Classificação Automática de Ritmos Cardíacos com *Vision Transformers* e Integração de Derivações

Douglas Blanc Pereira[1], **Taiane Coelho Ramos**[1]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)
Av. Gal. Milton Tavares de Souza, s/n - São Domingos, Niterói - RJ, 24210-310

{dpereira, taiane_ramos}@id.uff.br

Abstract. *Accurate interpretation of the electrocardiogram (ECG) is vital for the diagnosis of cardiovascular diseases, but manual analysis is complex and prone to errors. We propose an AI model based on Vision Transformers for the automatic analysis of cardiac signals. By integrating data from two leads, the system classifies five rhythms (one normal and four arrhythmias). The model achieved a 96.17% accuracy per patient, demonstrating its potential as a clinical decision support tool in the screening of cardiac anomalies.*

Resumo. *A interpretação precisa do eletrocardiograma (ECG) é vital para o diagnóstico de doenças cardiovasculares, mas sua análise manual é complexa e sujeita a erros. Propomos um modelo de IA baseado em Vision Transformers para análise automática do sinal cardíaco. Integrando dados de duas derivações, o sistema classifica cinco ritmos (um normal e quatro arritmias). O modelo atingiu 96,17% de acurácia por paciente, demonstrando seu potencial como suporte à decisão clínica na triagem de alterações cardíacas.*

1. Caracterização do Problema

As doenças cardiovasculares são a principal causa de morte no Brasil e uma das líderes em óbitos globais, tornando o diagnóstico precoce de anomalias como as arritmias cardíacas um fator crucial para a redução da mortalidade [Sociedade Brasileira de Cardiologia 2026, World Health Organization 2025]. O eletrocardiograma (ECG) é o exame primário para essa investigação, devido ao seu baixo custo e rapidez. Contudo, sua análise manual enfrenta desafios como a necessidade de interpretação por especialistas, a suscetibilidade a ruídos e a baixa sensibilidade para certas condições, o que pode levar a divergências diagnósticas [Breen et al. 2022].

Para mitigar essas limitações, sistemas automatizados baseados em aprendizado profundo (*Deep Learning*) têm demonstrado grande potencial [Ansari et al. 2023]. Arquiteturas como Redes Neurais Recorrentes (RNNs) e *Long Short-Term Memory* (LSTM) foram amplamente utilizadas para analisar dados sequenciais como o ECG. Modelos híbridos, que unem Redes Neurais Convolucionais (CNNs) com LSTMs ou Transformers, mostraram-se promissores por sua capacidade de capturar dependências de longo prazo e combinar a extração de características locais com a modelagem de relações temporais. Apesar do sucesso, modelos baseados em recorrência (como LSTMs) apresentam limitações relacionadas ao processamento sequencial, o que dificulta a paralelização durante o treinamento e a captura eficiente de dependências globais em sequências muito longas [Wen et al. 2022, Xiao et al. 2023, Alghieth 2025, Mahim et al. 2024].

Neste contexto, o uso de mecanismos de atenção, nativos da arquitetura *Transformer*, surge como uma alternativa eficaz. Diferentemente das CNNs, limitadas pelo campo receptivo local, e das LSTMs, restritas pela sequencialidade, os *Transformers* permitem a modelagem direta de dependências globais em todo o segmento do sinal cardíaco através do processamento paralelo. Esta característica não apenas acelera o treinamento em *hardware* moderno, mas também potencializa a identificação de padrões complexos em arritmias que se manifestam em diferentes escalas temporais [Vaswani et al. 2017, Xiao et al. 2023, Ji et al. 2024].

2. Motivação

Segundo Breen et al. (2022), apesar de seu uso disseminado, a interpretação manual do ECG é uma tarefa complexa, com estudos sugerindo que até 33% das análises realizadas por profissionais de saúde podem conter erros quando comparadas a especialistas. A dificuldade na aquisição dessa competência é incrementada pela variabilidade nos métodos de ensino e pela exigência de reconhecimento visual de padrões abstratos, o que demanda muita prática e retenção cognitiva. O estudo também alerta para os riscos associados à dependência de interpretações automatizadas convencionais que, devido a limitações de acurácia, podem comprometer a segurança do paciente, evidenciando a necessidade de sistemas de suporte à decisão mais confiáveis e precisos [Breen et al. 2022].

Embora algoritmos clássicos de processamento de sinais existam, eles frequentemente carecem de generalização para a variabilidade morfológica encontrada em grandes populações. Conforme a revisão sistemática de Prakash et al. (2025), os avanços em aprendizado profundo permitem superar as limitações das abordagens clássicas ao processar automaticamente a complexidade morfológica e a natureza não linear do sinal de ECG, garantindo maior precisão diagnóstica frente à variabilidade clínica entre pacientes.

