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

Reconhecimento de Emoções como ferramenta de apoio às terapias personalizadas

Emotion Recognition as a tool to support personalized therapies

Reconocimiento de Emociones como herramienta de apoyo para terapias personalizadas

Arianne Sarmento Torcate[1], Maíra Araújo de Santana[2], Juliana Carneiro Gomes[2], Ana Clara Gomes da Silva[3], Wellington Pinheiro dos Santos[4]

1 Mestra em Engenharia da Computação, Universidade de Pernambuco, Recife (PE), Brasil
2 Doutora em Engenharia da Computação, Universidade de Pernambuco, Recife (PE), Brasil
3 Mestra em Engenharia Biomédica, Universidade Federal de Pernambuco, Recife (PE), Brasil
4 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

Contexto: Em contextos terapêuticos, sistemas de reconhecimento de emoções podem ser uma ferramenta valiosa para pacientes com dificuldades de expressão emocional. Objetivo: Portanto, este trabalho tem como objetivo apresentar um comparativo entre arquiteturas híbridas para realizar reconhecimento de emoções em expressões faciais. Método: As arquiteturas propostas foram treinadas-validadas com a base de dados FER2013 e se baseiam na decomposição de *Wavelet* e em *Transfer Learning*. Diferentes configurações de pré-processamento dos dados também foram exploradas. Resultado: Como resultado, a arquitetura composta por uma VGG16 e um Random Forest, obteve 74,52% de acurácia no treinamento e 84,72% no teste, apenas com 27% dos atributos da VGG16. A arquitetura de DWNN, com 4 camadas e Random Forest, obteve 70,77% de acurácia no treinamento e 81,21% no teste, utilizando 34% dos atributos. Conclusão: A melhor arquitetura irá compor um sistema de reconhecimento de emoções para personalização de terapias.

Palavras-chave: Arquiteturas Híbridas; Reconhecimento de Emoções em Expressões Faciais; Terapias Personalizadas.

Abstract

Background: In therapeutic contexts, emotion recognition systems can be a valuable tool for patients with emotional expression difficulties. Objective: Therefore, this work aims to present a comparison between hybrid architectures to perform emotion recognition in facial expressions. Method: The proposed architectures were trained-validated with the FER2013 database and are based on Wavelet decomposition and Transfer Learning. Different data preprocessing configurations were also explored. Result: As a result, the architecture composed of a VGG16 and a Random Forest obtained 74.52% accuracy in training and 84.72% in testing, with only 27% of the attributes of VGG16. The DWNN architecture, with 4 layers and Random Forest, achieved 70.77% accuracy in training and 81.21% in testing, using 34% of the attributes. Conclusion: The best architecture will compose an emotion recognition system for personalizing therapies.

Keywords: Hybrid Architectures; Recognition of Emotions in Facial Expressions; Personalized Therapies.

Resumen

Antecedentes: En contextos terapéuticos, sistemas de reconocimiento de emociones pueden ser una herramienta valiosa para pacientes con dificultades de expresión emocional. Objetivo: Este trabajo tiene como objetivo presentar una comparación entre arquitecturas híbridas para realizar reconocimiento de emociones en expresiones faciales. Método: Las arquitecturas propuestas fueron entrenadas-validadas con la base de datos FER2013 y se basan en descomposición Wavelet y Transfer Learning. También se exploraron diferentes configuraciones de preprocesamiento de datos. Resultado: Como resultado, la arquitectura compuesta por un VGG16 y Random Forest obtuvo un 74,52% de precisión en el entrenamiento y un 84,72% en las pruebas, con 27% de los atributos de VGG16. La arquitectura DWNN, con 4 capas y Random Forest, logró un 70,77% de precisión en el entrenamiento y un 81,21% en las pruebas, utilizando 34% de atributos. Conclusión: La mejor arquitectura compondrá un sistema de reconocimiento de emociones para terapias personalizadas.

Descriptores: Arquitecturas Híbridas; Reconocimiento de Emociones en Expresiones Faciales; Terapias personalizadas.

Introdução

Pode-se definir as emoções como um processo abrangente que envolve diversos elementos distintos e inter-relacionados, como sentimentos (componentes subjetivos), mudanças corporais (respostas fisiológicas) e aspectos comportamentais (como expressões e ações) [1]. Todos esses elementos que compõem a resposta emocional influenciam diretamente na comunicação direta e indireta de indivíduos, além do desenvolvimento pessoal e social [2]. Sabendo que as emoções estão presentes no cotidiano e do aumento da interação humano-computador nos últimos anos, é importante considerar os sentimentos humanos no desenvolvimento dos aparelhos tecnológicos [3].

