Detecção Automática da Depressão Assistida por *Stacking DNNs* em Dados de Descritores de Características Visuais

Filipe F. de Almeida[1], **André C. B. Soares**[2], **Laurindo de S. B. Neto**[2], **Kelson R. T. Aires**[2]

[1]Departamento de Computação – Universidade Federal do Maranhão (UFMA)
CEP: 65080-805 – São Luís – MA – Brasil

[2]Departamento de Computação – Universidade Federal do Piauí (UFPI)
CEP: 64049-550 – Teresina – PI – Brasil

filipefontineli@gmail.com, {andre.soares, laurindoneto, kelson}@ufpi.edu.br

Abstract. *People experience more and more feelings of anguish, anxiety and sadness. These point, among other pathologies, to depression and, worse, thoughts of suicidal ideation. That said, computational techniques capable of identifying this disorder early become indispensable. The present work presents a Stacking Deep Neural Networks model for analysis of facial expressions and subsequent automatic detection of depression. The results obtained indicate a promising advance regarding the automatic detection of depression. The Stacking DNNs model achieves, in the test base, 78.5% recall and 62.8% F1-Score. Such values are 22% and 17% higher, respectively, than unimodal models that apply similar methods.*

Resumo. *Pessoas vivenciam cada vez mais sentimentos de angústia, ansiedade e tristeza. Esses apontam, entre outras patologias, à depressão e, pior, pensamentos de ideação suicida. Posto isso, técnicas computacionais capazes de apontar tal transtorno precocemente se tornam indispensáveis. O presente trabalho apresenta um modelo baseado em Stacking Deep Neural Networks para análise de expressões faciais e subsequente detecção automática da depressão. Os resultados obtidos indicam um avanço promissor quanto à detecção automática da depressão. O modelo Stacking DNNs atinge, na base de teste, 78,5% de Recall e 62,8% de F1-Score. Tais valores são 22% e 17% superiores, respectivamente, a modelos unimodais que aplicam métodos semelhantes.*

1. Introdução

A depressão é uma patologia psicossocial que produz alteração do humor caracterizada por tristeza profunda e forte sentimento de desesperança [Farias et al. 2020]. Em estágios avançados, pode tornar os pacientes propícios à ideação suicida, o que eleva a importância em definir métodos de detecção com diagnósticos precisos. Contudo, essa patologia segue negligenciada [WHO 2022].

Transtornos mentais são mais difíceis de detectar, se comparados a doenças físicas [Taschereau-Dumouchel et al. 2022]. Métodos convencionais como questionários estruturados são frequentemente aplicados em pacientes de modo a avaliar a gravidade dos sintomas depressivos. À vista disso, um procedimento bastante conhecido é o *Personal*

Health Questionnaire Depression Scale (PHQ-8) [Kroenke et al. 2009]. Esse, clinicamente validado pela comunidade científica, mede a gravidade dos sintomas de depressão em vários aspectos pessoais.

Especificamente, a depressão é uma forma de sofrimento psíquico que pode ser prevista por aspectos não verbais, como [Morales 2018]: ângulos mais descendentes da cabeça e do olhar; olhar distante; menos contato visual com outras pessoas; sorrisos menos intensos e durações médias mais curtas de sorriso; menos aceno de cabeça. As emoções podem ser efetivamente avaliadas pelo computador por meio do uso de descritores de expressões faciais. Um exemplo disso é o dicionário *Facial Action Coding System* (FACS) [Cohn et al. 2007]. Esse, modelado por uma equipe de psicólogos, combina um conjunto de movimentos musculares faciais que correspondem a uma emoção exibida.

Tal instrumento pode ser usado como uma ferramenta descritiva que detecta o estado emocional de uma pessoa por meio de análise facial. Nos últimos anos, novos trabalhos despontaram com o propósito de detectar a depressão automaticamente por meio de expressões faciais, áudios ou transcrições textuais. Isso se tornou mais evidente com o surgimento do conjunto de dados *Distress Analysis Interview Corpus – Wizard of Oz* (DAIC-WOZ) [Gratch et al. 2014]. Perante o exposto, o modelo elaborado por este estudo decorre à inspiração em metodologias semelhantes e que de alguma maneira agregaram valor à proposta, como os estudos citados a seguir.

Os trabalhos de Nasir *et al.* [Nasir et al. 2016] e Song *et al.* [Song et al. 2018] propuseram a detecção automática da depressão por meio das técnicas *Stochastic Gradient Descent, Support Vector Machine* (SGD-SVM) e *Convolutional Neural Network* (CNN), respectivamente. Ambos os trabalhos implementaram seus métodos com as características de movimentos de cabeça e olhar. Porém, o primeiro inclui os pontos de referências faciais e o segundo, as unidades de ação facial. Em comum, são relatadas a alta dimensionalidade do *dataset* junto aos dados de treinamento limitado.

