtido o sinal discretizado $x[n]$, a abordagem tradicional para a caracterização espectral consiste no emprego da Transformada Discreta de Fourier (TDF), comumente computada por meio do algoritmo da Transformada Rápida de Fourier (*Fast Fourier Transform* – FFT). A análise de Fourier clássica decompõe o sinal em uma combinação linear de senoides e cossenos, revelando a distribuição global de energia do registro em função da frequência.

Apesar de sua ampla utilização na engenharia, a formulação clássica de Fourier apresenta uma limitação intrínseca severa quando aplicada a sinais fisiológicos: ela considera todo o intervalo de observação de forma agregada. Por integrar o sinal ao longo de toda a sua extensão temporal, a FFT fornece com precisão *quais* frequências existem no registro, mas oculta completamente a informação de *quando* esses componentes ocorreram.

Como o fonocardiograma é um sinal essencialmente não estacionário — cujas propriedades acústicas variam de maneira cíclica entre sístoles, diástoles e ruídos transitórios —, determinar apenas o espectro global é insuficiente para a detecção de cardiopatias. Por exemplo, a presença de uma frequência anômala concentrada no intervalo sistólico (característica de um sopro) geraria uma resposta global na FFT indistinguível de um ruído externo breve ocorrido em outro momento da gravação. Torna-se indispensável, portanto, uma ferramenta que preserve simultaneamente a localização no tempo e a distribuição em frequência (**??**).

3.2.2 Transformada de Fourier de Tempo Curto

Para superar a perda da coordenada temporal, emprega-se a TFTC. O princípio da técnica consiste em fragmentar a série temporal em pequenos intervalos consecutivos por meio de uma janela deslizante, assumindo que, em fatias temporais curtas, o sinal comporta-se de maneira aproximadamente estacionária.

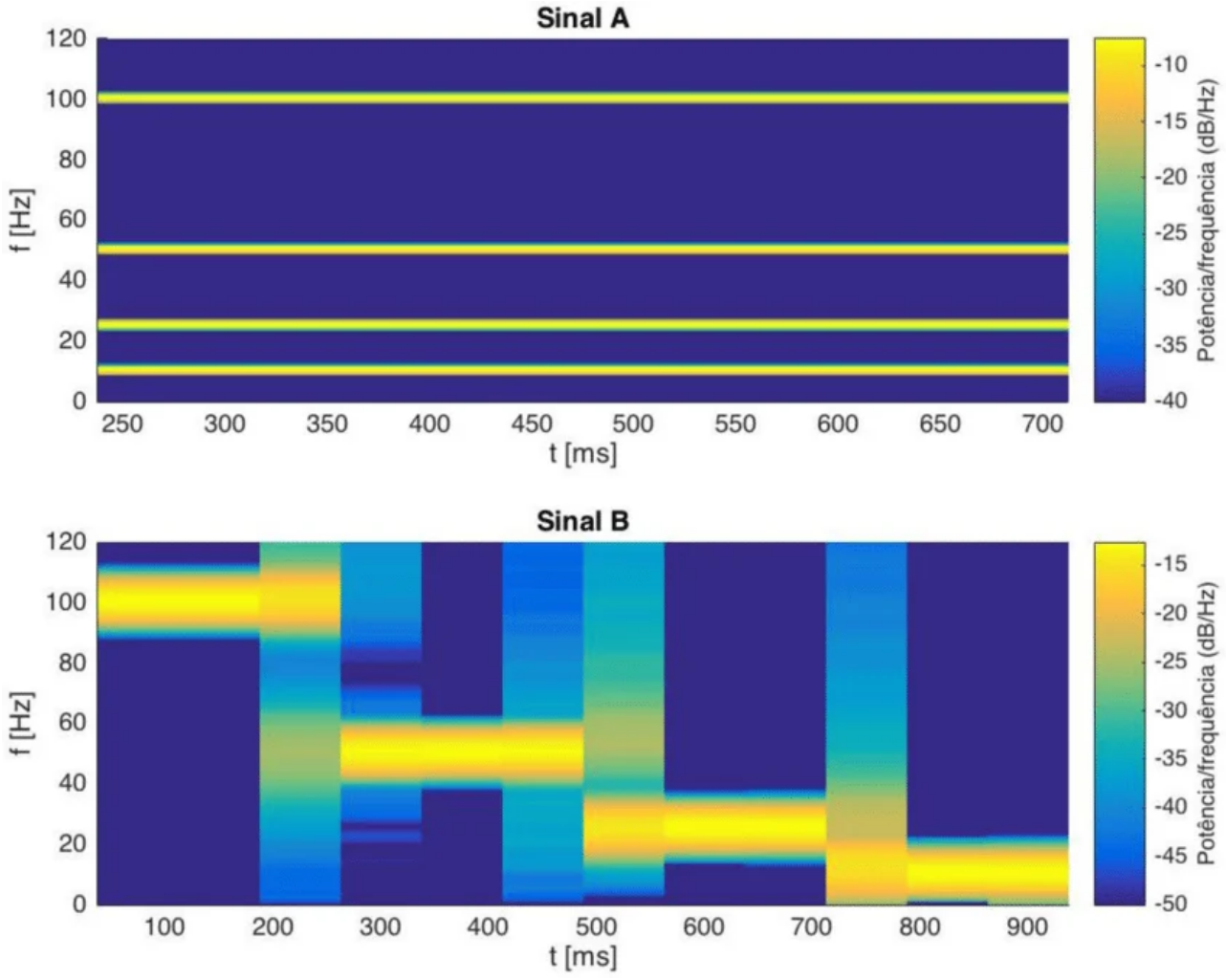
Formalmente, a TFTC de uma sequência discreta $x[n]$ é definida pela aplicação da Transformada de Fourier a cada segmento ponderado por uma função de janelamento $w[n]$:

$$X[m,k] = \sum_{n=-\infty}^{\infty} x[n]\, w[n - mR]\, e^{-j\frac{2\pi}{N}kn} \tag{2}$$

onde m representa o índice da janela ao longo do tempo, R é o deslocamento da janela (*hop size*) que dita a sobreposição temporal entre segmentos vizinhos, k corresponde ao índice discreto de frequência e $w[n]$ é uma janela simétrica (como a janela Hann), projetada para suavizar as descontinuidades nas extremidades do recorte (**??**).

Na prática, o algoritmo da TFTC desloca a janela passo a passo sobre o sinal de áudio, calculando a distribuição de frequências para cada trecho individual. A escolha da dimensão da janela impõe um compromisso físico: janelas menores produzem maior precisão sobre o instante exato em que os eventos ocorrem (alta resolução temporal), porém com menor nitidez na separação entre tons próximos (baixa resolução em frequência), e vice-versa.

Figura 1 – Representação tempo-frequência ilustrando a capacidade da STFT de localizar eventos no tempo: o Sinal A exibe componentes contínuas e constantes, enquanto o Sinal B mapeia transições espectrais em intervalos específicos.



Fonte: Queirós, Ramirez-Fernandez (2016) .

3.2.3 Representação por Espectrogramas como Entrada para Redes Neurais

A saída computada pela STFT é uma matriz bidimensional de números complexos. Para fins de visualização e extração de características em tarefas de aprendizado de máquina, avalia-se a magnitude

dessa resposta, gerando a representação denominada espectrograma:

$$S[m,k] = |X[m,k]| \tag{3}$$

Visto que as amplitudes sonoras cobrem uma faixa dinâmica bastante ampla, é praxe aplicar uma escala logarítmica (decibéis), aproximando a resposta matemática à percepção perceptual de intensidade e realçando componentes de baixa amplitude:

$$S_{\text{dB}}[m,k] = 20 \cdot \log_{10}\left(\frac{|X[m,k]|}{\max(|X|)}\right) \tag{4}$$

O resultado final dessa transformação é uma imagem bidimensional cujas dimensões ortogonais representam o tempo (eixo horizontal) e a frequência (eixo vertical), enquanto a cor ou o brilho de cada ponto reflete a intensidade da energia acústica.

Dessa forma, a conversão para espectrogramas atua como a ponte entre o processamento digital de áudio e a visão computacional: eventos pontuais como as bulhas fisiológicas manifestam-se como colunas verticais periódicas de baixa frequência, enquanto sopros e anomalias traduzem-se em faixas de textura visual contínua em faixas espectrais mais altas. Essa conformação permite tratar o problema da auscultação cardíaca diretamente por meio de Redes Neurais Convolucionais, viabilizando o reconhecimento automático de patologias a partir de seus padrões visuais característicos.

