Which Pretext Tasks Matter? A Comparative Study of Self-Supervised ECG-Based Emotion Recognition

Kevin Gustavo Montero Quispe[1], Daniel Mitsuaki da Silva Utyiama[1], Eduardo James Pereira Souto[1]

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Caixa Postal 69080-900 – Manaus – AM – Brasil

{kgmq,daniel.utyiama, esouto}@icomp.ufam.edu.br

Abstract. *The need for large volumes of labeled data limits the scalability of automatic emotion recognition systems based on biosignals. This study proposes a multitask self-supervised learning (SSL) approach applied to electrocardiogram (ECG) signals, aiming to reduce reliance on extensive manual annotations. A convolutional neural network was pre-trained to simultaneously discriminate six synthetic transformations and subsequently fine-tuned using reduced fractions (5%, 25%, 50%) of labeled data across four public datasets. Results show that combining four or five auxiliary tasks is sufficient to generate effective representations, yielding performance gains of up to 19 percentage points in the classification of valence, arousal, and stress. With only 25% of labeled data, the model reaches near-peak performance, demonstrating the practical viability of the proposed approach in scenarios with limited annotation resources.*

Resumo. *A necessidade de grandes volumes de dados rotulados limita a escalabilidade de sistemas automáticos de reconhecimento de emoções baseados em biossinais. Este estudo propõe uma abordagem multitarefa de aprendizado auto-supervisionado (SSL), aplicada a sinais de eletrocardiograma, visando reduzir a dependência de anotações manuais extensivas. Uma rede convolucional foi pré-treinada para discriminar simultaneamente seis transformações sintéticas, sendo posteriormente ajustada via fine-tuning supervisionado com frações reduzidas (5%, 25%, 50%) de rótulos em quatro bases de dados públicas. Os resultados mostram que a combinação de quatro ou cinco tarefas auxiliares é suficiente para gerar representações eficazes, com ganhos de desempenho de até 19 pontos percentuais na classificação de valência, excitação e estresse. Com apenas 25 % de rótulos, o modelo atinge desempenho próximo ao máximo, evidenciando a viabilidade da proposta em cenários com rotulagem limitada*.

1.Introdução

O reconhecimento automático de emoções tornou-se componente essencial em aplicações de saúde, educação e entretenimento, abrangendo desde videogames com jogabilidade adaptativa em tempo real até sistemas capazes de sinalizar estudantes com baixo engajamento ou monitorar transtornos emocionais em ambientes clínicos [Yang et al., 2018; Wan e Guo, 2020; Montesinos et al., 2019]. Embora promissores, esses sistemas

dependem fortemente de grandes bases rotuladas, cuja obtenção é onerosa e às vezes inviável em cenários sensíveis, comprometendo a capacidade de generalização dos modelos [Mehari e Strodthoff, 2021; Schmidt et al., 2018b; Sarkar e Etemad, 2020].

Entre as modalidades sensoriais utilizadas, o eletrocardiograma (ECG) destaca-se por refletir respostas involuntárias do sistema nervoso autônomo, servindo como marcador direto do estado afetivo [Fang et al., 2024; Wang & Wang, 2025]. Entretanto, ruídos de aquisição e a escassez de rótulos confiáveis dificultam a extração de representações discriminativas. Nesse contexto, o aprendizado auto-supervisionado (*Self-Supervised Learning* – SSL) emerge como alternativa promissora. Nesse paradigma, redes neurais são pré-treinadas por meio de tarefas auxiliares (chamadas *pretext tasks*), que são automaticamente geradas a partir dos próprios dados brutos. Esse processo elimina a necessidade de anotações manuais, permitindo que o modelo aprenda representações informativas que podem ser posteriormente refinadas em tarefas supervisionadas específicas por meio de ajuste fino (fine-tuning). Tarefas auxiliares como permutação temporal, adição de ruído e time-warping já foram exploradas de forma isolada em estudos anteriores [Vazquez-Rodriguez et al., 2022; Kan et al., 2023]. No entanto, ainda há pouco conhecimento sobre quais combinações dessas tarefas pretexto são mais eficazes para o reconhecimento de emoções a partir de sinais de ECG.

Para investigar essa lacuna, este trabalho propõe uma arquitetura convolucional multitarefa que, durante a fase auto-supervisionada, é treinada simultaneamente em seis tarefas auxiliares: adição de ruído, escalonamento, negação, inversão temporal, permutação e *time-warping*. O encoder convolucional resultante é então transferido para a tarefa-alvo de reconhecimento de emoções, por meio de uma etapa de *fine-tuning* supervisionado. A avaliação sistemática envolve quatro bases públicas reconhecidas (AMIGOS, DREAMER, SWELL e WESAD), examinando tanto a influência das diferentes combinações de tarefas auxiliares quanto o desempenho em cenários com proporções variadas de dados rotulados.