Já o trabalho de Wang *et al.* [Wang et al. 2020] propôs a detecção automática da depressão usando pontos de referência de expressões faciais. Esse método implementa uma rede *Long Short-Term Memory* (LSTM) e *global max pooling* para uma camada de *pool* com várias instâncias. Tal camada identifica as instâncias que indicam sintomas de depressão.

O estudo de Akbar *et al.* [Akbar et al. 2021] definiu um conjunto reduzido de recursos de unidades de ação facial utilizando *Particle Swarm Optmization* (PSO) para selecionar os melhores preditores e alimentá-los para redes neurais *feedforward* padrão otimizadas. Esse estudo analisou o comportamento facial para reconhecimento da depressão a partir de unidades de ação facial extraídas de uma sequência de *frames*.

Foi proposto por de Melo *et al.* [de Melo et al. 2021] uma arquitetura para explorar as variações da expressão facial em diferentes escalas temporais. Esse modelo de *Deep Learning* é composto por um bloco de maximização e um bloco de diferença. Dada uma entrada (sequência de imagens ou *frames*, exibidas ao longo de um tempo específico), o bloco de maximização é empregado para capturar transições suaves de estruturas faciais, enquanto o bloco de diferença codifica variações espaço-temporais repentinas. Esses blocos não dependem de filtros 3D e as características geradas são combinadas de uma forma que leva a uma representação de características robusta para detecção de depressão.

O trabalho de Guo *et al.* [Guo et al. 2022] propôs um método de detecção automática da depressão baseado em representação visual, especificamente, por pontos de referência facial 2D e movimentos da cabeça. O método possui dois módulos principais: *Temporal Dilated Convolution Network* (TDCN) e *Feature-Wise Attention* (FWA). O primeiro, extrai informações específicas da depressão completando as convoluções. O segundo, busca aumentar a capacidade de representação das características, atribuindo diferentes pesos aos canais destes.

O modelo proposto neste trabalho, diferentemente dos demais, oferece as perspectivas dos movimentos da cabeça, direções e ângulos do olhar, isto é, provém apenas esses tipos de aspectos não verbais. Com isso, foi implementado um modelo unimodal que gera uma arquitetura de menor complexidade, porém robusta, ao passo que atrela ao ambiente um conjunto de *Deep Neural Networks* (DNNs) por *Ensemble Stacking* [Moon et al. 2020]. Além disso, o trabalho se esforçou em reduzir possíveis vieses, desbalanceamentos, variâncias excessivas e alta dimensionalidade no conjunto de dados utilizado nesta pesquisa. Esse processo refletiu positivamente nos resultados apresentados na Seção 4. Antes, porém, a Seção 2 apresenta a metodologia aplicada por essa pesquisa. Em seguida, a Seção 3 explana os experimentos realizados. Por fim, a Seção 5 destaca a conclusão deste trabalho.

2. Materiais e método proposto

Propõe-se, neste trabalho, uma arquitetura guiada pela colaboração entre várias DNNs. Tal processo se tornou possível pelo emprego do método *Ensemble Machine Learning* (EML), especificamente pela técnica *Stacking*. A seguir, serão destacados, além do conjunto de dados utilizado, a explanação do modelo proposto. Esse possui quatro fases: pré-processamento de dados; fusão; Redes DNNs e *Stacking* DNNs. Tais informações estão descritas a seguir.

2.1. Dataset DAIC-WOZ

Aplicou-se neste modelo os dados do DAIC-WOZ, lançado como parte do Desafio de Emoções Audiovisuais de 2016 (AVEC2016) [Valstar et al. 2016]. O conjunto de dados foi projetado com o intuito de automatizar o diagnóstico de sofrimento psíquico como ansiedade, transtorno de estresse pós-traumático e depressão. Além disso, contém entrevistas semi-clínicas de 189 participantes.

As entrevistas foram divididas em três conjuntos principais: treinamento, validação e teste. O conjunto de treinamento possui 107 vídeos, no qual há 77 amostras de participantes não depressivos e 30 classificados como depressivos. Já no conjunto de validação, há 35 vídeos, 23 amostras de participantes não depressivos e 12 depressivos. Por fim, no conjunto de teste, existem 47 vídeos, 33 amostras de participantes não depressivos e 14 depressivos. A gravidade da depressão dos participantes foi validada através da escala PHQ-8, que forneceu valores binários de depressão ou não depressão para cada participante.

