CBIS'24 XX Congresso Brasileiro de Informática em Saúde
08/10 a 11/10 de 2024 - Belo Horizonte/MG - Brasil

Personalização de terapias pelo reconhecimento de emoções em biossinais

Therapies customization by emotion recognition in biosignals

Personalización de terapias mediante el reconocimiento de emociones en bioseñales

Maíra Araújo de Santana[1], Wellington Pinheiro dos Santos[2]

1 Doutora em Engenharia da Computação, Universidade de Pernambuco, Recife (PE), Brasil.
2 Professor do departamento de Engenharia Biomédica, Universidade Federal de Pernambuco, Recife (PE), Brasil.

Autor correspondente: Prof. Dr. Wellington Pinheiro dos Santos
E-mail: wellington.santos@ufpe.br

Resumo

Objetivo: Este estudo buscou desenvolver uma arquitetura de rede neural artificial híbrida para reconhecer estados de ânimo em biossinais de pessoas idosas, incluindo aquelas com demência leve a moderada, para ser utilizada como apoio à personalização de terapias. Método: O estudo empregou Transformada de Wavelet para converter sinais em imagens, que foram utilizadas como entrada para uma arquitetura híbrida formada por uma rede neural convolucional pré-treinada do tipo LeNet para extração de características e um algoritmo Random Forest com 450 árvores para classificação. O desempenho do algoritmo proposto foi avaliado em bases de dados pública de sinais de eletroencefalografia e voz, e posteriormente verificado em uma base de dados autoral de idosos com e sem demências. Resultados: A acurácia alcançada foi em torno de 71 a 73%. Conclusão: Essa tecnologia pode ser integrada em interfaces humano-máquina para personalizar terapias diversas, tais como a musicoterapia.

Descritores: Inteligência Artificial; Musicoterapia; Terapia Focada em Emoções

Abstract

Goal: This study aimed to develop a hybrid artificial neural network architecture to recognize mood states in biosignals from elderly individuals, including those with mild to moderate dementia, to support the personalization of therapies. Method: The study

employed Wavelet Transform to convert signals into images, used as input for a hybrid architecture comprising a pre-trained convolutional neural network of LeNet type for feature extraction, and a Random Forest algorithm with 450 trees for classification. The performance of the proposed algorithm was evaluated on publicly available databases of electroencephalography and voice signals, and subsequently validated on an in-house database of elderly individuals with and without dementia. Results: The achieved accuracy ranged from 71% to 73%. Conclusion: This technology can be integrated into human-machine interfaces to personalize various therapies, such as music therapy.

Keywords: Artificial Intelligence; Music Therapy; Emotion-Focused Therapy

Resumen

Objetivo: Este estudio tuvo como objetivo desarrollar una arquitectura híbrida de redes neuronales artificiales para reconocer estados de ánimo en biosensores de personas mayores, incluyendo aquellas con demencia leve a moderada, para apoyar la personalización de terapias. Método: El estudio utilizó la Transformada de Wavelet para convertir señales en imágenes, que fueron empleadas como entrada para una arquitectura híbrida compuesta por una red neuronal convolucional pre-entrenada del tipo LeNet para la extracción de características y un algoritmo de Bosque Aleatorio con 450 árboles para la clasificación. El rendimiento del algoritmo propuesto se evaluó en bases de datos públicas de señales de electroencefalografía y voz, y posteriormente se validó en una base de datos propia de personas mayores con y sin demencia. Resultados: Se alcanzó una precisión del 71% al 73%. Conclusión: Esta tecnología puede integrarse en interfaces humano-máquina para personalizar diversas terapias, como la musicoterapia.

Descriptores: Inteligencia Artificial; Musicoterapia; Terapia Centrada en la Emoción

Introdução

O envelhecimento populacional é um fenômeno global que apresenta desafios significativos na área da saúde pública. No caso específico do Brasil, a combinação de uma diminuição na taxa de natalidade e um aumento na expectativa de vida tem conduzido a um rápido envelhecimento demográfico [1,2]. Esta mudança demográfica trouxe consigo um aumento considerável na incidência de doenças relacionadas à

idade, como as síndromes demenciais, o que representa uma carga crescente para o sistema de saúde e a sociedade em geral [2].

A musicoterapia tem se consolidado como uma intervenção promissora no tratamento de doenças neurodegenerativas, como demências, a partir da desaceleração do progresso da doença [3,4]. Este tipo de terapia já é incorporado como uma das práticas integrativas do Sistema Único de Saúde (SUS) brasileiro e faz uso de ferramentas musicais para alcançar objetivos não musicais, tais como a recuperação de funcionalidades motoras e sociais [3]. A capacidade da música de ativar áreas cerebrais associadas à memória e ao senso de identidade a partir das emoções oferece uma via potencial para melhorar a qualidade de vida de idosos afetados por estas condições [3,4]. No entanto, o sucesso da musicoterapia depende em grande medida da capacidade do terapeuta para reconhecer e estimular emoções a partir de estímulos adequados e adaptados, o que pode ser particularmente desafiador na população idosa devido a diversos fatores crescentes com o avanço da idade, dentre eles a seletividade socioemocional, alterações no padrão vocal, o surgimento de marcas de expressão na face e a existência de patologias, fatores que afetam a manifestação das emoções [3].