O restante deste artigo está organizado da seguinte forma: na Seção 2 são revisados os principais trabalhos relacionados; a Seção 3 detalha a metodologia empregada, incluindo pré-processamento, tarefas auxiliares e arquitetura do modelo; a Seção 4 apresenta os resultados obtidos no pré-treinamento e na classificação de emoções; a Seção 5 conclui o estudo, destacando contribuições e sugestões para pesquisas futuras.

2. Trabalhos Relacionados

O reconhecimento automático de emoções pode explorar três principais fontes de informação: relatos subjetivos, sinais comportamentais (como fala, expressões faciais e gestos) e sinais fisiológicos. Embora questionários de autorrelato sejam amplamente utilizados, eles estão sujeitos a viés de memória e variabilidade individual. Da mesma forma, sinais comportamentais podem ser voluntariamente suprimidos ou requerem infraestrutura específica, como câmeras ou microfones em funcionamento contínuo [Zeng, 2008]. Em contraste, os biossinais como Eletroencefalograma (EEG), ECG, Atividade Eletrodérmica (EDA) e Eletromiogram (EMG) capturam respostas do sistema nervoso autônomo, oferecendo marcadores fisiológicos objetivos e involuntários do estado afetivo [Kreibig, 2010]. O ECG, em particular, destaca-se por sua facilidade de aquisição, podendo ser registrado com eletrodos torácicos ou dispositivos vestíveis,

embora ainda seja suscetível a ruídos induzidos por movimento. Já o EEG, por medir diretamente a atividade cortical, tem sido amplamente adotado em estudos recentes que empregam técnicas de aprendizado profundo.

Apesar do progresso alcançado com arquiteturas profundas, como redes convolucionais (CNNs), recorrentes (RNNs) e Transformers, na detecção de estados emocionais [Khattak, 2021], tais abordagens demandam grandes quantidades de dados rotulados, cuja obtenção é onerosa, especialmente em contextos fisiológicos. Como alternativa, tem ganhado destaque o aprendizado auto-supervisionado (Self-Supervised Learning – SSL), no qual modelos aprendem representações genéricas por meio de tarefas auxiliares (*pretext tasks*) geradas a partir dos próprios dados não rotulados [Chowdhury, 2021]. Entre as estratégias mais exploradas nesse paradigma, destacam-se aquelas baseadas no reconhecimento de transformações sintéticas aplicadas aos sinais: ruído, inversão temporal, permutação, *time-warping*, entre outras. Ao treinar a rede para identificar qual transformação foi aplicada, força-se a extração de representações robustas e invariantes às distorções artificiais, capazes de capturar características estruturais do sinal original.

Sarkar e Etemad (2020) inauguraram esse enfoque ao mostrar que seis transformações aplicadas ao ECG bastam para aprender representações transferíveis e superar modelos totalmente supervisionados. A partir de 2022, a literatura consolidou e expandiu esse paradigma, sempre mantendo as transformações como núcleo do SSL.

Vazquez-Rodriguez et al. (2022) pré-treinaram um transformer em ECGs não rotulados via reconhecimento de seis transformações e estabeleceram novo estado-da-arte usando a base de dados AMIGOS apenas com ECG. Usando o sinal EEG, Wang et al. (2023) comprovaram que uma CNN multitarefa treinada para distinguir seis perturbações no sinal, quando refinado com menos de 25% de rótulos, supera abordagens supervisionadas nos conjuntos SEED e DEAP. Wu et al. (2023) mostraram que aplicar cinco transformações no sinal ECG, EDA e temperatura e depois fundir as representações com um transformer multimodal gera ganhos robustos mesmo com apenas 10 % dos rótulos disponíveis. Em 2024, Wang et al. propuseram uma arquitetura em cascata na qual uma tarefa de reconstrução tempo-frequência precede o reconhecimento de transformações e um módulo contrastivo, resultando em melhor generalização sujeito-independente no EEG. Observa-se, portanto, que variantes e combinações de tarefas de transformação seguem dominando os métodos SSL, seja em modalidades únicas ou na fusão de múltiplos canais.

Apesar do progresso em SSL para reconhecimento de emoções, a literatura carece de um estudo comparativo e sistemático que investigue: quais transformações de ECG geram representações mais informativas; a quantidade ideal de transformações a serem combinadas para maximizar o ganho do SSL; o comportamento dessas escolhas em cenários de dados limitados (5–50% de rótulos) e em múltiplas bases de dados. Trabalhos existentes frequentemente avaliam conjuntos fixos de perturbações em uma única base de dados, com raras investigações sobre sinergias entre as tarefas.

O presente preenche essa lacuna ao apresentar uma análise abrangente de seis combinações de seis transformações clássicas de ECG, avaliadas em quatro bancos de dados públicos. Os resultados indicam que a combinação das seis tarefas auxiliares produz as representações mais robustas, especialmente em cenários com apenas 5% e

Tabela 1. Características das bases de dados utilizadas.