Nenhum arquivo de vídeo bruto é disponibilizado publicamente. Portanto, só é possível aproveitar as características visuais extraídas pelo provedor do *dataset*. O AVEC2016 forneceu características visuais de baixo nível, juntamente com características funcionais usando o *framework* OpenFace, este, com uma implementação Python livre e de código aberto para reconhecimento facial com DNNs [Baltrušaitis et al. 2016].

2.2. Pré-processamento dos dados e fusão

Os dados do DAIC-WOZ, como mencionados na subseção acima, foram disponibilizados em três conjuntos: treinamento, validação e teste. Para cada participante, foram gerados os dados (descritores visuais de baixo nível) com vetores de características para cada *frame*. As características de movimento de cabeça contêm 6 dimensões, incluindo Tx, Ty, Tz, Rx, Ry e Rz, em que Tx, Ty, Tz são as coordenadas de posição da cabeça, e Rx, Ry, Rz são as coordenadas de rotação da cabeça. As características do olhar registram as direções e ângulos do olhar e possuem 12 dimensões.

Processou-se os dados, levando em consideração que cada vetor de características, com os 18 valores mencionados acima, é um *frame* capturado. Assim, concatenou-se os dados (fusão) dos 12.447 primeiros *frames* (aproximadamente 7 minutos) de cada participante. Com isso, gerou-se novos vetores com 18 x 12.447 dados de cada paciente envolvido. Cada modelo de rede DNN, sequencial e densa, recebeu como entrada 18 valores dos movimentos de cabeça, direção e ângulos do olhar de cada *frame* selecionado.

2.3. Redes DNNs

As Redes Neurais Artificiais (RNAs) são inspiradas pela maneira como os neurônios biológicos enviam sinais uns para os outros. São compostas por neurônios artificiais e as seguintes camadas: dados de entrada (camada de entrada); uma ou mais camadas ocultas e uma camada de saída. Cada nó (neurônio artificial), nestas camadas, conecta-se ao outro nó da camada seguinte e possui um peso e um limite associados.

Quando RNAs possuem duas ou mais camadas ocultas entre a camada de entrada e a camada de saída, trata-se de DNN [Durstewitz et al. 2019]. Caso contrário, é apontada como RNA simples ou rasa. A Figura 1 apresenta essa diferença.

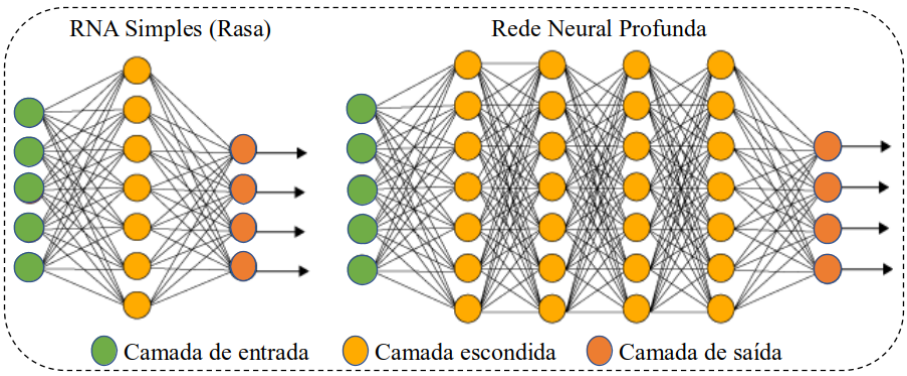

Figura 1. Diferença entre RNA simples e DNN

Foram implementadas à arquitetura 4 DNNs, cada qual com suas especificações. A Tabela 1 apresenta as características individuais das redes. Cada DNN possui 6 camadas escondidas, essas distribuídas com 32, 64 e 128 neurônios, respectivamente. Além disso, foram inseridas camadas de Dropout entre as 4 e 5 primeiras camadas escondidas, essas configuradas com taxas de 0,4. Dropout é uma técnica de regularização que busca prevenir o *overfitting*. As camadas escondidas foram implementadas com a função de ativação LeakyRelu, ELU ou ambas. O otimizador aplicado foi o Nadam. A arquitetura com as 4 DNNs implementadas é representada na Figura 2.