Em resposta a esse desafio, tem-se explorado o uso de ferramentas de Inteligência Artificial (IA) para dar apoio ao reconhecimento de emoções. Muitos dos estudos investem na identificação de emoções em expressões faciais, no entanto, tem-se observado que combinar outros tipos de dados que se relacionem com a manifestação de emoções, como padrões de escrita e biossinais fisiológicos e de fala, podem fornecer informações complementares, por este motivo diversas abordagens vêm utilizando análises multimodais [5]. Devido à complexidade associada a análise de dados de naturezas distintas, tem-se priorizado a utilização de redes neurais artificiais profundas, com mais camadas de processamento, pois apresentam bons desempenhos na integração e interpretação de informações de diferentes fontes [5].

Embora o reconhecimento de emoções seja uma área em ascensão nos últimos anos [5], poucos trabalhos atuais que se propõem a realizar esta tarefa alcançam desempenho de classificação acima de 85% [6,7]. Mais ainda, mesmo soluções já comercializadas apresentam acurácia na faixa de 48% a 62% [7].

Neste contexto surge a seguinte pergunta de pesquisa: como a Computação Afetiva pode auxiliar nos processos de geração de música customizada para tratamento de pacientes com Doença de Alzheimer ou outras demências em estágio inicial? A hipótese é que o reconhecimento das emoções por meio da análise de sinais diversos como os eletroencefalográficos e da fala pode contribuir para a construção de interfaces musicais humano-máquina que permitam auxiliar musicoterapeutas na determinação de gêneros musicais, estilos, ritmos e abordagens personalizadas no contexto do tratamento de demências.

O presente estudo concentra-se no desenvolvimento de uma arquitetura de rede neural híbrida para reconhecer estados de humor em sinais Eletroencefalográficos (EEG) e de voz de indivíduos idosos. A arquitetura híbrida foi adotada pois combina algoritmo de aprendizagem profunda, bom para extrair características relevantes de dados de naturezas distintas, e uma camada de classificação com método raso, que exige menos recursos computacionais e, portanto, favorece a implementação em diferentes configurações de máquinas. Essa arquitetura será utilizada como núcleo do módulo de reconhecimento de emoções por biossinais a ser incorporado em um sistema de apoio à personalização das intervenções terapêuticas, tornando-as mais adequadas aos gostos e desejos dos pacientes.

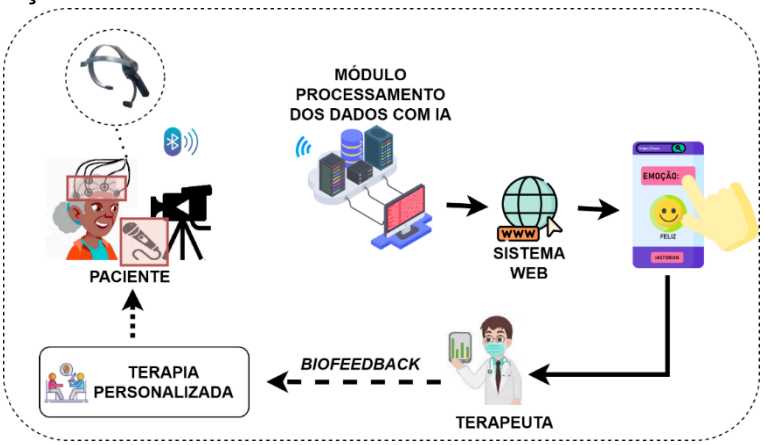
Este artigo está organizado como segue: após esta introdução é apresentada a seção de métodos, onde são descritas as etapas adotadas para o desenvolvimento do estudo. Na seção seguinte são apresentados e discutidos os resultados obtidos, seguidos da seção de conclusão, que encerra o artigo apresentando as principais contribuições, limitações e oportunidades advindas da pesquisa realizada.