Bases utilizadas	Fs (Hz)	# Participantes	Rótulos emocionais*	# Classes
AMIGOS	256	40	Valência/Excitação	9/9
DREAMER	128	23	Valência/Excitação	5/5
SWELL	2048	25	Valência/Excitação	9/9
WESAD	700	15	neutro, stress, diversão	3

(Fs) Frequência de Amostragem.
(*) Escalas originais discretizadas conforme os protocolos de cada base de dados.

25% dos rótulos disponíveis. Adicionalmente, o estudo oferece diretrizes práticas para a seleção dessas transformações em futuros sistemas de monitoramento emocional baseados em ECG.

3. Método

Esta seção descreve o método em quatro etapas: (1) pré-processamento do ECG; (2) pré-treino auto-supervisionado, onde uma CNN 1-D é otimizada em multitarefa para discriminar seis transformações sintéticas; (3) transferência dos pesos do encoder pré-treinado; e (4) ajuste supervisionado de um classificador que converte as representações em rótulos emocionais.

3.1. Preparação das bases de dados

Os sinais de ECG analisados foram extraídos de quatro bases públicas amplamente utilizadas em estudos de reconhecimento de emoções: AMIGOS [Miranda-Correa et al. 2018], DREAMER [Katsigiannis & Ramzan 2017], SWELL [Koldijk et al. 2014] e WESAD [Schmidt et al. 2018]. Cada conjunto de dados segue protocolo próprio, com variações na frequência de amostragem, na duração das sessões e na taxonomia emocional. Além do ECG, todos registram sinais adicionais como EDA, EMG e respiração, mas neste trabalho apenas o canal cardíaco foi considerado. A Tabela 1 descreve as quatro bases de dados utilizadas. A variedade de protocolos, taxas de amostragem e taxonomias emocionais possibilita avaliar a capacidade de generalização do método.

3.2. Pré-processamento e segmentação

Para normalizar as amostras de sinais obtidos das bases com protocolos distintos, aplicou-se um pré-processamento de três etapas nos dados. Primeiro, todos os sinais foram reamostrados para 256 Hz por interpolação cúbica, igualando a dimensionalidade de entrada apesar das diferentes taxas originais das bases. Na sequência, cada sinal foi filtrado por um passa-alta *Butterworth* de 2ª ordem com frequência de corte em 0,8 Hz, removendo ruídos de baixa frequência sem distorcer a morfologia do sinal. Por fim, realizou-se normalização *z-score* por participante, atenuando diferenças de ganho entre participantes e sessões.

Após o pré-processamento, os sinais de ECG foram segmentados em janelas fixas de 10 segundos sem sobreposição, evitando qualquer potencial de vazamento de dados. Logo cada segmento passou a constituir uma amostra independente para o treinamento e avaliação dos modelos subsequentes.

Tabela 2. Transformações auxiliares aplicadas aos segmentos de ECG durante o pré-treinamento auto-supervisionado.

Transformações do sinal	Parâmetros
T_0 Sem transformação	-
T_1 Ruído gaussiano	SNR (Signal-to-Noise Ratio) = 15 dB
T_2 Escalonamento de amplitude	fator de escala = 1,1
T_3 Negação de polaridade	× (–1)
T_4 Inversão temporal	-
T_5 Permutação de sub-janelas	20 blocos embaralhados
T_6 *Time-warping*	fator local = 1,05 em 20 blocos

3.3. Tarefas Auxiliares (*pretext tasks*)

Seis transformações sintéticas, amplamente citadas na literatura, constituem as tarefas auxiliares binárias resumidas na Tabela 2. Cada transformação é aplicada a 100 % dos segmentos da base de dados. Essas seis transformações, tal como parametrizadas, correspondem ao conjunto "padrão" adotado em trabalhos de referência sobre SSL em biossinais [Sarkar & Etemad 2020; Vazquez-Rodríguez et al. 2022].

3.4. Arquitetura do modelo

A Figura 1 apresenta uma visão geral do modelo arquitetural proposto, estruturado em duas fases: (①) aprendizagem auto-supervisionada, na qual o encoder é pré-treinado com as tarefas auxiliares de discriminação de transformações e (②) transferência de conhecimento supervisionada, na qual o encoder é utilizado para a tarefa de classificações de emoções.

O encoder é uma sequência de três blocos convolucionais 1-D que processa segmentos de ECG de 10 segundos e devolve um vetor latente de 128 componentes. Os detalhes de cada bloco desse encoder estão resumidos na Tabela 3. Na fase auto-supervisionada (① da Figura 1), o encoder alimenta sete cabeças (*heads)* classificadoras binárias, cada qual destinada a indicar a presença de uma transformação T_0 –T_6 (Seção 3.3). Cada cabeça segue a topologia `Dense 128 → ReLU → Dense 1→ Sigmoid` e é otimizada pela entropia-cruzada binária:

$\mathcal{L}_i = -[y_i \, log \, \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)]$, onde yi é o rótulo binário associado à transformação *i*. A perda global utilizada para atualizar os pesos do *encoder* é a soma das perdas individuais, $\mathcal{L}_{global} = \sum_{i=0}^{6} \mathcal{L}_i$.