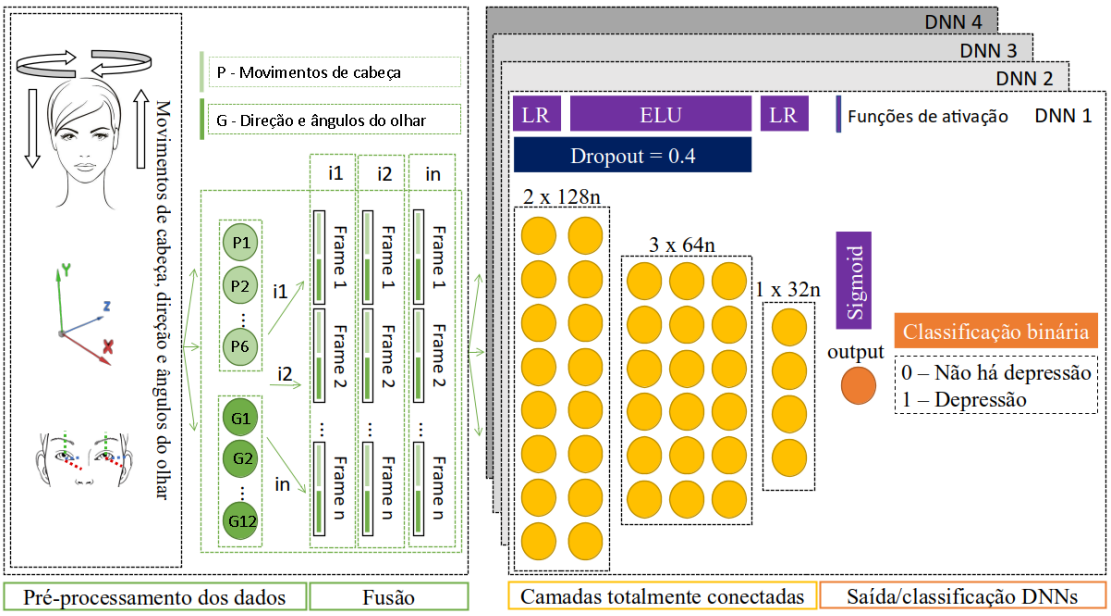

Figura 2. Arquitetura desenvolvida com quatro DNNs

Dada a natureza do problema, no qual é requerido ao modelo uma correta classificação entre pacientes com depressão e sem depressão, ou seja, uma classificação binária, a função de custo *binary crossentropy* foi definida. Esta calcula a perda de entropia cruzada entre a saída verdadeira e a saída prevista, isto é, busca contribuir para a minimização do erro durante o treinamento.

No treinamento das redes DNNs foram inseridos valores de época e *batch size* como 1.000 e 64, respectivamente. Todavia, a técnica *EarlyStopping* foi aplicada ao processo. Essa permite que o treinamento cesse quando uma métrica monitorada "para de melhorar" [Li et al. 2020]. O valor da acurácia na base de validação foi usado como parâmetro de parada. Além disso, foi implementado um método em que automatizou as execuções do treinamento conforme a acurácia definida no *EarlyStopping*. Como mencionado, este trabalho reflete a um problema de classificação binária, logo, implementou-se no modelo o sistema com apenas uma saída e a função de ativação Sigmoide.

Tabela 1. Especificações das DNNs implementadas

Modelo	Função de ativação	Camadas Escondidas	Dropout
DNN1	Leaky Relu e ELU	2x128, 3x64, 1x32	0,4 (5 primeiras camadas)
DNN2	ELU	3x128, 2x64, 1x32	0,4 (5 primeiras camadas)
DNN3	Leaky Relu	1x128, 3x64, 2x32	0,4 (4 primeiras camadas)
DNN4	Leaky Relu e ELU	2x128, 3x64, 1x32	0,4 (4 primeiras camadas)

Em todo desenvolvimento, incluindo implementação e testes, o serviço de nuvem gratuito Google Colab [Colab 2022] foi utilizado. Tal serviço contribuiu com a pesquisa ao embarcar os recursos de software e hardware necessários, como: infraestruturas das bibliotecas de software; *Graphics Processing Unit* (GPU); e memória RAM sob demanda.

2.4. *Stacking* DNN

As DNNs, embora alcancem resultados satisfatórios na base de validação, especialmente quando se aplica em conjunto as funções de ativação LeakyRelu e ELU, na base de testes não atenderam às expectativas. Diante disso, optou-se por aplicar o método *Stacking* junto às DNNs.

Dentre vários modelos conhecidos de EML, há o *Bagging*, *Boosting* e o *Stacking*. O último difere principalmente dos demais em dois aspectos. Enquanto o *Stacking* geralmente considera modelos "fracos" heterogêneos (diferentes algoritmos de aprendizado são combinados), o *Bagging* e o *Boosting* consideram principalmente modelos "fracos" homogêneos. O segundo aspecto remete ao fato de que o *Stacking* aprende a combinar os modelos básicos usando um meta-modelo, enquanto o *Bagging* e o *Boosting* os combinam seguindo algoritmos determinísticos [Mienye and Sun 2022].