É nesse cenário que surge a área da Computação Afetiva, com foco em reconhecimento de emoções, buscando investigar como os computadores podem reconhecer, interpretar as emoções e gerar cada vez mais respostas afetivas [4]. Mas, para que um sistema reconheça automaticamente emoções, dados de humanos devem ser utilizados. Estes dados podem se originar de fontes diversas, como Resposta Galvânica da Pele, Expressões Faciais, Voz, Eletrocardiogramas (ECG), sinais Eletroencefalográficos (EEG), entre outros [5]. Segundo Gong *et al.* (2024) [6], o reconhecimento de emoções através das expressões faciais (foco dessa pesquisa) está se tornando cada vez mais necessário devido às demandas atuais na área da saúde, onde o monitoramento e a intervenção a partir das emoções refletidas na face podem ter implicações positivas em tratamentos, sejam eles físicos ou psicológicos.

Sabendo disso, nos últimos anos a medicina personalizada tem se destacado devido a capacidade de adaptar o tratamento de indivíduos de acordo com suas características individuais, proporcionando abordagens cada vez mais precisas [7]. Fazendo um recorte específico para as práticas terapêuticas, os sistemas de reconhecimento de emoções através das expressões faciais podem se tornar uma ferramenta valiosa, principalmente para determinados públicos que possuem dificuldade de expressar as emoções, como idosos (acometidos por processos demenciais ou não) e crianças com transtorno do espectro autista [8, 9, 10]. É válido mencionar que a capacidade de expressar e reconhecer as emoções por meio da face é um estágio fundamental da comunicação básica. Por isso, não ser capaz de sinalizar emoções como raiva, tristeza ou nojo pode resultar em isolamento social e afetar negativamente a comunicação verbal ou não de indivíduos [11]. Em consequência disso, esses pacientes

podem ter dificuldades de comunicar mensagens importantes, como o desconforto associado a tratamentos. Portanto, identificar as emoções durante a terapia pode auxiliar o terapeuta a mudar/personalizar sua abordagem com base no *biofeedback* retornado.

É importante destacar que as aplicações que visam realizar o reconhecimento automático de emoções através das expressões faciais são diversas. Por exemplo, com o intuito de identificar expressões faciais emocionais de motoristas e prestar assistência imediata para fins de segurança, Sahoo *et al.* [12] desenvolveram métodos de *Transfer Learning* (TL) baseados nas redes neurais convolucionais (do inglês, *Convolutional Neural Networks* - CNNs) pré-treinadas AlexNet, SqueezeNet e VGG19. Dentre todas as arquiteturas testadas, a que obteve melhor desempenho na maioria das bases de dados foi a VGG19, com acurácia de 66,58%, 84,38%, 92,99%, 98,98%, 56,02% e 99,7% para as bases FER2013, JAFFE, KDEF, CK+, SFEW e KMU-FED, respectivamente. Podder *et al.* [13] também propõe uma CNN baseada em parâmetros mínimos e TL a fim de desenvolver um método com um custo computacional minimizado. A abordagem desenvolvida foi denominada de LiveEmoNet e treinada com as bases de dados FER2013, JAFFE e CK+. Os resultados obtidos apontam acurácia de 68,93%, 97,66% e 96,67% para as respectivas bases de dados.

Uma das arquiteturas híbridas proposta neste estudo utiliza uma *Deep-Wavelet Neural Network* (DWNN) para realizar a tarefa de extração de atributos das imagens de expressões faciais emocionais. Este modelo, também foi proposto e utilizado no estudo de De Freitas Barbosa [14], mas com o objetivo de identificar o câncer de mama em estágios iniciais. Então, os autores utilizaram a DWNN para extrair atributos de imagens de termografia mamária e um algoritmos simples, como SVM, MLP e ELM para detecção e classificação de lesão mamária nas imagens. Os resultados obtidos foram promissores, alcançando 0,95 de sensibilidade e 68,4% de acurácia, utilizando o algoritmo MLP. É válido mencionar que a primeira vez que a DWNN está sendo utilizada para fins de reconhecimento de emoções é nesta pesquisa.

Sabendo desse contexto, este trabalho tem como objetivo apresentar um comparativo entre duas abordagens de arquiteturas híbridas com diferentes configurações de pré-processamento dos dados. Pode-se compreender uma arquitetura híbrida como aquela que utiliza algoritmos distintos para executar diferentes tarefas que se complementam para atingir o objetivo final de um determinado sistema. A primeira arquitetura é baseada em *Transfer*