Métodos

Esse trabalho se dedica a construir uma base de dados e validar uma abordagem computacional a ser utilizada no desenvolvimento de uma interface humano-máquina para suporte à personalização de musicoterapia para idosos acometidos por demências a partir do *biofeedback* de emoções. A abordagem computacional proposta servirá como núcleo do módulo para reconhecimento automático de emoções expressas em dados humanos de diversas modalidades. O sistema em construção está esquematizado na Figura 1, onde os dados são capturados do paciente e processados pelo módulo de IA, que irá fornecer ao

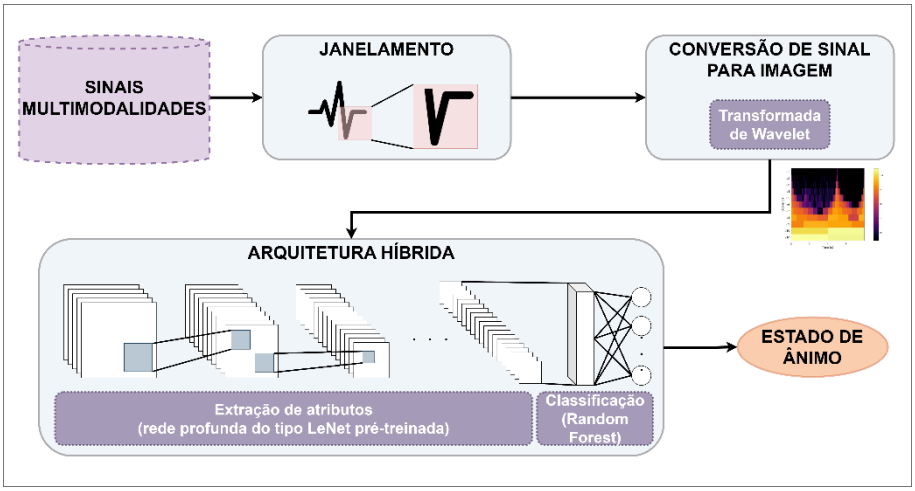
terapeuta a informação do seu estado emocional, a qual será utilizada para refinar a abordagem terapêutica que é entregue ao paciente. Nesse sentido, os próprios dados emocionais do indivíduo são utilizados para melhorar seu estado de ânimo no contexto terapêutico, funcionando como uma retroalimentação (*feedback*).

Figura 1 – Proposta do sistema para suporte terapêutico personalizado baseado em *biofeedback* de emoções



Nesse estudo em particular foram utilizados apenas dados que são adquiridos na forma de sinais (EEG e voz), destacados em vermelho na Figura 1. Foi, então, proposta uma arquitetura híbrida para atuar no reconhecimento de emoções a partir desses dados multimodais. Conforme ilustrado na Figura 2, no módulo proposto os dados de entrada passam por uma etapa de janelamento para possibilitar uma análise mais local do dado, com janela de 5s com sobreposição de 1s, escolhida empiricamente e cuja ordem de grandeza está alinhada com a literatura para soluções que envolvem o reconhecimento de emoções. Em seguida, cada janela do sinal é submetida a um processo de conversão para imagem a partir da transformada rápida de Wavelet, aplicando o algoritmo de Mallat com 10 bancos de filtros da família Daubechies 8 (db8), a qual é capaz de fornecer informações nos domínios de tempo e de frequência de sinais originalmente descritos no domínio do tempo [8]. Os canais dos sinais foram concatenados antes de aplicar a transformada, de maneira que foi gerada 1 imagem para cada janela do sinal, todas salvas em formato .jpeg. Essa etapa foi implementada em Octave, versão 5.1.0 [9]. Após essas etapas de pré-processamento, para classificar o estado emocional é proposta uma arquitetura híbrida que combina Rede Neural Convolucional (CNN, do inglês *Convolutional Neural Network*) pré-treinada para a extração de atributos e algoritmo supervisionado clássico para classificação.

Figura 2 – Módulo de reconhecimento de emoções baseado em uma arquitetura de rede neural artificial híbrida



Na abordagem proposta a CNN adotada foi uma LeNet, composta por 7 camadas, sendo 2 convolucionais, 2 de *pooling* de média e 3 totalmente conectadas, pré-treinada com a base de dados MNIST com a qual a LeNet apresenta altos desempenhos de classificação de imagem. Esta configuração de rede extrai 500 atributos de cada imagem de entrada. Na camada de classificação da arquitetura híbrida utilizou-se uma Random Forest com 450 árvores, algoritmo selecionado por seus altos desempenhos em problemas multiclasse e fácil explicabilidade, visto que consiste em um comitê de árvores de decisão, métodos hierárquicos de fácil interpretação e visualização gráfica.

Para avaliar o desempenho da arquitetura experimentada cada conjunto de dados utilizado foi dividido em duas subamostras aleatórias sem substituição e mantendo a distribuição de classes. Uma das partes (70%) foi utilizada como conjunto de treinamento-validação e a outra (30%) consistiu no conjunto de teste. O primeiro é utilizado para avaliar o desempenho de classificação durante a etapa de treinamento do modelo. Neste estudo, utilizou-se o método de validação cruzada k−*fold* para esta etapa, com k=10, no qual o conjunto foi subdividido em 10 partes que são iterativamente exploradas de maneira que 9 são utilizados para treinamento e 1 para validação. A adoção desse método de treinamento contribui para minimizar ocorrências de sobreajuste, ou *overfitting*, do classificador. Além disso, com o intuito de verificar de maneira mais detalhada o comportamento estatístico do modelo experimentado, foram executados 30 treinamentos completos. Já que durante o treino foi adotada a abordagem de validação cruzada essa metodologia adiciona uma etapa

de teste no próprio treinamento, cuja repetição permite uma análise mais aprofundada do desempenho, aumentando a confiabilidade dos resultados obtidos.