Especificamente, este trabalho implementa a técnica *Stacking* com uma série de DNNs como modelos "fortes". Cada DNN produz um resultado específico e, devido à complementaridade das redes, os resultados são mais robustos. Vários modelos são treinados usando técnicas e hiperparâmetros específicos para minimizar a taxa de erro e, assim, melhorar o desempenho geral do modelo. A Figura 3 apresenta a arquitetura de *Stacking* DNNs desenvolvida.

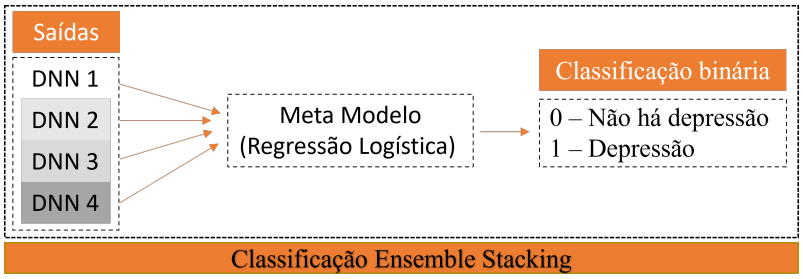

Figura 3. Arquitetura Stacking DNNs aplicada nesta proposta

Os resultados produzidos pelas DNNs foram incorporados como entrada para o meta-modelo que produziu a classificação final, isto é, apontou, de acordo com os dados de testes do conjunto de dados, os pacientes depressivos e não depressivos. Nos testes iniciais, a regressão logística foi a técnica de aprendizado de máquina definida para o meta-modelo.

A regressão logística é usada quando uma saída discreta é esperada, como a ocorrência de algum evento. Normalmente, esta técnica usa alguma função para apontar valores para um determinado intervalo. Uma função clássica utilizada nesse sentido é a "Sigmoide" que possui a curva em forma "S", utilizada para a classificação binária e discretizando os valores para 0 ou 1, por exemplo [Zou et al. 2019].

3. Experimentos

Durante o desenvolvimento do modelo, diversos testes foram realizados, resultando em novos *insights*. Destaca-se a importância de entender os dados fornecidos pelo *dataset*. Adicionalmente, o desempenho do modelo foi aprimorado com a aplicação das

técnicas SVM-SMOTE [Nguyen et al. 2011] e Principal Component Analysis (PCA) [Maćkiewicz and Ratajczak 1993], descritas a seguir.

O processo de reamostragem de dados SVM-SMOTE, técnica que combina o uso de Support Vector Machines (SVM) [Cortes and Vapnik 1995] e *Synthetic Minority Oversampling Technique* (SMOTE) [Chawla et al. 2002] para lidar com conjuntos de dados desequilibrados. O objetivo é aplicar o SMOTE para gerar amostras sintéticas para a classe minoritária e, em seguida, usá-las junto às amostras originais das classes minoritária e majoritária para treinar um classificador SVM [Ghorbani and Ghousi 2020].

Assim, foi possível avaliar a importância da análise de dados quanto aos vieses em desequilíbrio, ou seja, classes desbalanceadas para a representação do conhecimento em características visuais na base de dados DAIC-WOZ.

Já a técnica PCA consiste em identificar correlações e padrões em um conjunto de dados, para que ele possa ser transformado em um novo conjunto de dados de dimensionalidade significativamente menor sem a perda de nenhuma informação importante. É amplamente utilizada para diversas finalidades, como visualização de dados, extração de recursos e redução de ruído. Ademais, a redução de dimensionalidade ajuda a simplificar modelos, diminuir o tempo de treinamento e evitar *overfitting*.

Buscou-se diminuir o tempo de execução no treinamento devido ao grande volume de dados gerados no pré-processamento, o que levou muito tempo para ser concluído e devido à necessidade de realizar muitos testes.A partir das implementações baseadas nos experimentos propostos, isto é, aplicação do SVM-SMOTE e do PCA, foram realizados diversos testes e os resultados são apresentados na seção seguinte.

4. Resultados e discussões

Esta seção apresenta os resultados obtidos neste trabalho. Além disso, destaca uma análise comparativa entre o trabalho realizado e outros métodos unimodais de detecção automática da depressão de última geração.

4.1. Métricas de Avaliação

Modelos de classificação binária possuem como alvo apontar em qual classe uma nova observação pertence dentre duas classes possíveis. Geralmente, classe positiva (P) ou negativa (N). Quanto à aplicação neste trabalho, classifica-se um determinado participante ao apontar com depressão (P) ou sem depressão (N).

