Um Modelo Híbrido para ECG: Como Combinar Convoluções e Atenção

1st Turi Rezende
Dept. de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
turi@ufmg.br

2nd Wagner Meira Jr
Dept. de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
meira@dcc.ufmg.br

***Abstract*—Electrocardiograms (ECGs) play a crucial role in cardiovascular healthcare, requiring effective analytical models. ECG analysis is inherently hierarchical, involving multiple temporal scales from individual waveforms to intervals within heartbeats, and finally to the distances between heartbeats. Convolutional Neural Networks (CNNs) have demonstrated strong performance in ECG classification tasks due to their inductive bias toward local connectivity and translation invariance. In other domains, Transformers have emerged as powerful models for capturing long-range dependencies. In this regard, this paper introduces HiT-NeXt, a hybrid hierarchical model designed to capture both local morphological patterns and global temporal dependencies by combining CNNs with transformer blocks featuring restricted attention windows. The model incorporates ConvNeXt-based convolutional layers to extract local features and perform patch merging, enabling hierarchical representation learning. Transformer blocks are constrained with local attention windows and leverage relative contextual positional encoding to incorporate positional information effectively into embeddings, enhancing robustness to translations in ECG signal patterns. Experimental results demonstrate that HiT-NeXt outperforms state-of-the-art methods on tasks including ECG abnormality classification and cardiological age prediction, achieving superior performance compared to both existing models and cardiologist evaluations.**

***Index Terms*—ECG classification, age prediction, transformer model, hierarchical model, hybrid model.**

I. Introdução

As doenças cardiovasculares (DCVs) são a principal causa de morte no mundo, totalizando 17,9 milhões de óbitos em 2019 e representando 32% das mortes globais, segundo a Organização Mundial da Saúde (OMS) [1]. Eletrocardiogramas (ECGs), exames simples e não invasivos, desempenham papel crucial no diagnóstico e monitoramento de condições cardiovasculares. Seu uso ganhou ainda mais relevância na saúde digital com a adoção generalizada de ECGs digitais [2]. Nos últimos anos, a inteligência artificial (IA) em eletrocardiografia emergiu como ferramenta valiosa para classificação automática de anormalidades em ECG, predição de sexo e idade [3], segmentação de ondas [3] e previsão de eventos cardíacos [4]. Modelos de IA aplicados a ECGs frequentemente se baseiam em inovações de outros campos de aprendizado de máquina, como visão computacional, processamento de texto e reconhecimento de fala, demonstrando a adaptabilidade de técnicas entre domínios.

Inicialmente, as aplicações de aprendizado de máquina em análise de ECG focavam na extração de características significativas do sinal para melhorar o desempenho de classificação. Técnicas comuns incluíam a análise da morfologia do complexo QRS e dos intervalos RR [5], [6]. Nesse contexto, abordagens baseadas em processamento de sinais, como Transformada de Fourier, Transformada Discreta de Wavelet (DWT) e filtragem adaptativa, foram aplicadas para melhorar a qualidade do sinal, detectar pontos fiduciais e extrair características críticas como o segmento ST e a onda T. Por exemplo, a DWT foi amplamente utilizada para detectar complexos QRS e outros pontos fiduciais devido à sua capacidade de analisar sinais nos domínios de tempo e frequência [7], [8]. Subsequentemente, essas características elaboradas manualmente foram usadas para treinar classificadores tradicionais, como k-Nearest Neighbors (k-NN) e Support Vector Machines (SVM), possibilitando soluções eficazes e computacionalmente eficientes para detecção e classificação de arritmias [9], [10].

Com o advento do aprendizado profundo, houve uma mudança de paradigma nas metodologias de análise de ECG. Redes neurais convolucionais (CNNs) e redes neurais recorrentes (RNNs) vêm sendo cada vez mais utilizadas por sua capacidade de aprender características automaticamente a partir de dados brutos. CNNs aproveitam seus vieses indutivos—como localidade espacial e equivariância a translação—para extrair eficientemente características localizadas de sinais de ECG [11], [12]. RNNs, por sua vez, são adeptas a modelar dependências temporais, capturando de forma eficaz a natureza dinâmica dos ritmos cardíacos [13], [14]. Essas abordagens de aprendizado profundo avançaram significativamente a análise de ECG, alcançando níveis de desempenho que, em alguns casos, superam os de médicos especialistas [11]. Motivado por esses avanços, este trabalho propõe um novo modelo de aprendizado profundo para análise de ECG.

Nos últimos anos, modelos baseados em transformers impulsionaram uma mudança de paradigma em IA, avançando significativamente domínios como visão computacional [15], [16] e processamento de linguagem natural [17]. Apesar desse sucesso, a aplicação de arquiteturas transformer a dados de séries temporais, particularmente ECG, não produziu avanços comparáveis [18]. Essa discrepância pode ser atribuída a di-

versos desafios inerentes. O mecanismo global de autoatenção pode obscurecer padrões locais e temporais cruciais para uma análise precisa de ECG. Além disso, nas abordagens Vision Transformer (ViT), o processo de embedding linear de patches pode acarretar perda de informações temporais de alta resolução, prejudicando o desempenho do modelo [19]. Ademais, transformers tipicamente dependem de codificações posicionais absolutas para incorporar informações sequenciais; porém, essas codificações rompem propriedades de invariança—como a invariância a deslocamento—que poderiam ser úteis em análise de ECG [20]. Consequentemente, a efetividade dos transformers nesse domínio permanece limitada, ressaltando a necessidade de pesquisas adicionais para adaptar e aprimorar esses modelos para aplicações em séries temporais e desbloquear os potenciais benefícios dos transformers na análise de ECG.

De acordo com especialistas em cardiologia, um modelo desenvolvido para análise de ECG deve ser capaz de extrair e aprender as seguintes características: a morfologia local detalhada das ondas individuais, os intervalos entre ondas dentro de um único batimento, informações gerais sobre um batimento e as distâncias entre batimentos consecutivos. Isso reforça a necessidade de extração de características em múltiplas escalas temporais e de abstração. O modelo deve capturar informações locais e integrá-las de forma eficaz com informações extraídas do contexto global. Tal requisito sugere naturalmente a combinação de modelos que se destacam na extração de características locais, como CNNs, com modelos capazes de aprender características em contexto global, como transformers.

Esforços anteriores buscaram combinar modelos convolucionais e transformers seguindo uma das três abordagens:

Incorporação de viés convolucional em transformers: Essa linha de trabalho baseia-se em restringir o mecanismo de atenção para impor a extração de características locais em múltiplos níveis hierárquicos, conforme proposto em [16].

Incorporação de viés de transformer em CNNs: Essa abordagem integra propriedades próprias de transformers em modelos convolucionais ao empregar grandes tamanhos de kernel e bottlenecks invertidos, aumentando o campo receptivo e buscando aproximar o contexto global [21].