Após o treinamento, o modelo é utilizado para estimar as classes do conjunto de teste. Essa etapa é fundamental para verificar a capacidade de generalização do desempenho de classificação para novos dados, ou seja, dados que não participaram do treinamento do modelo e são, portanto, desconhecidos para o argoritmo. Esses experimentos de treinamento e teste foram realizados através das implementações em linguagem Java no *software* Weka (Waikato Environment for Knowledge Analysis), versão 3.8 [10].

A eficácia da classificação dos estados emocionais pela arquitetura proposta foi mensurada a partir das métricas: acurácia, índice kappa, sensibilidade, especificidade e área sob a curva ROC (AUC), cujas descrições matemáticas estão no Quadro 1, onde *VP* se refere ao número de verdadeiros positivos, *VN* de verdadeiros negativos, *FP* de falsos positivos e *FN* de falsos negativos, ρ_o é o valor observado e ρ_e é o valor esperado. A acurácia tem valor máximo 100%, enquanto a sensibilidade, especificidade e área sob a curva ROC possuem valores máximos iguais a 1. O índice kappa, por sua vez, fornece indícios sobre a correlação estatística entre os resultados obtidos e o esperado, esse índice pode assumir valores no intervalo [−1; 1], onde valores próximos de 1 indicam boa correlação entre os resultados esperados e obtidos.

Quadro 1 – Expressões matemáticas das métricas utilizadas para avaliar o desempenho da arquitetura de classificação proposta

Métricas
$Acurácia = \frac{VP+VN}{VP+VN+FP+FN}$
$Sensibilidade = \frac{VP}{VP+FN}$
$Especificidade = \frac{VN}{VN+FP}$
$AUC = \int VP\, d(FP)$
$Kappa = \kappa = \frac{\rho_o - \rho_e}{1-\rho_e}$, $onde\ \rho_e = \frac{(VP+FP)(VP+FN)+(FN+VN)(FP+VN)}{(VP+FP+FN+VN)^2}$

A validação da arquitetura foi inicialmente realizada a partir de experimentos de provas de conceito com bases de dados públicas contendo os sinais de interesse. Neste momento verificou-se a inexistência de bases de dados disponíveis publicamente envolvendo a expressão de emoções por pessoas idosas acometidas ou

não por processos demenciais. Portanto, a etapa de prova de conceito se deu paralelamente à elaboração do projeto para submissão aos comitês de ética responsáveis pela aprovação da coleta de dados para a construção de uma base de dados autoral com o público de interesse. Após a apreciação ética e aprovação da pesquisa com seres humanos, foram executadas as coletas.

Para as provas de conceito, avaliando a arquitetura híbrida proposta, foram utilizadas as bases de dados Multimodal Database for Affect Recognition and Implicit Tagging (MANOHB-HCI) [11], contendo dados fisiológicos centrais e periféricos associados a categorias de emoções, e a base Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [12], para reconhecimento de emoções em sinais de voz.

Da base MANOHB-HCI foram utilizados os dados provenientes de 38 canais, sendo 32 de EEG, 1 de Resposta Galvânica da Pele (GSR, do inglês *Galvanic Skin Response*), 3 de Eletrocardiograma (ECG), 1 para amplitude da respiração e 1 para temperatura cutânea, capturados de maneira sincronizada em 24 voluntários dos sexos feminino e masculino, com frequência de amostragem de 256Hz. Para a vinculação com estados emocionais os autores utilizaram estímulos visuais, na forma de imagens e vídeos, após os quais a pessoa realizou a autodeclaração do estado afetivo através de uma escala de Prazer-Ativação-Dominância. Por fim, com base nessas pontuações os autores rotularam os dados em 6 classes de emoções, culminando em 84 sinais de estado neutro, 65 de diversão, 50 de alegria, 36 de nojo, 27 de raiva e 14 de tristeza.

