Uma Técnica de Atenção Inter-Canais para Detecção de Depressão a partir de Expressões Faciais

Wheidima Carneiro de Melo[1], Waldir Sabino da Silva Júnior[2], Celso Barbosa Carvalho[2]
[1]Universidade do Estado do Amazonas (UEA), AM, Brasil
[2]Universidade Federal do Amazonas (UFAM), AM, Brasil
Emails: wmelo@uea.edu.br, waldirjr@ufam.edu.br, ccarvalho_@ufam.edu.br

***Resumo*— A detecção automática de depressão por meio das expressões faciais pode ser baseada nos modelos de aprendizado profundo. Entretanto, a quantidade limitada de dados para o treinamento de tais modelos permanece um desafio a ser superado. Neste artigo, propõe-se uma técnica de atenção inter-canais para facilitar a geração de representações discriminativas. A técnica é inserida no modelo ResNet-50 para explorar a saída da última camada de convolução. Os experimentos demonstram que a técnica é capaz de melhorar o desempenho do modelo ResNet-50 e que essa combinação produz resultados competitivos em relação ao estado da arte, enquanto requer menos recursos computacionais.**

***Palavras-Chave*— Detecção de depressão, aprendizado profundo, redes neurais convolucionais, técnica de atenção.**

***Abstract*— Automatic depression detection from facial expressions can be based on deep learning methods. However, the limited amount of data for the training of such methods remains a challenge. This paper proposes a cross-channel attention technique to facilitate the generation of discriminative representations. The technique is inserted on the ResNet-50 model to explore the output of the last convolutional layer. The experiments show that the technique is able to improve the performance of ResNet-50 and that this combination achieves competitive results in comparison with state-of-the-art, while requiring less computational resources.**

***Keywords*— Depression detection, deep learning, convolutional neural network, attention technique.**

I. Introdução

Depressão é classificada como um transtorno mental incapacitante e comum, com prevalência global estimada de mais de 300 milhões de pessoas [1]. Tal condição clínica pode causar alterações no apetite [2], distúrbios do sono [3], dificuldade de concentração [2], dor de cabeça [4], dor nas costas [4], dor de estômago [5], ansiedade [6], irritabilidade [7], tristeza [8], diminuição de prazer ou interesse em pessoas ou coisas [2]. Nos casos severos, a depressão pode levar ao abuso de substâncias e suicídio [9]. Além disso, essa condição pode aumentar as chances de adquirir estados clínicos graves, ou contribuir para a piora desses estados, tais como diabetes, câncer e doenças cardiovasculares [10].

Apesar da gravidade, existem tratamentos efetivos para depressão. Os procedimentos clínicos normalmente adotam antidepressivos e estabilizadores de humor [11]. Nesse processo, o correto diagnóstico de depressão e a sua severidade é fundamental para a redução das consequências negativas na vida do paciente. As avaliações de depressão podem ser caracterizadas como de natureza subjetiva. De fato, o diagnóstico depende do entendimento do relato do paciente por parte de um profissional de saúde ou do emprego de um instrumento de autoavaliação, como o Inventário de Depressão de Beck (*Beck Depression Inventory* – BDI) [14]. Estudos relatam que os clínicos enfrentam dificuldades em reconhecer os estados depressivos [12], [13] e que os erros na avaliação podem trazer sérias consequências para os pacientes [13].

Neste contexto, os sistemas automáticos de avaliação de depressão podem ser uma ferramenta fundamental para auxiliar no diagnóstico clínico, pois podem fornecer uma estimação de maneira objetiva e precisa. Tais sistemas podem explorar as expressões faciais para gerar uma estimação, as quais podem ser capturadas por câmeras de baixo custo. Esses sistemas são constituídos de três etapas: pré-processamento, extração de características e regressão. A etapa de pré-processamento é responsável por localizar e extrair a face a partir de uma imagem de entrada. A saída dessa etapa é explorada pela etapa de extração de características, a qual detecta e captura padrões relacionados à depressão presentes nas expressões faciais. É nessa etapa em que ocorre a geração de representações que permitem a distinção de um indivíduo saudável de um estando em depressão. Consequentemente, o desempenho dos sistemas de detecção de depressão depende fortemente das características extraídas (i.e., representações). Finalmente, a etapa de regressão explora as características geradas para produzir uma pontuação de depressão (ou seja, um valor numérico).