3.3 APRENDIZADO PROFUNDO E REDES NEURAIS ARTIFICIAIS

Ao contrário dos perceptrons multicamadas (MLP) tradicionais, cujas camadas densas conectam todos os nós de entrada a cada neurônio subsequente gerando uma rápida explosão na quantidade de parâmetros e descartando a estrutura espacial dos dados, as CNNs foram concebidas para processar dados com topologia de grade, como matrizes bidimensionais. Para atingir alta eficiência computacional e extrair características hierárquicas de forma robusta, as CNNs apoiam-se em três princípios arquitetônicos fundamentais: localidade espacial (campos receptivos locais), compartilhamento de pesos (*weight sharing*) e subamostragem espacial.

Uma arquitetura convolucional padrão organiza-se através da combinação sequencial e repetida das seguintes operações fundamentais:

1. **Camada de Convolução (*Convolutional Layer*):** Constitui o núcleo de extração de características (*feature extraction*) do modelo. A operação baseia-se no deslizamento sistemático de pequenos conjuntos de matrizes de pesos ajustáveis, denominados filtros ou *kernels*, sobre a totalidade da matriz de entrada. A cada passo desse deslizamento, calcula-se o produto interno (produto escalar) local entre os pesos do filtro e os valores dos pixels sob essa janela (o campo receptivo). O resultado é uma nova matriz bidimensional denominada mapa de características (*feature map*).

   Nas camadas iniciais da rede, estes filtros atuam como detectores de primitivas visuais de baixo nível (como linhas verticais, horizontais e transições bruscas de contraste). À medida que a rede se aprofunda, a combinação sucessiva destas primitivas permite que os filtros das camadas

superiores detectem padrões de textura e formas altamente complexas ,que, no contexto deste trabalho, traduzem-se nas assinaturas acústicas tempo-frequência específicas das patologias no espectrograma.

A dimensionalidade espacial da saída desta camada é controlada por dois hiperparâmetros essenciais:

- **Passo (*Stride*):** Define o tamanho do salto espacial (em pixels) que o filtro realiza ao deslizar. Passos maiores comprimem espacialmente o mapa de características resultante.
- **Preenchimento (*Padding*):** Adição de colunas e linhas de zeros nas margens da matriz de entrada (*zero-padding*). Permite controlar a redução do tamanho espacial após a convolução, garantindo, por exemplo, que a matriz de saída possua as mesmas dimensões espaciais da matriz de entrada (conhecido como *same padding*).

2. **Camada de Ativação Não Linear (*Non-linear Activation Layer*):** Como a operação matemática de convolução é estritamente linear, é necessário introduzir não-linearidade no sistema para que a rede consiga modelar e aproximar funções matemáticas complexas e mapeamentos intrincados. Imediatamente após cada convolução, aplica-se uma função de ativação não linear elemento a elemento sobre os mapas de características.

   A função predominante em arquiteturas profundas é a Unidade Linear Retificada (*Rectified Linear Unit* - ReLU), definida matematicamente como $f(x) = \max(0, x)$. A função ReLU propaga inalterados os valores positivos e anula os negativos. Além da sua extrema eficiência computacional, a ReLU mitiga severamente o problema do desvanecimento do gradiente (*vanishing gradient*) que assombrava funções de ativação clássicas (como a sigmoide ou a tangente hiperbólica), permitindo um treinamento mais rápido e estável em arquiteturas de dezenas de camadas.

3. **Camada de Agrupamento (*Pooling Layer*):** Intercaladas entre sucessivas camadas convolucionais, as camadas de agrupamento têm o propósito de subamostrar espacialmente as representações, condensando a informação e reduzindo progressivamente a resolução espacial dos mapas de características (altura e largura, mantendo a profundidade de canais intacta).

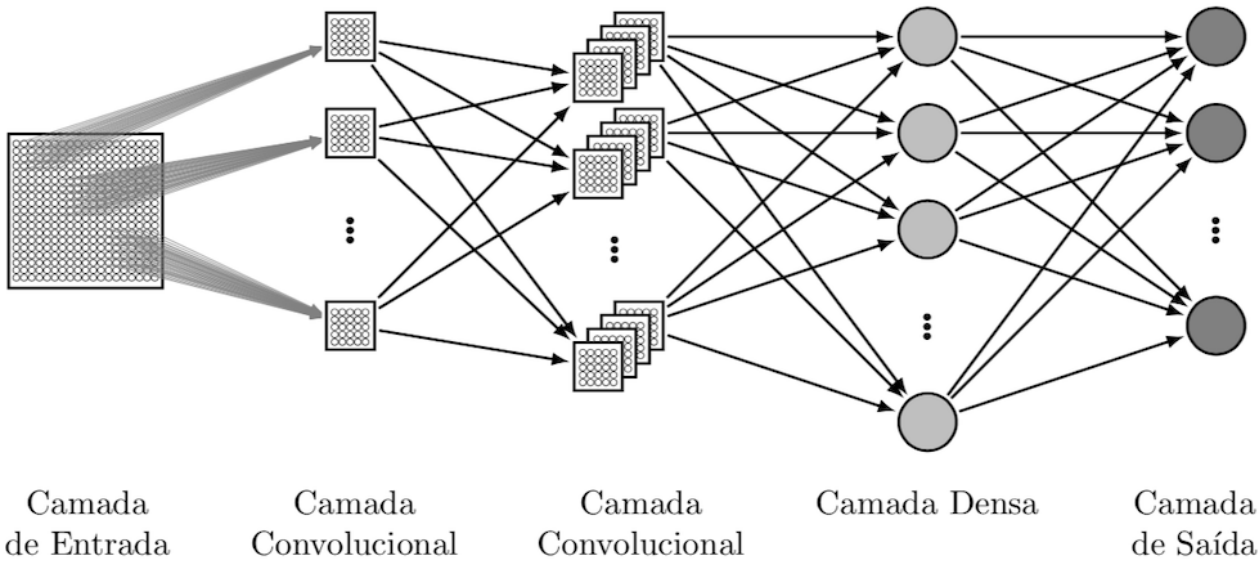
   A técnica mais utilizada é o agrupamento máximo (*Max-Pooling*), que particiona o mapa de características em pequenas regiões não sobrepostas (geralmente 2×2) e retém unicamente o valor de maior magnitude de cada região. Esta operação cumpre dois papéis vitais:

   - **Redução do Custo Computacional e *Overfitting*:** Ao diminuir a quantidade de parâmetros espaciais, reduz drasticamente o tempo de treinamento e o risco de a rede memorizar ruídos específicos do conjunto de treinamento.
   - **Invariância a Pequenas Translações:** Confere ao modelo tolerância geométrica, assegurando que o classificador continue a detectar um padrão relevante, mesmo que este sofra pequenas deslocações horizontais ou verticais na imagem de entrada.

4. **Camadas de Decisão e Classificação Final:** Após o processo de extração hierárquica pelo bloco convolucional, os mapas de características resultantes da última camada são tipicamente processados para gerar a decisão final. Nas redes mais antigas, os tensores bidimensionais eram achatados em um único vetor unidimensional (*Flatten*) e introduzidos em Camadas Totalmente Conectadas (*Fully Connected Layers*) tradicionais.

   Em arquiteturas mais modernas e profundas, aplica-se o Agrupamento Médio Global (*Global Average Pooling*), que extrai apenas a média de cada mapa de características, condensando todo o espaço bidimensional em um único valor por canal. Este vetor denso final alimenta a camada de saída (frequentemente utilizando a função *Softmax* para problemas multiclasse ou a função *Sigmoid* para classificação binária), que gera as probabilidades do diagnóstico pretendido (**??**).

Figura 2 – Representação esquemática da arquitetura de uma Rede Neural Convolucional.



Fonte: Sakurai, Rafael (2017).

3.3.1 O Problema da Degradação e o Aprendizado Residual

Com o avanço das técnicas de aprendizado profundo, consolidou-se a premissa de que o aumento na profundidade estrutural das redes neurais, isto é, o empilhamento sucessivo de camadas, amplia a capacidade do modelo de extrair representações hierárquicas mais ricas e abstratas dos dados. Entretanto, a simples adição de camadas introduziu dois obstáculos práticos de grande relevância:

- **Desvanecimento e Explosão de Gradientes (*Vanishing/Exploding Gradients*):** Durante o processo de retropropagação (**??**), as atualizações dos parâmetros dependem da multiplicação sucessiva de gradientes ao longo de toda a arquitetura. Em topologias muito profundas, essas magnitudes tendiam a decair exponencialmente em direção às primeiras camadas, impedindo o aprendizado de características primitivas, ou a crescer descontroladamente, desestabilizando o treinamento. O emprego de inicializações ponderadas e técnicas como a normalização em lote (*Batch Normalization*) atenuou grande parte dessa instabilidade, mas não solucionou o problema do desempenho em redes excessivamente profundas.