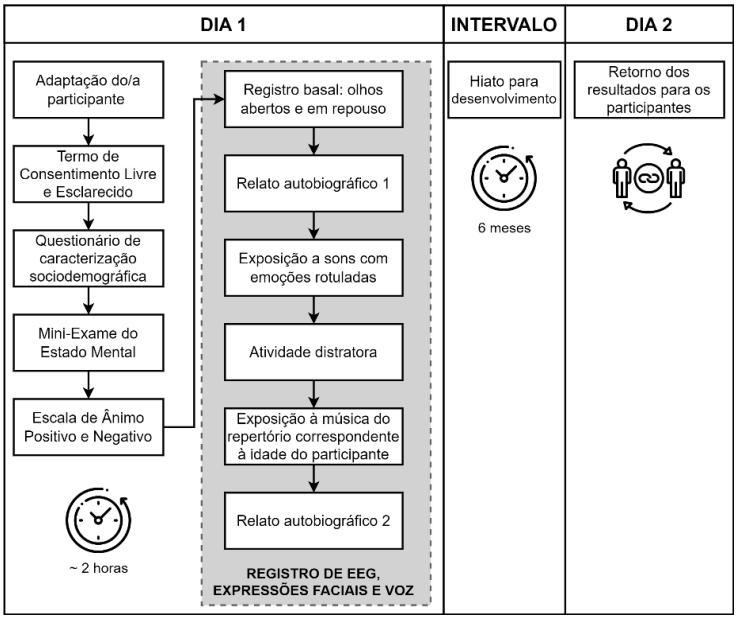
Já a base RAVDESS é composta por vozes de 24 atores profissionais, igualmente divididos nos sexos feminino e masculino, em idioma inglês com sotaque norte-americano. Para a aquisição dos sinais cada indivíduo se apresentou falando duas frases de modo a representar 8 emoções: neutro, calma, felicidade, tristeza, raiva, medo, surpresa e nojo, resultando em 1440 arquivos de voz em formato .WAV com frequência de amostragem de 48kHz, igualmente distribuídos nas classes.

Embora as bases de dados públicas utilizadas sejam de boa qualidade, amplamente conhecidas e divulgadas na literatura, além de terem sido relevantes para demonstrar a viabilidade da arquitetura proposta, os dados das modalidades distintas são provenientes de indivíduos diferentes em cenários experimentais diversos.

Ademais, os indivíduos envolvidos não são idosos e não há informação sobre a existência de patologias que podem interferir na expressão das emoções.

Neste sentido, foi proposto o protocolo de coleta de dados relacionados a aspectos emocionais de pessoas idosas com e sem demências. Conforme esquematizado na Figura 3, o contato com os indivíduos participantes se deu em 2 momentos, espaçados por aproximadamente 6 meses. No primeiro foram realizados todos os procedimentos experimentais de coleta de dados, dentre eles dados sociodemográficos, teste cognitivo mini-exame do estado mental (MEEM), além dos dados de EEG, voz e expressão facial durante as etapas listadas na porção destacada em cinza da Figura 3. O segundo momento consiste no retorno sobre o que foi realizado com os dados adquiridos. O protocolo foi aprovado pelos Comitês de Ética em Pesquisa (CEP) das instituições vinculadas ao processo de coleta: Universidade de Pernambuco (UPE), Universidade Federal de Pernambuco (UFPE) e Hospital das Clínicas da UFPE (HCPE), cujos registros são, respectivamente, CAAE/UPE/42381720.6.0000.5207; CAAE/UFPE/42381720.6.3001.5208; e CAAE/HCPE/42381720.6.3002.8807.

Figura 3 – Protocolo de coleta de dados



O sinal eletroencefalográfico foi adquirido por meio de 23 eletrodos de prata, dispostos com o auxílio de uma touca de neoprene e uso de gel condutor para reduzir a impedância pele-eletrodo. O equipamento utilizado foi o módulo de EEG do polissonógrafo Neurovirtual, modelo Brain Wave III, com taxa de amostragem de

300Hz. Para a aquisição os eletrodos foram posicionados de acordo com o sistema internacional 10-20, com eletrodo de referência posicionado na testa do participante. Os registros de expressões faciais e voz são simultaneamente realizados com uso de uma câmera fotográfica DSRL Canon EOS Rebel T7.

A coleta de dados se deu entre os meses de dezembro de 2022 e julho de 2023. A base de dados oriunda desse processo conta com dados de EEG, voz, expressão facial, demográficos e de diagnóstico de 39 participantes, sendo 21 do grupo controle (CTR) e 18 do grupo não-controle (TCM). Para a divisão por grupos o teste cognitivo considerado foi o MEEM com ponto de corte caso/não caso de 19/20 para indivíduos sem escolaridade e 23/24 para indivíduos alfabetizados, de acordo com critérios estabelecidos para população brasileira por Almeida (1998) [14]. Na Tabela 1 são apresentadas as características sociodemográficas e cognitivas associadas aos dados que constituem a base.