Uma métrica bem conhecida é a matriz de confusão, que indica quantos exemplos existem em cada grupo: falso positivo (FP), falso negativo (FN), verdadeiro positivo (VP) e verdadeiro negativo (VN). Com isso, é possível visualizar os exemplos classificados corretamente e erroneamente em cada classe. Logo, compreende-se a capacidade do modelo em favorecer uma classe em detrimento da outra.

Foram implementadas, neste trabalho, métricas baseadas na matriz de confusão que são frequentemente utilizadas em tarefas de classificação. Estas são explanadas a seguir:

- Acurácia: indica uma performance geral do modelo. Dentre todas as classificações, quantas o modelo classificou corretamente;

- Precisão: aponta a proporção de previsões P corretas em relação ao total de previsões P. Mede a capacidade do modelo de não classificar como P um exemplo N, evidenciando os FPs;
- Recall: é a proporção de previsões P corretas em relação ao total de exemplos positivos. Mede a capacidade do modelo de encontrar todos os exemplos positivos, evidenciando os FNs;
- F1-Score: média harmônica entre Precisão e Recall.

As equações que representam as métrica citadas são expostas abaixo:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{1}$$

$$Precisão = \frac{VP}{VP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = 2 * \frac{Precisão * Recall}{Precisão + Recall} \tag{4}$$

Posto isso, foram realizadas sequências de experimentos, no qual foram aplicadas, na base de treinamento, as técnicas PCA e SMOTE. Para análise comparativa, são apreciados três cenários:

- com redução de dimensionalidade e com a reamostragem dos dados (CRCS);
- sem a redução e com a reamostragem (SRCS);
- sem ambas as técnicas (SRSS).

Cada treinamento das DNNs gera resultados particulares, únicos a cada execução. Dessa forma, ao se utilizar o método *Ensemble Stacking DNNs* desenvolvido, as saídas de cada DNN foram incorporadas como entrada nesta estrutura. Essas saídas foram compiladas de modo a produzir um novo conjunto de dados no qual são apresentadas a um meta-modelo que por sua vez realiza a classificação final.

Todavia, o *Stacking* é flexível e pode ser usado com uma variedade de algoritmos de aprendizado de máquina. Como mencionado anteriormente, este permite a execução de diferentes combinações de modelos base e meta-modelos, o que pode levar a melhores resultados. Deste modo, ao longo do processo de execução de novos testes, foram incrementadas diversas técnicas ao meta-modelo. No qual se aplicou um processo de troca automática de meta-modelos, visando a avaliação qualitativa destes junto a *Stacking* DNNs.

Foram usadas, além da *Logistic Regression* (LR), as técnicas *Gaussian Naive Bayes* (GNB), SVM e *Multilayer perceptron* (MLP). Nos testes, todas as técnicas de aprendizado de máquina citadas acima foram aplicadas seguindo os parâmetros padrões estabelecidos pela biblioteca Scikit-learn [Pedregosa et al. 2011].

A Tabela 2 demonstra a eficiência da técnica *Stacking DNNs* desenvolvida. Foram considerados 4 meta-modelos. As taxas de Recall e F1-score são apresentadas de acordo com a execução de cada experimento, comparando os resultados na base de validação e teste.

Especificamente, o meta-modelo MLP com o experimento SRCS atingiu, respectivamente, 83% e 79% de Recall e 77% e 63% de F1-score nas bases de validação e teste. Já o meta-modelo LR com o experimento CRCS resultou em uma taxa de 83% e 71% de Recall e 87% e 61% de F1-score nas bases de validação e teste, respectivamente. Tais meta-modelos demonstraram superioridade maior em relação aos testes com GNB e SVM. Foi notado, ainda, que todos os meta-modelos testados junto ao experimento SRSS, obtiveram resultados abaixo do esperado nas bases testadas nesta pesquisa.

Tabela 2. Resultados das taxas de Recall e F1-Score de 4 meta-modelos Stacking DNNs. No qual (V) se refere a base de validação e (T), base de teste.

Meta Modelo	Recall						F1-Score					
	CRCS		SRCS		SRSS		CRCS		SRCS		SRSS	
	V	T	V	T	V	T	V	T	V	T	V	T
GNB	75%	50%	83%	71%	75%	36%	82%	42%	83%	59%	69%	36%
SVM	83%	57%	75%	43%	42%	21%	87%	46%	82%	46%	53%	26%
LR	**83%**	**71%**	83%	71%	50%	21%	**87%**	**61%**	83%	59%	60%	30%
MLP	83%	71%	**83%**	**79%**	58%	21%	83%	60%	**77%**	**63%**	67%	29%

Os resultados da Tabela 2 demonstram, portanto, a importância da análise dos dados e consequente implementação de procedimentos como o SVM-SMOTE e, não menos importante, a redução da dimensionalidade em bases de dados que apresentam classes desbalanceadas e com elevada variância nos dados.