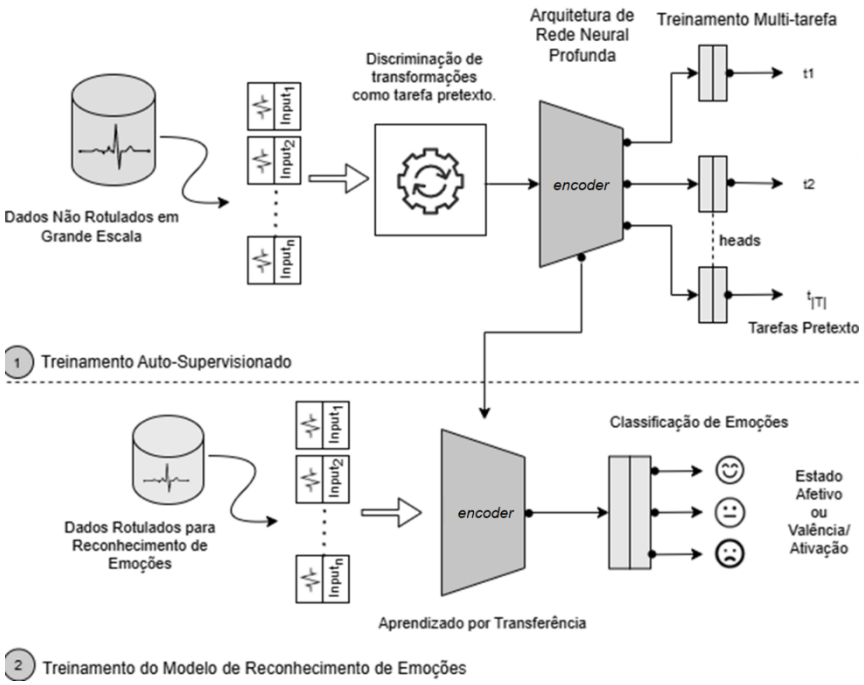

Figura 1 - Visão geral das duas fases: (①) aprendizagem auto-supervisionado e (②) transferência de aprendizado para classificação de emoções.

Após o pré-treino do encoder e as cabeças, os pesos dos blocos convolucionais são congelados e as cabeças auxiliares são descartadas. Na fase de transferência de aprendizado (② da Figura 2), conecta-se um bloco classificador (`Dense 128 → ReLU Dense 128 → Softmax`) que será treinado de forma supervisionada com segmentos rotulados de cada base de dados de emoções. Esse procedimento preserva as representações previamente aprendidas, reduz o custo de anotação e permite adaptar o modelo a particularidades de cada base de dados, proporcionando um reconhecimento de emoções eficaz mesmo em cenários de dados anotados escassos.

3.5. Detalhes de implementação

A arquitetura de rede neural proposta foi desenvolvida e implementada utilizando o framework TensorFlow, com aceleração via GPU Nvidia GeForce GTX 1080 Ti. O treinamento da rede voltada ao reconhecimento de transformações foi conduzido com o otimizador Adam, configurado com uma taxa de aprendizado de 0,001 e um tamanho de lote de 896 amostras (7×128). Esse valor foi definido de modo a assegurar que, em cada iteração, um segmento original e suas respectivas seis versões transformadas estivessem simultaneamente presentes no lote de treinamento, favorecendo o aprendizado auto-supervisionado. Contudo, para a fase de transferência de aprendizado, essa condição não foi necessária, e o tamanho do lote foi ajustado para 128.

A rede de reconhecimento de transformações foi treinada por até 150 épocas, enquanto a rede de reconhecimento de emoções por até 250 épocas. O número de épocas foi determinado para que o processo de aprendizado de ambas as redes atingisse um ponto de estabilidade, sendo também utilizado o mecanismo de Early Stopping com paciência de 10 para otimizar o treinamento e evitar overfitting.

Tabela 3. Detalhes da Arquitetura Implementada.

Camada	Especificação		Dimensões
Entrada	-		2560 × 1
Encoder	2 × (Conv1D, 1 × 32, 32, ReLu)		2560 × 32
	Maxpool, 1 × 8, *Stride* 2		1277 × 32
	2 × (Conv1D, 1 × 16, 64, ReLu)		1277 × 64
	Maxpool, 1 × 8, Stride 2		635 × 64
	2 × (Conv1D, 1 × 8, 128, ReLu)		635 × 128
	Global Max pooling		1 × 128
Camadas de Tarefas	7 *heads*	2 × (Dense, 128 units)	128
Saída		Sigmoid	2

Para avaliar o desempenho da arquitetura proposta em ambas as fases, adotamos a metodologia de validação cruzada com 10 partes (10-fold *cross-validation*), uma abordagem consistente com trabalhos na área, como os de Vazquez-Rodriguez et al. (2022) e Sarkar, P. et al. (2020).