- **Problema da Degradação do Desempenho:** Ao expandir a profundidade das redes além de um limiar crítico, observou-se empiricamente que a acurácia de treino atingia a saturação e passava a se degradar rapidamente. Essa degradação não decorria de sobreajuste (*overfitting*),caso em que o erro de treino seria mínimo e apenas o de validação subiria, mas sim da incapacidade dos algoritmos de otimização convencionais baseados em gradiente de ajustar os pesos de sequências muito extensas de transformações lineares e não lineares.

Para transpor essa barreira teórica e computacional, foi introduzido o paradigma do Aprendizado Residual. Em uma rede convolucional tradicional, um bloco de camadas é forçado a aprender um mapeamento completo e direto entre a entrada e a saída desejada. No modelo residual, as camadas são encarregadas de aprender apenas a diferença (o resíduo) necessária para transformar a entrada na representação pretendida.

Na prática estrutural da rede, esse conceito é materializado por meio das chamadas conexões de atalho (*shortcut* ou *skip connections*). Essas conexões consistem em desvios que transportam a entrada original diretamente para a saída do bloco de camadas, realizando uma simples soma elemento a elemento entre o sinal bruto de entrada e a transformação computada pelas convoluções.

A grande vantagem desse mecanismo reside na garantia de que o modelo consiga, no pior dos casos, simplesmente aprender uma função de identidade, bastando zerar os pesos do bloco convolucional residual para que o sinal de entrada passe inalterado. Além disso, as conexões de atalho criam verdadeiras vias diretas para a propagação dos gradientes durante a retropropagação, eliminando os gargalos de atenuação e viabilizando o treinamento estável e eficiente de arquiteturas extremamente profundas, como a ResNet-50.

3.3.2 A Arquitetura ResNet-50 e o Bloco *Bottleneck*

No contexto deste estudo, a adoção da ResNet-50 justifica-se pelo seu notável equilíbrio entre alta capacidade de abstração e eficiência computacional (**??**). A classificação de espectrogramas acústicos derivados de fonocardiogramas exige uma arquitetura profunda o suficiente para extrair assinaturas visuais complexas e sutis, essenciais para discriminar ruídos patológicos turbulentos de bulhas fisiológicas, sem sucumbir ao problema da degradação do desempenho. Além disso, a estrutura parametrizada enxuta proporcionada por seus blocos otimizados a torna ideal para a aplicação de técnicas de transferência de aprendizado (*transfer learning*), mitigando o risco de sobreajuste (*overfitting*) frente ao volume amostral intrinsecamente limitado das bases de dados clínicas.

A arquitetura ResNet-50 constitui uma variante com 50 camadas profundas parametrizadas concebida com blocos residuais otimizados conhecidos como blocos de gargalo (*Bottleneck blocks*). Para viabilizar a eficiência computacional em topologias profundas, o bloco residual básico de duas convoluções de filtro 3×3 é substituído por uma sequência hierárquica de três convoluções consecutivas:

1. Uma convolução 1×1, responsável por realizar a redução da profundidade (número de canais) do volume de características, reduzindo o custo aritmético das operações subsequentes;
2. Uma convolução espacial 3×3, encarregada de extrair padrões bidimensionais locais na resolução intermediária reduzida;
3. Uma convolução 1×1, que atua restaurando a dimensão original (ou quadruplicando-a) nos mapas de saída.

Em conjunto, a camada de normalização em lote (*Batch Normalization* – BN) é aplicada imediatamente após cada operação de convolução e antes da aplicação da não linearidade ReLU, assegurando distribuição controlada de médias e variâncias nos lotes e acelerando a convergência.

A estrutura residual pode operar de duas formas:

- **Atalho de Identidade Direta (*Identity Shortcut*):** Aplicado quando a dimensão espacial e o volume de canais de entrada x coincidem estritamente com os de $\mathcal{F}(x)$, somando x de maneira direta;
- **Atalho com Projeção Linear (*Projection Shortcut*):** Quando ocorre redução espacial (por meio de convoluções com *stride* 2) ou acréscimo de canais nas transições entre os estágios da ResNet, aplica-se uma convolução 1×1 no próprio vetor de entrada x para compatibilizar suas dimensões espaciais e matriciais antes da soma residual.

A topologia da ResNet-50 organiza-se em um estágio convolucional inicial de entrada (7×7 com *stride* 2), seguido de quatro macroestágios residuais que totalizam 16 blocos *bottleneck* encadeados, culminando em uma camada de *Global Average Pooling* que condensa os mapas espaciais em um vetor unidimensional de descritores antes da projeção na camada de decisão final.

3.3.3 Transferência de Aprendizado (*Transfer Learning*) e Ajuste Fino

O treinamento bem-sucedido de redes neurais com dezenas de milhões de parâmetros ajustáveis exige, teoricamente, volumes massivos de amostras rotuladas para evitar sobreajuste severo. Em tarefas médicas baseadas em sinais biológicos como fonocardiogramas, a disponibilidade de dados anotados por especialistas é frequentemente reduzida.

Para mitigar esse desafio amostral, adota-se amplamente a técnica de **Transferência de Aprendizado** (*Transfer Learning*). Nesse paradigma, uma rede neural é inicialmente treinada em uma base de dados externa de larga escala, tipicamente o *ImageNet*, composto por mais de 1,4 milhão de imagens distribuídas em mil classes visuais genéricas. Durante essa fase prévia, os núcleos dos filtros das camadas inferiores aprendem extratores de primitivas visuais invariantes, como detecção de bordas, transições de contraste, texturas e curvaturas espaciais.

Posteriormente, esses parâmetros pré-treinados são transferidos para o problema-alvo (neste contexto, a classificação de espectrogramas derivados de PCG). No processo de ajuste fino (*Fine-Tuning*):

- A camada densa final original do *ImageNet* é descartada e substituída por uma nova camada de saída compatível com a dimensionalidade do problema clínico (como uma camada de neurônio único para discriminação binária normal versus anormal);
- As camadas iniciais de representação convolucional podem ser temporariamente congeladas para preservar os filtros primitivos genéricos ou ajustadas simultaneamente com taxas de aprendizado (*learning rates*) reduzidas, adaptando as texturas tempo-frequência presentes nos espectrogramas acústicos sem destruir os pesos previamente aprendidos.

Dessa forma, o *transfer learning* viabiliza convergências mais rápidas durante as épocas de treino e proporciona maior robustez de generalização, mesmo perante cenários clínicos com limites amostrais intrínsecos.

3.4 MÉTRICAS DE DESEMPENHO E AVALIAÇÃO DE MODELOS CLÍNICOS

A avaliação quantitativa de modelos de aprendizado profundo aplicados ao domínio biomédico exige métricas estatísticas criteriosas. Diferentemente de aplicações convencionais de visão computacional, no apoio ao diagnóstico médico assistido por computador o custo associado a diferentes tipos de erro é assimétrico: classificar erroneamente um sinal patológico como normal (falso negativo) pode acarretar a ausência de tratamento oportuno para um paciente enfermo, enquanto rotular um sinal saudável como anômalo (falso positivo) induz à realização de exames confirmatórios adicionais de maior custo.

3.4.1 Matriz de Confusão

A base analítica para a computação das métricas de desempenho em tarefas de classificação binária fundamenta-se na **Matriz de Confusão**. Considerando uma convenção diagnóstica onde a classe positiva designa a presença de anomalia cardíaca (ou sopro patológico) e a classe negativa denota a normalidade auscultatória fisiológica, as predições do modelo distribuem-se em quatro quadrantes fundamentais:

- **Verdadeiro Positivo (VP):** Registros patológicos corretamente inferidos pelo modelo como anômalos;
- **Falso Negativo (FN):** Registros patológicos erroneamente classificados pelo modelo como normais (Erro Tipo II);
- **Falso Positivo (FP):** Registros saudáveis erroneamente categorizados como anômalos (Erro Tipo I);
- **Verdadeiro Negativo (VN):** Registros fisiologicamente saudáveis corretamente identificados como normais.