Os avanços em técnicas de aprendizado profundo têm permitido a geração de representações que facilitam a discriminação de diferentes classes em diversas aplicações, tais como classificação de imagens [15], detecção de objetos [16] e reconhecimento de ações humanas em vídeos [17]. Um fato comum nessas aplicações é a disponibilidade de grandes quantidades de dados de treinamento, favorecendo a capacidade de geração de representações pelos modelos baseados em aprendizado profundo. Entretanto, a base de dados de depressão possui uma quantidade pequena de dados de treinamento. Esse cenário aumenta as chances de ocorrência de *overfitting*, o que diminui o poder de generalização do modelo. Para minimizar esse problema, os modelos precisam conter estruturas que

facilitem o aprendizado de padrões de depressão.

Neste artigo, propõe-se uma técnica baseada em aprendizado profundo com o intuito de facilitar a geração de representações discriminativas para detecção de depressão a partir de expressões faciais. A técnica proposta explora as correlações dos elementos de um mapa de características, oriundos dos canais da última camada convolucional de um modelo, e atribui mais atenção para os elementos mais informativos em relação aos padrões de depressão. Em específico, insere-se a técnica proposta no modelo denominado Rede Residual (*Residual Network* – ResNet). Esse processo resulta no aumento de desempenho do modelo ResNet-50 na detecção de depressão.

II. Trabalhos Relacionados

Pessoas sofrendo de depressão exibem alterações no comportamento facial [18]. Os sistemas de detecção de depressão mapeiam essas alterações para reconhecer os estados depressivos. Com isso, representações de depressão são geradas de modo a permitir a discriminação do estado clínico de um indivíduo. Os métodos de geração de representações podem ser divididos em duas categorias: descritores tradicionais e algoritmos de aprendizado profundo.

A. Descritores Tradicionais

Os métodos baseados em descritores tradicionais extraem características usando um algoritmo predefinido manualmente, o qual é elaborado com base no conhecimento do especialista [19]. Essa abordagem foi muito empregada nos primeiros desenvolvimentos para detecção de depressão. Entre os descritores utilizados pode-se destacar o *Local Binary Pattern* (LBP) [20], o *Local Phase Quantization* (LPQ) [21], o *Edge Orientation Histograms* (EOH) [22], o *Local Gabor Binary Patterns from Three Orthogonal Planes* (LGBP-TOP) [23]. Usando os descritores LBP, LPQ e EOH, Jan *et al*. [24] extraem diferentes características e captura as suas variações usando o algoritmo denominado *Motion History Histogram* (MHH). Valstar *et al*. [25] empregam o descritor LPQ na etapa de extração de características e um método denominado *Support Vector Regression* (SVR) na etapa de regressão. Em [26], os autores definem o descritor LGBP-TOP para capturar as variações da informação de aparência ao longo do tempo. Kaya *et al*. [27] exploram as correlações de características geradas pelos descritores LGBP-TOP e LQP usando um método denominado *Canonical Correlation Analysis* (CCA). A grande desvantagem dos descritores tradicionais é a limitação na geração de representações discriminativas. Isso porque é difícil elaborar manualmente uma estratégia para detectar e extrair padrões de depressão.