De acordo com Zirpolo et al. (2025), embora sistemas baseados em aprendizado profundo como CNNs e LSTMs tenham avançado significativamente na extração automática de características, eles ainda enfrentam desafios de generalização devido à variabilidade morfológica e à necessidade de capturar dependências globais de forma mais assertiva. Os autores reforçam que, historicamente, o desenvolvimento de modelos confiáveis foi limitado pela escassez de dados rotulados, persistindo a necessidade de arquiteturas que superem as limitações de interpretação e sensibilidade na detecção de arritmias. Conforme discutido por Ji et al. (2024), persiste uma lacuna na investigação de arquiteturas que explorem plenamente a complementaridade de derivações específicas, uma vez que a maioria dos modelos ignora as relações espaciais e as dependências informativas cruciais entre os 12 eixos do ECG para a detecção precisa de arritmias.

3. Objetivo e Contribuições

O objetivo principal deste estudo é o desenvolvimento de uma arquitetura baseada em *Vision Transformers (ViT)* para a classificação automatizada de arritmias cardíacas com alta acurácia. Esta abordagem explora a capacidade intrínseca do modelo em processar dependências globais do sinal [Ji et al. 2024], com um mecanismo de integração tardia das derivações II e aVR.

As principais contribuições deste trabalho são:

- ***Pipeline Vision Transformer* para classificação de cinco ritmos cardíacos:** Desenvolvimento de *pipeline* que abrange a segmentação de batimentos baseada em picos R, a seleção de atributos relevantes e o treinamento de modelos com múltiplas derivações sobre a base de dados pública Chapman-Shaoxing com mais de 10.000 pacientes;
- **Fusão tardia das derivações selecionadas:** Implementação de uma estratégia de fusão de características que integra as representações das derivações II e aVR, selecionadas via *Random Forest*, permitindo a captura de dependências espaciais para a distinção de arritmias;
- **Avaliação de desempenho do modelo:** Validação experimental utilizando particionamento interpacientes, com análise de métricas de desempenho em cenários de classificação de cinco ritmos, demonstrando o desempenho da abordagem frente ao estado da arte.

4. Trabalhos Relacionados

Realizou-se uma pesquisa na literatura de acordo com as diretrizes de Kitchenham (2004), focando em artigos recentes sobre a classificação de arritmias com aprendizado profundo. A busca utilizou combinações de termos relacionados a arritmias, ECG e inteligência artificial para mapear o estado da arte.

A literatura de classificação de arritmias evoluiu de abordagens baseadas em extração manual de características para modelos de aprendizado profundo. Hannun et al. (2019) estabeleceram um marco ao demonstrar que uma rede neural profunda (DNN) de ponta a ponta pode processar dados brutos de ECG de derivação única para classificar doze ritmos cardíacos, indo muito além das detecções binárias tradicionais. O modelo atingiu uma pontuação F1 média de 0,837, superando a média de 0,780 de seis cardiologistas individuais quando avaliados contra um padrão-ouro estabelecido por um comitê de consenso de especialistas. Entretanto, o estudo aponta limitações importantes, como a seleção intencional de pacientes com ritmos raros para compor o conjunto de dados, o que compromete a generalização de métricas dependentes da prevalência, como a pontuação F1, para a população geral. Também aponta o tamanho reduzido do conjunto de teste (328 pacientes), que limitou a capacidade de realizar análises detalhadas de subgrupos. Yildirim et al. (2020) apresentaram um modelo de rede neural profunda (DNN) para a detecção automatizada de arritmias cardíacas, avaliado na base Chapman-Shaoxing. O modelo foi submetido a dois cenários de classificação: no primeiro, estruturado para identificar sete classes de ritmos cardíacos, obteve-se uma acurácia de 92,24%, todavia, fica claro que este número é elevado devido ao desbalanceamento de classes, já que o *F1-Score* cai para 80,04%; no segundo cenário, onde os sinais foram agrupados em quatro classes principais de ritmos, a acurácia subiu para 96,13%. As melhores performances gerais do modelo foram observadas ao utilizar os dados da derivação II do ECG.

Xiao et al. (2023) apresentam uma revisão sistemática abrangente sobre a classificação de arritmias em sinais de ECG utilizando técnicas de aprendizado profundo, com base na análise de 368 estudos. A revisão constata que as redes CNN são os modelos predominantes (usados em cerca de 58,7% dos estudos). Os autores destacam que, embora a precisão global reportada seja frequentemente elevada, a grande maioria dos trabalhos foca-se no paradigma de avaliação intrapaciente, ou seja, acabam trabalhando com dados contaminados (*data leakage*), o que prejudica a generalização dos modelos; quando

os modelos são submetidos ao cenário clínico mais realista e rigoroso do paradigma interpaciente, regista-se uma degradação significativa no desempenho, particularmente em métricas como a *F1-Score*, a sensibilidade e a precisão. O estudo conclui que a transição destas tecnologias para a prática clínica exigirá a utilização de bases de dados mais diversificadas, métodos de pré-processamento e aumento de dados mais avançados, a adoção de novas arquiteturas (como modelos híbridos e *Transformers*) e uma maior investigação sobre o paradigma interpaciente.