3.4.2 Métricas Estatísticas Derivadas

A partir das contagens discretas consolidadas na matriz de confusão, formalizam-se as métricas clássicas de avaliação de desempenho:

3.4.2.1 Acurácia (Acc):

Quantifica a proporção global de predições corretas em relação à totalidade de instâncias avaliadas:

$$Acc = \frac{VP + VN}{VP + VN + FP + FN} \tag{5}$$

Apesar de ser uma métrica intuitiva e amplamente difundida, a acurácia global torna-se estatisticamente enviesada perante bases de dados que exibem desbalanceamento acentuado entre classes, onde um modelo trivial que sempre preveja a classe majoritária alcançará alto índice numérico sem possuir capacidade diagnóstica real.

3.4.2.2 Sensibilidade ou Revocação (*Sensitivity* / *Recall*):

Também denominada Taxa de Verdadeiros Positivos (TVP), expressa a fração de pacientes verdadeiramente portadores de anomalias cardíacas que o modelo foi capaz de detectar com sucesso:

$$Sensibilidade = \frac{VP}{VP + FN} \tag{6}$$

Em triagens clínicas cardiovasculares preliminares, a sensibilidade é uma métrica prioritária, pois sua maximização corresponde diretamente à minimização das omissões diagnósticas ($FN \to 0$).

3.4.2.3 Especificidade (*Specificity*):

Reflete a fração de registros normais corretamente identificados como ausentes de patologia, sendo complementar à Taxa de Falsos Positivos ($TFP = 1 - Especificidade$):

$$Especificidade = \frac{VN}{VN + FP} \tag{7}$$

Valores elevados de especificidade asseguram que indivíduos saudáveis não sejam desnecessariamente encaminhados para condutas clínicas invasivas.

3.4.2.4 Precisão (*Precision*):

Mensura a proporção de casos verdadeiramente anômalos dentre todos aqueles que o classificador rotulou como positivos:

$$Precisão = \frac{VP}{VP + FP} \tag{8}$$

3.4.2.5 Pontuação F_1 (F_1-score):

Representa a média harmônica entre precisão e sensibilidade, penalizando modelos com desequilíbrio acentuado entre esses dois critérios:

$$F_1 = 2 \cdot \frac{Precisão \cdot Sensibilidade}{Precisão + Sensibilidade} = \frac{2 \cdot VP}{2 \cdot VP + FP + FN} \tag{9}$$

O F_1-score constitui uma métrica robusta e equilibrada para cenários médicos caracterizados pela distribuição desigual de amostras normais e patológicas.

4 METODOLOGIA

4.1 BASE DE DADOS E CARACTERIZAÇÃO AMOSTRAL

Para a condução deste trabalho, utilizou-se a base de dados pública *CirCor DigiScope Phonocardiogram Dataset*, disponibilizada no repositório PhysioNet e originalmente desenvolvida para o desafio internacional *George B. Moody PhysioNet Challenge 2022* (**??**). A base reúne registros acústicos de auscultação cardíaca voltados ao rastreamento e diagnóstico de cardiopatias na população pediátrica.

4.1.1 Origem e Protocolo de Coleta dos Dados

Os dados foram coletados em duas campanhas distintas de triagem médica em comunidades rurais do Nordeste do Brasil, abrangendo crianças e adolescentes com idades entre 0 e 21 anos. As gravações acústicas de FCG foram captadas por meio de estetoscópios eletrônicos digitais com tecnologia de transdutor piezoelétrico (*Littmann 3200*), operando a uma taxa de amostragem original de 4000 Hz com quantização de 16 bits em formato monoauricular (`.wav`).

Para cada paciente, foram registradas gravações sequenciais nos quatro principais focos auscultatórios cardíacos (aórtico, pulmonar, tricúspide e mitral), com duração típica variando entre 5 e 70 segundos por arquivo, como mostrado na figura 5 . Em determinados casos, o conjunto também inclui áudios complementares obtidos em outras posições do tórax.

4.1.2 Anotações Clínicas e Definição das Classes de Estudo

A totalidade dos sinais de áudio presentes na base foi rotulada e anotada independentemente por uma equipe de especialistas cardiologistas e pediatras, utilizando exames confirmatórios de ecocardiograma (*gold standard*) para validação do diagnóstico clínico definitivo.

No conjunto de metadados da base CirCor, cada gravação auscultatória é acompanhada pela caracterização da presença de sopros cardíacos (*murmurs*), categorizada originalmente em três condições diagnósticas:

- **Presente (*Present*):** Registros em que sopros sistólicos, diastólicos ou contínuos foram claramente identificados e caracterizados clínica e acusticamente;
- **Ausente (*Absent*):** Registros estritamente normais, compostos apenas pelo padrão fisiológico de bulhas regulares (S_1 e S_2), sem indícios de turbulência patológica;
- **Desconhecido / Inconclusivo (*Unknown*):** Gravações com qualidade acústica comprometida por interferências externas intensas ou nas quais os especialistas não puderam determinar conclusivamente a presença de anomalias.

Para viabilizar a formulação do problema como uma tarefa supervisionada de classificação binária pelo modelo de aprendizado profundo, foram selecionadas as gravações categorizadas nas classes **Normal** (derivada dos sinais ausentes de sopro) e **Anormal** (derivada dos sinais com presença

confirmada de sopros cardíacos). Registros classificados como inconclusivos foram excluídos da etapa de treinamento para mitigar a propagação de ruídos de rotulagem na otimização da rede neural convolucional.

4.2 AMBIENTE COMPUTACIONAL E FERRAMENTAS DE DESENVOLVIMENTO

O desenvolvimento deste trabalho foi conduzido em ambiente computacional baseado na linguagem de programação Python (versão 3.10), utilizando a plataforma em nuvem Google Colaboratory (Colab). A escolha dessa infraestrutura viabilizou o acesso a unidades de processamento gráfico (GPUs dedicadas, como NVIDIA T4) com suporte a aceleração por hardware CUDA, essencial para reduzir a latência de processamento durante o treinamento iterativo de modelos convolucionais profundos.

O fluxo metodológico — desde o particionamento temporal das séries temporais até a convergência dos classificadores — apoiou-se na integração das seguintes bibliotecas e pacotes especializados:

- **PyDub:** Empregada para a manipulação rápida e segmentação temporal dos arquivos brutos de áudio em formato `.wav`, viabilizando o corte preciso de intervalos temporais fixos;
- **Librosa:** Biblioteca de referência para processamento digital de áudio e música, utilizada na leitura das séries temporais discretizadas, na extração da Transformada de Fourier de Tempo Curto (STFT) e na conversão dos espectros para escala logarítmica de decibéis;
- **NumPy e Pandas:** Utilizados para a manipulação de tensores multidimensionais, operações de álgebra linear matricial e estruturação dos metadados clínicos associados aos registros da base CirCor;
- **Matplotlib:** Empregada na renderização gráfica e exportação automatizada das matrizes espectrais tempo-frequência na forma de imagens bidimensionais em formato `.png`;
- **PyTorch e Torchvision:** Framework de aprendizado profundo selecionado para a instanciação da arquitetura ResNet-50 com pesos pré-treinados, definição dos *DataLoaders*, cálculo da propagação para frente (*forward pass*), retropropagação de gradientes e otimização dos pesos;
- **Scikit-Learn:** Utilizada no particionamento estratificado dos conjuntos de dados e no cálculo analítico das métricas quantitativas de desempenho (matriz de confusão, acurácia, sensibilidade, especificidade e pontuação F_1).

4.3 PIPELINE DE PRÉ-PROCESSAMENTO DE SINAIS ACÚSTICOS

Para transformar as séries temporais acústicas brutas em representações adequadas para a arquitetura convolucional, desenvolveu-se um fluxo de processamento dividido em duas etapas fundamentais: segmentação temporal uniforme e geração de representações tempo-frequência via STFT.

4.3.1 Segmentação e Padronização Temporal

As gravações originais da base CirCor DigiScope apresentam durações variáveis (entre 5 e 70 segundos). Para mitigar ruídos de borda e impor uma dimensão temporal homogênea para todas as amostras, cada arquivo de áudio em formato `.wav` foi submetido a uma rotina de recorte temporal utilizando a biblioteca PyDub. Extraiu-se de cada sinal um segmento contínuo de 5 segundos, delimitado entre o quinto e o décimo segundo de gravação ($t \in [5\text{ s}, 10\text{ s}]$) para áudios ácima de 15s e para os demais, o recorte foi feito apenas nos 5 segundos iniciais. Este procedimento assegurou a eliminação das perturbações de contato iniciais e garantiu que cada segmento contivesse múltiplos ciclos cardíacos completos e regulares sob regime estacionário local.

