UNIVERSIDADE DE SÃO PAULO
ESCOLA POLITÉCNICA

Detecção de Sopro Cardíaco com Técnicas de Aprendizado de Máquina

Bolsista: João Henrique Vilarim Mendes de Melo
Orientador: Renato Candido
Coorientador: Magno Teófilo Madeira da Silva

Resumo

Este trabalho apresenta uma metodologia baseada em Redes Neurais Convolucionais (CNN) aliada a técnicas de processamento de áudio, como o Modelo Oculto de Markov (HMM) e os Coeficientes Cepstrais nas Frequências Mel (MFCC), visando a classificação da ausculta cardíaca em cenários binários e multiclasse. Para o desenvolvimento e avaliação do sistema, utilizaram-se os conjuntos de dados PhysioNet/CinC 2016 e PASCAL. O conjunto PhysioNet/CinC 2016 foi empregado na classificação binária por representar um estudo consolidado na literatura, reunindo registros coletados em ambientes clínicos reais com variabilidade acústica. O dataset PASCAL foi selecionado tanto para a abordagem binária quanto para a multiclasse devido à presença de rotulagens específicas para diferentes condições fisiológicas (Normal, Sopro e Extrassístole). Na classificação binária, o sinal fonocardiográfico foi decomposto em quatro bandas de frequência e integrado ao mapeamento temporal de estados (S_1, sístole, S_2 e diástole) obtido via HMM. Para a tarefa multiclasse, empregou-se a extração via MFCC. A arquitetura convolucional desenvolvida incorpora camadas de convolução, normalização por lote, *dropout* e *Global Average Pooling* (GAP). Os resultados obtidos demonstraram que a adição dos atributos de segmentação pelo HMM resultou em um ganho expressivo de desempenho, alcançando acurácia de 94,35% no PhysioNet e 99,41% no PASCAL para a classificação binária. Na abordagem multiclasse no PASCAL, o modelo atingiu acurácia global de 88,09%, com precisões individuais de 96% para a classe Normal, 97% para Sopro e 82% para Extrassístole. Conclui-se que a integração entre processamento de áudio e redes convolucionais constitui uma estrutura eficaz para a caracterização e classificação automatizada de auscultas cardíacas.

Palavras-chave: Fonocardiograma (PCG); Ausculta Cardíaca; Modelo Oculto de Markov; Coeficientes Cepstrais nas Frequências Mel; Rede Neural Convolucional; PhysioNet; PASCAL.

Sumário

1 Introdução **4**

2 Objetivos **5**

3 Capacitação e Estudos Prévios **6**

3.1 Implementação Prática e Estudo do Framework PyTorch 6

3.2 Estudo de Processamento Digital de Sinais (PDS) 6

4 Metodologia aplicada **7**

4.1 Sinal de Fonocardiograma 7

4.2 Base de Dados 8

4.3 Preparação do sinal 8

4.4 Faixas de frequência 9

4.5 Coeficientes Cepstrais De Frequência Mel (MFCCs) 9

4.5.1 Pré-ênfase 10

4.5.2 Enquadramento e janelas do sinal 10

4.5.3 Banco de filtros Mel 11

4.5.4 Transformada discreta de cosseno 12

4.6 Janelamento dos áudios 13

4.6.1 Energia de Shannon 13

4.6.2 Sincronia 14

4.7 Modelo Oculto de Markov (HMM) 14

5 Classificação de ausculta cardiaca **16**

5.1 Classificação de sopros usando CNN 16

5.1.1 2016 PhysioNet/CinC Challenge 18

5.1.2 PASCAL Classifying Heart Sounds Challenge 18

5.2 Medidas de Desempenho 19

6 Resultados Desempenho da arquitetura CNN **20**

6.1 Classificação Binária 20

6.2 Classificação Multiclasse 21

7 Conclusões **23**

APÊNDICE **24**

Artigo SBrT 24

Artigo SIICUSP 28

1 Introdução

As doenças cardiovasculares continuam sendo a principal causa de morbidade e mortalidade em todo o mundo. Uma das etapas preliminares na avaliação do sistema cardiovascular é a análise dos sons cardíacos [1]. Dentre as doenças cardiovasculares, as arritmias cardíacas são as mais comuns e a sua classificação precisa é de grande interesse em estudos biomédicos [2]. O estetoscópio analógico, apesar de ter mais de 200 anos, depende da audição subjetiva e da experiência do médico [3].

Em meio a este problema, o fonocardiograma (PCG) se torna uma alternativa para a uma detecção precisa de problemas cardíacos. O fonocardiograma é a representação gráfica e o registro visual dos sons produzidos pelo coração e pelos grandes vasos sanguíneos. Diferente da ausculta tradicional (auditiva), o PCG transforma vibrações mecânicas em sinais elétricos e, posteriormente, em registros visuais que podem ser armazenados e analisados. Ele permite a correlação precisa entre os sons (S1, S2, sopros) e as fases da sístole e diástole, servindo como uma ferramenta de documentação clínica superior à memória auditiva [4].

Neste contexto, destaca-se o uso de métodos computadorizados para a identificação automática das arritmias e sopros cardíacos. A partir da década de 1990, surgiram vários trabalhos em que soluções baseadas em redes neurais perceptron multicamada (multilayer perceptron – MLP) [5] um exemplo de uma rede basica para a resolução do problema.

Neste trabalho de iniciação científica, abordam-se a detecção e classificação de sopro cardíaco com o uso de sinal de fonocardiograma (PCG). Considera-se inicialmente, a aplicação de redes neurais convolucionais (CNN) com a utilização dos Coeficientes Cepstrais de Frequência Mel (MFCCs). Foram considerados sinais do banco de dados The Pascal Classifying Heart Sound [36].

2 Objetivos

Este trabalho tem como objetivo geral o desenvolvimento de um modelo neural para a classificação de sinais fonocardiográficos (PCG), visando a detecção de sopros e anomalias cardíacas em cenários ruidosos. A abordagem evoluiu para uma arquitetura robusta baseada em Redes Neurais Convolucionais, integradas ao processamento digital de sinais via Coeficientes Cepstrais nas Frequências Mel (MFCC) e segmentação temporal por Modelos Ocultos de Markov (HMM).

Os objetivos específicos deste projeto compreendem:

1. Expandir o treinamento do modelo para cenários de classificação binária e multi-classe.

2. Utilizar um conjunto de dados com maior volume de amostras na classificação binária, visando enriquecer o treinamento e mitigar o desbalanceamento das classes.

3. Avaliar o desempenho do classificador por meio das métricas de Acurácia, Sensibilidade (Recall), Especificidade, Precisão e F1-Score.

4. Avaliar o ganho de desempenho no treinamento ao comparar o sinal sem tratamento prévio.

3 Capacitação e Estudos Prévios

Nesta seção, apresenta-se uma síntese das atividades desenvolvidas previamente ao desenvolvimento direto do projeto de iniciação científica. Inicialmente, descreve-se a familiarização com o *framework* PyTorch para a construção dos modelos de classificação. Em seguida, detalham-se os estudos teóricos em Processamento Digital de Sinais (PDS) para a segmentação temporal dos sinais fonocardiográficos.

3.1 Implementação Prática e Estudo do Framework PyTorch

Paralelamente ao curso, foram realizadas atividades práticas de implementação para domínio do framework PyTorch, ferramenta escolhida para o desenvolvimento do projeto. Para validar o aprendizado, foram desenvolvidos dois experimentos:

- Classificação não-linear com o problema das meias-luas: Implementou-se uma rede MLP, com e sem PyTorch, para resolver o problema de classificação binária. Este exercício permitiu visualizar a capacidade da rede em gerar fronteiras de decisão não-lineares, validando o entendimento sobre funções de ativação (como ReLU e Sigmoid).

- Reconhecimento de dígitos (MNIST): Desenvolveu-se um classificador para o banco de dados MNIST. Esta etapa foi crucial para lidar com dados de maior dimensionalidade e praticar o fluxo completo de Machine Learning: pré-processamento de dados, definição da arquitetura, treinamento e validação.

3.2 Estudo de Processamento Digital de Sinais (PDS)

Dada a natureza do projeto, foi crucial estabelecer uma compreensão sólida sobre o tratamento de sinais no domínio do tempo e da frequência. Para tanto, conceitos fundamentais foram estudados no livro Discrete-Time Signal Processing de Alan V. Oppenheim [18]. A leitura focou nos capítulos referentes à análise no domínio da frequência e transformadas, com o objetivo de compreender a base matemática por trás da extração de características espectrais. Esse estudo foi fundamental para o entendimento teórico dos coeficientes MFCCs, a técnica de extração de características escolhida para representar os sinais de áudio na entrada da rede neural.

4 Metodologia aplicada

Para o treinamento de um modelo com o objetivo de classificar sopros cardíacos, é necessário analisar o sinal utilizado e entender do que ele é constituído. A seguir, será descrito a representação do sinal de fonocardiograma e suas características.

4.1 Sinal de Fonocardiograma

O sinal de fonocardiograma é um método de gravação em áudio dos sons produzidos pela ação mecânica do coração, foi desenvolvido para analisar informações quantitativas e qualitativas sobre sons cardíacos. Neste trabalho, ele é utilizado com o objetivo de classificar sopros cardíacos [6].

Em condições normais, o coração produz sons nos ciclos S1, S2, S3 e S4, sendo que os mais importantes são o S1 e S2, mostrados na Figura 1. Os sons cardíacos S1 e S2 são, por definição, sinais não estacionários e estão localizados na faixa de baixa frequência, aproximadamente entre 10 Hz e 300 Hz. O primeiro som cardíaco (S1) corresponde ao início da sístole ventricular e ocorre devido ao fechamento das válvulas atrioventriculares (movimento de contração do músculo cardíaco) e é registrado em frequências que variam de 10 Hz a 200 Hz. O som S2 é marcado pelo fim da sístole ventricular, significando o início da diástole (relaxamento do músculo cardíaco), é registrado em frequências de 20 Hz a 250 Hz e dura menos de 100 milissegundos [7].

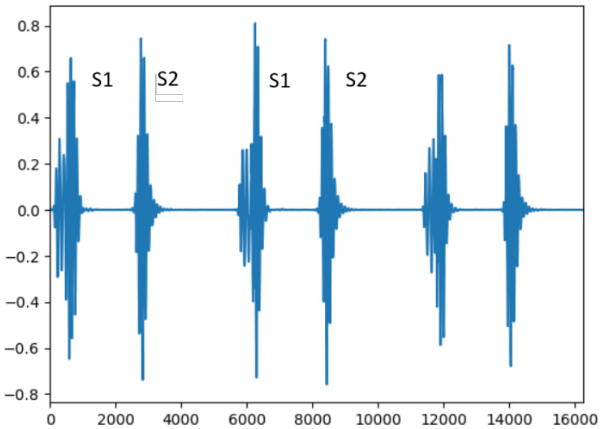


Figura 1: Representação do sinal de fonocardiograma, indicando sons S1 e S2.

