INSTITUTO LATINO-AMERICANO DE CIÊNCIAS DA VIDA E DA NATUREZA (ILACVN)

BIOTECNOLOGIA

VOZES DO ESQUECIMENTO: REVELANDO O ALZHEIMER COM A FALA E APRENDIZADO DE MÁQUINA

JULIO CESAR RIVEROS CARDUS

Foz do Iguaçu
2024

INSTITUTO LATINO-AMERICANO DE CIÊNCIAS DA VIDA E DA NATUREZA (ILACVN)

BIOTECNOLOGIA

VOZES DO ESQUECIMENTO: REVELANDO O ALZHEIMER COM A FALA E APRENDIZADO DE MÁQUINA

JULIO CESAR RIVEROS CARDUS

Trabalho de Conclusão de Curso apresentado ao Instituto Latino-Americano de Ciências da Vida e da Natureza da Universidade Federal da Integração Latino-Americana, como requisito parcial à obtenção do título de Bacharel em Biotecnologia.

Orientador: Prof. Dr. Willian Zalewski

Foz do Iguaçu
2024

JULIO CESAR RIVEROS CARDUS

VOZES DO ESQUECIMENTO: REVELANDO O ALZHEIMER COM A FALA E APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso apresentado ao Instituto Latino-Americano de Ciências da Vida e da Natureza da Universidade Federal da Integração Latino-Americana, como requisito parcial à obtenção do título de Bacharel em Biotecnologia.

BANCA EXAMINADORA

_____

Orientador: Prof. Dr. Willian Zalewski
UNILA

_____

Prof. Dr. Joylan Nunes Maciel
UNILA

_____

Prof. Dr. André Gustavo Maletzke
UNIOESTE

Foz do Iguaçu, 16 de outubro de 2024.

A Julio y Alice, mis padres.

AGRADECIMENTOS

Isaac Newton afirmou: "Se enxerguei mais longe, foi porque me apoiei sobre ombros de gigantes." Ao me aproximar do fim da minha jornada de graduação, é inevitável recordar e expressar minha gratidão a cada um desses gigantes que estiveram ao meu lado, oferecendo seu tempo, conhecimento e apoio.

Em primeiro lugar, expresso minha profunda gratidão ao meu professor orientador, cuja orientação contínua foi fundamental para o desenvolvimento deste trabalho. Sua paciência, atenção e confiança ao longo da pesquisa foram inestimáveis. Desde a iniciação científica e monitoria até este momento, tive o privilégio de compartilhar momentos de aprendizado. Sua influência foi decisiva para meu crescimento profissional, ensinando-me não apenas boas práticas de pesquisa e metodologia, mas também a importância da perseverança e do rigor científico. Sua mentoria foi, sem dúvida, um dos pilares que sustentaram esta trajetória.

Aos membros da banca, agradeço pelas orientações e contribuições. Suas críticas construtivas e sugestões foram essenciais para o aprimoramento deste trabalho e para a ampliação da minha visão sobre o tema estudado.

Aos colegas do LACA, meus sinceros agradecimentos pela colaboração, amizade e risadas ao longo dos anos. O ambiente de camaradagem e apoio mútuo foi crucial para que eu enfrentasse os desafios deste percurso com leveza e dedicação. Foi uma honra compartilhar este espaço com vocês.

Dedico um agradecimento à Thaynara, minha melhor amiga desde os primeiros passos na faculdade até este momento da defesa do TCC. Em cada riso compartilhado, cada conversa profunda e cada silêncio, construiu-se uma memória afetiva que transcende o tempo e enriquece não apenas minha jornada acadêmica, mas minha própria existência. Que estas palavras sejam um eterno tributo à sua generosidade e ao amor que fortaleceu minha trajetória, fazendo com que cada leitura desperte recordações dos momentos que vivemos juntos na graduação.

Por fim, agradeço à minha família, cujo suporte incondicional constituíram a base emocional que sustentou toda esta jornada. Cada um de vocês, de maneiras únicas, contribuiu para que eu chegasse até aqui, sou profundamente grato por isso.

A ignorância é uma bênção. Mas aqui estamos, em busca do oposto.
Inspirado por um café às 3 da manhã

RESUMO

A doença de Alzheimer (DA) é a forma mais comum de demência, um desafio global de saúde pública, sem cura definitiva. A detecção precoce da DA é crítica para intervenções oportunas, mas os métodos de diagnóstico atuais são invasivos e caros. Este estudo explora a fala espontânea como um potencial biomarcador não invasivo para detecção precoce da DA usando técnicas de aprendizado de máquina. Dado que a fala é uma tarefa cognitivamente exigente impactada por processos neurodegenerativos, levantamos a hipótese de que a análise automatizada de características da fala pode identificar indivíduos em risco de demência. Esta pesquisa utiliza o conjunto de dados ADReSS 2020, compreendendo amostras de fala de indivíduos com DA e controles saudáveis, e emprega três blocos experimentais: análise de características acústicas, compressão de texto e *embeddings* de texto. No primeiro bloco, características acústicas (eGeMAPS, ComParE, emobase) são extraídas e combinadas com métodos de seleção de características (SUC, ERC, SFM, SSC) para treinar classificadores (LDA, DT, NN, SVM, RF). O segunda bloco aplica algoritmos de compressão de texto para calcular a Distância de Compressão Normalizada (NCD) entre transcrições de áudio, usando kNN para classificação e previsão de pontuação do Mini-Exame do Estado Mental (MMSE). A terceira estrutura aproveita *embeddings* de texto (NV-Embed-v2, STELLA, GTE) para detectar DA e prever pontuações do MMSE. Os modelos foram avaliados por meio de Validação Cruzada Leave-One-Out (LOSO) e validação de holdout. Todas as três abordagens demonstraram a capacidade de diferenciar entre DA e indivíduos saudáveis. Notavelmente, os modelos baseados em *embeddings* produziram as maiores acurácias de classificação, com GTE e LEM atingindo 86,54% e 85,90%, respectivamente. Para tarefas de regressão, o embedding STELLA combinada com Random Forest produziu a menor Raiz do Erro Quadrático Médio (RMSE) de 5,15. As análises estatísticas demonstraram a inexistência de diferenças significativas entre o uso de seletores de características e a aplicação de todos os atributos dos conjuntos de dados, e que há diferenças significativas entre os algoritmos de compressão e entre os *embeddings*, tanto para classificação quanto para regressão. Essas descobertas sugerem que a análise automatizada de fala espontânea usando aprendizado de máquina pode servir como uma ferramenta poderosa para o diagnóstico precoce de DA. Com precisões que ultrapassam 85%, esses métodos têm o potencial de permitir triagem e monitoramento não invasivos e econômicos da doença de Alzheimer em larga escala.

Palavras-chave: doença de Alzheimer; fala espontânea; aprendizado de máquina, diagnóstico precoce; *embeddings* de texto.

RESUMEN

La enfermedad de Alzheimer (EA) es la forma más común de demencia, un desafío de salud pública global sin cura definitiva. La detección temprana de la EA es fundamental para realizar intervenciones oportunas, pero los métodos de diagnóstico actuales son invasivos y costosos. Este estudio explora el habla espontánea como un posible biomarcador no invasivo para la detección temprana de la EA mediante técnicas de aprendizaje automático. Dado que el habla es una tarea cognitivamente exigente afectada por procesos neurodegenerativos, planteamos la hipótesis de que el análisis automatizado de las características del habla podría identificar a las personas con riesgo de demencia. Esta investigación utiliza el conjunto de datos ADReSS 2020, que comprende muestras de voz de personas con EA y controles sanos, y emplea tres bloques experimentales: análisis de características acústicas, compresión de texto e incrustaciones de texto. En el primer bloque, se extraen características acústicas (eGeMAPS, ComParE, emobase) y se combinan con métodos de selección de características (SUC, ERC, SFM, SSC) para entrenar clasificadores (LDA, DT, NN, SVM, RF). El segundo bloque aplica algoritmos de compresión de texto para calcular la distancia de compresión normalizada (NCD) entre transcripciones de audio, utilizando kNN para la clasificación y la predicción de la puntuación del miniexamen del estado mental (MMSE). El tercer marco aprovecha las incrustaciones de texto (NV-Embed-v2, STELLA, GTE) para detectar DA y predecir puntuaciones MMSE. Los modelos se evaluaron mediante validación cruzada Leave-One-Out (LOSO) y validación de exclusión. Los tres enfoques demostraron la capacidad de diferenciar entre personas con EA e individuos sanos. En particular, los modelos basados en incrustación produjeron las precisiones de clasificación más altas, con GTE y LEM alcanzando 86,54% y 85,90%, respectivamente. Para las tareas de regresión, la incrustación de STELLA combinada con Random Forest produjo el error cuadrático medio (RMSE) más bajo de 5,15. Los análisis estadísticos demostraron que no existen diferencias significativas entre el uso de selectores de características y la aplicación de todos los atributos de los conjuntos de datos, y que existen diferencias significativas entre los algoritmos de compresión y las incrustaciones, tanto para clasificación como para regresión. Estos hallazgos sugieren que el análisis automatizado del habla espontánea mediante el aprendizaje automático podría servir como una herramienta poderosa para el diagnóstico temprano de la EA. Con precisiones superiores al 85%, estos métodos tienen el potencial de permitir la detección y el seguimiento de la enfermedad de Alzheimer a gran escala de forma no invasiva y rentable.

Palabras clave: enfermedad de Alzheimer; habla espontánea; aprendizaje automático, diagnóstico temprano; *embeddings* de texto.

ABSTRACT

Alzheimer’s disease (AD) is the most common form of dementia, a global public health challenge with no definitive cure. Early detection of AD is critical for timely interventions, but current diagnostic methods are invasive and expensive. This study explores spontaneous speech as a potential noninvasive biomarker for early detection of AD using machine learning techniques. Given that speech is a cognitively demanding task impacted by neurodegenerative processes, we hypothesize that automated speech feature analysis can identify individuals at risk of dementia. This research uses the ADReSS 2020 dataset, comprising speech samples from individuals with AD and healthy controls, and employs three experimental blocks: acoustic feature analysis, text compression, and text embeddings. In the first block, acoustic features (eGeMAPS, ComParE, emobase) are extracted and combined with feature selection methods (SUC, ERC, SFM, SSC) to train classifiers (LDA, DT, NN, SVM, RF). The second block applies text compression algorithms to calculate the Normalized Compression Distance (NCD) between audio transcripts, using kNN for classification and Mini-Mental State Examination (MMSE) score prediction. The third framework leverages text embeddings (NV-Embed-v2, STELLA, GTE) to detect AD and predict MMSE scores. The models were evaluated using Leave-One-Out (LOSO) Cross-Validation and holdout validation. All three approaches demonstrated the ability to differentiate between AD and healthy individuals. Notably, the embedding-based models yielded the highest classification accuracies, with GTE and LEM reaching 86.54% and 85.90%, respectively. For regression tasks, STELLA embedding combined with Random Forest produced the lowest Root Mean Square Error (RMSE) of 5.15. Statistical analyses demonstrated that there were no significant differences between using feature selectors and applying all attributes to the datasets, and that there were significant differences between compression algorithms and embeddings for both classification and regression. These findings suggest that automated spontaneous speech analysis using machine learning could serve as a powerful tool for early diagnosis of AD. With accuracies exceeding 85%, these methods have the potential to enable noninvasive and cost-effective screening and monitoring of Alzheimer’s disease on a large scale.