4.3.2 Geração dos Espectrogramas via STFT

A partir dos sinais temporais normalizados de 5 segundos, utilizou-se a biblioteca Librosa para converter as oscilações acústicas unidimensionais em matrizes bidimensionais de tempo-frequência por meio da Transformada de Fourier de Tempo Curto. O processamento espectral compreendeu as seguintes etapas:

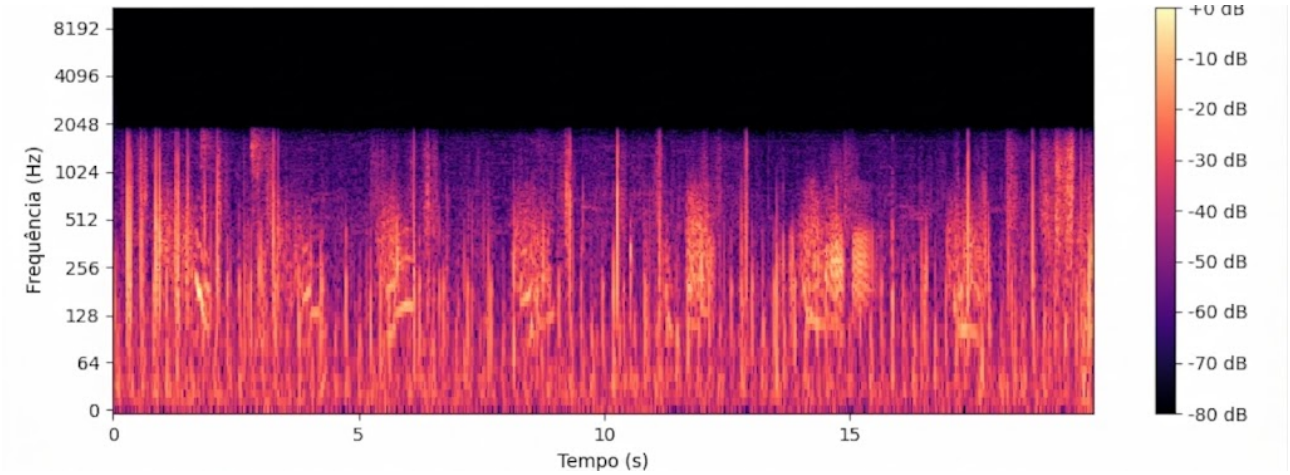
- **Cálculo da STFT Linear:** Foi executada a decomposição do sinal por meio da função `librosa.stft`, segmentando a série temporal em janelas móveis com função Hann e calculando a Transformada Rápida de Fourier (FFT) em cada trecho. Em seguida, extraiu-se a magnitude dos valores complexos gerados, consolidando a matriz do espectrograma linear de amplitude.

- **Conversão para Escala Logarítmica (Decibéis):** Para comprimir a ampla faixa dinâmica das amplitudes acústicas e acentuar padrões de menor intensidade associados a ruídos patológicos, as magnitudes lineares foram convertidas para a escala em decibéis utilizando a função `librosa.amplitude_to_db`, adotando a potência máxima de pico como referência normalizadora.

- **Renderização e Exportação das Imagens:** Empregando as rotinas gráficas de `librosa.display.specshow` em conjunto com a biblioteca Matplotlib, as matrizes espectrais foram renderizadas e exportadas como imagens no formato `.png`.

Por fim, os espectrogramas foram carregados com resolução espacial em canal único (escala de cinza). Esta conversão final justifica-se pelo fato de a arquitetura ResNet-50 ter sido pré-treinada com imagens do cotidiano (base ImageNet). Ao passarmos as imagens para escala de cinza, evitou-se que a rede criasse dependências ou interpretasse erroneamente os mapas de cores artificiais gerados computacionalmente pela biblioteca gráfica, focando-se estritamente nos padrões de textura e intensidade tempo-frequência.

Figura 3 – Exemplo de um Espectrograma de potência log-frequência gerado



Fonte: Elaborado pelo autor (2026).

4.4 CRITÉRIOS DE AVALIAÇÃO E VALIDAÇÃO EXPERIMENTAL

Para quantificar a eficácia da ResNet-50 na discriminação dos espectrogramas acústicos, o modelo foi avaliado sobre o conjunto de teste independente (*test split*), que não participou do ajuste de pesos durante o treinamento. A análise de desempenho apoiou-se no cômputo da matriz de confusão e na extração de métricas estatísticas clássicas de classificação binária via biblioteca *Scikit-Learn*.

4.4.1 Matriz de Confusão

Foi utilizada a matriz de confusão permite tabular a correspondência entre as categorias reais dos registros e as predições geradas pela rede neural a partir da função `argmax` aplicada sobre as probabilidades de saída como já detalhada na seção 3.4.

4.4.2 Métricas Estatísticas de Desempenho

A quantificação do desempenho preditivo do modelo foi conduzida mediante a extração computacional das métricas estatísticas detalhadas na Seção 3.4. A partir das probabilidades estimadas pela camada final de decisão da ResNet-50 no conjunto de teste independente (600 amostras), realizou-se a atribuição de classes pela operação argmax para consolidação dos rótulos inferidos. Com base na correspondência entre os rótulos preditos e os diagnósticos clínicos de referência da base CirCor, empregou-se a função `classification_report` da biblioteca Scikit-Learn. Por meio dessa rotina, computaram-se de forma automatizada a acurácia global, além da sensibilidade (recall), especificidade, precisão e pontuação F_1 para cada uma das classes de estudo (Normal e Anormal). Os resultados foram consolidados sob as abordagens de média macro e média ponderada pelo suporte amostral (weighted average), permitindo avaliar o comportamento do classificador sob as condições de desbalanceamento de classes inerentes ao cenário clínico e balizar as discussões comparativas com a triagem médica humana apresentadas no Capítulo 5.

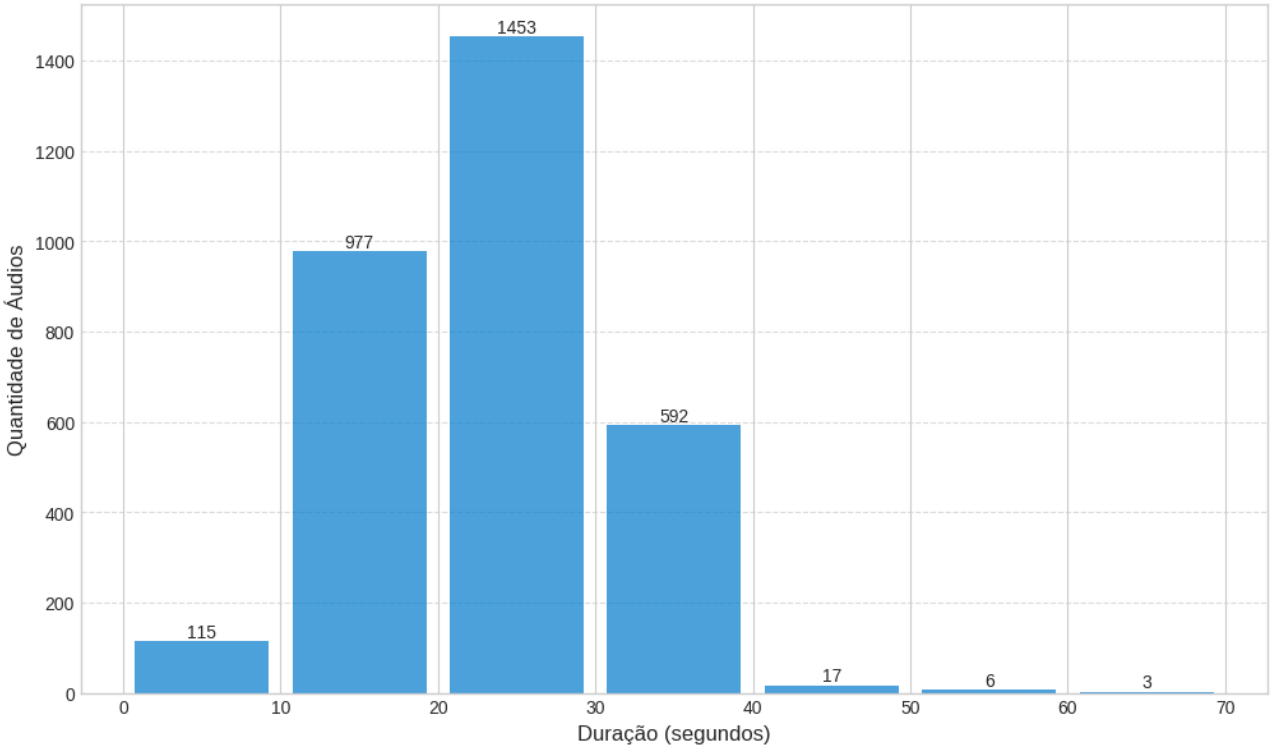
5 RESULTADOS E DISCUSSÕES

Neste capítulo, apresentam-se e discutem-se os resultados obtidos na classificação dos registros fonocardiográficos pela rede neural convolucional ResNet-50. A análise contempla o comportamento das funções de perda durante o treinamento e a avaliação quantitativa do classificador sobre o conjunto de teste independente, composto por 600 amostras (500 fisiológicas e 100 patológicas).