A identificação precisa de S1 e S2 é crítica, pois esses eventos delimitam as fases de sístole e diástole, permitindo a categorização temporal dos sopros. Além disso, anomalias na estrutura de S2 (como desdobramentos fixos) são marcadores patológicos diretos para

defeitos como a Comunicação Interatrial (ASD) [8]. No contexto computacional, S1 e S2 servem como fronteiras para algoritmos de segmentação, facilitando a extração de características em janelas de tempo relevantes.

4.2 Base de Dados

A etapa da seleção do banco de dados para essa primeira abordagem é crítica no desenvolvimento do modelo, pois a qualidade dos dados impacta diretamente a capacidade de generalização do sistema. Ao longo do trabalho foram avaliados dois bancos de dados, sendo um deles o conjunto de dados descrito por *2016 PhysioNet/CinC Challenge*, esse conjunto de dados reúne sete bases de dados independentes, geradas com diferentes instrumentos para captação de áudio [2] composta por 2575 amostras normais e 665 amostras rotuladas como sopros cardíacos. Outro banco analisado foi o *The PASCAL Classifying Heart Sounds Challenge (2011)* [36], este repositório é amplamente reconhecido na literatura por fornecer uma referência padronizada para a classificação automática de sons cardíacos.

O banco de dados PASCAL é dividido em dois conjuntos principais (A e B). O conjunto A foi coletado via aplicativo móvel (iStethoscope) em ambiente não controlado e o conjunto B foi obtido em ambiente clínico utilizando estetoscópios digitais. Esses conjuntos categorizam os sons em normal, murmur (sopro), extrasystole e extra heart sound. Neste estudo, utilizaram-se apenas os sinais das categorias normal e murmur (sopro) do conjunto B. Essa escolha justifica-se pela maior presença de dados para treinamento e validação e tem qualidade de som maior comparada ao dataset A.

4.3 Preparação do sinal

Para garantir a reprodutibilidade e a precisão do diagnóstico clínico, este trabalho propõe fluxos de processamentos que extrai dos sinais brutos as características para o aprendizado de máquina, para dois tópicos, o treinamento binário e treinamento multiclasse. Para classificação binária, a etapa de pré-processamento do sinal é explicado pela Figura 2, o processo é composto por três passos, a separação do áudio em faixas de frequência, a segmentação em cada um e o treinamento em um rede CNN.

Já a metodologia adotada para o treinamento multiclasse tem três etapas principais, a preparação do sinal bruto (pré-processamento), como esquematizado na Figura 3, a

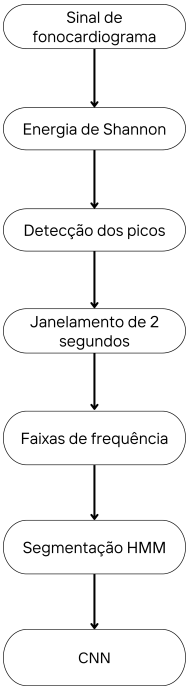


Figura 2: Diagrama de blocos do processamento dos sinais para o caso binário

extração de características acústicas fundamentais utilizando os MFCCs [10] e, por fim, a classificação por meio de uma rede CNN para o treinamento multiclasse. O aprofundamento dos métodos é descrito nas seções posteriores.

4.4 Faixas de frequência

No método de classificação binária, aplicou-se a decomposição do sinal de áudio em quatro bandas de frequência, visando otimizar a extração de características fonocardiográficas. Os sinais foram filtrados nas faixas de 25–45 Hz, 45–80 Hz, 80–200 Hz e 200–400 Hz, fornecendo atributos espectrais adicionais para a subsequente classificação pela rede neural, conforme ilustrado na Figura 4.

4.5 Coeficientes Cepstrais De Frequência Mel (MFCCs)

O sinal de fonocardiograma possui uma extensa quantidade de informações temporais e de frequências, porém, para a utilização em uma rede neural, é necessária a extração de características que capturem partes relevantes do sinal. A seguir são descritos alguns detalhes sobre os passos para o cálculo dos coeficientes MFCC

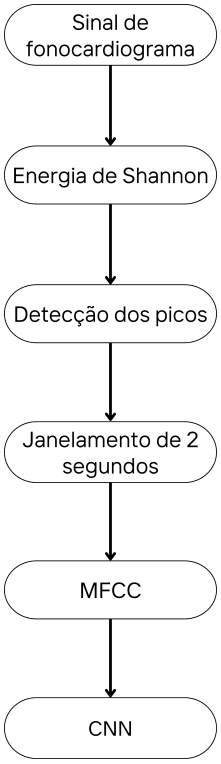


Figura 3: Diagrama de blocos do processamento dos sinais para o caso multiclasse.

4.5.1 Pré-ênfase

Inicialmente, o sinal de áudio x(n) é enviado para um filtro passa-alta, com equação de diferenças [43]

$$y(n) = x(n) - a \times x(n-1)$$

em que y(n) é a saída do sinal, x(n) é o sinal de entrada original e a é o coeficiente de pré-ênfase. Isso é importante porque em sons cardíacos a maior parte da energia está nas baixas frequências. A pré-ênfase ajuda a equalizar o espectro antes de calcular os MFCCs.

4.5.2 Enquadramento e janelas do sinal

O sinal de áudio é dividido em quadros de 20 a 30 milissegundos. Em cada quadro, é aplicado um janelamento para atenuar o sinal nas extremidades. Em geral, as janelas de Hanning e Hamming estão entre as opções mais conhecidas. Essas janelas podem realçar os harmônicos, suavizar as bordas e diminuir o efeito de borda ao calcular a DFT (Transformada Discreta de Fourier) do sinal [10].

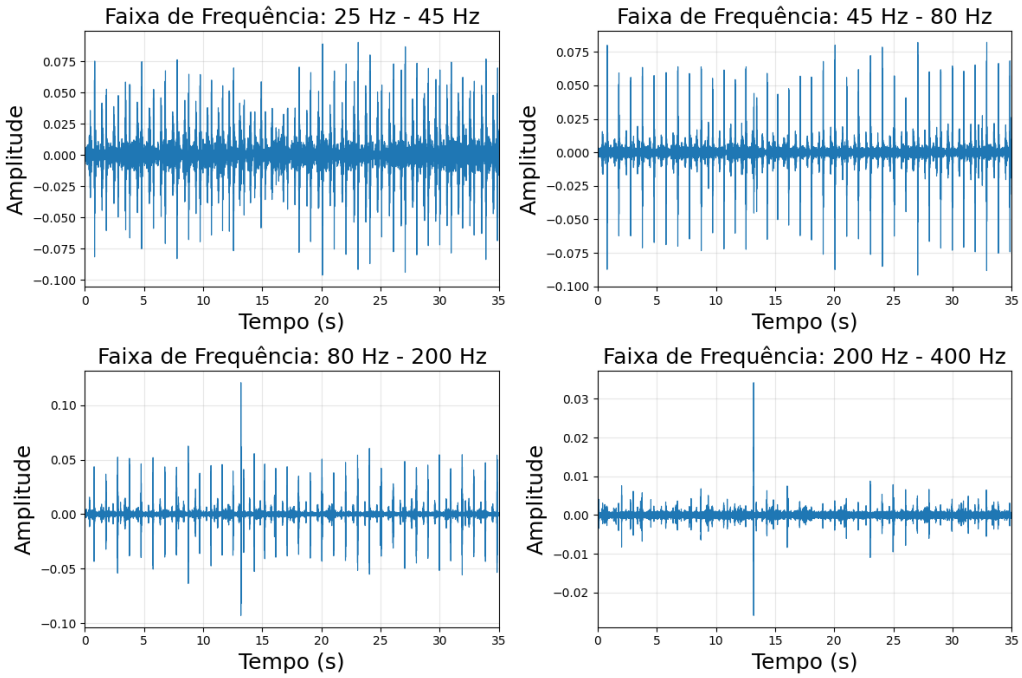


Figura 4: Exemplo de áudio separado em faixas de frequências.

4.5.3 Banco de filtros Mel

O banco de filtros passa-faixa Mel, mostrado na Figura 5, é construído com base na percepção de tom (pitch). O filtro Mel foi originalmente desenvolvido para análise de fala e, tal como a percepção auditiva humana, tem como objetivo extrair uma representação não linear do sinal de voz.

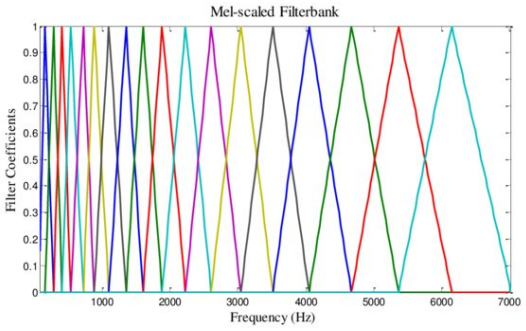


Figura 5: Banco de Filtros Mel.

A distribuição das frequências centrais desses filtros obedece à Escala Mel, uma escala logarítmica de frequências indicada na Figura 6, que é calculada como [11]

$$\text{Mel}(f) = 2595 \log\left(1 + \frac{f}{700}\right)$$

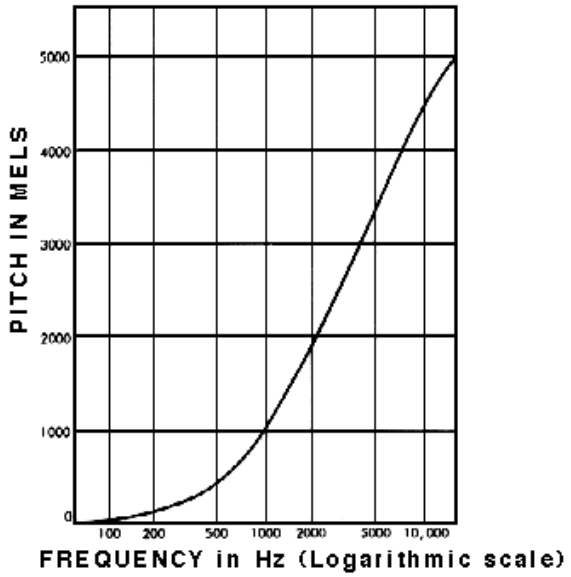


Figura 6: Representação da escala Mel.

No contexto da classificação de sons cardíacos, essa transformação é uma boa estratégia. Como a maior parte da energia dos sinais PCG (S1, S2 e sopros) se concentram em baixas frequências (entre 20 Hz e 500 Hz), a Escala Mel realça essa região, alocando mais filtros e extraindo características mais detalhadas onde a informação é crítica, ao passo que comprime as altas frequências, onde predominam ruídos e artefatos menos relevantes.