Concatenação de transformers com CNNs: Essa abordagem adota blocos convolucionais nas primeiras camadas para extrair características locais, seguidos de blocos transformer para aprender a contextualizar globalmente as características extraídas [22].

As duas primeiras abordagens enfrentam limitações inerentes a seus respectivos desenhos de modelo: o Swin Transformer tem dificuldade em replicar a extração eficiente de características alcançada por CNNs durante a redução de resolução [23], enquanto o grande campo receptivo do ConvNeXt, embora aprimore o contexto global, é insuficiente para capturar as ricas representações contextuais possibilitadas por mecanismos de atenção. A terceira abordagem, contudo, frequentemente leva à saturação de desempenho, já que implementações existentes de blocos convolucionais e transformer lutam para equilibrar eficiência e desempenho simultaneamente [24].

Neste trabalho, argumentamos que as três abordagens não são mutuamente exclusivas e podem ser combinadas de forma coesa para aproveitar os pontos fortes tanto dos transformers quanto das CNNs. Propomos o *HiT-NeXt*, um modelo hierárquico híbrido para análise de ECG que integra blocos convolucionais e transformer intercalados com janelas de atenção local restritas. O mecanismo de atenção local, combinado com aprendizagem hierárquica e multinível, permite que o HiT-NeXt extraia informações de sinais de ECG em diversos níveis de detalhe e escalas temporais, equilibrando efetivamente a extração de características globais via convoluções e o processamento contextual local por meio de atenção.

O modelo HiT-NeXt incorpora os seguintes conceitos técnicos e estruturais principais:

Extração de Características Aprimorada: Camadas convolucionais com aprimoramentos baseados no ConvNeXt, como bottlenecks invertidos e Global Response Normalization (GRN), possibilitam a extração e agregação eficaz de padrões locais em ECG.

Aprendizado de Representação Contextual: Blocos transformer com janelas de atenção restritas capturam dependências temporais ao transformar embeddings com base em relacionamentos contextuais locais.

Arquitetura Hierárquica: O HiT-NeXt combina blocos convolucionais e transformer em uma arquitetura multiestágio, possibilitando aprendizado em diversas escalas temporais e de abstração.

Codificação Posicional Relativa Contextual: Um mecanismo inédito que combina codificação posicional relativa e Contextual Position Encoding (CoPE) aprimora a informação de dinâmica temporal ao mesmo tempo em que oferece robustez a translações.

O HiT-NeXt supera os principais baselines do estado da arte em tarefas de classificação de anormalidades de ECG e predição de idade cardiológica, demonstrando desempenho superior nesses domínios desafiadores. Além disso, fornecemos uma discussão extensa detalhando o processo de desenvolvimento do modelo e os principais insights obtidos a partir de resultados intermediários. Essa análise abrangente documenta as melhorias iterativas alcançadas ao longo da fase de projeto, oferecendo insights valiosos para orientar desenvolvimentos futuros em modelagem preditiva baseada em ECG.

II. Trabalhos Relacionados

A aplicação de inteligência artificial (IA) na análise de ECG tem apresentado avanços notáveis nos últimos anos. Uma ampla gama de metodologias, desde técnicas tradicionais de aprendizado de máquina [10] até arquiteturas de aprendizado profundo de ponta [2], foi desenvolvida para aumentar a precisão e a confiabilidade da interpretação automática de ECGs. Nesta seção, revisamos trabalhos relevantes que exploram o uso de redes neurais convolucionais (CNNs), redes neurais recorrentes (RNNs) e transformers na análise de ECG, visando

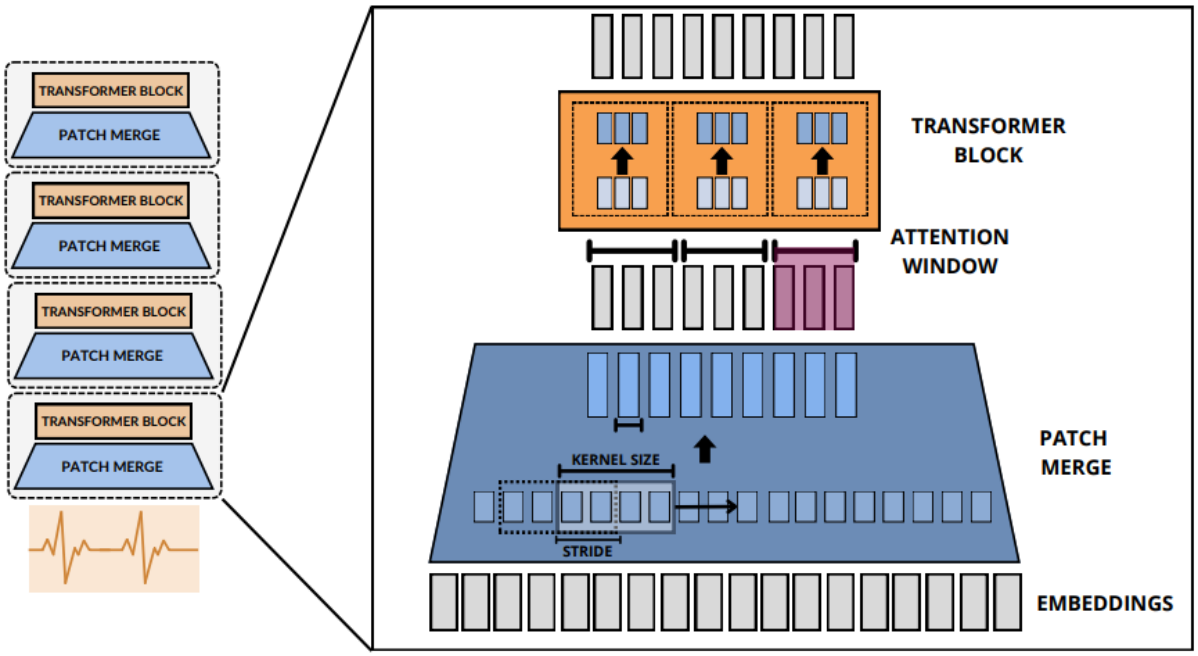

Fig. 1. Arquitetura de alto nível do modelo HiT-NeXt.

construir a trajetória técnica que fundamenta o presente trabalho.

Em [12], os autores introduziram uma arquitetura CNN profunda treinada de ponta a ponta em um conjunto de dados de grande escala, alcançando desempenho equivalente ao de cardiologistas na detecção de arritmias. O sucesso do modelo foi atribuído ao grande volume de dados rotulados e a escolhas criteriosas nas camadas convolucionais, priorizando kernels de grande tamanho para capturar dependências de longo alcance nos sinais de ECG e extrair características clinicamente relevantes.