B. Algoritmos de Aprendizado Profundo

Os algoritmos de aprendizado profundo têm demonstrado maior capacidade na geração de representações de depressão. Tais algoritmos automaticamente encontram essas representações a partir dos dados de entrada. A abordagem desses algoritmos é caracterizada pelo emprego de múltiplas transformações não-lineares para se alcançar representações mais abstratas [28], levando a facilitação da distinção dos diferentes estados em análise. Normalmente, emprega-se uma rede neural convolucional 2D (Convolutional Neural Network – 2D CNN) para explorar a informação de aparência das expressões faciais e um método adicional para melhorar o desempenho do algoritmo ou capturar a informação temporal. Usando essa abordagem, Kang *et al*. [29] empregam uma 2D CNN para gerar representações e um método denominado *Deep Transformation Learning* (DTL) para projetá-las em outro domínio. He *et al*. [30] utilizam uma 2D CNN e dois esquemas, um para explorar segmentos locais e um outro para explorar a informação global. Em [31], os autores empregam quatro modelos ResNet-50 para explorar regiões faciais e assim indicar as mais relevantes para a estimação. Os autores em [32] utilizam o modelo ResNet-50 em conjunto com uma técnica que permite o aprendizado de uma distribuição de depressão para cada imagem de entrada. O modelo ResNet-50 também é usado em [33], porém, desta vez, em cascata com um módulo de atenção que possui a função de combinar representações geradas para uma sequência de imagens.

Uma outra estratégia possível é a direta exploração da informação espacial e temporal. As 3D CNNs são uma opção para esse caso, pois analisam as informações entre uma sequência de imagens. Jazaery *et al*. [34] utilizam dois modelos de uma 3D CNN denominada *Convolutional 3D* (C3D) para aprender características espaço-temporais de regiões da face em duas distintas escalas. De maneira similar, os autores em [35] empregam duas redes C3D e um esquema de fusão para combinar a pontuação de depressão estimada por cada rede. Zhou *et al*. [36] utilizam a rede C3D para gerar características espaço-temporais e um método para aprender uma distribuição de depressão para cada videoclip de entrada. Existem outras formas de realizar a exploração da informação espacial e temporal. Em [37], [38], os autores empregam uma arquitetura denominada *two-stream network* que é formada por duas 2D CNNs, uma para explorar a aparência das expressões faciais, a outra para explorar a dinâmica delas. Similarmente, Uddin *al el*. [39] utiliza um descritor tradicional, denominado *Volume Local Directional Number* (VLDN), em cascata com uma 2D CNN para explorar a informação dinâmica e múltiplas 2D CNNs para explorar a informação de aparência. Já os autores em [40] empregam uma arquitetura que é composta por uma estrutura capaz de gerar características espaço-temporais em multiescala, a qual usa operadores de convolução espacial e funções para capturar a informação temporal.

Em contraste com os trabalhos existentes, o presente artigo introduz uma técnica de atenção capaz de melhorar o poder de geração de representações de um modelo. Para isso, a técnica procura dar mais atenção para os canais, da última camada de convolução de um modelo, que possuem uma informação que contribui mais para a detecção de padrões de depressão a partir de expressões faciais.

III. Técnica de Atenção Proposta

As CNNs são tipicamente empregadas nos sistemas de detecção de depressão. Tais redes são uma sequência de