4.5.4 Transformada discreta de cosseno

Como visto em [18], a Transformada Discreta de Cosseno (DCT) expressa uma sequência finita de dados como uma somatória de funções cosseno, oscilando em diferentes frequências. No processamento dos MFCCs, a DCT é aplicada no banco de filtro Mel para selecionar a maioria dos coeficientes de aceleração ou separar a relação nas magnitudes espectrais logarítmicas do banco de filtros [12].

Neste trabalho, os MFCCs calculados para cada áudio deram origem a uma imagem 15 x 16 como mostrada na Figura 7. Essa imagem foi então vetorizada para utilizar como entrada na rede CNN.

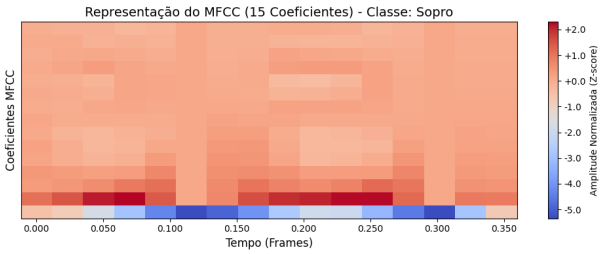


Figura 7: Imagem gerada após a extração dos MFCCs de um fonocardiograma (Com sopro).

4.6 Janelamento dos áudios

O comprimento de cada áudio apresenta uma elevada variação no banco de dados, podendo variar de 1 a 10 segundos. Portanto, para ser utilizada em uma rede CNN, é necessária que haja uma padronização desses comprimentos. Com base nisso, os áudios foram padronizados para 2 segundos, utilizando a detecção de picos e inserindo zeros para os áudios com comprimento menor.

4.6.1 Energia de Shannon

Para a sincronização dos dados que compõem sons dos batimentos cardíacos, adotou-se a Energia de Shannon como método principal para a extração da envoltória do sinal e localização de bulhas cardíacas (S1 e S2). A Energia de Shannon é calculada da forma

$$SE(n) = -x(n)^2 \cdot \log(x(n)^2)$$

em que onde x(n) é o sinal de PCG [15].

A vantagem de usar a energia de Shannon é a sua capacidade de enfatizar picos de média e baixa intensidade. Conforme mostrado em [16] na análise de sinais biomédicos, a Energia de Shannon consegue evidenciar os principais eventos cardíacos, atenuando ruídos irrelevantes de fundo, que são muito presentes no banco de dados escolhido para treinamento, sendo mais eficaz que a energia quadrática clássica [17]. Na Figura 8, é mostrado um sinal de PCG e a envoltória de Energia de Shannon correspondente.

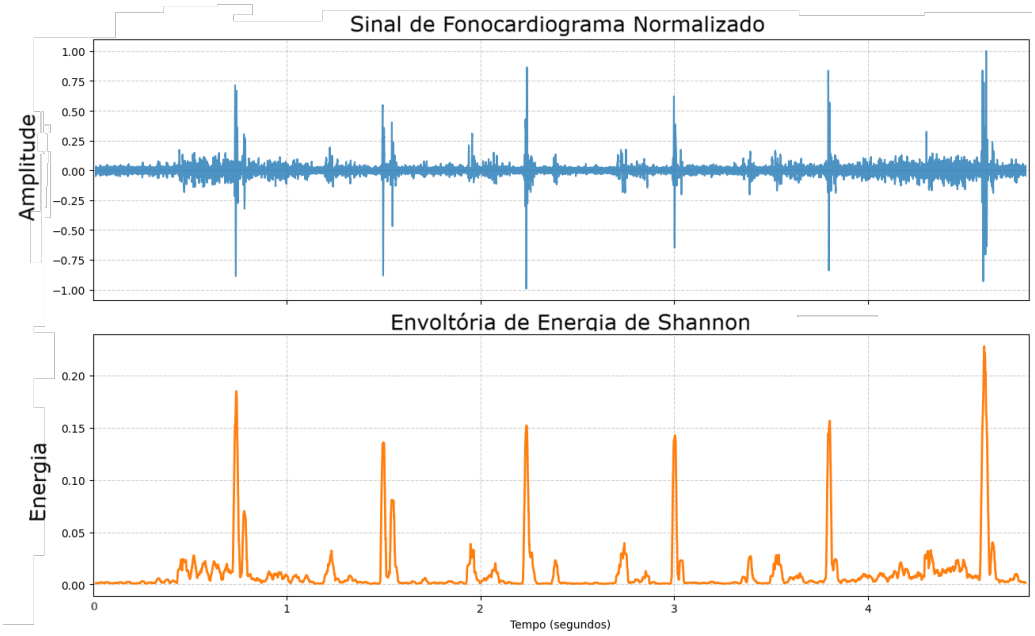


Figura 8: Exemplo de aplicação da Energia de Shannon.

A partir da aplicação do cálculo da Energia de Shannon, algoritmos de detecção de máximos locais (peak detection) foram empregados para identificar o instante do primeiro evento cardíaco de maior energia.

4.6.2 Sincronia

O primeiro evento cardíaco, pico detectado a partir da Energia de Shannon, serviu como ancoragem na sincronização dos áudios, a partir do qual se extraiu uma janela fixa de 2 segundos de duração, utilizando inserção de zeros se necessário, indicado pela Figura 9.

Posteriormente, os áudios janelados foram sumetidos à extração de características, gerando os vetores utilizados no treinamento da rede neural. Esse janelamento sincronizado assegurou que as imagens vetorizadas dos MFCCs, fornecidas como entrada para a arquitetura da CNN, possuíssem dimensões estritamente uniformes durante o treinamento do modelo.

4.7 Modelo Oculto de Markov (HMM)

Para a segmentação e caracterização temporal dos sinais fonocardiográficos (PCG), uma esteira de pré-processamento foi desenvolvida, inicialmente, o áudio bruto é submetido à extração do seu envelope homomórfico por meio de uma filtragem passa-baixas But-

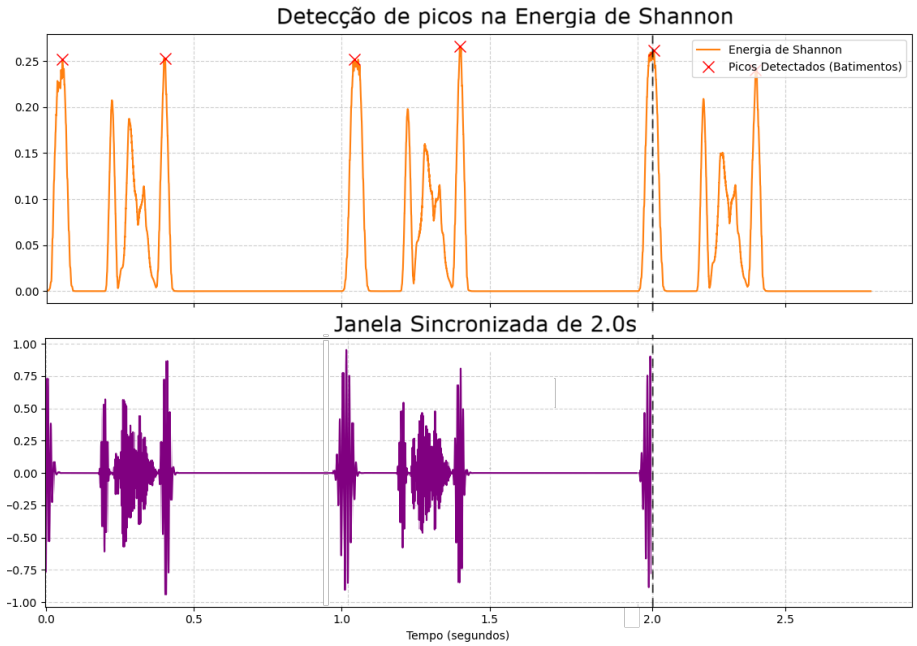


Figura 9: Detecção dos picos e extração dos dados de 2 segundos.

terworth de 2ª ordem com frequência de corte em 8 Hz aplicada sobre o logaritmo da amplitude. O envelope resultante é normalizado via *StandardScaler*.

Em seguida, emprega-se um Modelo Oculto de Markov (*HMM*) com 4 componentes, cujas matrizes de transição, médias e probabilidades iniciais foram customizadas para imitar a mudança das quatro fases do ciclo cardíaco: primeira bulha (S_1), sístole, segunda bulha (S_2) e diástole. A sequência de estados predita pelo HMM passa por um filtro mediano de suavização (com janela de 151 pontos) e por um algoritmo de correção lógica da sequência temporal, garantindo a coerência na duração relativa entre sístole e diástole.

Por fim, o sinal de saída do HMM é normalizado e concatenado aos sinais filtrados nas quatro sub-bandas de frequência de interesse (25–45 Hz, 45–80 Hz, 80–200 Hz e 200–400 Hz). Todos os canais são padronizados para uma dimensão máxima fixa (4.000 pontos), gerando um tensor de entrada final de dimensão 5×4000 (quatro sub-bandas de áudio mais o mapa de estados do HMM) fornecido diretamente à CNN 1D.

A eficácia da abordagem proposta na delimitação do sinal fonocardiográfico pode ser visualizada na Figura 10. Nela, é possível observar com clareza as mudanças de estado temporal do áudio ao longo do ciclo cardíaco, em que cada uma das quatro fases (S_1, sístole, S_2 e diástole) é identificada sobre o sinal de ausculta.

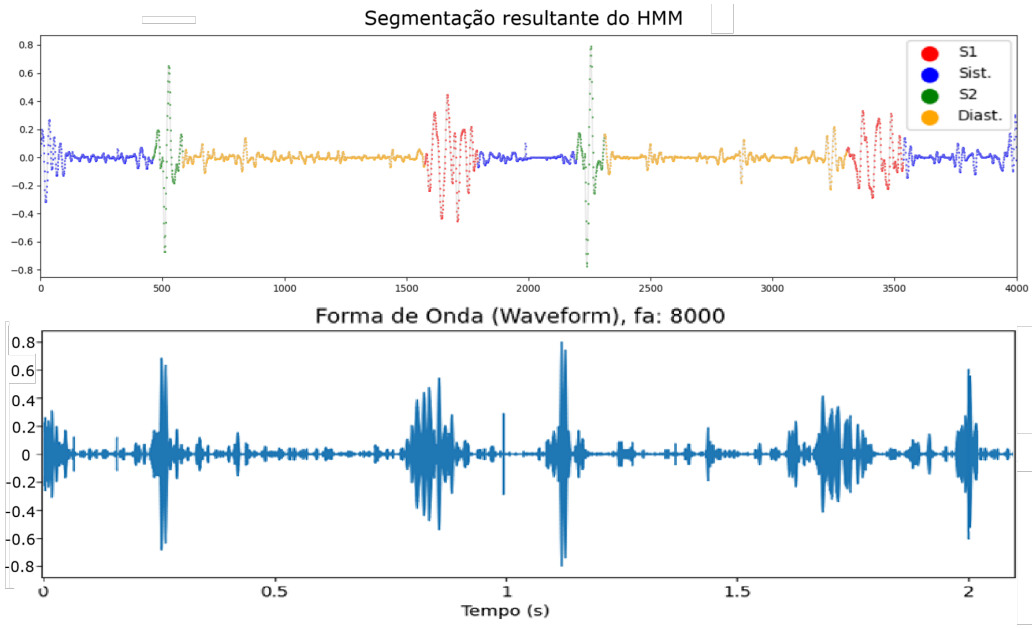


Figura 10: Resultado da segmentação por HMM

5 Classificação de ausculta cardiaca

Nesta seção, apresenta-se a avaliação do processo de treinamento e do desempenho dos modelos desenvolvidos para a classificação de sinais fonocardiográficos. As análises detalham o comportamento do sistema tanto na abordagem no conjunto de dados PhysioNet e PASCAL.