Na linha de arquiteturas baseadas em *Transformers* para análise de dados de séries temporais em processamento de sinais de ECG, Wen et al. (2022) fornecem uma revisão abrangente sobre a adaptação destes modelos para superar os desafios das séries temporais, categorizando as inovações ao nível da estrutura da rede (como novas codificações posicionais e módulos de atenção mais eficientes) e a sua aplicação em tarefas de previsão, classificação e detecção de anomalias. Em trabalhos que exploram esta vertente, Mahim et al. (2024) e Mohan et al. (2024) aplicam o uso de aprendizado profundo para a detecção de Fibrilação Atrial (AF) em ECG de derivação única, enfatizando não apenas a elevada precisão, mas também a interpretabilidade clínica dos modelos. Mahim et al. (2024) propõem o TransMixer-AF, uma arquitetura híbrida que combina *ConvMixer* e *Transformers* para extrair características locais e globais de sinais 1D, atingindo acurácia superior a 98% em bases de dados de referência (MIT-BIH e PhysioNet) e utilizando Grad-CAM++ para validar visualmente as decisões do modelo, sugerindo um forte potencial para monitorização em tempo real, porém utiliza dados intrapacientes. Complementarmente, Mohan et al. (2024) exploram a utilização de *Vision Transformers (ViT)* aplicados a representações 2D (espectrogramas) do sinal, alcançando uma exatidão de 95,25% e demonstrando, através de mapas de atenção, que o modelo foca corretamente características fisiológicas críticas como a onda P para justificar o seu diagnóstico.

Nossa proposta se insere nesta segunda onda de inovação *Transformers*, diferenciando-se por investigar a fusão tardia de derivações específicas como estratégia para potencializar a atenção global do modelo *Transformer*, sendo um aspecto não explorado nos trabalhos de Yildirim et al. (2020), Mahim et al. (2024) e Mohan et al. (2024), que focaram majoritariamente em uso de derivação única. O modelo proposto classifica múltiplas arritmias (campo não explorado pelos dois últimos trabalhos) com dados interpacientes, com objetivo de validar um modelo que possa ser generalizado, buscando solucionar um dos desafios apontados por Xiao et al. (2023). A Tabela 1 sumariza os resultados obtidos pelos trabalhos citados e que fizeram uso do paradigma interpaciente.

Tabela 1. Resultados dos principais trabalhos citados que adotam o paradigma interpaciente. O agrupamento de classes visa fundir ritmos clinicamente semelhantes para uma avaliação comparativa: AFIB (Fibrilação e Palpitação Atrial), SR (Ritmo Sinusal), SB (Bradicardia Sinusal) e GSVT (superclasse que engloba diversas Taquicardias Supraventriculares).

Trabalho	Método	Base	Derivações	Classes	F1 (%)	Acur. (%)
[Hannun et al. 2019]	DNN	MIT-BIH	II	12	83,70	-
[Yildirim et al. 2020]	DNN	Chapman-Shaoxing	II	7 4 (agrupadas)	80,04 95,57	92,24 96,13
[Mohan et al. 2024]	ResNet ViT	Chapman-Shaoxing	II	3	- -	86,13 92,46
Proposto	ViT	Chapman-Shaoxing	II e aVR	5 4 (agrupadas)	96,17 93,85	96,17 93,11

5. Metodologia

Na etapa experimental, implementou-se um modelo com arquitetura de aprendizado profundo fundamentada em *Vision Transformers* (ViT), adaptada para interpretar o sinal do ECG diretamente em sua forma sequencial. Ao tratar o sinal cardíaco como uma sequência de segmentos locais, o modelo utiliza mecanismos de atenção para identificar dependências temporais e morfológicas complexas entre diferentes fases do batimento, visando a classificação automática de arritmias cardíacas.

5.1. Base de Dados

Utilizamos a base de dados pública de 12 derivações da *Chapman University* e *Shaoxing People's Hospital* em sua versão pré-processada (*ECGDenoised*). Conforme descrito por Zheng et al. (2020), o tratamento original dos sinais para remoção de ruídos incluiu a aplicação sequencial de um filtro passa-baixa *Butterworth*, suavização por regressão polinomial local (LOESS) e filtragem por médias não-locais (NLM) para remoção de ruídos de alta frequência (acima de 50Hz) além de oscilações da linha de base.

Ela possui registros de 10 segundos de 10.646 pacientes, amostrados a 500Hz e classificados em 11 ritmos comuns rotulados por especialistas: Ritmo Sinusal (SR), Fibrilação Atrial (AFIB), Bradicardia Sinusal (SB), Taquicardia Sinusal (ST), Taquicardia Supraventricular (SVT), Palpitação Atrial (AF), Irregularidade Sinusal (SI), Taquicardia Atrial (AT), Taquicardia por Reentrada Nodal Atrioventricular (AVNRT), Taquicardia por Reentrada Atrioventricular (AVRT) e Ritmo Atrial Migratório do Nó Sinusal (SAAWR). Os rótulos foram atribuídos por um médico licenciado, revisados por um segundo médico e, para os casos de discordância entre eles, um terceiro médico sênior deu a decisão final. A versão pré-processada traz os arquivos de alguns pacientes com dados zerados, sendo 8 da classe ST, 3 da classe AFIB, 56 da classe SVT e 6 da classe AF, que foram descartados.

