PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ

LUCIANA TRINKAUS MENON

SELEÇÃO DINÂMICA DE MODALIDADE E REPRESENTAÇÃO DE DADOS PARA RECONHECIMENTO MULTIMODAL DE EMOÇÕES: UM ESTUDO SOBRE O IMPACTO DA AUSÊNCIA DE MODALIDADES

DOUTORADO EM
INFORMÁTICA
PUCPR

CURITIBA
2025

LUCIANA TRINKAUS MENON

Seleção Dinâmica de Modalidade e Representação de Dados para Reconhecimento Multimodal de Emoções: Um Estudo sobre o Impacto da Ausência de Modalidades

Tese apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de doutor em Informática.

Pontifícia Universidade Católica do Paraná - PUCPR

Programa de Pós-Graduação em Informática - PPGIa

Orientador: Prof. Dr. Alceu de Souza Britto Jr.

Coorientador: Prof. Dr. Alessandro Lameiras Koerich

Curitiba - PR, Brasil

2025

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central

Menon, Luciana Trinkaus
M547s Seleção dinâmica de modalidade e representação de dados para reconhecimento
2025 multimodal de emoções : um estudo sobre o impacto da ausência de modalidades /
Luciana Trinkaus Menon ; orientador: Alceu de Souza Britto Jr. ; coorientador:
Alessandro Lameiras Koerich. – 2025.
98 f. : il. ; 30 cm

Tese (doutorado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2025.
Bibliografia: f. 91-98

1. Informática. 2. Reconhecimento de padrões. 3. Aprendizado do computador.
4. Inteligência artificial. I. Britto Junior, Alceu de Souza. II. Koerich, Alessandro
Lameiras. III. Pontifícia Universidade Católica do Paraná. Programa de
Programa de Pós-Graduação em Informática. IV. Título.

CDD 20. ed. – 004

Bibliotecária: Luci Eduarda Wielganczuk – CRB 9/1118

Pontifícia Universidade Católica do Paraná
Escola Politécnica
Programa de Pós-Graduação em Informática

Curitiba, 11 de setembro de 2025.

98-2025

DECLARAÇÃO

Declaro para os devidos fins, que **LUCIANA TRINKAUS MENON** defendeu a tese de Doutorado intitulada "**Seleção Dinâmica de Modalidade e Representação de Dados para Reconhecimento Multimodal de Emoções: Um Estudo sobre o Impacto da Ausência de Modalidades"**, na área de concentração Ciência da Computação no dia 25 de abril de 2025, no qual foi aprovada.

Declaro ainda, que foram feitas todas as alterações solicitadas pela Banca Examinadora, cumprindo todas as normas de formatação definidas pelo Programa.

Por ser verdade, firmo a presente declaração.

gov.br Documento assinado digitalmente
JEAN PAUL BARDDAL
Data: 11/09/2025 20:48:46-0300
Verifique em https://validar.iti.gov.br

_____
Prof. Dr. Jean Paul Barddal
Coordenador do Programa de Pós-Graduação em Informática

Agradecimentos

Agradeço primeiramente à minha família, que me apoiou de forma incondicional durante toda a minha vida. À minha mãe, por me encorajar a buscar apoio, algo fundamental nesta reta final. Ao meu pai, que sempre foi um exemplo de acadêmico brilhante: desde que assisti à sua defesa de doutorado, passei a sonhar com o dia em que eu mesma concluiria o meu.

Ao meu marido, pela paciência e por ler e reler este documento inúmeras vezes, contribuindo para cada revisão. Aos meus orientadores, Alceu e Alessandro: ao Alceu, que me orientou pela primeira vez ainda na graduação, abriu portas para meu primeiro emprego e nunca deixou de acreditar no meu potencial; ao Alessandro, cujas palavras firmes me estimularam a trabalhar mais intensamente.

A todas as pessoas que me ouviram e apoiaram em momentos difíceis, em especial Fabiana, Juliana, Barbara, Eduardo e Andrea, que estiveram presentes nos momentos em que precisei de suporte e compreensão. Ao meu amigo Israel, parceiro de trabalho e de estudos, que ajudou na revisão deste trabalho e esteve ao meu lado desde o início.

Aos colegas Jean e Luiz, pelo trabalho conjunto que resultou na publicação no ICPR 2024, de maneira especial ao Luiz por apresentar nosso trabalho na conferência e ao Jean pela otimização do cálculo das distâncias dos vetores de características, sem a qual esta pesquisa seria inviável. Agradeço também a todos os professores que passaram pela minha vida acadêmica, contribuindo para a formação que me trouxe até aqui.

Este trabalho foi financiado pela UNIVISION INFORMÁTICA LTDA e pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), por meio das bolsas nº 306688/2018-2 e 142039/2019-5.

Por fim, meu sincero muito obrigada a cada um de vocês. Sem o incentivo, o exemplo e a ajuda de todos, nada disso teria sido possível.

Este trabalho é dedicado às crianças adultas que,
quando pequenas, sonharam em se tornar cientistas.

"Qualquer tecnologia suficientemente avançada é indistinguível de mágica."
(Arthur C. Clarke)
"Continua sendo mágica mesmo quando se sabe como é feito."
(Terry Pratchett)

Resumo

A compreensão das emoções humanas constitui uma área de estudo fundamental nas ciências sociais e cognitivas, impactando campos como psicologia, neurociência e, mais recentemente, inteligência artificial. O reconhecimento multimodal de emoções destaca-se por possibilitar que algoritmos integrem dados provenientes de múltiplos canais como expressões faciais, voz, texto e sinais fisiológicos para representar de maneira mais completa os estados emocionais humanos. No entanto, o reconhecimento multimodal de emoções enfrenta desafios técnicos consideráveis. Um obstáculo significativo é a indisponibilidade dos dados, uma situação comum no mundo real, em que uma ou mais modalidades, como visual, auditiva, textual ou outra, podem estar ausentes devido a falhas técnicas, condições ambientais desfavoráveis ou questões de privacidade. Este estudo avalia um método de seleção dinâmica de duas etapas, composto pela seleção da modalidade e da melhor representação intra-modal dos dados. O estudo compara o desempenho do método proposto com mecanismos de atenção cruzada e fusão estática das modalidades. O protocolo experimental foi estruturado para contemplar diferentes situações, incluindo cenários de ausência total e parcial de modalidades. Os resultados mostram que a seleção dinâmica supera os métodos de referência nesses cenários, evidenciando sua efetividade no contexto do reconhecimento multimodal de emoções. O estudo conclui enfatizando a complexa interação entre as modalidades de áudio e vídeo na predição de emoções e demonstrando a adaptabilidade dos métodos de seleção dinâmica no gerenciamento de modalidades ausentes.

Palavras-chave: Reconhecimento Multimodal de Emoções; Seleção Dinâmica; Modalidades Ausentes; Fusão de Modalidades; Aprendizado de Máquina; Inteligência Artificial.

Abstract

Understanding human emotions is a fundamental area of study in the social and cognitive sciences, impacting fields such as psychology, neuroscience, and, more recently, artificial intelligence. Multimodal emotion recognition stands out for enabling algorithms to integrate data from multiple channels, such as facial expressions, voice, text, and physiological signals, to more comprehensively represent human emotional states. However, multimodal emotion recognition faces significant technical challenges. One major obstacle is data unavailability, a common situation in the real world, where one or more modalities, such as visual, auditory, or textual, may be missing due to technical failures, unfavorable environmental conditions, or privacy concerns. This study evaluates a two-step dynamic selection based method, comprising the selection of the most appropriate modality and the best intra-modal data representation. The proposed method is compared with cross-attention mechanisms and static modality fusion approaches. The experimental protocol was designed to encompass different conditions, including scenarios with total or partial modality absence. Results show that dynamic selection outperforms baseline methods in these scenarios, highlighting its effectiveness in the context of multimodal emotion recognition. The study concludes by emphasizing the complex interaction between audio and video modalities in emotion prediction and demonstrating the adaptability of dynamic selection methods in handling missing modalities.

Keywords: Multimodal Emotion Recognition; Dynamic Selection; Missing Modalities; Modality Fusion; Machine Learning; Artificial Intelligence.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

IA	Inteligência Artificial
MER	Reconhecimento Multimodal de Emoções
eGeMAPS	Geneva Minimalistic Acoustic Parameter Set
MFCCs	Mel-Frequency Cepstral Coefficients
LSTM	Long Short-Term Memory
RECOLA	REmote COLlaborative and Affective interactions
SEWA	Sentiment Analysis in the Wild
DRS	Seleção Dinâmica de Regresssores
k-NN	K-vizinhos mais Próximos
MSE	Erro Quadrático Médio
MAE	Erro Absoluto Médio
DS	Seleção Dinâmica
DW	Ponderação Dinâmica
DWS	Ponderação Dinâmica com Seleção
RNN	Rede Neural Recorrente
Bi-LSTM	LSTM Bidirecional
Seq2Seq	Sequence-to-Sequence LSTM
Conv-LSTM	Convolutional LSTM
ANNEMO	ANNotating EMOtions
AVEC	Audio Visual Emotion Recognition Challenge
CCC	Coeficiente de Correlação de Concordância
LGBP-TOP	Local Gabor Binary Patterns from Three Orthogonal Planes
LBP	Local Binary Pattern

PCA	Análise de Componentes Principais
ECG	Eletrocardiograma
EDA	Atividade Eletrodérmica
HR	Frequência Cardíaca
HRV	Variabilidade Frequência Cardíaca
SCR	Condutância da Pele
SCL	Nível Basal de Condutância da Pele
SVM	Support Vector Machine
RF	Random Forest
DCNN	Redes Neurais Convolucionais Profundas
SIFT	Scale-Invariant Feature Transform
SVR	Support Vector Regression
MSDF	Multi-scale Dense SIFT
BoAW	Bag of Audio Words
BoVW	Bag of Video Words
BoTW	Bag of Text Words
CNN	Rede Neural Convolucional
DBLSTM-RNN	Deep Bidirectional Long Short-Term Memory RNN
MFB	Mel-Frequency Bank
3DMM	Modelos Morfáveis 3D
GRU	Unidade Recorrente com Portão
ST-GCN	Rede Convolucional de Grafo Espaço-Temporal
DBLSTM	Deep Bidirectional Long Short-Term Memory
DDAT	Dynamic Difficulty Awareness Training
MTL	Multi-Task Learning
AEG	Gerador baseado em Autoencoder

CD	Discriminador Condicional
I3D	3D-CNN
IF-MMIN	Invariant Feature aware Missing Modality Imagination Network
MLP	Multi-Layer Perceptron MLP

Sumário

1 Introdução

Em cenários complexos, a combinação de múltiplos modelos especializados via seleção dinâmica pode resultar em uma melhoria significativa na precisão preditiva (BRITTO ALCEU S.; SABOURIN; OLIVEIRA, 2014; CRUZ; SABOURIN; CAVALCANTI, 2018; MOURA; CAVALCANTI; OLIVEIRA, 2021). A seleção dinâmica pode ser caracterizada como o processo pelo qual as decisões são tomadas de forma adaptativa, considerando as características e condições específicas de uma situação ou ambiente. Diferentemente da seleção estática, que utiliza um conjunto fixo de modelos, a seleção dinâmica avalia, em tempo de execução, a competência de diferentes modelos e escolhe o modelo mais adequado para cada padrão de teste (MOURA; CAVALCANTI; OLIVEIRA, 2019).

Sistemas multimodais demonstram aumentar a precisão e a confiabilidade quando comparados a sistemas unimodais (GLADYS; VETRISELVI, 2023; LIU et al., 2023; D'MELLO; KORY, 2015a). Nesse contexto, o Reconhecimento Multimodal de Emoções (do inglês, Multimodal Emotion Recognition - MER) tem ganhado destaque, integrando dados de diferentes canais, como expressões faciais, voz, texto e sinais fisiológicos, para melhorar a precisão e fornecer uma visão mais completa dos estados emocionais. Esses canais oferecem informações complementares, permitindo que os modelos de MER capturem as nuances emocionais e avancem em aplicações como interação humano-computador, monitoramento de saúde mental e ambientes inteligentes (BALTRUSAITIS; AHUJA; MORENCY, 2019; LIU et al., 2023; ABDOLLAHI et al., 2023; LIAN et al., 2023a; ZHANG et al., 2024).

A maioria dos sistemas multimodais pressupõe a disponibilidade constante e a complementaridade das modalidades, o que leva a uma interpretação mais rica e abrangente dos dados subjacentes (LIU et al., 2023). Cada modalidade - visual, auditiva, textual ou outra - contribui com percepções únicas e, quando mescladas, essas percepções formam representações aprimoradas que são frequentemente mais informativas e precisas do que qualquer modalidade isolada (PRAVEEN; CARDINAL; GRANGER, 2023; D'MELLO; KORY, 2015a). No entanto, o cenário ideal em que todas as modalidades esperadas estão disponíveis para uma determinada tarefa nem sempre reflete a realidade, pois uma ou mais modalidades podem estar ausentes ou prejudicadas devido a problemas técnicos, condições ambientais desfavoráveis ou questões de privacidade (D'MELLO; KORY, 2015a; ZHU; SUN; ZHOU, 2023; VAZQUEZ-RODRIGUEZ et al., 2023; LIN; GONCALVES; BUSSO, 2023; LI et al., 2024).

1.1 Definição do Problema

Dados multimodais são complexos e heterogêneos, apresentando desafios significativos em aprendizado de máquina. Um dos principais desafios é a definição de mecanismos de fusão capazes de determinar, de forma eficiente, a contribuição relativa de cada modalidade para o desempenho geral do modelo (GLADYS; VETRISELVI, 2023). Além disso, a combinação eficaz de representações provenientes de diferentes modalidades é dificultada por suas diferenças no formato dos dados, escalas e características temporais (BALTRUSAITIS; AHUJA; MORENCY, 2019). Gerenciar modalidades ausentes também é um desafio crítico nas aplicações multimodais, demandando estratégias robustas para assegurar que a ausência de informações em uma ou mais modalidades não comprometa o desempenho do sistema (LIU et al., 2023).

Técnicas estáticas de fusão frequentemente subutilizam as informações disponíveis ao tratar todos os dados de maneira uniforme, desconsiderando as variações de relevância dos dados em diferentes contextos (MOURA; CAVALCANTI; OLIVEIRA, 2019). Para superar esses desafios, este trabalho avalia um modelo inovador de seleção dinâmica para fusão multimodal, combinando a escolha da modalidade mais informativa e a seleção intra-modal da representação de dados mais adequada para cada padrão de teste.

1.2 Objetivos

O principal objetivo desta pesquisa é avaliar um método de seleção dinâmica capaz de escolher a modalidade e a representação de dados mais apropriada para cada exemplo de teste, otimizando os resultados mesmo em cenários com modalidades ausentes. Para atingir esse objetivo, a pesquisa busca:

(i) Desenvolver um método de seleção dinâmica constituído de duas etapas: seleção da modalidade mais informativa e seleção intra-modal da melhor representação dos dados;

(ii) Comparar o desempenho do método proposto com abordagens de referência, incluindo mecanismo de atenção e fusão estática das modalidades (média e melhor individual);

(iii) Avaliar a sensibilidade do método proposto frente à ausência de modalidades, considerando cenários com diferentes níveis de indisponibilidade de dados.

1.3 Hipóteses

Com base nos objetivos propostos, as seguintes hipóteses foram formuladas para guiar esta pesquisa:

(i) Hipótese Principal (H_1): A utilização de técnicas de seleção dinâmica de modalidade e representação de dados melhora significativamente o desempenho de sistemas multimodais em comparação com abordagens de fusão estática.

(ii) Hipótese Secundária (H_2): A seleção dinâmica de modalidade e representação de dados é uma estratégia eficaz para manter a precisão em cenários com modalidades ausentes.

1.4 Método Proposto

Este trabalho investiga um método de seleção dinâmica em duas etapas com o objetivo de otimizar sistemas multimodais. A primeira etapa realiza a seleção da modalidade mais informativa, enquanto a segunda seleciona, dentro dessa modalidade, a representação de dados mais adequada. A abordagem proposta apresenta grande potencial no processo de seleção dinâmica, configurando-se como uma estratégia promissora para a fusão de informações multimodais.

O método proposto foi validado no contexto de MER para regressão de ativação e valência, considerando situações ideais com todas as modalidades disponíveis e ausência parcial ou total das modalidades. Para a modalidade de áudio, foram utilizadas três representações: Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), Mel-Frequency Cepstral Coefficients (MFCCs) e Mel espectrogramas. Para a modalidade de vídeo, foram extraídas características de aparência, características de geometria e características extraídas por meio de uma rede pré-treinada ResNet. Os resultados obtidos indicam que o método proposto é uma abordagem promissora para o MER, especialmente em cenários com modalidades ausentes.

1.5 Contribuições

A contribuição principal desta pesquisa inclui o desenvolvimento de um método inovador de seleção dinâmica para fusão multimodal. Até o melhor do nosso conhecimento, este é o primeiro trabalho a propor um modelo estruturado na seleção dinâmica da modalidade mais informativa, somado à seleção intra-modal da melhor representação dos dados.

O estudo compara o desempenho do método proposto com o mecanismo de atenção cruzada proposto por Praveen et al. (PRAVEEN et al., 2022) e fusão estática das modalidades, média e melhor regressor individual, avaliando como cada técnica se comporta em condições ideais, bem como diante da ausência total ou parcial de dados. Esta análise fornece informações valiosas sobre a robustez e eficácia de cada abordagem, fornecendo diretrizes para o desenvolvimento de sistemas mais resilientes e adaptáveis.

Embora a validação deste estudo tenha sido focada no MER, a natureza modular e escalável do método proposto permite sua aplicação em qualquer sistema multimodal, ampliando seu alcance e relevância. Optou-se pelo MER como estudo de caso em razão de sua relevância prática: sistemas capazes de identificar estados emocionais do usuário podem ser aplicados em múltiplos cenários, tais como saúde mental, educação e suporte ao usuário (BALTRUSAITIS; AHUJA; MORENCY, 2019; LIU et al., 2023; ABDOLLAHI et al., 2023; LIAN et al., 2023a; ZHANG et al., 2024). Além disso, a emoção é um fenômeno complexo, frequentemente mal capturado por uma única fonte de dados (unimodal), e a análise multimodal permite suprir lacunas e melhorar a robustez dos sistemas de reconhecimento de emoções (GLADYS; VETRISELVI, 2023; LIU et al., 2023; D'MELLO; KORY, 2015a). Por fim, o uso de MER apresenta contribuições sociais significativas, pois ao viabilizar sistemas de reconhecimento emocional mais resistentes a cenários reais, este estudo contribui para a criação de tecnologias que podem melhorar a qualidade de vida e promover um impacto positivo em áreas como saúde, educação e interação humano-computador.

1.5.1 Publicações

Este trabalho resultou na aprovação da seguinte publicação:

MENON, Luciana Trinkaus; NEDUZIAK, Luiz Carlos Ribeiro; BARDDAL, Jean Paul; KOERICH, Alessandro Lameiras; BRITTO, Alceu de Souza. Dynamic modality and view selection for emotion recognition: An experimental study on missing modality evaluation. ICPR 2024 International Workshops and Challenges. Pattern Recognition, Lecture Notes in Computer Science, vol 15617. 151–166 p. Disponível em: <http://dx.doi.org/10.1007/978-3-031-88217-3_11>.

1.6 Estrutura do Trabalho

O restante deste trabalho está organizado em 6 capítulos. O Capítulo 2 traz uma fundamentação teórica, revisando conceitos fundamentais para a compreensão do método proposto e dos trabalhos abordados no estado da arte. A revisão bibliográfica de trabalhos relevantes sobre reconhecimento de emoções por meio de múltiplas modalidades e o tratamento de modalidades ausentes durante a inferência é apresentada no Capítulo 3. A descrição detalhada do método proposto pode ser encontrada no Capítulo 4. O Capítulo 5 apresenta a análise dos resultados obtidos com a aplicação do método proposto no contexto de MER. Além disso, analisa a sensibilidade do método frente à ausência de modalidades e compara seu desempenho com abordagens baseadas em mecanismos de atenção e fusão estática das modalidades. Por fim, o Capítulo 6 discute os resultados e o Capítulo 7 finaliza o trabalho destacando sua relevância, contribuições e limitações, além de apresentar os trabalhos futuros.

2 Fundamentação Teórica

Este capítulo aborda os principais conceitos relacionados à seleção dinâmica de regressores, redes Long Short-Term Memory (LSTM) e mecanismos de atenção, que são fundamentais para a compreensão do método proposto e dos trabalhos abordados no estado da arte. Além disso, são apresentadas as definições e teorias de emoção, que fornecem o contexto no qual o método proposto foi validado. Por fim, são descritas as bases de dados RECOLA (REmote COLlaborative and Affective interactions) e SEWA (Sentiment Analysis in the Wild), empregadas na validação do método proposto.

2.1 Seleção Dinâmica de Regressores

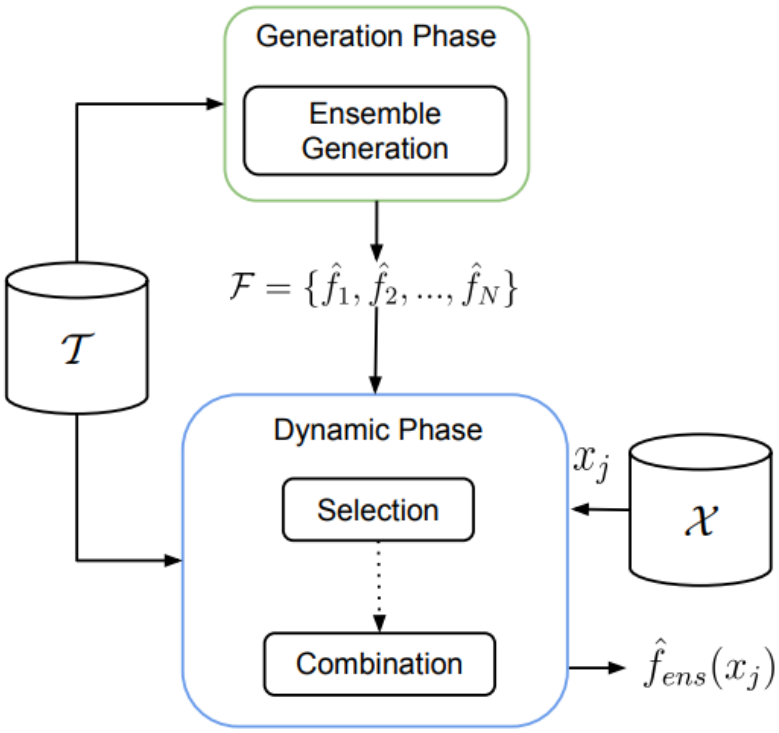
A motivação por trás da seleção dinâmica é identificar e selecionar os modelos mais competentes em uma região específica do espaço de características, conhecida como região de competência, visando fornecer a predição mais precisa possível para cada novo padrão de teste (BRITTO ALCEU S.; SABOURIN; OLIVEIRA, 2014; CRUZ; SABOURIN; CAVALCANTI, 2018; MOURA; CAVALCANTI; OLIVEIRA, 2021). Sistemas de seleção dinâmica são compostos por três fases principais, representadas na Figura 1: (a) geração do ensemble, (b) seleção e (c) combinação dos modelos selecionados. Na primeira fase, é gerado um conjunto de modelos; na segunda fase, um ou um subconjunto desses modelos é selecionado; enquanto, na última fase, a decisão final é tomada com base na(s) predição(ões) do(s) modelo(s) selecionado(s) (BRITTO ALCEU S.; SABOURIN; OLIVEIRA, 2014). É importante ressaltar que esta representação não é única, pois as etapas de seleção e combinação podem ser opcionais, havendo casos em que todo o conjunto de modelos é utilizado sem qualquer seleção, ou situações em que apenas um único modelo é escolhido, eliminando a necessidade da fase de integração.

Enquanto a seleção dinâmica é amplamente explorada em problemas de classificação, sua aplicação em problemas de regressão, conhecida como Seleção Dinâmica de Regressores (do inglês, Dynamic Regressor Selection - DRS), requer adaptações específicas. A DRS busca selecionar os regressores mais competentes de um ensemble para estimar o valor de uma variável-alvo em um problema de regressão.

2.1.1 Geração do Ensemble

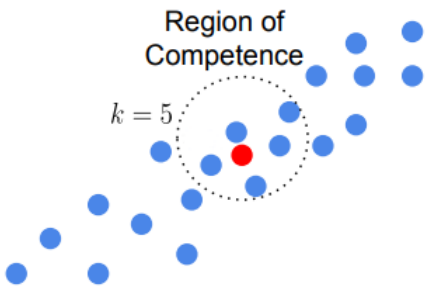
A fase de geração do ensemble envolve o treinamento de múltiplos regressores, denotado como $\mathcal{F} = \hat{f}_1, \hat{f}_2, \ldots, \hat{f}_N$, onde N é o número total de modelos. Quando um único algoritmo de aprendizado é utilizado para treinar todos os modelos, o conjunto é

Figura 1 – Visão geral da arquitetura de Seleção Dinâmica de Regressores, com três fases principais: geração do ensemble, seleção e combinação dos modelos selecionados. $\mathcal{T}$ e $\mathcal{X}$ são os conjuntos de treinamento e teste, respectivamente. $\mathcal{F}$ é o conjunto de regressores gerados na Fase de Geração, x_j é um padrão de teste e $\hat{f}_{\text{ens}}(x_j)$ é o resultado da estimativa do padrão de teste.



Fonte: Moura, Cavalcanti e Oliveira (2019)

Figura 2 – Ilustração da região de competência. Pontos azuis representam padrões do conjunto de treino. O ponto vermelho representa o padrão de teste x_j. k é o tamanho da região de competência.



Fonte: Moura, Cavalcanti e Oliveira (2019)

considerado homogêneo; entretanto, se múltiplos algoritmos são empregados, o conjunto é chamado de heterogêneo (MOURA; CAVALCANTI; OLIVEIRA, 2019). O objetivo desta fase é criar um conjunto de modelos diversificados que cometam erros diferentes e, consequentemente, apresentem algum grau de complementaridade (BRITTO ALCEU S.; SABOURIN; OLIVEIRA, 2014). Essa diversidade é importante pois aumenta a probabilidade de que pelo menos um dos modelos do ensemble seja competente em uma região específica do espaço de características.

2.1.2 Seleção Dinâmica de Modelos

Nesta etapa, os modelos são avaliados quanto à sua competência em prever um padrão de teste específico. Essa avaliação é realizada em uma região de competência, denotada como Ψ, definida como a vizinhança da amostra de teste no espaço de características do conjunto de treinamento. A técnica mais comum para encontrar esta região é o método dos K-vizinhos mais próximos (do inglês, K-Nearest Neighbors - k-NN), utilizando a distância euclidiana como medida de similaridade (MOURA; CAVALCANTI; OLIVEIRA, 2019). Já a competência dos modelos pode ser calculada com base em diversas medidas, como erros acumulados e desempenho passado em regiões semelhantes (CRUZ; SABOURIN; CAVALCANTI, 2018).

A competência de um classificador é frequentemente medida pela taxa de acerto ou precisão em uma classe específica. Em contrapartida, em problemas de regressão, as medidas de erro, como o Erro Quadrático Médio (do inglês, Mean Squared Error - MSE) ou o Erro Absoluto Médio (do inglês, Mean Absolute Error - MAE), são comumente utilizadas para avaliar a competência dos modelos. No entanto, estudos indicam que outras medidas também podem ser vantajosas e a escolha da medida de competência ideal pode variar conforme o problema. Além disso, os sistemas podem se beneficiar ao utilizar uma combinação de várias medidas, em vez de depender exclusivamente de uma única métrica (MOURA; CAVALCANTI; OLIVEIRA, 2019; BRITTO ALCEU S.; SABOURIN; OLIVEIRA, 2014; CRUZ; SABOURIN; CAVALCANTI, 2018).

2.1.3 Combinação dos Modelos Selecionados

Após a seleção, os modelos mais competentes são combinados para gerar a predição final. A fase dinâmica seleciona um subconjunto de regressores para cada padrão de teste x_j, podendo operar de três maneiras diferentes: (i) selecionar apenas um regressor do conjunto $\mathcal{F}$; (ii) realizar uma combinação ponderada de todos os regressores; e (iii) selecionar um subconjunto dos regressores e combiná-los. O resultado dessa fase é a predição do conjunto $\hat{f}_{\text{ens}}(x_j)$ (MOURA; CAVALCANTI; OLIVEIRA, 2019).

2.1.3.1 Seleção Dinâmica (DS)

A Seleção Dinâmica (do inglês, Dynamic Selection - DS) seleciona o regressor com o menor erro acumulado na região de competência. Os erros são ponderados pela distância entre os padrões da vizinhança e o padrão de teste. Apenas um único regressor é selecionado, sem a necessidade de combinar resultados. A estimativa do padrão de teste é o valor retornado pelo regressor selecionado.

2.1.3.2 Ponderação Dinâmica (DW)

A Ponderação Dinâmica (do inglês, Dynamic Weighting - DW) combina todos os regressores do ensemble usando uma média ponderada. Cada peso é calculado com base no desempenho do regressor na região de competência, garantindo que modelos mais precisos em um contexto específico contribuam mais para a predição final. Para cada padrão de teste x_j, sua região de competência Ψ é calculada. Para cada item em Ψ, um peso d_k é calculado usando a Equação 2.1, onde $dist_k$ é a medida de distância entre o item na região de competência $t_k \in \Psi$ e o padrão de teste x_j.

$$d_k = \frac{(\text{dist}_k)^{-1}}{\sum_{j=1}^{K} (\text{dist}_j)^{-1}} \tag{2.1}$$

O vetor $d_1, d_2, ..., d_k$ é utilizado para calcular o peso α_i do regressor f_i usando a Equação 2.2 onde N é o tamanho do conjunto de regressores, k representa o índice do vizinho e $sqe_{k,i}$ é o erro quadrático do regressor i calculado com o item $t_k \in \Psi$.

$$\alpha_i = \frac{\left[\sum_{k=1}^{K} (d_k * sqe_{k,i})\right]^{-1}}{\sum_{n=1}^{N} \left[\sum_{k=1}^{K} (d_k * sqe_{k,i})\right]^{-1}} \tag{2.2}$$

2.1.3.3 Ponderação Dinâmica com Seleção (DWS)

A Ponderação Dinâmica com Seleção (do inglês, Dynamic Weighting with Selection - DWS) combina um subconjunto dos regressores. O regressores com erro acumulado na metade superior do intervalo de erro $E_i > (E_{max} - E_{min})/2$ são descartados, garantindo que apenas os mais competentes contribuam para a predição. O método para calcular os pesos dos regressores e a estratégia para combinar os modelos são os mesmos do DW (Equações 2.1 e 2.2).

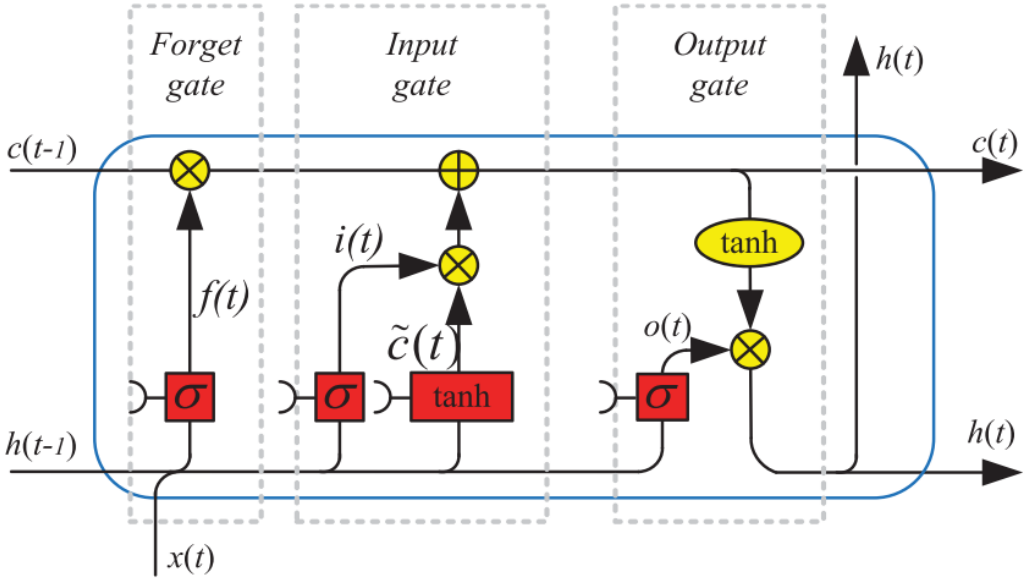
2.2 Long Short-Term Memory (LSTM)

A célula LSTM, introduzida por Hochreiter e Schmidhuber (1997), permite o gerenciamento eficiente de informações sequenciais, mantendo seletivamente informações relevantes ao longo do tempo através de mecanismos de gates. LSTMs foram desenvolvidas como solução para as limitações das Redes Neurais Recorrentes (do inglês, Recurrent Neural Networks - RNNs) tradicionais, especialmente em relação à captura de dependências de longo prazo, um desafio decorrente do problema de esquecimento de gradientes das RNNs (LINDEMANN et al., 2021).

2.2.1 Estrutura da LSTM

As células LSTM consistem em três principais componentes de controle: um gate de entrada (i_t), um gate de esquecimento (f_t) e um gate de saída (o_t), que permitem modificar o vetor de estado da célula, propagar informações de forma iterativa e capturar dependências de longo prazo (LINDEMANN et al., 2021). Esse fluxo controlado de informações dentro da célula permite à rede memorizar múltiplas dependências temporais com características distintas.

Figura 3 – Arquitetura de uma célula LSTM, composta por três gates principais: gate de entrada (i_t), gate de esquecimento (f_t) e gate de saída (o_t). O estado da célula $c(t)$ e o estado oculto $h(t)$ são atualizados iterativamente para capturar informações ao longo do tempo.



Fonte: Yu et al. (2019)

2.2.1.1 Gate de Entrada

O gate de entrada controla quais valores da nova entrada devem ser armazenados no estado da célula. Sua formulação é dada por:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.3}$$

onde i_t é o vetor de ativação do gate de entrada, h_{t-1} é o estado oculto da célula anterior, x_t é a entrada atual e W_i e b_i são os pesos e viéses do gate.

2.2.1.2 Gate de Esquecimento

O gate de esquecimento decide quais informações das etapas temporais anteriores devem ser descartadas. Ele é descrito pela fórmula:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.4}$$

onde f_t controla a extensão do esquecimento da informação anterior.

2.2.1.3 Gate de Saída

O gate de saída determina quais valores do estado da célula são utilizados para o próximo estado oculto e para a saída. Ele é descrito pela seguinte expressão:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.5}$$

ativando-se através da função tangente hiperbólica ($tanh$).

2.2.1.4 Armazenamento e Atualização do Estado da Célula

O estado da célula c_t é uma combinação do estado anterior c_{t-1}, controlada pelo gate de esquecimento f_t, e do estado candidato da célula $\tilde{c}_t$, ponderado pelo gate de entrada i_t:

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \tag{2.6}$$

$$\tilde{c}_t = \tanh(W_{\tilde{c}h} h_{t-1} + W_{\tilde{c}x} x_t + b_{\tilde{c}}) \tag{2.7}$$

$$h_t = o_t \cdot \tanh(c_t) \tag{2.8}$$

onde $\tilde{c}_t$ é o valor candidato que será incorporado ao estado atual da célula e h_t é o estado oculto atual, que serve como a saída ponderada do estado da célula.

2.2.2 Arquiteturas Variantes de LSTMs

As LSTMs evoluíram para diferentes variantes, cada uma com características que aprimoram o desempenho em tarefas específicas, como predição de séries temporais e tradução automática. Entre as variantes mais notáveis podemos citar as LSTMs Bidirecionais (Bi-LSTM), Seq2Seq (Sequence-to-Sequence) e Convolutional LSTMs (Conv-LSTM) (LINDEMANN et al., 2021).