5.1 Classificação de sopros usando CNN

Propôs-se a implementação de uma rede neural convolucional (CNN - Convolutional Neural Network), já que essa rede apresenta vantagens na extração de características com dependência espacial e estrutural, permitindo melhor generalização ao treinamento de imagens.

A arquitetura da Rede Neural Convolucional foi projetada para extrair características dos sinais fonocardiográficos em três camadas convolucionais, seguidas por uma etapa de classificação linear. A camada de entrada é dimensionada de acordo com o modelo de classificação adotado: para a abordagem binária, recebe um tensor de 5 canais (composto pelo sinal fonocardiográfico completo segmentado via HMM e pelas quatro faixas de frequência filtradas e igualmente segmentadas); para a abordagem multiclasse, recebe

a matriz de coeficientes MFCC com dimensão 15×32.

A primeira camada convolucional (*Conv1D*) utiliza 16 filtros com um *kernel* largo de tamanho 25, *stride* 2 e *padding* 12, permitindo capturar dependências de mais longo alcance no sinal, seguida da função de ativação ReLU e de um agrupamento máximo (*MaxPool1D*) de tamanho 4. A segunda camada emprega 16 filtros de *kernel* 3 com normalização por lote (*BatchNorm1D*) e aplicação de uma taxa de *dropout* de 6% (*dropout1*) para regularização intermediária. A terceira camada convolucional expande a representação para 32 filtros de *kernel* 3, também acompanhada de *BatchNorm1D* e de uma etapa adicional de agrupamento (*MaxPool1D* com *kernel* 2 e *stride* 3).

Após a extração de características, aplica-se a subamostragem média adaptativa (*Global Average Pooling* - GAP), reduzindo a representação ao vetor de características global de dimensão 32. Antes da classificação final, aplica-se uma taxa de *dropout* de 30% (*dropout*) sobre o vetor resultante para evitar o *overfitting*. Por fim, a camada totalmente conectada (*Linear*) projeta os 32 atributos diretamente na saída da rede. O modelo foi treinado com a função de perda de entropia cruzada (*Cross-Entropy Loss*) sem a aplicação explícita de função de ativação *Sigmoid* ao final, fornecendo as predições diretamente na forma de *logits*. A arquitetura completa está esquematizada na Figura 11.

Para a classificação binária, a entrada $X \times Y$ da rede, mostrado pela Figura 11, é constituída por 5 canais (as quatro faixas de frequência de áudio somadas à máscara de segmentação do HMM) ao longo de 4.000 amostras, com a saída final processada por uma função de ativação sigmoidal. Para a classificação multiclasse, a entrada consiste em uma matriz de coeficientes MFCC (15×32), fornecendo 5 saídas na camada final.

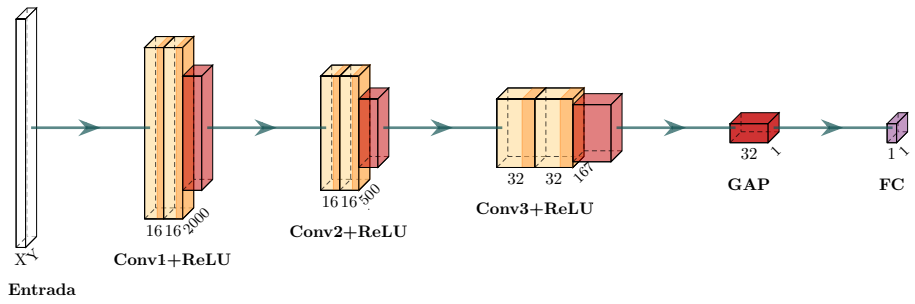


Figura 11: Arquitetura da CNN utilizada para classificação do fonocardiograma.

5.1.1 2016 PhysioNet/CinC Challenge

O conjunto de dados *PhysioNet/CinC Challenge 2016* foi empregado exclusivamente na etapa de classificação binária. O pré-processamento de cada sinal de áudio compreendeu a aplicação de um filtro passa-baixas e a decomposição em faixas de frequência. A base foi dividida na proporção de 80% para treinamento e 20% para teste. Na Fig. 12 é possível observar a evolução da função custo ao longo do treinamento.

O gráfico ilustra a convergência contínua do custo de treinamento ao longo das épocas. Em relação à curva de validação, embora se observe uma elevada variação nos valores, o que exige cuidado na análise, a verificação dos pontos de menor perda indica a manutenção de métricas de desempenho satisfatórias. Dessa maneira, avaliou-se que tais flutuações não comprometeram a qualidade geral da convergência do modelo.

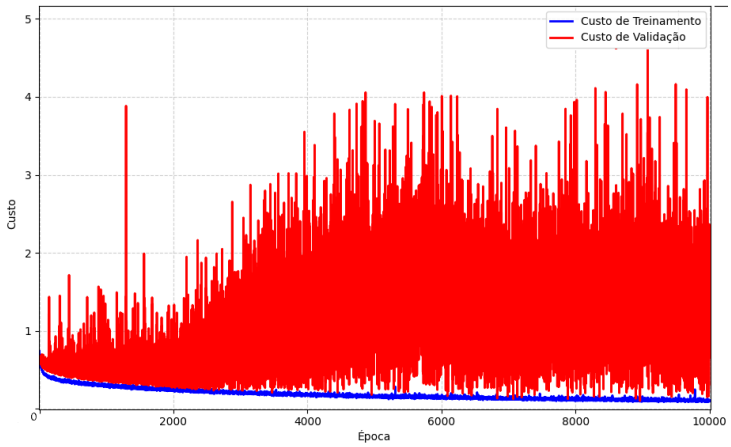


Figura 12: Evolução da função custo da rede - Binário (PhysioNet)

5.1.2 PASCAL Classifying Heart Sounds Challenge

A escolha do dataset PASCAL justifica-se por abranger mais de duas classes de patologias cardíacas. Por diretriz de projeto, essa base também foi utilizada como referência comparativa para a classificação binária, seguindo a mesma metodologia aplicada ao dataset PhysioNet. Para a abordagem multiclasse, aplicou-se o pré-processamento de áudio descrito anteriormente, seguido da extração de características via Coeficientes Cepstrais na Escala Mel (MFCC). O treinamento utilizou validação cruzada em 5 dobras (5-fold cross-validation), cujos resultados foram agregados ao final para permitir uma avaliação global do desempenho do modelo. Na classificação multiclasse, foram analisada as classes para batimento cardiaco normal, com sopro e extrassístole. A evolução do treinamento para o

caso binário é representada pela Fig. 13. Pode ser observada uma evolução de convergência equivalente à observada nos experimentos com a base PhysioNet. Consequentemente, aplicam-se a este cenário as mesmas considerações analíticas acerca da redução do custo de treinamento, da variância no conjunto de validação e da eficácia do *Early Stopping*.

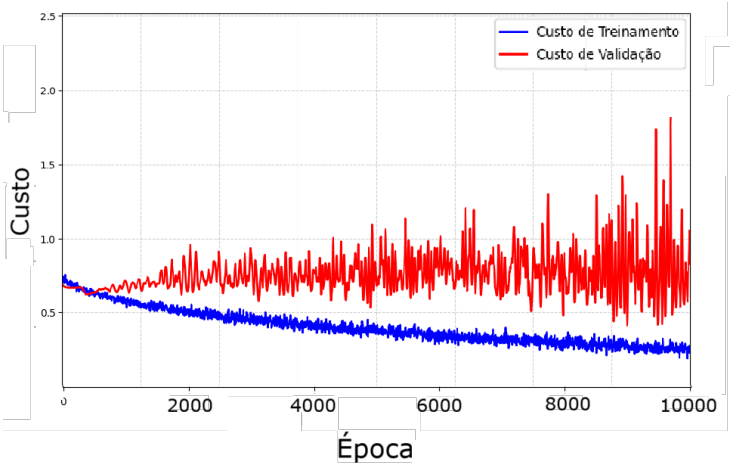


Figura 13: Evolução da função custo da rede - Binário (PASCAL)

5.2 Medidas de Desempenho

No estudo, métricas foram calculadas com base nas seguintes definições fundamentais:

- **Verdadeiro Positivo (VP):** Casos de sopro classificados corretamente como anormais;
- **Verdadeiro Negativo (VN):** Casos saudáveis classificados corretamente como normais;
- **Falso Positivo (FP):** Casos saudáveis classificados incorretamente como sopros (Alarme Falso);
- **Falso Negativo (FN):** Casos de sopro classificados incorretamente como normais (Erro Crítico).

Como métricas, foram consideradas a Acurácia, o Recall, a Especificidade, a Precisão e o F1-Score, calculados respectivamente como:

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}$$

$$\text{Recall} = \frac{VP}{VP + FN}$$

$$\text{Especificidade} = \frac{VN}{VN + FP}$$

$$\text{Precisão} = \frac{VP}{VP + FP}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}}$$

A Acurácia representa a taxa global de acertos do modelo. Como os dados de validação apresentam o mesmo tamanho, a acurácia representa uma boa medida de desempenho. O Recall mede a capacidade do modelo de detectar corretamente a presença de sopro. Um Recall alto indica que o sistema dificilmente deixa passar um caso negativo (baixo índice de Falsos Negativos). A Especificidade mede a capacidade do modelo de identificar corretamente os casos saudáveis. Uma baixa Especificidade implica uma alta taxa de alarmes falsos. Precisão a confiabilidade das predições positivas. Por fim, o F1-Score é a média harmônica entre a Precisão e o Recall. Esta métrica é fundamental para avaliar o equilíbrio do classificador, penalizando modelos que possuem uma grande disparidade entre Sensibilidade e Precisão, como modelos que classificam tudo como positivo para maximizar o Recall às custas da precisão.