5.1 ANÁLISE DO PRÉ-PROCESSAMENTO DOS DADOS

A análise preliminar do conjunto de dados evidenciou uma forte assimetria nas durações das gravações originais. Como ilustrado na Figura 4, a esmagadora maioria dos áudios concentra-se nas faixas de 10 a 20 segundos e de 20 a 30 segundos, validando a necessidade da etapa de segmentação e padronização temporal.

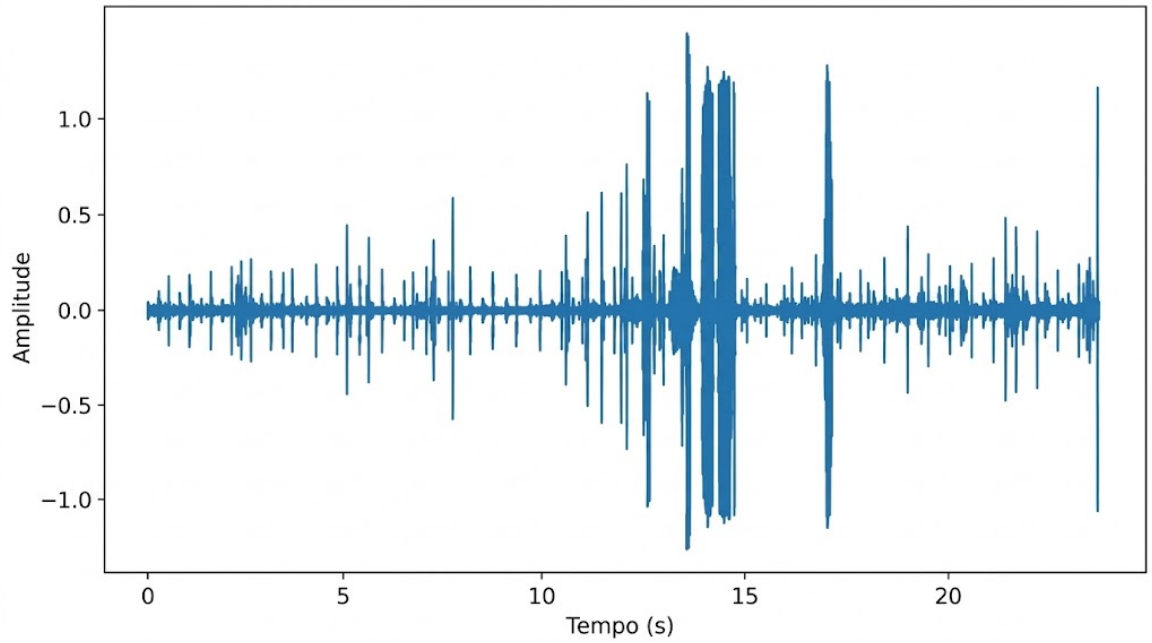
Figura 4 – Distribuição da duração dos áudios cardíacos na base de dados, justificando a necessidade de padronização.



Fonte: Elaborado pelo autor (2026).

Ademais, a inspeção dos sinais brutos no domínio do tempo revelou a presença frequente de transientes espúrios de alta amplitude nos instantes iniciais e finais da gravação. Estes artefatos, causados pelo atrito ou pela acomodação mecânica do estetoscópio na pele do paciente, podem ser visualizados na Figura 5, justificando o recorte padronizado entre o quinto e o décimo segundo de captação.

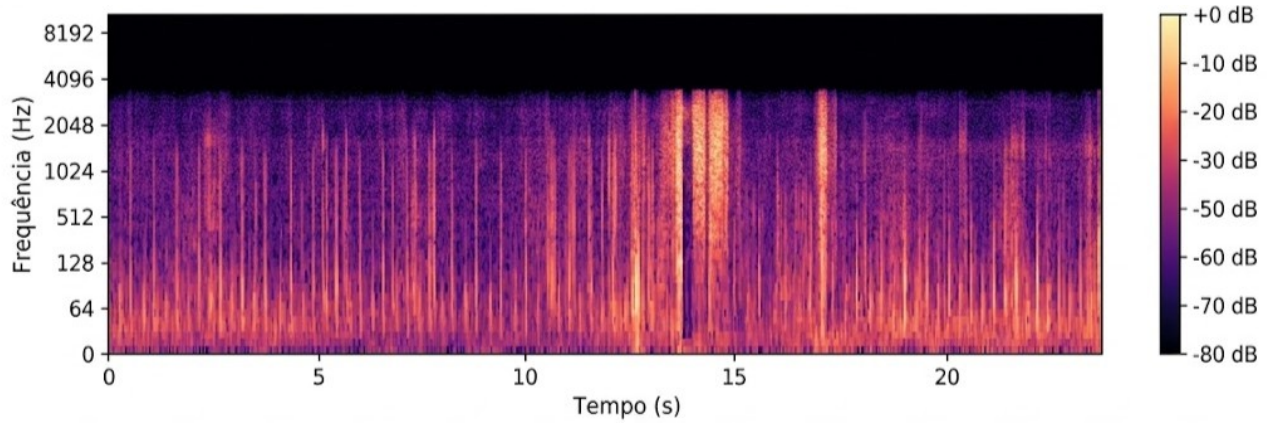
Figura 5 – Sinal acústico no domínio do tempo evidenciando artefatos e transientes de alta amplitude (Amplitude vs. Tempo em segundos).



Fonte: Elaborado pelo autor (2026).

A aplicação da Transformada de Fourier de Tempo Curto (STFT) permitiu traduzir as oscilações filtradas para o domínio tempo-frequência. A Figura 6 apresenta o espectrograma resultante em escala logarítmica (decibéis), mapeando a densidade de potência do sinal e evidenciando os padrões de turbulência acústica associados aos ciclos cardíacos.

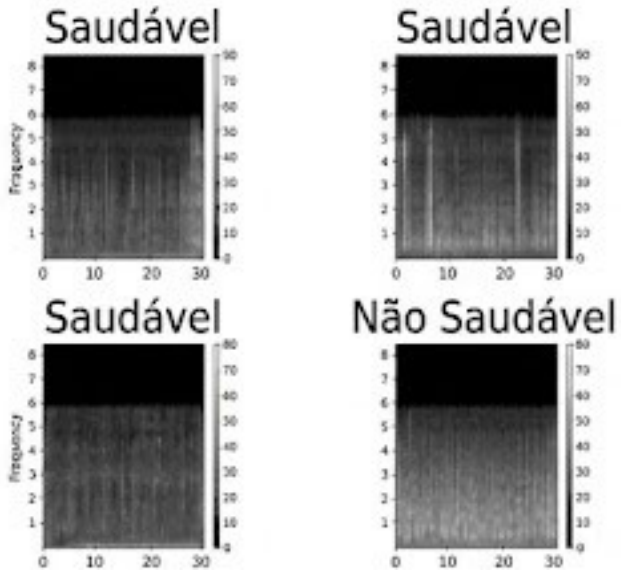
Figura 6 – Espectrograma de potência log-frequência de um sinal segmentado (Frequência em Hz vs. Tempo em segundos).



Fonte: Elaborado pelo autor (2026).

Finalmente, a Figura 7 demonstra o resultado da conversão destes espectrogramas para escala de cinza (canal único). Esta etapa garantiu a compatibilização adequada com a rede ResNet-50 pré-treinada. As imagens geradas pelo código apresentam as assinaturas acústicas consolidadas e anotadas (Healthy e Unhealthy), prontas para integrar os tensores de entrada do modelo.

Figura 7 – Amostra de lote de espectrogramas convertidos para escala de cinza, classificados como normais (Saudável) e anormais (Não Saudável), submetidos à rede neural.