5.2. Segmentação

O baixo volume de dados somado ao desequilíbrio entre as classes é um dos desafios centrais no treinamento de modelos de aprendizado profundo para ECG. Zhao et al. (2023) observam que o desempenho de trabalhos anteriores cai bruscamente ao lidar com conjuntos de dados desequilibrados. Xiao et al. (2023) destacam que a aplicação prática de modelos de aprendizado profundo em procedimentos médicos ainda é limitada, mencionando que 28% dos estudos analisados realizaram aumentos de dados (*data augmentation*) para lidar com a escassez de amostras.

Para enfrentar este desafio, adotou-se uma estratégia de segmentação baseada em batimentos cardíacos utilizando a biblioteca *NeuroKit2*. A derivação II foi selecionada como referência para a detecção dos picos R, onde para cada pico detectado, definiu-se uma janela temporal fixa de 2,5 segundos (1250 amostras a 500 Hz), centralizada nele. A janela compreende 1,25 segundos (625 amostras) anteriores e 1,25 segundos posteriores conforme ilustrado nas Figuras 1 e 2, onde cada faixa de cor marca o início e o fim de um segmento. Esta largura de janela foi escolhida para capturar não apenas a morfologia completa do batimento central, mas também o contexto dos batimentos adjacentes. A segmentação foi aplicada simultaneamente a todas as 12 derivações para garantir o alinhamento temporal, descartando-se segmentos nas extremidades do sinal que não preenchessem a janela completa.

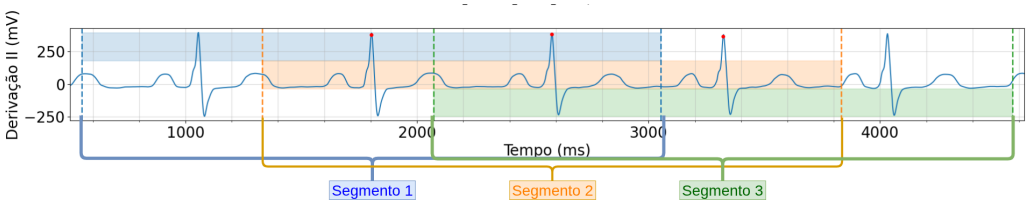

Figura 1. Segmentos de um paciente com ritmo SR.

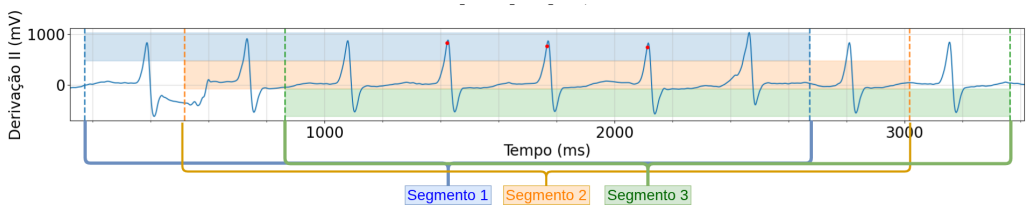

Figura 2. Segmentos de um paciente com ritmo SVT.

O subconjunto de dados selecionado para este estudo foi o das classes que possuíssem mais de 10.000 amostras de segmentos resultantes, sendo elas: SR, AFIB, SB, ST e SVT. A Tabela 2 apresenta a distribuição de pacientes e segmentos de toda a base de dados após a segmentação, ilustrando a representatividade de cada classe.

Tabela 2. Distribuição de pacientes e segmentos por classe na base de dados.

Classe	Pacientes	Batimentos
Bradicardia Sinusal (SB)	3889	**27065**
Taquicardia Sinusal (ST)	1560	**21768**
Fibrilação Atrial (AFIB)	1777	**20776**
Ritmo Sinusal (SR)	1826	**17356**
Taquicardia Supraventricular (SVT)	531	**10359**
Palpitação Atrial (AF)	439	5995
Irregularidade Sinusal (SI)	399	3645
Taquicardia Atrial (AT)	121	1718
Taquicardia por Reentrada Nodal AV (AVNRT)	16	328
Taquicardia por Reentrada AV (AVRT)	8	145
Ritmo Atrial Migratório do Nó Sinusal (SAAWR)	7	57

5.3. Modelo Proposto: Multi-Lead 1D-ViT

A arquitetura proposta (Figura 3), consiste em um modelo de aprendizado profundo que processa cada derivação de ECG independentemente através de ramos paralelos de blocos *Vision Transformer*. Para a definição das entradas do modelo, implementou-se uma seleção de derivações baseada na importância de atributos de um classificador *Random Forest*. Utilizando um subconjunto de 10.000 amostras do conjunto de dados, estratificados por paciente, os sinais brutos das 12 derivações foram concatenados em vetores de características. Após o treinamento do classificador com 100 árvores, calculou-se a importância acumulada de cada derivação somando-se os valores de importância de *Gini* de todos os seus pontos temporais, indicando as derivações II e aVR com as maiores capacidades discriminativas. Para garantir a integridade da avaliação e evitar o vazamento de

dados (*data leakage*), adotou-se um esquema de particionamento interpaciente. A divisão dos conjuntos de treino (70%), validação (20%) e teste (10%) foi realizada de forma determinística baseada no *hash* do identificador único do paciente, assegurando que todos os batimentos de um mesmo indivíduo pertençam exclusivamente a um único subconjunto.