Tabela 1 – Descrição de dados sociodemográficos e cognitivos dos participantes

Variáveis	Frequência	Média	Desvio-padrão
Gênero biológico			
Feminino	29 (74,36%)	-	-
Masculino	10 (25,64%)		
Idade	-	72	8,60
***Status* conjugal**			
Casado	18 (46,15%)		
Divorciado	5 (12,82%)	-	-
Viúvo	5 (12,82%)		
Solteiro	11 (28,20%)		
Escolaridade			
Nenhuma	3 (7,69%)		
Fundamental I incompleto	4 (10,26%)		
Fundamental I completo	5 (12,82%)		
Fundamental II incompleto	2 (5,13%)		
Fundamental II completo	5 (12,82%)	-	-
Médio incompleto	0 (0,00%)		
Médio completo	8 (20,51%)		
Superior incompleto	1 (2,56%)		
Superior completo	10 (25,64%)		
Não soube informar	1 (2,56%)		
Trabalho remunerado			
Sim	7 (17,95%)	-	-
Não	32 (82,05%)		
Experiência musical			
Sim	11 (28,20%)		
Não	27 (69,23%)	-	-
Não soube informar	1 (2,56%)		
Pontuação no MEEM	-	21,64	7,34

Após a coleta, os registros coletados foram organizados para experimentação, com foco em dados multimodais. Os trechos selecionados para análise neste estudo foram os relatos autobiográficos 1 e 2 (RAB1 e RAB2), nos quais sinais de EEG e voz são adquiridos simultaneamente. Os estados de ânimo foram associados às respostas dos indivíduos em escalas aplicadas antes de cada relato. Para rotular o ânimo no

RAB1, usou-se uma escala de ânimo positivo-negativo contendo 6 emoções de polaridades positivas e 8 negativas [3], com uma escala de frequência de ocorrência de 4 pontos (0-3) para cada uma. Assim, para cada emoção da escala foi registrada a nota de frequência atribuída pelo participante (np1-np6, para as emoções de polaridade positiva; e nn1-nn8, para as de polaridade negativa). No RAB2, usou-se uma escala de polaridade discreta simples em 5 pontos (1-5). Instrumentos distintos foram utilizados para favorecer a aplicação do protocolo, com vistas a não levar os participantes a exaustão, especialmente os acometidos por processos demenciais, por isso os relatos foram analisados separadamente. Os algoritmos da Figura 4 foram empregados para estabelecer os rótulos em cada trecho.

Figura 4 – Algoritmos adotados para definir as classes associadas aos trechos (a) RAB1 e (b) RAB2

```
início
        // notas de frequência atribuídas pelo participante para os ânimos positivos
        leia (np1, np2, np3, np4, np5, np6)
        // notas de frequência atribuídas pelo participante para os ânimos negativos
        leia (nn1, nn2, nn3, nn4, nn5, nn6, nn7, nn8)

        // cálculo das médias
        mediaPositivos = (np1 + np2 + np3 + np4 + np5 + np6) / 6
        mediaNegativos = (nn1 + nn2 + nn3 + nn4 + nn5 + nn6 + nn7 + nn8) / 8

        // definição da classe de emoção do RAB1 (rótulo do estado afetivo)
        se (mediaPositivos > mediaNegativos) e (mediaPositivos > 1) então
                escreva ("positivo")
        senão se (mediaNegativos > mediaPositivos) e (mediaNegativos > 1) então
                escreva ("negativo")
        senão
                escreva ("neutro")
        fim-se
fim
```

(a)

```
início
        // nota de polaridade atribuída pelo participante
        leia (nPolaridade)

        // definição da classe de emoção do RAB2 (rótulo do estado afetivo)
        se (nPolaridade > 3) então
                escreva ("positivo")
        senão se (nPolaridade < 3) então
                escreva ("negativo")
        senão
                escreva ("neutro")
        fim-se
fim
```

(b)

Após a rotulação por estado de ânimo, avaliou-se o desempenho da arquitetura proposta (Figura 2) na base de dados multimodal composta por sinais de EEG e de voz coletados simultaneamente, além de dados categóricos de idade, gênero (1 para feminino e 2 para masculino) e diagnóstico (0 para CTR e 1 para TCM). Assim, foram realizados experimentos computacionais associados ao reconhecimento automático de estados de ânimo das pessoas idosas que participaram da coleta utilizando a arquitetura validada anteriormente em bases de dados públicas.

Resultados e Discussão

Nesta seção são apresentados os resultados obtidos pela arquitetura híbrida avaliada neste estudo para a representação de sinais convertidos em imagens pela transformada de Wavelet e classificação de estados de ânimo nesses dados.