Key words: Alzheimer's disease; spontaneous speech; machine learning, early diagnosis; text embeddings.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

1NN	k-Nearest Neighbors com k = 1
3NN	k-Nearest Neighbors com k = 3
5NN	k-Nearest Neighbors com k = 5
Aβ	Beta-amilóide
Acc	Acurácia
ADR	Representação Ativa de Dados
APOE4	Alelo ε4 da apolipoproteína E
ASR	Reconhecimento Automático de Fala
ComParE	Computational Paralinguistics Evaluation
CRNN	Rede Neural Convolucional Recorrente
DA	Doença de Alzheimer
DT	Árvores de Decisão
eGeMAPS	extended Geneva Minimalistic Acoustic Parameter Set
ERC	Eliminação Recursiva de Características
FN	Falsos Negativos
FP	Falsos Positivos
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
ILACVN	Instituto Latino-Americano de Ciências da Vida e da Natureza
kNN	k-Nearest Neighbors
LCR	Líquido Cefalorraquidiano
LDA	Análise Discriminante Linear
LOSO	Leave-One-Subject-Out
LR	Logistic Regression
MAE	Erro Médio Absoluto
ML	Aprendizado de Máquina
MLP	Perceptron Multicamadas
Não-DA	Controles Saudáveis
NB	Naive Bayes
NCD	Normalized Compression Distance
NLP	Processamento de Linguagem Natural
NN	Redes Neurais Artificiais

PCA	Análise de Componentes Principais
PET	Tomografia por Emissão de Pósitrons
Prec.	Precisão
p-tau	Proteína tau hiperfosforilada
r	Coeficiente de correlação
RF	Random Forests
Rec	Recall
RM	Ressonância Magnética
RMSE	Raiz Do Erro Quadrático Médio
SA	Seleção de Atributos
SFM	Select From Model
SP	São Paulo
SSC	Seleção Sequencial de Características
SUC	Seleção Univariada de Características
SVM	Support Vector Machines
t-tau	Proteína tau total
UNILA	Universidade Federal da Integração Latino-Americana
VN	Verdadeiros Negativos
VP	Verdadeiros Positivos

1 INTRODUÇÃO

A demência é uma síndrome multifatorial caracterizada pelo declínio progressivo das funções cognitivas, impactando de forma significativa a memória, o raciocínio e a capacidade de realizar atividades cotidianas. Esta condição neurológica é considerada um dos maiores desafios de saúde pública do século XXI, com implicações profundas para os sistemas de saúde, as economias e as sociedades em geral (Verma, 2024).

A prevalência global de demência tem aumentado de forma alarmante nas últimas décadas, especialmente em decorrência do envelhecimento da população. Estima-se que, em 2024, cerca de 57,4 milhões de indivíduos vivam com demência em todo o mundo. Projeções indicam que esse número poderá crescer exponencialmente, alcançando 152,8 milhões até 2050 (Nichols et al., 2022). Este aumento não apenas afetará a qualidade de vida de milhões de pessoas e suas famílias, mas também acarretará um pesado ônus econômico para os sistemas de saúde globalmente. O Relatório Mundial de Alzheimer de 2024 (Alzheimer's Disease International, 2024) revela que quase 80% do público geral e 65% dos profissionais de saúde ainda acreditam que a demência é uma parte normal do envelhecimento. Essa crença pode atrasar o diagnóstico e o acesso a tratamentos adequados, reforçando a urgência de aumentar a conscientização sobre a demência como uma condição patológica.

Os custos associados à demência são substanciais. Em 2023, os custos globais estimados atingiram aproximadamente 1,3 trilhões de dólares por ano, com previsões indicando que esse valor poderá dobrar até 2050, ultrapassando 2 trilhões de dólares anuais (Kadhim et al., 2023). No Brasil, a situação é igualmente preocupante, com estimativas sugerindo que entre 1,5 a 1,8 milhões de pessoas vivem com demência, e uma incidência anual de novos casos variando entre 55.000 a 77.000 (Melo et al., 2020 e Bertola et al., 2023). Alarmantemente, cerca de 77% dos casos permanecem sem diagnóstico, enfatizando a necessidade de estratégias aprimoradas para detecção e manejo (Bertola et al., 2023). O custo anual por paciente com demência no Brasil foi estimado em $16.548,24 em 2018 (Ferretti et al., 2018), representando um impacto financeiro significativo tanto para as famílias quanto para o sistema de saúde.

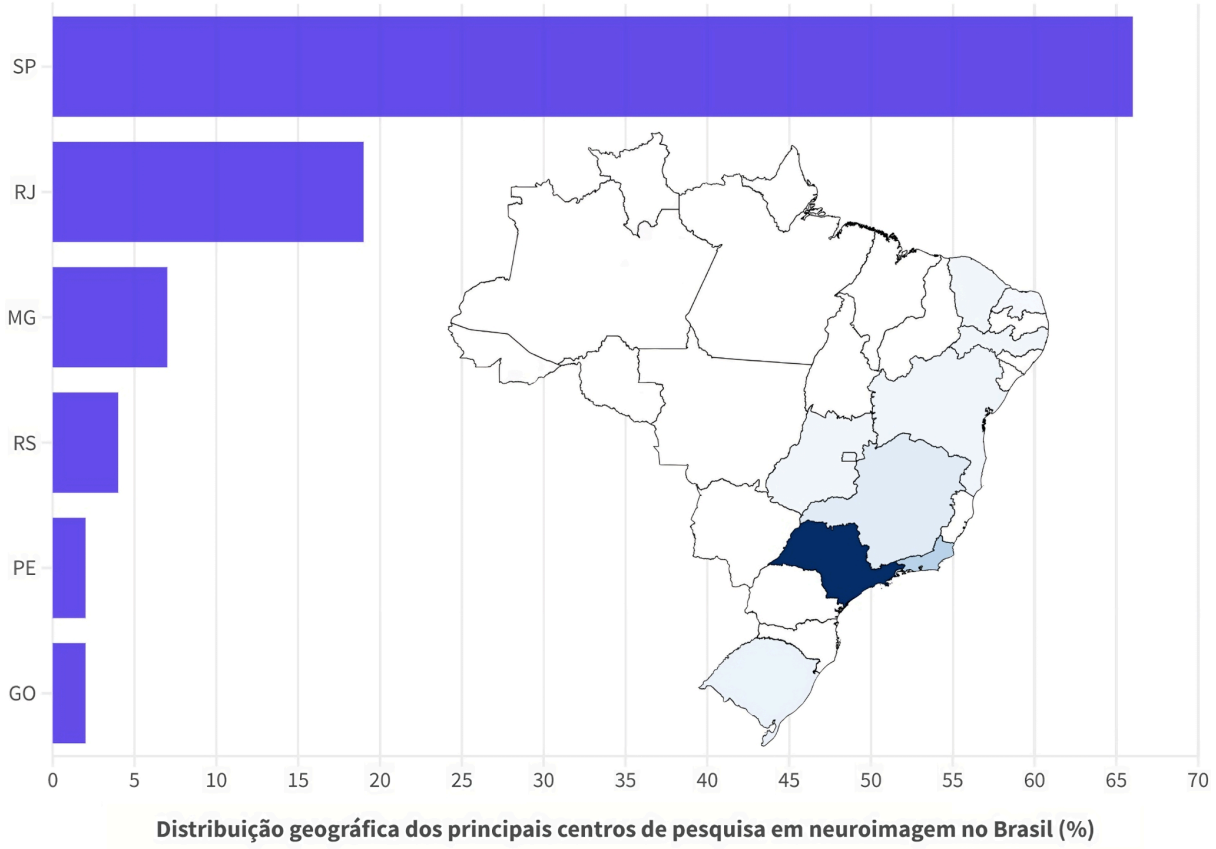
A detecção precoce da doença de Alzheimer é crucial para intervenções que possam retardar a progressão da doença e melhorar a qualidade de vida dos pacientes. Entretanto, o diagnóstico em estágios iniciais continua sendo um desafio. Métodos tradicionais, como exames de neuroimagem e avaliações cognitivas, podem ser dispendiosos, invasivos ou de difícil acesso em diversas regiões do Brasil e do Mundo.

Os métodos de neuroimagem desempenham um papel fundamental no diagnóstico e monitoramento da doença de Alzheimer, permitindo a visualização in vivo das alterações estruturais e funcionais do cérebro. Técnicas como Ressonância Magnética e Tomografia por Emissão de Pósitrons possibilitam a identificação de biomarcadores específicos, como atrofia hipocampal, alterações na substância branca e depósitos de proteína beta-amiloide, que são característicos da progressão da doença. Além disso, estas técnicas fornecem dados quantitativos que podem ser utilizados em estudos longitudinais e para avaliação da eficácia de tratamentos.

Na Figura 1 observa-se a distribuição dos principais centros de pesquisa em neuroimagem no Brasil, destacando a concentração dessas unidades na região sudeste, particularmente em São Paulo (SP). Esta visualização é útil para compreender a distribuição geográfica dos recursos de pesquisa em neuroimagem no país, o que pode ter implicações significativas para o diagnóstico e estudo da demência em diferentes regiões.

Nesse contexto, as análises de fala têm emergido como uma abordagem promissora. Métodos que utilizam padrões linguísticos podem oferecer uma alternativa não invasiva e acessível para identificar mudanças cognitivas associadas ao Alzheimer. A fala, como uma função cognitiva, envolve múltiplas regiões cerebrais e pode refletir alterações sutis antes que outros sintomas se tornem evidentes. Essas modificações podem indicar neuropatologias subjacentes nos sistemas motores e cognitivos (Ivanova, Martínez-Nicolás E García Meilán, 2023).

Figura 1 – Distribuição dos principais centros de pesquisa em neuroimagem no Brasil. O mapa destaca a concentração de unidades de pesquisa na região sudeste, particularmente em SP.



Fonte: Instituto Brasileiro de Geografia e Estatística (IBGE) e Rizzi, Aventurato e Balthazar, 2021.

Avanços recentes em tecnologias de processamento de linguagem natural e aprendizado de máquina possibilitaram o desenvolvimento de ferramentas automatizadas para análise de fala, as quais conseguem identificar padrões linguísticos e acústicos característicos da doença, mesmo em estágios iniciais. A capacidade de realizar triagens em larga escala utilizando gravações de fala abre novas perspectivas para a detecção precoce e o monitoramento da progressão da doença.

Nesse contexto, o objetivo deste estudo consiste na exploração e avaliação de diferentes métodos para a detecção automatizada da doença de Alzheimer, tais como: compressão de texto, seleção de características acústicas e *embeddings*.

Este trabalho está estruturado do seguinte modo:

Capítulo 2 - Fundamentação Teórica: Revisão sobre demência, Inteligência Artificial, Aprendizado de Máquina e análise de trabalhos que usam voz e processamento de linguagem para prever Alzheimer.

Capítulo 3 - Conjunto de Dados: Descreve o conjunto de dados utilizado, detalhando o desafio ADReSS 2020, as tarefas propostas e as limitações do conjunto de dados para a pesquisa em questão.

Capítulo 4 - Experimentos: É descrito o estudo experimental realizado neste estudo de predição de DA, o qual inclui três blocos: (1) descritores de áudio; (2) compressores textuais; e (3) *embeddings* NLP.