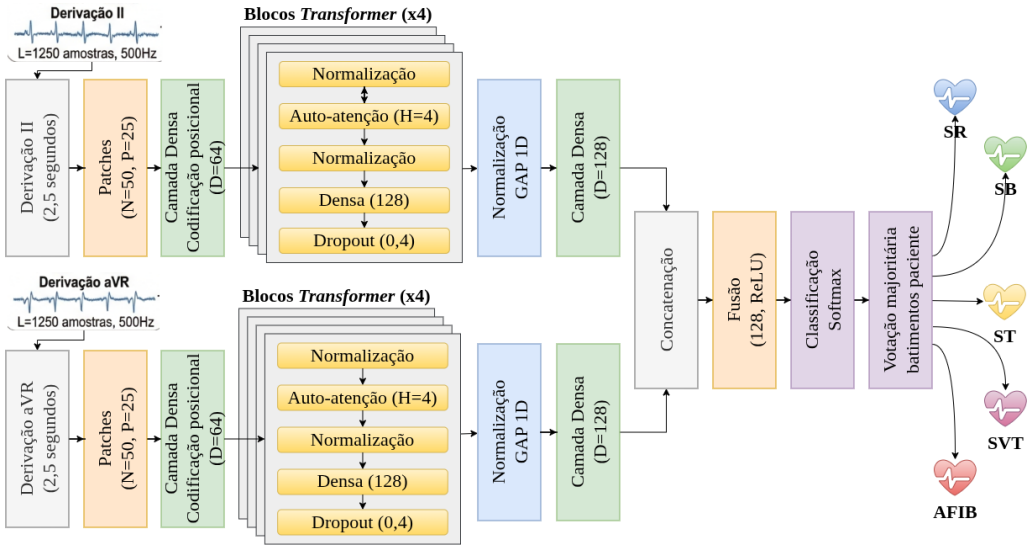

Figura 3. Arquitetura do modelo proposto. O sinal de cada derivação (II e aVR) é segmentado em *patches* e processado por um ramo *Transformer* independente. As características extraídas são então fundidas por concatenação para a classificação final do ritmo cardíaco.

Diferente da abordagem original do ViT para imagens 2D, o sinal de entrada é dividido em uma sequência de N *patches* não sobrepostos de tamanho P. Neste trabalho, definiu-se P = 25, resultando em N = 50 patches por segmento de L=1250 amostras. A escolha deste tamanho (equivalente a 50ms em 500Hz) permite capturar características, como a duração do complexo QRS e a presença de ondas P. Cada *patch* é mapeado para um vetor de dimensão latente D = 64 através de uma camada densa treinável, somado a uma codificação posicional aprendível. A escolha da dimensão foi feita de forma empírica, inspirada na arquitetura ResNet50 apresentada em Alghieth et. al. (2025). Antes do processamento pelos blocos *Transformer*, aplicou-se uma normalização por instância, garantindo que cada batimento tenha média zero e variância unitária independentemente. Para extrair as características, foram utilizadas 4 camadas de blocos *Transformer*, contendo Atenção Multi-Cabeça (H = 4) e MLP de 128 unidades. A saída de cada ramo é processada por *Layer Normalization*, *Global Average Pooling* 1D e uma camada densa de 128 unidades. A decisão final é obtida através de uma estratégia de fusão, onde os vetores resultantes de cada ramo (derivação II e aVR) são concatenados e processados por uma camada densa adicional (128 unidades, ReLU) e *Dropout* de 0,4, antes da camada de classificação final *Softmax* com 5 saídas.

Como o modelo realiza predições em nível de segmento (batimento), a classificação final do ritmo do paciente foi obtida por meio de uma votação majoritária das predições geradas para todos os seus batimentos no conjunto de teste, visando reduzir o impacto de falsos positivos isolados gerados por artefatos em segmentos específicos. O treinamento foi realizado utilizando a função de perda *Categorical Crossentropy* e o oti-

mizador *Adam* com taxa de aprendizado de 0,0001 e *clipnorm* de 1,0 para estabilidade dos gradientes. O tamanho do lote (*batch size*) foi configurado para 64. O modelo foi treinado por até 100 épocas, utilizando uma política de *Early Stopping* com paciência de 10 épocas, monitorando a perda no conjunto de validação para evitar o sobreajuste (*overfitting*) e restaurando os melhores pesos ao final.