Seguindo a sequência descrita na seção anterior, inicialmente foram obtidos os desempenhos de classificação durante a prova de conceito realizada com bases de

dados públicas contendo sinais fisiológicos e de voz. A Tabela 2 apresenta esses resultados por modalidade de sinal. Para os sinais fisiológicos, no treino foi obtida acurácia média de 80,62%, kappa de 0,77, além de sensibilidade (0,99), especificidade (0,97) e AUC (1,00) próximos dos valores máximos, todos com desvio-padrão baixos, demonstrando boa repetibilidade dos resultados. Esses altos desempenhos foram repetidos no teste, com um aumento na acurácia, chegando a 99,22%. O desempenho para os sinais de voz foi levemente inferior, com diferença mais expressiva na acurácia, onde foi alcançado valor médio 76,99% no treino e 82,22% no teste, atribuída principalmente ao aumento na quantidade de classes associadas, que era de 6 para os sinais fisiológicos e de 8 classes nos sinais de voz. Os valores de kappa (0,74 ± 0,03), sensibilidade (0,93 ± 0,06), especificidade (0,96 ± 0,02) e AUC (0,99 ± 0,01) no treino foram semelhantes aos de teste (0,79; 0,82; 0,97 e 0,98, respectivamente), ficando também na faixa dos resultados obtidos para os sinais fisiológicos nessas métricas.

Tabela 2 – Desempenho da arquitetura na etapa de prova de conceito

Modalidade	Conjunto	Acurácia	Kappa	Sens.	Esp.	AUC
Sinais fisiológicos	treino	80,62 ± 1,62	0,77 ± 0,02	0,99 ± 0,01	0,97 ± 0,01	1,00 ± 0,00
	teste	99,22	0,99	0,99	1,00	0,99
Sinais de voz	treino	76,99 ± 2,97	0,74 ± 0,03	0,93 ± 0,06	0,96 ± 0,02	0,99 ± 0,01
	teste	82,22	0,79	0,82	0,97	0,98

Conforme previamente mencionado, os resultados com bases de dados públicas se mostraram promissores e foram essenciais para a avaliar a viabilidade e desempenho geral da arquitetura proposta frente a dados das modalidades de interesse, no entanto, esses dados não são de um mesmo indivíduo nem de pessoas idosas, portanto houve a avaliação do desempenho do modelo nos dados da base construída. A Tabela 3 exibe esses resultados, com resultados semelhantes para ambos os trechos, o que dá indícios de consistência na performance do método proposto. No treino com o RAB1 a acurácia média foi de 73,14%, kappa de 0,60, considerado como moderado [13], e valores médios substanciais para sensibilidade (0,74), especificidade (0,89) e AUC (0,90). Houve uma queda nesse desempenho no teste em termos de acurácia, kappa e sensibilidade (respectivos valores de 63,97%, 0,30 e 0,64), contudo foi percebido aumento na especificidade (0,96) e na AUC (0,98). Nos dados do RAB2, no treino os valores médios de acurácia, kappa, sensibilidade e AUC foram de 71,58%, 0,57, 0,69, 0,80 e 0,85, respectivamente. O comportamento foi análogo ao obtido para o RAB1 no conjunto de teste. Em ambos os casos, com

exceção do kappa, que apontou pouca concordância entre os resultados esperado e obtido, as demais métricas se mantiveram dentro ou próximas da variabilidade encontrada durante o treinamento, se for considerado os desvios padrão calculados.

Tabela 3 – Desempenho da arquitetura nos dados de pessoas idosas

Trecho	Conjunto	Acurácia	Kappa	Sens.	Esp.	AUC
RAB1	treino	73,14 ± 8,05	0,60 ± 0,12	0,74 ± 0,14	0,89 ± 0,08	0,90 ± 0,06
	teste	63,97	0,30	0,64	0,96	0,98
RAB2	treino	71,58 ± 11,54	0,57 ± 0,17	0,69 ± 0,22	0,80 ± 0,13	0,85 ± 0,10
	teste	60,52	0,17	0,60	0,86	0,82

Embora tenha sido observado uma diminuição geral nos valores médios das métricas, quando comparado com os desempenhos obtidos durante a prova de conceito, é importante destacar que, especialmente para a acurácia e o kappa, se forem considerados os desvios-padrão os valores se assemelham ao desempenho anteriormente alcançado.

Os desempenhos na classificação de estados de ânimo aqui apresentados podem ser considerados realistas e contextualizados no estado-da-arte, dada a escassez de estudos consolidados em reconhecimento de emoções em idosos utilizando dados multimodais. As soluções comerciais para reconhecimento de emoções pela face, sem especificação de faixa etária, apresentam acurácia entre 48% e 62% [7], destacando a relevância dos resultados alcançados. A falta de investigações direcionadas para dados de idosos, especialmente os com demência, limita a comparação com outros estudos. Ainda assim, os desempenhos de classificação dos estados afetivos obtidos são inéditos, positivos e promissores no campo do reconhecimento de emoções em dados multimodais de pessoas idosas.