6 Resultados Desempenho da arquitetura CNN

Nesta seção, avalia-se o processo de treinamento e o desempenho da arquitetura CNN proposta para cada uma das abordagens de classificação adotadas. A análise está estruturada em duas etapas: inicialmente, apresenta-se a avaliação da classificação binária e, em seguida, os resultados obtidos na classificação multiclasse.

6.1 Classificação Binária

Para a análise dos resultados da classificação binária, foi proposta uma abordagem comparativa a fim de avaliar o impacto da segmentação baseada em Modelos Ocultos de Markov

(HMM). Inicialmente, estabeleceu-se uma linha de base realizando o treinamento do modelo com a base PhysioNet sem a aplicação de pré-processamento, apenas áudio puro (.wav). Em seguida, o mesmo treinamento foi executado submetendo todos os áudios de treino e teste ao pipeline proposto, composto pela filtragem por faixas de frequência, segmentação via HMM e classificação por Rede Neural Convolucional (CNN). Ao comparar os resultados do treinamento realizado sem o tratamento prévio dos sinais de áudio com aqueles obtidos após a segmentação, verifica-se um ganho significativo de desempenho em ambas as bases de dados (PhysioNet e PASCAL). Analisando os dados quantitativamente pela Tab. 1, o modelo sem tratamento apresentou uma acurácia de 91,69%, métrica consistente haja vista o balanceamento prévio do conjunto de treino (composto por 150 amostras normais e 151 com presença de sopro). Em contrapartida, a aplicação do pré-processamento com segmentação HMM promoveu um aumento de 2,7% na acurácia do conjunto PhysioNet (atingindo 94,35%) e obteve uma acurácia para 99,41% no conjunto PASCAL. Outras duas métricas de relevância no contexto médico são o F1-score e a Sensibilidade: os elevados valores obtidos para o F1-score indicam uma confiabilidade no resultado da acurácia global, enquanto a sensibilidade unitária (1,0000) evidencia a capacidade do modelo em mitigar a ocorrência de falsos negativos, fundamental na triagem cardíaca.

Métrica	**PASCAL**	**PhysioNet**	**Sem HMM (PhysioNet)**
Acurácia	0,9941	0,9435	0,9169
F1-Score	0,9655	0,9467	0,9196
Sensibilidade	1,0000	1,0000	0,9470
Especificidade	0,9936	0,8867	0,8867

Tabela 1: Comparação de desempenho da classificação binária entre os datasets e abordagens.

6.2 Classificação Multiclasse

Para a avaliação da classificação multiclasse, as predições obtidas em todas as partições (folds) de validação cruzada foram agregadas em um único conjunto global. A partir deste conjunto, foram calculadas a matriz de confusão e as métricas de desempenho por classe (Precisão, Recall, F1-Score e Especificidade). A Tabela 2 mostra que o modelo alcançou uma Precisão de 96% para a classe Normal, 97% para a classe Sopro (*Murmur*) e 82% para

a classe Extrassístole. O elevado F1-*Score* observado nas classes Normal (0,96) e Sopro (0,93) evidencia a alta confiabilidade do modelo na identificação das patologias de maior prevalência. A menor precisão observada na classe Extrassístole (82%) é compensada por uma alta especificidade (0,96), indicando uma taxa reduzida de falsos positivos para essa categoria.

Classe	**Precisão**	**Recall**	**F1-Score**	**Especificidade**
Normal	0,96	0,97	0,96	0,89
Murmur	0,97	0,89	0,93	0,98
Extrassístole	0,82	0,87	0,84	0,96

Tabela 2: Resultados classificação multiclase - PASCAL

Adicionalmente, gerou-se a matriz de confusão mostrado na Figura 14 para analisar os padrões de erro e o comportamento de dispersão das predições interclasses. A menor precisão verificada na classe Extrassístole (82%) justifica-se primordialmente pelo severo desbalanceamento de dados presente na base PASCAL, constituída por 320 amostras da classe Normal, 95 da classe Sopro e apenas 46 da classe Extrassístole. A análise detalhada da matriz de confusão revelou que a principal confusão do modelo ocorreu no direcionamento de sinais de Extrassístole para a classe Normal. Esse comportamento é esperado do ponto de vista fisiológico e acústico, visto que a extrassístole preserva morfologias espectrais semelhantes às batidas normais na maior parte do ciclo cardíaco, diferenciando-se apenas pelo intervalo temporal atípico entre os batimentos.

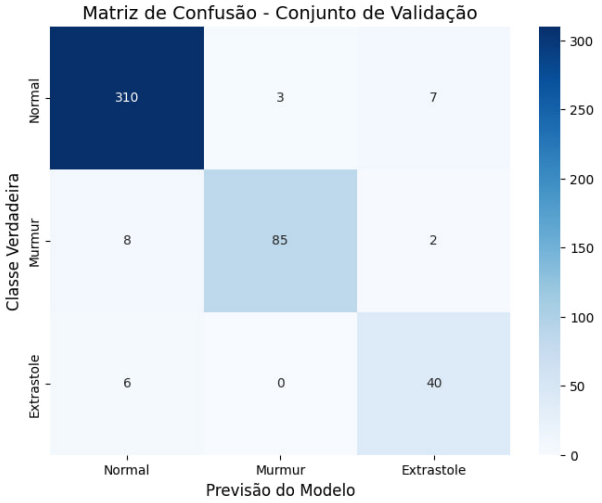


Figura 14: Matriz de confusão

7 Conclusões

Este trabalho apresentou um estudo sobre a aplicação de técnicas de Aprendizado de Máquina e Processamento Digital de Sinais na detecção automatizada de anomalias cardíacas a partir de sinais fonocardiográficos (PCG). A exploração de múltiplas abordagens de representação do sinal, desde o pré-processamento temporal e decomposição em faixas de frequência até a extração de coeficientes MFCC, demonstrou ser fundamental para a extração de atributos e construir modelos de diagnóstico.

Os resultados quantitativos comprovaram que a inclusão das características de segmentação geradas pelo Modelo Oculto de Markov (HMM) promoveu uma melhora expressiva na classificação binária. A incorporação do mapeamento do ciclo cardíaco como um canal de informação adicional permitiu que a CNN alcançasse uma acurácia de 94,35% na base PhysioNet e de 99,41% na base PASCAL, mantendo uma sensibilidade unitária (1,0000) e eliminando a ocorrência de falsos negativos na triagem das amostras.

Por outro lado, a análise de desempenho evidenciou que o desbalanceamento de dados no conjunto PASCAL (composto por 320 amostras da classe Normal frente a 95 de Sopro e 46 de Extrassístole) exerceu influência na precisão das classes minoritárias. Em contrapartida, o desempenho superior observado na base PhysioNet beneficiou-se do balanceamento prévio do conjunto de treinamento (150 amostras Normais para 151 de Sopro). Conclui-se que o uso de bases de dados mais extensas, como o conjunto PhysioNet, aliado à integração de novas estratégias de extração de características do áudio combinadas ao HMM ou MFCC, consolida-se como o caminho mais promissor para ampliar a capacidade de generalização e acurácia dos modelos de ausculta cardíaca.

APÊNDICE

Detecção de sopro cardíaco utilizando uma rede neural convolucional

João H. V. M. Melo, Magno T. M. Silva, Renato Candido

***Resumo*— Este trabalho apresenta um sistema para detecção automática de sopro cardíaco a partir de sinais de fonocardiograma utilizando técnicas de aprendizado de máquina. A abordagem proposta combina a extração de características acústicas com uma rede neural convolucional para realizar a classificação entre sinais normais e sinais com sopro cardíaco. Os resultados indicam que a metodologia proposta é promissora para aplicações de triagem cardíaca assistida por computador.**

***Palavras-Chave*— Aprendizado de máquina, classificação de sopros cardíacos, rede neural convolucional, fonocardiograma, coeficientes cepstrais de frequência mel.**

***Abstract*— This work presents an automatic heart murmur detection system based on phonocardiogram signals using machine learning techniques. The proposed approach combines acoustic feature extraction with a convolutional neural network to classify normal and murmur heart sounds. The results indicate that the proposed methodology is a promising approach for computer-aided cardiac screening applications.**

***Keywords*— Machine learning, heart murmur classification, convolutional neural network, phonocardiogram, mel-frequency cepstral coefficients.**

I. Introdução

Uma das etapas preliminares na avaliação do sistema cardiovascular é a análise dos sons cardíacos [1], [2]. Apesar de ter mais de 200 anos, o estetoscópio analógico depende da audição subjetiva e da experiência do médico [3]. Diante disso, o fonocardiograma (*phonocardiogram* – PCG) é uma alternativa para a detecção de problemas cardíacos. Diferente da ausculta tradicional (auditiva), o PCG transforma vibrações mecânicas em sinais elétricos e, posteriormente, em registros visuais que podem ser armazenados e analisados. Ele permite a correlação precisa entre os sons (S1, S2, sopros) e as fases da sístole e diástole, servindo como uma ferramenta de documentação clínica superior à memória auditiva [4].

Neste artigo, aborda-se a detecção de sopro cardíaco com o uso de sinais de PCG dos bancos de dados *The Pascal Classifying Heart Sound* [5] e de Yaseen e Kwon [10]. Como modelo, considerou-se uma rede neural convolucional (*convolutional neural network* – CNN) [6], alimentada com coeficientes cepstrais de frequência mel (*Mel-Frequency Cepstral Coefficients* – MFCCs) [7]–[9]. O artigo está organizado da seguinte forma. Na Seção II, descrevem-se os bancos de dados utilizados. Na Seção III, aborda-se a etapa de pré-processamento do sinal de fonocardiograma e na Seção IV, a arquitetura da CNN proposta. As Seções V e VI contêm respectivamente os resultados de simulação e as principais conclusões do trabalho.

João H. V. M. Melo, Magno T. M. Silva e Renato Candido, Escola Politécnica, Universidade de São Paulo, São Paulo, SP. Renato Candido também é afiliado ao SiDi Intelligence & Innovation Center. Emails: {joao.vilarim, magno.silva}@usp.br, renatocan@lps.usp.br. Este trabalho foi financiado pelo CNPq (PIBIC 2025-2321, 301743/2026-6 e 404081/2023-1), CAPES (código de financiamento 001) e pela FAPESP (2026/01471-7).