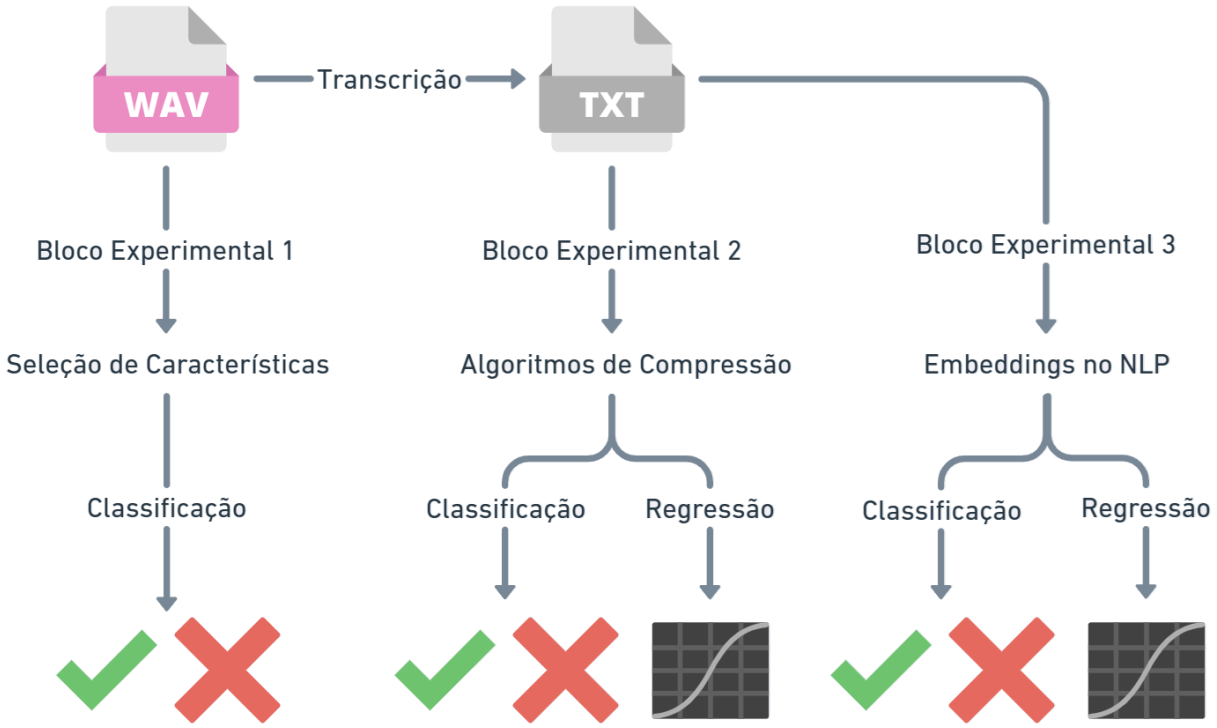
Capítulo 5 - Análise Comparativa dos Resultados: Apresenta uma síntese e análise estatística dos resultados obtidos nos três blocos experimentais, comparando e contrastando as diferentes abordagens utilizadas.

Capítulo 6 - Conclusão: Discute as implicações dos resultados para a detecção e predição da Doença de Alzheimer, aborda as limitações do estudo e propõe direções para pesquisas futuras nesta área.

A Figura 2 apresenta uma representação gráfica da estrutura metodológica adotada neste estudo, sintetizando os principais blocos experimentais e suas etapas. Inicialmente, são extraídas características acústicas diretamente dos áudios, permitindo a aplicação de métodos de seleção de atributos voltados para a classificação. Em seguida, os áudios são transcritos para texto, possibilitando a aplicação de duas abordagens textuais: (1) algoritmos de compressão e (2) embeddings baseados em Processamento de Linguagem Natural (NLP). O diagrama também diferencia as tarefas de classificação e regressão, evidenciando as estratégias avaliadas e os respectivos desempenhos em diferentes cenários experimentais.

.

Figura 2 – Representação Gráfica da Metodologia Proposta.



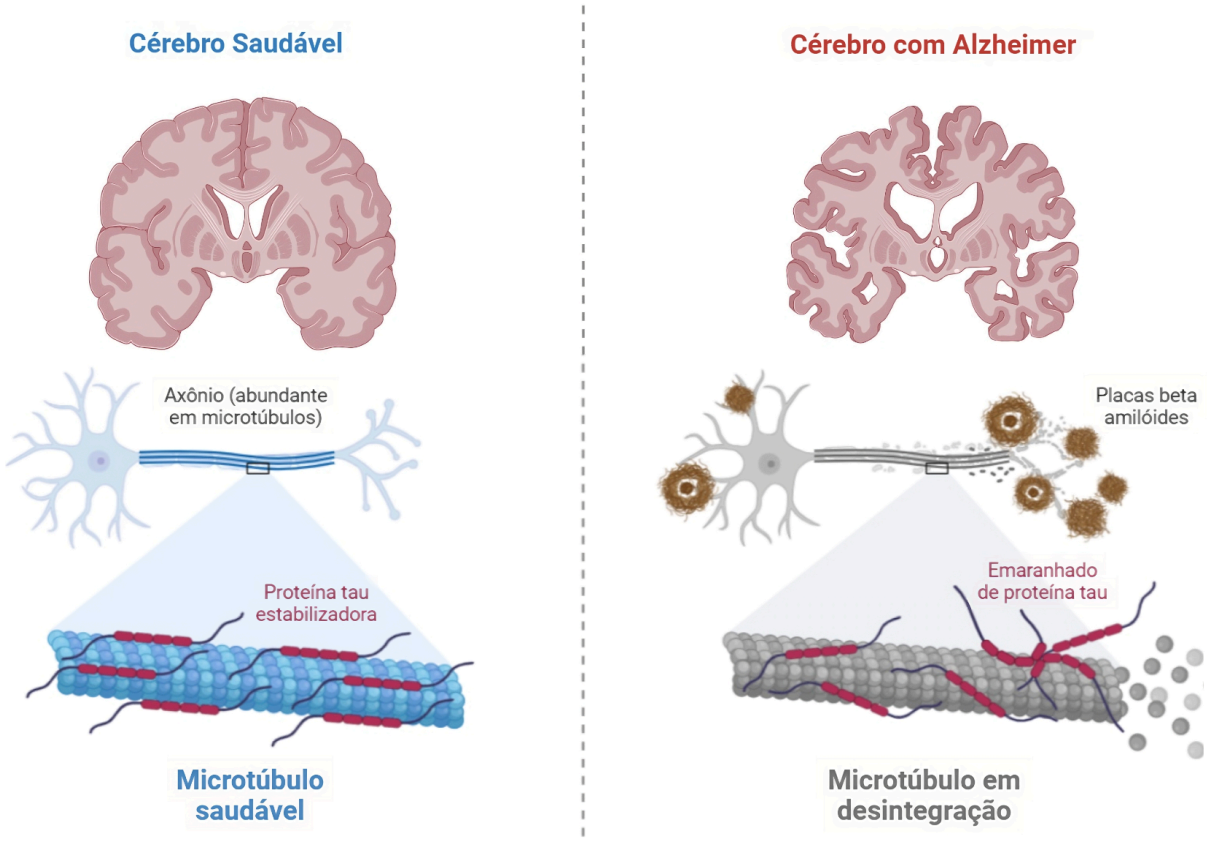
Fonte: Elaborado pelo autor (2024).

2 REVISÃO BIBLIOGRÁFICA

2.1 DEMÊNCIA E DOENÇA DE ALZHEIMER

A demência constitui um desafio crescente para a saúde global, com a Doença de Alzheimer (DA) representando sua etiologia predominante, responsável por até 60-70% dos casos de demência mundialmente (Nichols et al., 2019). Caracterizada por um declínio cognitivo progressivo, a demência afeta funções cognitivas superiores, incluindo memória, linguagem, raciocínio e capacidade de execução de atividades cotidianas (Hardy e Higgins, 1992). A DA, em particular, é definida pela presença de marcadores neuropatológicos específicos, notadamente placas amiloides e emaranhados neurofibrilares no parênquima cerebral (Montine et al., 2012). Estes últimos são compostos por proteína tau hiperfosforilada (p-tau), enquanto as placas amiloides resultam da agregação do peptídeo beta-amilóide (Aβ) (Jack Jr et al., 2018). Na Figura 3 observa-se a comparação entre um cérebro saudável e um afetado pela DA.

Figura 3 – Comparação entre um cérebro saudável (esquerda) e um cérebro afetado pela DA (direita). A imagem ilustra a atrofia cortical, a presença de placas amiloides e emaranhados neurofibrilares característicos da patologia de Alzheimer.



Fonte: Elaborado pelo autor com o auxílio da plataforma BioRender (2024).

É importante salientar que a demência constitui um termo guarda-chuva que abrange diversas condições neurodegenerativas, enquanto a DA se configura como uma entidade nosológica específica dentro deste espectro (Nichols et al., 2019). Esta distinção implica que, embora toda DA seja classificada como uma forma de demência, a recíproca não se verifica. Não obstante suas diferenças etiológicas e fisiopatológicas, ambas as condições convergem na manifestação de declínio cognitivo. A DA, em suas fases prodrômicas, frequentemente se apresenta com comprometimento mnêmico proeminente, ao passo que outras formas de demência podem exibir um espectro sintomatológico mais amplo, incluindo alterações comportamentais ou disfunções linguísticas. A elucidação da relação entre demência e DA reveste-se de importância capital para o desenvolvimento de ferramentas diagnósticas e abordagens terapêuticas mais eficazes. O diagnóstico precoce emerge como elemento crucial para intervenções que visem modificar o curso da doença e otimizar a qualidade de vida dos indivíduos afetados.

Na DA, as placas amiloides, constituídas predominantemente pelo peptídeo Aβ, depositam-se no compartimento extracelular do tecido neural. Em contrapartida, os emaranhados neurofibrilares, formados por agregados de p-tau, acumulam-se no compartimento intracelular dos neurônios. Essas alterações neuropatológicas, cuja gênese precede em aproximadamente 10 a 20 anos o início da sintomatologia clínica, desencadeiam uma cascata neurodegenerativa que resulta na perda progressiva de neurônios e sinapses, culminando no quadro de disfunção cognitiva característico da DA (Therriault et al., 2024).

As alterações fisiopatológicas observadas na DA são caracterizadas por sua complexidade e natureza multifatorial, englobando diversos processos interrelacionados:

- Neuroinflamação: A deposição de placas amiloides e a formação de emaranhados neurofibrilares induzem uma resposta inflamatória crônica no parênquima cerebral. Esta neuroinflamação, embora inicialmente possa desempenhar um papel neuroprotetor, subsequentemente contribui para a progressão da patologia ao longo do curso da doença (Webers, Heneka e Gleeson, 2020).

- Disfunção sináptica: A DA compromete significativamente a neurotransmissão, resultando em perda e disfunção sináptica. Esta alteração na plasticidade sináptica está intimamente associada aos déficits de memória e aprendizagem característicos da doença (Blumenfeld, 2024).
- Neurodegeneração: A progressão da DA é marcada por uma perda neuronal substancial em regiões cerebrais críticas para a cognição, notadamente o hipocampo e o neocórtex (Therriault, 2024). Esta neurodegeneração progressiva manifesta-se macroscópicamente como atrofia cerebral e clinicamente como deterioração das funções cognitivas.