4. Resultados

Os experimentos foram organizados em dois eixos principais: efeito da combinação de tarefas auxiliares no pré-treino auto-supervisionado e impacto da proporção de dados rotulados no desempenho final de reconhecimento de emoções.

4.1. Combinação de tarefas auxiliares

Nesta seção, avaliamos o impacto de diferentes combinações de tarefas auxiliares no pré-treinamento do modelo. Para isso, criamos seis grupos cumulativos (G_1–G_6) de transformações (T_0 –T_6). A Tabela 4 detalha a acurácia média e o desvio padrão para cada tarefa auxiliar em cada grupo.

Observamos uma evolução positiva na acurácia média, que subiu de 97,79% em G_1 para 98,59% em G_6, com desvios padrão consistentemente abaixo de 1,3 p.p., indicando treinamento estável. As altas acurácias gerais (muitas acima de 97%, algumas atingindo 100%) demonstram a capacidade do modelo de aprender os padrões intrínsecos das tarefas auxiliares.

Tarefas como T3 e T4 mostraram desempenho robusto, alcançando quase 100% de acurácia já em G3 e mantendo esse nível. Transformações mais complexas, como T5 (permutação de sub-janelas) e T6 (time-warping), também atingiram altas acurácias (99,7% e 99,5%, respectivamente) sem prejudicar o desempenho das tarefas existentes.

Analisando a progressão, os Grupos G3, G4 e G5 exibiram uma melhora geral na acurácia média conforme mais tarefas eram adicionadas (G5 atingindo a máxima de 98,86%). Isso sugere que a inclusão de tarefas auxiliares específicas pode otimizar as representações aprendidas.

Tabela 4. Resultados do treinamento auto-supervisionado para as tarefas auxiliares combinadas.

Grupo	Tarefas Auxiliares Acurácia							
	T_0	T_1	T_2	T_3	T_4	T_5	T_6	Média
G_1	96,06 ± 2,40	99,52 ± 0,25	-	-	-	-	-	97,79 ± 1,32
G_2	96,93 ± 1,95	99,55 ± 0,22	97,29 ± 1,90	-	-	-	-	97,92 ± 1,36
G_3	96,94 ± 1,90	99,56 ± 0,20	97,98 ± 1,81	99,99 ± 0,01	-	-	-	98.62 ± 0,98
G_4	97,00 ± 1,78	99,56 ± 0,25	97,32 ± 1,75	99,99 ± 0,01	99,95 ± 0,05	-	-	98,76 ± 0,77
G_5	96,81 ± 2,17	99,57 ± 0,43	97,17 ± 2,08	100 ± 0,03	99,92 ± 0,47	99,66 ± 0,68	-	98,86 ± 0,98
G_6	97,04 ± 2,28	96,64 ± 0,21	97,27 ± 2,20	100 ± 0,01	99,95 ± 0,25	99,74 ± 0,61	99,52 ± 1,56	98,59 ± 1,02

4.2. Transferência para classificação de emoções

Avaliamos os modelos pré-treinados na tarefa de classificação de emoções por meio da estratégia de aprendizagem por transferência. Os pesos obtidos na fase autossupervisionada foram transferidos e refinados em um contexto supervisionado, utilizando as bases de dados com anotações emocionais (Tabela 1).

O objetivo desta etapa foi investigar a capacidade de generalização das representações extraídas na fase auto-supervisionada e seu impacto no desempenho da classificação de emoções em diversas bases e escalas emocionais. As Tabelas 5 sumarizam os resultados, apresentando as métricas de acurácia com seus respectivos desvios padrões para cada conjunto de dados avaliado.

Para a base de dados AMIGOS, observamos uma rápida saturação de ganhos. Apenas duas ou três transformações já foram suficientes, com a acurácia de excitação subindo de 70,50% (G_1) para 72,58% (G_2), e a de valência atingindo 64,14% em G_4. Ganhos superiores a 2 p.p. não foram notados após esse ponto.

Em DREAMER, o impacto das transformações foi mais pronunciado. A acurácia de excitação melhorou de 72,87% (G_1) para 77,06% (G_5), enquanto a de valência alcançou 74,36% em G_4. A inclusão de permutação e inversão temporal (T4 e T5) a partir de G_4 foi crucial para explorar as variações rítmicas desse conjunto de dados.

Em SWELL, o método demonstrou alta robustez, com a acurácia de excitação avançando de 92,82% (G_1) para 95,68% (G_5), e a de valência de 89,56%(G_1) para 93,18%(G_5). Contudo, após quatro transformações (G_4), o ganho marginal diminuiu, sugerindo que as invariâncias adicionais saturam rapidamente.