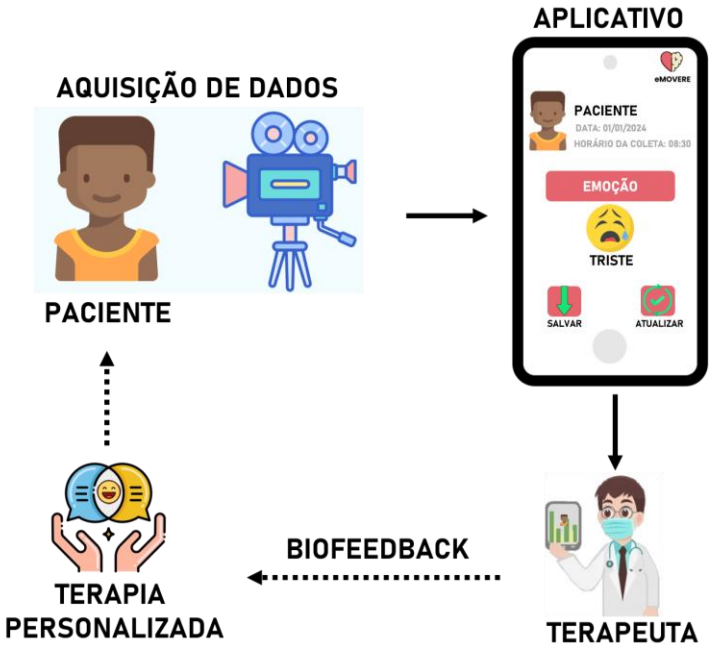
Learning, onde são utilizadas redes neurais profundas pré-treinadas (para realizar extração de atributos) e o algoritmo *Random Forest* (usado para etapa de classificação). A segunda arquitetura apresenta uma *Deep-Wavelet Neural Network* para extração de atributos e um *Random Forest* para classificação. O intuito é que a melhor arquitetura e configuração de pré-processamento de dados componha um sistema para reconhecimento de emoções através das Expressões Faciais, que será utilizado durante sessões de terapia para retornar *biofeedbacks* dos estados afetivos dos pacientes, auxiliando terapeutas para a personalização de forma mais assertiva no tratamento. Nossas arquiteturas foram treinadas e validadas com a base de dados FER2013 [15], amplamente utilizada na literatura.

Métodos

Proposta de Aplicação

Este trabalho faz parte de um projeto mais amplo, onde busca-se desenvolver o aplicativo "eMOVERE" para realizar o reconhecimento automático de emoções, que será utilizado para auxiliar terapeutas. A motivação de desenvolvimento deste projeto surge diante de pelo menos três contextos, que são: 1) Dificuldade de determinados públicos em expressar suas emoções durante tratamentos, resultando em isolamento, desconfortos e ineficácia do mesmo; 2) Terapias personalizadas estão sendo adotadas como estratégia para engajar cada vez mais os pacientes e tornar as abordagens mais assertivas e, 3) Sistemas de reconhecimento de emoções tem se popularizado cada vez mais diante do aumento da interação humano-computador, se tornando uma ferramenta útil para a área da saúde. Como é possível visualizar na Figura 1, na primeira etapa teremos a aquisição de dados (expressões faciais) que posteriormente serão pré-processados num sistema embarcado. Em seguida, as informações sobre os estados afetivos dos pacientes serão exibidas num aplicativo para os terapeutas. O objetivo é que o sistema possa ser utilizado para avaliar e personalizar a abordagem terapêutica.

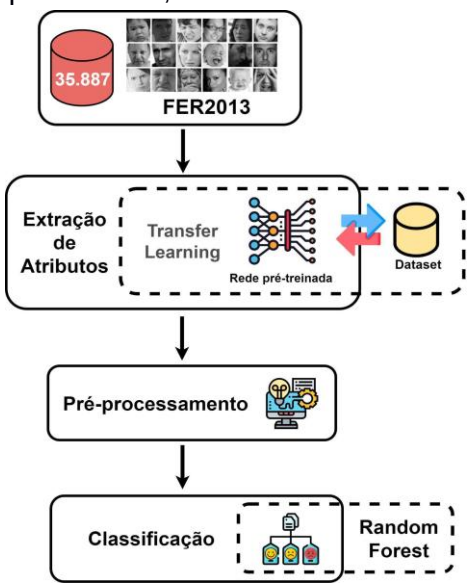
Figura 1 – Diagrama referente a motivação e proposta de aplicação do presente trabalho.