Dando continuidade a esse avanço, [11] propuseram uma nova arquitetura ResNet para analisar ECGs de 12 derivações no diagnóstico automático de diversas anormalidades cardíacas. Além disso, essa arquitetura foi utilizada para prever a idade do paciente a partir dos sinais de ECG, demonstrando que idades preditas significativamente maiores do que a idade cronológica estão correlacionadas a maior mortalidade [3].

No tocante à classificação de arritmias, [25] apresentaram uma abordagem híbrida que combina CNNs com camadas bidirecionais de memória de longo curto prazo (BiLSTM). A arquitetura utiliza CNNs para extração de características espaciais e BiLSTMs para capturar dependências temporais, fornecendo solução robusta para detecção precisa de fibrilação atrial.

Para integrar arquiteturas baseadas em transformers na análise de ECG, trabalhos recentes investigaram a sinergia entre camadas convolucionais para extração inicial de características e transformers para modelagem de relacionamentos globais. Por exemplo, em [22], foi proposto o modelo *ECG DETR*, que combina um backbone CNN com um transformer para reformular a classificação de arritmias como tarefa de detecção de objetos, permitindo localização e classificação simultâneas de batimentos cardíacos. Essa abordagem elimina a necessidade de segmentação explícita e aproveita o mecanismo de autoatenção para capturar dependências entre batimentos de forma eficaz. De maneira semelhante, foi desenvolvido em [26] um modelo híbrido CNN-transformer, onde camadas CNN extraem características espaciais e um codificador transformer aprende relações temporais ao longo de segmentos de ECG.

Baseando-se no sucesso de [16], os autores de [27] introduziram uma abordagem inovadora que explora a estrutura periódica inerente aos sinais de ECG utilizando um framework alinhado a batimentos (*Beat-Aligned Transformer* – BaT). O BaT emprega mecanismos de atenção local e processamento progressivo: o ECG de entrada é decomposto em segmentos ou "batimentos", aplica-se autoatenção local dentro de cada janela de batimento e, em seguida, mesclam-se progressivamente essas representações para capturar características locais e globais. Entretanto, o modelo apresenta a limitação de depender de etapas complexas de pré-processamento dos dados. Alinhamento e segmentação precisos de batimentos são necessários para o funcionamento do modelo, o que pode introduzir vieses ou erros, já que variações de morfologia, ruído ou batimentos irregulares podem comprometer a segmentação e o alinhamento corretos.

Aproveitando os avanços dos modelos anteriores de ECG e os recentes progressos em aprendizado profundo, este trabalho propõe uma nova abordagem para análise de eletrocardiogramas. O modelo adotado utiliza uma arquitetura transformer

hierárquica, incorporando conceitos de última geração oriundos de CNNs e um mecanismo contextual de codificação posicional.

III. Métodos

A análise eficaz de sinais de eletrocardiograma (ECG) exige a captação de informações em múltiplas escalas temporais e níveis de abstração. A natureza hierárquica dos sinais de ECG compreende três aspectos fundamentais: (1) a morfologia das ondas individuais (por exemplo, ondas P, complexo QRS e onda T), (2) os intervalos entre ondas dentro de um batimento (por exemplo, intervalo PR e intervalo QT) e (3) as distâncias entre batimentos consecutivos (intervalo RR), que podem indicar arritmia. Assim, uma análise abrangente de ECG deve extrair características nesses diversos níveis de detalhe para identificar padrões morfológicos granulares e tendências globais entre batimentos.

A. Arquitetura HiT-NeXt

O modelo *HiT-NeXt* foi projetado para capturar esses padrões hierárquicos em dados de ECG por meio de uma arquitetura híbrida que intercala blocos convolucionais e blocos transformer em múltiplos estágios. O modelo opera em diversas escalas temporais e de abstração, aprendendo, de forma progressiva, padrões locais e globais a partir dos dados.

A Figura 1 apresenta uma visão geral da arquitetura do HiT-NeXt. Em cada estágio, blocos convolucionais — denominados *patch merging blocks* — extraem e agregam características locais enquanto reduzem a dimensionalidade do sinal. Esses blocos capturam informações morfológicas finas de pequenos segmentos de onda, como o formato do complexo QRS ou das ondas P. Após a extração de características locais, blocos transformer modelam representações contextuais ao aprender as relações entre esses segmentos. Para manter o foco do modelo em padrões localizados relevantes para a análise de ECG, o mecanismo de atenção é restrito a uma janela fixa, garantindo que os embeddings sejam contextualizados apenas com base em segmentos vizinhos.

Esse processo de alternância entre camadas convolucionais e transformer se repete em múltiplos estágios. Cada estágio sucessivo reduz a resolução temporal enquanto expande o campo receptivo, capturando progressivamente padrões de ordem superior, como as relações entre batimentos completos. Essa abordagem hierárquica assegura que características de forma de onda locais e padrões globais de frequência cardíaca sejam modelados de forma eficaz para uma análise precisa de ECG.

O modelo recebe dados de entrada no formato `(B, seq_len, num_leads)` e os transforma progressivamente, preservando características locais e globais relevantes, onde `B` é o tamanho do *batch*, `seq_len` representa o número de amostras temporais do sinal de ECG e `num_leads` corresponde ao número de derivações.

Em cada bloco convolucional (*patch merging*), a dimensão temporal é reduzida por um fator 4 e o número de canais é dobrado. Os blocos transformer não alteram a dimensionalidade. Após quatro estágios, aplica-se *global average pooling*, seguido por um MLP que gera *logits* no formato `(B, num_classes)`.

1) Patch Merging: O processo de *patch merging* emprega camadas convolucionais para agregar informação local e reduzir a dimensionalidade dos dados. Cada bloco de *patch merging* é composto por dois sub-blocos residuais e não lineares. Ambos seguem as ideias apresentadas em [21], [23], que estabeleceram o novo estado da arte em classificação de imagens, superando modelos baseados em transformers. Isso é alcançado principalmente mediante o uso de uma arquitetura *inverted bottleneck* combinada a kernels de grande tamanho, capturando dependências de longo alcance de forma eficaz. Diferentemente da estrutura *bottleneck* tradicional (como em ResNet), onde o fluxo de dados segue compressão-expansão, o *inverted bottleneck* primeiro expande as dimensões do tensor para enriquecer a representação, aplica uma convolução com kernel grande a fim de captar padrões espaciais mais amplos e, por fim, comprime novamente com uma convolução 1×1. Essa estrutura permite aprendizagens mais ricas, pois as representações intermediárias operam em um espaço de maior dimensão.