Conclusão

O estudo avaliou uma arquitetura de rede neural profunda para reconhecimento de emoções em sinais de voz e eletroencefalográficos, visando personalizar a musicoterapia para demências. Utilizou-se Transformada de Wavelet para converter sinais em imagens, processadas por uma arquitetura híbrida com uma rede neural convolucional pré-treinada do tipo LeNet para extração de atributos, seguida por um algoritmo Random Forest com 450 árvores para classificação emocional. A partir da literatura acessada observou-se uma escassez de bases de dados públicas específicas para idosos com patologias. Provas de conceito com dados públicos mostraram taxas de acerto em torno de 80% para sinais fisiológicos e 76% para voz,

mas analisados separadamente. Para superar essas limitações, uma nova base de dados foi construída, envolvendo EEG, voz, expressão facial e dados demográficos de idosos, mostrando acurácias de aproximadamente 71% e 73% na identificação de estados emocionais.

Os resultados dos experimentos destacam que a arquitetura híbrida proposta teve sucesso no reconhecimento de emoções, mesmo diante de desafios como desbalanceamentos e múltiplas classes. A análise emocional é complexa devido à sua natureza subjetiva, o que dificulta a rotulação e o treinamento de algoritmos para identificação automática. No entanto, a ampliação dos estudos na área pode mitigar esses desafios, reduzindo a dependência da autodeclaração de emoções. A base de dados criada preenche uma lacuna na pesquisa de emoções em idosos, combinando diversas modalidades de dados. Além disso, essa base possibilita investigações computacionais imediatas, como métodos evolucionários para avaliar a relevância de cada modalidade de dado e estudos sobre as regiões cerebrais associadas a manifestação emocional nos dados de EEG coletados. Também pode ser explorada a influência de estímulos sonoros e musicais nos sinais envolvidos, comparando os estados emocionais antes e após tais estímulos. Este trabalho não apenas promove o campo das Interfaces Musicais Humano-Máquina como ferramenta terapêutica, mas também destaca a possibilidade de utilização de ferramentas de IA aplicadas ao reconhecimento de emoções de maneira responsável para otimizar os processos de musicoterapia e outras abordagens terapêuticas, visando melhorar a qualidade de vida de pessoas com distúrbios cognitivos, motores e comportamentais.

Agradecimentos

À Fundação de Amparo à Ciência e Tecnologia de Pernambuco (FACEPE) pelo financiamento parcial da pesquisa. Aos colaboradores do Hospital das Clínicas da UFPE, em especial ao Dr. Breno Barbosa, do Ambulatório de Neurologia. A cada pessoa que se voluntariou para a coleta de dados.

Referências

1. Vivas EN, Rocha SF. The Brazilian population aging and its contemporary challenges. MOJ Gerontol Ger. 2020;5(5):165-8.

2. Bloom DE, Canning D, Lubet A. Global population aging: Facts, challenges, solutions & perspectives. Daedalus. 2015;144(2):80-92.
3. Silva-Júnior JD. Memórias Autobiográficas e Música em Idosos. Campinas: Editora Alínea. 2018.
4. Peixoto CT da S. Saúde mental: um enfoque voltado à prevenção da demência de alzheimer. JHMReview [Internet]. 2021;7(3). Disponível em: https://ijhmreview.org/ijhmreview/article/view/276
5. Santana MA, Lima CL, Torcate AS, Fonseca FS, Santos WP. Affective computing in the context of music therapy: a systematic review. RSD [Internet]. 2021;10(15): e392101522844. Disponível em: https://rsdjournal.org/index.php/rsd/article/view/22844
6. Veltmeijer EA, Gerritsen C, Hindriks KV. Automatic emotion recognition for groups: a review. IEEE Transactions on Affective Computing. 2021;14(1):89-107.
7. Dupré D, Krumhuber EG, Küster D, McKeown GJ. A performance comparison of eight commercially available automatic classifiers for facial affect recognition. Plos one. 2020;15(4):e0231968.
8. Mallat SG. Multifrequency channel decompositions of images and wavelet models. IEEE Transactions on Acoustics, speech, and signal processing. 1989;37(12):2091-110.
9. Eaton JW, Bateman D, Hauberg S. GNU Octave version 3.0. 1 manual: a high-level interactive language for numerical computations. Whales: Network Theory Ltd. 2007.
10. Witten IH, Frank E, Hall MA. Data Mining: Practical machine learning tools and techniques. Burlington: Morgan Kaufmann Publishers. 2011.
11. Soleymani M, Lichtenauer J, Pun T, Pantic M. A multimodal database for affect recognition and implicit tagging. IEEE transactions on affective computing. 2011;3(1):42-55.
12. Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one. 2018;13(5):e0196391.
13. McHugh ML. Interrater reliability: the kappa statistic. Biochemia medica. 2012;22(3):276-82.
14. Almeida OP. Mini exame dos estado mental e o diagnóstico de demência no Brasil. Arquivos de Neuro-psiquiatria. 1998;56:605-12.