Experimentos

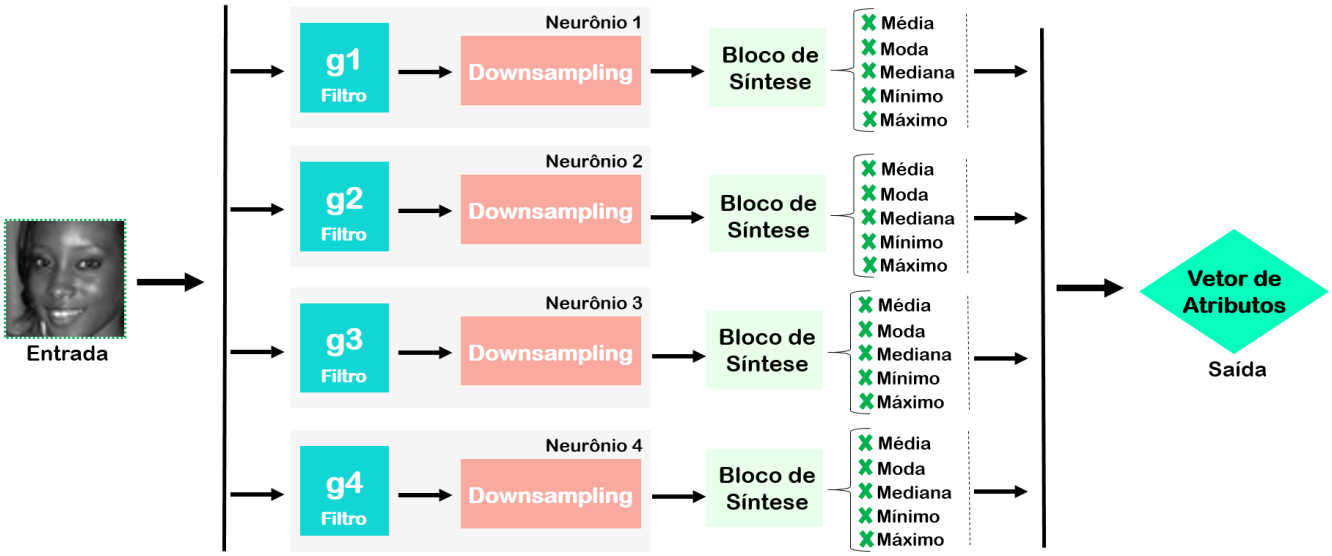
Neste trabalho, foram exploradas diferentes arquiteturas híbridas. Para isso, dois experimentos com diferentes configurações de pré-processamento dos dados foram executados. No primeiro experimento, como ilustrado na Figura 2, a base de dados FER2013 passou pelo processo de extração de atributos através de redes pré-treinadas, como DenseNet, ResNet50, SqueezeNet, LeNet e VGG16, com o conjunto de dados ImageNet, com exceção da LeNet, que foi pré-treinada com o conjunto MNIST. Em seguida, o conjunto de dados foi dividido em 70% para treinamento-validação e 30% para teste. Os conjuntos de treinamento-validação passaram por etapas de pré-processamento, como balanceamento de classes e seleção de atributos. Por fim, um algoritmo *Random Forest* (RF) com configuração de 400 árvores foi utilizado para realizar a tarefa de classificação das emoções. A escolha por esse algoritmo se deu por ele ser um dos mais simples, eficientes e explicativos.

Figura 2 – Diagrama referente ao experimento 1, utilizando Transfer Learning.



Como é possível visualizar na Figura 3, no segundo experimento o método utilizado para extração de atributos das imagens da base FER2013 foi baseado na decomposição de Wavelet, onde nomeamos a abordagem de *Deep-Wavelet Neural Network* (DWNN). A decomposição de *Wavelet* é baseada no algoritmo de Mallat [16], onde filtros passa-baixa e passa-alta são aplicados a uma imagem, resultando em um conjunto de quatro imagens (em casos de 1 camada, como é exemplificado na Figura 3), sendo uma imagem de aproximação e três imagens de detalhes horizontais, verticais e diagonais. Na DWNN um neurônio é composto pela combinação de um filtro com a operação de *downsampling*. Todos os filtros utilizados na DWNN são mantidos fixos (referem-se às famílias das Wavelets do tipo Haar) durante todo o processo e formam um banco de filtros. Então, se o banco possui *g* filtros, uma determinada imagem de entrada será submetida a *g* neurônios, formando a primeira camada da nossa rede neural. Na segunda camada, as imagens resultantes da primeira serão submetidas ao mesmo banco de filtros e ao *downsampling*. Esse processo se repete para a terceira camada e assim sucessivamente. Por fim, o bloco de síntese, considerado a camada de saída da DWNN, é responsável por extrair a média, moda, mediana, mínimo e máximo das imagens resultantes do processo. As informações extraídas formam um vetor de atributos utilizado para o algoritmo *Random Forest* de 400 árvores realizar a tarefa de classificação. Para este experimento, 4 camadas da DWNN foram utilizadas.

Figura 3 – Diagrama do experimento 2, ilustrando o funcionamento da DWNN com uma camada.



Como já mencionado, para ambos os experimentos, a fim de identificar a melhor configuração de pré-processamento dos dados, diferentes configurações foram aplicadas. Então, para cada experimento, as seguintes abordagens foram exploradas:

- **Abordagem 1:** conjunto de dados de treinamento com as classes desbalanceadas;
- **Abordagem 2:** conjunto de dados de treinamento com as classes balanceadas utilizando o método SMOTE (*Synthetic Minority Over-sampling TEchnique*) [17], com um número de *k* vizinhos igual a 3;
- **Abordagem 3:** conjunto de dados com as classes balanceadas (SMOTE) e com seleção de atributos relevantes utilizando o método PSO (*Particle Swarm Optimization*) [18], com o número de população e iterações igual a 50.

Por fim, com o intuito de obter dados estatísticos, cada configuração testada foi executada com 30 repetições. Além disso, o método de validação cruzada com 10 *folds* também foi aplicado. Para avaliar os resultados obtidos, cinco métricas foram utilizadas: Acurácia, Índice de Kappa, Sensibilidade, Especificidade e Área sob a curva ROC (AUC).