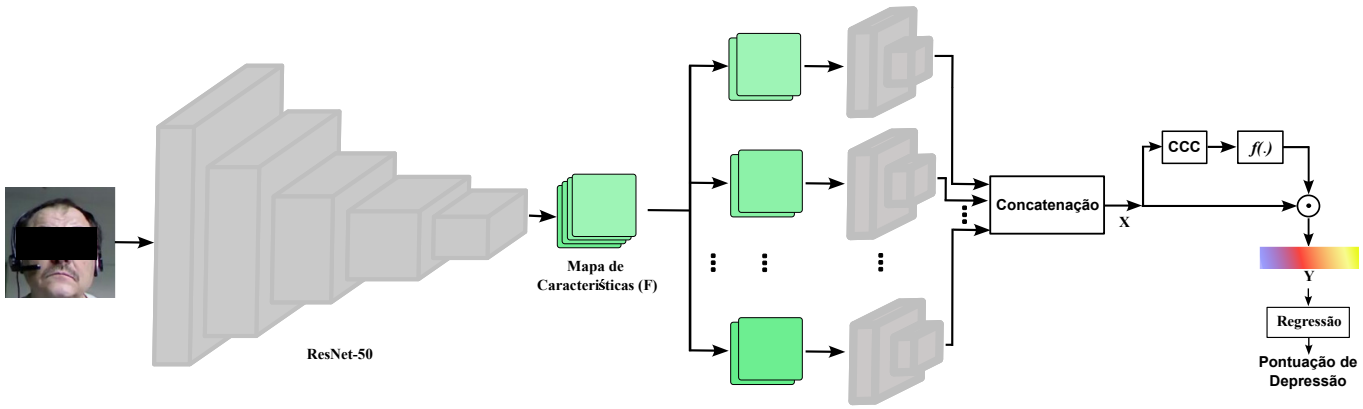

Fig. 1. Uma visão geral do sistema de detecção de depressão onde a técnica de atenção proposta é inserida. O mapa de características da última camada convolucional do modelo ResNet-50 é a entrada da técnica proposta. A saída da técnica é passada para a etapa de regressão, a qual gera uma pontuação de depressão. CCC refere-se à Camada Completamente Conectada. A função *sigmoid* é representada por f.

múltiplos pares constituídos por uma camada convolucional e uma camada de agrupamento. Na camada convolucional, os filtros, também denominados *kernels*, percorrem a entrada realizando o produto interno. Os parâmetros dos filtros são normalmente inicializados de maneira aleatória e otimizados durante o processo de treinamento. Após a operação de convolução, uma transformação não-linear é aplicada para ativação dos neurônios, e.g., a função *Rectified Linear Unit* (ReLU). A saída de uma camada convolucional é denominada mapa de características. Esse mapa pode ser visto como um vetor cuja os elementos contém diferentes características aprendidas. A camada de agrupamento é responsável pela redução da dimensionalidade dos elementos do mapa de características. Para realizar essa tarefa, pode-se por exemplo utilizar o valor máximo das regiões do elemento do mapa. Para gerar uma pontuação de depressão, uma rede pode empregar, após a última camada de convolucional, um método de agrupamento baseado na média de todos os valores do elemento do mapa, também conhecido como *Global Average Pooling* (GAP), em cascata com uma camada totalmente conectada. Neste artigo, substitui-se esse método de agrupamento pela técnica de atenção proposta.

Suponha que a saída da última camada convolucional de um modelo seja o mapa de características $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$, onde H e W denotam as dimensões espaciais e C é o número de canais. Esse mapa pode ser visto como um vetor definido por:

$$\mathbf{F} = [\mathbf{F_1}, \mathbf{F_2}, ..., \mathbf{F_C}] \tag{1}$$

onde o número total de elementos no vetor $\mathbf{F}$ é igual ao número de canais (ou seja, é igual a C). O primeiro passo da técnica proposta é dividir o vetor $\mathbf{F}$ em $C/2$ novos vetores, formados pela concatenação de dois consecutivos elementos de $\mathbf{F}$. Por exemplo, um novo vetor seria formado pela junção de $\mathbf{F_1}$ com $\mathbf{F_2}$.

Na sequência, cada novo vetor é enviado para uma sequência de duas camadas convolucionais. Esse processo pode ser matematicamente definido por:

$$\mathbf{X_i} = h_i([\mathbf{F_i}, \mathbf{F_{i+1}}]; \theta_i) \tag{2}$$

onde h_i denota a i-ésima sequência de camadas convolucionais e θ_i são os parametros da sequência. Esse processo permite a direta exploração das correlações de dois canais adjacentes.