Em WESAD, observamos um crescimento gradual na acurácia, de 91,43% (G_1) para 93,96% (G_5). A inclusão da sexta transformação (time-warping, em G_6) teve pouco

Tabela 5. Classificação de emoções: acurácias (%) obtidas a partir da validação cruzada dos seis grupos de combinações de pre-tasks.

Grupo	AMIGOS		DREAMER		SWELL		WESAD
	Excitação	Valência	Excitação	Valência	Excitação	Valência	Estresse
G_1	70,50± 1,26	63,17± 1,61	72,87± 3,26	71,42± 3,02	92,82± 1,00	89,56± 1,28	91,43± 1,17
G_2	**72,58**± 1,34	62,93± 1,73	75,56± 2,99	73,43± 2,42	94,39± 0,84	92,01± 1,04	92,60± 0,81
G_3	71,06± 1,83	63,48± 2,39	75,97± 2,42	72,70± 2,68	94,49± 0,76	91,96± 1,21	92,83± 0,91
G_4	70,64± 1,95	**64,14**± 2,48	76,82± 2,67	**74,36**± 2,80	95,29± 0,32	92,68± 0,95	93,05± 1,38
G_5	71,65± 1,64	63,79± 2,39	**77,06**± 2,99	73,15± 2,52	**95,68**± 0,53	**93,18**± 0,90	**93,96**± 0,87
G_6	71,12± 1,71	62,42± 1,74	74,03± 2,01	72,94± 3,06	95,47± 0,45	93,15± 0,32	93,36± 1,19

efeito, com acurácia de 93,36%, indicando que cinco perturbações são suficientes para este conjunto de dados.

Em síntese, os grupos com quatro ou cinco transformações (G_4–G_5) apresentaram o melhor equilíbrio entre benefício e custo computacional, adicionando tipicamente de 2 a 4 p.p. de ganho em relação ao grupo base (G_1). A introdução de uma sexta transformação, entretanto, resultou em ganhos mínimos ou neutros, e em alguns casos, até em uma ligeira regressão.

4.3. Análise do impacto da fração de dados rotulados

Nesta seção, investiga-se a influência da quantidade de dados rotulados disponíveis para *fine-tuning* no desempenho do modelo. O objetivo é quantificar a eficiência de dados da abordagem de treinamento auto-supervisionado em comparação com um treinamento puramente supervisionado. Para tal, os modelos foram treinados com frações de 5%, 25% e 50% do conjunto de treinamento, mantendo-se o conjunto de teste integral para a avaliação. Os resultados comparativos são apresentados na Figura 2.

Os resultados demonstram a acentuada superioridade da abordagem auto-supervisionada em grande parte das configurações experimentais, abrangendo as tarefas de classificação de excitação, valência e estresse. Tal desempenho corrobora a hipótese de que o pré-treinamento, ao explorar a estrutura intrínseca dos sinais fisiológicos, induz a extração de características latentes mais robustas e generalizáveis. Notavelmente, essa robustez é particularmente relevante em cenários com disponibilidade limitada de dados rotulados, nos quais o modelo precisa alcançar desempenho satisfatório utilizando apenas uma fração dos dados anotados provenientes de uma base extensiva de sinais de ECG.