II. Bancos de dados

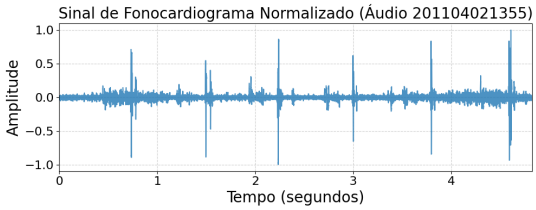
O banco de dados PASCAL [5] é dividido em dois conjuntos principais (A e B). O conjunto A foi coletado via aplicativo móvel (iStethoscope) em ambiente não controlado e o conjunto B foi obtido em ambiente clínico utilizando estetoscópios digitais. Esses conjuntos categorizam os sons em normal, *murmur* (sopro), *extrasystole* e *extra heart sound*. Neste estudo, utilizaram-se apenas os sinais das categorias normal e sopro do conjunto B. Essa escolha justifica-se pela maior quantidade de arquivos já que o conjunto B contém 320 arquivos de áudios no formato wav, categorizados como normal e 95 categorizados como sopro, amostrados com 4 kHz. Para a avaliação do modelo, adotou-se o divisão de treino e teste, utilizando 80% dos dados para treino e 20% para teste. Essa estratégia permite avaliar a capacidade de generalização e resiliência do classificador. Também foi utilizado o banco de dados Yaseen [10], contendo 200 arquivos classificados como normal e 800 como diferentes tipos de sopro (AS, MR, MS e MVP), amostrados com 8 kHz.

III. Etapa de pré-processamento do sinal

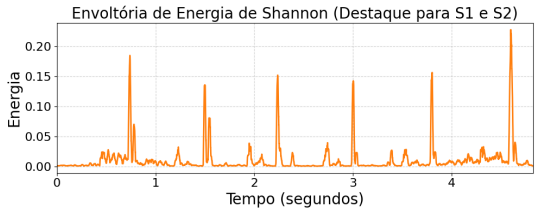
No conjunto de dados Yaseen, os sinais foram reamostrados para uma frequência de 4 KHz. Para sincronizar os registros, empregou-se a energia de Shannon na extração da envoltória do sinal e na localização das bulhas cardíacas (S1 e S2, principais eventos presentes em sinais de fonocardiograma) [11]. A energia de Shannon foi escolhida por sua capacidade de realçar picos de média e baixa intensidade, evidenciando os principais eventos cardíacos e atenuando ruídos de fundo presentes nos registros, característica particularmente importante para os sinais do banco de dados PASCAL [12], [13]. A Fig. 1 apresenta um exemplo de um sinal de PCG e sua respectiva envoltória obtida por meio da energia de Shannon.

O comprimento dos áudios apresenta variação entre os registros, podendo assumir valores de 1 a 10 segundos. Para padronizar o vetor de entrada da rede neural, os sinais de ambos os bancos de dados foram limitados a segmentos de 2 segundos, duração suficiente para abranger aproximadamente dois ciclos cardíacos e evitar que o zero *padding* influenciasse significativamente o processo de treinamento. Para a seleção dos segmentos, foram empregados algoritmos de detecção de máximos locais (*peak detection*) para identificar o instante correspondente ao primeiro evento cardíaco de maior energia. A partir desse ponto de referência, o áudio foi segmentado, utilizando-se apenas o trecho subsequente de 2 segundos como

entrada para o treinamento do modelo, conforme ilustrado na Fig. 2. Depois de sincronizar e limitar os sinais de áudio, os MFCCs foram calculados considerando 15 coeficientes, janela e passo de 2048 e 512 amostras respectivamente, dando origem a uma matriz de dimensões 15×16, como ilustrado na Fig. 3.

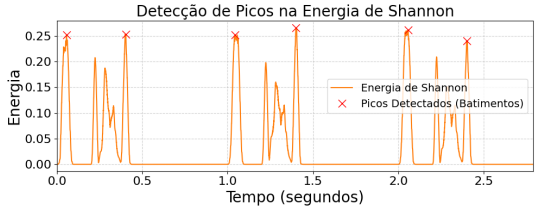


(a)

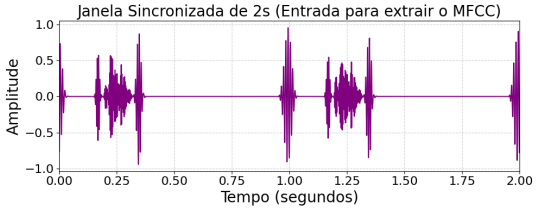


(b)

Fig. 1. Comparação temporal evidenciando picos sistólicos e diastólicos: (a) Sinal de fonocardiograma normalizado no domínio do tempo; (b) Envoltória da Energia de Shannon destacando as bulhas cardíacas principais (S1 e S2).



(a)



(b)

Fig. 2. Processo de sincronização e segmentação do sinal de fonocardiograma: (a) Detecção dos picos cardíacos utilizando a envoltória gerada pela energia de Shannon; (b) Janela sincronizada de 2,0 segundos, segmentada a partir do pico detectado e pronta para a extração dos MFCCs.

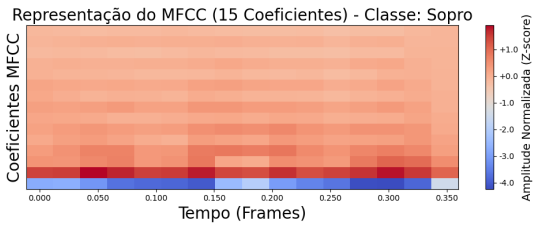


Fig. 3. Imagem gerada após a extração dos MFCCs de um fonocardiograma contendo sopro cardíaco.

IV. Rede Neural Convolucional

A arquitetura CNN proposta está mostrada na Fig. 4. Ela é composta por dois blocos de extração de características (camadas convolucionais) seguidos por uma etapa de classificação (um neurônio). A camada de entrada recebe a matriz de MFCCs com dimensões 15×16. Nela, é aplicada uma convolução com 8 filtros 3×3, utilizando *stride* igual a 1 e *padding* de 1. O resultado passa pela função de ativação ReLU, seguida de uma operação de subamostragem (*MaxPooling*) com janela 2×2 e passo 2, responsável por reduzir a dimensionalidade do mapa de características. A segunda camada convolucional recebe os 8 mapas de características gerados e aplica uma nova convolução com 16 filtros 3×3, passando por uma função de ativação ReLU e por uma segunda camada de *MaxPooling* 2×2. Após a extração, a estrutura tridimensional resultante é vetorizada. Esse vetor de dimensão 192×1 alimenta um único neurônio de saída com função de ativação sigmoidal.

O processo de aprendizado foi executado ao longo de 1000 épocas, dividindo o conjunto de dados em lotes (*batches*) de tamanho 16. A atualização dos pesos da rede foi realizada com o otimizador Adam (*Adaptive Moment Estimation*) com taxa de aprendizado $0,001$, $\beta_1 = 0,9$ e $\beta_2 = 0,999$. Além disso, como a finalidade do problema consiste em uma classificação binária, adotou-se a entropia cruzada binária (*binary cross-entropy* – BCE) como função custo [6].

V. Resultados

Para avaliar o processo de aprendizado e a capacidade de generalização da arquitetura proposta, monitorou-se a evolução da função custo ao longo do treinamento. As curvas da entropia cruzada para os conjuntos de treinamento e validação são apresentadas na Fig. 5. Observa-se que, após um determinado número de épocas, a função custo do conjunto de treinamento continua diminuindo, enquanto a do conjunto de validação deixa de apresentar melhorias, caracterizando o início de *overfitting*. Para mitigar esse efeito, adotou-se a estratégia de *early stopping*, selecionando o modelo correspondente à época que apresentou o menor erro de validação.

O desempenho do classificador foi avaliado por meio das métricas Acurácia, *Recall*, Especificidade, Precisão e F_1-Score, calculadas respectivamente como Acurácia $= (VP + VN)/(VP + VN + FP + FN)$, *Recall* $= VP/(VP + FN)$, Especificidade $= VN/(VN + FP)$, Precisão $= VP/(VP + FP)$ e F1-Score $= 2 \cdot (\text{Precisão} \cdot \textit{Recall})/(\text{Precisão} + \textit{Recall})$, em que VP, VN, FP e FN denotam os números de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, respectivamente.

A CNN foi treinada exclusivamente com os sinais das classes normal e sopro do conjunto B do banco de dados PASCAL [5] e posteriormente avaliada tanto em um conjunto de teste proveniente da própria base quanto no conjunto de dados Yaseen. Os resultados obtidos utilizando apenas dados do PASCAL são apresentados na Tabela I. Todas as métricas de desempenho alcançadas pela arquitetura proposta foram superiores a 90%, evidenciando elevado desempenho na distinção entre sinais normais e sinais com sopro.

Com o objetivo de avaliar a capacidade de generalização sob mudança de domínio, o modelo treinado no banco PASCAL foi aplicado diretamente ao banco de dados Yaseen, sem qualquer etapa de retreinamento ou adaptação. Os resultados obtidos são apresentados na Tabela II. Observa-se uma redução significativa do desempenho, com Acurácia de 54,10% e Especificidade de 43,50%, indicando que diferenças nas condições de aquisição, nas características acústicas e na distribuição dos

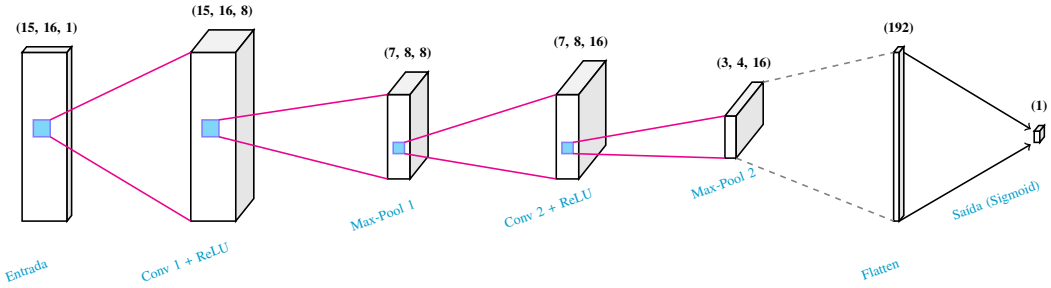


Fig. 4. Arquitetura detalhada da CNN proposta para classificação da matriz MFCC.

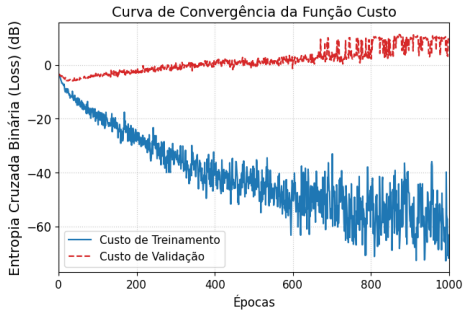


Fig. 5. Curva de convergência da função de custo (Entropia Cruzada Binária) para os conjuntos de treinamento e validação ao longo de 1000 épocas.

dados entre as duas bases afetam significativamente a capacidade de generalização da CNN. Em contrapartida, a Precisão permaneceu relativamente elevada (80,08%), sugerindo que, quando o modelo identifica a presença de sopro, essa decisão tende a ser confiável, embora a elevada incidência de falsos positivos e falsos negativos comprometa o desempenho global.