2.2.2.1 LSTM Bidirecional (Bi-LSTM)

A LSTM tradicional é amplamente utilizada em tarefas que requerem a modelagem de dependências de longo prazo, como séries temporais e processamento de sinais, devido ao seu controle detalhado sobre o fluxo de informações por meio dos gates de entrada, saída e esquecimento. A Bi-LSTM, propaga o vetor de estado não apenas no sentido direto, mas também no sentido reverso. Isso permite que dependências em ambas as direções temporais sejam consideradas. Dessa forma, correlações futuras podem ser incorporadas nos outputs atuais da rede, devido à propagação reversa do estado. Assim, as redes LSTM bidirecionais conseguem detectar e extrair mais dependências temporais do que as redes LSTM unidirecionais (LINDEMANN et al., 2021; SCHUSTER; PALIWAL, 1997).

2.2.2.2 LSTM Convolucional (Conv-LSTM)

As Conv-LSTMs são uma extensão das LSTMs que integram operações de convolução, tornando-as especialmente úteis para problemas de processamento de dados espaciais e temporais, como na análise de vídeos e detecção de movimento. A camada convolucional aplicada à entrada permite que a Conv-LSTM capture dependências de longo alcance, ao processar de forma mais eficiente as sequências de entrada e projetar os dados em um espaço de características de menor dimensão (LINDEMANN et al., 2021).

2.2.2.3 Redes Sequence-to-Sequence (Seq2Seq)

A arquitetura Seq2Seq consiste em duas partes principais: um codificador e um decodificador. O codificador recebe a sequência de entrada e a transforma em uma representação vetorial fixa, que encapsula as informações mais relevantes da entrada. Essa representação vetorial é então passada ao decodificador, que gera uma sequência de saída baseada na informação armazenada no vetor do codificador, permitindo a criação de uma sequência de saída de comprimento variável em relação à entrada (LINDEMANN et al., 2021; SUTSKEVER; VINYALS; LE, 2014).

2.3 Mecanismos de Atenção

Os mecanismos de atenção surgiram como uma alternativa para capturar dependências em tarefas de modelagem sequenciais, permitindo que o modelo foque seletivamente em partes relevantes da entrada. Essa abordagem possibilita o aprendizado de relações complexas sem levar em conta a distância entre os elementos na sequência e é especialmente útil em aplicações como tradução automática, síntese de fala e processamento de linguagem natural (BRAUWERS; FRASINCAR, 2023). Com a introdução do modelo Transformer, Vaswani et al. (2017) revolucionaram o uso de mecanismos de atenção ao eliminar a dependência de redes recorrentes e convolucionais, demonstrando que uma arquitetura baseada exclusivamente em atenção pode alcançar resultados de alto nível em diversas tarefas. A Figura 4 detalha a arquitetura Transformer.

2.3.1 Atenção Própria

O conceito de atenção própria permite que cada elemento de uma sequência preste atenção em outros elementos da mesma sequência. Ao calcular uma representação ponderada dos elementos da entrada em relação ao elemento atual, a atenção própria possibilita que o modelo capture relações entre os dados, independentemente da sua posição na sequência. Cada elemento é transformado em três vetores: consulta (*query*), chave e valor, conforme apresentado na Figura 5. As pontuações entre chave e valor determinam a importância relativa de cada elemento para a representação final. Em uma camada de atenção, o vetor de saída é obtido como uma combinação ponderada, onde os pesos são dados pela similaridade entre as consultas e as chaves. Esse processo é expresso pela fórmula:

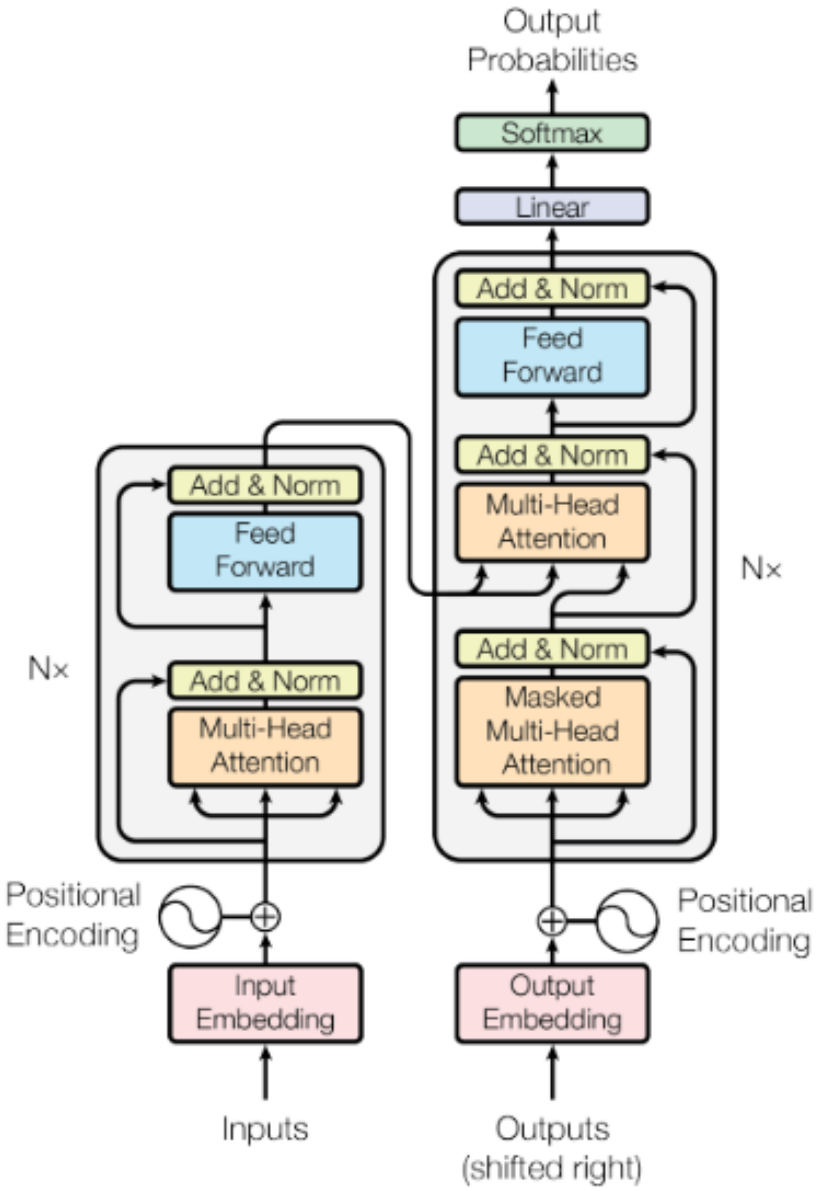
$$\text{Attention}(Q, C, V) = \text{softmax}\left(\frac{QC^T}{\sqrt{d_c}}\right)V, \tag{2.9}$$

onde Q, C e V representam os vetores de consultas, chaves e valores, respectivamente, e (d_c) é a dimensão dos vetores chave, utilizada para normalizar as pontuações (VASWANI et al., 2017).

2.3.2 Atenção Multi-Cabeça

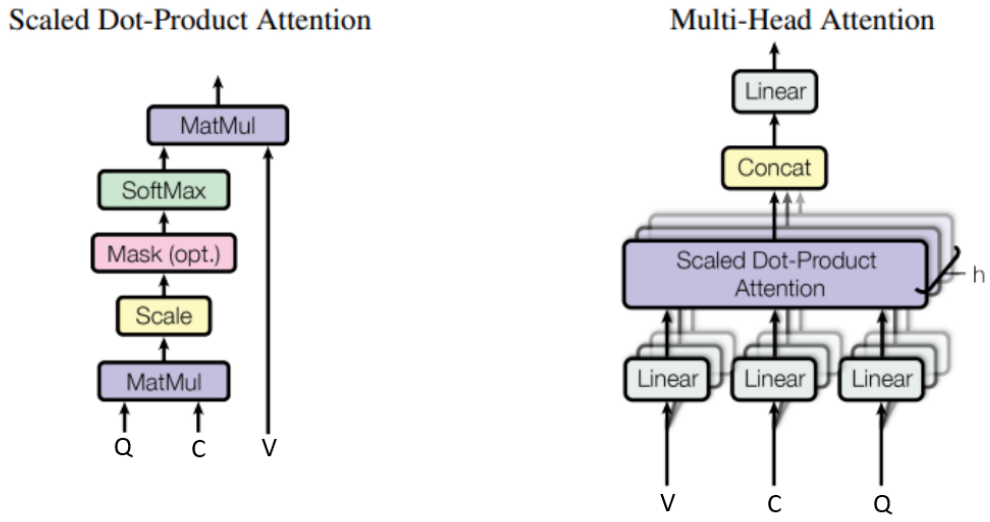
O mecanismo de atenção multi-cabeça é uma extensão da atenção própria, que aplica múltiplas cabeças de atenção paralelamente. Cada cabeça usa diferentes conjuntos de pesos para as consultas, chaves e valores, permitindo que o modelo capture vários tipos de relações entre os elementos da sequência simultaneamente. A saída de cada cabeça é concatenada e transformada em uma única representação, permitindo que o modelo tenha múltiplas perspectivas sobre a mesma entrada e capture relacionamentos mais complexos.

Figura 4 – Arquitetura do modelo Transformer, composta por duas partes principais: o codificador e o decodificador. Cada bloco do codificador e decodificador é composto por múltiplas camadas de atenção e camadas totalmente conectadas. O codificador recebe uma sequência de entrada com embeddings posicionais para preservar a ordem dos dados, enquanto o decodificador processa a sequência de saída deslocada, também com embeddings posicionais. No decodificador, uma camada de atenção mascarada impede o acesso a futuras posições da sequência durante o treinamento. A camada final aplica uma função softmax para gerar as probabilidades de saída.



Fonte: Vaswani et al. (2017)

Figura 5 – Mecanismos de atenção. O mecanismo de Atenção por Produto Interno Escalado (do inglês, Scaled Dot-Product Attention) multiplica os vetores de consulta (Q) e chave (C), escala o resultado, aplica uma máscara opcional e normaliza com softmax antes de multiplicar pelos valores (V). O mecanismo de Atenção Multi-Cabeça (do inglês, Multi-Head Attention) aplica múltiplas instâncias da atenção escalada em diferentes subespaços de Q, C e V, combinando as saídas para capturar relações complexas na sequência.



Fonte: Vaswani et al. (2017)

Este processo pode ser descrito como:

$$\text{MultiHead}(Q, C, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O, \tag{2.10}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, CW_i^C, VW_i^V) \tag{2.11}$$

onde W^O é uma matriz de projeção que transforma a saída final na dimensão original (VASWANI et al., 2017). Essa abordagem permite que cada cabeça aprenda representações específicas da entrada e que o modelo atenda simultaneamente a informações de diferentes subespaços de representação em diferentes posições.

2.4 Definições e Teorias de Emoções

Emoção é um estado psicológico complexo que envolve uma resposta subjetiva a um estímulo, que pode ser interno ou externo. Em um sentido amplo, as emoções podem ser definidas como estados afetivos que influenciam o comportamento humano e variam em intensidade e qualidade (GLADYS; VETRISELVI, 2023).

Emoções são fundamentais para a interpretação e interação em contextos sociais e podem ser percebidas por meio de uma combinação de sinais visuais, auditivos e fisiológicos, como alterações nas expressões faciais, variação na entonação vocal e ritmo cardíaco. Para compreender de forma aprofundada as emoções e seu impacto nos comportamentos, diversas teorias de categorização foram propostas. As teorias são frequentemente divididas em duas abordagens principais: a abordagem discreta, que trata as emoções como categorias qualitativas distintas, e a abordagem contínua, que as representa em dimensões contínuas de variação.

2.4.1 Categorização Discreta de Emoções

As teorias de emoções discretas propõem que as emoções são categorias distintas e qualitativamente diferentes entre si. Cada emoção é vista como uma entidade separada, com características definidas e sinais específicos.

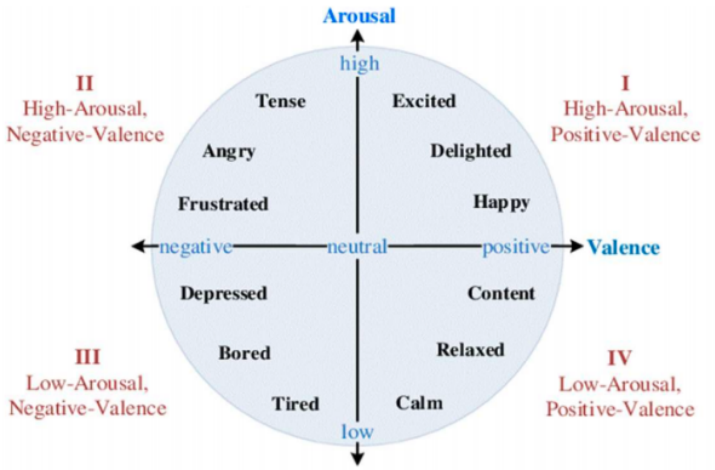
Ekman (1992) foi um dos pioneiros dessa linha de pensamento ao identificar seis emoções básicas: alegria, tristeza, medo, raiva, surpresa e nojo. Segundo ele, as emoções são universais entre culturas e apresentam expressões faciais específicas e reconhecíveis. Outro modelo influente é o de Plutchik (1982), que propôs a "Roda das Emoções", identificando oito emoções primárias (alegria, confiança, medo, surpresa, tristeza, aversão, raiva e antecipação) organizadas em pares de opostos. Segundo o modelo de Plutchik as emoções primárias podem combinar-se para formar emoções mais complexas.

2.4.2 Classificação Contínua de Emoções

As teorias contínuas de emoções representam os estados emocionais como pontos em um espaço multidimensional, oferecendo uma alternativa à categorização rígida das abordagens discretas. No Modelo Circumplexo de Afeto de Russell (1980), as emoções são descritas em duas dimensões principais: ativação (arousal) e valência (positividade versus negatividade). Em uma representação circular do espaço emocional, o eixo vertical representa o nível de ativação na emoção e o eixo horizontal representa o nível de valência, como apresentado na Figura 6. Ambas as dimensões podem ser normalizadas no intervalo entre -1 e +1 (RUSSELL, 1980), o que permite que os sistemas de IA posicionem emoções ao longo dos eixos de ativação e valência, facilitando a rotulação contínua e o reconhecimento de estados emocionais variados.

A valência denota o espectro das emoções, desde muito negativas até muito positivas, enquanto a ativação reflete a intensidade das emoções, variando de estados muito passivos a muito ativos. Emoções com alta valência e ativação, ou seja, positivas e ativas, estão localizadas no quadrante superior direito, enquanto emoções de baixa valência e ativação, ou seja, negativas e passivas, situam-se no quadrante inferior esquerdo.

Figura 6 – Circumplexo de Russell, representação das emoções nos eixos de ativação (do inglês, arousal) e valência (do inglês, valence).



Fonte: Praveen et al. (2022)

2.4.3 Aplicações Práticas

As teorias de emoção têm aplicações diretas em sistemas de reconhecimento de emoções. As teorias de categorização das emoções como Ekman (1992) e Plutchik (1982) são particularmente úteis para a classificação de emoções discretas, enquanto o modelo circumplexo de Russell (1980) se adapta eficientemente a sistemas de regressão de emoções, onde as emoções são descritas como níveis contínuos de ativação e valência.

2.5 Bases de Dados

Esta seção detalha as bases de dados utilizadas para a validação do método proposto, cada uma essencial para testar diferentes aspectos da eficácia do modelo em reconhecer emoções em contextos variados. A base de dados RECOLA (RINGEVAL et al., 2013) é utilizada para examinar interações colaborativas em um ambiente controlado. Por outro lado, a base SEWA (KOSSAIFI et al., 2021) permite a análise de emoções em ambientes não controlados, refletindo desafios encontrados em aplicações do mundo real. A escolha dessas bases de dados é estratégica para abranger um espectro amplo de cenários emocionais e garantir que o método desenvolvido seja robusto e adaptável a diversas condições ambientais e de interação.

A métrica oficial para avaliar o desempenho MER é o Coeficiente de Correlação de Concordância (CCC) (VALSTAR et al., 2016), que captura a relação de co-variação entre predições e rotulação verdade e leva em conta qualquer desvio. Como resultado, oferece

uma representação precisa do alinhamento entre predições e rotulação (LIAN et al., 2023b). Valores mais altos de CCC significam excelente desempenho em termos de consistência e precisão. No contexto deste trabalho, o CCC é calculado separadamente para as duas dimensões emocionais analisadas: ativação e valência. Dessa forma, são obtidos dois valores distintos da métrica, refletindo o desempenho do modelo em cada uma dessas dimensões afetivas. O processo de cálculo para o CCC será definido a seguir:

$$CCC = \frac{2 * \rho * \sigma_y * \sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2} \tag{2.12}$$

onde ρ é o coeficiente de correlação de Pearson, σ_y e $\sigma_{\hat{y}}$ são os desvios padrão e μ_y e $\mu_{\hat{y}}$ são as médias do estado emocional real e previsto.

2.5.1 RECOLA

A base de dados RECOLA (Remote Collaborative and Affective Interactions) (RINGEVAL et al., 2013) compreende 9,5 horas de gravações de áudio, vídeos e atividades fisiológicas de discussões online entre 46 participantes de língua francesa. As emoções presentes nesta base são naturais, expressadas pelos participantes ao resolver uma tarefa em conjunto. O processo de rotulação da base foi feito por seis anotadores, três homens e três mulheres, utilizando a ferramenta ANNEMO (ANNotating EMOtions).

A rotulação dos cinco primeiros minutos de interação de 18 participantes está disponível para treinamento e validação, já a anotação dos dados da partição de teste não está disponível. Este conjunto de dados foi utilizado para benchmarking de sistemas de reconhecimento de emoção em várias edições da competição AVEC (Audio Visual Emotion Recognition Challenge): AVEC'15 (RINGEVAL et al., 2015), AVEC'16 (VALSTAR et al., 2016) e AVEC'18 (RINGEVAL et al., 2018).

A base de dados RECOLA destaca-se pela riqueza e diversidade de suas modalidades de dados, e sua estrutura bem documentada facilita o avanço das pesquisas de aprendizado de máquina e reconhecimento multimodal de emoções. O caráter natural das interações também contribui para a robustez e a generalização dos modelos desenvolvidos. A Figura 7 apresenta um exemplo de alguns frames de vídeo da base de dados RECOLA e as Figuras 8 e 9 são exemplos da rotulação de ativação e valência para um vídeo da base.

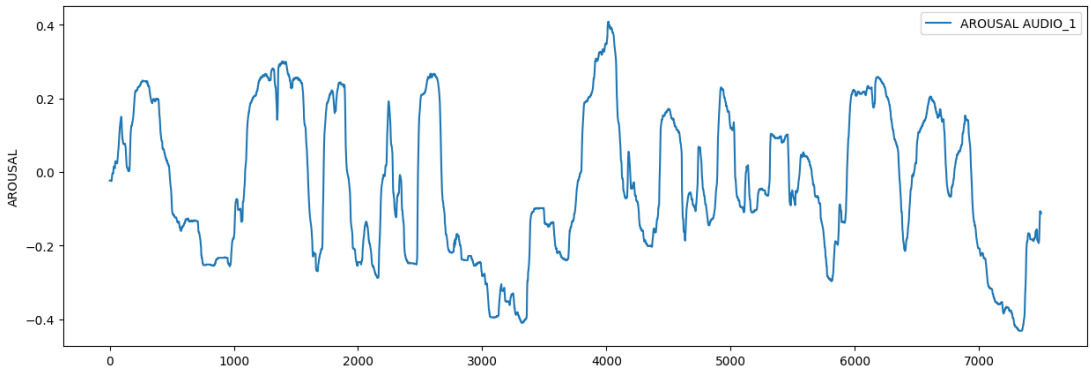
A base de dados RECOLA inclui um conjunto de características pré-extraídas, que são utilizadas como baseline nas competições e são divididas em três grandes categorias: áudio, vídeo e dados fisiológicos. Cada uma dessas modalidades fornece informações cruciais que ajudam a captar as nuances dos estados emocionais dos participantes.

Para o áudio, as características foram extraídas utilizando o conjunto eGeMAPS, implementado através da biblioteca openSMILE (EYBEN; WöLLMER; SCHULLER, 2010).

Figura 7 – Exemplo frames de vídeo RECOLA.

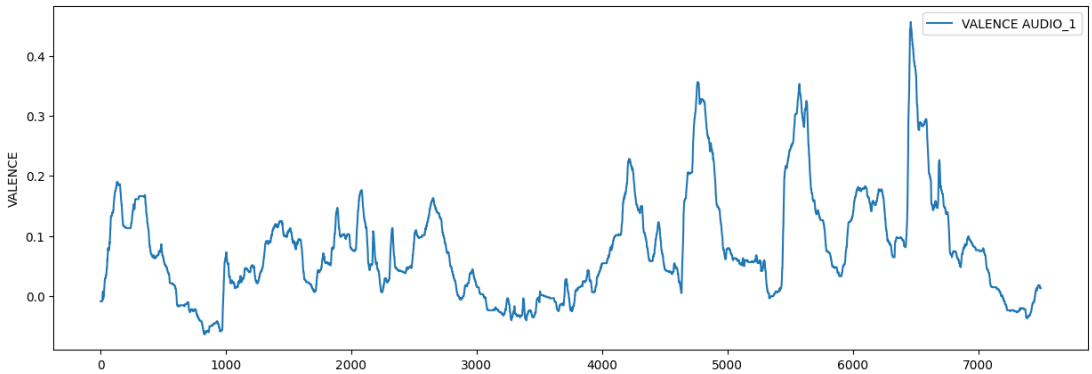
Fonte: Ringeval et al. (2013)

Figura 8 – Exemplo de rotulação da dimensão de ativação na base de dados RECOLA. O eixo vertical representa os valores contínuos da dimensão de ativação (arousal), enquanto o eixo horizontal indica o tempo, medido em frames. Os dados referem-se a modalidade de áudio do participante T1 da base de dados RECOLA.



Fonte: Autoria Própria

Figura 9 – Exemplo de rotulação da dimensão de valência na base de dados RECOLA.O eixo vertical representa os valores contínuos da dimensão de valência (valence), enquanto o eixo horizontal indica o tempo, medido em frames. Os dados referem-se a modalidade de áudio do participante T1 da base de dados RECOLA.

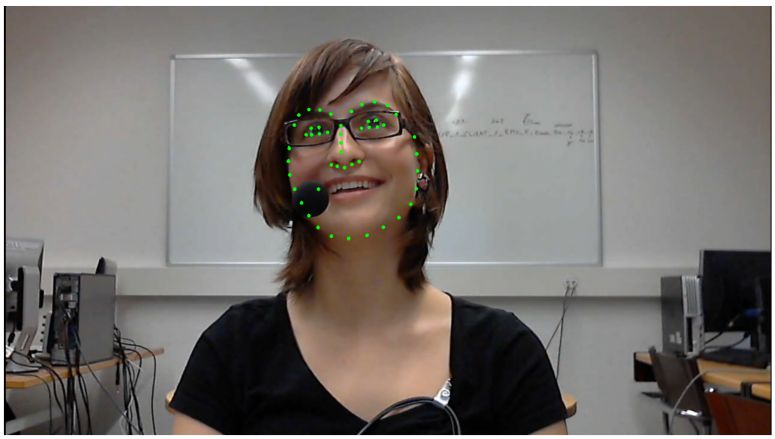


Fonte: Autoria Própria

Este conjunto de parâmetros foi escolhido por sua eficácia na captura de emoções e inclui informações prosódicas, espectrais e de qualidade vocal. No aspecto prosódico, a frequência fundamental da voz (*pitch*) e a intensidade (*loudness*) são consideradas. Além disso, são analisadas a estabilidade na frequência (*jitter*) e na amplitude (*shimmer*), indicadores importantes para identificar mudanças emocionais sutis. Em termos de qualidade de voz, a análise dos harmônicos e dos formantes permite capturar características da voz que estão relacionadas ao timbre e à expressividade. A representação dos dados da modal áudio disponibilizada pela base de dados contém 88 características (VALSTAR et al., 2016).

No vídeo, as características faciais foram extraídas em duas frentes: aparência e geometria. A análise de aparência usou o método LGBP-TOP (Local Gabor Binary Patterns from Three Orthogonal Planes), que examina a textura facial ao longo de três planos ortogonais (x-y, x-t e y-t). Esse método divide cada frame em pequenos volumes e aplica filtros de Gabor 2D em cada um deles para capturar padrões de textura, que são depois processados com a técnica de Local Binary Patterns (LBP). A Análise de Componentes Principais (do inglês, Principal Component Analysis - PCA) foi utilizada para reduzir a dimensionalidade dos dados, resultando numa representação dos dados com 168 características. Já a análise geométrica utilizou os 49 marcos faciais, pontos específicos do rosto, como olhos, sobrancelhas e boca, ilustrados na Figura 10. Estes marcos foram detectados frame a frame utilizando um detector de face baseado em regressão de gradiente em cascata (KAZEMI; SULLIVAN, 2014). Após a detecção, foram calculadas distâncias e ângulos entre os marcos faciais, fornecendo uma representação precisa das variações nas expressões faciais ao longo do tempo. A representação final dos dados geométricos contém 632 características (VALSTAR et al., 2016).

Figura 10 – Exemplo de detecção dos 49 marcos faciais utilizados na extração das características geométricas e de aparência.



Fonte: Autoria Própria

Os dados fisiológicos foram coletados através de sinais de ECG (eletrocardiograma) e EDA (atividade eletrodérmica), ambos conhecidos por estarem relacionados a respostas

emocionais. O ECG passou por um processo de filtragem e dele foram extraídas 19 características, incluindo taxas de cruzamento em zero, momentos estatísticos e entropia espectral. Além disso, foi extraída a frequência cardíaca (do inglês, Heart Rate - HR) e sua medida de variabilidade (do inglês, Heart Rate Variability - HRV) do sinal de ECG filtrado, cada uma com 10 características. O EDA, por sua vez, foi dividido em resposta de condutância da pele (do inglês, Skin Conductance Response - SCR) e nível basal de condutância (do inglês, Skin Conductance Level - SCL). Cada uma dessas divisões forneceu 8 características, incluindo medidas estatísticas e derivadas temporais (VALSTAR et al., 2016).

Para o reconhecimento das emoções, os autores utilizaram Support Vector Machine (SVM) para regressão, com otimização de parâmetros baseada em um grid search para maximizar o CCC. Os modelos foram combinados por regressão linear dos resultados unimodais para produzir a predição final multimodal, que se beneficiou da fusão de todas as modalidades para uma análise emocional robusta. A Tabela 1 apresenta os resultados base em termos de CCC para o reconhecimento contínuo de emoções no conjunto de testes da base de dados RECOLA para representações de áudio, vídeo (aparência e geometria) e dados fisiológicos (ECG, HRHRV, EDA, SCL e SCR), e sua fusão multimodal.

Tabela 1 – Resultados base para reconhecimento de emoções na base RECOLA para representações de áudio, vídeo (aparência e geometria) e dados fisiológicos (ECG, HRHRV, EDA, SCL e SCR), e sua fusão multimodal. O desempenho é medido em Coeficiente de Correlação de Concordância (CCC).

Modalidade	Ativação (CCC)	Valência (CCC)
Áudio	0,648	0,375
Aparência Vídeo	0,343	0,486
Geometria Vídeo	0,272	0,507
ECG	0,158	0,121
HRHRV	0,334	0,198
EDA	0,075	0,228
SCL	0,066	0,215
SCR	0,065	0,145
Multimodal	**0,683**	**0,639**

Fonte: Valstar et al. (2016)

A análise monomodal mostrou que o áudio teve a melhor performance em detectar ativação, com suas características prosódicas e de qualidade de voz capturando bem a intensidade emocional. O vídeo, por outro lado, destacou-se na predição da valência, onde expressões faciais são cruciais para capturar sentimentos de prazer ou desagrado. Quando as modalidades foram integradas, houve um aumento na precisão, evidenciando que a fusão dos diferentes modais e representações de dados melhora significativamente a precisão na regressão contínua de emoções.

2.5.2 SEWA

A base de dados SEWA (Sentiment Analysis in the Wild) (KOSSAIFI et al., 2021) foi projetada para superar limitações comumente observadas em outras bases de dados existentes, que frequentemente dependem de condições controladas de coleta, baixo espectro demográfico e anotações limitadas a poucas dimensões afetivas. Essa base é composta por mais de 30 horas de dados audiovisuais de 398 participantes provenientes de seis culturas diferentes: Britânica, Alemã, Húngara, Grega, Sérvia e Chinesa. Destes, 50% são mulheres, e os participantes são uniformemente distribuídos em cinco faixas etárias (18-29, 30-39, 40-49, 50-59, 60+).

Os dados foram coletados em dois contextos principais: enquanto os participantes assistiam a vídeos publicitários projetados para provocar estados emocionais variados, como prazer, desgosto e confusão; e enquanto discutiam suas opiniões sobre os vídeos em uma conversa por videochamada. Essa abordagem permitiu registrar comportamentos espontâneos em ambientes não controlados, utilizando webcams e microfones comuns. Este conjunto de dados foi utilizado para benchmarking de sistemas de reconhecimento de emoção cross-culturais nas competições: AVEC'17 (RINGEVAL et al., 2017), AVEC'18 (RINGEVAL et al., 2018) e AVEC'19 (RINGEVAL et al., 2019).

Ao longo dos anos, a base SEWA foi expandida e aprimorada. Temos acesso à versão utilizada no AVEC'17, que inclui rotulações detalhadas da interação de 48 participantes. A base de dados oferece anotações em várias dimensões, como ativação, valência, apreciação, sinais faciais, vocais e sociais. Essas características tornam a SEWA um recurso único para a análise multicultural e multimodal de emoções. A base de dados SEWA apresenta diferenças substanciais em relação à base RECOLA, destacando-se por sua coleta em ambientes não controlados ("in-the-wild") e pela inclusão de diversas culturas. A Figura 11 apresenta alguns frames de vídeo da base de dados SEWA e as Figuras 12 e 13 são exemplos da rotulação de ativação e valência para um vídeo da base.

Os experimentos baseline da SEWA avaliaram o reconhecimento de ativação e valência utilizando quatro abordagens principais: SVM, RNN, Random Forest (RF) e redes neurais convolucionais profundas (do inglês, Deep Convolutional Neural Networks - DCNN). Cada modelo foi aplicado a conjuntos de dados contendo informações extraídas das modalidades de áudio, vídeo e a combinação de ambas.

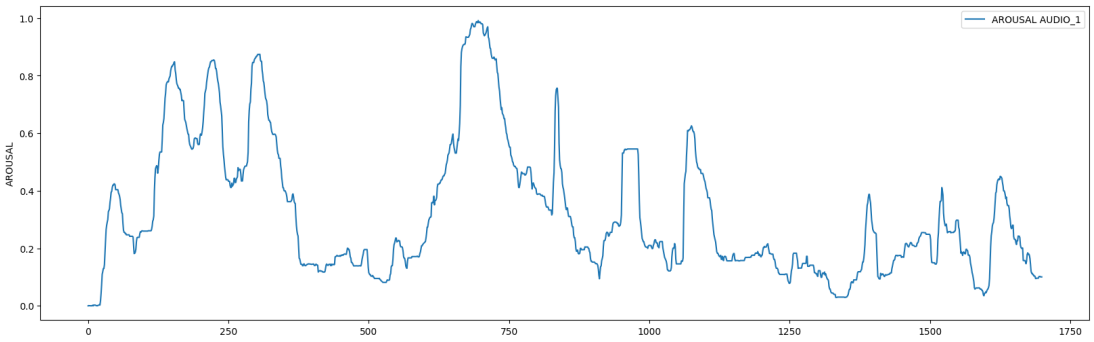
Para o áudio, as mesmas 88 características da base de dados RECOLA foram extraídas utilizando o conjunto eGEMAPS, implementado através da biblioteca openSMILE. Para o vídeo, as características faciais foram extraídas em duas frentes: aparência e geometria. A análise de aparência utilizou Dense SIFT (Scale-Invariant Feature Transform), que captura padrões locais de textura em regiões específicas do rosto. Foram extraídas características Dense SIFT a partir de janelas de 11×11 pixels ao redor dos 49 marcos fa-

ciais, extraindo informações que refletem variações sutis na aparência. A dimensionalidade dessas características foi reduzida por meio de PCA, retendo os 300 componentes mais relevantes. A representação geométrica considera distâncias e ângulos entre os pares de marcos faciais, capturando mudanças na forma e nos movimentos faciais. Essa abordagem permite representar expressões dinâmicas, como sorrisos, movimentos de sobrancelhas e abertura da boca. A representação dos dados geométricos contém 121 características (KOSSAIFI et al., 2021).

Figura 11 – Exemplo frames de vídeo SEWA.

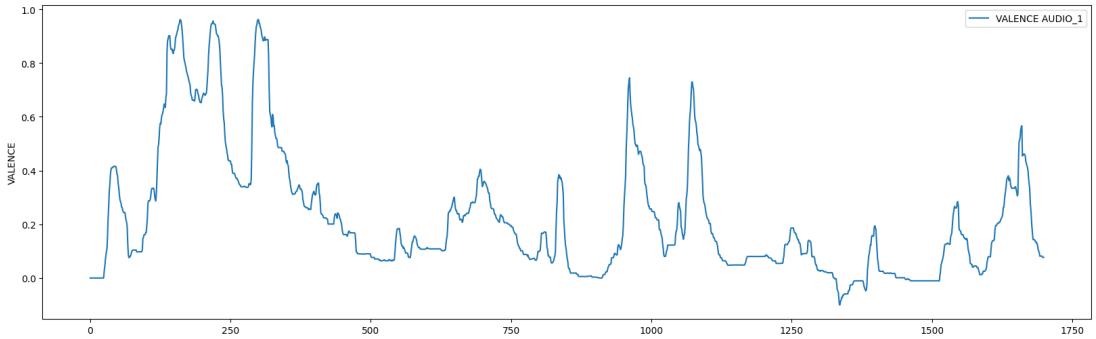
Fonte: Kossaifi et al. (2021)

Figura 12 – Exemplo de rotulação da dimensão de ativação na base de dados SEWA. O eixo vertical representa os valores contínuos da dimensão de ativação (arousal), enquanto o eixo horizontal indica o tempo, medido em frames. Os dados referem-se a modalidade de áudio do participante T1 da base de dados SEWA.



Fonte: Autoria Própria

Figura 13 – Exemplo de rotulação da dimensão de valência na base de dados SEWA. O eixo vertical representa os valores contínuos da dimensão de valência (valence), enquanto o eixo horizontal indica o tempo, medido em frames. Os dados referem-se a modalidade de áudio do participante T1 da base de dados SEWA.



Fonte: Autoria Própria

Para representar a modalidade de vídeo, ambas as abordagens, de aparência e de geometria, foram concatenadas para fornecer uma descrição robusta do comportamento facial ao longo do tempo. A integração entre áudio e vídeo foi realizada através de fusão antecipada (concatenação das características, resultando em um único vetor de 509 características), criando representações multimodais que combinam informações complementares de ambas as modalidades. Os modelos foram avaliados com base no CCC, conforme apresentado na Tabela 2.

Tabela 2 – Resultados base para reconhecimento de emoções na base SEWA para representações de áudio, vídeo e sua fusão multimodal. O desempenho é medido em Coeficiente de Correlação de Concordância (CCC).

Modalidade	Modelo	Ativação (CCC)	Valência (CCC)
Áudio	SVM	0,262	0,196
	RF	0,223	0,078
	LSTM-RNN	0,234	0,082
Vídeo	SVM	0,172	0,194
	RF	0,127	0,192
	LSTM-RNN	0,119	0,153
	DCNN (ResNet-18)	0,290	0,350
Multimodal	SVM	0,175	0,199
	RF	0,125	0,137
	LSTM-RNN	0,162	0,195

Fonte: Kossaifi et al. (2021)

Os resultados da tabela demonstram que, nas modalidades isoladas de áudio e vídeo, modelos específicos como o SVM (para áudio) e o DCNN (para vídeo) alcançaram os melhores desempenhos. O SVM obteve um CCC de 0,262 para ativação e 0,196 para valência na modalidade áudio, enquanto o DCNN destacou-se na modalidade vídeo com um CCC de 0,290 para ativação e 0,350 para valência. Contudo, na modalidade multimodal,

onde se esperaria um aumento na eficácia devido à integração de áudio e vídeo, os modelos não demonstraram melhorias significativas. Os resultados foram menos expressivos em comparação com a base de dados RECOLA, o que pode ser atribuído à complexidade inerente da base SEWA, que contém dados coletados em ambientes não controlados e abrange diversas culturas.