Na última etapa da técnica proposta, concatenam-se os resultados da etapa anterior, gerando, assim, um vetor $\mathbf{X}$. Então, emprega-se esse resultado para calcular os coeficientes que irão ponderar os elementos de próprio vetor $\mathbf{X}$. Matematicamente, os coeficientes são calculados pela seguinte equação:

$$\alpha_\mathbf{i} = f(\mathbf{X}, \mathbf{W_i}) \tag{3}$$

onde $\mathbf{W_i}$ denota os pesos do i-ésimo neurônio e f é a função *sigmoid*. A saída final da técnica é definada por:

$$\mathbf{Y} = \mathbf{X} \odot [\alpha_1, \alpha_2, ..., \alpha_n] \tag{4}$$

onde $\odot$ representa multiplicação elemento a elemento. Essa operação permite dar mais atenção para um elemento de $\mathbf{X}$. Consequentemente, pode-se atribuir um grau maior de importância para alguns canais provenientes da última camada convolucional de um modelo. Especificamente, a técnica proposta é inserida no modelo ResNet-50. Para gerar a pontuação final de depressão, utiliza-se o vetor $\mathbf{Y}$ como entrada da etapa de regressão, a qual é composta por uma camada totalmente conectada. Na Figura 1, ilustra-se o sistema de detecção de depressão com a técnica proposta inserida.

IV. Procedimento Experimental

A. Base de Dados de Depressão

O desempenho da abordagem proposta é avaliado utilizando-se a base de dados de depressão denominada AVEC 2014 [26]. Essa base contém um total de 300 vídeos que possuem duração entre 6 e 248 segundos. A base é dividida em três partições: treinamento, desenvolvimento e teste. Cada partição contém 100 vídeos. A taxa de quadros dos vídeos é igual a 30 quadros por segundo com resolução de 640×480. Cada vídeo é rotulado com um pontuação de depressão de acordo com o inventário BDI. O valor mínimo da pontuação é 0 e o valor máximo é 63.

TABELA I
COMPARAÇÃO DO MODELO RESNET-50 COM A TÉCNICA DE ATENÇÃO PROPOSTA FRENTE A OUTRAS DUAS ESTRATÉGIAS.

Método	MAE	RMSE
Extrator de características	8.62	10.54
GAP+CCC	6.37	8.71
Técnica proposta	**6.13**	**8.05**

B. Detalhes do Treinamento e Métricas de Avaliação

Como a base AVEC 2014 possui uma pequena quantidade de dados, o sistema proposto é primeiro pré-treinado utilizando-se a base de dados denominada VGGFace2 [41]. Após esse processo, o aprendizado é transferido utilizando-se a técnica de *fine-tuning*. Para isso, emprega-se o algoritmo de otimização Adam [42] com uma taxa de aprendizado inicial igual a 0.001 e decaimento da taxa para cada *epoch*. Além disso, durante o treinamento, a função L2 é empregada como função de perda.

As métricas de avaliação empregadas neste artigo são *Mean Absolute Error* (MAE) e *Root Mean Square Error* (RMSE). Tais métricas são comumente empregadas, o que permite uma direta comparação com o estado da arte presente na literatura.

C. Resultados e Discussão

1) Desempenho do modelo ResNet-50 com a técnica proposta: Para demonstrar a capacidade da técnica de atenção proposta, geram-se resultados utilizando o modelo ResNet-50 em três diferentes casos: como extrator de características, com o uso do método GAP seguido por uma Camada Completamente Conectada (CCC), e com a técnica proposta. No primeiro, o modelo ResNet-50 é primeiramente treinado usando a base de dados VGGFace2. Depois, os pesos da etapa de extração de características são congelados e então empregados na detecção de depressão. No segundo caso, o modelo é primeiro treinado empregando a base de dados VGGFace2 e então usa-se o processo de *fine-tuning* para o aprendizado de características de depressão. Nos dois primeiros casos, após a saída da última camada de convolução do modelo, usa-se o método GAP seguido da etapa de regressão, a qual é uma CCC. No último caso, utiliza-se o modelo com a técnica de atenção proposta. Na Tabela I, apresentam-se os resultados gerados para os três casos. Como pode ser observado, a utilização do modelo como extrator de características gera os piores resultados. O motivo é que o modelo está utilizando características relacionadas ao reconhecimento de faces para detecção de depressão. Ao realizar o processo de *fine-tuning*, o modelo passa a detectar e extrair padrões relacionados à depressão, levando a um desempenho melhor em comparação com o caso em que se usa o modelo como extrator de características. Os melhores resultados são alcançados quando emprega-se o modelo com a técnica de atenção proposta. Tais resultados demonstram que a técnica de atenção facilita a detecção e a extração de padrões de depressão a partir das expressões faciais.