TABELA I

Desempenho da CNN - PASCAL	
Métrica	**Valor (%)**
Acurácia	96,08
Recall	90,01
Especificidade	97,50
Precisão	90,91
F_1-Score	90,91
Desempenho da MLP de [14] (4 K-Folds)	
Acurácia (%)	63,54

TABELA II

Desempenho da CNN - PASCAL vs Yaseen	
Métrica	**Valor (%)**
Acurácia	54,10
Recall	56,75
Especificidade	43,50
Precisão	80,08
F_1-Score	66,42

Para validar a arquitetura proposta, realizou-se ainda uma comparação com [14], no qual foi empregada uma rede *multilayer perceptron* (MLP) alimentada por MFCCs e avaliada por meio de validação cruzada com quatro partições [6]. Essa comparação é particularmente relevante por utilizar características acústicas semelhantes às adotadas neste trabalho, diferindo principalmente na arquitetura do classificador. Embora os autores tenham reportado acurácias de até 87,5% em partições específicas da validação cruzada, a acurácia média global obtida foi de 63,54%, conforme apresentado na Tabela I. Em contraste, a CNN proposta atingiu 96,08% de acurácia utilizando apenas a divisão convencional entre treinamento e teste no banco PASCAL, indicando que a arquitetura convolucional foi mais eficaz na extração de padrões discriminantes dos coeficientes MFCC.

VI. Conclusões

A arquitetura proposta demonstrou elevado desempenho na detecção de sopros cardíacos a partir de coeficientes MFCC, alcançando acurácia de 96,08% e especificidade de 97,50% no banco de dados PASCAL. Esses resultados evidenciam a eficácia da CNN na extração de características discriminantes para a classificação de fonocardiogramas, superando o desempenho médio reportado por trabalhos correlatos baseados em MLP. Entretanto, a avaliação cruzada entre os bancos PASCAL e Yaseen revelou uma redução significativa no desempenho, evidenciando que a robustez frente à transferência de domínio ainda representa um desafio. Como perspectiva futura, pretende-se investigar estratégias de treinamento com bases de dados mais diversificadas, bem como técnicas de adaptação de domínio, visando aumentar a capacidade de generalização do modelo para diferentes condições de aquisição e populações.

Referências

[1] T. Chakrabarti *et al.*, "Phonocardiogram signal analysis - practices, trends and challenges: A critical review," *Proc. Int. Conf. and Workshop Comput. Commun. (IEMCON)*, pp. 1-4, 2015.
[2] S. K. Berkaya *et al.*, "A survey on ECG analysis," *Biomed. Signal Process. Control*, vol. 43, pp. 216–235, 2018.
[3] A. D. Levin *et al.*, "A Machine Learning Approach to Heart Murmur Detection and Classification," *Proc. Int. Congr. Image Signal Process., Biomed. Eng. Informatics (CISP-BMEI)*, pp. 521-525, 2020.
[4] M. R. Montinari and S. Minelli, "The first 200 years of cardiac auscultation and future perspectives," *J. Multidiscip. Healthc.*, vol. 12, pp. 183-189, 2019.
[5] E. Gomes, P. Bentley, M. Coimbra, E. Pereira and Y. Deng, "Classifying heart sounds: Approaches to the PASCAL challenge," *Proc. Int. Conf. Health Informatics (HEALTHINF)*, pp. 337-340, 2013.
[6] C. M. Bishop, *Deep Learn.: Found. Concepts*. Springer, 2024.
[7] Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review," *IEEE Access*, vol. 10, pp. 122136-122158, 2022.
[8] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.
[9] Y. M. Ali *et al.* "Analysis of Accent-Sensitive Words in Multi-Resolution Mel-Frequency Cepstral Coefficients for Classification of Accents in Malaysian English," *Int. J. Aut. Mech. Eng.*, vol. 7, pp. 1053-1073, 2013.
[10] S. G.-Y. Yaseen and S. Kwon, "Classification of heart sound signal using multiple features," *Applied Sciences*, vol. 8, pp. 2444, 2018.
[11] H. Liang, S. Lukkarinen and I. Hartimo, "Heart sound segmentation algorithm based on heart sound envelogram," *Computers in Cardiology*, pp. 105-108, 1997.
[12] A. Moukadem *et al.*, "A robust heart sounds segmentation module based on S-transform," *Biomed. Signal Process. Control*, vol. 8, pp. 273–281, 2013.
[13] Z. Zidelmal *et al.*, "QRS detection using S-Transform and Shannon energy," *Comput. Methods Programs Biomed.*, vol. 116, pp. 1–9, 2014.
[14] C. Lubis and F. Gondawijaya, "Heart sound diagnose system with BFCC, MFCC, and backpropagation neural network," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 508, pp. 012119, 2019.

DETECÇÃO DE SOPRO CARDÍACO COM TÉCNICAS DE APRENDIZADO DE MÁQUINA

João Henrique V. M. Melo, Magno T.M. Silva e Renato Candido
Escola Politécnica, USP, SP
joao.vilarim@usp.br, magno.silva@usp.br, renatocan@lps.usp.br

Objetivos

Neste trabalho, investiga-se a aplicação de técnicas de aprendizado de máquina para a classificação de sons cardíacos da base de dados PhysioNet 2016. A metodologia proposta integra modelos ocultos de Markov (*hidden Markov model* – HMM) para a segmentação temporal do áudio. O mapeamento desses dados é realizado por uma rede neural convolucional (*convolutional neural network* – CNN). Adicionalmente, a robustez e a eficácia do modelo são validadas estatisticamente por meio do cálculo de intervalos de confiança via *bootstrapping*.

Métodos e Procedimentos

Para o treinamento do modelo foi usado o conjunto de dados descrito em *2016 PhysioNet/CinC Challenge*. Esse conjunto de dados reúne sete bases de dados independentes, geradas com diferentes instrumentos para captação de áudio [2]. Para o treinamento, utilizou-se 2575 amostras normais e 665 amostras rotuladas como sopros cardíacos. Por se tratar de um banco de dados desbalanceado, utilizaram-se funções de perdas ponderadas para penalizar mais erros da classe minoritária. O conjunto de validação totaliza 301 amostras, sendo 150 referentes à classe normal e 151 à classe de Sopro.

O HMM é um modelo probabilístico capaz de mapear a sequência do ciclo cardíaco [1], como a primeira bulha (S1), sístole, segunda bulha (S2) e diástole. Como as atividades do coração ocorrem em uma ordem temporal rigorosa, esse modelo interpreta essas quatro fases fisiológicas como estados sequenciais [1]. A partir dessas observações, o HMM calcula as probabilidades de transição entre os estados, estimando a sequência temporal mais provável do ciclo e permitindo delimitar, com alta precisão, o exato momento em que cada evento cardíaco começa e termina no áudio.

Cada áudio foi decomposto em quatro faixas de frequência utilizando filtros passa-faixa de Butterwoth [3]: 25 - 45, 45 - 80, 80 - 200, 200 - 400 (valores em Hz). Em cada faixa, foi aplicado o HMM, segmentando o áudio nos estados do ciclo cardíaco. A predição do HMM foi extraída a partir da envoltória homomórfica do áudio original, utilizando um filtro passa-baixa com frequência de corte 8 Hz.

Após a extração, os sinais de cada faixa de frequência e o vetor de estados do HMM foram padronizados para um comprimento fixo de 4000 amostras, empregando técnicas de truncamento ou preenchimento com zeros (*zero-padding*). Por fim, essas informações foram empilhadas para entrar no modelo. Para verificar a vantagem do pré-processamento com o HMM, a rede também foi treinada com os áudios originais.

O modelo é dado por uma CNN com 5 canais na entrada. A arquitetura foi construída com 3 camadas convolucionais unidimensionais, utilizando uma função de ativação *ReLU* em cada uma e uma etapa de regularização *dropout* de 6% na segunda convolução. Essa estrutura é seguida por uma camada de *average pooling*. Antes da etapa de decisão, é aplicado um *dropout* de 30%, para evitar o *overfitting*. Ao final, foi construída uma camada totalmente conectada com um neurônio, responsável pela classificação binária (saída com *logits*). A arquiterura da CNN utilizada está esquematizada na Figura 1.

O treinamento foi realizado com 10.000 épocas, processando os dados em lotes (*batches*) de 128 amostras. A atualização e otimização dos pesos da rede foram realizadas pelo algoritmo Adam (passo $0{,}0001$, $\beta_1 = 0{,}9$ e $\beta_2 = 0{,}999$). O cálculo do erro fundamentou-se na função de perda de entropia cruzada binária com logits (BCEWithLogitsLoss).

Resultados

Para validar os resultados, foi utilizado o método estatístico de reamostragem *bootstrap* [4], com o objetivo de comparar o desempenho da rede em dois estágios distintos do treinamento. Essa análise permitiu avaliar a generalização do modelo tanto durante o treinamento quanto no ponto ótimo de aprendizado, correspondente ao menor valor da função de custo.

As Tabelas 1 e 2 apresentam as métricas de desempenho dos modelos, acompanhadas de intervalos de confiança de 95%, permitindo avaliar não apenas os valores obtidos, mas também sua variabilidade estatística. Na Tabela 1, observa-se uma melhora consistente entre as épocas 1841 e 8963, com a acurácia aumentando de 0,8405 para 0,9435 e o F1-*score* de 0,8431 para 0,9467. A sensibilidade também apresentou aumento, atingindo 1 na época 8963, enquanto a especificidade passou de

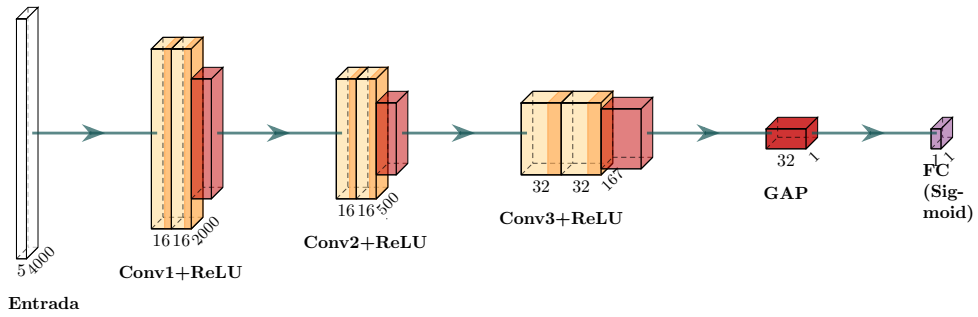


Figura 1: Arquitetura da CNN 1D utilizada para classificação do fonocardiograma.