Outro aspecto importante é a heterogeneidade da DA, que se manifesta através de diversas apresentações clínicas e padrões de progressão. Fatores genéticos, como a presença do alelo ε4 da apolipoproteína E (APOE4), exercem uma influência significativa sobre o risco de desenvolvimento e a taxa de progressão da doença (Blumenfeld, 2024). Ademais, a DA frequentemente coexiste com outras entidades neuropatológicas, um fenômeno que pode modular a apresentação clínica e representar um desafio diagnóstico adicional (Devi, 2023).

A compreensão integrada destes processos fisiopatológicos é fundamental para o desenvolvimento de estratégias diagnósticas mais precisas e abordagens terapêuticas potencialmente modificadoras da doença. Esta complexidade fisiopatológica se traduz em um espectro sintomatológico igualmente diverso e multifacetado, que se manifesta clinicamente através de alterações cognitivas e comportamentais características. Os sintomas da demência, particularmente na DA, podem ser categorizados em manifestações cognitivas e comportamentais, refletindo diretamente as alterações neuropatológicas subjacentes.

Manifestações Cognitivas:

- Comprometimento mnêmico: Este sintoma é próprio da DA, afetando inicialmente a memória de curto prazo. Os indivíduos acometidos frequentemente apresentam dificuldades na retenção e evocação de eventos recentes, conteúdo de interações sociais e informações recém-adquiridas (McKhann et al., 2011).

- Disfunção linguística: Com a progressão da DA, observa-se um declínio nas habilidades linguísticas, manifestando-se como anomia, dificuldades na compreensão de discurso complexo e na produção de fala coerente. A expressão verbal pode se tornar progressivamente desorganizada e ininteligível (Kaltsa et al., 2023).
- Desorientação: A DA pode comprometer a orientação espacial e temporal. Os indivíduos podem apresentar desorientação topográfica em ambientes previamente familiares, amnésia para datas significativas ou dificuldade na compreensão da passagem do tempo (Peters-Founshtein et al., 2024).
- Déficits atencionais e disfunção executiva: A DA pode afetar adversamente a capacidade de concentração, planejamento, tomada de decisões e execução de tarefas simultâneas. Os indivíduos podem demonstrar reduzida flexibilidade cognitiva e dificuldades na adaptação a novos contextos ou na alternância entre diferentes tarefas cognitivas (Gaubert, Borg e Chainay, 2022).

Manifestações Comportamentais:

- Alterações tímicas e de personalidade: A DA pode induzir modificações significativas no humor e na personalidade. Os indivíduos podem manifestar apatia, sintomas depressivos, ansiedade, irritabilidade, agitação psicomotora ou comportamento social inapropriado, conforme observado por Jamieson et al. (2016) e Sörensen e Conwell (2011).
- Fenômenos psicóticos: Alguns indivíduos com DA podem experienciar alucinações (visuais, auditivas ou táteis) e delírios. Evidências empíricas sugerem que estes sintomas, particularmente as alucinações visuais, estão associados a um declínio cognitivo e funcional mais acentuado em pacientes com DA (Pezzoli et al., 2022).
- Comportamentos repetitivos: Os indivíduos com DA podem exibir comportamentos perseverativos, manifestando-se como repetições verbais, estereotipias comportamentais complexas ou estereotipias motoras simples. Estes fenômenos apresentam diferentes correlatos cognitivos subjacentes, variando de acordo com o estágio da demência (Polin et al., 2023).

- Distúrbios no ciclo de sono: A DA pode ocasionar perturbações nos padrões de sono, resultando em insônia, hipersonia diurna ou comportamento agitado noturno (Morrone et al., 2023).

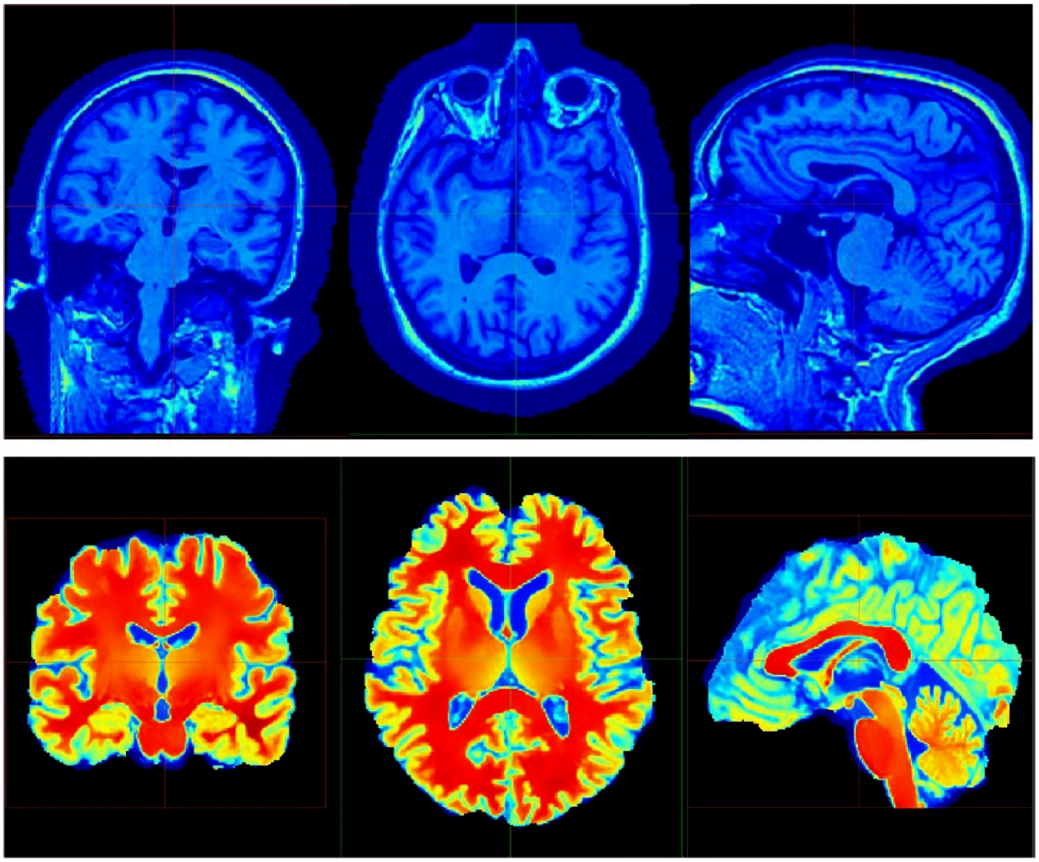
A compreensão abrangente deste espectro sintomatológico, que abarca desde alterações cognitivas sutis até manifestações comportamentais complexas, é fundamental para o diagnóstico precoce, manejo clínico adequado e desenvolvimento de intervenções terapêuticas direcionadas na DA. No entanto, a heterogeneidade e a natureza progressiva destes sintomas apresentam desafios significativos para o processo diagnóstico, especialmente nos estágios iniciais da doença.

Diante desta complexidade sintomatológica, o diagnóstico da demência, particularmente em seus estágios iniciais, requer uma abordagem multifacetada e criteriosa. A variabilidade na apresentação clínica, influenciada por fatores como reserva cognitiva, comorbidades e a própria heterogeneidade da DA, demanda uma estratégia diagnóstica que integre múltiplas modalidades de avaliação. Tradicionalmente, este processo baseia-se em uma abordagem multidisciplinar que engloba diversos componentes, cada um visando capturar aspectos específicos da sintomatologia e patologia subjacente:

- Avaliação Cognitiva: Instrumentos padronizados são amplamente utilizados para a identificação de déficits cognitivos. Embora essas avaliações forneçam uma medida global da função cognitiva, sua especificidade para a DA é limitada, e os resultados podem ser influenciados por variáveis confundidoras, como idade e nível educacional (Scheffels et al., 2023).
- Exames de Neuroimagem: Modalidades de neuroimagem estrutural, como a ressonância magnética (RM), desempenham um papel crucial na detecção de alterações cerebrais associadas à DA. A comparação entre imagens de RM antes e após o pré-processamento é mostrada na Figura 4, evidenciando a capacidade desta técnica de detectar atrofias em áreas cerebrais específicas, como o hipocampo, um biomarcador distintivo da DA. Adicionalmente, técnicas de imagem molecular, como a tomografia por emissão de pósitrons (PET) com ligantes específicos para amiloide e tau, permitem a visualização in vivo e quantificação da carga dessas

proteínas no parênquima cerebral (Dong et al., 2023). Contudo, a acessibilidade e os custos associados a estas modalidades avançadas de neuroimagem podem representar barreiras à sua implementação generalizada.

Figura 4 – Comparação de imagens de RM Cerebral Antes e Após o Pré-Processamento.



Fonte: Ahanger et al. (2024).

- Biomarcadores: A análise do Líquido Cefalorraquidiano (LCR) para quantificação de biomarcadores específicos, incluindo proteína tau total (t-tau), p-tau e peptídeos Aβ, tornou-se um elemento crítico no diagnóstico da DA. Alterações nos níveis destes biomarcadores no LCR refletem os processos neuropatológicos subjacentes à DA, com elevações de p-tau e t-tau indicativas de neurodegeneração, e redução de Aβ42 sugestiva de deposição amilóide cerebral (Abukuri, 2024). No entanto, a natureza invasiva da punção lombar necessária para obtenção do LCR limita a aplicabilidade generalizada desta abordagem diagnóstica.

2.2 APRENDIZADO DE MÁQUINA

A inteligência artificial (IA) é um campo interdisciplinar de estudo que visa o desenvolvimento de algoritmos computacionais capazes de replicar processos cognitivos humanos (Alpaydin, 2020). A IA integra conhecimentos de diversas disciplinas, incluindo probabilidade e estatística, ciência da computação, teoria da informação, neuropsicologia cognitiva, além das próprias técnicas de IA (El Naqa e Murphy, 2015).

O aprendizado de máquina (*machine learning*, ML) constitui um subcampo dinâmico da IA, centrado no desenvolvimento de algoritmos que aprimoram seu desempenho ao longo do tempo, a partir de dados. Diferente dos métodos tradicionais de programação, nos quais as instruções são rigidamente codificadas, os algoritmos de ML são projetados para ajustar automaticamente seus parâmetros internos por meio de um processo de treinamento (El Naqa e Murphy, 2015). Esse treinamento utiliza dados de entrada para otimizar a execução de uma tarefa específica. Este processo de adaptação contínua permite que os algoritmos não apenas aprendam a partir dos dados disponíveis, mas também generalizem para novos dados, até então não observados.

A importância do treinamento em ML reside na capacidade dos algoritmos de se aprimorarem progressivamente, semelhante ao aprendizado humano contínuo. Essa característica permite que os algoritmos realizem tarefas complexas de forma eficiente e adaptativa, superando as limitações do processamento convencional de números, ao aprender e adaptar-se a partir de exemplos repetidos.

No contexto do ML, os algoritmos são frequentemente classificados com base na natureza dos dados fornecidos durante o treinamento: supervisionado, não supervisionado e semi-supervisionado. A aprendizagem supervisionada envolve a estimativa de um mapeamento desconhecido entre entradas e saídas a partir de amostras rotuladas. Em contraste, a aprendizagem não supervisionada opera sobre dados não rotulados, realizando tarefas como agrupamento e estimativa de densidade de probabilidade. A aprendizagem semi-supervisionada combina aspectos de ambas as abordagens, utilizando um subconjunto rotulado para inferir informações sobre os dados não rotulados, sendo amplamente aplicada em sistemas de recuperação de texto e imagem.