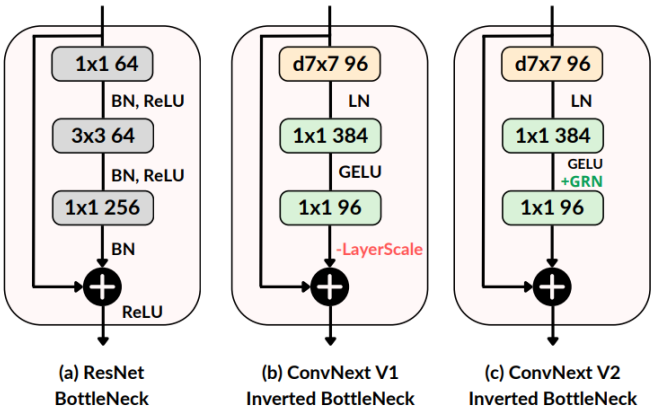

Fig. 2. Comparação entre o *bottleneck* clássico da ResNet e o *inverted bottleneck*, primeiras e segundas versões do ConvNeXt. Os valores ilustrados são arbitrários para fins didáticos; em nosso trabalho, fizemos pequenas modificações nos tamanhos de kernel, e as dimensões de embedding mudam em cada estágio do modelo hierárquico. A ideia central de expansão-compressão, contudo, permanece.

A arquitetura completa do processo de *patch merging* é mostrada na Figura 3. O primeiro sub-bloco possui um ramo principal formado por camadas convolucionais intercaladas com *Layer Normalization*, ativação GELU e *dropout*. O processo inicia com uma convolução de kernel 10, passo 4 e *padding* 4, reduzindo o número de embeddings por um fator 4. Essa camada recebe embeddings de dimensão *in_c* e produz embeddings de dimensão *out_c*, onde $out\_c = 2 \times in\_c$, seguida de *Layer Normalization* para estabilizar o treinamento. Em seguida, uma convolução 1×1 expande a dimensão das características em um fator 4, alinhando-se ao princípio do

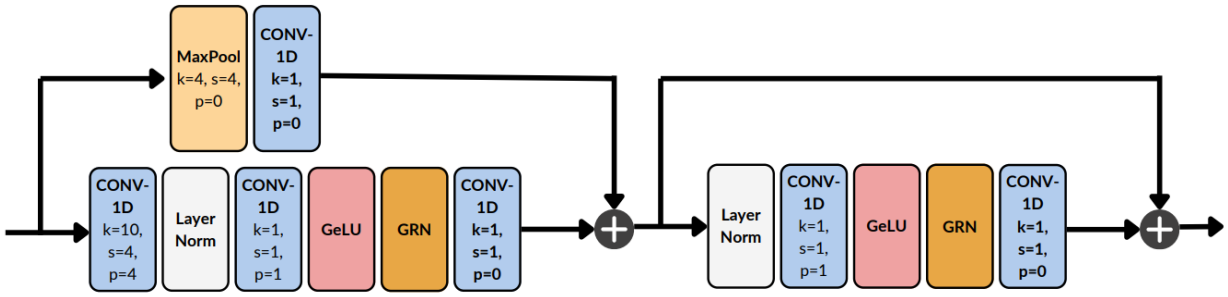

Fig. 3. Arquitetura convolucional proposta para *patch merging*.

inverted bottleneck. Essa expansão é sucedida por ativação GELU e *dropout*.

Conforme [21], incorporamos uma camada *Global Response Normalization* (GRN) ao modelo. A GRN promove diversidade de características mediante: (1) *agregação global de características*, onde se calcula a norma L2 de cada canal; (2) *normalização divisiva*, ajustando essas magnitudes; e (3) *calibração de respostas*, ponderando as características de entrada conforme as magnitudes normalizadas:

1) **Agregação global**: $G(X)_i = \|X_i\|_2$
2) **Normalização divisiva**: $N(G(X)_i) = \frac{\|X_i\|_2}{\sum_{j=1}^{\dim} \|X_j\|_2}$
3) **Calibração**: $X_i = \gamma \times X_i \times N(G(X)_i) + \beta + X_i$

em que $X \in \mathbb{R}^{(B,\text{seq\_len},\dim)}$.

Por fim, uma última convolução retorna os embeddings à dimensão *out_c*, concluindo a fase de compressão do *inverted bottleneck*. O ramo residual do primeiro sub-bloco contém uma camada *MaxPool* (redução $\times 4$) e uma convolução 1×1 que mapeia *in_c* $\rightarrow$ *out_c*. A soma dos ramos principal e residual segue o padrão de redes residuais.

O segundo sub-bloco preserva o *inverted bottleneck*, mas não reduz o número de embeddings: apenas executa as convoluções 1×1 de expansão e compressão. Seu ramo residual é identidade, atuando como "corredor de gradiente".

2) Blocos Transformer: Os blocos transformer seguem a arquitetura padrão de codificador transformer, porém com modificações essenciais. Primeiramente, o mecanismo de atenção é restrito a uma janela fixa, conforme [16]. Esse mecanismo pode ser visto como uma transformação contextual: valores de um embedding são ajustados pelos valores de embeddings vizinhos na mesma janela, guiados por pontuações de atenção (produto escalar). Restringindo a atenção, o modelo gera embeddings mais detalhados, baseados apenas em informação local.

Intercalando blocos transformer com blocos de *patch merging*, aumentamos gradualmente a escala temporal de contexto: estágios iniciais focam em sub-batimentos; o segundo estágio cobre aproximadamente um batimento completo; estágios posteriores englobam múltiplos batimentos. Implementamos a estratégia de *janela com deslocamento cíclico*, garantindo robustez a translações — fundamental em ECGs, pois deslocamentos de batimentos entre exames não devem alterar o diagnóstico. Assim, dispensamos etapas heurísticas de pré-processamento (segmentação de batimentos ou detecção de ondas) e o modelo aprende diretamente dos dados brutos.

3) Codificação Posicional Relativa: O mecanismo de autoatenção é invariante a permutações e não possui, intrinsecamente, noção da ordem sequencial. Em transformers tradicionais, usa-se *absolute positional encoding* (APE), adicionando um vetor específico de posição ao embedding de entrada. Contudo, isso cria dependência de posições absolutas, comprometendo a robustez a translações. Adotamos, portanto, *relative positional encoding* (RPE), incorporando relações posicionais diretamente na atenção:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + E\right) V \qquad (1)$$

em que $E \in \mathbb{R}^{B \times H \times L \times L}$ codifica a relação posicional relativa. Empregamos duas formas complementares de RPE: um *relative position bias* aprendível (RPB) e o *Contextual Position Encoding* (CoPE) [28].

a) Relative Position Bias (RPB).: Para uma janela de tamanho M, inicializamos $\hat{B} \in \mathbb{R}^{2M-1}$. Para posições i, j, com diferença relativa $d = i - j$:

$$B_{ij} = \hat{B}_{d+M-1}.$$

O vetor $\hat{B}$ aprende a importância de cada distância relativa, mas ignora o conteúdo entre posições.

b) Contextual Position Encoding (CoPE).: O CoPE introduz dependências contextuais usando uma função *gate*:

$$g_{ij} = \sigma(q_i^\top k_j), \qquad j < i.$$

O valor $g_{ij} \in (0, 1)$ indica quão semelhante k_j é a q_i. O valor posicional contextual entre i e j é então

$$p_{ij} = \sum_{k=j}^{i} g_{ik}.$$

Por meio de interpolação, obtém-se o embedding posicional contínuo $e[p_{ij}]$, sensível ao conteúdo (por exemplo, quantidade de ondas significativas entre dois pontos).