0,8267 para 0,8867. A Tabela 2 apresenta os resultados com a CNN treinada com os dados entrada sem o pré-processamento com HMM. Observa-se que neste caso, o modelo atingiu uma acurácia de 0,9169, F1-Score de 0,9196, sensibilidade de 0,9470 e especificidade de 0,8867. Os intervalos de confiança de 95% reforçam a análise estatística dos resultados e fornecem uma indicação da incerteza associada às métricas obtidas. Comparando os resultados com a Tabela 1, verifica-se que o uso do HMM é benéfico para o treinamento do modelo.

Tabela 1: Métricas da CNN treinada com sinais pré-processados com HMM e intervalos de confiança (IC) de 95%.

Métrica	Época 1841	Época 8963
Acurácia (IC 95%)	0,8405 (0,7973 - 0,8804)	0,9435 (0,9169 - 0,9668)
F1-Score (IC 95%)	0,8431 (0,7958 - 0,8842)	0,9467 (0,9200 - 0,9702)
Sensibilidade (IC 95%)	0,8543 (0,7950 - 0,9079)	1,0000 (1,0000 - 1,0000)
Especificidade (IC 95%)	0,8267 (0,7639 - 0,8841)	0,8867 (0,8333 - 0,9338)

Tabela 2: Métricas da CNN treinada com sinais sem pré-processamento com HMM e IC de 95%.

Métrica	CNN sem HMM
Acurácia (IC 95%)	0,9169 (0,8837 - 0,9468)
F1-Score (IC 95%)	0,9196 (0,8860 - 0,9494)
Sensibilidade (IC 95%)	0,9470 (0,9085 - 0,9801)
Especificidade (IC 95%)	0,8867 (0,8344 - 0,9346)

Conclusões

Os resultados demonstram que a segmentação temporal dos sinais por HMM contribui para melhorar o desempenho da CNN na classificação automática de sopros cardíacos em fonocardiogramas, proporcionando um aumento de aproximadamente 2,7% na acurácia. Além disso, o modelo com HMM alcançou sensibilidade de 100% no conjunto de validação, característica relevante em cenários de triagem médica por reduzir a ocorrência de falsos negativos e auxiliar no encaminhamento de pacientes para exames confirmatórios. Por fim, a análise por meio do método de *bootstrap* indicou que o elevado desempenho observado nas etapas finais do treinamento não foi resultado de uma variação estatística isolada. Os resultados também evidenciaram uma melhora progressiva na generalização ao longo das épocas, reforçando a confiabilidade do modelo para a classificação de dados inéditos.

Os autores declaram não haver conflito de interesses. João concebeu o estudo, realizou as simulações e redigiu o resumo. Magno e Renato planejaram o estudo e revisaram o resumo. Todos os autores aprovaram a versão final do resumo. João recebeu bolsa PIBIC-CNPq.

Referências Bibliográficas

[1] S. E. Schmidt *et al.*, "Segmentation of heart sound recordings by a duration-dependent hidden Markov model," *Physiological Measurement*, vol. 31, no. 4, pp. 513–529, Apr. 2010.

[2] C. Liu *et al.*, "An open access database for the evaluation of heart sound algorithms," *Physiological Measurement*, vol. 37, no. 12, p. 2181, Nov. 2016.

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

[4] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 1995, pp. 1137–1145.

Referências bibliográficas

[1] T. Chakrabarti, S. Saha, S. Roy, and I. Chel, "Phonocardiogram signal analysis - practices, trends and challenges: A critical review,"in *2015 IEEE International Conference and Workshop on Computing and Communication (IEMCON)*, Vancouver, BC, Canada, 2015, pp. 1–4.

[2] S. K. Berkaya *et al.*, "A survey on ECG analysis,"*Biomedical Signal Processing and Control*, vol. 43, pp. 216–235, 2018.

[3] A. D. Levin, A. Ragazzi, S. L. Szot, and T. Ning, "A Machine Learning Approach to Heart Murmur Detection and Classification,"in *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Chengdu, China, 2020, pp. 521–525.

[4] M. R. Montinari and S. Minelli, "The first 200 years of cardiac auscultation and future perspectives,"*Journal of Multidisciplinary Healthcare*, vol. 12, pp. 183–189, 2019.

[5] L. Edenbrandt, B. Devine, and P. W. Macfarlane, "Neural networks for classification of ECG ST-T segments,"*Journal of Electrocardiology*, vol. 25, pp. 167–173, 1992.

[6] S. M. Debbal, "Analysis of the four heart sounds statistical study and spectro-temporal characteristics,"*Clinical Case Reports Journal*, vol. 1, no. 4, pp. 1–13, 2020.

[7] A. Noponen *et al.*, "Phono-spectrographic analysis of heart murmur in children,"in *Advances in Multimedia Information Processing – PCM 2004*, K. Aizawa, Y. Nakamura, and S. Satoh, Eds. Berlin: Springer, 2007, pp. 566–574.

[8] C. Lubis and F. Gondawijaya, "Heart sound diagnose system with BFCC, MFCC, and backpropagation neural network,"*IOP Conference Series: Materials Science and Engineering*, vol. 508, no. 1, p. 012119, 2019.

[9] Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review,"*IEEE Access*, vol. 10, pp. 122136–122158, 2022.

[10] G. Strang, "The discrete cosine transform,"*SIAM Review*, vol. 41, no. 1, pp. 135–147, 1999.

[11] Y. Mohd Ali, P. M. Paulraj, S. Yaacob, R. Yusuf, and S. Abu Bakar, "Analysis of Accent-Sensitive Words in Multi-Resolution Mel-Frequency Cepstral Coefficients for Classification of Accents in Malaysian English,"*International Journal of Automotive and Mechanical Engineering*, vol. 7, pp. 1053–1073, 2013.

[12] B. Truax, Ed., *Handbook for Acoustic Ecology*, 2nd ed. Burnaby: Simon Fraser University, 1999.

[13] M. Z. Suboh, R. Jaafar, N. A. Nayan, and N. H. Harun, "Shannon Energy Application for Detection of ECG R-peak using Bandpass Filter and Stockwell Transform Methods,"*Advances in Electrical and Computer Engineering*, vol. 20, no. 3, pp. 41–48, 2020.

[14] A. Moukadem, A. Dieterlen, N. Hueber, and C. Brandt, "A robust heart sounds segmentation module based on S-transform,"*Biomedical Signal Processing and Control*, vol. 8, no. 3, pp. 273–281, 2013.

[15] Z. Zidelmal, A. Amirou, D. Ould-Abdeslam, A. Moukadem, and A. Dieterlen, "QRS detection using S-Transform and Shannon energy,"*Computer Methods and Programs in Biomedicine*, vol. 116, no. 1, pp. 1–9, 2014.

[16] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Upper Saddle River, NJ: Pearson, 2010.

[17] R. Yu, W. Yu, and X. Wang, "Kan or mlp: A fairer comparison,"*arXiv preprint arXiv:2407.16674*, 2024.

[18] H. Taud and J.-F. Mas, "Multilayer perceptron (MLP),"in *Geomatic Approaches for Modeling Land Change Scenarios*, Cham: Springer, 2017, pp. 451–455.

[19] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. New York: Pearson Education, 2009.

[20] GeeksforGeeks, "What is Perceptron - The Simplest Artificial Neural Network,"2025. [Online]. Available: https://www.geeksforgeeks.org

[21] S. Srivastava, "Understanding the Difference Between ReLU and Sigmoid Activation Functions in Deep Learning,"*Medium*, 2024.

[22] Deep Learning Book Brasil, "Capítulo 23 – Como Funciona o Dropout?", 2025. [Online]. Available: https://www.deeplearningbook.com.br

[23] kinguistics, "Heartbeat Sounds,"*Kaggle*, 2016. [Online]. Available: https://www.kaggle.com

[24] Z. Li *et al.*, "A survey of convolutional neural networks: analysis, applications, and prospects,"*IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, 2021.

[25] R. Yamashita *et al.*, "Convolutional neural networks: an overview and application in radiology,"*Insights into Imaging*, vol. 9, no. 4, pp. 611–629, 2018.

[26] K. O'Shea and R. Nash, "An introduction to convolutional neural networks,"*arXiv preprint arXiv:1511.08458*, 2015.

[27] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge: MIT Press, 2016.

[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning,"*Nature*, vol. 521, pp. 436–444, 2015.

[29] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition,"in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, Haifa, Israel, 2010, pp. 111–118.

[30] H. Yingge, I. Ali, and K.-Y. Lee, "Deep Neural Networks on Chip - A Survey,"in *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2020, pp. 589–592.

[31] A. Hidaka and T. Kurita, "Consecutive Dimensionality Reduction by Canonical Correlation Analysis for Visualization of Convolutional Neural Networks,"in *Proceedings of the ISCIE International Symposium on Stochastic Systems Theory and its Applications*, 2017, pp. 160–167.

[32] E. Gomes, P. Bentley, M. Coimbra, E. Pereira, and Y. Deng, "Classifying heart sounds: Approaches to the PASCAL challenge,"in *HEALTHINF 2013 - Proceedings of the International Conference on Health Informatics*, 2013, pp. 337–340.

[33] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in Python*. Cham: Springer, 2023.

[34] S. Dixit and D. Nagaria, "LMS adaptive filters for noise cancellation: A review,"*International Journal of Electrical and Computer Engineering*, vol. 7, no. 5, pp. 2520–2529, 2017.

[35] L. Datta, "A survey on activation functions and their relation with xavier and he normal initialization,"*arXiv preprint arXiv:2004.06632*, 2020.

[36] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[37] C. Ricciardi *et al.*, "Linear discriminant analysis and principal component analysis to predict coronary artery disease,"*Health Informatics Journal*, vol. 26, no. 3, pp. 2181–2192, 2020.

[38] A. Ng, "Deep Learning Specialization,"Coursera / deeplearning.ai, 2026. [Online]. Available: https://www.coursera.org/specializations/deep-learning

[39] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs: Prentice Hall, 1993.

[40] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification,"*IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[41] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines,"in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, Haifa, Israel, 2010, pp. 807–814.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization,"*arXiv preprint arXiv:1412.6980*, 2014.

[43] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[44] N. Srivastava *et al.*, "Dropout: a simple way to prevent neural networks from overfitting,"*The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[45] C. Liu *et al.*, "An open access database for the evaluation of heart sound algorithms,"*Physiological Measurement*, vol. 37, no. 12, pp. 2181–2213, 2016.

[46] S. E. Schmidt *et al.*, "Segmentation of heart sound recordings by a duration-dependent hidden Markov model,"*Physiological Measurement*, vol. 31, no. 4, pp. 513–529, 2010.

[47] C. Potes, S. Parvaneh, A. Rahman, and B. Conroy, "Ensemble of Feature-based and Deep learning-based Classifiers for Detection of Abnormal Heart Sounds,"in *Computing in Cardiology (CinC)*, Vancouver, Canada, 2016, pp. 621–624.