No contexto da aprendizagem supervisionada, dois tipos principais de algoritmos se destacam: classificadores e regressores. Os classificadores são projetados para atribuir rótulos discretos a dados de entrada, categorizando-os em classes predefinidas. Por exemplo, na análise de imagens cerebrais, um classificador pode determinar a presença ou ausência de indicadores patológicos específicos. Em contrapartida, os regressores são utilizados para prever valores contínuos, como escores cognitivos ou taxas de progressão da doença, estabelecendo relações funcionais entre variáveis de entrada e saída.

A avaliação do desempenho destes algoritmos é realizada através de métricas específicas para cada tipo de tarefa. Para classificação, a acurácia mede a proporção total de predições corretas, enquanto a precisão indica a fração de identificações positivas que estão corretas, e o recall quantifica a capacidade do modelo de identificar todos os casos positivos relevantes. Estas métricas são calculadas a partir de uma matriz de confusão que contabiliza Verdadeiros Positivos (VP), Verdadeiros Negativos (VN), Falsos Positivos (FP) e Falsos Negativos (FN).

Para tarefas de regressão, são comumente utilizadas três métricas principais: o Erro Médio Absoluto (MAE), o Raiz do Erro Quadrático Médio (RMSE) e o coeficiente de correlação de Pearson (r). O MAE calcula a média das diferenças absolutas entre as predições e os valores reais, fornecendo uma medida diretamente interpretável do erro médio nas unidades originais da variável. O RMSE, por sua vez, calcula a raiz quadrada da média dos erros ao quadrado, penalizando mais fortemente erros maiores devido à operação de quadrado. Esta métrica é particularmente útil quando erros maiores são especialmente indesejáveis, pois atribui mais peso a eles no cálculo final.

Já o r é uma medida estatística que avalia a força e direção da relação linear entre os valores previstos e reais, variando de -1 a +1. Um valor de r próximo a +1 indica uma forte correlação positiva, onde aumentos nos valores previstos correspondem a aumentos proporcionais nos valores reais. Um r próximo a -1 sugere uma forte correlação negativa, enquanto valores próximos a 0 indicam ausência de correlação linear. Esta métrica é particularmente útil para avaliar se o modelo está capturando adequadamente as tendências nos dados, independentemente da escala absoluta dos erros.

As técnicas de ML têm sido aplicadas com sucesso em uma vasta gama de campos, incluindo reconhecimento de padrões, visão computacional, engenharia aeroespacial, finanças, entretenimento, biologia computacional e aplicações médicas. Recentemente, tem-se observado um crescente interesse na aplicação de técnicas de IA para automatizar e otimizar o processo diagnóstico da demência e da DA (Rehan et al., 2025). A implementação de algoritmos de ML no campo da neurologia clínica oferece um potencial promissor para o desenvolvimento de ferramentas diagnósticas e prognósticas inovadoras. Particularmente, a capacidade do ML de analisar extensos conjuntos de dados multidimensionais e identificar padrões complexos abre novas perspectivas para a predição precoce e acurada da demência, potencialmente integrando dados de múltiplas modalidades diagnósticas.

A integração de abordagens baseadas em ML com os métodos diagnósticos convencionais pode proporcionar uma estratégia mais robusta e precisa para a identificação precoce e caracterização da demência, facilitando intervenções terapêuticas mais oportunas e personalizadas.

2.3 TRABALHOS RELACIONADOS

Os estudos que investigam o uso de processamento de fala e linguagem natural para o diagnóstico automático da DA evidenciam que a DA compromete de maneira significativa as habilidades linguísticas e acústicas dos pacientes, mesmo em estágios iniciais. Esse impacto tem motivado pesquisadores a explorar como essas alterações na fala podem ser aproveitadas para a identificação e avaliação da progressão da doença.

Diversos trabalhos têm aplicado técnicas de processamento de fala e ML no diagnóstico da DA, com especial atenção à acurácia das classificações—isto é, à capacidade de distinguir pacientes com DA de indivíduos saudáveis—e à minimização do RMSE na predição das pontuações do Mini-Exame do Estado Mental (MMSE) com base em características extraídas da fala.

Entre os trabalhos mais relevantes, destacam-se aqueles que alcançaram resultados notáveis tanto em termos de acurácia quanto de RMSE. Por exemplo, Wang et al. (2022a) relataram uma acurácia de 91,7%, utilizando modelos de linguagem pré-treinados como BERT e RoBERTa, combinados com Support Vector Machines (SVM). De maneira similar, Sarawgi et al. (2020) atingiram uma acurácia de 88,0% e um RMSE de 4,60, empregando características de disfluência e prosódia, integradas a um Perceptron Multicamadas (MLP).

Adicionalmente, Martinc et al. (2021) e Wang et al. (2022b) sobressaem-se por obterem uma acurácia de 93,8%, ao combinarem técnicas como Bag-of-n-grams e fusão de modelos com Random Forests (RF) e SVM, respectivamente.

Importa destacar, contudo, que os resultados desses estudos variam significativamente em função dos conjuntos de dados, métodos de extração de características e modelos de ML utilizados. Essa variabilidade sublinha a necessidade contínua de pesquisas adicionais que validem e aprimorem a robustez desses sistemas, com vistas à sua futura aplicação em contextos clínicos.

No Quadro 1 são apresentados mais detalhes dos principais estudos, incluindo modalidades, características, classificadores, otimizações, resultados e conjuntos de dados.

Quadro 1 – Resumo dos Principais Estudos sobre Diagnóstico da DA Utilizando Processamento de Fala e ML.

Referência	Modalidade	Características	Classificador	Otimização	Acurácia	RMSE	*Dataset*
Wang et al., 2022a	Texto (ASR)	BERT, RoBERTa	SVM	Melhoria de ASR, Fusão de modelos	91,7%	-	Pitt
Sarawgi et al., 2020	Áudio, Texto	Disfluência, ComParE, Intervenções	MLP	Características de prosódia, Fusão de modelos	88,0%	4,60	
Ye et al., 2021	Texto (ASR)	BERT	SVM	Melhoria de ASR	88,0%	-	
Wang et al., 2022b	Texto (ASR)	BERT, RoBERTa	SVM	Fusão de modelos	93,8%	-	ADReSS 20
Martinc et al., 2021	Texto	ADR, Bag-of-n-gram	Agrupamento k-means, RF	Recursos ADR, Fusão de modelos	93,8%	-	
Yuan et al., 2020	Texto	Pausas	ERNIE	Recursos específicos de tarefas	89,6%	-	
Pan et al., 2021	Texto (ASR)	Hipótese ASR, Pontuação de confiança	BERT	Recursos ASR	84,5%	-	ADReSS 21
Syed et al., 2021	Texto (ASR)	BERT	LR	Fusão de modelos	84,5%	4,35	
Rohanian et al., 2021	Áudio, Texto (ASR)	Acústicas, GloVe, Disfluência, Pausas	LSTM com gating	Recursos de prosódia	84,0%	4,26	

Fonte: Elaborado pelo autor (2024).

3 CONJUNTO DE DADOS

O conjunto de dados utilizado neste estudo é oriundo do Desafio ADReSS, realizado como parte da conferência INTERSPEECH 2020. O desafio teve como principal objetivo a detecção automatizada da DA a partir de dados de fala espontânea, com o intuito de promover avanços na identificação precoce do declínio cognitivo (Luz et al., 2021).

3.1 VISÃO GERAL DO DESAFIO ADRESS 2020

Com a participação de 34 equipes de pesquisa de diferentes regiões do mundo, o Desafio ADReSS focou em duas tarefas primordiais:

- Classificação binária: Identificação de falas de indivíduos com DA e controles saudáveis (Não-DA).
- Determinação do MMSE: Predição das pontuações do MMSE, uma medida amplamente utilizada na avaliação do comprometimento cognitivo em pacientes com DA (Martinc & Pollak, 2020). O MMSE é um teste de rastreio que avalia diferentes domínios cognitivos, incluindo orientação temporal e espacial, memória imediata e de evocação, atenção, cálculo, linguagem e capacidade construtiva visual. A pontuação do teste varia de 0 a 30 pontos, onde pontuações mais baixas indicam maior comprometimento cognitivo. Em geral, pontuações abaixo de 24 são consideradas indicativas de possível comprometimento cognitivo, embora este ponto de corte possa variar de acordo com fatores como idade e escolaridade do paciente.

O desafio foi estruturado para minimizar vieses comumente observados em estudos de detecção de Alzheimer, como o balanceamento inadequado de dados e a repetição de amostras de fala de participantes sob diferentes condições experimentais. Esse esforço de padronização foi essencial para facilitar a comparabilidade entre as diversas abordagens desenvolvidas pelos competidores.

As gravações de fala passaram por um processo de pré-processamento com o objetivo de assegurar a qualidade e a consistência das amostras. As etapas principais incluíram a normalização do volume, que seria o controle das variações de volume causadas por diferentes condições de gravação, como a distância e posição do microfone e remoção de amostras repetidas e ajuste da qualidade acústica para manter a homogeneidade das gravações.

3.2 ESTRUTURA DO CONJUNTO DE DADOS

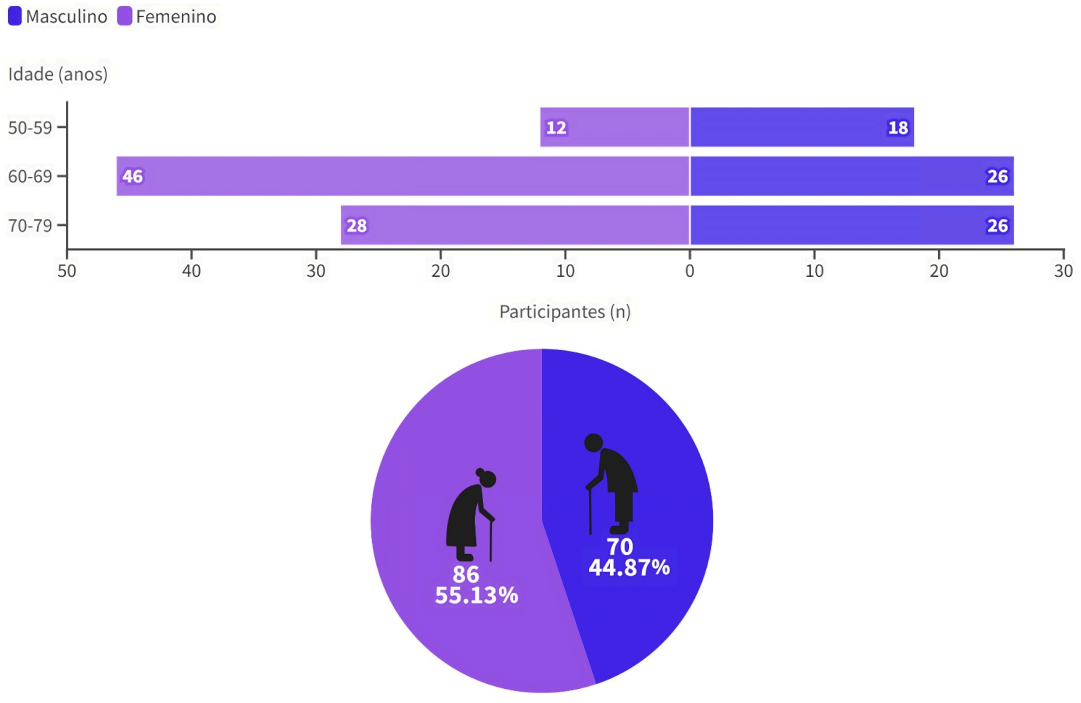
O conjunto de dados é composto por gravações de fala espontânea de 156 participantes, divididos em dois grupos iguais: 78 diagnosticados com DA e 78 controles (saudáveis). A distribuição demográfica dos participantes foi cuidadosamente balanceada em termos de gênero e idade, assegurando a robustez dos modelos desenvolvidos. O Quadro 2 e a Figura 5 permitem sumarizar a distribuição dos participantes por faixa etária, gênero e diagnóstico:

Quadro 2 – Distribuição de Pacientes DA e Não-DA por Faixa Etária e Gênero.