2.6 Considerações Finais

A vantagem da seleção dinâmica é sua capacidade de adaptar a escolha de modelos às características locais dos dados, melhorando a precisão preditiva em comparação com métodos estáticos. Diversos estudos têm explorado a seleção dinâmica em diferentes contextos e destacam a sua eficácia em tarefas de classificação e regressão (BRITTO ALCEU S.; SABOURIN; OLIVEIRA, 2014; CRUZ; SABOURIN; CAVALCANTI, 2018; MOURA; CAVALCANTI; OLIVEIRA, 2021; MOURA; CAVALCANTI; OLIVEIRA, 2019). Esses trabalhos reforçam a ideia de que a seleção dinâmica pode superar as abordagens de seleção estática, especialmente em cenários onde a variabilidade dos dados é alta.

As redes LSTM e suas variantes são ferramentas utilizadas em problemas de modelagem sequencial, onde é necessário capturar e gerenciar dependências temporais de curto e longo prazo. Com seu controle detalhado de fluxo de informações através dos gates de entrada, esquecimento e saída, LSTMs têm se postado como uma solução robusta para tarefas de predição de séries temporais e processamento de dados sequenciais (LINDEMANN et al., 2021; YU et al., 2019; ZHANG et al., 2019; ALBADAWY; KIM, 2018; TZIRAKIS et al., 2017). Isso reforça seu valor em uma ampla variedade de aplicações, que abrangem desde predições de séries temporais até traduções automáticas e reconhecimento de emoções.

Por sua capacidade de capturar relações complexas nos dados, os mecanismos de atenção tornaram-se um componente essencial na modelagem de sequências, sendo aplicados em áreas como geração de texto, resumo automático e análise de sentimentos, mostrando sua capacidade de trabalhar com diferentes tipos de dados sequenciais (BRAUWERS; FRASINCAR, 2023; VAZQUEZ-RODRIGUEZ et al., 2023; NGUYEN et al., 2023; LI et al., 2024; CHENG et al., 2024; LIU et al., 2024). Em particular, o modelo Transformer, baseado inteiramente em mecanismos de atenção, tem alcançado resultados promissores em benchmarks de tradução e em outras tarefas de processamento de linguagem natural, consolidando-se como uma arquitetura altamente eficiente (VASWANI et al., 2017).

Este capítulo também explorou as principais teorias de emoção e apresentou as bases de dados RECOLA e SEWA, que são amplamente utilizadas para a validação de modelos MER. A base RECOLA destaca-se por sua abrangência em múltiplas modalidades, incluindo áudio, vídeo e sinais fisiológicos. Por outro lado, a SEWA traz como diferencial sua coleta em ambientes naturais (“in-the-wild”), com dados mais diversificados.

3 Estado da Arte

O campo do MER tem sido objeto de pesquisa devido à sua aplicabilidade em diversos domínios, como saúde mental, interação humano-computador, educação, entretenimento e sistemas de vigilância (PRAVEEN; CARDINAL; GRANGER, 2023; CARDINAL et al., 2015; ORTEGA et al., 2018; ORTEGA; CARDINAL; KOERICH, 2019; PRAVEEN et al., 2022; GLADYS; VETRISELVI, 2023; ASLAM et al., 2023; LIU et al., 2023; CHENG et al., 2024). Este capítulo examina trabalhos sobre o reconhecimento de emoções utilizando múltiplas modalidades e o tratamento de modalidades ausentes durante a inferência.

3.1 Reconhecimento Multimodal de Emoções

Inicialmente, a compreensão das emoções humanas concentrou-se em abordagens unimodais, em que os sistemas inferiam estados emocionais a partir de uma única fonte de dados, como áudio ou imagens. Embora eficazes em contextos específicos, essas abordagens apresentaram limitações na precisão e na captura de nuances emocionais (UDAHEMUKA; DJOUANI; KURIEN, 2024). Por exemplo, modelos baseados em áudio podem capturar entonação e volume da fala, mas perdem informações sutis expressas através de gestos faciais. Analogamente, modelos visuais conseguem identificar expressões faciais, mas falham em captar emoções latentes, que podem ser detectadas através de sinais fisiológicos (D'MELLO; KORY, 2015a).

A comunicação humana engloba tanto a expressão verbal quanto a não verbal das emoções, destacando nossa dependência de pistas multimodais. Cada modalidade contribui de maneira única, aumenta a robustez do sistema e permite que o MER capture tanto emoções explícitas quanto emoções latentes (BALTRUSAITIS; AHUJA; MORENCY, 2019). Quando pronunciada com ênfase em uma voz profunda, uma simples palavra pode transmitir uma emoção mais intensa ao ouvinte. Emoções complexas, como o medo, são reconhecidas mais prontamente através da interação de expressões faciais e variações de tom de voz, em vez de se basear apenas nas transcrições das palavras faladas (GLADYS; VETRISELVI, 2023).

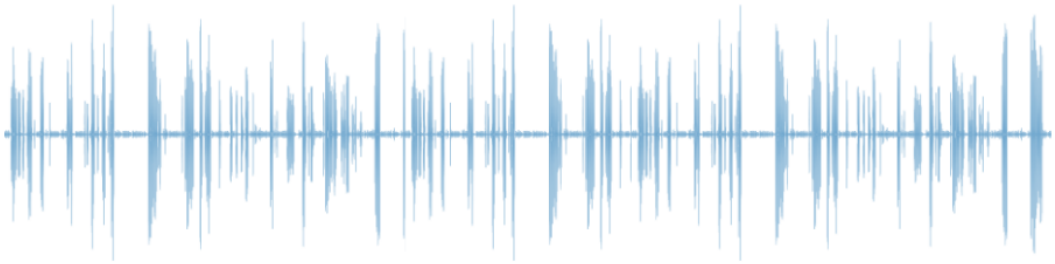
Modelos multimodais demonstraram aumentar a precisão e a confiabilidade quando comparados a modelos unimodais (GLADYS; VETRISELVI, 2023; LIU et al., 2023; D'MELLO; KORY, 2015a). Nesse contexto, o MER tem ganhado destaque, integrando dados de diferentes canais, como expressões faciais, voz, texto e sinais fisiológicos, para melhorar a precisão e fornecer uma visão mais completa dos estados emocionais. Entretanto, os dados multimodais apresentam um alto grau de complexidade e heterogeneidade, o que impõe desafios técnicos significativos ao seu uso em aprendizado de máquina. Entre esses

desafios, destacam-se a escolha das representações de dados mais adequadas, a correlação de informações intramodais e intermodais, o desalinhamento entre elementos de diferentes modalidades e a definição de estratégias eficazes de fusão (GLADYS; VETRISELVI, 2023).

3.1.1 Características e Limitações de Cada Modalidade

A análise de áudio no MER concentra-se em características prosódicas, como entonação, ritmo, intensidade da fala, duração e pausas, que refletem o nível de ativação emocional. Contudo, essa modalidade enfrenta desafios significativos. Em ambientes ruidosos ou em situações em que a expressividade vocal é limitada, a precisão dos modelos baseados em áudio diminui, requerendo algoritmos avançados de filtragem de ruído e técnicas de aprendizado que se adaptem à ausência total ou parcial dos dados.

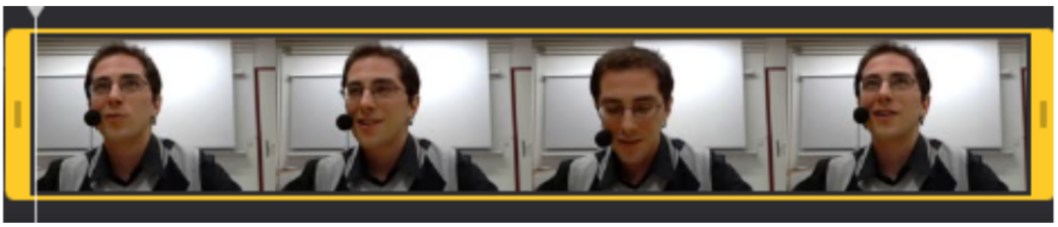
Figura 14 – Exemplo de sinal de áudio.



Fonte: Autoria Própria

A modalidade visual é uma fonte rica de dados emocionais, capturando desde expressões faciais até gestos e microexpressões. No entanto, a análise visual é suscetível a problemas como oclusões faciais e variações de iluminação, que reduzem a confiabilidade da detecção de emoções. Em situações onde o rosto está parcialmente coberto ou a iluminação é desfavorável, a modalidade visual pode não fornecer dados consistentes para o reconhecimento emocional, limitando o alcance de seus resultados.

Figura 15 – Exemplo de sinal de vídeo.



Fonte: Autoria Própria

Já dados fisiológicos, como ECG e EDA, fornecem dados internos menos suscetíveis a manipulações conscientes e são especialmente úteis em monitoramentos clínicos. Esses sinais são cruciais para capturar emoções em contextos onde as reações são sutis ou moduladas socialmente. Contudo, a coleta desses dados depende de dispositivos específicos, dificultando sua aplicação em larga escala e em contextos de uso diário.

Figura 16 – Exemplo de sinal de eletrocardiograma.

Fonte: Autoria Própria

3.2 Abordagens para o Reconhecimento Multimodal de Emoções

Técnicas como redes convolucionais e recorrentes e técnicas baseadas em mecanismos de atenção têm sido amplamente aplicadas em MER, cada uma com características específicas para tratar a temporalidade e a multimodalidade dos dados. As próximas seções examinam as principais abordagens para a modelagem temporal de emoções, com foco especial para experimentos realizados nas bases de dados RECOLA e SEWA. Um resumo de todos os trabalhos analisados, com as características utilizadas, arquitetura e resultados alcançados é apresentado nas Tabelas 3 e 4.

3.2.1 Modelos Clássicos de Aprendizado de Máquina

Weber et al. (2016) apresentam uma abordagem baseada nas características de áudio, vídeo (geometria e aparência) e dados fisiológicos (ECG, HRHRV, EDA, SCL e SCR) disponíveis na base RECOLA, complementados com informações referentes a pose da cabeça e assinaturas de expressão facial para o reconhecimento contínuo de emoções. As características adicionais propostas pelos autores permitem capturar movimentos faciais de maneira contínua e precisa, melhorando a compreensão das variações emocionais ao longo do tempo. O modelo de regressão escolhido foi o SVM linear. O processo se inicia com o treinamento de um regressor unimodal individual para cada sujeito da base de treinamento em cada modalidade. Em seguida, as predições geradas são combinadas através de um regressor SVM de fusão, treinado para mapear essas predições para as dimensões emocionais finais. Na base RECOLA, o modelo obteve um CCC de 0,663 para ativação e 0,645 para valência.

Sun et al. (2016) apresentam um método que envolve a extração de diversas características visuais complementares para um modelo de Support Vector Regression (SVR). As modalidades utilizadas foram áudio, vídeo (geometria e aparência) e sinais fisiológicos (ECG, EDA, SCL, SCR e HRV). Para o vídeo, os autores extraíram características adicionais como MSDF (Multi-scale Dense SIFT) e características baseadas nas arquiteturas AlexNet e ResNet, pré-treinadas no conjunto de dados FER e, em seguida, otimizadas na base RECOLA. As características extraídas foram usadas como entrada para o SVR, que previu ativação e valência. O processo de fusão multimodal foi realizado em nível de

decisão, combinando as predições unimodais através de uma regressão linear múltipla. O modelo alcançou CCCs de 0,683 para ativação e 0,642 para valência na base de dados RECOLA.

Ringeval et al. (2017) apresentam o baseline do desafio AVEC'2017, que utiliza a base SEWA para reconhecimento de emoções nas dimensões de ativação e valência. São utilizadas as características BoAW (Bag of Audio Words), BoVW (Bag of Video Words) e BoTW (Bag of Text Words) com um comprimento de segmento de 6 segundos e arquitetura SVR. Como as anotações contínuas sofrem um certo atraso, para compensar o tempo de reação dos anotadores, as características são deslocadas para frente por um atraso variável de 0 a 5 segundos. Na base SEWA, o modelo obteve um CCC de 0,306 para ativação e 0,466 para valência.

3.2.2 Modelos Baseados em Redes Convolucionais e Recorrentes

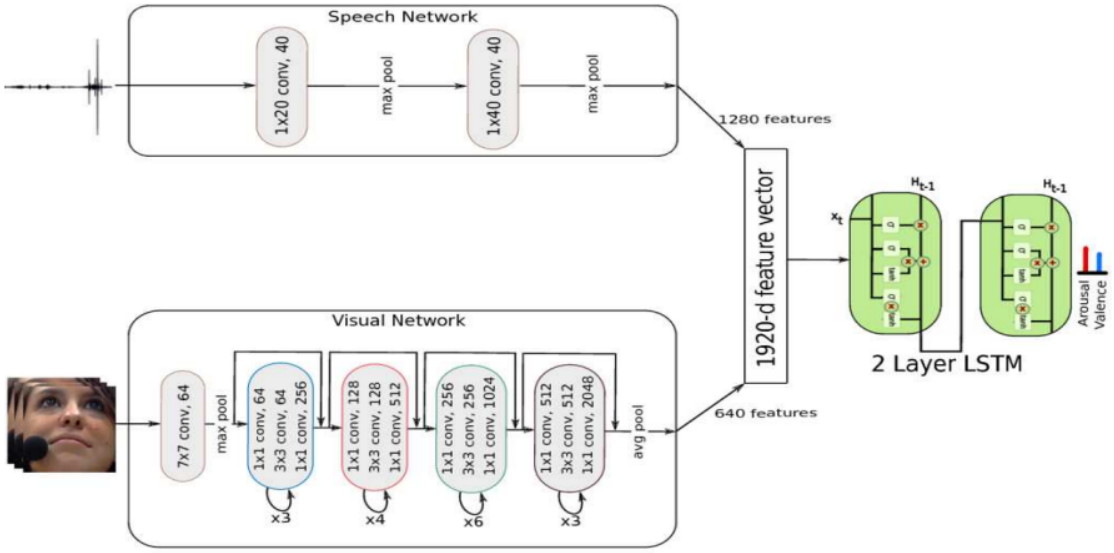
Tzirakis et al. (2017) apresentam uma abordagem de aprendizado end-to-end para reconhecimento emocional multimodal que permite processar as modalidades de áudio e vídeo de maneira integrada. O método utiliza a extração de características a partir dos sinais brutos de áudio e vídeo, com uma Rede Neural Convolucional (do inglês, Convolutional Neural Network - CNN) processando as informações auditivas e uma ResNet-50 processando os sinais visuais. O método utiliza a extração de características a partir dos sinais brutos de áudio e vídeo, com uma Rede Neural Convolucional (do inglês, Convolutional Neural Network – CNN) responsável pelo processamento das informações auditivas, enquanto uma ResNet-50 processa os sinais visuais.

Essas redes, inicialmente treinadas de forma independente, são combinadas em um estágio de treinamento de ponta a ponta em que as saídas das redes são concatenadas e passadas para duas camadas LSTM. Essas camadas são responsáveis por modelar as dependências temporais dos dados e fazer predições contínuas dos estados emocionais de ativação e valência. O modelo inclui a otimização do CCC como função objetivo durante o treinamento, em vez do tradicional MSE, para melhorar a precisão das predições. A Figura 17 exibe a arquitetura proposta pelos autores. Na base RECOLA, o modelo obteve um CCC de 0,714 para ativação e 0,612 para valência.

De forma similar, Keren et al. (2017) propõem um modelo de aprendizado end-to-end, um modelo no qual todas as etapas do processo, desde a extração de características até a predição final, são aprendidas simultaneamente durante o treinamento. Os autores utilizam CNNs para extrair representações diretamente dos dados brutos de ECG e EDA, seguidas RNNs para capturar dependências temporais de longo prazo. A combinação de CNNs e RNNs oferece uma vantagem ao sistema em dados temporais contínuos, onde o aprendizado direto dos dados brutos facilita a captura de complexidades emocionais que poderiam ser perdidas com um pré-processamento excessivo. O modelo alcançou CCCs de

0,430 para ativação e 0,407 para valência na base de dados RECOLA.

Figura 17 – Método proposto por Tzirakis et al. (2017). A rede é composta por duas partes: a parte de extração de características multimodais e a parte de processamento dos dados sequenciais. A parte multimodal extrai características de sinais brutos de áudio e vídeo. As características extraídas são concatenadas e usadas para alimentar 2 camadas LSTM, utilizadas para capturar a informação contextual nos dados.



Fonte: Tzirakis et al. (2017)

He et al. (2015b) exploram uma rede neural Deep Bidirectional Long Short-Term Memory RNNs (DBLSTM-RNN) para prever estados afetivos contínuos de ativação e valência a partir de entradas de áudio, vídeo e sinais fisiológicos. A natureza bidirecional do modelo garante uma visão completa das sequências temporais, permitindo a predição contínua e precisa de ativação e valência. Foram utilizadas as características de áudio, vídeo, ECG e EDA disponíveis na base de dados RECOLA, além de outros descritores de baixo nível de áudio extraídos pelos autores por meio da ferramenta YAAFE, características de quantização de fase local em três planos ortogonais dos vídeos (LPQ-TOP) e características complementares relacionadas aos dados de ECG e EDA. Os recursos utilizados incluem 4684 características de áudio, 768 características de vídeo e características fisiológicas que totalizam 106 para ECG e 82 para EDA. O método empregado consiste no treinamento das redes DBLSTM-RNN, seguido por uma suavização Gaussiana das predições. As predições suavizadas são então passadas para uma segunda camada de redes DBLSTM-RNN para refinar a predição final. Na base RECOLA, o modelo obteve um CCC de 0,747 para ativação e 0,609 para valência.

Chao et al. (2015) utilizam um modelo LSTM-RNN com uma arquitetura composta por uma camada oculta, uma camada de pooling temporal, duas camadas LSTM e uma camada linear de predição. O modelo é aprimorado por uma função de perda ϵ-insensitive, que permite ignorar pequenas flutuações nas predições de emoções, reduzindo a sensibilidade do modelo a ruído nos rótulos, problema comum em anotações de emoções dimensionais. Outro ponto interessante do trabalho é o uso de temporal pooling, que acumula informações contextuais de múltiplos quadros, consolidando-as em uma única predição e representando um estado emocional estável. Essa técnica ajuda a suavizar transições bruscas e melhora a robustez das predições. Foram utilizadas características de áudio, vídeo (geometria e aparência) e sinais fisiológicos (ECG e EDA). O modelo alcançou CCCs de 0,716 para ativação e 0,618 para valência na base de dados RECOLA.

Khorram, McInnis e Provost (2021) apresentam uma nova rede neural convolucional (denominada Multi-delay Sinc Network) para predição conjunta de anotações contínuas de emoção a partir de características log Mel-Frequency Bank (MFB). O método aborda o problema de dessincronização entre os rótulos de emoção e o sinal de fala devido aos atrasos causados pelo tempo de reação inerente às avaliações humanas. A rede proposta é composta por camadas convolucionais empilhadas, seguidas por uma rede de alinhamento. O componente chave é uma nova camada convolucional, chamada de camada sinc atrasada. Esta camada implementa um filtro passa-baixa (sinc) deslocado no tempo que aprende um único atraso usando um algoritmo baseado em gradiente. O método foi avaliado em dois conjuntos de dados de emoção, RECOLA e SEWA, aprendendo atrasos variáveis no tempo enquanto prediz descritores dimensionais de ativação e valência. Os resultados obtidos foram de 0,688 e 0,492 de CCC para ativação e valência na base RECOLA e 0,412 e 0,379 de CCC para ativação e valência na base SEWA.

Cruz et al. (2024) apresentam um método de aumento de dados baseado em Modelos Morfáveis 3D (do inglês, 3D Morphable Models - 3DMM). O método de aumento de dados cria sequências sintéticas de expressões faciais para valores sub-representados no espaço ativação-valência do conjunto de dados SEWA. Para a predição é utilizada uma rede bidirecional de Unidade Recorrente com Portão (do inglês, Gated Recurrent Unit - GRU) com duas camadas de tamanho 128. A entrada do modelo é uma sequência de 100 quadros (dois segundos), e a saída é uma tupla de dois valores representando a ativação e a valência da sequência. O modelo é treinado usando uma combinação de MSE, coeficiente de correlação produto-momento de Pearson e CCC como função de perda. No conjunto de dados SEWA, o modelo proposto alcança um CCC de 0,799 para ativação e 0,788 para valência.

Chen et al. (2019) investigam características espaço-temporais profundas. Para vídeo, os autores extraem características de aparência espaço-temporais combinando uma 2D-CNN (ResNet50 pré-treinada no conjunto de dados FER+) e uma 1D-CNN para

capturar informações temporais. Além disso, características geométricas espaço-temporais são extraídas dos landmarks faciais usando uma Rede Convolucional de Grafo Espaço-Temporal (do inglês, Spatial-Temporal Graph Convolutional Network - ST-GCN). Para áudio, mel espectrogramas são processados por uma rede VGGish pré-treinada, seguida por uma 1D-CNN para capturar a dinâmica temporal. Uma rede DBLSTM (Deep Bidirectional Long Short-Term Memory) de 2 camadas com 128 células é usada para prever afeto a partir de fluxos de características individuais (aparência, geometria, espectro de áudio) e da fusão antecipada das características. Finalmente, as saídas dos DBLSTMs unimodais e de fusão antecipada são alimentadas em um DBLSTM de fusão de decisão (2 camadas com 64 células cada) para obter as predições finais. No conjunto de teste SEWA, o método alcança um CCC de 0,513 para ativação e 0,515 para valência.

Zhao et al. (2018) defendem a ideia de que o estado emocional de uma pessoa é influenciado pelo comportamento do interlocutor. O método proposto pelos autores emprega fusão antecipada, concatenando as características de diferentes modalidades antes de alimentá-las em uma rede LSTM para capturar informações temporais de longo prazo. As características acústicas são extraídas usando o modelo VGGish, que é treinado em um grande conjunto de dados e pode aprender representações de áudio mais ricas. Para as características visuais, são comparados dois tipos de características faciais extraídas de diferentes CNNs pré-treinadas no conjunto de dados de reconhecimento de expressão facial FER+. As características textuais são extraídas de modelos de word embedding pré-treinados. Várias estratégias de interação multimodal são propostas para modelar os padrões de interação diádica, considerando as informações do interlocutor de diferentes modalidades (áudio, expressão facial e texto). A estratégia que se mostrou mais eficaz combina as características acústicas do interlocutor quando este está falando e ignora as informações faciais do interlocutor ("ATFAT Interaction"). Na partição alemã da base de dados SEWA, esta estratégia alcançou um CCC de 0,704 para ativação e 0,783 para valência.

3.2.3 Modelos Baseados em Autoencoders

Zhang et al. (2019) propõe um método inovador chamado Dynamic Difficulty Awareness Training (DDAT), onde um autoencoder é utilizado como camada de pré-processamento para melhorar a qualidade dos dados em ambientes ruidosos. O DDAT se baseia em dois tipos de indicadores de dificuldade: Erro de reconstrução do autoencoder e percepção de incerteza (nível de discordância entre anotadores). As características de dificuldade são concatenadas aos dados originais, enriquecendo o conjunto de entradas com um vetor de dificuldade que muda ao longo do tempo. A abordagem Multi-Task Learning (MTL) é usada para treinar redes LSTM simultaneamente na tarefa principal (reconhecimento de emoções) e nas tarefas auxiliares (reconstrução de entrada ou predição

de incerteza). Os autores não apresentaram resultados multimodais, apenas para as representações específicas de áudio e vídeo (geometria e aparência).

De forma similar AlBadawy e Kim (2018) propõe uma arquitetura DBLSTM, uma solução multi-task que otimiza ao mesmo tempo as perdas de classificação e regressão de emoções. O modelo de classificação utiliza rótulos discretizados gerados via clustering k-means, enquanto o modelo de regressão usa rótulos contínuos de ativação e valência. Na base RECOLA, o modelo obteve um CCC de 0,699 para ativação e 0,617 para valência.

Aspandi et al. (2020) exploram o uso de características latentes, representações internas de alto nível aprendidas automaticamente a partir dos dados brutos, extraídas de forma adversarial para melhorar as estimativas de modelos relacionados a afeto. O método proposto, denominado Latent-Based Adversarial Neural Networks, utiliza duas sub-redes principais: um Gerador baseado em Autoencoder (do inglês, Autoencoder-Based Generator - AEG) e um estimador de afeto baseado em Discriminador Condicional (do inglês, Conditional Discriminator - CD). O AEG recebe imagens ruidosas como entrada e gera versões limpas dessas imagens, enquanto extrai características latentes robustas. O CD, por sua vez, recebe as imagens limpas, as características latentes do AEG e características de áudio como entrada. Ele é treinado para discriminar entre imagens reais e falsas (geradas pelo AEG) e, simultaneamente, estimar os valores de ativação e valência. As características de áudio são extraídas usando os descritores de baixo nível do conjunto de características eGeMAPS e combinadas com as características faciais através de fusão tardia. As características latentes e de áudio são concatenadas com os kernels intermediários do CD. O CD utiliza um classificador adicional em cima do classificador principal para detectar o status real/falso e estimar os valores de ativação e valência. O modelo é treinado progressivamente, primeiro treinando o gerador e o discriminador juntos e, em seguida, adicionando as características latentes e de áudio. Na SEWA, o modelo alcançou um CCC de 0,430 para ativação e 0,405 para valência.

Zhao et al. (2019) utilizam uma abordagem de adaptação adversarial de domínio não supervisionada para reduzir a lacuna entre culturas no reconhecimento de emoções. O método utiliza uma arquitetura que consiste em um codificador de características (E), um regressor de emoções (R) e um classificador de cultura (C). O codificador E, implementado com uma rede LSTM, transforma as características multimodais (acústicas, visuais e textuais) em uma representação latente. O regressor R, baseado em camadas totalmente conectadas, prediz as emoções a partir da representação latente. O classificador C, também baseado em camadas totalmente conectadas, classifica a cultura da representação latente. O treinamento é realizado de forma adversarial, onde E é treinado para melhorar o desempenho do regressor R e confundir o classificador C, resultando em uma representação latente mais focada em informações emocionais e menos em informações culturais. As características multimodais utilizadas incluem VGGish, um modelo pré-treinado em um

grande conjunto de dados de áudio, que extrai representações de áudio ricas. Características faciais extraídas de redes CNN pré-treinadas em um conjunto de dados de reconhecimento de expressões faciais (FER+), incluindo VGGFace e DenseFace. Vetores de palavras (word embeddings) pré-treinados em diferentes modelos, incluindo "de.word2vec", "en.word2vec", "en.gensim.word2vec"e "cn.word2vec". A função de perda para E é uma combinação de MSE e CCC. Os resultados obtidos na partição alemã da base de dados SEWA foram de 0,689 e 0,766 de CCC para ativação e valência.

3.2.4 Mecanismos de Atenção

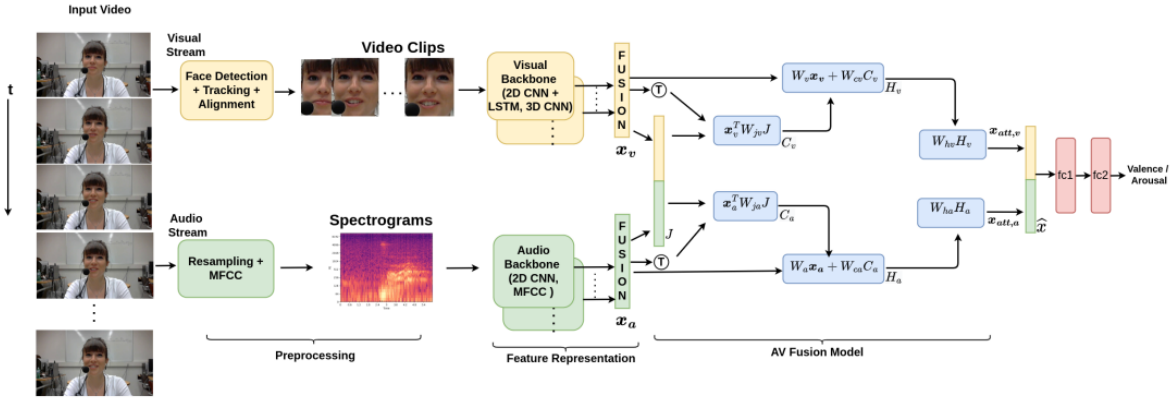
Praveen, Cardinal e Granger (2023) propõe um modelo de fusão audio-visual (A-V) baseado em atenção cruzada para o reconhecimento de emoções em termos de ativação e valência. O modelo utiliza características extraídas de redes neurais profundas para as modalidades de áudio e vídeo, que são, posteriormente, integradas em um módulo de atenção cruzada. A Figura 18 exibe a arquitetura proposta pelos autores. Inicialmente, as redes são treinadas separadamente para extrair características profundas de cada modalidade: para o vídeo, foi utilizada uma rede Inflated 3D-CNN (I3D) treinada com dados de expressões faciais; para o áudio, redes 2D-CNN foram aplicadas em espectrogramas. As matrizes de correlação cruzada entre a representação conjunta das características A-V e as características individuais de áudio e vídeo ajudam a destacar a relevância semântica entre as modalidades, melhorando a robustez da fusão e a captura de relações complementares. O resultado final é a concatenação das características ponderadas pela atenção, seguida por camadas totalmente conectadas para prever os valores contínuos de ativação e valência. O modelo alcançou CCCs de 0,842 para ativação e 0,728 para valência na base de dados RECOLA.

Em termos gerais, o modelo de atenção cruzada, proposto por Praveen, Cardinal e Granger (2023) é configurado para receber duas sequências de dados das modalidades de áudio e vídeo, que são combinadas e processadas para gerar uma única predição de ativação ou valência.

Sejam os conjuntos de vetores de características X_a e X_v extraídos das modalidades de áudio (A) e vídeo (V) de uma subsequência de tamanho fixo S, onde $X_a = \{x_a^1, x_a^2, \ldots, x_a^L\} \in \mathbb{R}^{d_a \times L}$ e $X_v = \{x_v^1, x_v^2, \ldots, x_v^L\} \in \mathbb{R}^{d_v \times L}$. Aqui, L denota o número de clipes não sobrepostos retirados uniformemente de S. Por sua vez, d_a e d_v representam as dimensões das características de áudio e vídeo, respectivamente, onde x_a^l e x_v^l são vetores de áudio e vídeo para os clipes $l = 1, 2, \ldots, L$.

A representação conjunta das características de áudio e vídeo ($\mathbf{J}$) é obtida a partir da concatenação dos vetores de características de áudio e vídeo $\mathbf{J} = [\mathbf{X_a}; \mathbf{X_v}] \in \mathbb{R}^{d \times L}$, onde $d = d_a + d_v$ denota a dimensão das características concatenadas. A representação conjunta $\mathbf{J}$ de uma subsequência $\mathbf{S}$ é usada para focar a atenção nas representações unimodais $\mathbf{X_a}$ e

Figura 18 – Método proposto por Praveen, Cardinal e Granger (2023). A rede é composta por duas partes principais: a parte de extração de características multimodais e o módulo de fusão com mecanismo de atenção cruzada. A parte multimodal extrai características dos sinais brutos de áudio e vídeo usando backbones específicos (CNNs 2D/3D e MFCC). As características extraídas são combinadas em uma representação conjunta e processadas pelo módulo de atenção cruzada, que captura relações inter e intra-modais. As características ponderadas resultantes são concatenadas e passadas por camadas totalmente conectadas para prever os valores de ativação e valência.



Fonte: Praveen, Cardinal e Granger (2023)

$\mathbf{X_v}$. Nesse sentido, a matriz de correlação conjunta para as características de áudio $\mathbf{C_a}$ entre as características de áudio $\mathbf{X_a}$ e a representação $\mathbf{J}$ será dada pela expressão:

$$\mathbf{C}_\mathrm{a} = \tanh\left(\frac{\mathbf{X}_a^T \mathbf{W}_{ja} \mathbf{J}}{\sqrt{d}}\right) \tag{3.1}$$

onde $\mathbf{W}_{ja}$ representa a matriz de pesos treinável de dimensão $L \times L$. Da mesma forma, a matriz de correlação conjunta para as características de vídeo ($\mathbf{C_v}$) será dada pela expressão:

$$\mathbf{C}_\mathrm{v} = \tanh\left(\frac{\mathbf{X}_v^T \mathbf{W}_{jv} \mathbf{J}}{\sqrt{d}}\right) \tag{3.2}$$

As matrizes C_a e C_v representam a correlação conjunta entre os vetores de entrada $\mathbf{X}_a$ e $\mathbf{X}_v$ e a representação conjunta $\mathbf{J}$. É importante destacar que as matrizes de correlação C_a e C_v têm um significado semântico onde valores mais altos de C_a e C_v implicam alta correlação entre as modalidades de áudio e vídeo e dentro de cada modalidade.

Por sua vez, a modalidade de áudio $\mathbf{X}_a$ e a modalidade de vídeo $\mathbf{X}_v$ são combinadas com as matrizes de correlação conjunta C_a e C_v para calcular os mapas de atenção H_a e

H_v, respectivamente:

$$\mathbf{H}_{\mathrm{a}} = \mathrm{ReLU}(\mathbf{W}_a \mathbf{X}_a + \mathbf{W}_{ca} C_a^t) \tag{3.3}$$

$$\mathbf{H}_{\mathrm{v}} = \mathrm{ReLU}(\mathbf{W}_v \mathbf{X}_v + \mathbf{W}_{cv} C_v^t) \tag{3.4}$$

Os mapas de atenção são utilizados para capturar a atenção em cada modalidade de acordo com as seguintes expressões:

$$\mathbf{X}_{\mathrm{att},a} = \mathbf{W}_{ha} \mathbf{H}_a + \mathbf{X}_a \tag{3.5}$$

$$\mathbf{X}_{\mathrm{att},v} = \mathbf{W}_{hv} \mathbf{H}_v + \mathbf{X}_v \tag{3.6}$$

Finalmente, as matrizes de atenção $\mathbf{X}_{\mathrm{att},a}$ e $\mathbf{X}_{\mathrm{att},v}$ são concatenadas para obter a matriz de atenção dada por:

$$\mathbf{X}_{\mathrm{att}} = [\mathbf{X}_{\mathrm{att},a}; \mathbf{X}_{\mathrm{att},v}] \tag{3.7}$$

que é alimentada em uma camada densamente conectada que preverá os valores de ativação ou valência.

De forma similar, Chen e Jin (2017) propõe uma estratégia de fusão multimodal para predição de emoções dimensionais utilizando redes neurais recorrentes LSTM como base para capturar dependências de longo prazo juntamente com um método de fusão condicional por atenção, onde os pesos atribuídos às modalidades variam dinamicamente com base nos dados de entrada atuais e no histórico recente. As saídas das LSTMs para as modalidades de áudio e vídeo são combinadas por meio de um mecanismo de atenção, que recebe como entrada a concatenação das saídas e dos dados de entrada das modalidades. Esse mecanismo permite que o modelo adapte seus pesos de atenção para priorizar a modalidade mais confiável em tempo real. O treinamento inclui uma função de perda que leva em consideração a energia acústica e a presença de detecção facial, ajudando a guiar a atenção do modelo para a modalidade mais relevante em diferentes contextos. Foram utilizadas as características de áudio e vídeo (geometria e aparência) da base de dados RECOLA e o modelo obteve um CCC de 0,694 para valência; resultados de ativação não foram reportados pelos autores.

Tzirakis et al. (2021) propõem uma estrutura para reconhecimento de emoção da fala que utiliza informações semânticas e paralinguísticas. Para a extração de características semânticas, os autores treinam modelos Word2Vec e Speech2Vec e alinham seus espaços de representação. Características paralinguísticas são extraídas usando uma rede convolucional

recorrente composta por três camadas CNN 1-D com ReLU como função de ativação e operações de max-pooling entre elas. As características semânticas e paralinguísticas são então fundidas usando um mecanismo de atenção "desembaraçada". Esse mecanismo projeta os dois conjuntos de características no mesmo espaço vetorial e os funde usando atenção. Três camadas totalmente conectadas com ativação linear são aplicadas à saída fundida para desembaraçar o fluxo de informações por dimensão afetiva. Finalmente, outra camada de atenção funde as informações desembaraçadas para produzir uma representação unificada da entrada. Esta representação é então passada por um módulo LSTM para capturar a dinâmica temporal no sinal antes da predição final. No conjunto de dados SEWA, o modelo atinge um CCC de 0,429 para ativação e 0,503 para valência.