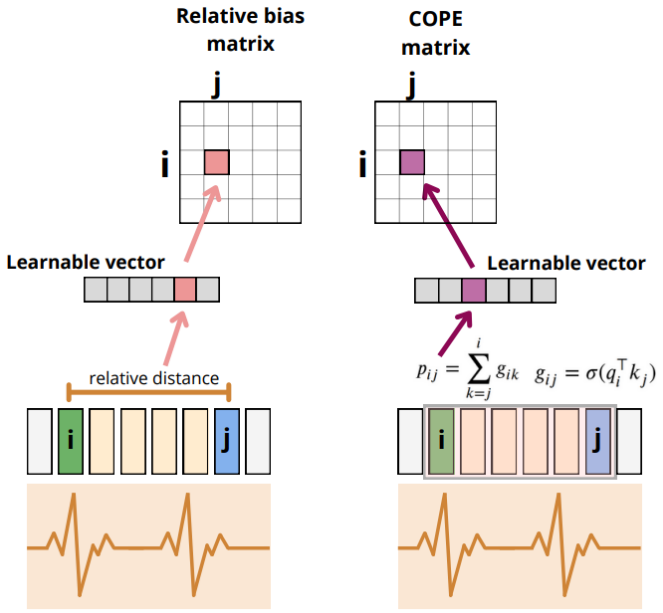

Fig. 4. Ilustração de alto nível da codificação posicional relativa que combina RPB e CoPE.

c) Combinação de RPB e CoPE.: Introduzimos um vetor de pesos aprendível $\boldsymbol{\alpha} \in \mathbb{R}^2$:

$$\mathbf{E}_{\text{combined}} = \alpha_1 \mathbf{E}_{\text{CoPE}} + \alpha_2 \mathbf{E}_{\text{RPE}},$$

$$\boldsymbol{\alpha}_{\text{norm}} = \frac{\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|_2}, \quad \mathbf{E}_{\text{combined}} = \alpha_{\text{norm},1} \mathbf{E}_{\text{CoPE}} + \alpha_{\text{norm},2} \mathbf{E}_{\text{RPE}}.$$

Dessa forma, exploramos as vantagens complementares de ambos os métodos. Utilizando janelas com deslocamento cíclico e somente codificações relativas (sem APE), o modelo mantém robustez a translações — crucial em ECGs. Apesar da complexidade técnica, o HiT-NeXt opera fim-a-fim, dispensando pré-processamento manual e engenharia de atributos.

IV. Configuração Experimental

A. Conjuntos de dados

Para o desenvolvimento e treinamento do modelo, utilizamos a amostra pública de 15% do conjunto CODE (*Clinical Outcomes in Digital Electrocardiography*) [29], denominada **CODE-15**. O CODE é uma coorte retrospectiva que reúne mais de 2 milhões de eletrocardiogramas (ECGs) digitais pareados a registros de mortalidade e internações hospitalares do estado de Minas Gerais, Brasil. Uma equipe de cardiologistas da Rede de Teleassistência de Minas Gerais (TNMG) analisou os ECGs para identificar seis anormalidades cardíacas comuns em exames eletrocardiográficos: bloqueio atrioventricular de primeiro grau (1ª AVB), bloqueio de ramo direito (RBBB), bloqueio de ramo esquerdo (LBBB), bradicardia sinusal (SB), fibrilação atrial (AF) e taquicardia sinusal (ST). Essas anormalidades são clinicamente relevantes, pois se associam a maior risco de eventos cardiovasculares e podem exigir intervenções específicas e acompanhamento regular.

O CODE-15 contém 345 779 exames de 233 770 pacientes e tem sido amplamente adotado na pesquisa em ECG, servindo como *benchmark* para desenvolvimento e avaliação de modelos de aprendizado profundo [11], [30].

Para avaliar o desempenho do nosso modelo, empregamos o conjunto **CODE-TEST**, também coletado pela TNMG. Esse conjunto inclui 827 exames de ECG rotulados por consenso rigoroso de dois ou três especialistas em cardiologia, contemplando as mesmas seis anormalidades citadas acima. Os rótulos de alta qualidade do CODE-TEST fornecem referência confiável para a avaliação dos modelos.

Para o desenvolvimento e validação do HiT-NeXt, adotamos quatro subconjuntos não sobrepostos:

- **Treino**: 90% do CODE-15, usado para treinar o modelo.
- **Validação**: 5% do CODE-15, usado para *early stopping* e prevenção de sobre-ajuste.
- **Desenvolvimento**: 5% do CODE-15, usado para ajuste de hiperparâmetros, decisões de arquitetura e estudos de ablação.
- **Teste**: conjunto completo CODE-TEST, empregado na avaliação final contra os métodos-base.

A partição foi realizada aleatoriamente a partir dos IDs de pacientes, garantindo que múltiplos exames de um mesmo paciente não aparecessem em subconjuntos diferentes. A Tabela I apresenta a distribuição das seis anormalidades no CODE-15, bem como a distribuição de idade e sexo. Observe que um paciente pode apresentar mais de uma anormalidade simultaneamente ou nenhuma das condições avaliadas.

B. Modelos de referência e métricas de avaliação

Comparámos o HiT-NeXt a um conjunto de modelos-base pertencentes a diferentes famílias arquiteturais, incluindo CNNs tradicionais e arquiteturas baseadas em transformers, assegurando uma avaliação rigorosa entre paradigmas distintos. Os baselines foram implementados a partir dos códigos originais, com configurações de treinamento recomendadas pelos respectivos autores. Todos os modelos foram treinados no mesmo conjunto de Treino, validados no conjunto de Validação e avaliados no conjunto de Teste.

Empregamos métricas clássicas de classificação: acurácia, F1-score, precisão e revocação. Tais métricas foram calculadas individualmente para cada condição cardíaca e de forma agregada, proporcionando visão detalhada do desempenho em diferentes patologias.