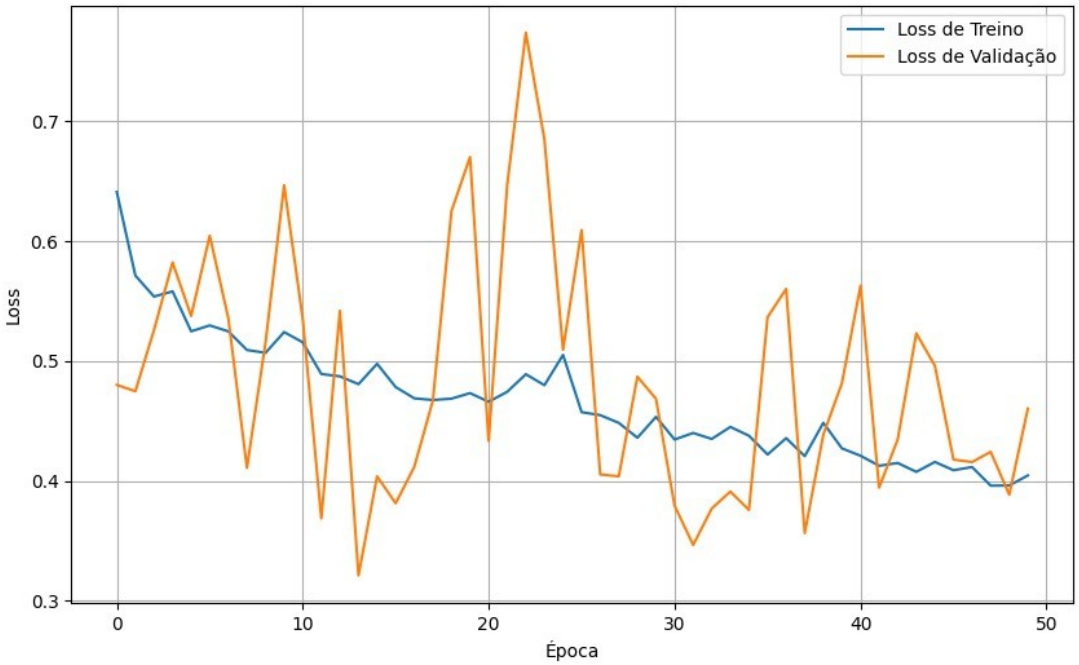


Fonte: Elaborado pelo autor (2026).

5.2 CONVERGÊNCIA E DINÂMICA DE TREINAMENTO

O comportamento do modelo durante a etapa de otimização foi acompanhado pelo monitoramento da função de perda (*loss*) de treino e de validação ao longo de 50 épocas, conforme ilustrado na Figura 8.

Figura 8 – Evolução da função de perda (*loss*) de treinamento e de validação por época.



Fonte: Elaborado pelo autor (2026).

A curva de perda de treinamento exibiu decaimento contínuo e estável, partindo de aproxima-

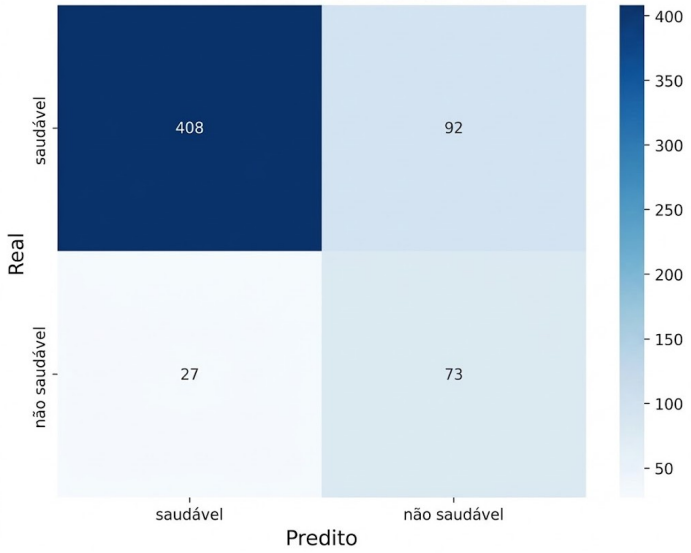
damente $0{,}64$ e estabilizando-se em torno de $0{,}40$ nas épocas finais. Esse comportamento atesta a convergência adequada do otimizador Adam sob a taxa de aprendizado adotada (10^{-4}), demonstrando a assimilação gradual das características dos espectrogramas pelas camadas densas superiores acopladas.

Por outro lado, a curva de validação apresentou oscilações de maior amplitude ao longo do processo, alternando picos e vales entre $0{,}32$ e $0{,}77$, convergindo nas épocas finais para valores próximos a $0{,}45$. Essa variabilidade decorre primariamente do desbalanceamento natural da base e da heterogeneidade acústica intrínseca aos registros clínicos pediátricos (como ruídos fisiológicos e de respiração). Não obstante as flutuações pontuais, a perda de validação manteve uma média assintótica condizente com a curva de treino, não caracterizando divergência acentuada por sobreajuste severo (*overfitting*).

5.3 MATRIZ DE CONFUSÃO NO CONJUNTO DE TESTE

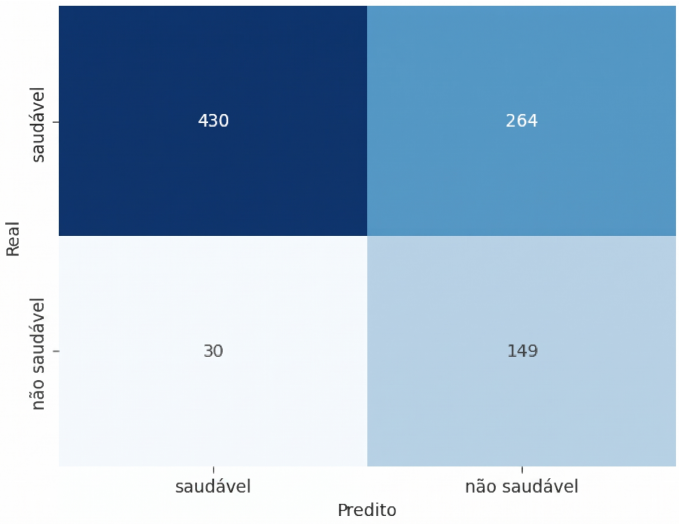
A capacidade de discriminação diagnóstica da ResNet-50 sobre as 600 amostras inéditas de teste é descrita pela matriz de confusão apresentada na Figura 9.

Figura 9 – Matriz de confusão resultante da avaliação do modelo no conjunto de teste.



Fonte: Elaborado pelo autor (2026).

Figura 10 – Matriz de confusão refletindo o desempenho do diagnóstico clínico (médico) na base CirCor.



Fonte: Elaborado pelo autor (2026).

A distribuição numérica entre predições e rótulos reais consolida-se em:

- **Verdadeiros Negativos (VN):** 408 registros normais foram classificados corretamente como saudáveis (*healthy*);
- **Falsos Positivos (FP):** 92 registros fisiológicos foram erroneamente classificados como doentes (*unhealthy*), incorrendo em Erro Tipo I;
- **Falsos Negativos (FN):** 27 registros com anomalias auscultatórias confirmadas foram categorizados equivocadamente como saudáveis, incorrendo em Erro Tipo II;

- **Verdadeiros Positivos (VP):** 73 registros patológicos foram assertivamente detectados pela rede como doentes.

Para contextualizar o desempenho do modelo computacional frente à prática clínica, introduziu-se uma avaliação comparativa com a matriz de confusão referente ao diagnóstico humano (erro médico) na triagem de sopros da base CirCor, apresentada na Figura 10.

A matriz clínica evidencia a seguinte distribuição empírica de acertos e erros médicos em um universo de 873 avaliações:

- **Verdadeiros Negativos (VN):** 430;
- **Falsos Positivos (FP):** 264;
- **Falsos Negativos (FN):** 30;
- **Verdadeiros Positivos (VP):** 149.

Como os conjuntos avaliados apresentam dimensões amostrais distintas (600 inferências para a ResNet-50 contra 873 para a triagem clínica), a comparação direta de valores absolutos induz a vieses interpretativos. Portanto, o contraste entre os métodos deve fundamentar-se nas proporções relativas (taxas e percentuais) dentro de cada classe real.