Base de dados

A base de dados FER2013 foi introduzida no *Challenges Representation Learning*, por Goodfellow *et al.* (2013) [15]. Ao todo, a base possui 35.887 imagens de expressões faciais emocionais, todas em escala de cinza e com redimensionamento de 48x48 *pixels*. As imagens estão distribuídas em sete classes de emoções, que são: Raiva (4.953), Nojo (587), Medo (5.121), Feliz (8.989), Triste (6.077), Surpreso (4.002) e Neutro (6.198).

Resultados e Discussões

Conforme é possível visualizar na Tabela 1, a arquitetura que obteve melhor desempenho na abordagem de pré-processamento dos dados (classes com dados desbalanceados) no experimento 1 (utilizando redes pré-treinadas para extração de atributos) foi a rede DenseNet com 50,29% de acurácia média, seguida pela rede VGG16, com 46,78%. Um ponto relevante que deve ser considerado é que para todas as arquiteturas a métrica de índice kappa se manteve baixa, indicando discordância na classificação. Pois, quanto mais próximo de 1, maior é a concordância. Entretanto, quanto mais próximo de 0 ou valores negativos, maior é a probabilidade de concordância aleatória. Sabendo deste fato e considerando que o índice kappa não é uma métrica sensível ao desbalanceamento, a abordagem 1 testada neste experimento não alcançou bons resultados.

Tabela 1 – Resultados do experimento 1 com a abordagem 1.

Arquitetura	Acurácia	Kappa	Sensibilidade	Especificidade	AUC
DenseNet + RF	**50,29 ± 0,81**	**0,3812 ± 0,0104**	**0,8399 ± 0,0140**	**0,6849 ± 0,0110**	**0,8623 ± 0,0079**
ResNet50 + RF	41,21 ± 0,80	0,2570 ± 0.0104	0,8414 ± 0.0151	0,5011 ± 0,0113	0,7653 ± 0,0111
SqueezeNet + RF	44,39 ± 0,84	0,3086 ± 0.0106	0,7468 ± 0.0172	0,6417 ± 0.0112	0,7759 ± 0.0101
LeNet + RF	43,82 ± 0,77	0,2945 ± 0.0100	0,8224 ± 0.0151	0,5678 ± 0.0111	0,7828 ± 0.0100
VGG16 + RF	46,78 ± 0,81	0,3357 ± 0.0104	0,7903 ± 0.0143	0,6291 ± 0.0099	0,8013 ± 0.0090

Os resultados da abordagem 2 (classes balanceadas com o método SMOTE) do experimento 1 são apresentados na Tabela 2. Todas as arquiteturas tiveram seu desempenho melhorado significativamente, inclusive o índice kappa aumentou, indicando concordância na

classificação. Dentre todas as arquiteturas, destacou-se a rede LeNet com 75,35% de acurácia, seguida pela VGG16, com 75,08%.

Tabela 2 – Resultados do experimento 1 com a abordagem 2.

Arquitetura	Acurácia	Kappa	Sensibilidade	Especificidade	AUC
DenseNet + RF	72,29 ± 0,67	0,6767 ± 0,0078	0,6403 ± 0,0191	0,9379 ± 0,0043	0,9049 ± 0,0055
ResNet50 + RF	72,39 ± 0,68	0,6779 ± 0,0079	0,5486 ± 0,0192	0,9204 ± 0,0047	0,8534 ± 0,0070
SqueezeNet + RF	68,05 ± 0,63	0,6272 ± 0.0074	0,4918 ± 0.0182	0,9235 ± 0.0042	0,8381 ± 0.0082
LeNet + RF	**75,35 ± 0,62**	**0,7124 ± 0.0072**	**0,5853 ± 0.0188**	**0,9284 ± 0.0041**	**0,8794 ± 0.0066**
VGG16 + RF	75,08 ± 0,61	0,7093 ± 0.0071	0,5648 ± 0.0185	0,9379 ± 0.0038	0,8771 ± 0.0068

A Tabela 3 apresenta os resultados da abordagem 3 (com classes balanceadas e seleção de atributos) no experimento 1. Entre as variações de arquiteturas, destacou-se a rede VGG16 em relação a acurácia (74,52%), kappa (0,7027), sensibilidade (0,5655), especificidade (0,9355) e AUC (0,8749). Como em todos as abordagens a VGG16 se manteve entre os melhores resultados e com poucas variações, assumimos que para a abordagem de *Transfer Learning* (experimento 1), a melhor arquitetura é composta por um VGG16 e um *Random Forest*, com classes balanceadas e seleção de atributos.

Tabela 3 – Resultados do experimento 1 com a abordagem 3.