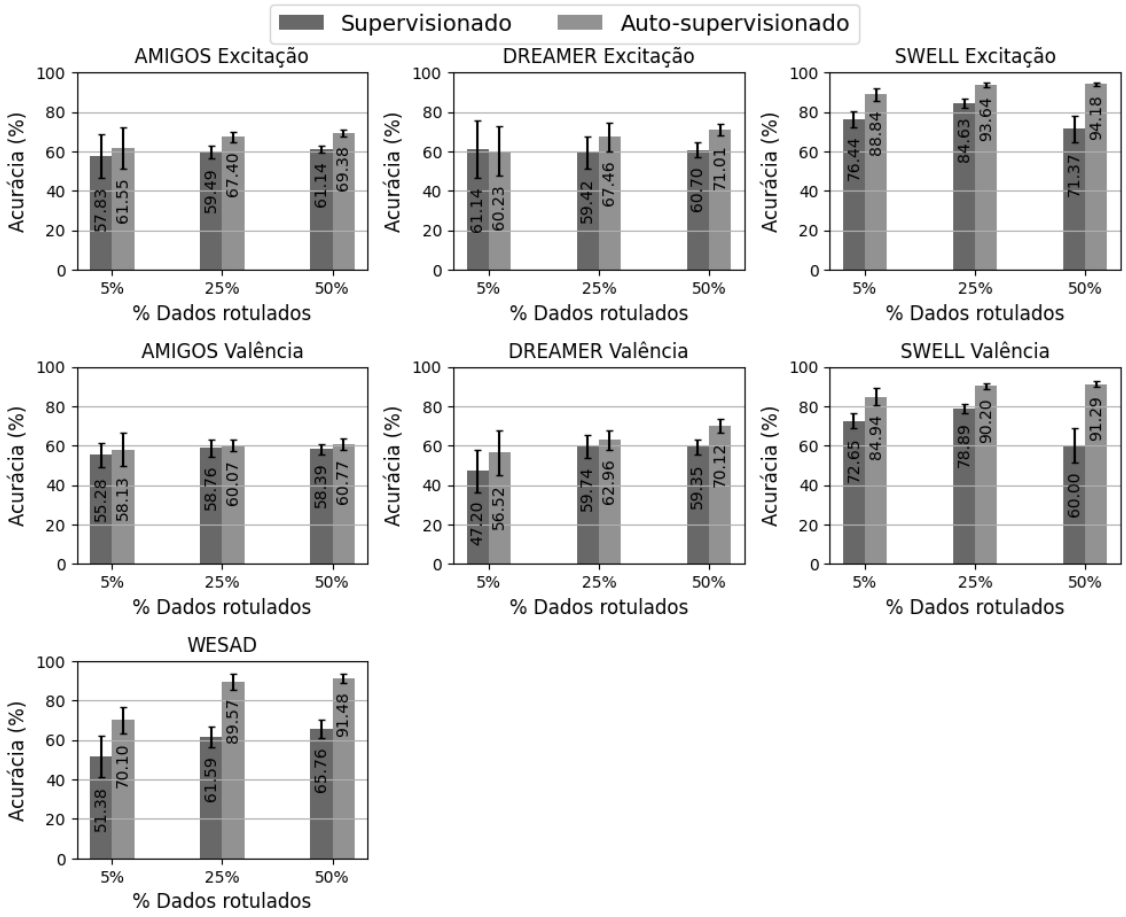

Figura 2 - Desempenho Comparativo do Aprendizado Supervisionado e Auto-supervisionado para Reconhecimento de Emoções em Sinais de ECG sob Diferentes Proporções de Dados Rotulado.

Em WESAD, o modelo auto-supervisionado (70,18%) supera o modelo supervisionado (51,38%) em 18,72 pontos percentuais (p.p.), demonstrando um ganho com apenas 5% de dados rotulados. Para 25% e 50% de dados rotulados, o modelo obteve um ganho de 27,98 p.p. e 25,72 p.p., respectivamente.

Para a base de dados SWELL (Excitação), o modelo auto-supervisionado com 5% de dados (88,84%) não apenas exibe um ganho de 12,4 p.p. sobre o modelo supervisionado (76,44%), como também ultrapassa o desempenho do modelo supervisionado treinado com 50% dos dados (84,63%). Esse achado é particularmente relevante, pois demonstra que a representação obtida por meio do pré-treinamento é capaz de mitigar a escassez de exemplos anotados, ao mesmo tempo em que reduz a suscetibilidade a erros de classificação decorrentes de variações individuais presentes nos sinais de ECG. Tendências análogas, com ganhos significativos, são observadas para as bases de dados DREAMER (e.g., +9,3 p.p. em Valência) e AMIGOS (e.g., +3,7 p.p. em Excitação).

A curva de aprendizado em função da quantidade de dados rotulados revela um padrão de retornos decrescentes para o modelo pré-treinado. Os maiores ganhos ocorreram ao aumentar a fração de rótulos de 5% para 25% (e.g., WESAD: +19,5 p.p.; SWELL Valência: +5,3 p.p.). Subsequentemente, o aumento para 50% resultou em

ganhos marginais, tipicamente inferiores a 2 p.p. para as bases de dados de maior desempenho. Este padrão sugere que 25% dos dados rotulados pode representar um ponto de equilíbrio ótimo entre o custo de anotação e a acurácia para a metodologia proposta.

6. Conclusão

Este estudo evidenciou que a combinação de tarefas auxiliares no pré-treinamento auto-supervisionado de sinais de ECG é uma estratégia eficaz para a extração de representações latentes robustas e generalizáveis. A aplicação de quatro a cinco transformações sintéticas mostrou-se suficiente para melhorar o desempenho em tarefas de classificação de valência, excitação e estresse, mesmo com quantidades reduzidas de dados rotulados. Notavelmente, o uso de apenas 25 % dos rótulos revelou-se um ponto de equilíbrio entre desempenho e custo de anotação, reforçando o potencial dessa abordagem em cenários com disponibilidade limitada de dados anotados. Embora o pré-treinamento multitarefa exija maior investimento computacional, esse custo é amplamente compensado pela expressiva redução na necessidade de anotações manuais, permitindo a obtenção de modelos eficientes com frações mínimas de rótulos.