6. Resultados

A avaliação do modelo proposto foi realizada no conjunto de teste, composto por dados de pacientes não vistos durante o treinamento, garantindo a validade das métricas apresentadas quanto à capacidade de generalização e evitando o vazamento de dados (*data leakage*). O desempenho global alcançou uma acurácia de 96,17% e um F1-Score macro de 94,87% para a classificação dos cinco ritmos cardíacos alvo. A Tabela 3 detalha o desempenho por classe. O modelo atingiu uma sensibilidade superior a 93% em quase todas as classes, destacando a SB, que apresentou 98,47% de sensibilidade. A classe de AFIB, crítica para o diagnóstico clínico devido ao seu alto risco associado a acidentes vasculares cerebrais (AVC) e insuficiência cardíaca, foi detectada com 96,41% de sensibilidade, que foi superior à precisão (91,48%). No contexto de uma ferramenta de triagem médica, esse comportamento é altamente desejável, indicando que o modelo prioriza a identificação correta da patologia (minimizando falsos negativos), mesmo que ao custo de um leve aumento de falsos positivos que seriam posteriormente revisados por um especialista.

Tabela 3. Métricas de desempenho do modelo por classe no conjunto de teste.

Classe	Precisão (%)	Sensibilidade (%)	F1-Score (%)	Suporte
SR	98,80	93,71	96,19	175
AFIB	91,48	96,41	93,88	167
SB	97,97	98,47	98,22	392
ST	95,33	95,33	95,33	150
SVT	92,45	89,09	90,74	55
Média Macro	95,21	94,60	94,87	939
Média Ponderada	96,22	96,17	96,17	939

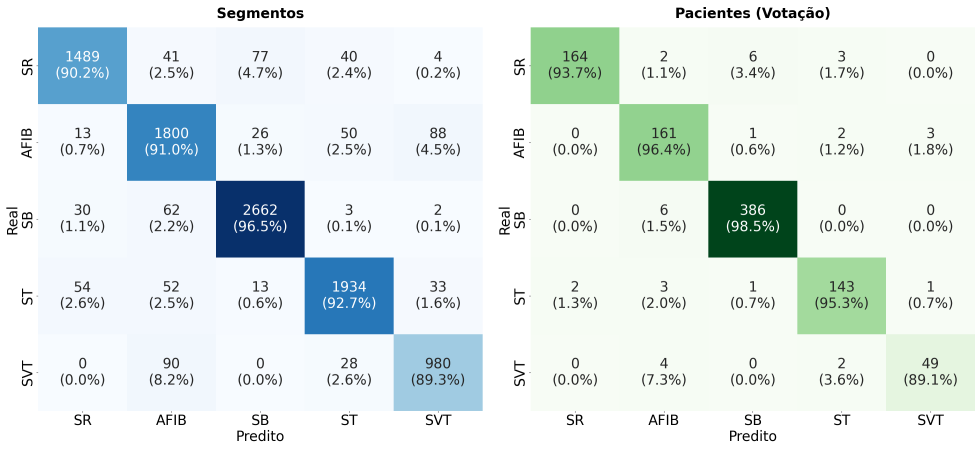

Figura 4. Matriz de confusão do modelo *Multi-Lead 1D-ViT* cinco classes. À esquerda, em nível de segmentos, à direita, em nível de paciente.

Para um melhor entendimento dos resultados, a matriz de confusão é apresentada na Figura 4, onde é possível compreender os padrões de erro do classificador proposto.

6.1. Cenário de Agrupamento de Classes

Para fins de comparação direta com o trabalho de Yildirim et al. (2020), o modelo foi reavaliado em um cenário de quatro classes. Este agrupamento segue a metodologia proposta por Zheng et al. (2020), que fundiram ritmos raros ou clinicamente relacionados. As superclasses são: AFIB, que une AFIB e AF; GSVT, que agrupa todas as taquicardias supraventriculares (SVT, AT, SAAWR, ST, AVNRT, AVRT); SB, contendo apenas ela própria; SR, que combina SR e SI. Conforme a Tabela 4, o modelo manteve uma performance competitiva neste cenário, com acurácia de 93,11%.

Tabela 4. Métricas de desempenho do modelo por classe no conjunto de teste.

Classe	Precisão (%)	Sensibilidade (%)	F1-Score (%)	Suporte
AFIB	91,30	87,50	89,36	216
GSVT	89,91	89,50	89,70	219
SB	96,74	98,47	97,60	392
SR	91,30	92,65	91,97	204
Média Macro	92,31	92,03	92,16	1031
Média Ponderada	93,08	93,11	93,08	1031

A Figura 5 traz a matriz de confusão do modelo para o cenário de quatro classes agrupadas.