Faixa Etária (anos)	DA (Masculino/Feminino)	Não-DA (Masculino/Feminino)
[50 - 55)	2/0	2/0
[55 - 60)	7/6	7/6
[60 - 65)	4/9	4/9
[65 - 70)	9/14	9/14
[70 - 75)	9/11	9/11
[75 - 80)	4/3	4/3

Fonte: Luz et al., 2021.

Figura 5 – Distribuição Demográfica dos Participantes por Faixa Etária e Gênero.



Fonte: Adaptado de Luz et al., 2021.

O conjunto de dados está dividido em conjuntos de treino (70% dos dados) e teste (30% dos dados), seguindo a estrutura estabelecida pela competição ADReSS. Esta divisão padronizada tem sido amplamente adotada na literatura relacionada, permitindo uma comparação justa e direta entre diferentes abordagens metodológicas.

Para ilustrar as diferenças linguísticas características entre os grupos DA e Não-DA, apresentamos abaixo duas transcrições representativas extraídas do conjunto de dados, onde é possível observar padrões típicos de fala de cada grupo:

- Exemplo 1 - Participante sem DA (Controle):
    - "okay. it was summertime and mother and the children were working in the kitchen. and the window was open and there was a slight breeze blowing in. &-um mother was daydreaming and forgot and left the water in the sink running and it was overflowing. &-um the children were hungry and because they knew mother was distracted doing the &+di dishes they did something they probably should not have done. and they got the kitchen stool and moved it under the cupboard where the cookie jar was. and the young boy climbed up to get a cookie. &-um because he did not place his weight correctly on the stool he's about to fall and probably hurt himself (be)cause his head's gonna hit the kitchen cupboards. &-um it looks like the house is set in the country. and it's a large house but [//] &+e either that or you're seeing another house or the [//] a wing of the house. there's grass growing <a little> [/] a little path. mother looks pretty laid back there &=laughs."
- Exemplo 2 - Participante com DA:
    - "mhm. +< alright. there's &-um a young boy that's getting a cookie jar. and it [//] he's &-uh in bad shape because &-uh the thing is fallin(g) over. and in the picture the mother is washin(g) dishes and doesn't see it. and so <is the> [//] the water is overflowing in the sink. and the dishes might <get falled [* +ed] over if you don't> [//] fell [//] fall over there [/] there if you don't get it. and it [//] there [//] it's a picture of a kitchen window. and the curtains are very &-uh distinct. but the water is &+flow still flowing."

Ao analisar estes exemplos, observam-se diferenças significativas nos padrões linguísticos entre os dois grupos. O participante do grupo controle demonstra uma narrativa estruturada temporalmente, começando com a contextualização da cena ("it was summertime") e desenvolvendo uma sequência lógica de eventos. Sua fala apresenta complexidade sintática através do uso de conectivos elaborados para estabelecer relações causais (como em "because they knew mother was distracted"), além de incluir interpretações que vão além da descrição literal da cena (por exemplo, ao inferir que "mother was daydreaming"). A riqueza de detalhes contextuais, como a descrição do ambiente ("house is set in the country"), evidencia uma capacidade preservada de observação e expressão.

Em contraste, o participante com DA apresenta uma narrativa caracterizada por fragmentações e reformulações frequentes, evidenciadas pelas múltiplas correções no discurso (marcadas por "[//]"). Sua fala é marcada por hesitações ("&-um", "&-uh") e estruturas sintáticas mais simples, frequentemente iniciadas pela conjunção "and". Observa-se também dificuldade na precisão lexical e construção gramatical, como em "get falled [* +ed]", bem como uma tendência a focar em elementos isolados da cena sem estabelecer relações temporais ou causais complexas entre eles. Estas diferenças linguísticas são relevantes para a análise computacional desenvolvida nos blocos experimentais subsequentes, pois representam padrões que podem ser identificados e quantificados através das diferentes abordagens metodológicas propostas.

4 EXPERIMENTOS

Neste Capítulo são apresentados os três estudos experimentais desenvolvidos neste trabalho: (1) Análise baseada em Seleção de Características; (2) Análise baseada em Compressão de texto; e (3) Análise baseada em *embeddings* NLP. Para fins de reprodutibilidade dos resultados apresentados, o código-fonte[1] dos experimentos realizados está disponível na internet publicamente.

4.1 BLOCO EXPERIMENTAL 1: ANÁLISE BASEADA EM SELEÇÃO DE CARACTERÍSTICAS

Neste bloco experimental, realizamos uma análise com base na seleção de características acústicas para a classificação de pacientes com DA, utilizando amostras de fala do conjunto de dados ADReSS 2020. Partimos da hipótese de que o uso de algoritmos de compressão poderia reduzir a dimensionalidade dos dados e aumentar a eficiência computacional, mantendo ou até mesmo melhorando os resultados em comparação com o uso do conjunto completo de características.

Este trabalho foi apresentado no 5º Congresso de Ingenierías y Ciencias Aplicadas de las Tres Fronteras (MEC3F), realizado entre os dias 17 e 20 de setembro de 2024, na Universidade Federal da Integração Latino-Americana (UNILA), na Unidade Itaipu Parquetec, em Foz do Iguaçu, Paraná, Brasil (Cardus, Maciel e Zalewski, 2024).

4.1.1 Método

Para essa investigação, foram três conjuntos de atributos acústicos distintos: eGeMAPS (extended Geneva Minimalistic Acoustic Parameter Set), que contém 88 características acústicas (Eyben et al., 2016); ComParE (Computational Paralinguistics Evaluation), com 6.373 características acústicas (Schuller et al., 2019); e emobase, que inclui 988 características relacionadas a aspectos emocionais e expressivos da fala (Eyben, Wöllmer e Schuller, 2010). A extração dessas características foi realizada com o auxílio da biblioteca openSMILE v3.0.

Nesse contexto, foram aplicadas quatro técnicas de seleção de atributos, que incluem Seleção Univariada de Características (SUC), Eliminação Recursiva de

[1] https://github.com/WoolierBrooks/DementiaDetect-Speech

Características (ERC), Select From Model (SFM) e Seleção Sequencial de Características (SSC). Para cada técnica, foram selecionados os 10%, 20% e 30% melhores atributos. Para a classificação, foram utilizados cinco algoritmos de Aprendizado de Máquina, sendo eles: Análise Discriminante Linear (LDA), Árvores de Decisão (DT), Redes Neurais Artificiais (NN), Support Vector Machines (SVM) e Random Forests (RF). Os modelos foram empregados com seus hiperparâmetros padrão (default) conforme implementados na biblioteca scikit-learn, permitindo a análise isolada do impacto dos atributos selecionados sobre o desempenho preditivo. Além disso, a separação entre treino e teste foi realizada, visando garantir a reprodutibilidade dos experimentos e a comparabilidade entre os diferentes métodos avaliados.

4.1.2. Resultados

Os resultados mais relevantes para cada conjunto de características foram:

- emobase: A combinação de SFM com NN alcançou 79,17% de acurácia com 20% dos atributos.
- ComParE: O RF com SFM atingiu 77,08% de acurácia usando 10% dos atributos.
- eGeMAPS: O RF com SUC obteve 70,83% de acurácia utilizando 20% dos atributos.

A análise da eficácia das técnicas de seleção de características revelou que:

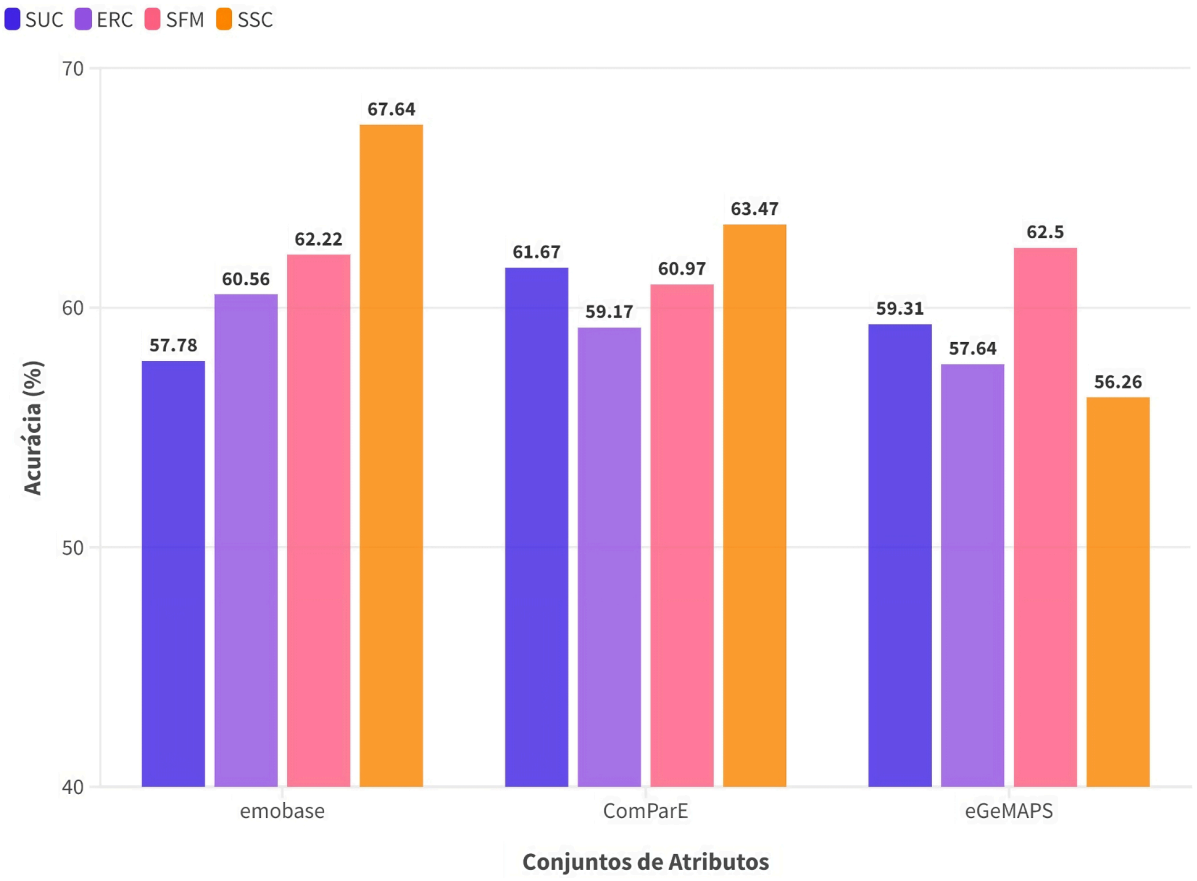
- SUC apresentou desempenho satisfatório em todos os conjuntos, destacando-se em eGeMAPS e ComParE.
- ERC foi eficaz em conjuntos com maior número de atributos, como ComParE e emobase.
- SFM demonstrou ser uma das técnicas mais consistentes, especialmente em conjuntos de características maiores.
- SSC provou ser eficaz, atingindo uma acurácia de 75,00% em ComParE ao utilizar 20% das características.