2) Comparação com o estado da arte: Na Tabela II, compara-se os resultados da abordagem proposta com os resultados de diversos métodos para detecção de depressão.

TABELA II
COMPARAÇÃO DA ABORDAGEM PROPOSTA FRENTE AO ESTADO DA ARTE PARA DETECÇÃO DE DEPRESSÃO.

Método	MAE	RMSE
LGBP-TOP (Valstar *et al*. [26])	8.86	10.86
MHH (Jan *et al*. [24])	8.44	10.50
LGBP-TOP+LPQ (Kaya *et al*. [27])	8.20	10.27
DTL (Kang *et al*. [29])	7.74	9.43
Two-stream network (Zhu *et al*. [37])	7.47	9.55
2×C3D (Jazaery *et al*. [34])	7.22	9.20
VLDN+CNN (Uddin *et al*. [39])	6.86	8.78
DJ-LDML (Zhou *et al*. [36])	6.59	8.30
2×C3D (de Melo *et al*. [35])	6.59	8.31
ResNet-50+*pooling* (Zhou *et al*. [33])	6.37	8.43
4×ResNet-50 (Zhou *et al*. [31])	6.21	8.39
Two-stream network (Chen *et al*. [38])	6.16	8.13
ResNet-50+DL (de Melo *et al*. [32])	6.15	8.23
MDN (de Melo *et al*. [40])	**6.06**	**7.65**
ResNet-50+técnica proposta	<u>6.13</u>	<u>8.05</u>

O melhor resultado de cada coluna é destacado em negrito, e o segundo melhor resultado é sublinhado.

Como esperado, o modelo ResNet-50 em conjunto com a técnica de atenção proposta obtém melhores resultados que os métodos baseados em descritores tradicionais [24], [26], [27]. Os métodos em [29], [31], [32], [33] empregam 2D CNNs em cascata com um esquema para melhorar a exploração da informação de aparência ou para capturar a informação temporal. Por exemplo, os autores em [31] utilizam quatro modelos ResNet-50 para explorar múltiplas regiões faciais. Pode-se observar que a união do modelo ResNet-50 com a técnica de atenção proposta gera melhores resultados que esses métodos. Esse fato destaca a capacidade da técnica de atenção proposta de aumentar o desempenho do modelo ResNet-50. Os métodos em [34], [35], [36], [37], [38], [39] exploram diretamente a informação espacial e temporal. Por exemplo, o método em [34] emprega dois modelos C3D para gerar características espaço-temporais a partir de duas sequências de imagens faciais em diferentes escalas. Os valores de erro (i.e., MAE e RMSE) gerados pelo modelo ResNet-50 com a técnica de atenção proposta são inferiores aos gerados por esses métodos. Isso demonstra a habilidade da abordagem proposta em gerar representações discriminativas pois a combinação do modelo ResNet-50 com a técnica de atenção proposta somente explora a informação espacial. Finalmente, a abordagem proposta alcança resultados competitivos em relação ao modelo MDN [40], o qual explora a informação espacial e temporal em múltiplas escalas. A grande vantagem do modelo ResNet-50 em conjunto com a técnica proposta em relação ao modelo MDN é a sua complexidade computacional. O modelo MDN requer 52M paramêtros e 107.72G *Floating Point Operations* (FLOPs), enquanto que a abordagem proposta requer 23.8M parâmetros e 3.9G FLOPs.

V. CONCLUSÕES

Os sistemas de detecção de depressão a partir de expressões faciais tem o potencial de serem uma importante ferramenta de auxílio no diagnóstico clínico. Neste artigo, propôs-se um avanço nesses sistemas através da introdução de uma técnica de atenção inter-canais para facilitar a geração de

representações discriminativas por um modelo de aprendizado profundo. A técnica explora as correlações de elementos adjacentes presentes em um mapa de características e atribui mais importância para elementos mais informativos. O modelo ResNet-50 é escolhido para gerar o mapa de características a ser explorado pela técnica de atenção. Resultados experimentais mostram que a técnica de atenção promove uma melhora na geração de representações pelo modelo ResNet-50 e que essa combinação alcança um desempenho competitivo em relação ao estado da arte, com a vantagem de possuir uma complexidade computacional inferior.