3.2.5 Filtros de Kalman

Somandepalli et al. (2016) introduzem um método onde as predições unimodais de ativação e valência são inicialmente obtidas usando regressores SVR e filtros de Kalman realizam a fusão tardia das predições unimodais, permitindo o rastreamento contínuo dos estados afetivos. O modelo também aproveita as correlações entre as dimensões de ativação e valência, utilizando a predição de ativação como uma característica adicional para melhorar a precisão da predição de valência. O método se baseia em um sistema dinâmico linear onde as predições unimodais são tratadas como observações ruidosas para os filtros de Kalman. Quatro tipos de filtros de Kalman são desenvolvidos: apenas áudio, apenas vídeo, áudio e vídeo, e apenas dados fisiológicos. A seleção do filtro apropriado em cada momento é feita com base em indicadores da presença de áudio e da visibilidade do rosto, permitindo que o sistema escolha filtros específicos conforme a disponibilidade das modalidades. Quando a voz ou o rosto não são detectados, o sistema alterna para um filtro de Kalman que não depende dessas entradas. Foram utilizadas as características de áudio, vídeo (geometria e aparência) e dados fisiológicos (SCL, SCR, HR e HRV) disponíveis na base de dados RECOLA, além de outros descritores como probabilidade de face e voz. Na base RECOLA, o modelo obteve um CCC de 0,703 para ativação e 0,681 para valência.

De forma similar, Brady et al. (2016) propuseram uma abordagem que combina dados de áudio, vídeo e sinais fisiológicos, onde cada modalidade foi processada de forma independente através de modelos de regressão baseados em redes LSTM e combinados através de um filtro de Kalman. Na base RECOLA, o modelo obteve um CCC de 0,770 para ativação e 0,687 para valência.

Tabela 3 – Resumo trabalhos relacionados, incluindo referência, características utilizadas, arquitetura e resultados na partição de testes da base de dados RECOLA. O resultado é apresentado em termos de CCC.

Referência	Características	Arquitetura	Resultados (CCC)
Weber et al. (2016)	Áudio, vídeo (geometria e aparência) e dados fisiológicos (ECG, HRHRV, EDA, SCL e SCR)	SVR	Ativação: 0,663 Valência: 0,645
Sun et al. (2016)	Áudio, vídeo (geometria e aparência) e dados fisiológicos (ECG, EDA, SCL, SCR e HRV), MSDF + características de vídeo extraídas a partir de redes pré-treinadas	SVR	Ativação: 0,683 Valência: 0,642
Tzirakis et al. (2017)	Dados brutos áudio e vídeo	CNN + LSTM	Ativação: 0,714 Valência: 0,612
Keren et al. (2017)	Dados brutos ECG e EDA	CNN + RNN	Ativação: 0,430 Valência: 0,407
He et al. (2015b)	Áudio, vídeo (geometria e aparência) e sinais fisiológicos (ECG e EDA) + características complementares extraídas pelos autores	DBLSTM-RNN	Ativação: 0,747 Valência: 0,609
Chao et al. (2015)	Áudio, vídeo (geometria e aparência) e sinais fisiológicos (ECG e EDA)	LSTM	Ativação: 0,716 Valência: 0,618
Khorram, McInnis e Provost (2021)	Áudio (Log MFB)	Multi-delay Sinc Network	Ativação: 0,688 Valência: 0,492
Zhang et al. (2019)	Áudio e vídeo (geometria e aparência)	LSTM	-
AlBadawy e Kim (2018)	Áudio e vídeo (geometria e aparência)	D-BLSTM	Ativação: 0,699 Valência: 0,617
Praveen, Cardinal e Granger (2023)	Dados brutos áudio e vídeo	I3D + CNN + Atenção Cruzada	Ativação: 0,842 Valência: 0,728
Cheng et al. (2024)	Áudio e vídeo (geometria e aparência)	LSTM-RNN + Atenção Cruzada	Valência: 0,664

Continuação da Tabela 3

Referência	Características	Arquitetura	Resultados (CCC)
Somandepalli et al. (2016)	Áudio, vídeo (geometria e aparência) e dados fisiológicos (SCL, SCR, HR e HRV) características complementares extraídas pelos autores	SVR + filtro de Kalman	Ativação: 0,663 Valência: 0,645
Brady et al. (2016)	Áudio e dados fisiológicos (ECG, SCL e SCR) + características complementares extraídas pelos autores	LSTM + filtro de Kalman	Ativação: 0,770 Valência: 0,687

Tabela 4 – Resumo trabalhos relacionados, incluindo referência, características utilizadas, arquitetura e resultados na base de dados SEWA. O resultado é apresentado em termos de CCC.

Referência	Características	Arquitetura	Resultados (CCC)
Ringeval et al. (2017)	BoAW, BoVW e BoTW	SVR	Ativação: 0,306 Valência: 0,466
Khorram, McInnis e Provost (2021)	Áudio (Log MFB)	Multi-delay Sinc Network	Ativação: 0,412 Valência: 0,379
Cruz et al. (2024)	Vídeo (aumento de dados)	GRU	Ativação: 0,799 Valência: 0,788
Chen et al. (2019)	Características de áudio e vídeo extraídas a partir de redes pré-treinadas	DBLSTM	Ativação: 0,513 Valência: 0,515
Zhao et al. (2018)	Características áudio, vídeo e texto extraídas a partir de redes pré-treinadas	LSTM	Ativação: 0,704 Valência: 0,783
Aspandi et al. (2020)	Áudio e vídeo	AEG	Ativação: 0,405 Valência: 0,430
Zhao et al. (2019)	Características áudio, vídeo e texto extraídas a partir de redes pré-treinadas	LSTM	Ativação: 0,689 Valência: 0,766
Tzirakis et al. (2021)	Áudio	RNN + Atenção	Ativação: 0,429 Valência: 0,503

3.3 Impacto da Ausência de Modalidades e Desafios

A ausência de uma ou mais modalidades durante o reconhecimento emocional é um desafio recorrente em sistemas multimodais. Essa falta de dados pode ocorrer devido a falhas técnicas, restrições de privacidade ou limitações ambientais. A falta de modalidades específicas, como vídeo ou áudio, afeta consideravelmente a precisão do MER, limitando a quantidade de informações disponíveis para a análise emocional e comprometendo a capacidade do sistema de gerar inferências precisas (D'MELLO; KORY, 2015b).

O tratamento de modalidades ausentes em abordagens multimodais frequentemente envolve três estratégias para se adaptar a tais situações (ASLAM et al., 2023; LIN; GONCALVES; BUSSO, 2023; LIN; HU, 2023; CHENG et al., 2024; VAZQUEZ-RODRIGUEZ et al., 2023): (i) imputar ou preencher modalidades ausentes usando imputação de dados ou aproveitando informações de outras modalidades disponíveis; (ii) projetar modelos que possam gerenciar graciosamente cenários onde certas modalidades estão ausentes, potencialmente aprendendo a depender mais das modalidades disponíveis; (iii) desenvolver modelos que possam se adaptar a modalidades variáveis ou a diferentes distribuições de dados, permitindo uma melhor generalização quando confrontados com modalidades ausentes.

Para enfrentar esses desafios, estratégias como imputação de dados, seleção dinâmica e atenção cruzada têm sido investigadas. A imputação de dados insere valores ausentes com base nas informações disponíveis, mas pode introduzir ruídos que afetam a precisão do modelo (ZHU; SUN; ZHOU, 2023). Mecanismos de atenção cruzada oferecem uma solução promissora ao focar seletivamente nas modalidades ativas, distribuindo o peso do processamento para as combinações mais informativas e aprimorando a precisão em cenários com dados faltantes (PRAVEEN; CARDINAL; GRANGER, 2023). Em contraste, a seleção dinâmica permite que o sistema identifique e priorize as modalidades disponíveis que oferecem as informações mais relevantes, compensando a ausência de dados específicos (MOURA; CAVALCANTI; OLIVEIRA, 2021).

3.3.1 Autoencoders Variacionais para Dados Incompletos

Silva-Filarder et al. (2021) afirmam que uma propriedade crítica dos modelos multimodais é ter abordagens eficientes para trabalhar com modalidades ausentes e permitir a geração cruzada. A geração cruzada envolve o uso de um subconjunto de modalidades de entrada para gerar modalidades de saída que estão ausentes no subconjunto de entrada. Os autores propõem dois métodos: geração cruzada exaustiva e eliminação de componentes latentes. A eliminação de componentes latentes é baseada em randomização e simula modalidades ausentes aplicando uma máscara de dropout a elementos individuais das variáveis latentes, uma para cada modalidade. Em seguida, ela recorre a um especialista

anterior para os elementos selecionados pela máscara. A geração cruzada exaustiva é um método baseado em força bruta que considera todos os subconjuntos possíveis de modalidades.

Zhu, Sun e Zhou (2023) propuseram uma característica invariante para uma rede de imaginação de modalidades ausentes (do inglês, Invariant Feature aware Missing Modality Imagination Network - IF-MMIN). Este método depende de dois codificadores: o codificador de especificidade, responsável por extrair características de alto nível dos recursos de entrada brutos, e o codificador de invariância, que recebe as características específicas da modalidade obtidas pelo codificador de especificidade como entrada para extrair características invariantes à modalidade. Além disso, o método incorpora um módulo de imaginação baseado em características invariantes (IF-IM) que prevê as características específicas da modalidade ausente com base na modalidade disponível e gera uma representação conjunta.

3.3.2 Mecanismos de Atenção e Ajuste Dinâmico

Vazquez-Rodriguez et al. (2023) propuseram uma arquitetura baseada em transformers para reconhecer ativação e valência de maneira contínua no tempo, mesmo com modalidades de entrada ausentes. Um transformer multimodal é utilizado como codificador para obter representações das diferentes modalidades, e um transformer decodificador é utilizado para processar essas representações e fazer predições. Um mecanismo de atenção cruzada do decodificador é utilizado para ponderar a importância de diferentes modalidades. O decodificador é autoregressivo, considerando valores preditos anteriores ao fazer a inferência atual, o que é essencial ao realizar predições contínuas no tempo.

Li et al. (2024) também abordam a questão das modalidades ausentes. Segundo os autores, esses dados ausentes prejudicam a extração de características dos dados multimodais, resultando em uma queda no desempenho do modelo e em resultados imprecisos. Portanto, é proposta uma rede de atenção múltipla com várias cabeças baseada em um codificador com modalidades ausentes (MMAN-M2). A atenção múltipla representa a modalidade com base na sequência inteira, e o mapeamento cruzado é utilizado para obter a relação entre as modalidades. Modalidades ausentes aleatórias são codificadas e combinadas com um módulo de otimização para aprimorar a associação entre modalidades ausentes e não ausentes. O módulo codificador e decodificador obtém informações globais e mapeia informações globais para vários espaços.

Outros trabalhos interessantes apresentam uma estrutura computacional multimodal baseada na arquitetura de transformers e mecanismos de atenção para o reconhecimento de emoções com dados incompletos (LIN; GONCALVES; BUSSO, 2023; CHENG et al., 2024; LIU et al., 2024; NGUYEN et al., 2023; PRAVEEN; CARDINAL; GRANGER, 2023).

3.3.3 Seleção Dinâmica

Menon et al. (2025) propõem um método de seleção dinâmica em duas etapas para otimizar o reconhecimento multimodal de emoções. A primeira etapa envolve a seleção da modalidade mais informativa, enquanto a segunda realiza uma seleção intra-modal da melhor representação dos dados. O método de seleção dinâmica de modalidades e visualizações baseia-se na avaliação de diferentes regressores, que são treinados individualmente para cada modalidade (áudio e vídeo). Na fase de seleção da modalidade mais informativa, um meta-classificador escolhe a melhor modalidade e, em seguida, a seleção dinâmica intra-modal é realizada para avaliar cada instância de teste. Resultados indicam que o método proposto é uma abordagem efetiva para o MER, mesmo em cenários com modalidades ausentes.

3.4 Considerações Finais

Em MER, a continuidade temporal é fundamental para modelar a evolução das emoções ao longo do tempo, refletindo como os estados emocionais mudam e se ajustam a diferentes estímulos. Em relação às técnicas utilizadas para resolver a tarefa de reconhecimento de emoções, redes LSTM têm se mostrado eficazes na captura de dependências temporais, aspectos críticos para o reconhecimento de emoções (TZIRAKIS et al., 2017; CHAO et al., 2015; ZHANG et al., 2019; ALBADAWY; KIM, 2018; BRADY et al., 2016; ZHAO et al., 2019). Sua capacidade de reter dependências de longo prazo as torna adequadas para modelar sequências de expressões emocionais ao longo do tempo, sendo amplamente utilizadas em problemas de regressão de ativação e valência.

Os mecanismos de atenção se destacam pela flexibilidade em integrar múltiplas modalidades e por focar seletivamente em dados relevantes em tempo real, aumentando a eficiência na análise, mesmo em condições não ideais. Contudo, essas vantagens vêm acompanhadas de alta complexidade computacional e dependência de grandes volumes de dados para atingir o desempenho ideal (PRAVEEN; CARDINAL; GRANGER, 2023; CHENG et al., 2024). Os filtros de Kalman criam representações compartilhadas e facilitam a inferência entre diferentes tipos de dados. Essa técnica proporciona resiliência e adapta-se bem a cenários com dados incompletos, mas também demanda consideráveis recursos computacionais (SOMANDEPALLI et al., 2016; BRADY et al., 2016).

Por outro lado, métodos que dependem de dados completos, como as redes neurais sem adaptação específica e as fusões que utilizam bottleneck features, são mais suscetíveis à perda de precisão na ausência de uma modalidade. Para melhorar a confiabilidade de sistemas multimodais em contextos reais, é essencial adotar estratégias de adaptação e fusão dinâmica, minimizando o impacto dos dados faltantes nas predições emocionais (D'MELLO; KORY, 2015b). Seleção dinâmica, mecanismos de atenção e as estratégias de

imputação de dados são métodos eficazes para gerenciar a ausência de dados em uma ou mais modalidades. Essas estratégias permitem que o modelo mantenha predições confiáveis e precisas mesmo em condições de dados incompletos, tornando-se indispensáveis para aplicações robustas em ambientes dinâmicos e desafiadores.

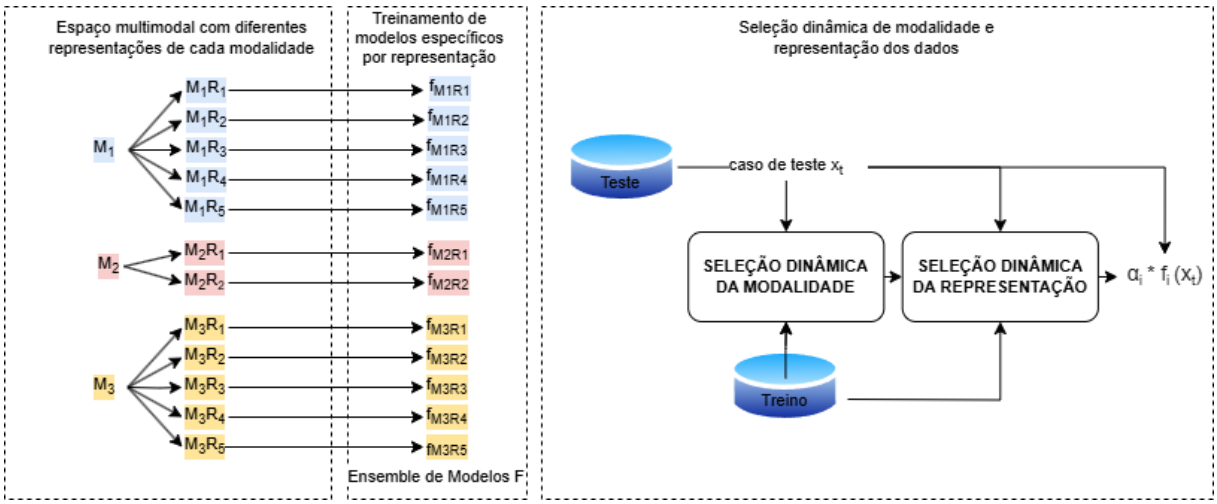
Apesar do desempenho impressionante das abordagens baseadas em mecanismos de atenção, o cálculo dos pesos de atenção nessas técnicas não considera a relação complementar entre as características de áudio e vídeo (PRAVEEN et al., 2022). Além disso, estudos recentes mostram que Transformers têm dificuldade em capturar dependências sequenciais em séries temporais (ZENG et al., 2023). Métodos baseados em seleção dinâmica fornecem melhor interpretabilidade e são computacionalmente mais eficientes do que a estrutura de atenção, que pode envolver todos os elementos da sequência. Finalmente, a seleção dinâmica é menos dependente de grandes conjuntos de dados para o treinamento.

A escolha da abordagem ideal para MER depende de um equilíbrio entre precisão, eficiência computacional e a capacidade de adaptação a dados ausentes ou incompletos. Cada técnica apresenta vantagens e desafios, e a decisão deve considerar as necessidades específicas do projeto e as restrições de recursos.

4 Método Proposto

Em sistemas multimodais, os dados são representados por diferentes modalidades M_1, M_2, ..., M_K, cada uma com suas características e contribuições específicas para a tarefa em questão. Dentro de cada modalidade M_i, podem existir múltiplas representações M_iR_1, M_iR_2, ..., M_iR_N, que capturam diferentes aspectos dos dados. O problema fundamental reside em selecionar, de maneira eficiente e adaptativa, a modalidade e as representações que maximizem o desempenho do sistema. Para isso, o método proposto consiste em duas etapas principais: seleção da modalidade M e seleção intra-modal da melhor representação dos dados $\mathcal{R}$. Esse processo é realizado de forma independente para cada uma das dimensões emocionais, ativação e valência. A Figura 19 apresenta uma visão geral do método proposto.

Figura 19 – Método de seleção dinâmica de modalidade e representação de dados. Todos os modelos são treinados separadamente e um conjunto de modelos é obtido. A melhor modalidade é selecionada na fase de seleção dinâmica da modalidade, e depois, na seleção dinâmica intra-modal da representação, cada modelo recebe um peso para avaliar cada caso de teste de acordo com sua assertividade na zona de competência.



Fonte: Autoria Própria

Esta abordagem em duas etapas permite que uma única modalidade M seja selecionada, enquanto uma ou mais representações associadas a ela $\mathcal{R}$ podem ser ponderadas, utilizando pesos α, para otimizar o desempenho geral do sistema. O objetivo do problema pode ser formalizado como a identificação da melhor combinação (M, $\mathcal{R}$), onde:

$$M \in M_1, M_2, ..., M_K \tag{4.1}$$

$$\mathcal{R} = \{(\mathrm{MR}_1, \alpha_1), (\mathrm{MR}_2, \alpha_2), \ldots, (\mathrm{MR}_N, \alpha_N)\} \tag{4.2}$$

com α_j representando o peso atribuído a cada representação R_j dentro da modalidade selecionada M. Esses pesos refletem a competência relativa de cada representação no contexto dos dados de entrada $\mathbf{x}_j^{\mathrm{ts}}$. O problema é, então, resolver:

$$\max_{\mathrm{M},\mathcal{R}} P(\mathrm{M}, \mathcal{R}, \mathbf{x}_j^{\mathrm{ts}}) \tag{4.3}$$

onde $P(\mathrm{M}, \mathcal{R}, \mathbf{x}_j^{\mathrm{ts}})$ é a medida de desempenho da combinação dos modelos treinados para a modalidade M e o subconjunto de representações $\mathcal{R}$, avaliada no contexto dos dados de entrada $\mathbf{x}_j^{\mathrm{ts}}$.

4.1 Seleção Dinâmica de Modalidade e Representação

O método de seleção dinâmica proposto envolve duas etapas principais: (i) escolha da modalidade mais adequada (M) e (ii) escolha da melhor representação dentro da modalidade selecionada (R). No entanto, para possibilitar essas etapas, é necessário, inicialmente, treinar um ensemble de modelos, que será utilizado como suporte para a tomada de decisão. Todo o processo é realizado de forma independente para cada dimensão emocional, com ensembles distintos e seleções específicas para ativação e valência.

4.1.1 Geração do Ensemble de Modelos ($\mathcal{F}$)

Inicialmente, um conjuntos de modelos $\mathcal{F} = \{f_1, f_2, \ldots, f_N\}$ é gerado, onde cada modelo é treinado com uma das combinações possíveis de modalidades e representações $\mathrm{M}_1\mathrm{R}_1$, $\mathrm{M}_1\mathrm{R}_2$, ..., $\mathrm{M}_K\mathrm{R}_N$.

Sem perda de generalização, definimos cada modelo f como um regressor, que recebe como entrada um vetor $\mathbf{x} \in \mathbb{R}^d$, contendo d características, e retorna como saída um valor escalar $y \in \mathbb{R}$. Assim, formalmente:

$$f : \mathbb{R}^d \to \mathbb{R}, \quad f(\mathbf{x}) = y \tag{4.4}$$

onde $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ representa o vetor de entrada com d características extraídas de uma representação específica e y é o valor estimado pelo modelo.

4.1.2 Seleção Dinâmica da Modalidade M

Na seleção dinâmica da modalidade, conforme descrito no Algoritmo 1, um meta-classificador é responsável por inferir, para cada instância de entrada, a modalidade mais apropriada a ser utilizada pelo sistema. Este classificador atua sobre o espaço de decisões

gerado pelas diferentes combinações de modalidades e representações e é formalmente definido como uma função:

$$h : \mathbb{R}^P \to \mathcal{M}, \quad h(\mathbf{z}) = \mathrm{M} \tag{4.5}$$

onde $\mathbf{z} = (z_1, z_2, \ldots, z_P) \in \mathbb{R}^P$ representa um vetor de entrada de P dimensões, contendo as predições individuais dos modelos associados a diferentes combinações de modalidades e representações $(\mathrm{M}_1\mathrm{R}_1, \mathrm{M}_1\mathrm{R}_2, ..., \mathrm{M}_K\mathrm{R}_N)$, $\mathcal{M}$ é o conjunto de modalidades possíveis, e a saída $\mathrm{M} \in \mathcal{M}$ é a modalidade selecionada para inferência final.

O meta-classificador adotado foi uma rede neural multicamadas (Multi-Layer Perceptron, MLP), configurada com uma única camada oculta de 50 neurônios e função de ativação softmax. Para prevenir overfitting, utilizou-se regularização ℓ_2 com coeficiente $\alpha = 10^{-5}$, e o otimizador Adam foi empregado na atualização dos pesos. O treinamento foi realizado com early stopping e taxa de aprendizado adaptativa, sendo interrompido após 100.000 iterações ou ausência de melhora na função de perda.

Além do MLP, avaliou-se o desempenho de um meta-classificador Naive Bayes (NB) como baseline simples e interpretável. Os resultados experimentais obtidos na base RECOLA, apresentados na Tabela 5, demonstram que o MLP supera consistentemente o NB na tarefa de escolha da modalidade mais apropriada.

Tabela 5 – Acurácia do meta-classificador responsável pela seleção dinâmica da modalidade mais apropriada, avaliado no conjunto de treino da base RECOLA. São considerados três cenários de disponibilidade de modalidades: áudio (A) e vídeo (V) disponíveis, áudio indisponível e vídeo indisponível.

	Ativação		Valência	
Modalidades	MLP	NB	MLP	NB
A e V disponíveis	0,99	0,67	0,72	0,70
100% V indisponível	0,90	0,73	0,76	0,62
100% A indisponível	0,79	0,61	0,89	0,72

Após o treinamento do meta-classificador no conjunto de treinamento $\mathcal{T}_r$, o conjunto de testes $\mathcal{T}_e$ é submetido a esse modelo, que infere a modalidade mais adequada para cada amostra de teste, denotada por M_j.

Algoritmo 1 Seleção Dinâmica da Modalidade M

Entrada: Conjunto de modalidades $\mathcal{M}$; Conjunto de modelos $\mathcal{F}$; conjunto de treinamento $\mathcal{T}_r$; conjunto de testes $\mathcal{T}_e$
Saída: M_j, a modalidade mais promissora para cada caso de teste $\mathbf{x}_j^{\text{ts}} \in \mathcal{T}_e$
Inicialize as listas $X \leftarrow []$ e $Y \leftarrow []$
foreach $\mathbf{x}_i^{tr}$ *em* $\mathcal{T}_r$ **do**
  Submeta $\mathbf{x}_i^{\text{tr}}$ a todos os modelos em $\mathcal{F}$
  Obtenha as predições de todos os modelos e armazene-as em X
  Calcule a concordância entre cada predição em X e o rótulo de $\mathbf{x}_i^{\text{tr}}$
  foreach M *em* $\mathcal{M}$ **do**
    Calcule a assertividade da modalidade M
  end
  Adicione a modalidade mais informativa a Y
end
Treine um meta-classificador para selecionar a modalidade mais informativa usando X e Y
Submeta o conjunto de testes $\mathcal{T}_e$ ao classificador treinado e obtenha a predição da modalidade mais promissora M_j para cada caso de teste $\mathbf{x}_j^{\text{ts}}$
return M_j

4.1.3 Seleção Dinâmica da Representação R

Dentro da modalidade M selecionada, as representações MR_1, MR_2, ..., MR_N são avaliadas com base em métricas de desempenho ou adequação à entrada $\mathbf{x}_j^{\text{ts}}$. A seleção intra-modal da melhor representação dos dados pode selecionar o modelo com o menor erro acumulado na região de competência Ψ, combinar todos os modelos ou um subconjunto deles, utilizando uma média ponderada, onde cada modelo f_i recebe um peso α_i proporcional ao seu desempenho na região Ψ.

A região de competência Ψ é definida como o conjunto formado pelos K-vizinhos mais próximos de $\mathbf{x}_j^{\text{ts}}$ no conjunto de treinamento. A medida de distância utilizada para encontrar Ψ é definida como a distância Euclidiana entre dois vetores de características. Para cada instância de teste $\mathbf{x}_j^{\text{ts}}$, determinamos Ψ calculando as distâncias $dist(\mathbf{x}_j^{\text{ts}}, \mathbf{x}_i^{\text{tr}})$ entre $\mathbf{x}_j^{\text{ts}}$ e todas as instâncias $\mathbf{x}_i^{\text{tr}}$ do conjunto de treinamento. Em seguida, essas distâncias são ordenadas em ordem crescente, e Ψ é definido como o conjunto dos K vizinhos mais próximos de $\mathbf{x}_j^{\text{ts}}$, conforme descrito no Algoritmo 2.

$$\text{dist}(\mathbf{x}_j^{\text{ts}}, \mathbf{x}_i^{\text{tr}}) = \sqrt{\sum_{k=1}^{K} (x_{j,k}^{\text{ts}} - x_{i,k}^{\text{tr}})^2} \tag{4.6}$$

$$\Psi_j = \{(\mathbf{x}_{(1)}^{\text{tr}}, \text{dist}_{(1)}), (\mathbf{x}_{(2)}^{\text{tr}}, \text{dist}_{(2)}), \ldots, (\mathbf{x}_{(K)}^{\text{tr}}, \text{dist}_{(K)})\} \tag{4.7}$$

Algoritmo 2 Cálculo da Região de Competência Ψ

Entrada: Instância de teste $\mathbf{x}_j^{\text{ts}}$; Conjunto de treino $\mathcal{T}_r$; Número de vizinhos K
Saída: Ψ_j, o conjunto de K vizinhos mais próximos de $\mathbf{x}_j^{\text{ts}}$
Inicialize a lista $\mathcal{D} \leftarrow []$
foreach $\mathbf{x}_i^{tr}$ *em* $\mathcal{T}_r$ **do**
  Calcule a distância $d_i \leftarrow \text{dist}(\mathbf{x}_j^{\text{ts}}, \mathbf{x}_i^{\text{tr}})$
  Adicione o par $(\mathbf{x}_i^{\text{tr}}, d_i)$ à lista $\mathcal{D}$
end
Ordene $\mathcal{D}$ em ordem crescente pelos valores d_i
Defina Ψ_j como os K primeiros elementos de $\mathcal{D}$
return Ψ_j

O conjunto de modelos $\mathcal{F} = \{f_1, f_2, \ldots, f_N\}$, previamente treinado, é utilizado para realizar a predição de $\mathbf{x}_j^{\text{ts}}$. A predição final realizada pelo ensemble f_{ens} representa uma combinação ponderada dos modelos pertencentes a $\mathcal{F}$. Para a modalidade M selecionada e suas representações R, a predição final é realizada como:

$$f_{\text{ens}}(\mathbf{x}) = \sum_{j=1}^{N} \alpha_j \cdot f_{MR_j}(\mathbf{x}) \tag{4.8}$$

onde N é o número total de modelos, $f_{MR_j}(\mathbf{x})$ é a predição gerada pelo modelo associado à representação MR_j e α_j o peso associado a esta representação.

O valor de α_j é determinado a partir da região de competência Ψ_j, que corresponde ao conjunto dos K vizinhos mais próximos da instância x_j^{ts} no conjunto de treinamento. Essa região é obtida com base na distância Euclidiana entre a instância de teste e as instâncias de treinamento, como descrito no Algoritmo 2. A ideia central da seleção dinâmica é avaliar o desempenho de cada modelo considerando o seu erro sobre essas instâncias da região de competência Ψ_j, e então atribuir maior peso aos modelos com melhor desempenho local.

Técnicas tradicionais de seleção dinâmica podem ser adaptadas para atender às especificidades dos problemas multimodais, como por exemplo seleção dinâmica (DS), ponderação dinâmica (DW) e seleção dinâmica com ponderação (DWS), conforme descrito no Algoritmo 3. Cada técnica utiliza a região de competência e calcula o peso dos modelos de forma distinta:

- Seleção Dinâmica (DS): seleciona, dentre todos os modelos relacionados à modalidade previamente escolhida, aquele que apresenta o menor erro acumulado ao prever as instâncias pertencentes à região de competência. Esse único modelo é então utilizado para gerar a predição.
- Ponderação Dinâmica (DW): considera todos os modelos da modalidade pré-selecionada, mas atribui a cada um deles um peso α_j proporcional ao seu desempenho local. Para

cada padrão de teste $\mathbf{x}_j^{\text{ts}}$, sua região de competência Ψ_j é calculada. Para cada item em Ψ_j, um peso d_k é calculado usando a Eq. 4.9, onde $\text{dist}(\mathbf{x}_j^{\text{ts}}, \mathbf{x}_k^{\text{tr}})$ é a distância Euclidiana entre o item na região de competência $\mathbf{x}_k^{\text{tr}} \in \Psi_j$ e o padrão de teste $\mathbf{x}_j^{\text{ts}}$. O vetor $(d_1, d_2, ..., d_k)$ é utilizado para calcular o peso α_i do modelo f_i usando a Eq. 4.10, onde N é o tamanho do conjunto de modelos da modalidade selecionada, k representa o índice do vizinho e $sqe_{k,i}$ é o erro quadrático do modelo i para $\mathbf{x}_k^{\text{tr}} \in \Psi_j$.

- Seleção Dinâmica com Ponderação (DWS): combina os dois princípios anteriores. Primeiro, seleciona um subconjunto de modelos, descartando os modelos com erro acumulado na metade superior do intervalo de erro $E_i > (E_{max} - E_{min})/2$. Em seguida, aplica a ponderação dinâmica apenas sobre esse subconjunto, utilizando os mesmos critérios de cálculo dos pesos da técnica DW (Eqs. 4.9 e 4.10).

Essas estratégias permitem que o sistema se adapte dinamicamente ao contexto local da instância de entrada, valorizando os modelos que demonstram maior competência em situações semelhantes observadas durante o treinamento e estão descritas com mais detalhes na Seção 2.1.3 "Combinação dos Modelos Selecionados", Fundamentação Teórica.

$$d_k = \frac{\left(\text{dist}(\mathbf{x}_j^{\text{ts}}, \mathbf{x}_k^{\text{tr}})\right)^{-1}}{\sum_{i=1}^{K} \left(\text{dist}(\mathbf{x}_j^{\text{ts}}, \mathbf{x}_i^{\text{tr}})\right)^{-1}} \tag{4.9}$$

$$\alpha_i = \frac{\left[\sum_{k=1}^{K} d_k \cdot sqe_{k,i}\right]^{-1}}{\sum_{n=1}^{N} \left[\sum_{k=1}^{K} d_k \cdot sqe_{k,i}\right]^{-1}} \tag{4.10}$$

Algoritmo 3 Seleção Dinâmica da Representação de Dados M_R^*

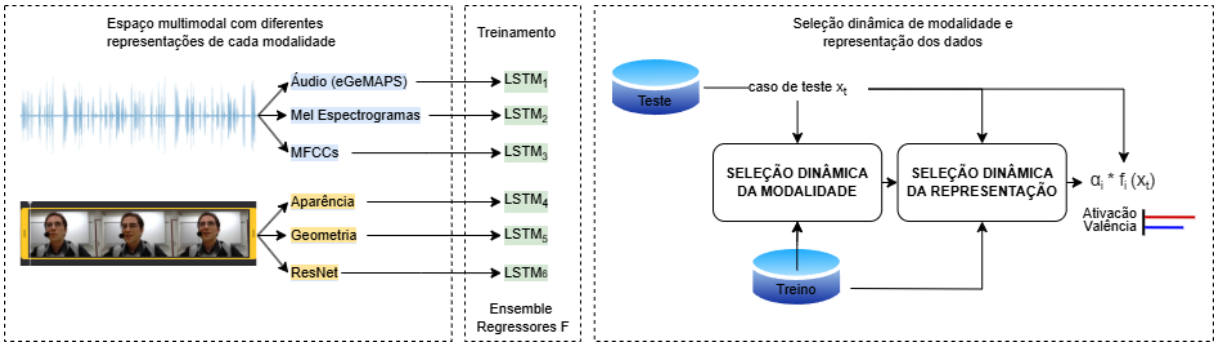
Entrada: Instância de teste $\mathbf{x}_j^{\text{ts}} \in \mathcal{T}_e$; modalidade mais promissora para cada caso de teste M_j; conjunto de modelos $\mathcal{F}$; conjunto de treino $\mathcal{T}_r$; tamanho da vizinhança K; método de seleção dinâmica a ser calculado (DS, DW ou DWS)
Saída: $f_{ens}(\mathbf{x}_j^{\text{ts}})$, a predição mais promissora para o cada caso de teste $\mathbf{x}_j^{\text{ts}} \in \mathcal{T}_e$
Calcule a região de competência Ψ_j usando o Algoritmo 2 com os parâmetros $\mathbf{x}_j^{\text{ts}}$, $\mathcal{T}_r$ e K
foreach $\mathbf{x}_k^{tr}$ *em* Ψ_j **do**
  Calcule o peso d_k usando a Eq. 4.9
end
foreach f_i *em* $\mathcal{F} \in M_j$ **do**
  Calcule o peso α_i para o modelo f_i usando a Eq. 4.10
  Calcule o erro acumulado $E_i = \sum_{k=1}^{K} d_k \cdot sqe_{k,i}$
end
if *método* == *DS* **then**
  Selecione o modelo f_i com o menor erro acumulado E_i em Ψ_j
  Defina $f_{\text{ens}}(\mathbf{x}_j^{\text{ts}}) \leftarrow f_i(\mathbf{x}_j^{\text{ts}})$
end
if *método* == *DW* **then**
  Defina $f_{\text{ens}}(\mathbf{x}_j^{\text{ts}}) \leftarrow \sum_{i=1}^{N} \alpha_i \cdot f_i(\mathbf{x}_j^{\text{ts}})$
end
if *método* == *DWS* **then**
  Selecione subconjunto S com $E_i \leq (E_{\max} - E_{\min})/2$
  Defina $f_{\text{ens}}(\mathbf{x}_j^{\text{ts}}) \leftarrow \sum_{i=1}^{S} \alpha_i \cdot f_i(\mathbf{x}_j^{\text{ts}})$
end
return $f_{\text{ens}}(\mathbf{x}_j^{\text{ts}})$

Quando a variação do ensemble ocorre por meio de alterações na representação dos dados, como em soluções multimodais, certas adaptações tornam-se necessárias. Cada modelo $f \in \mathcal{F}$ é treinado com uma representação específica dos dados no espaço de características $\mathrm{M}_i\mathrm{R}_j$, requerendo que o cálculo de Ψ seja realizado dentro deste espaço de características. Embora representações distintas apresentem diversidade, tanto em número de características quanto em escala e distribuições, as distâncias são ponderadas para que os valores sejam normalizados, garantindo uma comparação coerente entre as diferentes representações.