A Figura 6 contém a acurácia média das técnicas de seleção de características nos diferentes conjuntos de atributos, calculada a partir da média das acurácias dos classificadores utilizados no estudo.

Observa-se que:

- SUC apresentou performance moderada em todos os conjuntos, com acurácia média variando entre 57,78% (emobase) e 61,67% (ComParE).
- ERC mostrou performance mais sensível ao conjunto de dados, variando de 57,64% (eGeMAPS) a 60,56% (emobase).
- SFM demonstrou comportamento consistente, alcançando 62,22% (emobase) e 60,97% (ComParE).
- SSC destacou-se como uma das técnicas mais eficazes no conjunto ComParE (63,47%) e emobase (67,64%).

Figura 6 – Comparação da Acurácia Média (%) das Técnicas de Seleção de Características em Diferentes Conjuntos de Atributos.



Fonte: Adaptado de Luz et al., 2021.

Na avaliação dos algoritmos de classificação, NN e RF destacaram-se consistentemente como os melhores classificadores, atingindo altos níveis de acurácia em diversos cenários. SVM também apresentou bom desempenho em várias condições, enquanto LDA e DT tiveram desempenhos mais modestos.

A análise da influência da quantidade de características nos algoritmos de classificação mostrou que diferentes modelos reagem de maneiras diversas às mudanças na quantidade de atributos utilizados, reforçando a importância da seleção adequada de características para otimizar o desempenho dos algoritmos de classificação.

4.1.3. Análise Estatística

No primeiro bloco experimental, foi realizada uma análise estatística para avaliar a existência de diferenças significativas entre o uso de seletores de características e a utilização integral dos conjuntos de dados emobase, ComParE e eGeMAPS. Para isso, foram definidos cinco grupos de comparação:

- As médias dos conjuntos ComParE, eGeMAPS e emobase para os seletores de características LDA, NN, SVM, RF e DT, considerando 100% das características.
- As médias dos conjuntos ComParE, eGeMAPS e emobase para os seletores LDA, NN, SVM, RF e DT, utilizando a seleção de características de 10%, 20% e 30% por meio do método SUC.
- As médias dos conjuntos ComParE, eGeMAPS e emobase para os seletores LDA, NN, SVM, RF e DT, utilizando a seleção de características de 10%, 20% e 30% através do método ERC.
- As médias dos conjuntos ComParE, eGeMAPS e emobase para os seletores LDA, NN, SVM, RF e DT, utilizando a seleção de características de 10%, 20% e 30% por meio do método SFM.
- As médias dos conjuntos ComParE, eGeMAPS e emobase para os seletores LDA, NN, SVM, RF e DT, utilizando a seleção de características de 10%, 20% e 30% através do método SSC.

Inicialmente, aplicou-se o teste de normalidade de Shapiro-Wilk para determinar a escolha estatística mais adequada para este bloco. Observou-se que três dos cinco conjuntos não apresentaram distribuição normal, optando-se, portanto, pelo teste de Friedman.

O teste resultou em um p-valor = 0,6032. A hipótese nula postula a ausência de diferenças significativas entre os grupos, ou seja, entre os diferentes seletores de características e a utilização integral das características. Estes achados sugerem que a utilização de seletores de características produz resultados comparáveis ao uso integral das características. Tal constatação é relevante, pois indica a possibilidade de redução da dimensionalidade e aumento da eficiência computacional sem comprometer significativamente o desempenho da classificação.

4.2 BLOCO EXPERIMENTAL 2: ANÁLISE BASEADA EM COMPRESSÃO DE TEXTO

As técnicas de compressão de dados desempenham um papel essencial na redução da dimensionalidade, permitindo a representação mais concisa dos dados originais enquanto preservam as características mais relevantes para a tarefa em questão. Estas técnicas podem ser baseadas em princípios estatísticos, como a Análise de Componentes Principais (PCA), ou em arquiteturas neurais mais complexas, como autoencoders, que aprendem automaticamente representações comprimidas otimizadas para diferentes aplicações.

Aproveitando o potencial dessas técnicas de compressão, realizamos neste segundo bloco experimental uma análise de compressão de texto com transcrições de áudios de pacientes com e sem Alzheimer, provenientes da mesma base de dados do primeiro bloco. A hipótese para o uso de compressores de texto baseia-se na premissa de que estes algoritmos podem identificar características linguísticas relevantes para a predição da DA, agrupando textos semelhantes durante o processo de compressão. Esta abordagem foi inspirada no trabalho de Jiang et al. (2023), que propuseram o uso de compressores de texto em tarefas de NLP, combinando-os com o classificador k-nearest-neighbor (kNN). Este método se destaca por ser simples, leve e não requerer pré-processamento ou treinamento, além de ser independente de recursos de GPU. Fundamentando-se no princípio de que compressores capturam regularidades, e que objetos da mesma categoria compartilham mais regularidades entre si do que aqueles de categorias diferentes, o método utiliza a Normalized Compression Distance (NCD), derivada da complexidade de Kolmogorov, como métrica de similaridade entre as transcrições comprimidas, permitindo assim a classificação via kNN.

4.2.1 Método

Nesta abordagem, ao invés de analisar os áudios diretamente, foram utilizadas as transcrições no formato texto (.cha), focando exclusivamente nas falas dos pacientes. Antes de proceder à compressão, foi realizado um pré-processamento nas transcrições, que consistiu na eliminação das falas dos entrevistadores, preservando apenas as falas dos pacientes. Além disso, símbolos que não representam informações úteis para a análise, como os caracteres '%', '|',

'_', etc e números presentes nas transcrições, foram removidos. Esses símbolos são utilizados pelo formato .cha para estabelecer sincronização com os áudios no software CLAN, mas não eram relevantes para a nossa análise de compressão textual. Em seguida, foi aplicada a compressão dos textos utilizando diversos algoritmos de compressão. Os algoritmos aplicados foram LZ4, Zstd, LZ77, gzip, bzip2, Snappy, LZMA, Brotli, Huffman, SF, fpzip, Rice, pySmaz e RLE. Cada um desses algoritmos foi utilizado para comprimir as transcrições de forma a calcular a NCD (Cohen e Vitányi, 2014).entre pares de transcrições. O cálculo da NCD foi realizado conforme a equação:

$$NCD(x, x2) = \frac{C(x+x2) - min(C(x), C(x2))}{max(C(x), C(x2))}$$

onde $C(x)$ e $C(x2)$ representam os comprimentos das transcrições comprimidas x e $x2$, respectivamente, e $C(x + x2)$ representa o comprimento da compressão concatenada dos dois textos.

Após o cálculo dos valores de NCD para cada par de transcrições, prosseguimos com as tarefas de classificação e regressão do Desafio ADReSS 2020. Para ambas as tarefas, empregamos o algoritmo *k-Nearest Neighbors* (kNN) com três variações do parâmetro k: k = 1 (1NN), k = 3 (3NN) e k = 5 (5NN).

Na tarefa de classificação, a métrica principal utilizada foi a acurácia, escolhida para garantir consistência com o primeiro bloco experimental e permitir a comparação direta entre os resultados obtidos. Além disso, foram introduzidas as métricas de recall (Rec.), definida como a razão entre o número de verdadeiros positivos e a soma do número de verdadeiros positivos com o número de falsos negativos, e precisão (Prec.), calculada como a razão entre o número de verdadeiros positivos e a soma do número de verdadeiros positivos com o número de falsos positivos. Para a tarefa de regressão, empregaram-se as métricas de MAE, RMSE e o r entre os valores observados e preditos, com o objetivo de avaliar a qualidade das previsões geradas pelo modelo.

Um aspecto metodológico relevante deste trabalho é a adoção do procedimento Leave-One-Subject-Out (LOSO) no conjunto de treino para seleção de parâmetros, em contraste com a abordagem convencional de holdout (divisão entre treino e teste), comumente utilizada na literatura. Além disso, introduziu-se uma variação do procedimento LOSO, aplicando-o ao conjunto combinado de treino e teste, uma abordagem que não era previamente estabelecida no contexto da competição ADReSS 2020. Essa metodologia permite uma avaliação mais robusta dos modelos, reduzindo o risco de superestimação do desempenho e proporcionando uma análise mais generalizável para novos dados.

Inicialmente, realizou-se o procedimento LOSO no conjunto de treino, tanto para a tarefa de classificação quanto para a de regressão. A partir desse procedimento, selecionou-se o melhor valor de k com base na acurácia obtida para a classificação e no RMSE para a regressão, optando pelo menor k em caso de empate. Após a seleção do melhor valor de k, aplicou-se o modelo ao conjunto de teste. Por fim, repetiu-se o procedimento LOSO considerando tanto o conjunto de treino quanto o de teste, com o objetivo de avaliar a generalização do modelo nos dados completos.

Vale ressaltar que, durante a tarefa de regressão, um paciente controle foi excluído da análise devido à ausência de registro do escore MMSE correspondente.

4.2.2 Resultados e Discussão

Tarefa de Classificação: Avaliação no Conjunto de Treinamento

Os resultados apresentados na Tabela 1 ilustram o desempenho dos algoritmos de compressão utilizando o procedimento de LOSO no conjunto de treino, com diferentes valores de k (1, 3 e 5) para o algoritmo kNN. A tabela mostra três métricas principais: Rec., Prec. e acurácia (Acc.), para cada algoritmo de compressão avaliado.

Os resultados indicam que o desempenho dos algoritmos de compressão variou com o valor de k. No 1NN, LZ4 e bzip obtiveram as melhores Acc. (80,56%). Com 3NN, LZ4 e Zstd atingiram 86,11%, e, no 5NN, o LZ4 alcançou a maior Acc. (88,89%) e a alta Prec. (95,65%). Em contraste, algoritmos como pySmaz e RLE

mostraram baixo desempenho (~50%), sugerindo que suas técnicas de compressão não foram eficazes em distinguir entre Alzheimer e controles.

De maneira geral, os algoritmos LZ4, Zstd e LZ77 apresentaram os melhores resultados, especialmente para 3NN e 5NN, sugerindo que são mais adequados para a compressão de transcrições de pacientes com Alzheimer e controles no contexto da classificação por compressão. A escolha do melhor valor de k variou entre os algoritmos, sendo que 3NN mostrou um bom equilíbrio entre Rec., Prec. e Acc. para a maioria dos algoritmos, enquanto 5NN favoreceu os algoritmos LZ4 e LZ77.

Tabela 1 – Resultados do procedimento de LOSO no conjunto de treino para diferentes algoritmos de compressão e valores de k.