AGRADECIMENTOS

Parte dos resultados apresentados neste trabalho foram patrocinados pela ENVISION Indústria de Produtos Eletrônicos LTDA, nos termos da Lei Federal Brasileira nº 8.387/91 (SUFRAMA). Esta pesquisa foi conduzida em parceria com a UFAM/CETELI, UEA e a Envision (Grupo TPV).

REFERÊNCIAS

[1] M. Marcus, M. T. Yasamy, M. van Ommeren, D. Chisholm, and S. Saxena, "Depression: A global public health concern," 2012.

[2] A. Pampouchidou et al., "Automatic assessment of depression based on visual cues: A systematic review," in *IEEE Transactions on Affective Computing*, vol. 10, pp. 445-470, 2019.

[3] M. Murphy and M. Peterson, "Sleep disturbances in depression," in *Sleep Med. Clin.*, vol. 10, pp. 17-23, 2015.

[4] S. Borgman, I. Ericsson, E. K. Clausson and P. Garmy, "The relationship between reported pain and depressive symptoms among adolescents," in *J. Sch. Nurs.*, vol. 36, pp. 87-93, 2018.

[5] J. P. Lépine and M. Briley, "The epidemiology of pain in depression," in *Hum. Psychopharmacol.*, vol. 19, no. S1, pp. S3-S7, 2004.

[6] M. Jansson-Fr and K. Lindblom, "A bidirectional relationship between anxiety and depression and insomnia? A prospective study in the general population," in *J. Psychosomatic Res.*, vol. 64, pp. 443-449, 2008.

[7] P. Vidal-Ribas, and A. Stringaris, "How and why are irritability and depression linked?," in *Child Adolesc. Psychiatr. Clin. N. Am.*, vol. 30, pp. 401-414, 2021.

[8] A. T. Beck, and B. A. Alford, "Depression: Causes and treatment," *University of Pennsylvania Press*, 2009.

[9] Z. A. E. Sarhan, H. A. E. Shinnawy, M. E. Eltawil, Y. Elnawawy, W. Rashad and M. S. Mohammed, "Global functioning and suicide risk in patients with depression and comorbid borderline personality disorder," in *Neurol. Psychiatry Brain Res.*, vol. 31, pp. 37-42, 2019.

[10] J. L. Sotelo and C. B. Nemeroff, "Depression as a systemic disease," in *Personalized Med. Psychiatry*, vol. 1, pp. 11-25, 2017.

[11] C. U. Correll, J. Detraux, J. De Lepeleire, and M. De Hert, "Effects of antipsychotics, antidepressants and mood stabilizers on risk for physical diseases in people with schizophrenia, depression and bipolar disorder," in *World Psychiatry*, vol. 14, pp. 119-136, 2015.

[12] A. J. Mitchell, A. Vaze and S. Rao, "Clinical diagnosis of depression in primary care: A meta-analysis," in *Lancet*, vol. 374, pp. 609-619, 2009.

[13] M. J. Bostwick, "Recognizing mimics of depression: The '8 Ds'," in *Curr. Psychiatry*, vol. 11, pp. 31-36, 2012.

[14] A. T. Beck, R. A. Steer, and G. K. Brown, "Manual for the beck depression inventory-II." The Psychological Corporation, 1996.

[15] M. B. Sariyildiz, Y. Kalantidis, D. Larlus and K. Alahari, "Concept generalization in visual representation learning," in *IEEE/CVF International Conference on Computer Vision*, pp. 9609-9619, 2021.

[16] Z. Zou, K. Chen, Z. Shi, Y. Guo and J. Ye, "Object detection in 20 years: A survey," in *Proc. of the IEEE*, vol. 111, no. 3, pp. 257-276, 2023.