Na avaliação dos Erros Tipo I (Falsos Positivos), analisa-se a proporção de alarmes falsos gerados dentro do total de pacientes que eram verdadeiramente saudáveis. No diagnóstico médico, dos 694 pacientes saudáveis, 264 foram classificados erroneamente como doentes, o que representa uma Taxa de Falsos Positivos de 38,0% (especificidade de 62,0%). Em contrapartida, a ResNet-50 avaliou 500 pacientes saudáveis e errou em 92 casos, resultando em uma Taxa de Falsos Positivos substancialmente menor, de 18,4% (especificidade de 81,6%). Isso demonstra que o modelo computacional é significativamente mais específico, reduzindo pela metade a proporção de diagnósticos falso-positivos, o que na prática mitigaria a sobrecarga de encaminhamentos e ecocardiogramas desnecessários no sistema de saúde.

Quanto aos Erros Tipo II (Falsos Negativos), avalia-se a proporção de patologias omitidas dentro do total de pacientes verdadeiramente doentes. No diagnóstico clínico, dos 179 pacientes com anomalias, o médico não detectou 30 casos, resultando em uma Taxa de Falsos Negativos de 16,8% (sensibilidade de 83,2%). Já o modelo de inteligência artificial, de um total de 100 pacientes doentes, omitiu 27 casos, gerando uma Taxa de Falsos Negativos de 27,0% (sensibilidade de 73,0%).

Conclui-se que a triagem humana apresenta maior sensibilidade, deixando passar proporcionalmente menos casos patológicos do que a rede neural. Contudo, o modelo ResNet-50 oferece um equilíbrio diagnóstico valioso: ele sacrifica uma pequena margem de sensibilidade em troca de um ganho massivo em especificidade, demonstrando grande potencial como ferramenta de filtragem inicial para auxiliar a tomada de decisão médica.

5.4 MÉTRICAS CONSOLIDADAS DE DESEMPENHO

As métricas calculadas a partir da inferência no conjunto de teste estão organizadas na Tabela 1.

Tabela 1 – Métricas estatísticas de desempenho obtidas pela ResNet-50 no conjunto de teste.

Classe	**Precisão**	**Sensibilidade (Recall)**	**Pontuação F_1**	**Suporte**
Saudável (*Healthy*)	0,94	0,82	0,87	500
Doente (*Unhealthy*)	0,44	0,73	0,55	100
Acurácia Global		0,80 **(**80,0%**)**		
Média Macro	0,69	0,77	0,71	600
Média Ponderada	0,86	0,80	0,82	600

Fonte: Elaborado pelo autor (2026).

A análise detalhada dos dados consolidados revela facetas importantes sobre o comportamento decisório da arquitetura computacional frente ao desbalanceamento inerente aos dados clínicos. Observa-se que o classificador apresenta um desempenho altamente robusto na identificação de instâncias da classe Saudável, alcançando uma precisão de 0,94 e uma pontuação F_1 de 0,87. Este alto valor de precisão indica que, quando a rede neural prediz a ausência de anomalias, o grau de confiabilidade da predição é excelente.

Por outro lado, as métricas associadas à classe Doente (patológica) refletem os desafios impostos pelo suporte amostral reduzido, com apenas 100 amostras patológicas contra 500 fisiológicas. A precisão de 0,44 nesta categoria é matematicamente tracionada para baixo pelo volume de falsos positivos gerados, mas a sensibilidade de 0,73 demonstra que uma parcela representativa dos casos verdadeiramente anômalos continua sendo capturada com sucesso pela rede.

Adicionalmente, a divergência entre os métodos de cálculo das médias evidencia o peso da classe majoritária no sistema global. A média macro, que avalia as categorias independentemente da quantidade de amostras, resulta num F_1-score de 0,71. Este valor expõe a dificuldade residual do modelo em mapear a classe minoritária (anormal) com a mesma eficácia da majoritária.

6 CONCLUSÃO

Este trabalho investigou com sucesso a aplicação de técnicas de aprendizado profundo, especificamente a arquitetura convolucional ResNet-50, na detecção e classificação automática de anomalias em sinais de fonocardiograma (PCG). Através da utilização da base de dados pediátrica CirCor DigiScope, o estudo validou a viabilidade da metodologia de pré-processamento, que consistiu na segmentação temporal dos áudios e em sua posterior conversão para espectrogramas no domínio tempo-frequência via Transformada de Fourier de Tempo Curto (STFT). A avaliação do modelo no conjunto de teste independente demonstrou resultados consistentes, alcançando uma acurácia global de 80,0% e uma pontuação F_1 ponderada de 0,82. O destaque do desempenho da rede neural residiu na sua elevada especificidade diagnóstica quando comparada proporcionalmente à triagem médica humana. Enquanto o diagnóstico clínico apresentou uma Taxa de Falsos Positivos de 38,0%, o classificador computacional conseguiu reduzir essa métrica para 18,4%. Apesar de a avaliação clínica tradicional ter mantido uma sensibilidade superior em relação à rede (83,2% contra 73,0%), conclui-se que o modelo proposto oferece um excelente e valioso equilíbrio diagnóstico. Ao reduzir expressivamente os alarmes falsos, a inteligência artificial atua como um filtro inicial promissor, capaz de mitigar encaminhamentos médicos e exames ecocardiográficos desnecessários, aliviando a sobrecarga do sistema de saúde sem comprometer o rastreio patológico. Para trabalhos futuros, sugere-se a investigação de técnicas de aumento de dados (data augmentation) e estratégias para lidar com o desbalanceamento intrínseco das classes, visando aprimorar a sensibilidade do modelo. Adicionalmente, a exploração de outras arquiteturas convolucionais recentes e a integração de dados multimodais, combinando os espectrogramas acústicos com os metadados clínicos dos pacientes, representam caminhos promissores para elevar ainda mais a precisão e a robustez do diagnóstico assistido por computador.

REFERÊNCIAS

ESTEVA, A. *et al*. A guide to deep learning in healthcare. **Nature medicine**, Nature Publishing Group US, v. 25, n. 1, p. 24–29, 2019.

GOODFELLOW, I. *et al*. **Deep learning**. [*S. l.*]: MIT press Cambridge, 2016. v. 1.

HAYKIN, S. **Neural networks**: a comprehensive foundation. [*S. l.*]: Prentice hall PTR, 1994.

HE, K. *et al*. Deep residual learning for image recognition. *In*: THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2016, Las Vegas. **Proceedings** [...].[*S. l: s. n*], 2016. p. 770–778.

LENG, S. *et al*. The electronic stethoscope. **BioMedical Engineering OnLine**, Springer, v. 14, n. 1, p. 66, 2015.

OLIVEIRA, J. *et al*. The CirCor DigiScope dataset: From murmur detection to murmur classification. **IEEE Journal of Biomedical and Health Informatics**, IEEE, v. 26, n. 6, p. 2524–2535, 2021.

OPPENHEIM, A. V.; SCHAFER, R. W. **Discrete-time signal processing**. 3rd. ed. [*S. l.*]: Pearson, 2010.

QUEIRÓS, Ricardo; RAMIREZ-FERNANDEZ, Francisco. **Técnicas de processamento digital de sinais acústicos**: Caracterização do Meio Submarino. Disponível em: https://www.researchgate.net/figure/Figura-8-Resultado-da-aplicao-da-Transformada-de-Fourier-de-Tempo-Curto-STFT-aos_fig5_309558290. Acesso em: 28 set. 2026.

RANGAYYAN, R. M. **Biomedical signal analysis**. 2nd. ed. [*S. l.*]: John Wiley & Sons, 2015. Disponível em: https://www.researchgate.net/figure/Figura-8-Resultado-da-aplicao-da-Transformada-de-Fourier-de-Tempo-Curto-STFT-aos_fig5_309558290. Acesso em: 28 set. 2026.

RUBIN, J. *et al*. Recognizing abnormal heart sounds using deep learning. *In*:INTERNATIONAL JOINT CONFERENCE ON ARTIFICAL INTELLIGENCE (IJCAI-17), 26, 2017, Melbourne. **Proceedings** [...]. [*S. l.: s. n*],2017. p. 2401–2407.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, Nature Publishing Group UK London, v. 323, n. 6088, p. 533–536, 1986.

SAKURAI, R. **Implementando a estrutura de uma rede neural convolucional utilizando o MapReduce do Spark**. 20 dez. 2017. Disponível em:https://www.sakurai.dev.br/cnn-mapreduce/. Acesso em: 28 set. 2026.

WHITE, F. M. **Mecânica dos fluidos**. [*S. l.*]: McGraw Hill Brasil, 1988.

WORLD HEALTH ORGANIZATION. **Cardiovascular diseases**. 2025. Disponível em: https://www.who.-int/health-topics/cardiovascular-diseases. Acesso em: 28 set. 2026.