4.2 Validação usando MER

O método proposto de seleção dinâmica de modalidade e representação de dados foi validado no contexto de MER. Inspirado nos trabalhos de Zhang et al. (2019), AlBadawy e Kim (2018), Cheng et al. (2024) foram consideradas características de áudio e vídeo. Como mostrado na Fig. 20, as características de áudio incluem características acústicas, MFCCs e Mel espectrogramas. No lado do vídeo, incorporamos características de aparência, características geométricas e características extraídas via ResNet.

Figura 20 – Método de seleção dinâmica de modalidade e representação de dados aplicado ao contexto de reconhecimento multimodal de emoções contínuas. As características de áudio incluem características acústicas, MFCCs e Mel espectrogramas. As características de vídeo incluem características de aparência, características geométricas e características extraídas via ResNet. Todos os regressores são treinados separadamente, e um conjunto de regressores é obtido. A melhor modalidade é selecionada na fase de seleção dinâmica da modalidade, e depois, na seleção dinâmica intra-modal da representação, cada modelo recebe um peso para avaliar cada caso de teste de acordo com sua assertividade na zona de competência.



Fonte: Autoria Própria

4.2.1 Modalidades e Representações

Para o áudio, foram utilizadas as 88 características relacionadas a eGeMAPS disponibilizadas em Valstar et al. (2016) e (KOSSAIFI et al., 2021). Além disso, para assegurar a existência de múltiplas representações da modalidade de áudio, foram extraídos conjuntos de características baseados em Mel-Frequency Cepstral Coefficients (MFCCs) e Mel espectrogramas, extrações realizadas pela autora deste trabalho utilizando a biblioteca Librosa (MCFEE et al., 2024).

Os MFCCs representam a envoltória espectral do áudio e capturam características relacionadas ao timbre. Os 10 principais coeficientes MFCC (n_mfcc=10) foram extraídos utilizando a função librosa.feature.mfcc. Em seguida, foram calculadas suas derivadas de primeira e segunda ordem, visando capturar de forma mais eficaz as variações temporais do sinal. Todos os coeficientes foram concatenados para formar a matriz final com (30,5) características MFCC, que foi reduzida para um vetor de 30 elementos através da média de cada linha.

O Mel espectrograma mede a distribuição da energia do sinal de áudio ao longo do tempo em uma escala de frequências perceptual, a escala Mel. Para sua extração, o sinal áudio foi processado por uma Transformada de Fourier de Curto Tempo (STFT) utilizando librosa.stft, seguido pela conversão para uma escala logarítmica. O processo resultou em uma matriz final contendo (65,5) características, que foi simplificada para um vetor de 65 elementos ao calcular a média de cada linha.

Para o vídeo, foram utilizadas as características de aparência e geometria disponibilizadas em Valstar et al. (2016) e Kossaifi et al. (2021). Na RECOLA, a análise de aparência processa a textura facial para extrair 168 características, e a análise geométrica calcula 632 características com base em distâncias e ângulos entre os marcos faciais. Na SEWA, são consideradas 3000 características relacionadas à aparência facial e 121 características geométricas. Para obter o mesmo número de representações das modalidades áudio e vídeo foi utilizada uma rede pré-treinada como extratora de características das faces recortadas de cada frame do vídeo. Inspirado nos trabalhos de Chen et al. (2019), Zhao et al. (2018), Zhao et al. (2019) foram realizados testes com as redes ResNet (HE et al., 2015a), Densenet (HUANG et al., 2016) e VGG (SIMONYAN; ZISSERMAN, 2014) pré-treinadas na Imagenet (DENG et al., 2009). ResNet apresentou melhores resultados e foi a rede selecionada para extrair 2048 características faciais. Os maiores vetores de características, incluindo as características faciais extraídas pela ResNet, características de aparência SEWA e características de geometria RECOLA foram reduzidos em dimensionalidade para considerar apenas as 150 características mais informativas via PCA, valor definido empiricamente.

Tabela 6 – Dimensionalidade dos vetores de características de cada espaço de representação, abrangendo representações das modalidades de áudio (A) e vídeo (V), com características acústicas, MFCCs, Mel espectrogramas, características de aparência, geometria e características extraídas via ResNet. As dimensionalidades são apresentadas antes e depois da redução de dimensionalidade via PCA. NA indica que a redução de dimensionalidade via PCA não se aplica àquela representação.

	Dimensionalidade Original		Dimensionalidade Após PCA	
Espaço de Representação	RECOLA	SEWA	RECOLA	SEWA
Áudio (A)	88	88	NA	NA
MFCCs (A)	30	30	NA	NA
Mel Espectrogramas (A)	65	65	NA	NA
Aparência (V)	168	3000	NA	150
Geometria (V)	632	121	150	NA
ResNet (V)	2048	2048	150	150

4.2.2 Geração do Ensemble de Regressores

Inicialmente, dois conjuntos de modelos são gerados: $\mathcal{F}_{\mathrm{A}} = \{f_1^A, f_2^A, \ldots, f_N^A\}$ para a dimensão de ativação, e $\mathcal{F}_{\mathrm{V}} = \{f_1^V, f_2^V, \ldots, f_N^V\}$ para a dimensão de valência. Cada ensemble de modelos $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$ é construído a partir das diferentes combinações de modalidades e representações, definidas na Seção 4.2.1. As representações são

normalizadas para uma distribuição com média zero e desvio padrão unitário.

Para representar os padrões temporais presentes nas séries de dados analisadas, empregamos redes neurais recorrentes do tipo LSTM. Cada regressor f_i recebe como entrada uma sequência temporal de características $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, onde $\mathbf{x}_t \in \mathbb{R}^d$ representa o vetor de características no instante t, e retorna uma sequência de predições $\mathbf{y} = \{y_1, y_2, \ldots, y_T\}$, onde cada $y_t \in \mathbb{R}$ representa a saída correspondente ao instante t. Formalmente, temos:

$$f : \mathbb{R}^{T \times d} \to \mathbb{R}^T, \quad f(\mathbf{X}) = \mathbf{y} \tag{4.11}$$

onde T representa o número de instantes temporais considerados na entrada da LSTM, d é a dimensão do vetor de características em cada instante t e $\mathbf{y} = \{y_1, y_2, \ldots, y_T\}$ é a sequência de predições ao longo do tempo.

Os modelos foram treinados usando uma abordagem de busca em grade para otimizar hiperparâmetros do modelo LSTM. Os dados foram segmentados em sequências de 6 segundos, correspondendo a 150 passos de tempo (frames) na base RECOLA e 60 passos na SEWA a uma taxa de amostragem de 16kHz. A busca dos hiperparâmetros foi realizada utilizando a biblioteca Keras Tuner (O'MALLEY et al., 2019), que permitiu testar diferentes configurações do modelo, conforme os seguintes parâmetros e estratégias:

- Número de Unidades nas Camadas LSTM: O número de unidades em cada camada foi ajustado no intervalo de 32 a 256, com incrementos de 32.
- Taxa de Dropout: Variada entre 0 e 0,5, em passos de 0,1, para introduzir regularização e evitar overfitting.
- Número de Camadas LSTM: A estrutura permitiu entre 1 a 4 camadas de LSTM, ajustando a profundidade da rede conforme o valor ótimo para cada tentativa.

A camada de saída foi definida com uma ativação linear, o que é adequado para problemas de regressão. O modelo foi compilado usando o otimizador 'RMSprop' e uma função de perda personalizada baseada no valor de CCC para maximizar a concordância entre as predições e os valores reais. Uma estratégia de interrupção antecipada foi utilizada, monitorando a perda no conjunto de validação e interrompendo o treinamento após 20 épocas sem melhora, reduzindo o risco de overfitting. O treinamento foi realizado por até 300 épocas com lotes de tamanho 32.

A escolha da LSTM se deve à sua capacidade de capturar dependências de longo prazo e à sua consolidação como método de referência na modelagem de sequências temporais. No contexto de reconhecimento de emoções, apresenta resultados promissores em tarefas de regressão de ativação e valência, conforme discutido nas Seções 2.5 e 3.2.2.

4.2.3 Protocolo Experimental

As bases de dados RECOLA (RINGEVAL et al., 2013) e SEWA (KOSSAIFI et al., 2021) foram utilizadas para validar o método proposto. Todos os experimentos foram realizados de forma independente para as dimensões de ativação e valência, com ensembles de modelos distintos e avaliações individualizadas para cada dimensão emocional.

Um protocolo experimental baseado em holdout com múltiplas repetições foi implementado para aumentar a confiança na avaliação dos resultados. A base de dados RECOLA compreende dados rotulados de cinco minutos de interações audiovisuais de dezoito indivíduos; para equilibrar os conjuntos de treinamento, teste e validação, três indivíduos foram alocados para o conjunto de teste, três para validação e os participantes restantes foram destinados ao treinamento. A base de dados SEWA, por outro lado, contém dados rotulados de quarenta e sete indivíduos com interações que variam entre 46 segundos e 175 segundos; para equilibrar os conjuntos de treinamento, teste e validação, os vídeos foram recortados em 170 segundos (dois minutos e cinquenta segundos) e os vídeos menores que este valor foram descartados. Sobraram 40 indivíduos, oito indivíduos foram alocados para o conjunto de teste, oito para validação e vinte e quatro para o conjuntos de treinamento. Nas duas bases, o experimento foi repetido dez vezes, e em cada iteração os participantes foram embaralhados e redistribuídos de maneira aleatória entre os subconjuntos de treinamento, teste e validação. A Tabela 7 resume as principais características dos dados utilizados, incluindo a duração total das amostras e o número de indivíduos nos subconjuntos de treino, validação e teste.

Tabela 7 – Resumo dos dados utilizados no protocolo experimental.

Base de dados	Duração (s)	Frames (6s)	#Treino	#Validação	#Teste
RECOLA	300	150	12	3	3
SEWA	170	60	24	8	8

A otimização dos modelos foi realizada apenas na primeira repetição, onde foi identificada a melhor configuração de hiperparâmetros. Esta configuração foi então aplicada nas demais repetições para garantir uma comparação consistente entre as execuções. As Tabelas 8 a 19 apresentam os modelos otimizados nas modalidades de áudio e vídeo, abrangendo os modelos baseados em eGeMAPS, MFCCs, Mel espectrogramas, características de aparência, características de geometria e características extraídas via ResNet, para as bases de dados RECOLA e SEWA.

Tabela 8 – Arquitetura do modelo baseado em eGeMAPS para a predição de ativação e valência na base de dados RECOLA.

Ativação

Camada (tipo)	Formato de Saída
lstm (LSTM)	150, 160
lstm1 (LSTM)	150, 192
dense (Dense)	150, 1

Valência

Camada (tipo)	Formato de Saída
lstm (LSTM)	150, 64
lstm1 (LSTM)	150, 64
dense (Dense)	150, 1

Tabela 9 – Arquitetura do modelo baseado em MFCCs para a predição de ativação e valência na base de dados RECOLA.

Ativação

Camada (tipo)	Formato de Saída
lstm (LSTM)	150, 224
dense (Dense)	150, 1

Valência

Camada (tipo)	Formato de Saída
lstm (LSTM)	150, 160
lstm (LSTM)	150, 192
dense (Dense)	150, 1

Tabela 10 – Arquitetura do modelo baseado em Mel espectrogramas para a predição de ativação e valência na base de dados RECOLA.

Ativação

Camada (tipo)	Formato de Saída
lstm (LSTM)	150, 96
lstm1 (LSTM)	150, 128
lstm2 (LSTM)	150, 128
lstm3 (LSTM)	150, 128
dense (Dense)	150, 1

Valência

Camada (tipo)	Formato de Saída
lstm (LSTM)	150, 160
lstm1 (LSTM)	150, 256
lstm2 (LSTM)	150, 256
lstm3 (LSTM)	150, 256
dense (Dense)	150, 1

Tabela 11 – Arquitetura do modelo baseado em características de aparência para a predição de ativação e valência na base de dados RECOLA.

Ativação

Camada (tipo)	Formato de Saída
lstm (LSTM)	150, 96
lstm1 (LSTM)	150, 160
dense (Dense)	150, 1

Valência

Camada (tipo)	Formato de Saída
lstm (LSTM)	150, 96
lstm1 (LSTM)	150, 160
dense (Dense)	150, 1

Tabela 12 – Arquitetura do modelo baseado em características de geometria para a predição de ativação e valência na base de dados RECOLA.

Ativação

Camada (tipo)	Formato de Saída
lstm (LSTM)	150, 32
lstm1 (LSTM)	150, 224
lstm2 (LSTM)	150, 224
lstm3 (LSTM)	150, 224
dense (Dense)	150, 1

Valência

Camada (tipo)	Formato de Saída
lstm (LSTM)	150, 192
lstm1 (LSTM)	150, 32
dense (Dense)	150, 1

Tabela 13 – Arquitetura do modelo baseado em características extraídas via ResNet para a predição de ativação e valência na base de dados RECOLA.

Ativação

Camada (tipo)	Formato de Saída
lstm (LSTM)	150, 160
lstm1 (LSTM)	150, 192
dense (Dense)	150, 1

Valência

Camada (tipo)	Formato de Saída
lstm (LSTM)	150, 64
lstm1 (LSTM)	150, 64
dense (Dense)	150, 1

Tabela 14 – Arquitetura do modelo baseado em eGeMAPS para a predição de ativação e valência na base de dados SEWA.

Ativação

Camada (tipo)	Formato de Saída
lstm (LSTM)	60, 128
dense (Dense)	60, 1

Valência

Camada (tipo)	Formato de Saída
lstm (LSTM)	60, 32
dense (Dense)	60, 1

Tabela 15 – Arquitetura do modelo baseado em MFCCs para a predição de ativação e valência na base de dados SEWA.

Ativação

Camada (tipo)	Formato de Saída
lstm (LSTM)	60, 128
dense (Dense)	60, 1

Valência

Camada (tipo)	Formato de Saída
lstm (LSTM)	60, 32
dense (Dense)	60, 1

Tabela 16 – Arquitetura do modelo baseado em Mel espectrogramas para a predição de ativação e valência na base de dados SEWA.

Ativação

Camada (tipo)	Formato de Saída
lstm (LSTM)	60, 160
lstm1 (LSTM)	60, 256
lstm2 (LSTM)	60, 256
lstm3 (LSTM)	60, 256
dense (Dense)	60, 1

Valência

Camada (tipo)	Formato de Saída
lstm (LSTM)	60, 160
lstm1 (LSTM)	60, 192
dense (Dense)	60, 1

Tabela 17 – Arquitetura do modelo baseado em características de aparência para a predição de ativação e valência na base de dados SEWA.

Ativação

Camada (tipo)	Formato de Saída
lstm (LSTM)	60, 160
lstm1 (LSTM)	60, 1192
dense (Dense)	60, 1

Valência

Camada (tipo)	Formato de Saída
lstm (LSTM)	60, 160
lstm1 (LSTM)	60, 256
lstm2 (LSTM)	60, 256
lstm3 (LSTM)	60, 256
dense (Dense)	60, 1

Tabela 18 – Arquitetura do modelo baseado em características de geometria para a predição de ativação e valência na base de dados SEWA.

Ativação

Camada (tipo)	Formato de Saída
lstm (LSTM)	60, 160
lstm1 (LSTM)	60, 256
lstm2 (LSTM)	60, 256
lstm3 (LSTM)	60, 256
dense (Dense)	60, 1

Valência

Camada (tipo)	Formato de Saída
lstm (LSTM)	60, 160
lstm1 (LSTM)	60, 192
dense (Dense)	60, 1

Tabela 19 – Arquitetura do modelo baseado em características extraídas via ResNet para a predição de ativação e valência na base de dados SEWA.

Ativação

Camada (tipo)	Formato de Saída
lstm (LSTM)	60, 160
lstm1 (LSTM)	60, 256
lstm2 (LSTM)	60, 256
lstm3 (LSTM)	60, 256
dense (Dense)	60, 1

Valência

Camada (tipo)	Formato de Saída
lstm (LSTM)	60, 160
lstm1 (LSTM)	60, 256
lstm2 (LSTM)	60, 256
lstm3 (LSTM)	60, 256
dense (Dense)	60, 1

4.3 Ausência de Modalidade

Para avaliar a robustez e a adaptabilidade do método proposto em condições reais e cenários não ideais, a ausência de modalidade é tratada simulando cenários em que uma ou mais fontes de dados se tornam indisponíveis durante a inferência. Essa abordagem foi inspirada em métodos de análise de sensibilidade utilizados em (NAIK; KIRAN, 2021; NUNES et al., 2004). Foram testadas proporções de ausência de dados de 25%, 50% e 100%. Em termos práticos, isso significa que, para cada instância do conjunto de teste, uma determinada proporção dos segmentos (janela temporal) teve os vetores de características de uma modalidade específica substituídos por zeros, simulando a ausência completa da modalidade neste intervalo.

A Figura 21 apresenta a predição de ativação de todos os modelos para um mesmo caso de teste, considerando cenários em que as modalidades de áudio e vídeo estão integralmente disponíveis quanto situações em que cada modalidade está 50% indisponível. Em condições ideais, com áudio e vídeo disponíveis, observa-se que todos os modelos exibem um padrão consistente, com predições semelhantes. No entanto, em determinados intervalos, um modelo pode se aproximar mais do padrão de referência, enquanto em outros, outro modelo apresenta desempenho superior. Quando uma das modalidades está parcialmente indisponível, há um declínio perceptível no desempenho dos modelos que dependem das representações correspondentes a esta modalidade.

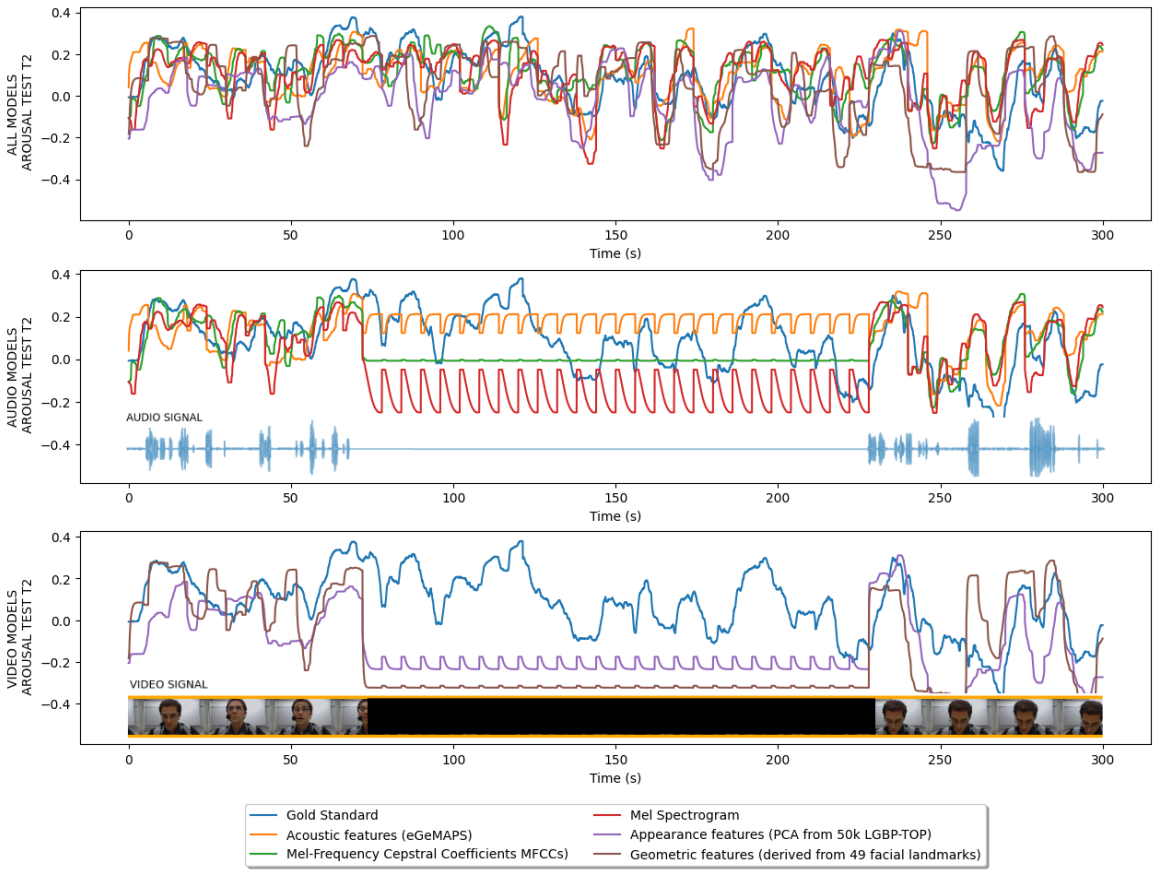
Para investigar a robustez do sistema diante dessa limitação, em vez de simplesmente remover ou zerar os sinais no início ou no fim da sequência, optou-se por suprimi-los no trecho central, onde se encontra o ápice da emoção. Dessa forma, avalia-se o desempenho do modelo justamente nos instantes mais críticos de ausência de dados, verificando se o método proposto consegue adaptar-se efetivamente às restrições impostas pelo ambiente de teste.

4.4 Considerações Finais

Neste capítulo, foi apresentado o método proposto para seleção dinâmica de modalidade e representação dos dados. Inicialmente, um conjunto de modelos é treinado com diferentes combinações de modalidades e suas respectivas representações. A seleção dinâmica é implementada em duas fases complementares: primeiramente, identifica-se a modalidade mais apropriada para cada caso de teste; em seguida, seleciona-se a representação mais eficaz dentro da modalidade escolhida, conforme as características do dado em análise. Esta abordagem integrada não só melhora a precisão do sistema, mas também permite a sua adaptação às particularidades do problema e ao ambiente em que será aplicado.

No próximo capítulo, apresentaremos os resultados experimentais alcançados com a aplicação deste método no contexto de MER, incluindo uma análise comparativa com métodos já existentes, mecanismo de atenção cruzada proposto por Praveen et al. (PRAVEEN et al., 2022) e fusão estática das modalidades (média e melhor regressor individual).

Figura 21 – Padrão-ouro de ativação e predição de modelos baseados em características acústicas, MFCCs, Mel espectrogramas, características de aparência e características geométricas. A imagem foi gerada usando o caso de teste T2 (segunda pessoa do conjunto de testes) do segundo conjunto de validação cruzada (k=2), base de dados RECOLA. De cima para baixo, temos (i) predição com todas as modalidades ativas, (ii) predições dos modelos de áudio com 50% de ausência da modalidade de áudio e o respectivo sinal de áudio, e (iii) predições dos modelos de vídeo com 50% de ausência da modalidade de vídeo e a sequência correspondente de quadros de vídeo.



Fonte: Autoria Própria

5 Resultados

Este capítulo apresenta uma análise detalhada dos resultados obtidos com a aplicação do método proposto de seleção dinâmica de modalidade e representação de dados para o contexto de reconhecimento multimodal de emoções contínuas, rotuladas em termos de ativação e valência. Para fins de comparação, também foi avaliado o desempenho do método proposto em relação a abordagens de referência, incluindo mecanismos de atenção e fusão estática de modalidades — especificamente, a média das saídas dos regressores e o melhor regressor individual. Foram considerados tanto cenários ideais, com todas as modalidades disponíveis, quanto situações de indisponibilidade total ou parcial das modalidades de áudio e vídeo.

5.1 Cenário com Informações Completas

Esta seção apresenta os resultados obtidos nas condições ideais, com todas as modalidades disponíveis.

5.1.1 RECOLA

A Tabela 20 exibe os resultados de ativação e valência na base RECOLA, em termos de CCC, para o conjunto de regressores $\mathcal{F}$, abrangendo modelos baseados nas representações de características acústicas, MFCCs, Mel espectrogramas, características de aparência, geometria e características extraídas via ResNet. Os resultados revelam que a modalidade de áudio (A) representa melhor a dimensão de ativação, com características acústicas, MFCCs e Mel Espectrogramas alcançando valores de CCC de $0,69$, $0,64$ e $0,68$, respectivamente. Por outro lado, valência é representada com mais precisão pela modalidade de vídeo (V), com suas representações relacionadas à aparência, geometria e características extraídas via ResNet alcançando valores de CCC de $0,45$, $0,49$ e $0,39$, respectivamente. Os valores reportados correspondem à média dos resultados, considerando as 10 repetições do experimento.

Em condições ideais, com todas as modalidades disponíveis — áudio e vídeo —, os resultados apresentados na Tabela 21 mostram que o maior desempenho de ativação (CCC = 0,73) foi alcançado utilizando as técnicas DW, DWS e a média simples das saídas de todos os regressores. As técnicas DW e DWS apresentaram os melhores resultados na dimensão de valência, com um CCC de 0,54, superando tanto a média das saídas dos regressores quanto o melhor regressor individual, que obtiveram um CCC de 0,49. Os resultados da atenção cruzada apresentaram um CCC de 0,46 para ativação e 0,41 para valência.

Tabela 20 – CCC para ativação e valência na base de dados RECOLA, abrangendo modelos baseados nas modalidades de áudio (A) e vídeo (V), com características acústicas, MFCCs, Mel espectrogramas, características de aparência, geometria e características extraídas via ResNet.

Espaço de Representação	**Ativação**	**Valência**
Áudio (A)	$\mathbf{0{,}69 \pm 0{,}06}$	$0,20 \pm 0,07$
MFCCs (A)	$0,64 \pm 0,06$	$0,35 \pm 0,08$
Mel Espectrogramas (A)	$0,68 \pm 0,06$	$0,23 \pm 0,09$
Aparência (V)	$0,42 \pm 0,09$	$0,45 \pm 0,06$
Geometria (V)	$0,42 \pm 0,09$	$\mathbf{0{,}49 \pm 0{,}14}$
ResNet (V)	$0,15 \pm 0,13$	$0,39 \pm 0,14$

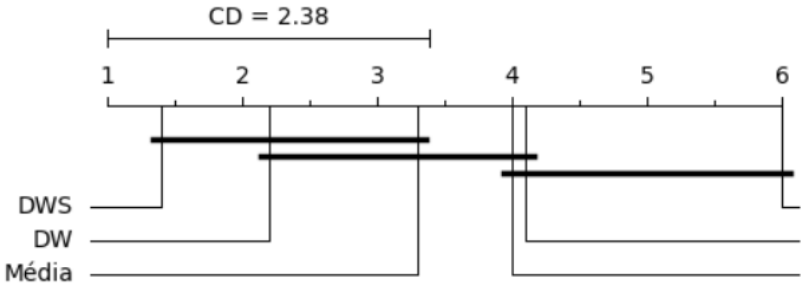
Tabela 21 – CCC para ativação e valência na base de dados RECOLA, abrangendo a média das saídas dos regressores, melhor regressor indivial, seleção dinâmica (DS, DW, DWS) e atenção cruzada. Os resultados de DS, DW e DWS foram gerados com K = 100.

Método	**Ativação**	**Valência**
DS	$0,68 \pm 0,06$	$0,46 \pm 0,08$
DW	$\mathbf{0{,}73 \pm 0{,}04}$	$\mathbf{0{,}54 \pm 0{,}10}$
DWS	$\mathbf{0{,}73 \pm 0{,}04}$	$\mathbf{0{,}54 \pm 0{,}10}$
Média	$\mathbf{0{,}73 \pm 0{,}04}$	$0,49 \pm 0,08$
Melhor Individual	$0,69 \pm 0,06$	$0,49 \pm 0,14$
Atenção Cruzada	$0,46 \pm 0,13$	$0,41 \pm 0,17$

Os métodos DW e DWS não apenas apresentam os melhores resultados médios para ativação, mas também mantêm uma baixa dispersão, o que indica maior consistência de resultados. O método de atenção cruzada apresenta um valor médio menor e maior dispersão, sugerindo uma instabilidade em seu desempenho para ativação. Ao analisar a valência, é possível verificar uma situação semelhante, com DW e DWS alcançando os melhores resultados e a atenção cruzada apresentando um desempenho levemente inferior. A dispersão é mais pronunciada para valência em comparação com ativação.

Para verificar se há diferença significativa entre os resultados obtidos, foi utilizado o protocolo descrito por Demšar (2006), baseado no teste estatístico não paramétrico de Friedman (FRIEDMAN, 1937), seguido do teste post-hoc de Nemenyi (NEMENYI, 1963). Aplicando o teste de Friedman com nível de significância $\alpha = 0,05$ concluiu-se que há diferença significativa entre os métodos analisados na dimensão de ativação ($p < 0{,}05$), enquanto na dimensão de valência as diferenças não são significativas ($p = 0{,}093$). Considerando que uma diferença significativa foi identificada na dimensão de ativação em pelo menos um dos métodos analisados, aplicou-se o teste post-hoc de Nemenyi (NEMENYI, 1963) para determinar quais métodos apresentam diferenças significativas entre si. Os resultados desta comparação estão apresentados na Figura 22.

Figura 22 – Diagrama de Nemenyi para comparações de classificações médias (mean rank) dos métodos em termos de desempenho CCC na dimensão de ativação, base de dados RECOLA. A linha horizontal conecta os métodos cuja diferença na classificação média não é estatisticamente significativa ao nível de confiança de 0,05, considerando uma distância crítica (CD) de 2,38. A tabela à direita apresenta as classificações médias para cada método.



	Mean Rank
DS	4
DW	2,2
DWS	1,4
Melhor Individual	4,1
Média	3,3
Atenção Cruzada	6

Fonte: Autoria Própria

O teste de Nemenyi é uma análise de comparações múltiplas que avalia todas as variáveis observadas de forma pareada e identifica se uma variável é significativamente diferente da outra ao comparar a diferença entre suas classificações médias (*mean rank*). No teste de Nemenyi, a significância é determinada pela diferença crítica (do inglês, Critical Difference - CD). A CD é um valor calculado com base no número de variáveis e no número de observações, e representa a menor diferença necessária entre as classificações médias para que duas variáveis sejam consideradas significativamente diferentes. Quando a diferença entre as classificações médias de duas variáveis é maior que a CD, isso indica que elas são estatisticamente diferentes com o nível de confiança estabelecido (DEMsAR, 2006).

O teste de Nemenyi confirma que, em condições ideais com áudio e vídeo disponíveis, os métodos DW, DWS e média simples apresentam os melhores resultados para ativação. Além disso, não foi identificada diferença significativa entre os desempenhos desses métodos.

5.1.2 SEWA

A Tabela 22 exibe os resultados de ativação e valência na base SEWA, em termos de CCC, para o conjunto de regressores $\mathcal{F}$. Na base SEWA, a dimensão de ativação também foi melhor representada pela modalidade de áudio (A), com características acústicas, MFCCs e Mel Espectrogramas alcançando valores de CCC de $0,30$, $0,35$ e $0,36$, respectivamente. Valência foi representada com mais precisão pela modalidade de vídeo (V), com suas representações de aparência, geometria e características extraídas via ResNet alcançando

valores de CCC de $0,35$, $0,43$ e $0,44$, respectivamente. Os valores reportados correspondem à média dos resultados obtidos, considerando as 10 repetições do experimento.

Em condições ideais, com todas as modalidades disponíveis — áudio e vídeo —, o maior desempenho de ativação (CCC = 0,46) foi alcançado utilizando as técnicas DW, DWS e a média simples das saídas de todos os regressores. A técnica DWS apresentou o melhor resultado na dimensão de valência, com um CCC de 0,55. Os resultados estão disponíveis na Tabela 23.

Para verificar se há diferença significativa entre os resultados obtidos, foi aplicado o teste estatístico de Friedman (FRIEDMAN, 1937) com nível de significância $\alpha = 0,05$. Concluiu-se que há diferença significativa entre os métodos analisados nas dimensões de ativação e valência ($p < 0,05$). O teste post-hoc de Nemenyi (NEMENYI, 1963) foi aplicado para determinar quais métodos apresentam diferenças significativas entre si. Os resultados desta comparação estão apresentados na Figura 23.

O teste de Nemenyi confirma que, em condições ideais com áudio e vídeo disponíveis, os métodos DW, DWS e média simples apresentam os melhores resultados nas duas dimensões, ativação e valência. No entanto, não foi identificada diferença significativa entre os desempenhos desses métodos.

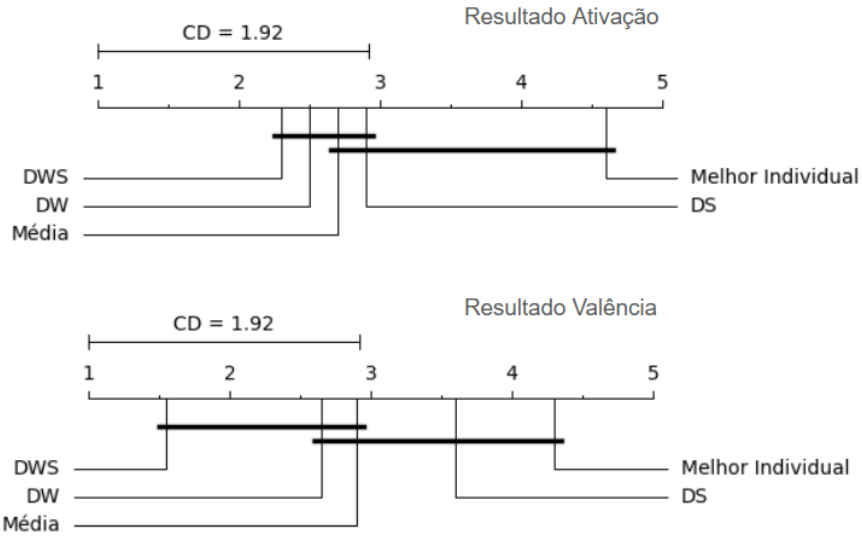
Tabela 22 – CCC para ativação e valência na base de dados SEWA, abrangendo modelos baseados nas modalidades de áudio (A) e vídeo (V), com características acústicas, MFCCs, Mel espectrogramas, características de aparência, geometria e características extraídas via ResNet.

Espaço de Representação	**Ativação**	**Valência**
Áudio (A)	$0,30 \pm 0,08$	$0,27 \pm 0,07$
MFCCs (A)	$0,35 \pm 0,08$	$0,38 \pm 0,06$
Mel Espectrogramas (A)	$\mathbf{0,36 \pm 0,08}$	$0,39 \pm 0,14$
Aparência (V)	$0,25 \pm 0,11$	$0,35 \pm 0,10$
Geometria (V)	$0,23 \pm 0,08$	$0,43 \pm 0,08$
ResNet (V)	$0,36 \pm 0,10$	$\mathbf{0,44 \pm 0,09}$

Tabela 23 – CCC para ativação e valência na base de dados SEWA, abrangendo a média das saídas dos regressores, melhor regressor indivial e seleção dinâmica (DS, DW, DWS). Os resultados de DS, DW e DWS foram gerados com K = 100.

Método	**Ativação**	**Valência**
DS	$0,46 \pm 0,07$	$0,49 \pm 0,08$
DW	$\mathbf{0,46 \pm 0,04}$	$0,53 \pm 0,07$
DWS	$\mathbf{0,46 \pm 0,04}$	$\mathbf{0,55 \pm 0,08}$
Média	$0,46 \pm 0,07$	$0,52 \pm 0,08$
Melhor Individual	$0,36 \pm 0,08$	$0,44 \pm 0,09$

Figura 23 – Diagrama de Nemenyi para comparações de classificações médias (mean rank) dos métodos em termos de desempenho CCC nas dimensões de ativação e valência, base de dados SEWA. A linha horizontal conecta os métodos cuja diferença na classificação média não é estatisticamente significativa ao nível de confiança de 0,05, considerando uma distância crítica (CD) de 1,92. A tabela à direita apresenta as classificações médias para cada método.



	Mean Rank
DS	2,9
DW	2,5
DWS	2,3
Média	2,7
Melhor Individual	4,6

	Mean Rank
DS	3,6
DW	2,65
DWS	1,55
Média	2,9
Melhor Individual	4,3

Fonte: Autoria Própria

5.2 Impacto da Ausência de Modalidades

Esta seção apresenta os resultados obtidos no contexto de ausência total ou parcial das modalidades de áudio e vídeo, com as proporções de dados ausentes de 25%, 50% e 100%.

5.2.1 RECOLA

As Tabelas 24 e 25 apresentam os resultados de ativação e valência, em termos de CCC, para todos os métodos comparados, incluindo a média das saídas dos regressores, o melhor regressor individual, as técnicas de seleção dinâmica (DS, DW, DWS) e a abordagem de atenção cruzada. A Figura 24 compara o padrão-ouro de ativação, a predição com todas as modalidades ativas e com a ausência total de cada modalidade.