Trabalhos futuros podem explorar a adoção de tarefas auxiliares adaptativas, a integração de esquemas de aprendizado contrastivo e a extensão da metodologia para sinais fisiológicos multimodais, como EEG e GSR, visando ampliar ainda mais a aplicabilidade e a eficácia dos modelos auto-supervisionados no reconhecimento de estados emocionais.

Agradecimentos

O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (AUXPE-CAPES-PROEX) – Código de Financiamento 001. Adicionalmente, este trabalho foi parcialmente financiado pela Fundação de Amparo à Pesquisa do Estado do Amazonas – FAPEAM – por meio do projeto PDPG-CAPES e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq.

Referências

Chowdhury, A., Rosenthal, J., Waring, J. and Umeton, R. (2021). Applying Self-Supervised Learning to Medicine: Review of the State of the Art and Medical Implementations. Informatics, vol. 8, no. 3, p. 59.

Fang, A., Pan, F., Yu, W., Yang, L. and He, P. (2024). ECG-Based Emotion Recognition Using Random Convolutional Kernel Method. Biomedical Signal Processing and Control, vol. 86, art. 105907.

Kan, H., Yu, J., Huang, J., Liu, Z. and Zhou, H. (2023). Self-supervised Group Meiosis Contrastive Learning for EEG-Based Emotion Recognition. Applied Intelligence, vol. 53, pp. 27207–27225.

Katsigiannis, S., Ramzan, N. (2017). DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. IEEE journal of biomedical and health informatics, v. 22, n. 1, p. 98-107.

Khattak, A. M., Asghar, M. Z., Ali, M. and Batool, U. (2021). An Efficient Deep Learning Technique for Facial Emotion Recognition. Multimedia Tools and Applications, vol. 81, no. 2, pp. 1649-1683.

Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M. A., Kraaij, W. (2014). The swell knowledge work dataset for stress and user modeling research. Proceedings of the 16th international conference on multimodal interaction. p. 291-298.

Kreibig, S. D. (2010). Autonomic Nervous System Activity in Emotion: A Review. Biological Psychology, vol. 84, no. 3, pp. 394-421.

Mehari, T., Strodthoff, N. (2021). Self-supervised representation learning from 12 lead ECG data. arXiv preprint arXiv:2103.12676.

Miranda-Correa, Juan A, Abadi, M. K., Sebe, N., Patras, I. (2018). Amigos: A dataset for affect, personality and mood research on individuals and groups. IEEE transactions on affective computing, v. 12, n. 2, p. 479-493.

Montesinos, V., Dell'Agnola, F., Arza, A., Aminifar, A., Atienza, D. (2019). Multi-Modal Acute Stress Recognition Using Off-the-Shelf Wearable Devices. Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), pp. 2196–2201.

Sarkar, P., Etemad, A. (2020). Self-Supervised ECG Representation Learning for Emotion Recognition. IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1541-1554, doi: 10.1109TAFFC.2020.3014842.

Schmidt, P., Reiss, A., Durichen R., Laerhoven K. (2018). Labelling affective states "in the wild": Practical guidelines and lessons learned. Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, pp. 654–659.

Vazquez-Rodriguez, J., Lefebvre, G., Cumin, J., Crowley, J.(2022). Transformer-Based Self-Supervised Learning for Emotion Recognition. 26th International Conference on Pattern Recognition (ICPR), pp. 2605-2612, doi: 10.1109/ICPR56361.2022.9956027.

Wan, B., Guo, J. (2020). Learning Immersion Assessment Model Based on Multidimensional Physiological Characteristics. Proceedings of 2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), pp. 87–90, doi: 10.1109/ICPICS50287.2020.9202208.

Wang, Z. and Wang, Y. (2025). Emotion Recognition Based on Multimodal Physiological Electrical Signals. Frontiers in Neuroscience, vol. 19, art. 1512799.

Wang, X., Ma, Y., Cammon J., Fang, F., Gao, Y., Zhang Y. (2023). Self-supervised EEG Emotion Recognition Models Based on CNN. IEEE Transactions on Neural System and Rehabilitation Engineering, vol 21, pp. 1952-1962, doi: 10.1109/TNSRE.2023.3263570.

Wang H., Chen T., Song L. (2024). Cascaded Self-supervised Learning for Subject-independent EEG-based Emotion Recognition. arXiv preprint arXiv:2403.04041.

Wu, Y., Daoudi M. (2023). Transformer-based Self-supervised Multimodal Representation Learning for Wearable Emotion Recognition. IEEE Transaction on Affective Computing, vol 15, no. 1, pp. 157-172, doi: 10.1109/TAFFC.2023.3263907.

Yang, W., Rifqi, M., Marsala C., Pinna A. (2018). Physiological-Based Emotion Detection and Recognition in a Video Game Context. Proceedings of the International Joint Conference on Neural Networks (IJCNN), pp. 1–8, doi: 10.1109/IJCNN.2018.8489125.

Zeng, Z., Pantic, M., Roisman, G. I. and Huang, T. S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 1, pp. 39-58.