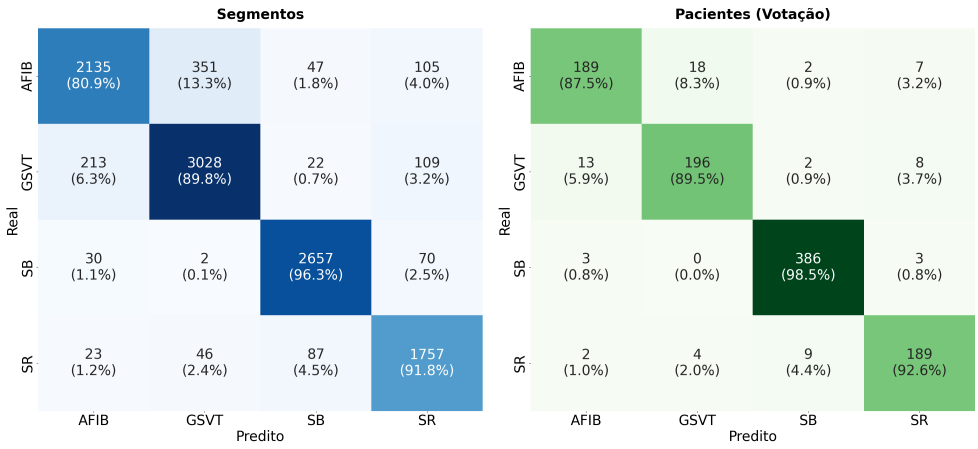

Figura 5. Matriz de confusão do modelo *Multi-Lead 1D-ViT* quatro classes agrupadas. À esquerda, em nível de segmentos, à direita, em nível de paciente.

7. Discussão

Os resultados obtidos corroboram a hipótese de que arquiteturas baseadas em *Vision Transformers*, quando alimentadas com subconjunto de múltiplas derivações, superam as limitações de modelos que dependem de uma única visão elétrica do coração ou que utilizam processamento estritamente sequencial. A acurácia de 96,17% no cenário de cinco classes representa um avanço significativo frente a abordagens tradicionais, especialmente considerando o protocolo de particionamento interpaciente adotado. Conforme

alertado pela revisão sistemática de Xiao et al. (2023), grande parte da literatura reporta acurácias inflacionadas devido à avaliação intrapaciente. Ao isolar os pacientes no nosso conjunto de teste, garantimos que o modelo aprendeu características generalizáveis das arritmias, e não peculiaridades específicas do ruído ou do sinal de um único indivíduo.

Nosso modelo para predição de cinco classes demonstrou vantagens aos trabalhos relacionados. Enquanto Mohan et al. (2024) alcançaram 92,46% utilizando ViT em três classes a partir de espectrogramas (2D), nossa abordagem de processamento sequencial 1D (em N *patches*) mostrou que a inclusão da derivação aVR e a segmentação temporal direta enriquecem a representação latente do sinal de forma mais eficiente. Em relação a Yildirim et al. (2020), que alcançaram 96,13% em quatro classes com uma DNN profunda, nosso modelo obteve desempenho competitivo (93,11%) com uma arquitetura que favorece a paralelização e a interpretabilidade através dos mecanismos de atenção.

A alta sensibilidade na detecção de AFIB (96,41%) é particularmente relevante. Como destacam Breen et al. (2022) [Breen et al. 2022], a interpretação manual do ECG é complexa e suscetível a erros de análise que podem ter consequências perigosas, ou até fatais, para os doentes. Sendo a fibrilação atrial uma condição com elevado risco clínico associado, a capacidade do modelo proposto de discernir estas anomalias com alta confiabilidade sugere o seu potencial como ferramenta de triagem em ambientes de urgência, onde a rapidez e a precisão são críticas.

A escolha arquitetural de fundir as derivações II e aVR demonstrou ser um diferencial. Enquanto modelos recentes de fronteira, como o MSGformer proposto por Ji et al. (2024), focam no processamento simultâneo das 12 derivações utilizando grades multiescala complexas, nossa abordagem baseada em *Feature Importance* demonstra que é possível obter alta precisão com custo computacional menor. O alto desempenho de modelos baseados em atenção na presença de ruído também vai ao encontro dos achados de Alghieth et al. (2025), que destacam a superioridade dos *Transformers* na captura de dependências globais em tempo real, superando as tradicionais LSTM.

Apesar dos resultados promissores, este estudo possui limitações que devem ser consideradas. O desempenho do modelo foi validado em uma única base de dados, sendo necessária a validação cruzada em conjuntos de dados externos para assegurar a generalização do modelo em diferentes populações e equipamentos. Embora a seleção de classes tenha mitigado o desbalanceamento extremo, ainda persiste uma disparidade no número de amostras entre as classes, e a interpretabilidade das decisões do ViT permanece como uma importante direção para investigações futuras.

Para mitigar essas limitações e aproximar a ferramenta da prática clínica, trabalhos futuros deverão focar em avanços arquiteturais e de explicabilidade. No âmbito da arquitetura, planeja-se explorar a integração de mecanismos de atenção cruzada (*cross-attention*) diretamente entre as derivações na camada de *embedding*, em vez de uma fusão tardia. Isso permitirá que o modelo pondere dinamicamente a importância de cada canal elétrico a cada instante do batimento. Em paralelo, a aplicação de técnicas de Inteligência Artificial Explicável (XAI) será crucial. Pretende-se extrair e projetar os mapas de atenção do modelo sobre o sinal original do ECG, demonstrando visualmente aos cardiologistas quais elementos fisiológicos motivaram a classificação da rede. Outro passo essencial será a validação cruzada do modelo em bases de dados externas e heterogê-

neas, confirmando sua capacidade de generalização em diferentes populações, protocolos clínicos e *hardwares* de aquisição de dados.