C. Detalhes de implementação

O treinamento utilizou o otimizador AdamW [31], com *scheduler* de taxa de aprendizado de decaimento cosseno (*cosine annealing*) [32]. A taxa inicial foi 10^{-4} e decaiu de forma cossenoidal até 10^{-5}. Implementamos *early stopping*, encerrando o treinamento se o erro de validação não diminuísse por sete épocas consecutivas. Os experimentos foram executados em paralelo em 4 GPUs NVIDIA V100.

- **Classificação**: utilizou-se a *Binary Cross-Entropy with Logits* (*BCEWithLogitsLoss*) com `num_classes` neurônios de saída—um por classe—permitindo

classificação multi-rótulo e multi-classe simultaneamente. Se nenhum rótulo estiver presente, todos os logits permanecem abaixo do limiar após a aplicação da função sigmóide.

- **Predição de idade**: empregou-se a perda de Erro Quadrático Médio (MSE) entre idade predita e idade real.

TABLE I
DISTRIBUIÇÃO DAS SEIS ANORMALIDADES DE ECG, IDADE E SEXO NOS CONJUNTOS UTILIZADOS.

Categoria	Variável	Treino/Val/Dev (n = 345 779)	Teste (n = 827)
Anormalidade	1ª AVB	5 716 (1,7%)	28 (3,4%)
	RBBB	9 672 (2,8%)	34 (4,1%)
	LBBB	6 026 (1,7%)	30 (3,6%)
	SB	5 605 (1,6%)	16 (1,9%)
	AF	7 033 (2,0%)	13 (1,6%)
	ST	7 584 (2,2%)	36 (4,4%)
Faixa etária	16–25	32 820 (9,5%)	43 (5,2%)
	26–40	66 729 (19,3%)	122 (14,8%)
	41–60	100 072 (28,9%)	340 (41,1%)
	61–80	112 181 (32,4%)	278 (33,6%)
	≥ 81	33 957 (9,8%)	44 (5,3%)
Sexo	Masculino	206 576 (59,7%)	321 (38,8%)
	Feminino	139 203 (40,3%)	506 (61,2%)

V. RESULTADOS

O método foi testado em dois cenários: identificação e classificação de anormalidades em ECGs e predição de idade.

A. Identificação de anormalidades em ECG

Avaliamos primeiramente o desempenho do HiT-NeXt na tarefa de classificacao das seis categorias previamente definidas: 1ª AVB, RBBB, LBBB, SB, AF e ST. A Tabela II apresenta os resultados obtidos. Cabe ressaltar que o conjunto de dados é desbalanceado, com frequencias distintas para cada condicao, o que ajuda a explicar a alta acuracia observada em todos os modelos. Assim, metricas como precisao, revocacao e, sobretudo, o F1-score — que combina precisao e revocacao em uma unica medida — tornam-se fundamentais para uma avaliacao mais robusta.

O modelo proposto, HiT-NeXt, obteve o maior F1-score entre todos os modelos avaliados. Apesar de apresentar uma precisao ligeiramente inferior a de alguns baselines, o HiT-NeXt alcancou revocacao significativamente superior (0,862 contra 0,799 do BAT, o segundo melhor valor). Isso indica que o modelo identificou uma proporcao maior de casos reais entre os ECGs com determinada anormalidade. Tal ganho em revocacao nao ocorreu a custa de uma queda acentuada em precisao, como evidenciado pelo F1-score geral superior. Em aplicacoes medicas, a revocacao e critica, pois reflete a capacidade de detectar casos positivos dentre os individuos afetados — fator essencial para a assistencia ao paciente.

A Tabela III apresenta o F1-score por classe. O HiT-NeXt atingiu o melhor desempenho na maioria das classes. Nos poucos casos em que não ficou em primeiro lugar, seu desempenho permaneceu muito próximo e competitivo, demonstrando robustez e consistência. Destaca-se a classe 1ª AVB, em que o HiT-NeXt obteve F1-score substancialmente superior aos baselines, possivelmente graças ao seu mecanismo hierárquico de atenção local — capaz de capturar padrões morfológicos finos — combinado à codificação posicional relativa, que permite aprender distâncias temporais como o intervalo PR.

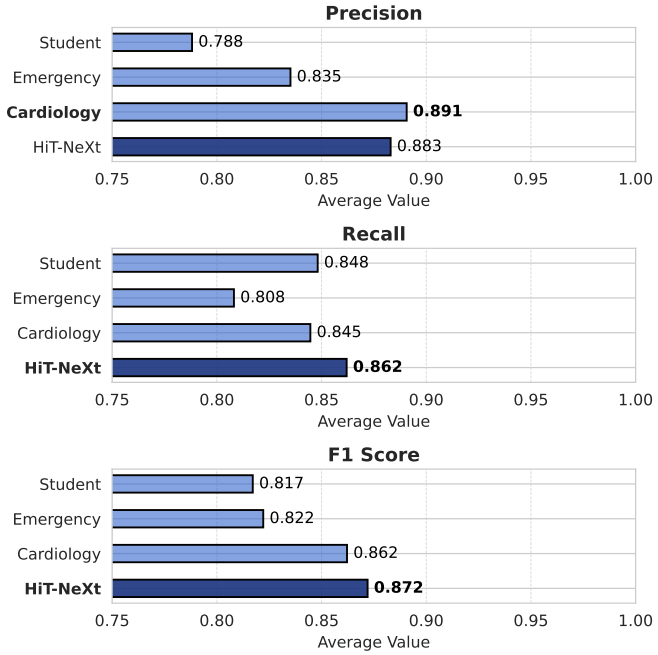

Fig. 5. Comparação média de Precisão, Revocação e F1-score entre o modelo proposto e as avaliações de residentes de cardiologia, residentes de emergência e estudantes de medicina.

O conjunto CODE-TEST contém rótulos fornecidos por profissionais em diferentes estágios de formação: (i) residentes de cardiologia do 4º ano (cardio.), (ii) residentes de emergência do 3º ano (emerg.) e (iii) estudantes de medicina do 5º ano (stud.). Como referência-verdade utilizou-se o consenso de três cardiologistas sêniores. Em relação a essa referência, o HiT-NeXt superou todos os grupos de avaliadores em revocação e F1-score; em precisão, ficou abaixo dos residentes de cardiologia, mas acima dos estudantes e dos residentes de emergência.

B. Predição de idade

A predição de idade a partir do ECG ganhou destaque por possibilitar a estimativa de um marcador de idade biológica associado à saúde cardiovascular [3], [33]. Um delta de idade elevado (idade predita maior que a cronológica) está associado a maior mortalidade e eventos cardiovasculares [3].

Na Tabela IV apresentamos o erro absoluto médio (MAE) e o erro quadrático médio (MSE) para a tarefa de predição de idade. O HiT-NeXt novamente demonstrou desempenho superior aos baselines, evidenciando sua robustez e flexibilidade para diferentes tarefas.