Ao se comparar os resultados obtidos por esse trabalho com propostas relacionadas, foi possível notar um ganho significativo aos demais. A Tabela 3 apresenta vários métodos semelhantes no qual foram aplicados dados de Áudio (A), Características Visuais (V) e Texto (T). Além disso, foram utilizadas características específicas dos dados de V como Unidade de Ação Facial (AUs), Gaze (G), Pose (P) e Pontos de Referência Facial 2D (2DL) ou 3DL. A apreciação dos resultados se dá pela análise na base de validação. As métricas comparadas são Acurácia, Precisão, Recall e F1-score.

É possível notar que a partir do experimento CRCS, os métodos empregados nesta pesquisa, superam em Acurácia, Precisão e F1-Score (*Stacking DNNs*, junto ao meta-modelo LR), os métodos unimodais avaliados.

Tabela 3. Resultados obtidos a partir da execução na base de validação DAIC-WOZ.

Método	Características	Acurácia	Precisão	Recall	F1-Score
[Song et al. 2018]	AUs+G+P	–	63,6%	58,3%	60,9%
[Nasir et al. 2016]	3DL+G+P	–	56%	71%	63%
Stacking (SRSS) (LR)	G+P	77,1%	62,4%	83,3%	71,4%
[Haque et al. 2018]	A+T+3DL	–	71,4%	83,3%	76,9%
[Wei et al. 2022]	A + V + T	85,1%	89%	57%	70%
Stacking (SRCS) (MLP)	G+P	82,8%	71,4%	83,3%	76,9%
[Wang et al. 2020]	2DL	–	81,8%	75%	78,3%
[Guo et al. 2022]	2DL + P	85,7%	76,9%	83,3%	79,9%
Stacking (CRCS) (LR)	**G+P**	**91,4%**	**90,9%**	**83,3%**	**86,9%**

Embora os modelos multimodais processem uma quantidade maior de características avaliadas, transmitindo uma robustez maior ao ambiente, ainda possuem

limitações quanto à generalização pois expõem-se de maneira mais incisiva a vieses do conjunto de dados, alta dimensionalidade e variância.

De qualquer maneira, os métodos unimodais são igualmente relevantes à análise de detecção automática da depressão. Isso ocorre pois concentram-se apenas em uma ou mais características, como as visuais, por exemplo. Desta forma, potencializam as chances de um apontamento específico mais adequado.

Levando-se em consideração os resultados na base de teste, foi observado que o método *Stacking* SRCS (MLP) alcançou taxas de 72,3%, 52,3%, 78,5% e 62,8% em Acurácia, Precisão, Recall e F1-Score, respectivamente. Tal método representa o melhor desempenho alcançado em todos os testes executados por essa pesquisa. A Tabela 4 apresenta os resultados obtidos a partir da execução de diversos modelos na base de teste.

Tabela 4. Resultados obtidos a partir da execução dos modelos na base de teste DAIC-WOZ

Método	Característica	Acurácia	Precisão	Recall	F1-Score
[Guo et al. 2022]	2D Landmarks + P	66%	45%	64,3%	53%
Stacking SRSS (LR)	G+P	70,2%	49,9%	21,4%	29,9%
Stacking CRCS (MLP)	G+P	72,3%	52,6%	71,4%	60,6%
Stacking SRCS (MLP)	**G+P**	**72,3%**	**52,3%**	**78,5%**	**62,8%**

Os resultados demonstram a competência da *Ensemble Stacking* implementada a partir de um conjunto de DNNs. Contribuem, assim como a análise e consequente validação dos experimentos CRCS e SRCS, na diminuição da variância extrema, vieses embarcados ao conjunto de dados empregado e alta dimensionalidade. Além disso, os resultados evidenciam a capacidade da arquitetura proposta em classificar a depressão por características visuais como movimentos de cabeça e olhar.

5. Conclusão

Este trabalho propôs a detecção automática da depressão baseada em descritores de características visuais, essencialmente os movimentos de cabeça e olhar. A aplicação se baseou no método *ensemble machine learning* que inspirou a arquitetura *Stacking DNNs*.