No contexto da ativação, os métodos DW e DWS apresentam o melhor desempenho médio (CCC = 0,41) em cenários hipotéticos com 100% de ausência da modalidade de áudio. DS também se destaca, apresentando uma média ligeiramente inferior a DW e DWS, mas ainda com resultados superiores aos demais métodos (CCC = 0,35). DW e DWS mostram uma dispersão menor e melhores resultados em comparação aos outros métodos, sugerindo consistência no desempenho. É interessante observar que o melhor regressor individual para ativação é o baseado em características de áudio, portanto, numa

situação onde a modalidade de áudio está indisponível, seu resultado é bastante impactado, chegando a um CCC próximo de zero.

Para valência, ainda na condição de ausência total da modalidade de áudio, o desempenho de DW e DWS é superior aos outros métodos, ambos com CCC médio de 0,59. DW e DWS mantêm um desempenho elevado, com uma distribuição de resultados mais concentrada e consistente, indicando uma robustez maior para valência. Em contraste, a média, a atenção cruzada e o melhor regressor individual apresentam os piores desempenhos, com CCCs médios de 0,32, 0,40 e 0,49 respectivamente. É importante destacar que o método de atenção cruzada também possui uma dispersão elevada, sugerindo uma menor eficácia para a regressão de valência.

Nos cenários de 100% de ausência da modalidade visual, os métodos DW e DWS apresentam o melhor desempenho médio (CCC = 0,71) para ativação, com o método DS também se destacando com um CCC médio de 0,67. O melhor regressor individual também apresentou um resultado alto com CCC de 0,69. O melhor regressor para ativação é baseado em características de áudio e não é impactado pela ausência do vídeo. Os métodos DS, DW, DWS e o melhor regressor individual apresentam uma dispersão relativamente menor e resultados consistentes quando comparados a outros métodos, evidenciando uma performance robusta. Em contraste, a atenção cruzada apresenta um valor médio menor (CCC = 0,49) e maior variação, indicando menor efetividade na ausência da modalidade visual.

Para valência, ainda na condição de ausência total da modalidade visual, DW, DWS e DS apresentam valores próximos (CCC = 0,30, 0,30 e 0,28, respectivamente), mas apresentam dispersão considerável quando comparados com a média (CCC = 0,27). O melhor regressor individual para valência é baseado em características visuais geométricas e é fortemente impactado quando a modalidade de vídeo está indisponível, resultando em um CCC próximo a zero.

O teste de Friedman (FRIEDMAN, 1937) com um nível de significância $\alpha = 0,05$ revelou que, nas dimensões de ativação e valência, há uma diferença significativa na aplicação de pelo menos um dos métodos analisados em todos os cenários avaliados, que contemplam condições com ausência total das modalidades de áudio e de vídeo. O teste post-hoc de Nemenyi (NEMENYI, 1963) foi utilizado para ranquear os métodos e esclarecer quais métodos apresentam diferenças significativas entre si, apresentando os resultados comparativos nas Figuras 25 e 26.

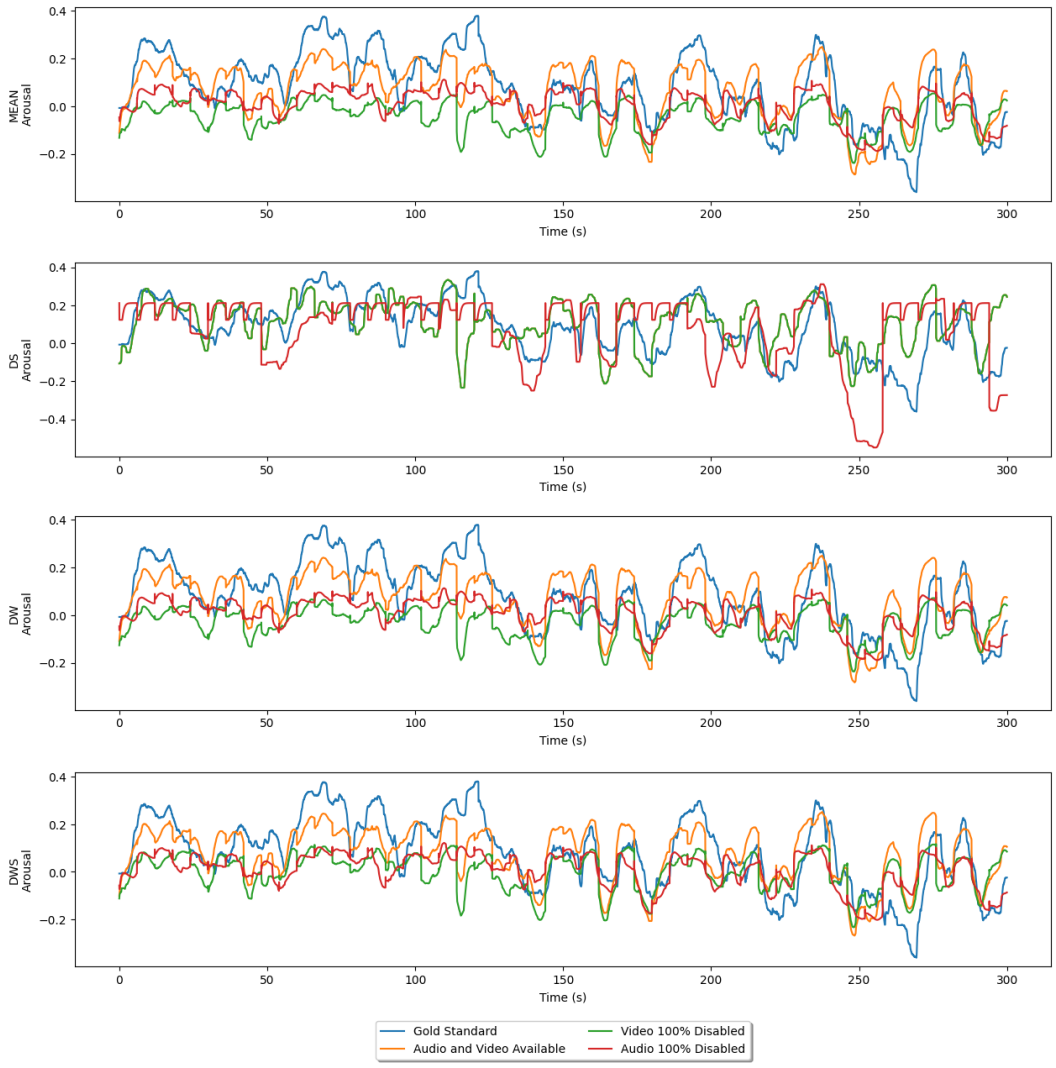
Tabela 24 – Resultados de ativação, em termos de CCC, base de dados RECOLA, abrangendo a média das saídas dos regressores, melhor regressor individual, seleção dinâmica (DS, DW, DWS) e atenção cruzada. Os resultados são apresentados nos seguintes cenários: áudio (A) e vídeo (V) disponíveis, áudio indisponível e vídeo indisponível. As proporções de dados ausentes analisadas foram de 25%, 50% e 100%.

Modalidades	**DS**	**DW**	**DWS**	**Média**	**Melhor individual**	**Atenção Cruzada**
A e V disponíveis	$0,68 \pm 0,06$	$\mathbf{0,73 \pm 0,04}$	$\mathbf{0,73 \pm 0,04}$	$\mathbf{0,73 \pm 0,04}$	$0,69 \pm 0,06$	$0,46 \pm 0,13$
25% V indisponível	$0,62 \pm 0,03$	$0,65 \pm 0,05$	$0,65 \pm 0,05$	$0,68 \pm 0,05$	$\mathbf{0,69 \pm 0,05}$	$0,42 \pm 0,11$
25% A indisponível	$0,51 \pm 0,07$	$0,54 \pm 0,07$	$0,54 \pm 0,08$	$\mathbf{0,60 \pm 0,04}$	$0,54 \pm 0,07$	$0,36 \pm 0,15$
50% V indisponível	$0,65 \pm 0,03$	$0,68 \pm 0,04$	$0,68 \pm 0,04$	$0,66 \pm 0,06$	$\mathbf{0,69 \pm 0,05}$	$0,41 \pm 0,09$
50% A indisponível	$0,48 \pm 0,07$	$\mathbf{0,52 \pm 0,07}$	$0,51 \pm 0,07$	$0,50 \pm 0,04$	$0,37 \pm 0,08$	$0,21 \pm 0,14$
100% V indisponível	$0,67 \pm 0,06$	$\mathbf{0,71 \pm 0,05}$	$\mathbf{0,71 \pm 0,05}$	$0,61 \pm 0,09$	$0,69 \pm 0,06$	$0,49 \pm 0,12$
100% A indisponível	$0,35 \pm 0,15$	$\mathbf{0,41 \pm 0,17}$	$\mathbf{0,41 \pm 0,17}$	$0,31 \pm 0,05$	$0,001 \pm 0,0003$	$0,23 \pm 0,12$

Tabela 25 – Resultados de valência, em termos de CCC, base de dados RECOLA, abrangendo a média das saídas dos regressores, melhor regressor individual, seleção dinâmica (DS, DW, DWS) e atenção cruzada. Os resultados são apresentados nos seguintes cenários: áudio (A) e vídeo (V) disponíveis, áudio indisponível e vídeo indisponível. As proporções de dados ausentes analisadas foram de 25%, 50% e 100%.

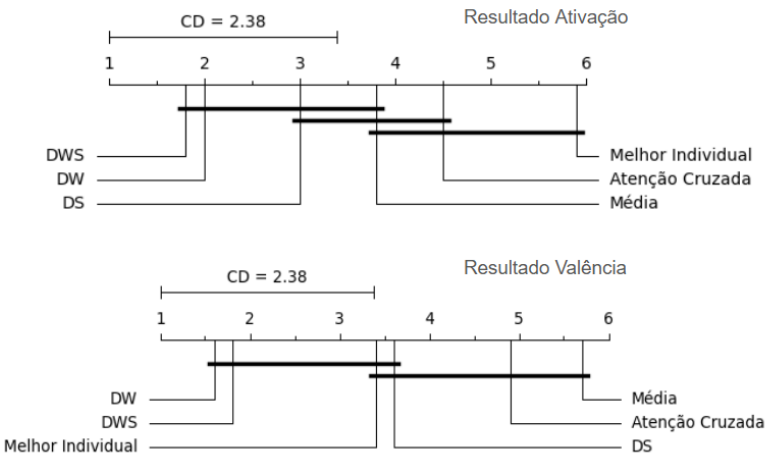
Modalidades	**DS**	**DW**	**DWS**	**Média**	**Melhor individual**	**Atenção Cruzada**
A e V disponíveis	$0,46 \pm 0,08$	$\mathbf{0,54 \pm 0,10}$	$\mathbf{0,54 \pm 0,10}$	$0,46 \pm 0,08$	$0,49 \pm 0,14$	$0,41 \pm 0,17$
25% V indisponível	$0,41 \pm 0,07$	$\mathbf{0,46 \pm 0,06}$	$\mathbf{0,46 \pm 0,06}$	$0,44 \pm 0,07$	$0,36 \pm 0,10$	$0,34 \pm 0,11$
25% A indisponível	$0,46 \pm 0,07$	$\mathbf{0,52 \pm 0,07}$	$\mathbf{0,52 \pm 0,07}$	$0,44 \pm 0,08$	$0,49 \pm 0,13$	$0,37 \pm 0,14$
50% V indisponível	$0,40 \pm 0,08$	$\mathbf{0,42 \pm 0,07}$	$\mathbf{0,42 \pm 0,07}$	$0,40 \pm 0,07$	$0,26 \pm 0,06$	$0,25 \pm 0,07$
50% A indisponível	$0,45 \pm 0,06$	$\mathbf{0,52 \pm 0,07}$	$0,51 \pm 0,07$	$0,39 \pm 0,07$	$0,49 \pm 0,13$	$0,38 \pm 0,16$
100% V indisponível	$0,28 \pm 0,18$	$\mathbf{0,30 \pm 0,17}$	$\mathbf{0,30 \pm 0,17}$	$0,27 \pm 0,08$	$0,0002 \pm 0,0006$	$0,13 \pm 0,15$
100% A indisponível	$0,49 \pm 0,09$	$\mathbf{0,59 \pm 0,09}$	$\mathbf{0,59 \pm 0,09}$	$0,32 \pm 0,06$	$0,49 \pm 0,14$	$0,40 \pm 0,14$

Figura 24 – Comparação do padrão-ouro de ativação, predição com todas as modalidades ativas, predição com a ausência da modalidade de áudio e predição com a ausência da modalidade de vídeo da média das saídas dos regressores e métodos baseados em seleção dinâmica (DS, DW, DWS). A imagem foi gerada usando o caso de teste T2 (segunda pessoa do conjunto de testes) do segundo conjunto de validação cruzada (k=2), base de dados RECOLA.



Fonte: Autoria Própria

Figura 25 – Diagrama de Nemenyi para comparações de classificações médias (mean rank) dos métodos em termos de desempenho CCC nas dimensões de ativação e valência, base de dados RECOLA, caso com 100% de ausência da modalidade de áudio. A linha horizontal conecta os métodos cuja diferença na classificação média não é estatisticamente significativa ao nível de confiança de 0,05, considerando uma distância crítica (CD) de 2,38. A tabela à direita apresenta as classificações médias para cada método.

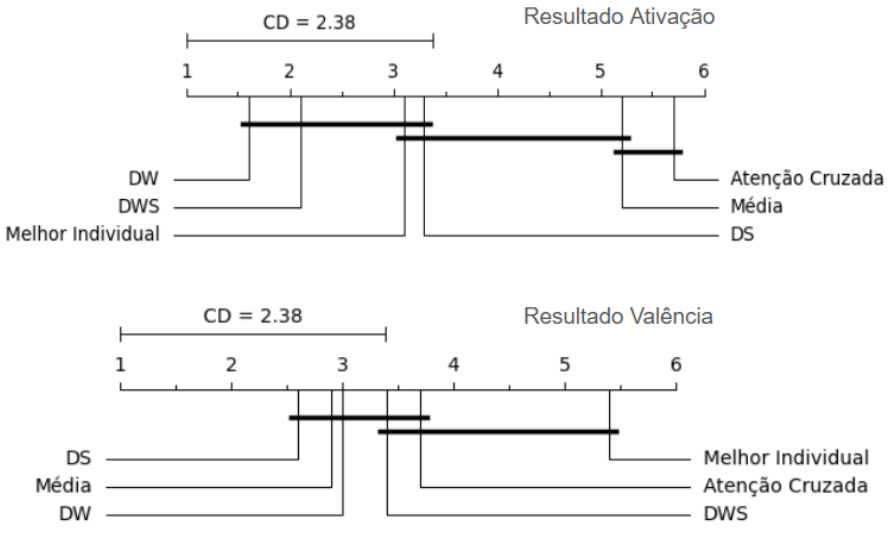


	Mean Rank
DS	3
DW	2
DWS	1,8
Média	3,8
Melhor Individual	5,9
Atenção Cruzada	4,5

	Mean Rank
DS	3,6
DW	1,6
DWS	1,8
Média	5,7
Melhor Individual	3,4
Atenção Cruzada	4,9

Fonte: Autoria Própria

Figura 26 – Diagrama de Nemenyi para comparações de classificações médias (mean rank) dos métodos em termos de desempenho CCC nas dimensões de ativação e valência, base de dados RECOLA, caso com 100% de ausência da modalidade visual. A linha horizontal conecta os métodos cuja diferença na classificação média não é estatisticamente significativa ao nível de confiança de 0,05, considerando uma distância crítica (CD) de 2,38. A tabela à direita apresenta as classificações médias para cada método.



	Mean Rank
DS	3,3
DW	1,6
DWS	2,1
Média	5,2
Melhor Individual	3,1
Atenção Cruzada	5,7

	Mean Rank
DS	2,6
DW	3
DWS	3,4
Média	2,9
Melhor Individual	5,4
Atenção Cruzada	3,7

Fonte: Autoria Própria

Na condição de ausência da modalidade de áudio, DW e DWS apresentam os melhores resultados, como discutido anteriormente. No entanto, para ativação não há diferença significativa entre DW, DWS, DS e média. Já para valência, não há diferença significativa entre DW, DWS, DS e melhor regressor individual.

Na condição de ausência da modalidade visual, DW e DWS também apresentam os melhores resultados para ativação e não há diferença significativa entre os resultados destes métodos quando comparados com DS e o melhor regressor individual. Já para valência, apenas o melhor regressor individual apresenta resultado significativamente inferior; os demais métodos estão empatados.

5.2.2 SEWA

As Tabelas 26 e 27 apresentam os resultados de ativação e valência, em termos de CCC, para todos os métodos comparados, incluindo a média das saídas dos regressores, o melhor regressor individual e as técnicas de seleção dinâmica (DS, DW, DWS). A Figura 27 compara o padrão-ouro de ativação, a predição com todas as modalidades ativas e com a ausência total de cada modalidade.

No contexto da ativação, DS apresenta o melhor desempenho médio (CCC = 0,45) em cenários hipotéticos com 100% de ausência da modalidade de áudio. DW e DWS também se destacam, apresentando uma média ligeiramente inferior a DS, mas ainda com resultados superiores aos demais métodos (CCC = 0,43 e 0,44). É interessante observar que o melhor regressor individual para ativação é o baseado em características de áudio, portanto, numa situação onde a modalidade de áudio está indisponível, seu resultado é bastante impactado, chegando a um CCC próximo de zero.

Para valência, ainda na condição de ausência total da modalidade de áudio, o desempenho de DW é superior aos outros métodos, com CCC médio de 0,50. DW e DWS também mantêm um desempenho elevado, com CCCs médios de 0,46. Em contraste, a média apresenta o pior desempenho, com CCC médio de 0,24.

Nos cenários de 100% de ausência da modalidade visual, o desempenho do melhor regressor individual é superior aos outros métodos, com CCC médio de 0,36. DWS, DW e DS mantêm um desempenho elevado, com CCCs médios de 0,31, 0,30 e 0,30, respectivamente. Em contraste, a média apresenta o pior desempenho, com CCC médio de 0,24.

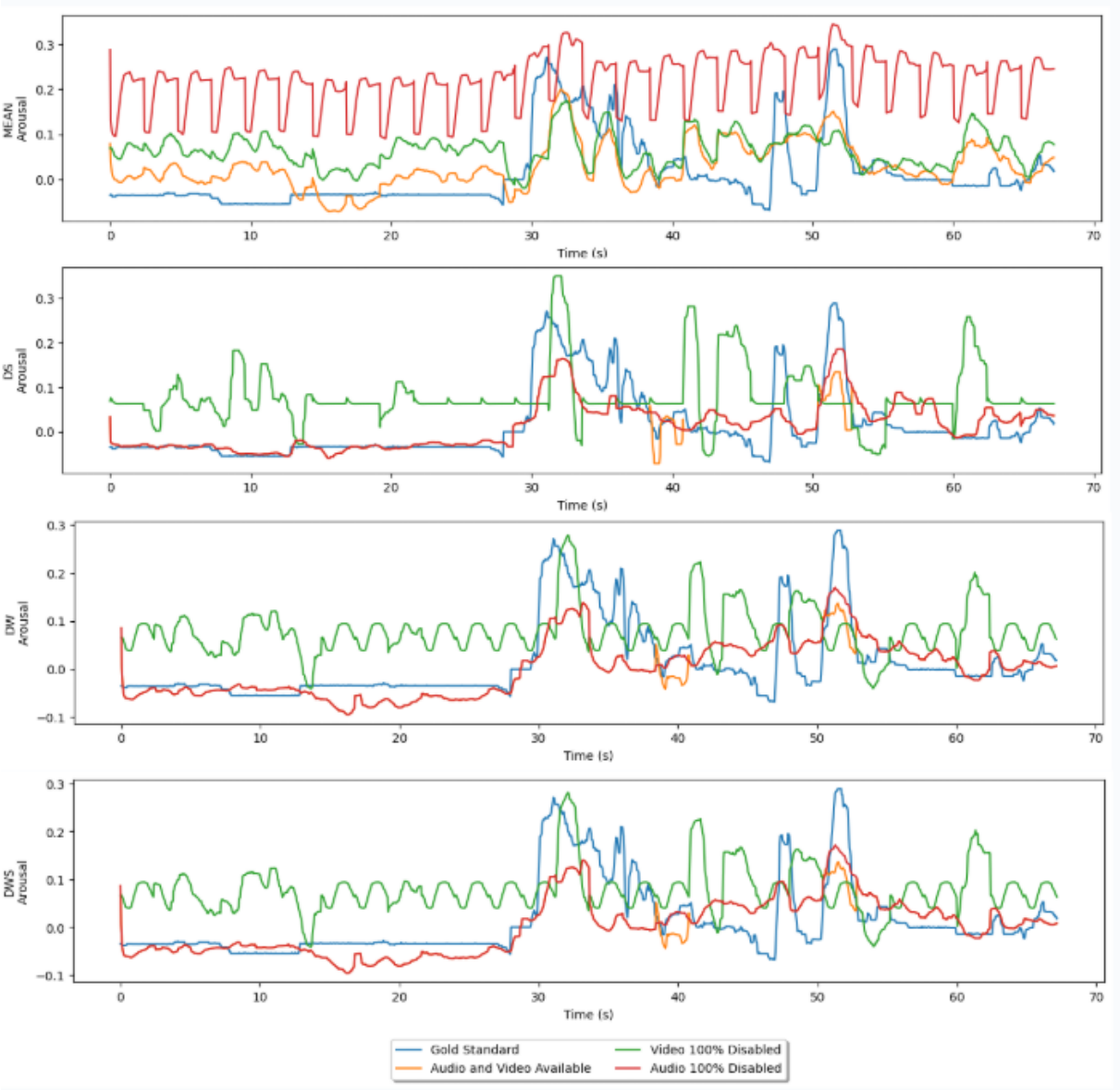
Tabela 26 – Resultados de ativação, em termos de CCC, base de dados SEWA, abrangendo a média das saídas dos regressores, melhor regressor individual e seleção dinâmica (DS, DW, DWS). Os resultados são apresentados nos seguintes cenários: áudio (A) e vídeo (V) disponíveis, áudio indisponível e vídeo indisponível. As proporções de dados ausentes analisadas foram de 25%, 50% e 100%.

Modalidades	**DS**	**DW**	**DWS**	**Média**	**Melhor individual**
A e V disponíveis	$0,46 \pm 0,07$	$\mathbf{0,46 \pm 0,04}$	$\mathbf{0,46 \pm 0,04}$	$0,46 \pm 0,07$	$0,36 \pm 0,08$
25% V indisponível	$0,36 \pm 0,06$	$\mathbf{0,43 \pm 0,05}$	$\mathbf{0,43 \pm 0,05}$	$0,41 \pm 0,07$	$0,36 \pm 0,10$
25% A indisponível	$0,44 \pm 0,08$	$\mathbf{0,44 \pm 0,04}$	$\mathbf{0,44 \pm 0,04}$	$0,40 \pm 0,05$	$0,26 \pm 0,08$
50% V indisponível	$0,34 \pm 0,07$	$\mathbf{0,40 \pm 0,05}$	$0,40 \pm 0,06$	$0,36 \pm 0,05$	$0,36 \pm 0,10$
50% A indisponível	$0,45 \pm 0,09$	$0,44 \pm 0,05$	$\mathbf{0,45 \pm 0,05}$	$0,34 \pm 0,06$	$0,19 \pm 0,06$
100% V indisponível	$0,30 \pm 0,10$	$0,30 \pm 0,10$	$0,31 \pm 0,10$	$0,24 \pm 0,05$	$\mathbf{0,36 \pm 0,10}$
100% A indisponível	$\mathbf{0,45 \pm 0,08}$	$0,43 \pm 0,07$	$0,44 \pm 0,07$	$0,23 \pm 0,05$	$0,0007 \pm 0,0046$

Tabela 27 – Resultados de valência, em termos de CCC, base de dados SEWA, abrangendo a média das saídas dos regressores, melhor regressor individual e seleção dinâmica (DS, DW, DWS). Os resultados são apresentados nos seguintes cenários: áudio (A) e vídeo (V) disponíveis, áudio indisponível e vídeo indisponível. As proporções de dados ausentes analisadas foram de 25%, 50% e 100%.

Modalidades	**DS**	**DW**	**DWS**	**Média**	**Melhor individual**
A e V disponíveis	$0,49 \pm 0,08$	$0,53 \pm 0,07$	$\mathbf{0,55 \pm 0,08}$	$0,52 \pm 0,08$	$0,44 \pm 0,09$
25% V indisponível	$0,48 \pm 0,07$	$0,50 \pm 0,08$	$\mathbf{0,52 \pm 0,09}$	$0,46 \pm 0,06$	$0,34 \pm 0,05$
25% A indisponível	$0,46 \pm 0,10$	$0,50 \pm 0,10$	$\mathbf{0,52 \pm 0,11}$	$0,45 \pm 0,07$	$0,44 \pm 0,09$
50% V indisponível	$0,46 \pm 0,06$	$0,48 \pm 0,07$	$\mathbf{0,49 \pm 0,07}$	$0,41 \pm 0,05$	$0,23 \pm 0,05$
50% A indisponível	$0,46 \pm 0,11$	$0,49 \pm 0,10$	$\mathbf{0,52 \pm 0,11}$	$0,40 \pm 0,07$	$0,44 \pm 0,09$
100% V indisponível	$0,41 \pm 0,06$	$0,42 \pm 0,07$	$\mathbf{0,43 \pm 0,08}$	$0,31 \pm 0,06$	$0,001 \pm 0,008$
100% A indisponível	$0,46 \pm 0,12$	$0,46 \pm 0,10$	$\mathbf{0,50 \pm 0,13}$	$0,24 \pm 0,09$	$0,44 \pm 0,09$

Figura 27 – Comparação do padrão-ouro de ativação, predição com todas as modalidades ativas, predição com a ausência da modalidade de áudio e predição com a ausência da modalidade de vídeo da média das saídas dos regressores e métodos baseados em seleção dinâmica (DS, DW, DWS). A imagem foi gerada usando o caso de teste T2 (segunda pessoa do conjunto de testes) do segundo conjunto de validação cruzada (k=2), base de dados SEWA.



Fonte: Autoria Própria

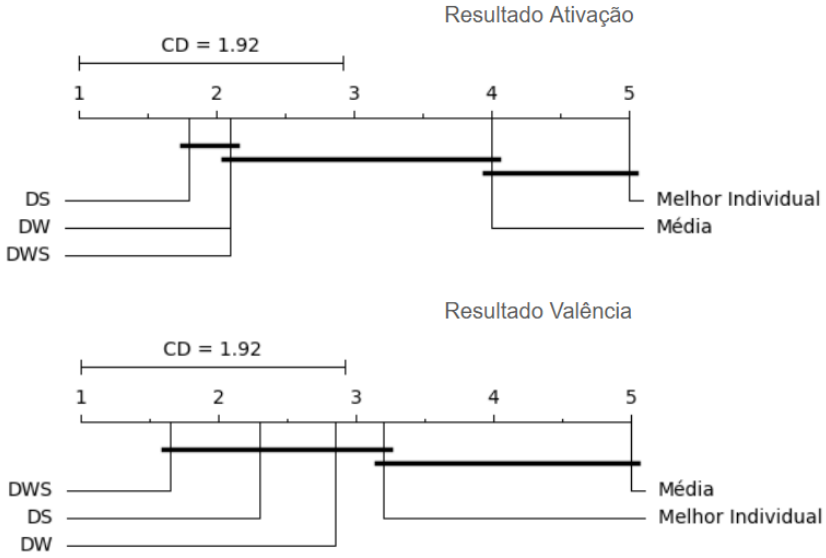
Para valência, ainda na condição de ausência total da modalidade de vídeo, o desempenho de DWS é superior aos outros métodos, com CCC médio de 0,43. DW e DS também mantêm um desempenho elevado, com CCCs médios de 0,42 e 0,41, respectivamente. O melhor regressor individual para valência é o baseado em características de vídeo; portanto, numa situação onde a modalidade de vídeo está indisponível, seu resultado é bastante impactado, chegando a um CCC próximo de zero.

O teste de Friedman (FRIEDMAN, 1937) com um nível de significância $\alpha = 0,05$ revelou que, nas dimensões de ativação e valência, há uma diferença significativa na aplicação de pelo menos um dos métodos analisados em todos os cenários avaliados, que contemplam condições com ausência total das modalidades de áudio e de vídeo. O teste post-hoc de Nemenyi (NEMENYI, 1963) foi utilizado para ranquear os métodos e esclarecer quais métodos apresentam diferenças significativas entre si, apresentando os resultados comparativos nas Figuras 28 e 29.

Na condição de ausência da modalidade de áudio, DS, DW e DWS apresentam os melhores resultados para ativação, como discutido anteriormente. No entanto, não há diferença significativa entre estes métodos. Já para valência, não há diferença significativa entre DS, DW, DWS e melhor regressor individual.

Na condição de ausência da modalidade visual, o melhor regressor individual, DS, DW e DWS também apresentam os melhores resultados para ativação e não há diferença significativa entre os resultados destes métodos. Já para valência, apenas a média e o melhor regressor individual apresentam resultado significativamente inferior; os demais métodos estão empatados.

Figura 28 – Diagrama de Nemenyi para comparações de classificações médias (mean rank) dos métodos em termos de desempenho CCC nas dimensões de ativação e valência, base de dados SEWA, caso com 100% de ausência da modalidade de áudio. A linha horizontal conecta os métodos cuja diferença na classificação média não é estatisticamente significativa ao nível de confiança de 0,05, considerando uma distância crítica (CD) de 1,92. A tabela à direita apresenta as classificações médias para cada método.

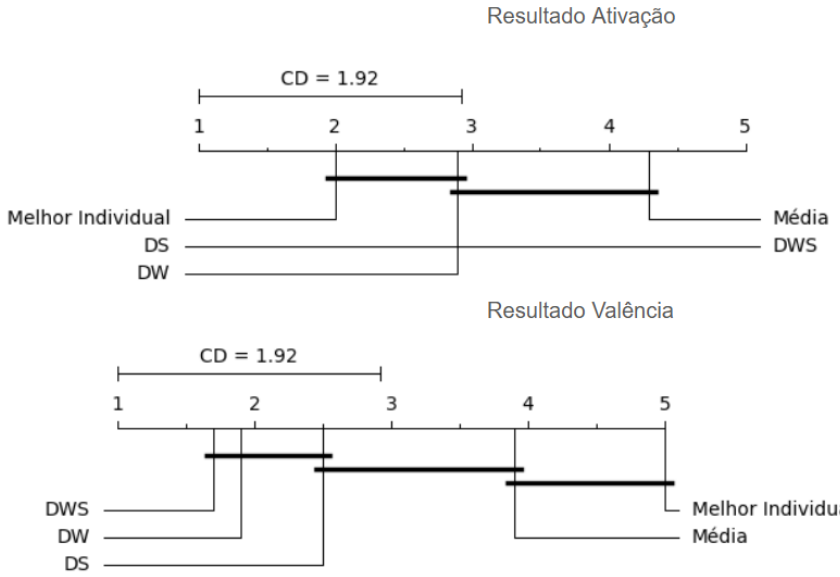


	Mean Rank
DS	1,8
DW	2,1
DWS	2,1
Média	4
Melhor Individual	5

	Mean Rank
DS	2,3
DW	2,85
DWS	1,65
Média	5
Melhor Individual	3,2

Fonte: Autoria Própria

Figura 29 – Diagrama de Nemenyi para comparações de classificações médias (mean rank) dos métodos em termos de desempenho CCC nas dimensões de ativação e valência, base de dados SEWA, caso com 100% de ausência da modalidade visual. A linha horizontal conecta os métodos cuja diferença na classificação média não é estatisticamente significativa ao nível de confiança de 0,05, considerando uma distância crítica (CD) de 1,92. A tabela à direita apresenta as classificações médias para cada método.



	Mean Rank
DS	2,9
DW	2,9
DWS	2,9
Média	4,3
Melhor Individual	2

	Mean Rank
DS	2,5
DW	1,9
DWS	1,7
Média	3,9
Melhor Individual	5

Fonte: Autoria Própria

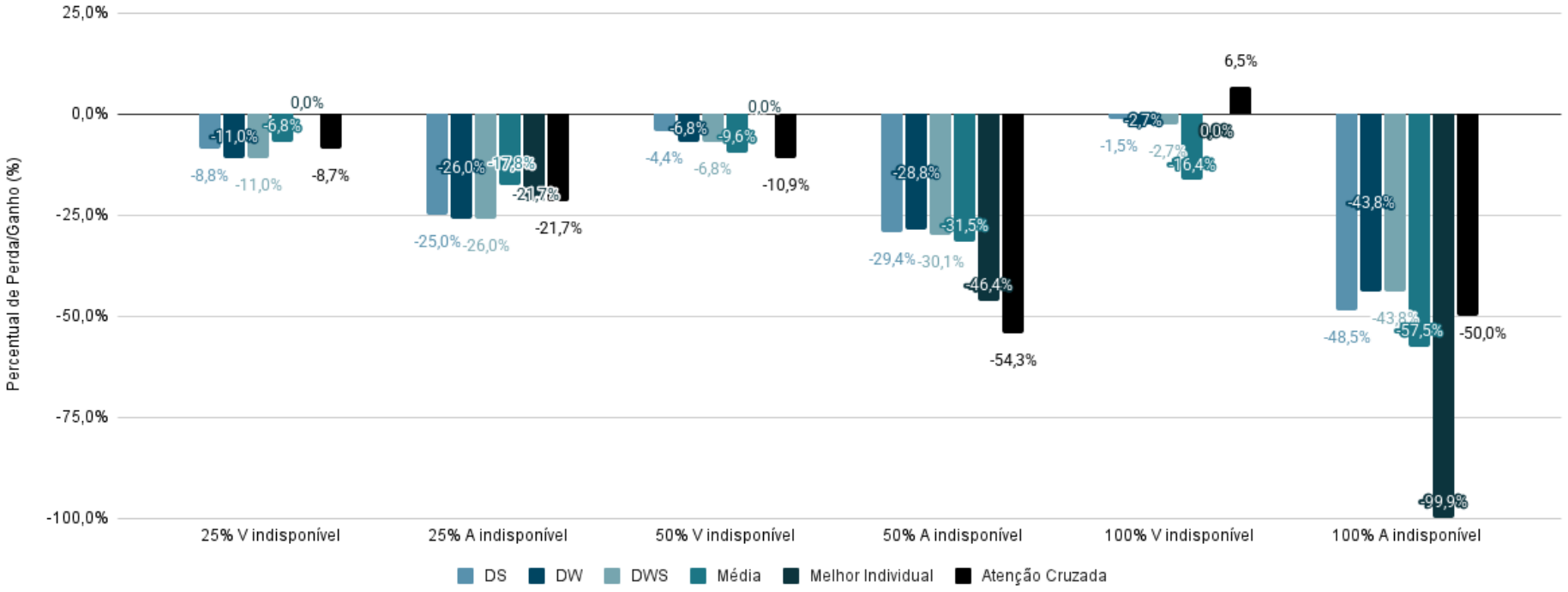
5.3 Análise de Sensibilidade a Modalidade Ausente

Nesta seção, discutiremos a sensibilidade dos métodos analisados, comparando os resultados (perdas/ganhos) obtidos nas condições de indisponibilidade de dados com aqueles obtidos na condição ideal, em que áudio e vídeo estão disponíveis.