8. Conclusão

Apresentamos uma arquitetura baseada em *Vision Transformers* (ViT) para a classificação automática de arritmias, abordando desafios como a variabilidade morfológica interpaciente e a necessidade de generalização. Ao integrar a segmentação de batimentos com a fusão tardia das derivações II e aVR, o modelo alcançou uma acurácia de 96,17% e um F1-Score macro de 94,87% na distinção de cinco ritmos cardíacos na ampla base de dados *Chapman-Shaoxing*, demonstrando a eficácia dos mecanismos de atenção na captura de dependências globais do sinal, mesmo com um conjunto reduzido de entradas.

A excelente sensibilidade alcançada para condições clínicas de alto risco, notadamente 96,41% para a Fibrilação Atrial (AFIB), destaca o potencial desta ferramenta como sistema de suporte à decisão clínica. Ao garantir o isolamento estrito de pacientes entre as fases de treino e teste, o modelo provou não sofrer do fenômeno de vazamento de dados que compromete a aplicabilidade clínica de sistemas de IA na cardiologia.

Conclui-se que a combinação de arquiteturas modernas de aprendizado profundo com estratégias de seleção de atributos baseadas em dados oferece um caminho promissor para mitigar a subjetividade e os erros na interpretação manual do ECG. Como trabalhos futuros, pretende-se investigar a interpretabilidade das decisões do modelo através da visualização dos mapas de atenção e realizar a validação cruzada com bases de dados externas, visando consolidar a confiabilidade necessária para o suporte à decisão clínica.

Agradecimentos

Este estudo foi financiado pela FAPERJ - Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, Processo SEI E-26-210.369/2024.

Referências

Alghieth, M. (2025). Deepecg-net: A hybrid transformer-based deep learning model for real-time ecg anomaly detection. *Scientific Reports*, 15(1):20714.

Ansari, Y., Mourad, O., Qaraqe, K., and Serpedin, E. (2023). Deep learning for ecg arrhythmia detection and classification: an overview of progress for period 2017–2023. *Frontiers in Physiology*, 14:1246746.

Breen, C., Kelly, G., and Kernohan, W. (2022). Ecg interpretation skill acquisition: A review of learning, teaching and assessment. *Journal of electrocardiology*, 73:125–128.

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., and Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65–69.

Jaya Prakash, A., Nasreddine Belkacem, A., Elfadel, I. M., Jelinek, H. F., and Atef, M. (2025). Advances in machine and deep learning for ecg beat classification: a systematic review. *Frontiers in Digital Health*, 7:1649923.

Ji, C., Wang, L., Qin, J., Liu, L., Han, Y., and Wang, Z. (2024). Msgformer: A multi-scale grid transformer network for 12-lead ecg arrhythmia detection. *Biomedical Signal Processing and Control*, 87:105499.

Kitchenham, B. et al. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.

Mahim, S., Hossen, M. E., Al Hasan, S., Islam, M. K., Iqbal, Z., Alibakhshikenari, M., Collotta, M., and Miah, M. S. (2024). Transmixer-af: advanced real-time detection of atrial fibrillation utilizing single-lead electrocardiogram signals. *IEEE Access*, 12:143149–143162.

Mohan, A., Elbers, D., Zilbershot, O., Afghah, F., and Vorchheimer, D. (2024). Deciphering heartbeat signatures: a vision transformer approach to explainable atrial fibrillation detection from ecg signals. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–6. IEEE.

Sociedade Brasileira de Cardiologia (2026). Cardiômetro: Mortes por Doenças Cardiovasculares no Brasil. `http://www.cardiometro.com/default.asp`. Acessado em: 23 de fevereiro de 2026.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., and Sun, L. (2022). Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*.

World Health Organization (2025). Cardiovascular diseases (CVDs). `https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`. Acessado em: 23 de fevereiro de 2026.

Xiao, Q., Lee, K., Mokhtar, S. A., Ismail, I., Pauzi, A. L. b. M., Zhang, Q., and Lim, P. Y. (2023). Deep learning-based ecg arrhythmia classification: A systematic review. *Applied Sciences*, 13(8):4964.

Yildirim, O., Talo, M., Ciaccio, E. J., San Tan, R., and Acharya, U. R. (2020). Accurate deep neural network model to detect cardiac arrhythmia on more than 10,000 individual subject ecg records. *Computer methods and programs in biomedicine*, 197:105740.

Zhao, Y., Ren, J., Zhang, B., Wu, J., and Lyu, Y. (2023). An explainable attention-based tcn heartbeats classification model for arrhythmia detection. *Biomedical Signal Processing and Control*, 80:104337.

Zheng, J., Zhang, J., Danioko, S., Yao, H., Guo, H., and Rakovski, C. (2020). A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7(1):48.

Zírpolo, A. S., Mesquita, E. T., and Ramos, T. C. (2025). Um modelo explicável para classificação de arritmias cardíacas utilizando a rede lstm. In *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, pages 55–60. SBC.