Implementou-se métodos em que, a partir da análise de vieses no conjunto de dados DAIC-WOZ, suprimiram os efeitos nocivos como amostras desbalanceadas e com variância extrema. À vista disso, os experimentos e consequentes resultados apontaram uma melhora significativa ao aplicar tais métodos à arquitetura proposta.

Para trabalhos futuros, pretende-se realizar novas rodadas de testes com essa arquitetura em novos dados de características visuais, tal como pontos de referências faciais 2D e 3D. Junto a isso, pretende-se desenvolver um modelo multimodal, baseado em *Stacking*, que atenda as características de áudio, vídeo e transcrições de texto e de modo a comparar com o modelo unimodal já proposto. Por fim, há ideia também de aplicar esses modelos a outras patologias como ansiedade e transtorno de estresse pós-traumático.

Referências

Akbar, H., Dewi, S., Rozali, Y. A., Lunanta, L. P., Anwar, N., and Anwar, D. (2021). Exploiting facial action unit in video for recognizing depression using metaheuristic

and neural networks. In *2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)*, volume 1, pages 438–443. IEEE.

Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Cohn, J. F., Ambadar, Z., and Ekman, P. (2007). Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment*, 1(3):203–221.

Colab, G. (2022). Google colaboratory. url:https://colab.research.google.com/.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20:273–297.

de Melo, W. C., Granger, E., and Lopez, M. B. (2021). Mdn: A deep maximization-differentiation network for spatio-temporal depression detection. *IEEE Transactions on Affective Computing*.

Durstewitz, D., Koppe, G., and Meyer-Lindenberg, A. (2019). Deep neural networks in psychiatry. *Molecular psychiatry*, 24(11):1583–1598.

Farias, M., Gusmão, R., and Gusmão, C. (2020). Mineração de dados aplicada à saúde mental de estudantes universitários: Uma revisão sistemática. *Anais do XX Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 49–59.

Ghorbani, R. and Ghousi, R. (2020). Comparing different resampling methods in predicting students' performance using machine learning techniques. *IEEE Access*, 8:67899–67911.

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., et al. (2014). The distress analysis interview corpus of human and computer interviews. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES.

Guo, Y., Zhu, C., Hao, S., and Hong, R. (2022). Automatic depression detection via learning and fusing features from visual cues. *arXiv preprint arXiv:2203.00304*.

Haque, A., Guo, M., Miner, A. S., and Fei-Fei, L. (2018). Measuring depression symptom severity from spoken language and 3d facial expressions. *arXiv preprint arXiv:1811.08592*.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., and Mokdad, A. H. (2009). The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173.

Li, M., Soltanolkotabi, M., and Oymak, S. (2020). Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International conference on artificial intelligence and statistics*, pages 4313–4324. PMLR.

Maćkiewicz, A. and Ratajczak, W. (1993). Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342.

Mienye, I. D. and Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access*, 10:99129–99149.

Moon, J., Jung, S., Rew, J., Rho, S., and Hwang, E. (2020). Combination of short-term load forecasting models based on a stacking ensemble approach. *Energy and Buildings*, 216:109921.

Morales, M. R. (2018). *Multimodal depression detection: An investigation of features and fusion techniques for automated systems*. City University of New York.

Nasir, M., Jati, A., Shivakumar, P. G., Nallan Chakravarthula, S., and Georgiou, P. (2016). Multimodal and multiresolution depression detection from speech and facial landmark features. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 43–50.

Nguyen, H. M., Cooper, E. W., and Kamei, K. (2011). Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1):4–21.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Song, S., Shen, L., and Valstar, M. (2018). Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 158–165. IEEE.

Taschereau-Dumouchel, V., Michel, M., Lau, H., Hofmann, S. G., and LeDoux, J. E. (2022). Putting the "mental" back in "mental disorders": a perspective from research on fear and anxiety. *Molecular Psychiatry*, 27(3):1322–1330.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., and Pantic, M. (2016). Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10.

Wang, Y., Ma, J., Hao, B., Hu, P., Wang, X., Mei, J., and Li, S. (2020). Automatic depression detection via facial expressions using multiple instance learning. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1933–1936. IEEE.

Wei, P.-C., Peng, K., Roitberg, A., Yang, K., Zhang, J., and Stiefelhagen, R. (2022). Multi-modal depression estimation based on sub-attentional fusion. *arXiv preprint arXiv:2207.06180*.

WHO (2022). World health organization - depression. url:https://bityli.com/EqHbyv.

Zou, X., Hu, Y., Tian, Z., and Shen, K. (2019). Logistic regression model optimization and case analysis. In *2019 IEEE 7th international conference on computer science and network technology (ICCSNT)*, pages 135–139. IEEE.