[17] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 200-210, 2020.

[18] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati and D. P. Rosenwald, "Social risk and depression: Evidence from manual and automatic facial expression analysis," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pp. 1-8, 2013.

[19] L. Nanni, S. Ghidoni and S. Brahnam, "Handcrafted vs. non-handcrafted features for computer vision classification," in *Pattern Recognition*, vol. 71, pp. 158-172, 2017.

[20] T. Ojala, M. Pietikainen and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 971-987, 2002.

[21] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *International Conference Image and Signal Processing*, pp. 236-243, 2008.

[22] K. Levi and Y. Weiss, "Learning object detection from a small number of examples: the importance of good features," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

[23] T. R. Almaev and M. F. Valstar, "Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 356-361, 2013.

[24] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh, "Automatic depression scale prediction using facial expression dynamics and regression," in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge*, pp. 73–80, 2014.

[25] M. F. Valstar *et al*., "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emot. Challenge*, pp. 3–10, 2013.

[26] M. Valstar *et al*., "AVEC 2014: 3D dimensional affect and depression recognition challenge," in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge*, pp. 3–10, 2014.

[27] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble CCA for continuous emotion prediction," in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge*, pp. 19–26, 2014.

[28] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning," in *Nature*, vol. 521, pp. 436-444, 2015.

[29] Y. Kang, X. Jiang, Y. Yin, Y. Shang and X. Zhou, "Deep transformation learning for depression diagnosis from facial images," in *Proc. Chinese Conf. Biometric Recognit.*, pp. 13-22, 2017.

[30] L. He, J. C. W. Chan, Z. Wang, "Automatic depression recognition using cnn with attention mechanism from videos," in *Neurocomputing*, vol. 422, pp. 165–175, 2021.

[31] X. Zhou, K. Jin, Y. Shang, G. Guo, "Visually interpretable representation learning for depression recognition from facial images," in *IEEE Transactions on Affective Computing*, vol. 11, pp. 542–552, 2020.

[32] W. C. de Melo, E. Granger and A. Hadid, "Depression Detection Based on Deep Distribution Learning," in *IEEE International Conference on Image Processing*, pp. 4544-4548, 2019.

[33] X. Zhou, P. Huang, H. Liu, S. Niu, "Learning content-adaptive feature pooling for facial depression recognition in videos," in *Electronics Letters*, vol. 55, pp. 648–650, 2019.

[34] M. Al Jazaery and G. Guo, "Video-Based Depression Level Analysis by Encoding Deep Spatiotemporal Features," in *IEEE Transactions on Affective Computing*, vol. 12, pp. 262-268, 2021.

[35] W. C. de Melo, E. Granger and A. Hadid, "Combining global and local convolutional 3D networks for detecting depression from facial expressions," in *14th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 1-8, 2019.

[36] X. Zhou, Z. Wei, M. Xu, S. Qu and G. Guo, "Facial depression recognition by deep joint label distribution and metric learning," in *IEEE Transactions on Affective Computing*, vol. 13, pp. 1605-1618, 2022.

[37] Y. Zhu, Y. Shang, Z. Shao and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," in *IEEE Transactions on Affective Computing*, vol. 9, pp. 578-584, 2018.

[38] Q. Chen, I. Chaturvedi, S. Ji, and E. Cambria, "Sequential fusion of facial appearance and dynamics for depression recognition," in *Pattern Recognition Letters*,"vol. 150, pp. 115–121, 2012.

[39] M. A. Uddin, J. B. Joolee, Y. K. Lee, "Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm," in *IEEE Transactions on Affective Computing*, vol. 13, pp. 864–870, 2022.

[40] W. C. de Melo, E. Granger, M. B. López, "Mdn: A deep maximization-differentiation network for spatio-temporal depression detection," in *IEEE Transactions on Affective Computing*, vol. 14, pp. 578–590, 2023.

[41] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 67–74, 2018.

[42] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *International Conference on Learning Representations*, 2015.