Arquitetura	Acurácia	Kappa	Sensibilidade	Especificidade	AUC
DenseNet + RF	71,55 ± 0,69	0,6681 ± 0,0081	0,6421 ± 0,0186	0,9364 ± 0,0040	0,9050 ± 0,0054
ResNet50 + RF	71,78 ± 0,66	0,6707 ± 0,0077	0,5556 ± 0,0180	0,9175 ± 0,0046	0,8521 ± 0,0072
SqueezeNet + RF	66,31 ± 0,66	0,6069 ± 0,0077	0,4750 ± 0,0188	0,9198 ± 0,0042	0,8286 ± 0,0083
LeNet + RF	73,64 ± 0,65	0,6925 ± 0,0076	0,5886 ± 0,0196	0,9226 ± 0,0044	0,8761 ± 0,0066
VGG16 + RF	**74,52 ± 0,62**	**0,7027 ± 0,0072**	**0,5655 ± 0,0185**	**0,9355 ± 0,0038**	**0,8749 ± 0,0069**

Na Tabela 4 é possível visualizar os resultados do experimento 2 (utilizando DWNN para extração dos atributos) com a abordagem 1. Nesse experimento foram testadas 4 camadas, dentre elas destacou-se a camada 3 com 37,54% de acurácia e a camada 4, com 37,21%. Assim como no experimento 1 utilizando a abordagem 1, o índice kappa nesse experimento e com essa abordagem também não obteve bons resultados, indicando discordância na classificação.

Tabela 4 – Resultados do experimento 2 com a abordagem 1.

Arquitetura	Acurácia	Kappa	Sensibilidade	Especificidade	AUC
Camada 1 + RF	29.81 ± 0,82	0.1221 ± 0.0103	0.5715 ± 0.0193	0.5295 ± 0.0113	0.5774 ± 0.0121
Camada 2 + RF	35.13 ± 0,77	0.1731 ± 0.0099	0.7817 ± 0.0176	0.4025 ± 0.0103	0.6537 ± 0.0118
Camada 3 + RF	**37.54 ± 0,84**	**0.2156 ± 0.0108**	**0.7098 ± 0.0185**	**0.5366 ± 0.0113**	**0.6840 ± 0.0115**
Camada 4 + RF	37.21 ± 0,80	0.2076 ± 0.0103	0.7365 ± 0.0182	0.4965 ± 0.0111	0.6777 ± 0.0117

Os resultados do experimento 2 com a abordagem 2 são apresentados na Tabela 5. A camada 4 obteve melhor resultado, com 71,19% de acurácia, seguida pela camada 3, com 70,30%.Nessa abordagem o índice kappa melhorou, indicando concordância na classificação.

Tabela 5 – Resultados do experimento 2 com a abordagem 2.

Arquitetura	Acurácia	Kappa	Sensibilidade	Especificidade	AUC
Camada 1 + RF	61.27 ± 0,66	0.5481 ± 0.0077	0.2562 ± 0.0171	0.9232 ± 0.0045	0.7241 ± 0.0092
Camada 2 + RF	68.90 ± 0,64	0.6372 ± 0.0074	0.3820 ± 0.0178	0.9277 ± 0.0042	0.8088 ± 0.0077
Camada 3 + RF	70.30 ± 0,69	0.6535 ± 0.0081	0.4149 ± 0.0179	0.9228 ± 0.0047	0.8070 ± 0.0081
Camada 4 + RF	**71.19 ± 0.59**	**0.6638 ± 0.0069**	**0.4091 ± 0.0188**	**0.9256 ± 0.0043**	**0.8056 ± 0.0086**

A Tabela 6 apresenta os resultados do experimento 2 com a abordagem 3. Como é possível visualizar, a Camada 4 se manteve com bons resultados em relação a acurácia (70,77%), kappa (0,6590), sensibilidade (0,4073), especificidade (0,9256) e AUC (0,8049). Diante dos resultados, para o experimento 2, escolhemos a abordagem 3 como a camada 4 como melhor configuração de arquitetura híbrida.

Tabela 6 – Resultados do experimento 2 com a abordagem 3.

Arquitetura	Acurácia	Kappa	Sensibilidade	Especificidade	AUC
Camada 1 + RF	53.71 ± 0,70	0.4600 ± 0.0082	0.2244 ± 0.0159	0.9096 ± 0.0048	0.6635 ± 0.0100
Camada 2 + RF	64.64 ± 0,69	0.5875 ± 0.0080	0.3951 ± 0.0190	0.9064 ± 0.0054	0.7840 ± 0.0082
Camada 3 + RF	68.56 ± 0,70	0.6332 ± 0.0082	0.4062 ± 0.0187	0.9172 ± 0.0046	0.7949 ± 0.0083
Camada 4 + RF	**70.77 ± 0,66**	**0.6590 ± 0.0077**	**0.4073 ± 0.0187**	**0.9256 ± 0.0042**	**0.8049 ± 0.0078**

As arquiteturas híbridas de ambos os experimentos conseguiram ter um bom desempenho com a abordagem 3, utilizando menor quantidade de atributos. Para fins

comparativos, a VGG16 possui 4.098 atributos em sua versão original. Após a seleção de atributos com o método PSO na etapa de pré-processamento, a VGG16 ficou com 1.093 atributos. Em relação à camada 4 da DWNN, a quantidade original de atributos é de 1.280. Com a aplicação do PSO, 437 foram selecionados como relevantes para a classificação.