TABLE II
COMPARAÇÃO DO DESEMPENHO DO HIT-NEXT COM MÉTODOS DE ESTADO DA ARTE.

Métrica	ResNet-1	ResNet-2	ECG-Transform	BAT	ECG-DETR	HiT	HiT-NeXt
Acurácia	0,991	0,989	0,981	0,991	0,984	0,991	**0,993**
Precisão	0,875	0,908	0,711	**0,918**	0,777	0,909	0,883
Revocação	0,778	0,743	0,687	0,799	0,661	0,798	**0,862**
F1-Score	0,814	0,811	0,677	0,848	0,699	0,841	**0,872**

TABLE III
F1-SCORE POR CLASSE NO CONJUNTO DE TESTE.

Anormalidade	ResNet-1	ResNet-2	ECG-Transform	BAT	ECG-DETR	HiT	HiT-NeXt
1ª AVB	0,661	0,719	0,489	0,689	0,631	0,682	**0,750**
RBBB	0,924	0,890	0,909	0,922	0,747	0,886	**0,941**
LBBB	0,927	0,843	0,886	0,945	0,826	0,909	**0,966**
SB	0,767	0,821	0,535	**0,836**	0,588	0,824	0,750
AF	0,703	0,758	0,478	0,818	0,563	0,833	**0,880**
ST	0,897	0,833	0,763	0,870	0,838	0,914	**0,944**
F1 médio	0,814	0,811	0,677	0,848	0,699	0,841	**0,872**

TABLE IV
MAE E MSE NA PREDIÇÃO DE IDADE.

Métrica	ResNet-1	ResNet-2	ECG-Transform	BAT	ECG-DETR	HiT	HiT-NeXt
MAE	11,7	10,86	10,8	13,5	11,4	10,34	**9,5**
MSE	214,6	195,2	191,3	271,9	205,6	199,8	**147,6**

VI. ROTEIRO DE DESENVOLVIMENTO DO MODELO

Nesta seção, apresentamos um estudo de *ablation* estruturado como um roteiro que descreve a contribuição de cada componente do HiT-NeXt para o desempenho final. Escolhemos o formato de roteiro pela clareza com que expõe resultados intermediários e os insights obtidos ao longo do desenvolvimento do modelo, constituindo base para trabalhos futuros em cardiologia.

Todos os experimentos desta seção foram avaliados no *development set*. Esse conjunto apresentou desempenho sistematicamente inferior ao de teste, conforme também relatado em [11]. Os valores reportados correspondem à média de 10 000 reamostragens por *bootstrap*.

A. Modelo baseado em Vision Transformer (ViT)

Inicialmente, partimos de um Vision Transformer (ViT). O sinal de ECG foi dividido em *patches* de 12 canais (as 12 derivações) e, para cada *patch*, aplicou-se projeção linear obtendo-se embeddings iniciais. Esses embeddings foram processados por blocos transformer com atenção global; utilizou-se codificação posicional absoluta senoidal. Cada *patch* continha 16 amostras, resultando em 160 *patches* para sinais de 2560 amostras.

O modelo alcançou F1 médio de 0,535, com desempenho insatisfatório em 1ª-AVB (0,098) e FA (0,337). Tais condições envolvem detalhes sutis, como o prolongamento do intervalo PR ou a ausência de ondas P bem definidas, o que sugere que a atenção global do ViT não capturou informações em janelas temporais pequenas.

B. Estrutura hierárquica com atenção restrita

Para contornar a limitação do ViT, adotamos uma estrutura hierárquica inspirada em [16]. Em cada bloco transformer, um estágio de *patch merging* (inicialmente via *pooling*) agregava informações e reduzia a dimensionalidade, permitindo operar em múltiplas escalas temporais. A atenção foi limitada a janelas locais e aplicou-se deslocamento cíclico.

Essas mudanças elevaram o F1 médio para 0,653, com fortes ganhos em 1ª-AVB (0,485) e FA (0,641).

1) Bloco convolucional na projeção de patches*:* Melhoramos a projeção de *patches* substituindo a projeção linear por um bloco residual convolucional baseado em [11]. O F1 médio passou a 0,656, ressaltando a relevância de embeddings robustos.

2) Blocos convolucionais no patch merging*:* Em seguida, trocamos o *pooling* por (i) uma única convolução e (ii) blocos

convolucionais residuais não lineares na fusão de *patches*. A única convolução atingiu F1 de 0,678, enquanto a versão completa chegou a 0,695, mostrando a importância de blocos robustos na agregação de informações.

3) Arquitetura baseada em ConvNeXt: Mantendo a estrutura residual, redesenhamos os blocos para incluir *inverted bottlenecks*, *kernels* grandes, *Layer Norm* e *Global Response Normalization*, conforme [21], [23]. Também substituímos a codificação posicional absoluta por relativa baseada apenas em distâncias. O F1 médio subiu de 0,695 para 0,704.

C. Codificação Posicional Contextual Combinada

O modelo usava inicialmente apenas o *Relative Position Bias* (RPB) de [16]. Inserimos o *Contextual Positional Encoding* (CoPE) [28] e combinamos ambos linearmente. O F1 médio aumentou para 0,710.

A Tabela V resume o impacto das diferentes variantes de RPE, com e sem codificação posicional absoluta (APE).

TABLE V
COMPARAÇÃO DOS TIPOS DE RPE COM E SEM APE (F1-SCORE MÉDIO).

Tipo de RPE	Com APE	Sem APE
Somente RPB	0,698	0,704
Somente CoPE	0,706	0,706
Combinada	0,703	**0,710**

Modelos sem APE apresentaram melhor desempenho geral, possivelmente devido à maior robustez a translações. O melhor resultado foi obtido pela codificação combinada sem APE.

VII. CONCLUSÕES

Apresentamos o HiT-NeXt, método para análise de ECGs que integra blocos convolucionais e transformer, incorporando avanços recentes de *deep learning* para lidar com padrões locais e globais dos sinais. O método foi avaliado em duas tarefas — classificação de seis anormalidades e predição de idade — superando todos os baselines em F-measure e MSE, respectivamente, e ultrapassando o desempenho de estudantes de medicina na classificação.

Como trabalho futuro, propomos o pré-treinamento auto-supervisionado do HiT-NeXt em grandes conjuntos não rotulados de ECG, seguido de ajuste fino, o que pode melhorar a generalização, especialmente para anomalias raras. Além disso, técnicas de adaptação de domínio podem ampliar a aplicabilidade clínica a populações diversas e diferentes condições de aquisição. Outra linha promissora consiste na implementação de uma CNN modificada, cujo kernel convolucional incorpora cabeças de atenção, permitindo a captura explícita de informações contextuais dentro dos vizinhos espaciais definidos pelo kernel. Espera-se que essa modificação melhore a capacidade do modelo de extrair características contextuais e detalhes morfológicos finos, potencializando ainda mais seu desempenho clínico e versatilidade em diferentes tarefas diagnósticas.