5.3.1 RECOLA

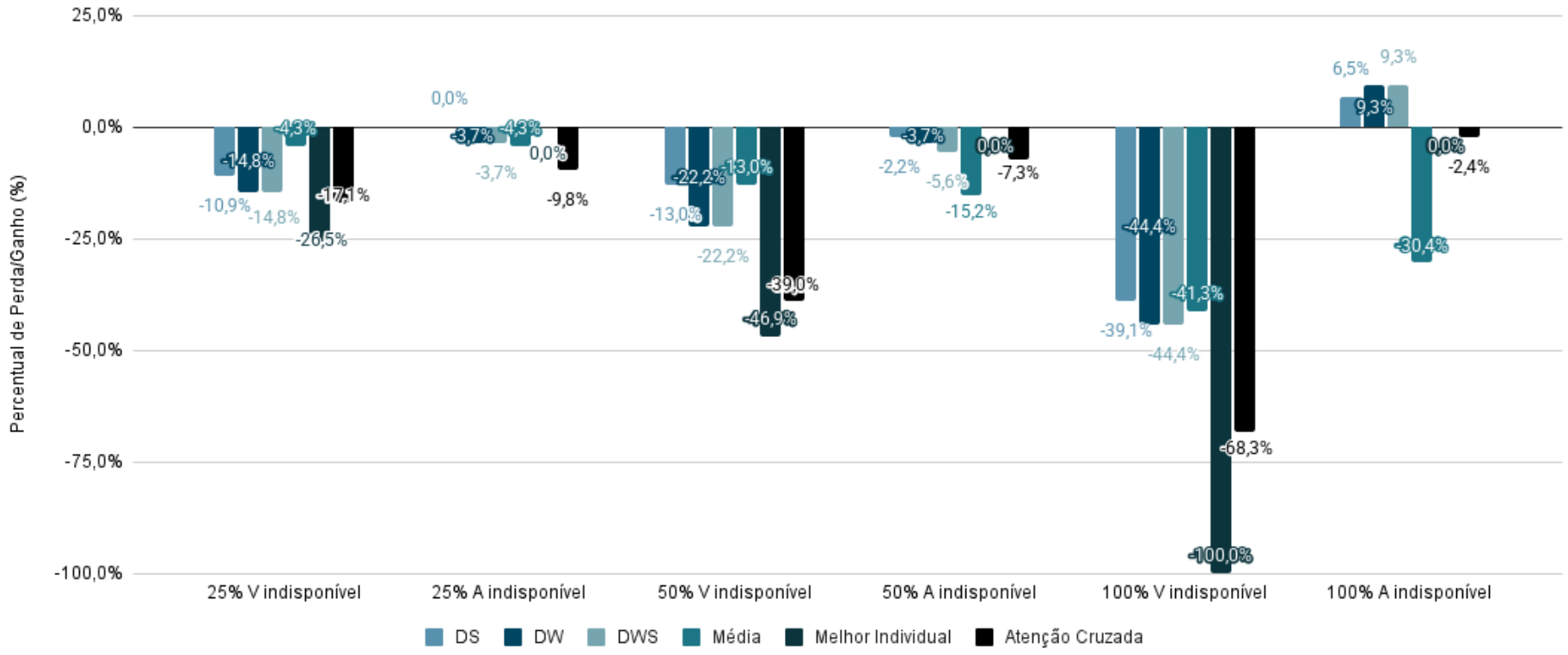
As Figuras 30 e 31 mostram o percentual de perda ou ganho de desempenho na base de dados RECOLA para os métodos DS, DW, DWS, Média, Melhor Individual e Atenção Cruzada, nos cenários de indisponibilidade de vídeo (V indisponível) e de áudio (A indisponível) em relação ao cenário de referência com ambas as modalidades disponíveis.

Figura 30 – Gráfico comparativo mostrando o percentual de perda ou ganho de desempenho para diferentes métodos de classificação multimodal (DS, DW, DWS, Média, Melhor Individual e Atenção Cruzada) para regressão de ativação nos cenários de indisponibilidade de vídeo (V indisponível) e de áudio (A indisponível) em relação ao cenário de referência com ambas as modalidades disponíveis, base de dados RECOLA.



Fonte: Autoria Própria

Figura 31 – Gráfico comparativo mostrando o percentual de perda ou ganho de desempenho para diferentes métodos de classificação multimodal (DS, DW, DWS, Média, Melhor Individual e Atenção Cruzada) para regressão de valência nos cenários de indisponibilidade de vídeo (V indisponível) e de áudio (A indisponível) em relação ao cenário de referência com ambas as modalidades disponíveis, base de dados RECOLA.



Fonte: Autoria Própria

No contexto da ativação, o método de atenção cruzada demonstrou maior robustez em termos de sensibilidade, exibindo um aumento de 6,52% no CCC quando a modalidade de vídeo estava totalmente ausente, em comparação com o cenário ideal em que tanto as modalidades de áudio quanto de vídeo estavam disponíveis. Já o melhor regressor individual para ativação, baseado em características de áudio, manteve seu desempenho inalterado, sem qualquer impacto decorrente da ausência da modalidade visual. Contrariamente, os métodos restantes mantiveram seu desempenho ou experimentaram alguma perda. Entre os métodos baseados em seleção dinâmica, o DS exibiu redução mínima de 1,47% no desempenho, enquanto DW e DWS mostraram uma diminuição de 2,74%. O método da média observou o declínio mais substancial, registrando uma perda significativa de 16,44%.

Perdas de desempenho mais pronunciadas foram observadas quando a modalidade de áudio estava totalmente ausente. Diversas abordagens apresentaram uma queda superior a 50% no desempenho, o que é justificável, considerando que a modalidade de áudio representa de maneira mais eficaz a dimensão de ativação. DW e DWS emergiram como as abordagens mais robustas em cenários sem áudio, experimentando uma queda de desempenho de 43,84% em comparação com o cenário com todas as modalidades disponíveis. O método DS apresentou desempenho ligeiramente inferior, com uma redução de 48,53% no CCC. Em contrapartida, os métodos de média e o melhor regressor individual exibiram as maiores quedas de desempenho, registrando perdas significativas de 57,53% e 99,86%, respectivamente, evidenciando uma fragilidade em cenários de ausência da modalidade de áudio.

Na dimensão de valência, observou-se um padrão contrastante, em que a ausência total da modalidade de áudio resultou em desempenhos superiores. Quando a modalidade de vídeo foi desativada, o método DS demonstrou ser a abordagem menos sensível, com uma queda de desempenho de 39,13%. No mesmo cenário, o melhor regressor individual não apresentou nenhuma perda relacionada à ausência da modalidade de áudio. Depois dele, DW, DWS e DS emergiram como os métodos mais robustos, mostrando um aumento de 9,26%, 9,26% e 6,52% no CCC, respectivamente.

No método de atenção cruzada, em relação à ativação, houve um aumento notável de 6,52% no CCC quando a modalidade de vídeo estava ausente, mas uma queda substancial de 50% no desempenho quando a modalidade de áudio foi desativada. Em relação à valência, resultados promissores foram observados quando a modalidade de vídeo foi desativada, com uma redução de desempenho de apenas 2,44%. No entanto, a desativação da modalidade de vídeo resultou em uma queda significativa de 68,29% no desempenho quando comparada ao cenário ideal, em que as modalidades de áudio e vídeo estavam disponíveis.

Cenários com 25% e 50% de dados ausentes foram identificados padrões semelhantes. Uma observação interessante é que, para ativação, quanto maior a proporção de áudio

indisponível, pior é o desempenho obtido. Esses resultados reforçam a importância do áudio na predição dessa dimensão. Por outro lado, à medida que aumentamos a proporção de vídeo desativado, os resultados melhoram, pois a solução passa a depender mais das características de áudio. Para a valência, observou-se um padrão inverso: o desempenho piorou à medida que a proporção de vídeo indisponível aumentou, enquanto a redução da disponibilidade de áudio resultou em melhorias nos resultados.

5.3.2 SEWA

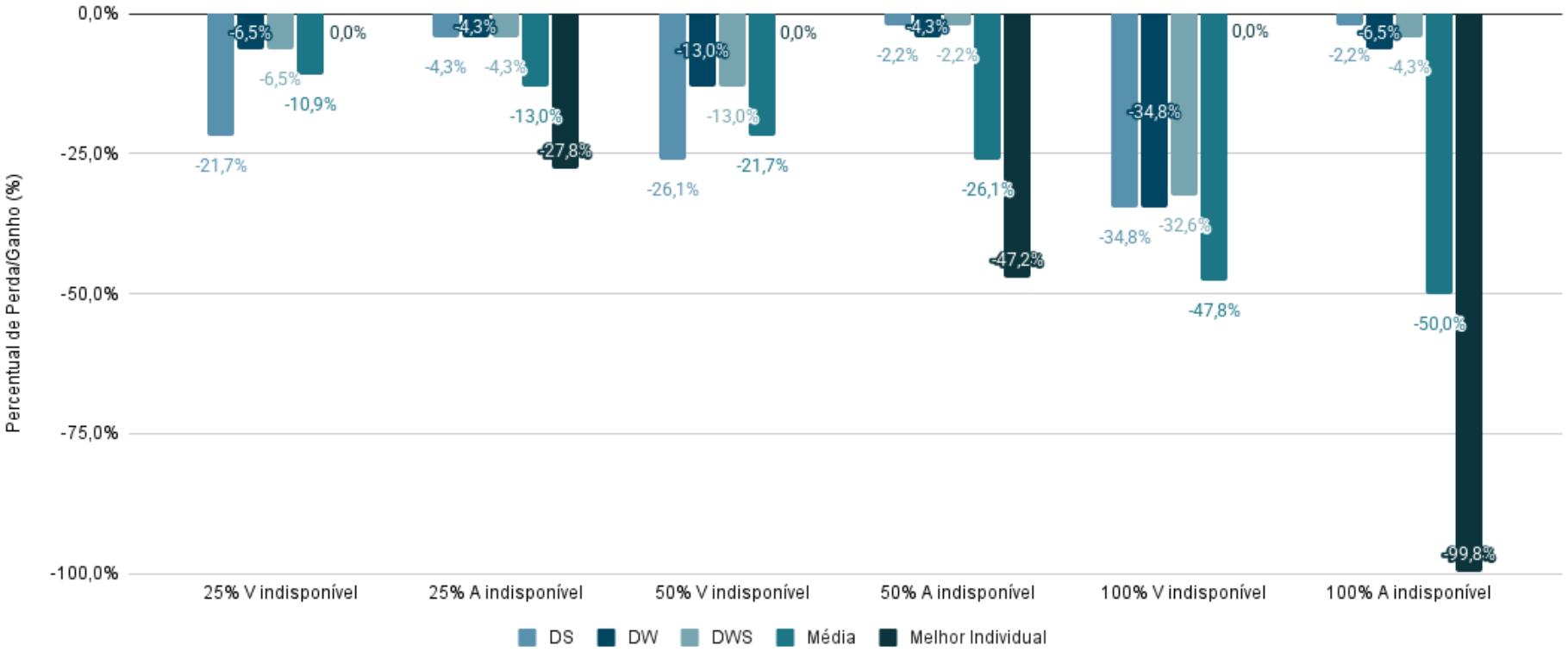
As Figuras 32 e 33 mostram o percentual de perda ou ganho de desempenho na base de dados SEWA para os métodos DS, DW, DWS, Média e Melhor Individual, nos cenários de indisponibilidade de vídeo (V indisponível) e de áudio (A indisponível) em relação ao cenário de referência com ambas as modalidades disponíveis.

No contexto da ativação, o melhor regressor individual, baseado em características de áudio, demonstrou a maior robustez em termos de sensibilidade decorrente da ausência da modalidade visual, mantendo seu desempenho inalterado. Contrariamente, os métodos restantes experimentaram alguma perda. Entre os métodos baseados em seleção dinâmica, o DWS exibiu uma redução de 32,61% no desempenho, enquanto DS e DW mostraram uma diminuição de 34,78%. O método da média observou o declínio mais substancial, registrando uma perda significativa de 47,83%.

Perdas de desempenho mais pronunciadas foram observadas quando a modalidade de áudio estava totalmente ausente, com quedas superiores a 50% nos métodos de fusão estática das modalidades, melhor regressor individual e da média das saídas dos regressores. DS e DWS emergiram como as abordagens mais robustas em cenários sem áudio, experimentando uma queda de desempenho de 2,17% e 4,35% em comparação com o cenário com todas as modalidades disponíveis. O método DW apresentou desempenho ligeiramente inferior, com uma redução de 6,52% no CCC.

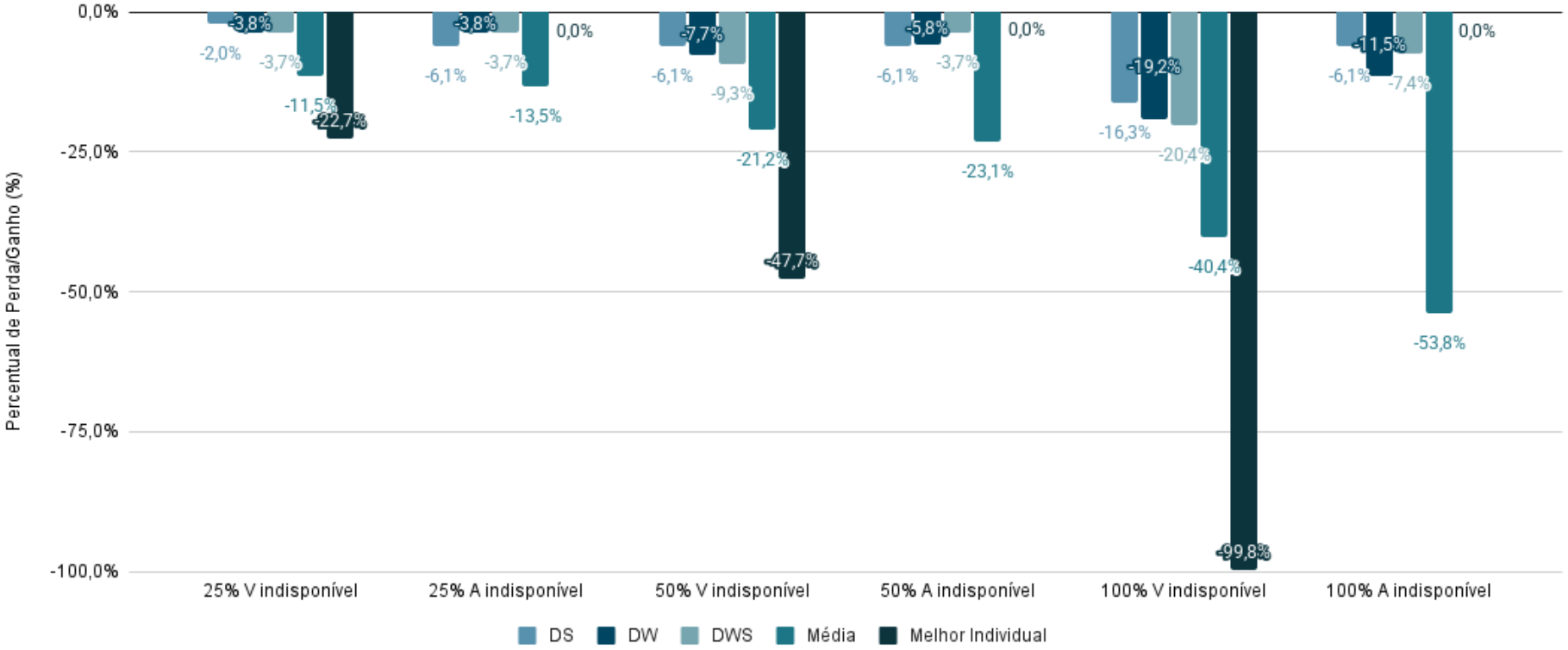
Na dimensão de valência, observou-se um padrão contrastante, em que a ausência total da modalidade de áudio resultou em desempenhos superiores. Quando a modalidade de vídeo foi desativada, o método DS demonstrou ser a abordagem menos sensível, com uma queda de desempenho de 16,33%. No mesmo cenário, o melhor regressor individual não apresentou nenhuma perda relacionada à ausência da modalidade de áudio. Depois dele, DS, DWS e DW emergiram como os métodos mais robustos, mostrando quedas de 6,12%, 7,41% e 11,54% no CCC, respectivamente.

Figura 32 – Gráfico comparativo mostrando o percentual de perda ou ganho de desempenho para diferentes métodos de classificação multimodal (DS, DW, DWS, Média, Melhor Individual e Atenção Cruzada) para regressão de ativação nos cenários de indisponibilidade de vídeo (V indisponível) e de áudio (A indisponível) em relação ao cenário de referência com ambas as modalidades disponíveis, base de dados SEWA.



Fonte: Autoria Própria

Figura 33 – Gráfico comparativo mostrando o percentual de perda ou ganho de desempenho para diferentes métodos de classificação multimodal (DS, DW, DWS, Média, Melhor Individual e Atenção Cruzada) para regressão de valência nos cenários de indisponibilidade de vídeo (V indisponível) e de áudio (A indisponível) em relação ao cenário de referência com ambas as modalidades disponíveis, base de dados SEWA.



Fonte: Autoria Própria

Cenários com 25% e 50% de dados ausentes foram identificados padrões semelhantes. Para ativação, quanto maior a proporção de áudio indisponível, pior é o desempenho obtido. Para a valência, o padrão é inverso: o desempenho piora à medida que a proporção de vídeo indisponível aumenta. Esses resultados indicam que a modalidade de áudio exerce maior influência na detecção da ativação e a modalidade de vídeo exerce maior influência na detecção da valência.

Esses resultados reforçam a robustez dos métodos de seleção dinâmica, em particular DW e DWS, que, mesmo em condições de perda parcial de dados, conseguem manter uma performance elevada em comparação aos métodos de fusão estática das representações de dados, que se mostram menos adaptáveis à ausência de informações de áudio ou vídeo.

5.4 Distribuição Geral de Modalidades e Representações

Esta seção apresenta os resultados obtidos em condições ideais e sob diferentes níveis de indisponibilidade das modalidades, destacando como o sistema adapta suas escolhas de modalidade (áudio/vídeo) e representação (Aparência, Geometria, ResNet, Características Acústicas, MFCCs e Mel Espectrogramas) conforme o cenário.

5.4.1 RECOLA

A Tabela 28 apresenta a distribuição percentual de escolha das modalidades de áudio e vídeo na base de dados RECOLA. No contexto da ativação, com ambas as modalidades disponíveis, o áudio domina (83,67%), sugerindo que a prosódia e a entonação vocal são fortes indicadores de ativação. À medida que o vídeo se torna parcial ou totalmente indisponível, o áudio alcança até 91,84%. Caso o áudio seja a modalidade ausente, a preferência pelo modal visual aumenta, mas mantém-se em torno de 43,54% quando não há áudio algum. Na dimensão de valência, em condições ideais, o vídeo (56,46%) supera ligeiramente o áudio (43,54%). Conforme o vídeo é reduzido, o sistema recorre cada vez mais à modalidade acústica (até 67,35%), e, se o áudio estiver indisponível, o vídeo passa a ser a única fonte de informações (100%).

Esses números confirmam a importância do áudio para ativação, consonante com a literatura que relaciona intensidade emocional à fala. Para valência, as pistas visuais (expressões faciais, movimentos sutis) ganham maior destaque, embora o áudio contribua de forma relevante quando o vídeo está comprometido.

A Tabela 29 apresenta as preferências por representações em cada uma das abordagens, DS, DW e DWS. No contexto de DS, certas representações podem ser totalmente descartadas, privilegiando aquelas que demonstram melhor correlação com a emoção alvo. Para ativação, MFCCs (A) e Mel Espectrogramas (A) tendem a surgir com maior frequência, ao passo que, para valência, observa-se ênfase em Geometria (V) e ResNet (V).

Em DW, todos os extratores permanecem ativos, mas com pesos distintos. No caso de ativação, o áudio pode chegar a superar 60% no total (considerando MFCC, Mel Espectrogramas e características acústicas), ao mesmo tempo em que as representações de vídeo recebem participação menor. Já para valência, ocorre o inverso, com atribuição de maior peso para a modalidade visual.

Por fim, DWS reúne elementos de ambas as abordagens, resultando em um modelo flexível que pode privilegiar MFCC (A) ou Mel Espectrogramas (A) na ativação e, simultaneamente, reduzir a participação de certos extratores visuais. A valência apresenta predominância de Geometria (V) e ResNet (V), evidenciando a capacidade do método em adaptar-se às condições de cada modalidade.

5.4.2 SEWA

No contexto da ativação, com ambas as modalidades disponíveis, o vídeo domina (66,96%), como é apresentado na Tabela 30. Mesmo com 25% de indisponibilidade na modalidade visual, a modal de vídeo mantém-se acima de 67%. Somente ao ampliar essa indisponibilidade (50% ou 100%), o áudio passa a assumir frações consideravelmente maiores, atingindo, por exemplo, 70,54% quando o vídeo está totalmente ausente. Para a valência, o vídeo apresenta um leve predomínio (53,13%) em condições ideais. À medida que o vídeo sofre reduções de disponibilidade, o áudio cresce, mas não chega a ser a modalidade dominante (atinge no máximo 34,38% quando o vídeo está completamente indisponível). Se o áudio estiver totalmente ausente (100% indisponibilidade), o vídeo prevalece em 88,39% dos casos.

Esses resultados contrastam com o observado na RECOLA, pois, na SEWA, as pistas visuais se mostram mais marcantes tanto para a ativação quanto para a valência — possivelmente devido a expressões faciais mais intensas ou maior variabilidade gestual, além de fatores como menor entonação vocal ou ruídos que prejudicam a clareza da modal acústica.

A Tabela 31 apresenta as preferências por representações na base de dados SEWA. Em cenários de boa qualidade de vídeo, a abordagem DS tende a destacar representações visuais como Geometria (V) e ResNet (V), chegando a descartar completamente certos extratores de áudio. No entanto, quando o vídeo se torna ausente ou possui restrições severas, MFCCs (A) e Mel Espectrogramas (A) tornam-se fundamentais para a análise de ativação e valência.

O método DW mantém todas as representações ativas, porém com pesos diferentes. Na SEWA, há uma clara ênfase na modalidade visual tanto para ativação quanto para valência, ocupando entre 60% e 70% da composição total em condições ideais. Ainda assim, o áudio permanece em uso, com pesos menores (cerca de 10% a 15% para cada extrator

acústico).

Por fim, DWS combina a lógica de seleção parcial com ajustes de peso, resultando em um comportamento equilibrado. Quando o vídeo está disponível e em boa qualidade, ele tende a responder pela maior parcela do processo de estimação (entre 60% e 70%). Se ocorre indisponibilidade significativa da modalidade visual, o áudio assume maior relevância, demonstrando flexibilidade na presença de diferentes graus de restrição.

5.5 Considerações Finais

Na base RECOLA, características acústicas se mostram especialmente relevantes para ativação, validando observações prévias de que esta dimensão tende a correlacionar-se com parâmetros de entonação e intensidade no áudio. Já expressões faciais aparecem como principais indicativos de valência, embora o sistema se adapte ao áudio quando a modalidade visual falha. Em cenários de indisponibilidade total ou parcial de alguma das modalidades, o modelo realoca o foco para a modalidade restante, atestando a eficácia da seleção dinâmica.

Na base SEWA, o domínio visual é decisivo para a ativação, em contraste com a RECOLA, onde o áudio prevalece para essa dimensão. Para a valência, as expressões faciais novamente se destacam, enquanto o áudio atua como recurso complementar, especialmente quando o canal de vídeo se encontra comprometido. A diversidade de participantes (incluindo diferentes etnias) e de métodos de captura faz com que o sistema precise se adaptar a cenários mais heterogêneos, evidenciando a relevância de abordagens dinâmicas (DS, DW, DWS) para operar com condições mais próximas do ambiente real.

Tabela 28 – Distribuição Geral de Modalidades (A/V), base de dados RECOLA.

		A e V disp.	25% V indisp.	25% A indisp.	50% V indisp.	50% A indisp.	100% V indisp.	100% A indisp.
Ativação	Áudio (A)	**83,67%**	**87,07%**	**69,39%**	**86,39%**	**68,71%**	**91,84%**	**56,46%**
	Vídeo (V)	16,33%	12,93%	28,57%	13,61%	31,29%	8,16%	43,54%
Valência	Áudio (A)	43,54%	**56,46%**	29,93%	**61,90%**	22,45%	**67,35%**	0,00%
	Vídeo (V)	**56,46%**	43,54%	**70,07%**	38,10%	**77,55%**	32,65%	**100,00%**

Tabela 29 – Preferência por Representações (DS, DW, DWS), base de dados RECOLA.

			A e V disp.	25% V indisp.	25% A indisp.	50% V indisp.	50% A indisp.	100% V indisp.	100% A indisp.
Ativação	DS	Áudio (A)	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
		Aparência (V)	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
		Geometria (V)	3,40%	12,93%	6,12%	13,61%	4,76%	8,16%	12,93%
		Resnet (V)	12,93%	0,00%	24,49%	0,00%	26,53%	0,00%	30,61%
		MFCCs (A)	41,50%	40,82%	**45,58%**	40,14%	31,29%	42,86%	0,00%
		Mel Espec (A)	**42,18%**	**46,26%**	23,81%	**46,26%**	**37,41%**	**48,98%**	**56,46%**
	DW	Áudio (A)	26,37%	27,45%	22,26%	27,23%	22,28%	28,95%	18,72%
		Aparência (V)	5,30%	4,29%	9,95%	4,52%	10,17%	2,72%	14,16%
		Geometria (V)	5,50%	4,35%	10,30%	4,57%	10,52%	2,73%	14,66%
		Resnet (V)	5,53%	4,28%	10,37%	4,51%	10,61%	2,71%	14,71%
		MFCCs (A)	28,61%	29,76%	**23,81%**	29,52%	**23,34%**	31,38%	18,71%
		Mel Espec (A)	**28,70%**	**29,87%**	23,32%	**29,64%**	23,08%	**31,50%**	**19,04%**
	DWS	Áudio (A)	25,14%	26,20%	22,07%	26,00%	22,55%	27,65%	19,57%
		Aparência (V)	4,70%	4,26%	8,85%	4,49%	9,03%	2,71%	**12,66%**
		Geometria (V)	5,83%	4,37%	10,93%	4,58%	11,18%	2,71%	15,60%
		Resnet (V)	5,80%	4,30%	10,83%	4,53%	11,08%	2,74%	15,28%
		MFCCs (A)	**29,42%**	**30,60%**	**24,75%**	**30,35%**	**23,55%**	**32,24%**	17,53%
		Mel Espec (A)	29,11%	30,27%	22,57%	30,05%	22,61%	31,95%	19,36%

Table 29 continued from previous page

			A e V disp.	25% V indisp.	25% A indisp.	50% V indisp.	50% A indisp.	100% V indisp.	100% A indisp.
Valência	DS	Áudio (A)	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
		Aparência (V)	0,00%	4,76%	0,00%	4,08%	0,00%	0,00%	0,00%
		Geometria (V)	8,84%	38,10%	12,93%	34,01%	13,61%	32,65%	23,13%
		Resnet (V)	**47,62%**	0,68%	**57,14%**	0,00%	**63,95%**	0,00%	**76,87%**
		MFCCs (A)	0,68%	0,68%	14,29%	0,68%	12,24%	1,36%	0,00%
		Mel Espec (A)	42,86%	**55,78%**	15,65%	**61,22%**	10,20%	**65,99%**	0,00%
	DW	Áudio (A)	14,12%	18,29%	9,75%	20,06%	7,31%	21,82%	0,00%
		Aparência (V)	18,07%	14,45%	22,42%	12,66%	24,81%	10,86%	32,00%
		Geometria (V)	19,03%	14,62%	23,64%	12,76%	26,16%	10,90%	33,78%
		Resnet (V)	**19,36%**	14,47%	**24,01%**	12,68%	**26,58%**	10,89%	**34,22%**
		MFCCs (A)	14,57%	**18,90%**	10,08%	**20,72%**	7,56%	**22,53%**	0,00%
		Mel Espec (A)	14,85%	19,27%	10,10%	21,13%	7,58%	22,99%	0,00%
	DWS	Áudio (A)	13,54%	17,50%	9,40%	19,19%	7,07%	20,87%	0,00%
		Aparência (V)	15,96%	14,02%	19,74%	12,26%	21,84%	10,50%	28,17%
		Geometria (V)	20,04%	16,37%	24,95%	14,32%	27,63%	12,25%	35,62%
		Resnet (V)	**20,46%**	13,15%	**25,38%**	11,52%	**28,08%**	9,90%	**36,20%**
		MFCCs (A)	15,39%	19,99%	10,70%	21,92%	8,01%	23,82%	0,00%
		Mel Espec (A)	14,61%	**18,98%**	9,83%	**20,80%**	7,38%	**22,65%**	0,00%

Tabela 30 – Distribuição Geral de Modalidades (A/V), base de dados SEWA.

		A e V disp.	25% V indisp.	25% A indisp.	50% V indisp.	50% A indisp.	100% V indisp.	100% A indisp.
Ativação	Áudio (A)	33,04%	32,14%	13,39%	41,07%	9,82%	**70,54%**	0,00%
	Vídeo (V)	**66,96%**	**67,86%**	**86,61%**	**58,93%**	**90,18%**	29,46%	**100,00%**
Valência	Áudio (A)	46,88%	**45,98%**	33,93%	42,86%	29,91%	34,38%	11,61%
	Vídeo (V)	**53,13%**	54,02%	**66,07%**	**57,14%**	**70,09%**	**65,63%**	**88,39%**

Tabela 31 – Preferência por Representações (DS, DW, DWS), base de dados SEWA.

			A e V disp.	25% V indisp.	25% A indisp.	50% V indisp.	50% A indisp.	100% V indisp.	100% A indisp.
Ativação	DS	Áudio (A)	0,45%	0,00%	0,45%	0,00%	0,45%	0,45%	0,00%
		Aparência (V)	0,00%	21,43%	0,00%	**28,57%**	0,00%	29,46%	0,00%
		Geometria (V)	0,45%	**44,64%**	0,45%	**28,57%**	0,45%	0,00%	0,45%
		Resnet (V)	**66,52%**	1,79%	**86,16%**	1,79%	**89,73%**	0,00%	**99,55%**
		MFCCs (A)	11,61%	11,16%	4,02%	16,07%	1,34%	24,11%	0,00%
		Mel Espec (A)	20,98%	20,98%	8,93%	25,00%	8,04%	**45,98%**	0,00%
	DW	Áudio (A)	10,72%	10,41%	4,36%	13,31%	3,19%	**22,87%**	0,00%
		Aparência (V)	21,38%	22,45%	27,63%	**19,58%**	28,76%	9,91%	31,90%
		Geometria (V)	22,25%	**23,03%**	28,77%	**19,85%**	29,96%	9,71%	33,23%
		Resnet (V)	**23,33%**	22,38%	**30,20%**	19,50%	**31,46%**	9,84%	**34,88%**
		MFCCs (A)	11,13%	10,83%	4,51%	13,85%	3,30%	23,77%	0,00%
		Mel Espec (A)	11,18%	10,90%	4,52%	13,92%	3,32%	23,89%	0,00%
	DWS	Áudio (A)	10,00%	9,72%	4,12%	12,35%	3,01%	21,31%	0,00%
		Aparência (V)	20,75%	22,93%	26,78%	**20,22%**	27,85%	10,57%	30,89%
		Geometria (V)	**23,25%**	**23,75%**	**30,10%**	20,09%	**31,32%**	9,29%	**34,70%**
		Resnet (V)	22,96%	21,17%	29,73%	18,62%	31,01%	9,61%	34,41%
		MFCCs (A)	11,51%	11,22%	4,56%	14,38%	3,36%	**24,67%**	0,00%
		Mel Espec (A)	11,53%	11,21%	4,71%	14,34%	3,45%	24,55%	0,00%

Tabela 31 continuação da página anterior

			A e V disp.	25% V indisp.	25% A indisp.	50% V indisp.	50% A indisp.	100% V indisp.	100% A indisp.
Valência	DS	Áudio (A)	0,00%	0,00%	0,00%	0,00%	0,00%	0,45%	0,00%
		Aparência (V)	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
		Geometria (V)	34,38%	**53,57%**	**41,96%**	**56,70%**	**45,09%**	**65,63%**	**58,04%**
		Resnet (V)	18,75%	0,45%	24,11%	0,45%	25,00%	0,00%	30,36%
		MFCCs (A)	8,48%	8,48%	7,59%	6,70%	5,80%	4,02%	0,00%
		Mel Espec (A)	**38,39%**	37,50%	26,34%	36,16%	24,11%	29,91%	11,61%
	DW	Áudio (A)	15,41%	15,10%	11,17%	14,07%	9,88%	11,28%	3,88%
		Aparência (V)	16,61%	17,66%	20,68%	18,76%	21,93%	21,68%	27,65%
		Geometria (V)	**18,32%**	**18,54%**	**22,77%**	**19,48%**	**24,16%**	**22,14%**	**30,48%**
		Resnet (V)	18,19%	17,82%	22,63%	18,91%	24,00%	21,80%	30,26%
		MFCCs (A)	15,64%	15,33%	11,32%	14,28%	9,96%	11,44%	3,83%
		Mel Espec (A)	15,83%	15,56%	11,44%	14,51%	10,08%	11,66%	3,90%
	DWS	Áudio (A)	15,29%	14,97%	11,10%	13,95%	9,83%	11,19%	3,89%
		Aparência (V)	13,91%	17,14%	17,40%	18,55%	18,45%	22,03%	23,28%
		Geometria (V)	**20,64%**	**19,70%**	**25,55%**	**20,17%**	**27,10%**	**22,04%**	**34,19%**
		Resnet (V)	18,58%	17,18%	23,13%	18,42%	24,54%	21,56%	30,93%
		MFCCs (A)	15,14%	14,83%	10,96%	13,81%	9,63%	11,08%	3,71%
		Mel Espec (A)	16,44%	16,18%	11,87%	15,10%	10,44%	12,11%	4,01%

6 Discussão

O método proposto de seleção dinâmica de modalidade e representação de dados foi efetivo no contexto de MER. Em condições ideais, utilizando as modalidades de áudio e vídeo, as melhores performances foram obtidas com DWS, alcançando um CCC de 0,73 para ativação e 0,54 para valência na base RECOLA e CCC de 0,46 para ativação e 0,55 para valência na base SEWA.

As abordagens de seleção dinâmica se destacaram pela robustez frente à ausência de modalidades, apresentando desempenho superior a todas as outras abordagens testadas, independentemente do grau de ausência dos dados (25%, 50% ou 100%). Para ativação, perdas de desempenho mais pronunciadas foram observadas quando a modalidade de áudio estava indisponível, enquanto para valência, as maiores perdas foram observadas com a indisponibilidade da modalidade visual.

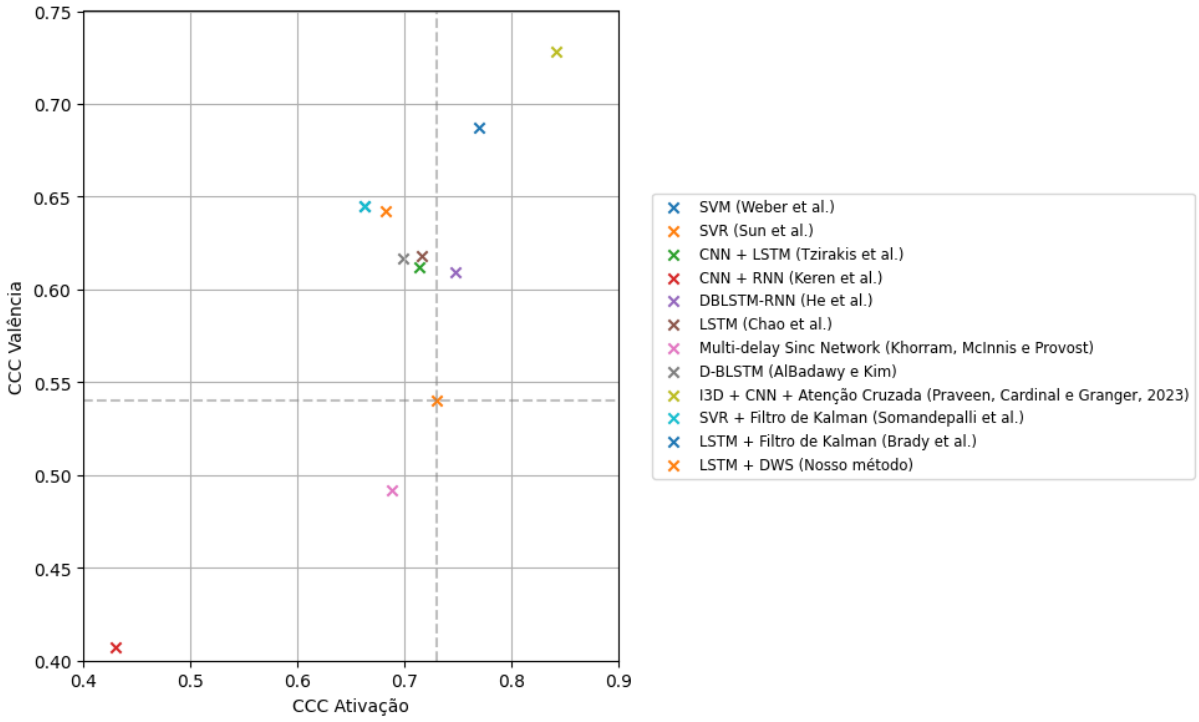
Estudos anteriores demonstram de forma consistente que características acústicas são mais informativas para a predição de ativação, enquanto que recursos visuais tendem a ser mais preditivos para valência (SOMANDEPALLI et al., 2016; KEREN et al., 2017; HAN et al., 2019). Essa diferença pode ser atribuída ao fato de que a ativação, associada à intensidade emocional, é mais claramente capturada por elementos como tom de voz, volume e taxa de fala, mesmo na ausência de pistas visuais (VALSTAR et al., 2016; HE et al., 2015b; BRADY et al., 2016). A entonação, por exemplo, desempenha um papel fundamental na comunicação, não apenas pela clareza na transmissão da mensagem, mas também por refletir o tom emocional e a intenção por trás das palavras (MIGUEL, 2015).