Após analisar a etapa de treinamento-validação dos modelos, foi realizada a etapa de teste apenas com as duas melhores arquiteturas híbridas (VGG16 + RF e DWNN com Camada 4 + RF) e a configuração de pré-processamento dos dados (classes balanceadas com SMOTE e seleção de atributos com PSO), utilizando apenas o conjunto de teste. É válido ressaltar que o conjunto de teste não foi utilizado durante o treinamento-validação dos modelos. Como resultado desta etapa, a abordagem de VGG16 e um *Random Forest* se destaca como melhor arquitetura híbrida, alcançando 84,72% de acurácia, 0,815 de índice kappa, 0,847 de sensibilidade, 0,967 de especificidade e 0,981 de AUC. Em relação a DWNN com camada 4 e um RF, a acurácia obtida foi de 81,21%, com kappa de 0,773, sensibilidade de 0,812, especificidade de 0,956 e AUC de 0,971. A seguir, na Tabela 7 é possível comparar o desempenho das arquiteturas híbridas propostas nessa pesquisa com alguns trabalhos da literatura, utilizando apenas a métrica de acurácia.

Tabela 7 – Comparação de resultados com trabalhos da literatura.

Trabalho	Método	Base de dados	Acurácia
Sahoo *et al*. (2023) [12]	VGG19	FER2013	66,58%
Podder *et al*. (2022) [13]	LiveEmoNet	FER2013	68,93%
Yang *et al.* (2021) [19]	Inception-ResNet-v1 e SVM	FER2013	68,1%
Ab Wahab *et al*. (2021) [20]	CNN-KNN	FER2013	72,26%
Gunawan *et al.* 2020 [21]	CNN	FER2013	57,4%
Arquitetura proposta 1	**VGG16-RF**	**FER2013**	**74,52%**
Arquitetura proposta 2	**DWNN Camada 4-RF**	**FER2013**	**70,77%**

A base de dados FER2013 é amplamente utilizada na área de reconhecimento de emoções através de expressões faciais, apesar de apresentar muitos desafios, como ambiguidade nos rótulos, desbalanceamento de classes, qualidade das imagens e expressões faciais muito superficiais. Todos esses fatores influenciam no desempenho dos modelos citados

na Tabela 9. As arquiteturas híbridas que estamos propondo, em relação aos trabalhos citados, se destacam positivamente. A VGG16-RF se sobressai com 74,52% de acurácia e logo em seguida tem-se a DWNN com camada 4-RF com 70,77%, ficando atrás para o desempenho apenas do modelo de Ab Wahab *et al.* (2021) [19]. É válido destacar que um dos grandes achados é que tanto a arquitetura híbrida de VGG16-RF quanto a de DWNN camada 4-RF que estamos propondo possuem bom desempenho, mesmo utilizando apenas 27% e 34% dos atributos, respectivamente. Esse fato contribui para um custo computacional minimizado. Além disso, ambas as arquiteturas possuem um bom desempenho utilizando o conjunto de dados de teste, demonstrando boa capacidade de generalização. Então, é importante mencionar que a abordagem de DWNN apresentada não passa pelo processo de treinamento, diferentemente das redes CNNs apresentadas, tornando o processo de aplicação da arquitetura mais rápido.

Conclusão

Este trabalho teve como intuito analisar um comparativo entre arquiteturas híbridas baseadas em *Transfer Learning* e Decomposição de *Wavelet* com diferentes configurações de pré-processamento dos dados. O objetivo é que a melhor arquitetura componha um sistema de reconhecimento de emoções através de expressões faciais que será utilizado por terapeutas para retornar *biofeedbacks* dos estados afetivos dos pacientes, auxiliando para a customização do tratamento, tornando-o mais assertivo.

A arquitetura híbrida baseada em VGG16 e *Random Forest* se mostrou ideal, apresentando bons resultados tanto na etapa de treinamento-validação (com 74,52% de acurácia) quanto na etapa de teste (com 84,72% de acurácia), utilizando apenas 27% dos atributos da VGG16. Em seguida, destacamos o potencial da arquitetura baseada em transformada de *Wavelet*, a DWNN com apenas 4 camadas e um *Random Forest* obteve 70,77% de acurácia na etapa de treinamento-validação e 81,21% no teste, utilizando somente 34% dos atributos. Além disso, a contribuição desse trabalho também é apresentar a DWNN como uma alternativa às redes neurais convolucionais convencionais, pois esta abordagem não possui ajustes constantes de pesos, mas sim um banco de filtros que são fixos, agilizando a aplicação da rede e alcançando resultados quase equivalentes com menos atributos sendo extraídos das imagens.