VIII. REFERÊNCIAS

[1] World Health Organization. Cardiovascular diseases, 2024.

[2] Xinwen Liu, Huan Wang, Zongjin Li, and Lang Qin. Deep learning in ecg diagnosis: A review. *Knowledge-Based Systems*, 227:107187, 2021.

[3] Emilly M Lima, Antônio H Ribeiro, Gabriela MM Paixão, Manoel Horta Ribeiro, Marcelo M Pinto-Filho, Paulo R Gomes, Derick M Oliveira, Ester C Sabino, Bruce B Duncan, Luana Giatti, et al. Deep neural network-estimated electrocardiographic age as a mortality predictor. *Nature communications*, 12(1):5117, 2021.

[4] Zahra Ebrahimi, Mohammad Loni, Masoud Daneshtalab, and Arash Gharehbaghi. A review on deep learning methods for ecg arrhythmia classification. *Expert Systems with Applications: X*, 7:100033, 2020.

[5] Chun-Cheng Lin and Chun-Min Yang. Heartbeat classification using normalized rr intervals and morphological features. *Mathematical Problems in Engineering*, 2014(1):712474, 2014.

[6] Ahmad Khoureich Ka. Ecg beats classification using waveform similarity and rr interval. *arXiv preprint arXiv:1101.1836*, 2011.

[7] Juan Pablo Martínez, Rute Almeida, Salvador Olmos, Ana Paula Rocha, and Pablo Laguna. A wavelet-based ecg delineator: evaluation on standard databases. *IEEE Transactions on biomedical engineering*, 51(4):570–581, 2004.

[8] Nobuaki Fujita, Akira Sato, and Masatoshi Kawarasaki. Performance study of wavelet-based ecg analysis for st-segment detection. In *2015 38th International Conference on Telecommunications and Signal Processing (TSP)*, pages 430–434. IEEE, 2015.

[9] Roshan Joy Martis, U Rajendra Acharya, and Lim Choo Min. Ecg beat classification using pca, lda, ica and discrete wavelet transform. *Biomedical Signal Processing and Control*, 8(5):437–448, 2013.

[10] Muhammad Wasimuddin, Khaled Elleithy, Abdel-Shakour Abuzneid, Miad Faezipour, and Omar Abuzaghleh. Stages-based ecg signal analysis from traditional signal processing to machine learning approaches: A survey. *IEEE Access*, 8:177782–177803, 2020.

[11] Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1760, 2020.

[12] P Rajpurkar, AY Hannun, M Haghpanahi, C Bourn, and AY Ng. Cardiologist-level arrhythmia detection with convolutional neural networks. arxiv 2017. *arXiv preprint arXiv:1707.01836*, 2011.

[13] Sumanta Kuila, Namrata Dhanda, and Subhankar Joardar. Ecg signal classification and arrhythmia detection using elm-rnn. *Multimedia Tools and Applications*, 81(18):25233–25249, 2022.

[14] Giulio Ruffini, David Ibanez, Marta Castellano, Stephen Dunne, and Aureli Soria-Frisch. Eeg-driven rnn classification for prognosis of neurodegeneration in at-risk patients. In *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part I 25*, pages 306–313. Springer, 2016.

[15] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[17] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[18] S Ahmed, IE Nielsen, A Tripathi, S Siddiqui, G Rasool, and RP Ramachandran. Transformers in time-series analysis: A tutorial. arxiv 2022. *arXiv preprint arXiv:2205.01138*, 2022.

[19] Pedro Robles Dutenhefner, Turi Andrade Vasconcelos Rezende, Gisele Lobo Pappa, Gabriela Miana de Matos Paixão, Antônio Luiz Pinho Ribeiro, and Wagner Meira Jr. Um transformer hierárquico para classificação e diagnóstico de eletrocardiograma. *Journal of Health Informatics*, 16(Especial), 2024.

[20] Binqiang Chen, Yang Li, and Nianyin Zeng. Centralized wavelet multiresolution for exact translation invariant processing of ecg signals. *IEEE Access*, 7:42322–42330, 2019.

[21] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16133–16142, 2023.

[22] Rui Hu, Jie Chen, and Li Zhou. A transformer-based deep neural network for arrhythmia detection using continuous ecg signals. *Computers in Biology and Medicine*, 144:105325, 2022.

[23] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[24] Shunfeng Li, Chunxue Wu, and Naixue Xiong. Hybrid architecture based on cnn and transformer for strip steel surface defect classification. *Electronics*, 11(8):1200, 2022.

[25] J Wang and W Li. Atrial fibrillation detection and ecg classification based on cnn-bilstm. arxiv 2020. *arXiv preprint arXiv:2011.06187*.

[26] Hany El-Ghaish and Emadeldeen Eldele. Ecgtransform: Empowering adaptive ecg arrhythmia classification framework with bidirectional transformer. *Biomedical Signal Processing and Control*, 89:105714, 2024.

[27] Xiaoyu Li, Chen Li, Yuhua Wei, Yuyao Sun, Jishang Wei, Xiang Li, and Buyue Qian. Bat: Beat-aligned transformer for electrocardiogram classification. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 320–329. IEEE, 2021.

[28] Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. Contextual position encoding: Learning to count what's important. *arXiv preprint arXiv:2405.18719*, 2024.

[29] Antonio Luiz P Ribeiro, Gabriela MM Paixao, Paulo R Gomes, Manoel Horta Ribeiro, Antonio H Ribeiro, Jessica A Canazart, Derick M Oliveira, Milton P Ferreira, Emilly M Lima, Jermana Lopes de Moraes, et al. Tele-electrocardiography and bigdata: the code (clinical outcomes in digital electrocardiography) study. *Journal of electrocardiology*, 57:S75–S78, 2019.

[30] Diogo Tuler, Pedro Robles Dutenhefner, Jose Geraldo Fernandes, Turi Rezende, Gabriel Lemos, Gisele L Pappa, Gabriela Paixao, Antônio Ribeiro, and Wagner Meira Jr. Leveraging cardiologists prior-knowledge and a mixture of experts model for hierarchically predicting ecg disorders.

[31] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[32] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[33] Pedro Robles Dutenhefner, Gabriel Lemos, Turi Rezende, Jose Geraldo Fernandes, Diogo Tuler, Gisele Lobo Pappa, Gabriela Miana Paixao, Antônio Luiz Pinho Ribeiro, and Wagner Meira Jr. Ecg-resnext: Age prediction in pediatric electrocardiograms and its correlations with comorbidities. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 49–60. SBC, 2024.