A valência, associada à positividade ou negatividade das emoções, é frequentemente refletida em expressões faciais, apresenta nuances mais sutis e complexas que dificilmente podem ser discernidas apenas a partir do áudio (VALSTAR et al., 2016; SUN et al., 2016; TZIRAKIS et al., 2017). Pistas visuais, como expressões faciais, são cruciais para identificar a valência emocional, tornando o vídeo a modalidade mais informativa para essa dimensão. Estudos mostram que, ao tentar reconhecer emoções, participantes humanos tiveram maior sucesso em identificar nuances de valência quando puderam observar expressões faciais comparando com julgamentos baseados apenas em sinais de voz, destacando a relevância das pistas visuais na interpretação desta dimensão (ZAMBELI; LIRA; CASSOL, 2024).

A comparação dos resultados do método proposto com o estado da arte nas bases RECOLA e SEWA evidencia sua competitividade, especialmente na dimensão de ativação (Figuras 34 e 35). Praveen et al. (2022) alcançaram os melhores resultados entre os artigos analisados da base RECOLA empregando 2D-CNNs para processar espectrogramas de áudio e Inflated 3D-CNNs para analisar diretamente as faces extraídas de cada quadro dos

vídeos. Essa abordagem explora a capacidade das redes profundas de extrair representações diretamente dos sinais brutos. Outros seguem a mesma linha, utilizando redes neurais como extratoras de características diretamente sobre os sinais brutos de áudio e vídeo (TZIRAKIS et al., 2017; KEREN et al., 2017). O melhor resultado da base da SEWA foi alcançado por Cruz et al. (2024) e um método de aumento de dados baseado que cria sequências sintéticas de expressões faciais para valores sub-representados no espaço ativação-valência do conjunto de dados. A adoção de representações mais sofisticadas e a criação de sequências de dados sintéticas poderia potencialmente ampliar a capacidade do método proposto em generalizar e identificar padrões emocionais com maior precisão, especialmente na dimensão de valência.

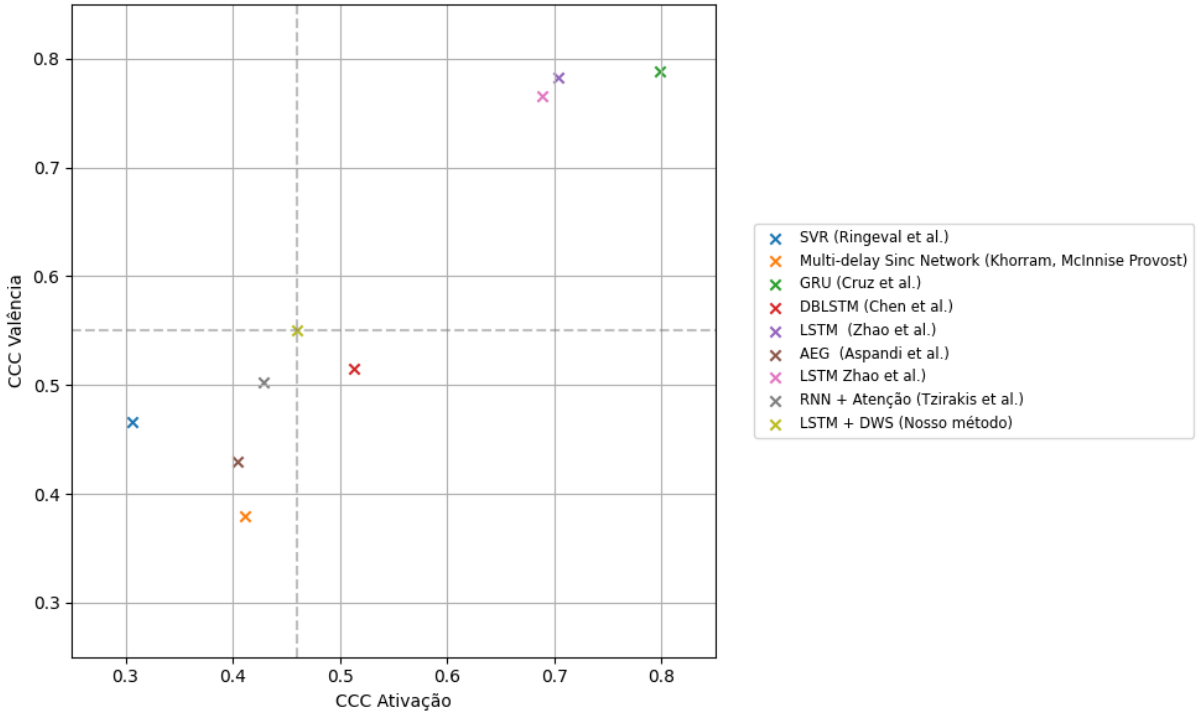
Figura 34 – Comparação entre diferentes métodos em termos de CCC para ativação e valência, base de dados RECOLA. As linhas horizontais e verticais destacam os valores de referência para o método proposto, permitindo uma análise clara de seu desempenho em relação aos demais métodos.



Fonte: Autoria Própria

A hipótese principal (H1) do presente trabalho sugere que a utilização de técnicas de seleção dinâmica de modalidade e representação de dados melhora significativamente o desempenho de sistemas multimodais em comparação com abordagens de fusão estática. Nos experimentos realizados nas bases RECOLA e SEWA, os métodos DW e DWS se destacaram, apresentando desempenho superior ao do melhor regressor individual e da média simples. No entanto, o empate entre os métodos DW e DWS com a média simples sugere que, embora a seleção dinâmica melhore o desempenho em relação ao melhor

Figura 35 – Comparação entre diferentes métodos em termos de CCC para ativação e valência, base de dados SEWA. As linhas horizontais e verticais destacam os valores de referência para o método proposto, permitindo uma análise clara de seu desempenho em relação aos demais métodos.



Fonte: Autoria Própria

regressor individual, ela não supera abordagens de agregação simples, como a média, neste cenário específico. Isso indica que H1 é parcialmente válida.

Por fim, a hipótese secundária (H2) sugere que a seleção dinâmica de modalidade e representação de dados é uma estratégia eficaz para manter a precisão em cenários com modalidades ausentes. Nos experimentos realizados no contexto de MER, os métodos baseados em seleção dinâmica (DS, DW e DWS) apresentaram consistentemente os melhores desempenhos, ocupando as primeiras posições no ranking médio em todos os cenários de ausência de modalidades avaliados. Dessa forma, os resultados confirmam a validade de H2.

7 Conclusão

Este trabalho apresentou um método de seleção dinâmica de modalidades e representações, com foco em reconhecimento contínuo de emoções, cujos resultados experimentais, conduzidos nas bases RECOLA e SEWA, evidenciaram ganhos de desempenho em comparação a abordagens de fusão estática e mecanismos de atenção. Tais ganhos sustentam a Hipótese Principal (H1), pois a adoção da seleção dinâmica levou a melhorias no desempenho multimodal. Além disso, o método demonstrou robustez diante da ausência parcial ou total de modalidades, o que corrobora a Hipótese Secundária (H2), ao manter níveis elevados de precisão em condições adversas.

Além dos benefícios de adaptação e resiliência, há de se considerar o aumento de complexidade inerente às abordagens de seleção dinâmica. No entanto, a capacidade de operar com dados ausentes e de potencializar a complementaridade de diferentes canais justifica o custo adicional em aplicações que demandem elevada precisão e adaptabilidade, como interação humano-computador, monitoramento de saúde mental e sistemas educacionais.

Em termos de contribuições, destacam-se: (i) o framework inovador de seleção dinâmica estruturado na escolha progressiva de modalidade e representação de dados; (ii) a análise sistemática de seu comportamento em condições adversas (por exemplo, ausência total ou parcial de dados) e (iii) a comparação com métodos de fusão estática e mecanismos de atenção, evidenciando cenários em que a abordagem dinâmica supera as alternativas.

Apesar das contribuições significativas, este trabalho apresenta algumas limitações que merecem destaque. Primeiramente, o treinamento de múltiplos regressores, somado ao uso de k-NN para a busca baseada em similaridade, pode limitar a escalabilidade em aplicações que exijam processamento de grandes volumes de dados ou restrições de tempo de resposta. Para mitigar esse impacto, foi utilizada a biblioteca `cuML` da RAPIDS AI, que permite acelerar operações de aprendizado de máquina em GPU, utilizando estruturas compatíveis com CUDA (RASCHKA; PATTERSON; NOLET, 2020). O cálculo de distâncias foi totalmente paralelizado, o que garantiu desempenho adequado mesmo com conjuntos de validação de alta dimensionalidade. A paralelização do processo de inferência por meio de GPU contribuiu significativamente para reduzir os tempos de resposta da etapa de seleção, tornando a aplicação viável em contextos quase em tempo real.

Além disso, o método proposto foi validado exclusivamente no contexto de MER, utilizando as bases de dados RECOLA e SEWA, o que pode restringir sua generalização para outros problemas com características distintas. Outro aspecto relevante é a ausência de uma análise mais aprofundada em cenários onde os dados não estão completamente

indisponíveis, apenas apresentam ruídos ou inconsistências, que são situações comuns em aplicações práticas.

Como linhas futuras de investigação, propõe-se avaliar o método em domínios diversos, como diagnóstico médico multimodal e sistemas de recomendação, a fim de ampliar sua aplicabilidade e comprovar seu potencial de generalização. Outra linha de pesquisa interessante seria a integração de novas modalidades, ainda no contexto de MER, como dados fisiológicos ou contextuais, para enriquecer ainda mais o processo de reconhecimento de emoções. Além disso, a integração com arquiteturas emergentes, tais como modelos fundacionais e sistemas neurais adaptativos, sugere oportunidades de ganhos adicionais em robustez e desempenho, beneficiando aplicações em cenários mais dinâmicos e de maior complexidade.

Referências

ABDOLLAHI, Hojjat; MAHOOR, Mohammad H.; ZANDIE, Rohola; SIEWIERSKI, Jarid; QUALLS, Sara H. Artificial emotional intelligence in socially assistive robots for older adults: A pilot study. *IEEE Transactions on Affective Computing*, v. 14, n. 3, p. 2020–2032, 2023.

ALBADAWY, Ehab A.; KIM, Yelin. *Joint Discrete and Continuous Emotion Prediction Using Ensemble and End-to-End Approaches*. ACM, 2018. 366–375 p. Disponível em: <http://dx.doi.org/10.1145/3242969.3242972>.

ASLAM, Muhammad Haseeb; ZEESHAN, Osama; PEDERSOLI, Marco; KOERICH, Alessandro L; BACON, Simon; GRANGER, Eric. Privileged knowledge distillation for dimensional emotion recognition in the wild. In: *CVPRw 2023: IEEE/CVF CVPR, Vancouver, Canada*. [S.l.: s.n.], 2023.

ASPANDI, Decky; MALLOL-RAGOLTA, Adria; SCHULLER, Bjorn; BINEFA, Xavier. *Latent-Based Adversarial Neural Networks for Facial Affect Estimations*. IEEE, 2020. 606–610 p. Disponível em: <http://dx.doi.org/10.1109/FG47880.2020.00053>.

BALTRUSAITIS, Tadas; AHUJA, Chaitanya; MORENCY, Louis-Philippe. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Institute of Electrical and Electronics Engineers (IEEE), v. 41, n. 2, p. 423–443, fev 2019. Disponível em: <http://dx.doi.org/10.1109/TPAMI.2018.2798607>.

BRADY, Kevin; GWON, Youngjune; KHORRAMI, Pooya; GODOY, Elizabeth; CAMPBELL, William; DAGLI, Charlie; HUANG, Thomas S. *Multi-Modal Audio, Video and Physiological Sensor Learning for Continuous Emotion Prediction*. ACM, 2016. Disponível em: <http://dx.doi.org/10.1145/2988257.2988264>.

BRAUWERS, Gianni; FRASINCAR, Flavius. *A General Survey on Attention Mechanisms in Deep Learning*. Institute of Electrical and Electronics Engineers (IEEE), 2023. 3279–3298 p. Disponível em: <http://dx.doi.org/10.1109/TKDE.2021.3126456>.

BRITTO ALCEU S., Jr.; SABOURIN, Robert; OLIVEIRA, Luiz E.S. *Dynamic selection of classifiers—A comprehensive review*. Elsevier BV, 2014. 3665–3680 p. Disponível em: <http://dx.doi.org/10.1016/j.patcog.2014.05.003>.

CARDINAL, P.; DEHAK, N.; KOERICH, A. L.; ALAM, J.; BOUCHER, P. ETS system for AV+EC 2015 challenge. In: *ACM Multimedia Conference*. [S.l.: s.n.], 2015. p. 17–23.

CHAO, Linlin; TAO, Jianhua; YANG, Minghao; LI, Ya; WEN, Zhengqi. *Long Short Term Memory Recurrent Neural Network based Multimodal Dimensional Emotion Recognition*. ACM, 2015. 65–72 p. Disponível em: <http://dx.doi.org/10.1145/2808196.2811634>.

CHEN, Haifeng; DENG, Yifan; CHENG, Shiwen; WANG, Yixuan; JIANG, Dongmei; SAHLI, Hichem. *Efficient Spatial Temporal Convolutional Features for Audiovisual Continuous Affect Recognition*. ACM, 2019. 19–26 p. Disponível em: <http://dx.doi.org/10.1145/3347320.3357690>.

CHEN, Shizhe; JIN, Qin. Multi-modal conditional attention fusion for dimensional emotion prediction. arXiv, 2017. Disponível em: <https://arxiv.org/abs/1709.02251>.

CHENG, Cheng; FAN, Zhaoxin; FENG, Lin; JIA, Ziyu. A novel transformer autoencoder for multi-modal emotion recognition with incomplete data. *Neural Networks*, p. 106111, 2024. ISSN 0893-6080.

CRUZ, Christian Arzate; SECHAYK, Yotam; IGARASHI, Takeo; GOMEZ, Randy. Data augmentation for 3dmm-based arousal-valence prediction for hri. arXiv, 2024. Disponível em: <https://arxiv.org/abs/2410.00349>.

CRUZ, Rafael M.O.; SABOURIN, Robert; CAVALCANTI, George D.C. *Dynamic classifier selection: Recent advances and perspectives*. Elsevier BV, 2018. 195–216 p. Disponível em: <http://dx.doi.org/10.1016/j.inffus.2017.09.010>.

DEMsAR, Janez. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, JMLR.org, v. 7, p. 1–30, dez. 2006. ISSN 1532-4435.

DENG, Jia; DONG, Wei; SOCHER, Richard; LI, Li-Jia; LI, Kai; FEI-FEI, Li. *ImageNet: A large-scale hierarchical image database*. IEEE, 2009. Disponível em: <http://dx.doi.org/10.1109/CVPR.2009.5206848>.

D'MELLO, Sidney K.; KORY, Jacqueline. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys*, v. 47, n. 3, p. 1–36, 2015.

D'MELLO, Sidney K.; KORY, Jacqueline. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys*, Association for Computing Machinery (ACM), v. 47, n. 3, p. 1–36, abr 2015. Disponível em: <http://dx.doi.org/10.1145/2682899>.

EKMAN, Paul. An argument for basic emotions. *Cognition and Emotion*, v. 6, n. 3, p. 169–200, 1992.

EYBEN, Florian; WöLLMER, Martin; SCHULLER, Björn. Opensmile: the munich versatile and fast open-source audio feature extractor. In: *Proceedings of the 18th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2010. (MM '10), p. 1459–1462. ISBN 9781605589336. Disponível em: <https://doi.org/10.1145/1873951.1874246>.

FRIEDMAN, Milton. *The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance*. Informa UK Limited, 1937. 675–701 p. Disponível em: <http://dx.doi.org/10.1080/01621459.1937.10503522>.

GLADYS, A. Aruna; VETRISELVI, V. Survey on multimodal approaches to emotion recognition. *Neurocomputing*, v. 556, p. 126693, 2023.

HAN, Jing; ZHANG, Zixing; REN, Zhao; SCHULLER, Björn. Emobed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings. *arXiv*, arXiv, 2019. Disponível em: <https://arxiv.org/abs/1907.10428>.

HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing; SUN, Jian. Deep residual learning for image recognition. arXiv, 2015. Disponível em: <https://arxiv.org/abs/1512.03385>.

HE, Lang; JIANG, Dongmei; YANG, Le; PEI, Ercheng; WU, Peng; SAHLI, Hichem. *Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks*. ACM, 2015. 73–80 p. Disponível em: <http://dx.doi.org/10.1145/2808196.2811641>.

HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. *Long Short-Term Memory*. MIT Press, 1997. 1735–1780 p. Disponível em: <http://dx.doi.org/10.1162/neco.1997.9.8.1735>.

HUANG, Gao; LIU, Zhuang; MAATEN, Laurens van der; WEINBERGER, Kilian Q. Densely connected convolutional networks. arXiv, 2016. Disponível em: <https://arxiv.org/abs/1608.06993>.

KAZEMI, Vahid; SULLIVAN, Josephine. One millisecond face alignment with an ensemble of regression trees. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2014. p. 1867–1874.

KEREN, Gil; KIRSCHSTEIN, Tobias; MARCHI, Erik; RINGEVAL, Fabien; SCHULLER, Bjorn. *End-to-end learning for dimensional emotion recognition from physiological signals*. IEEE, 2017. 985–990 p. Disponível em: <http://dx.doi.org/10.1109/ICME.2017.8019533>.

KHORRAM, Soheil; MCINNIS, Melvin G; PROVOST, Emily Mower. *Jointly Aligning and Predicting Continuous Emotion Annotations*. Institute of Electrical and Electronics Engineers (IEEE), 2021. 1069–1083 p. Disponível em: <http://dx.doi.org/10.1109/taffc.2019.2917047>.

KOSSAIFI, Jean; WALECKI, Robert; PANAGAKIS, Yannis; SHEN, Jie; SCHMITT, Maximilian; RINGEVAL, Fabien; HAN, Jing; PANDIT, Vedhas; TOISOUL, Antoine; SCHULLER, Bjorn; STAR, Kam; HAJIYEV, Elnar; PANTIC, Maja. *SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild*. Institute of Electrical and Electronics Engineers (IEEE), 2021. 1022–1040 p. Disponível em: <http://dx.doi.org/10.1109/TPAMI.2019.2944808>.

LI, Jiayao; LI, Li; SUN, Ruizhi; YUAN, Gang; WANG, Shufan; SUN, Shulin. Mman-m2: Multiple multi-head attentions network based on encoder with missing modalities. *Pattern Recognition Letters*, v. 177, p. 110–120, 2024. ISSN 0167-8655.

LIAN, Hailun; LU, Cheng; LI, Sunan; ZHAO, Yan; TANG, Chuangao; ZONG, Yuan. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, v. 25, n. 10, 2023. ISSN 1099-4300. Disponível em: <https://www.mdpi.com/1099-4300/25/10/1440>.

LIAN, Hailun; LU, Cheng; LI, Sunan; ZHAO, Yan; TANG, Chuangao; ZONG, Yuan. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, v. 25, n. 10, 2023. ISSN 1099-4300.

LIN, Ronghao; HU, Haifeng. MissModal: Increasing Robustness to Missing Modality in Multimodal Sentiment Analysis. *Transactions of the Association for Computational Linguistics*, v. 11, p. 1686–1702, 12 2023. ISSN 2307-387X.

LIN, Wei-Cheng; GONCALVES, Lucas; BUSSO, Carlos. Enhancing resilience to missing data in audio-text emotion recognition with multi-scale chunk regularization. In: *25th Intl Conf on Multimodal Interaction*. [S.l.: s.n.], 2023. p. 207–215. ISBN 9798400700552.

LINDEMANN, Benjamin; MüLLER, Timo; VIETZ, Hannes; JAZDI, Nasser; WEYRICH, Michael. *A survey on long short-term memory networks for time series prediction*. Elsevier BV, 2021. 650–655 p. Disponível em: <http://dx.doi.org/10.1016/j.procir.2021.03.088>.

LIU, Shuai; GAO, Peng; LI, Yating; FU, Weina; DING, Weiping. Multi-modal fusion network with complementarity and importance for emotion recognition. *Inf. Sci.*, v. 619, p. 679–694, 2023.

LIU, Zhizhong; ZHOU, Bin; CHU, Dianhui; SUN, Yuhang; MENG, Lingqiang. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Inf. Fusion*, v. 101, p. 101973, 2024.

MCFEE, Brian; MCVICAR, Matt; FARONBI, Daniel; ROMAN, Iran; GOVER, Matan; BALKE, Stefan; SEYFARTH, Scott; MALEK, Ayoub; RAFFEL, Colin; LOSTANLEN, Vincent; NIEKIRK, Benjamin van; LEE, Dana; CWITKOWITZ, Frank; ZALKOW, Frank; NIETO, Oriol; ELLIS, Dan; MASON, Jack; LEE, Kyungyun; STEERS, Bea; HALVACHS, Emily; THOMé, Carl; ROBERT-STöTER, Fabian; BITTNER, Rachel; WEI, Ziyao; WEISS, Adam; BATTENBERG, Eric; CHOI, Keunwoo; YAMAMOTO, Ryuichi; CARR, CJ; METSAI, Alex; SULLIVAN, Stefan; FRIESCH, Pius; KRISHNAKUMAR, Asmitha; HIDAKA, Shunsuke; KOWALIK, Steve; KELLER, Fabian; MAZUR, Dan; CHABOT-LECLERC, Alexandre; HAWTHORNE, Curtis; RAMAPRASAD, Chandrashekhar; KEUM, Myungchul; GOMEZ, Juanita; MONROE, Will; MOROZOV, Viktor Andreevitch; ELIASI, Kian; NULLMIGHTYBOFO; BIBERSTEIN, Paul; SERGIN, N. Dorukhan; HENNEQUIN, Romain; NAKTINIS, Rimvydas; BEANTOWEL; KIM, Taewoon; ÅSEN, Jon Petter; LIM, Joon; MALINS, Alex; HEREñú, Darío; STRUIJK, Stef van der; NICKEL, Lorenz; WU, Jackie; WANG, Zhen; GATES, Tim; VOLLRATH, Matt; SARROFF, Andy; XIAO-MING; PORTER, Alastair; KRANZLER, Seth; VOODOOHOP; GANGI, Mattia Di; JINOZ, Helmi; GUERRERO, Connor; MAZHAR, Abduttayyeb; TODDRME2178; BARATZ, Zvi; KOSTIN, Anton; ZHUANG, Xinlu; LO, Cash TingHin; CAMPR, Pavel; SEMENIUC, Eric; BISWAL, Monsij; MOURA, Shayenne; BROSSIER, Paul; LEE, Hojin; PIMENTA, Waldir. *librosa/librosa: 0.10.2.post1*. Zenodo, 2024. Disponível em: <https://doi.org/10.5281/zenodo.11192913>.

MENON, Luciana Trinkaus; NEDUZIAK, Luiz Carlos Ribeiro; BARDDAL, Jean Paul; KOERICH, Alessandro Lameiras; BRITTO ALCEU, Jr. de Souza. *Dynamic Modality and View Selection for Emotion Recognition: An Experimental Study on Missing Modality Evaluation*. Springer Nature Switzerland, 2025. 151–166 p. Disponível em: <http://dx.doi.org/10.1007/978-3-031-88217-3_11>.

MIGUEL, Fabiano Koich. *Psicologia das emoções: uma proposta integrativa para compreender a expressão emocional*. FapUNIFESP (SciELO), 2015. 153–162 p. Disponível em: <http://dx.doi.org/10.1590/1413-82712015200114>.

MOURA, Thiago J.M.; CAVALCANTI, George D.C.; OLIVEIRA, Luiz S. *MINE: A framework for dynamic regressor selection*. Elsevier BV, 2021. 157–179 p. Disponível em: <http://dx.doi.org/10.1016/j.ins.2020.07.056>.

MOURA, Thiago J. M.; CAVALCANTI, George D. C.; OLIVEIRA, Luiz S. *Evaluating Competence Measures for Dynamic Regressor Selection*. IEEE, 2019. Disponível em: <http://dx.doi.org/10.1109/IJCNN.2019.8851835>.

NAIK, Dayakar; KIRAN, Ravi. A novel sensitivity-based method for feature selection. *Journal of Big Data*, v. 8, 10 2021.

NEMENYI, P. *Distribution-free Multiple Comparisons*. Princeton University, 1963. Disponível em: <https://books.google.com.br/books?id=nhDMtgAACAAJ>.

NGUYEN, Cam-Van Thi; NGUYEN, Huy-Sang; LE, Duc-Trong; HA, Quang-Thuy. Multimodal learning with incompleteness towards multimodal sentiment analysis and emotion recognition task. In: BINH, Huynh Thi Thanh; HOANG, Van-Thuc; NGUYEN, Le-Minh; LE, SyVinh; VU, Thi-Dao; PHAM, Duy Trung (Ed.). *15th Intl Conf on Knowledge Syst Eng, Hanoi, Vietnam*. [S.l.: s.n.], 2023. p. 1–6.

NUNES, C.M.; BRITTO, Ad.S.; KAESTNER, C.A.A.; SABOURIN, R. An optimized hill climbing algorithm for feature subset selection: evaluation on handwritten character recognition. In: *9th Intl Workshop on Frontiers in Handwriting Recognition*. [S.l.: s.n.], 2004. p. 365–370.

O'MALLEY, Tom; BURSZTEIN, Elie; LONG, James; CHOLLET, François; JIN, Haifeng; INVERNIZZI, Luca et al. *KerasTuner*. 2019. <https://github.com/keras-team/keras-tuner>.

ORTEGA, J. D. S.; CARDINAL, P.; KOERICH, A. L. Emotion recognition using fusion of audio and video features. In: *IEEE Intl Conf. Syst., Man, Cybern.* [S.l.: s.n.], 2019. p. 3827–3832.

ORTEGA, J D S; SENOUSSAOUI, M; GRANGER, E; PEDERSOLI, M; CARDINAL, P; KOERICH, A L. Multimodal fusion with deep neural networks for audio-video emotion recognition. *arXiv preprint 1907.03196*, 2018.

PLUTCHIK, Robert. *A psychoevolutionary theory of emotions*. SAGE Publications, 1982. 529–553 p. Disponível em: <http://dx.doi.org/10.1177/053901882021004003>.

PRAVEEN, R. Gnana; CARDINAL, Patrick; GRANGER, Eric. Audio–visual fusion for emotion recognition in the valence–arousal space using joint cross-attention. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, v. 5, n. 3, p. 360–373, 2023.

PRAVEEN, R Gnana; MELO, Wheidima Carneiro de; ULLAH, Nasib; ASLAM, Haseeb; ZEESHAN, Osama; DENORME, Théo; PEDERSOLI, Marco; KOERICH, Alessandro L.; BACON, Simon; CARDINAL, Patrick; GRANGER, Eric. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In: *2022 IEEE/CVF CVPRw*. [S.l.: s.n.], 2022. p. 2485–2494.

RASCHKA, Sebastian; PATTERSON, Joshua; NOLET, Corey. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *arXiv preprint arXiv:2002.04803*, 2020.

RINGEVAL, Fabien; SCHULLER, Bjoern; VALSTAR, Michel; COWIE, Roddy; PANTIC, Maja. *AVEC 2015*. ACM, 2015. Disponível em: <http://dx.doi.org/10.1145/2733373.2806408>.

RINGEVAL, Fabien; SCHULLER, Björn; VALSTAR, Michel; GRATCH, Jonathan; COWIE, Roddy; SCHERER, Stefan; MOZGAI, Sharon; CUMMINS, Nicholas; SCHMITT, Maximilian; PANTIC, Maja. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In: *Proceedings of the 7th Annual Workshop on*

Audio/Visual Emotion Challenge. New York, NY, USA: Association for Computing Machinery, 2017. (AVEC '17), p. 3–9. ISBN 9781450355025. Disponível em: <https://doi.org/10.1145/3133944.3133953>.

RINGEVAL, Fabien; SCHULLER, Björn; VALSTAR, Michel; COWIE, Roddy; KAYA, Heysem; SCHMITT, Maximilian; AMIRIPARIAN, Shahin; CUMMINS, Nicholas; LALANNE, Denis; MICHAUD, Adrien; CIFTçI, Elvan; GüLEç, Hüseyin; SALAH, Albert Ali; PANTIC, Maja. *AVEC 2018 Workshop and Challenge*. ACM, 2018. Disponível em: <http://dx.doi.org/10.1145/3266302.3266316>.

RINGEVAL, Fabien; SCHULLER, Björn; VALSTAR, Michel; CUMMINS, Nicholas; COWIE, Roddy; PANTIC, Maja. Avec'19: Audio/visual emotion challenge and workshop. In: *Proceedings of the 27th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2019. (MM '19), p. 2718–2719. ISBN 9781450368896. Disponível em: <https://doi.org/10.1145/3343031.3350550>.

RINGEVAL, Fabien; SONDEREGGER, Andreas; SAUER, Juergen; LALANNE, Denis. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In: IEEE. *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. [S.l.], 2013. p. 1–8.

RUSSELL, James A. A circumplex model of affect. *Journal of personality and social psychology*, American Psychological Association, v. 39, n. 6, p. 1161, 1980.

SCHUSTER, M.; PALIWAL, K.K. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, v. 45, n. 11, p. 2673–2681, 1997.

SILVA-FILARDER, Matthieu Da; ANCORA, Andrea; FILIPPONE, Maurizio; MICHIARDI, Pietro. Multimodal variational autoencoders for sensor fusion and cross generation. In: WANI, M. Arif; SETHI, Ishwar K.; SHI, Weisong; QU, Guangzhi; RAICU, Daniela Stan; JIN, Ruoming (Ed.). *20th IEEE Intl Conf on Machine Learning and Applications, Pasadena, CA, USA*. [S.l.: s.n.], 2021. p. 1069–1076.

SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. arXiv, 2014. Disponível em: <https://arxiv.org/abs/1409.1556>.

SOMANDEPALLI, Krishna; GUPTA, Rahul; NASIR, Md; BOOTH, Brandon M.; LEE, Sungbok; NARAYANAN, Shrikanth S. *Online Affect Tracking with Multimodal Kalman Filters*. ACM, 2016. 59–66 p. Disponível em: <http://dx.doi.org/10.1145/2988257.2988259>.

SUN, Bo; CAO, Siming; LI, Liandong; HE, Jun; YU, Lejun. *Exploring Multimodal Visual Features for Continuous Affect Recognition*. ACM, 2016. 83–88 p. Disponível em: <http://dx.doi.org/10.1145/2988257.2988270>.

SUTSKEVER, Ilya; VINYALS, Oriol; LE, Quoc V. Sequence to sequence learning with neural networks. arXiv, 2014. Disponível em: <https://arxiv.org/abs/1409.3215>.

TZIRAKIS, Panagiotis; NGUYEN, Anh; ZAFEIRIOU, Stefanos; SCHULLER, Bjorn W. *Speech Emotion Recognition Using Semantic Information*. IEEE, 2021. 6279–6283 p. Disponível em: <http://dx.doi.org/10.1109/ICASSP39728.2021.9414866>.

TZIRAKIS, Panagiotis; TRIGEORGIS, George; NICOLAOU, Mihalis A.; SCHULLER, Bjorn W.; ZAFEIRIOU, Stefanos. *End-to-End Multimodal Emotion Recognition Using Deep Neural Networks*. Institute of Electrical and Electronics Engineers (IEEE), 2017. 1301–1309 p. Disponível em: <http://dx.doi.org/10.1109/JSTSP.2017.2764438>.

UDAHEMUKA, Gustave; DJOUANI, Karim; KURIEN, Anish M. *Multimodal Emotion Recognition Using Visual, Vocal and Physiological Signals: A Review*. MDPI AG, 2024. 8071 p. Disponível em: <http://dx.doi.org/10.3390/app14178071>.

VALSTAR, Michel; GRATCH, Jonathan; SCHULLER, Björn; RINGEVAL, Fabien; LALANNE, Denis; TORRES, Mercedes Torres; SCHERER, Stefan; STRATOU, Giota; COWIE, Roddy; PANTIC, Maja. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. New York, NY, USA: Association for Computing Machinery, 2016. (AVEC '16), p. 3–10. ISBN 9781450345163. Disponível em: <https://doi.org/10.1145/2988257.2988258>.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N; KAISER, Ł ukasz; POLOSUKHIN, Illia. Attention is all you need. In: GUYON, I.; LUXBURG, U. Von; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.

VAZQUEZ-RODRIGUEZ, Juan; LEFEBVRE, Grégoire; CUMIN, Julien; CROWLEY, James L. Accommodating missing modalities in time-continuous multimodal emotion recognition. *CoRR*, abs/2311.10119, 2023.

WEBER, Raphaël; BARRIELLE, Vincent; SOLADIé, Catherine; SéGUIER, Renaud. *High-Level Geometry-based Features of Video Modality for Emotion Prediction*. ACM, 2016. 51–58 p. Disponível em: <http://dx.doi.org/10.1145/2988257.2988262>.

YU, Yong; SI, Xiaosheng; HU, Changhua; ZHANG, Jianxun. *A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures*. MIT Press, 2019. 1235–1270 p. Disponível em: <http://dx.doi.org/10.1162/neco_a_01199>.

ZAMBELI, João Gabriel Antunes; LIRA, Antonio Alexandre de Medeiros; CASSOL, Mauriceia. *Reconhecimento de emoções pela voz e expressão facial por estudantes de medicina*. FapUNIFESP (SciELO), 2024. Disponível em: <http://dx.doi.org/10.1590/2317-6431-2023-2889pt>.

ZENG, Ailing; CHEN, Muxi; ZHANG, Lei; XU, Qiang. *Are Transformers Effective for Time Series Forecasting?* [S.l.]: Association for the Advancement of Artificial Intelligence (AAAI), 2023. 11121–11128 p.

ZHANG, Yuan; TAO, Xiaomei; AI, Hanxu; CHEN, Tao; GAN, Yanling. *Multimodal Emotion Recognition by Fusing Video Semantic in MOOC Learning Scenarios*. 2024. Disponível em: <https://arxiv.org/abs/2404.07484>.

ZHANG, Zixing; HAN, Jing; COUTINHO, Eduardo; SCHULLER, Bjorn. *Dynamic Difficulty Awareness Training for Continuous Emotion Prediction*. Institute of Electrical and Electronics Engineers (IEEE), 2019. 1289–1301 p. Disponível em: <http://dx.doi.org/10.1109/TMM.2018.2871949>.

ZHAO, Jinming; LI, Ruichen; CHEN, Shizhe; JIN, Qin. *Multi-modal Multi-cultural Dimensional Continues Emotion Recognition in Dyadic Interactions*. ACM, 2018. 65–72 p. Disponível em: <http://dx.doi.org/10.1145/3266302.3266313>.

ZHAO, Jinming; LI, Ruichen; LIANG, Jingjun; CHEN, Shizhe; JIN, Qin. *Adversarial Domain Adaption for Multi-Cultural Dimensional Emotion Recognition in Dyadic Interactions*. ACM, 2019. 37–45 p. Disponível em: <http://dx.doi.org/10.1145/3347320.3357692>.

ZHU, Yizhe; SUN, Xin; ZHOU, Xi. Exploiting multi-modal fusion for robust face representation learning with missing modality. In: ILIADIS, Lazaros; PAPALEONIDAS, Antonios; ANGELOV, Plamen; JAYNE, Chrisina (Ed.). *Artificial Neural Networks and Machine Learning – ICANN 2023*. [S.l.: s.n.], 2023. p. 283–294. ISBN 978-3-031-44210-0.