Como perspectivas de trabalhos futuros, pretende-se: 1) Realizar novos experimentos com ambas as arquiteturas híbridas, mas utilizando diferentes bases de dados, inclusive com uma autoral no contexto de idosos (que se encontra em processo de pré-processamento); 2) Testar em novos experimentos diferentes redes pré-treinadas; 3) Realizar novos experimentos com a DWNN, mas utilizando diferentes configurações de redimensionamento das imagens; 4) Testar outros algoritmos além do *Random Forest* para realizar a tarefa de classificação das emoções e, por fim, 5) Desenvolver um sistema *web* ou *mobile* para ser utilizado por terapeutas em contextos reais, contribuindo para personalização de tratamentos e, consequentemente, melhorando a qualidade de vida dos pacientes e de sua rede de apoio.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Referências

1. Adyapady, R. Rashmi; Annappa, B. A comprehensive review of facial expression recognition techniques. Multimedia Systems, v. 29, n. 1, p. 73-103, 2023.

2. Khateeb, Muhammad; Anwar, Syed Muhammad; Alnowami, Majdi. Multi-domain feature fusion for emotion classification using DEAP dataset. Ieee Access, v. 9, p. 12134-12142, 2021.

3. Leong, Sze Chit et al. Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing. Computer Science Review, v. 48, p. 100545, 2023.

4. Torcate, Arianne Sarmento; De Santana, Maíra Araújo; Dos Santos, Wellington Pinheiro. Emotion Recognition to Support Personalized Therapy: An Approach Based on a Hybrid Architecture of CNN and Random Forest. In: 2023 IEEE Latin American Conference on Computational Intelligence, 2023.

5. González, Eduardo J. Santos; Mcmullen, Kyla. The design of an algorithmic modal music platform for eliciting and detecting emotion. In: 2020 8th international winter conference on brain-computer interface (bci). IEEE, 2020. p. 1-3.

6. Gong, Weijun et al. Enhanced spatial-temporal learning network for dynamic facial expression recognition. Biomedical Signal Processing and Control, v. 88, p. 105316, 2024.

7. Motadi, Lesetja et al. Ai as a novel approach for exploring ccfnas in personalized clinical diagnosis and prognosis: Providing insight into the decision-making in precision oncology. In: Artificial

Intelligence and Precision Oncology: Bridging Cancer Research and Clinical Decision Support. Cham: Springer Nature Switzerland, 2023. p. 73-91.

8. Ferreira, Cyntia Diógenes; Torro-Alves, Nelson. Reconhecimento de emoções faciais no envelhecimento: uma revisão sistemática. Universitas Psychologica, v. 15, p. 1-12, 2016.

9. Teh, Elizabeth J.; Yap, Melvin J.; Rickard Liow, Susan J. Emotional processing in autism spectrum disorders: Effects of age, emotional valence, and social engagement on emotional language use. Journal of autism and developmental disorders, v. 48, p. 4138-4154, 2018.

10. Bernieri, G., & Duarte, J. C. (2023). Identifying Alzheimer's Disease Through Speech Using Emotion Recognition. Journal of Health Informatics, 15 (Especial). https://doi.org/10.59681/2175-4411.v15.

11. Grondhuis, Sabrina N. et al. Having difficulties reading the facial expression of older individuals? Blame it on the facial muscles, not the wrinkles. Frontiers in Psychology, v. 12, p. 620768, 2021.

12. Sahoo, Goutam Kumar; Das, Santos Kumar; Singh, Poonam. Performance comparison of facial emotion recognition: a transfer learning-based driver assistance framework for in-vehicle applications. Circuits, Systems, and Signal Processing, v. 42, n. 7, p. 4292-4319, 2023.

13. Podder, Tanusree; Bhattacharya, Diptendu; Majumdar, Abhishek. Time efficient real time facial expression recognition with CNN and transfer learning. Sādhanā, v. 47, n. 3, p. 177, 2022.

14. De Freitas Barbosa, Valter Augusto et al. Deep-wavelet neural networks for breast cancer early diagnosis using mammary termographies. In: Deep learning for data analytics. Academic Press, 2020. p. 99-124.

15. Goodfellow, Ian J. et al. Challenges in representation learning: A report on three machine learning contests. In: Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20. Springer berlin heidelberg, 2013. p. 117-124.

16. Mallat, Stephane G. Multifrequency channel decompositions of images and wavelet models. IEEE Transactions on Acoustics, speech, and signal processing, v. 37, p. 2091-2110, 1989.

17. Chawla, Nitesh V.. et al. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, v. 16, p. 321-357, 2002.

18. Kennedy, James; Eberhart, Russell. Particle swarm optimization. In: Proceedings of ICNN'95-international conference on neural networks. ieee, 1995. p. 1942-1948.

19. Yang, Lei et al. Facial expression recognition based on transfer learning and SVM. In: Journal of Physics: Conference Series. IOP Publishing, 2021. p. 01.

20. Ab Wahab, Mohd Nadhir et al. Efficient net-lite and hybrid CNN-KNN implementation for facial expression recognition on raspberry pi. IEEE Access, v. 9, p. 134065-134080, 2021.

21. Gunawan, Teddy Surya et al. Development of video-based emotion recognition using deep learning with Google Colab. TELKOMNIKA (Telecommunication Computing Electronics and Control), v. 18, n. 5, p. 2463-2471, 2020.