Compressor	1NN			3NN			5NN		
	Rec.	**Prec.**	**Acc.**	**Rec.**	**Prec.**	**Acc.**	**Rec.**	**Prec.**	**Acc.**
LZ4	72,22%	86,67%	80,56%	75,93%	95,35%	86,11%	81,48%	95,65%	88,89%
Zstd	72,22%	84,78%	79,63%	77,78%	93,33%	86,11%	74,07%	95,24%	85,19%
LZ77	75,93%	82,00%	79,63%	77,78%	85,71%	82,41%	83,33%	91,84%	87,96%
gzip	74,07%	83,33%	79,63%	79,63%	87,76%	84,26%	74,07%	90,91%	83,33%
bzip2	81,48%	80,00%	80,56%	75,93%	87,23%	82,41%	74,07%	88,89%	82,41%
Snappy	74,07%	81,63%	78,70%	79,63%	79,63%	79,63%	81,48%	88,00%	85,19%
LZMA	81,48%	78,57%	79,63%	74,07%	80,00%	77,78%	74,07%	88,89%	82,41%
Brotli	75,93%	71,93%	73,15%	68,52%	67,27%	67,59%	75,93%	71,93%	73,15%
Huffman	73,08%	66,07%	66,67%	71,79%	69,81%	69,44%	79,49%	66,67%	65,74%
SF	70,37%	64,41%	65,74%	62.96%	59,65%	60,19%	59,26%	61,54%	61,11%
fpzip	59,26%	58,18%	58,33%	62,96%	64,15%	63,89%	62,96%	64,15%	63,89%
Rice	50,00%	61,36%	59,26%	46,30%	56,82%	55,56%	35,19%	47,50%	48,15%
pySmaz	31,48%	51,52%	50,93%	31,48%	47,22%	48,15%	37,04%	51,28%	50,93%
RLE	100%	0,00%	50,00%	100%	0,00%	50,00%	0,00%	0,00%	50,00%

Fonte: Elaborado pelo autor (2024).

Tarefa de Classificação: Avaliação no Conjunto de Teste (Competição)

Após a realização do procedimento de LOSO no conjunto de treino e a identificação do valor de k que maximizou o desempenho para cada algoritmo de compressão, procedeu-se à avaliação do desempenho desses modelos no conjunto de teste (*holdout*). A Tabela 2 apresenta os resultados, destacando as métricas de Rec,, Prec. e Acc. para cada algoritmo de compressão, utilizando o valor de k otimizado no treino. Foram identificados padrões importantes e algumas divergências em relação ao desempenho observado no conjunto de treino.

O algoritmo LZ4 apresentou o melhor desempenho no conjunto de teste, com Acc. de 85,42%. O gzip obteve a segunda maior Acc. (83,33%), demonstrando equilíbrio entre as classes. Algoritmos como RLE (50,00%) e Rice (58,33%) tiveram baixo desempenho, reforçando que métodos mais simples não capturam bem as características linguísticas necessárias para a classificação.

Tabela 2 – Resultados da avaliação *holdout* no conjunto de teste para diferentes algoritmos de compressão.

Compressor	Rec.	Prec.	Acc.
LZ4 (5NN)	83,33%	86,96%	85,42%
Zstd (3NN)	62,50%	83,33%	75,00%
LZ77 (5NN)	62,50%	88,24%	77,08%
gzip (3NN)	79,17%	86,36%	83,33%
bzip2 (3NN)	75,00%	85,71%	81,25%
Snappy (5NN)	79,17%	82,61%	81,25%
LZMA (5NN)	70,83%	77,27%	75,00%
Brotli (1NN)	66,67%	69,57%	68,75%
Huffman (3NN)	79,17%	65,52%	68,75%
SF (1NN)	62,50%	75,00%	70,83%
fpzip (3NN)	75,00%	69,23%	70,83%
Rice (1NN)	62,50%	57,69%	58,33%
pySmaz (1NN)	50,00%	75,00%	66,67%
RLE (1NN)	0,00%	0,00%	50,00%

Fonte: Elaborado pelo autor (2024).

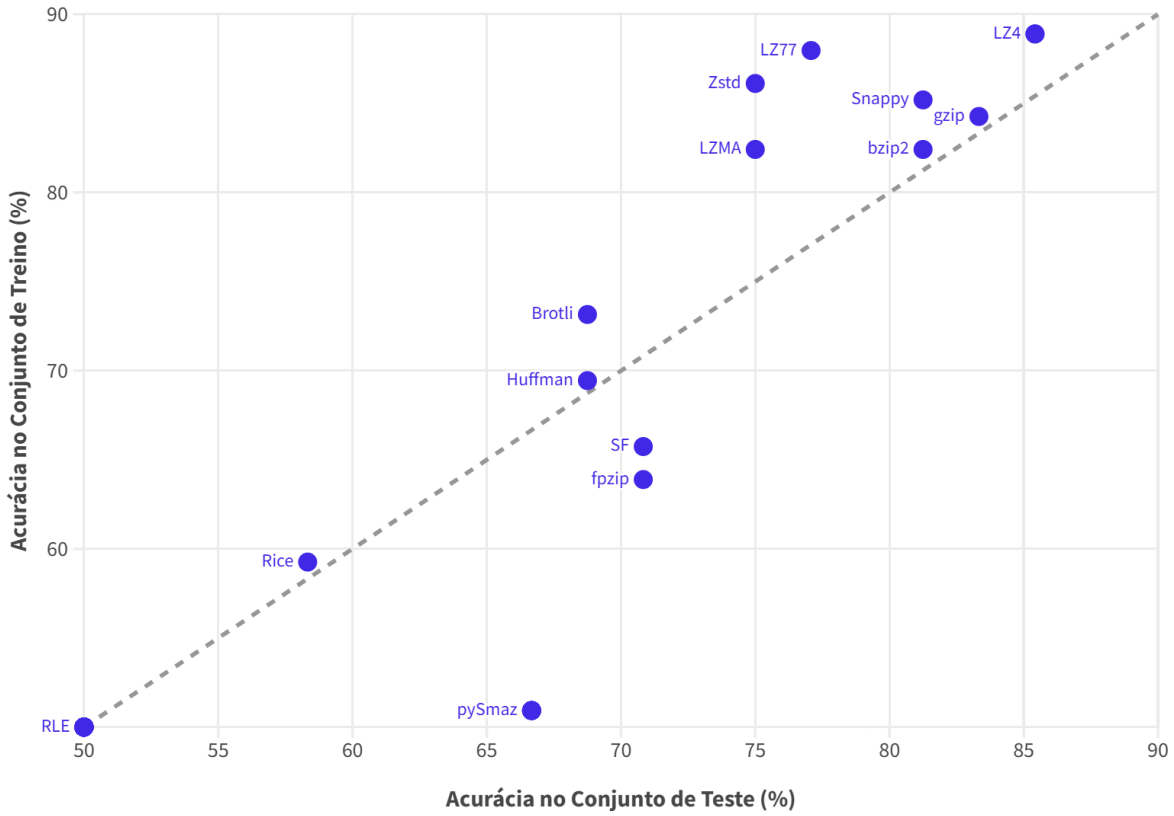
Os resultados da avaliação *holdout* (70/30) no conjunto de teste confirmam a eficácia de alguns algoritmos de compressão na tarefa de classificação de Alzheimer, com destaque para os algoritmos LZ4, gzip, bzip2 e Snappy, que demonstraram boa capacidade de generalização ao manterem um desempenho robusto em dados não utilizados no treinamento. No entanto, observou-se variações no desempenho entre o conjunto de treino e o de teste, ressaltando a importância de uma avaliação rigorosa e a necessidade de considerar cuidadosamente a escolha do algoritmo de compressão mais apropriado para a tarefa em questão.

A robustez e a capacidade de generalização dos modelos baseados em compressão são avaliadas ao comparar o desempenho nos conjuntos de treino e teste. A comparação direta entre os resultados obtidos no procedimento de LOSO no conjunto de treino e os resultados do holdout no conjunto de teste para cada algoritmo de compressão está representada na Figura 7.

A Figura 7 oferece uma análise visual da capacidade de generalização dos algoritmos de compressão testados. O gráfico de dispersão compara a Acc. obtida no conjunto de treino (eixo y) com a Acc. no conjunto de teste (eixo x) para cada algoritmo. A linha diagonal tracejada indica o ponto de equivalência entre os desempenhos nos dois conjuntos, facilitando a identificação de casos de overfitting e underfitting.

Algoritmos como LZ4, gzip, Snappy e bzip2 demonstram desempenho uniforme entre os conjuntos de treino e teste, indicando boa capacidade de generalização. Em contrapartida, LZ77, Zstd e LZMA mostram sinais de overfitting, com Acc. significativamente superiores no treino, sendo que o Zstd apresenta uma queda de 11,11% entre os conjuntos. Rice, pySmaz e RLE têm baixo desempenho em ambos os conjuntos, evidenciando que suas abordagens de compressão não capturam os padrões linguísticos necessários para a classificação.

Figura 7 – Comparação da Acc. entre conjuntos de treino e teste para diferentes algoritmos de compressão.



Fonte: Elaborado pelo autor (2024).

Os algoritmos SF e fpzip exibem uma performance levemente superior no teste em relação ao treino, sugerindo uma capacidade única de generalização ou benefícios de características específicas dos dados de teste. Há uma variação significativa na robustez dos algoritmos: enquanto LZ4 e gzip mantêm desempenho estável, LZ77 e Zstd sofrem quedas acentuadas. Esses resultados ressaltam a importância de avaliar a generalização dos modelos na seleção do algoritmo de compressão mais adequado para a classificação da Doença de Alzheimer a partir de transcrições de fala. O LZ4 se destaca como o algoritmo mais promissor, equilibrando alto desempenho e boa generalização, embora todos os algoritmos de melhor desempenho apresentem algum grau de overfitting.

Tarefa de Classificação: Avaliação Geral

A análise experimental foi ampliada para incluir um procedimento de avaliação LOSO nos conjuntos de treinamento e teste combinados, proporcionando uma avaliação mais robusta do desempenho dos algoritmos de compressão. A Tabela 3 ilustra esses resultados para diferentes valores de k (1, 3 e 5) no algoritmo kNN.

Os resultados da avaliação LOSO no conjunto de dados combinado revelam diversos padrões interessantes. O bzip2 destacou-se como o melhor desempenho em todos os valores de k, mostrando uma melhoria consistente à medida que k aumenta. Com 5NN, alcança a maior Acc. geral de 85,26%, com uma sólida Prec. de 91,04%. Isso sugere que o bzip2 é particularmente eficaz em capturar os padrões linguísticos associados à DA.

LZ4 e Snappy também demonstram um desempenho interessante e consistente em todos os valores de k. O LZ4 atinge seu desempenho máximo em 5NN, com uma Acc. de 83,97%, enquanto o Snappy obtém seu melhor resultado em 5NN, com uma Acc. de 84,62%. Ambos os algoritmos apresentam um bom equilíbrio entre Rec. e Prec., indicando sua robustez em distinguir entre pacientes com DA e controles.

O algoritmo LZMA, que teve um bom desempenho no conjunto de treinamento, mantém uma performance sólida no conjunto combinado, atingindo seu pico em 5NN, com uma Acc. de 82,05%. Isso sugere que a eficácia do LZMA observada no conjunto de treinamento se generaliza bem para o conjunto de dados maior. O algoritmo Zstd apresenta desempenho consistente em todos os valores de k, com sua melhor Acc. de 81,41% em 3NN, indicando que o algoritmo de compressão é menos sensível às variações no número de vizinhos considerados.

O algoritmo gzip, que teve um bom desempenho no conjunto de teste, apresenta resultados mais variáveis no conjunto combinado. Seu desempenho atinge o pico em 3NN, com uma Acc. de 79,49%, mas apresenta uma leve queda em 5NN. Essa variabilidade sugere que o desempenho do gzip pode depender mais das características específicas do conjunto de dados ao qual é aplicado.

Tabela 3 – Resultados da avaliação LOSO no conjunto combinado (treinamento + teste) para diferentes algoritmos de compressão e valores de k.

Compressor	1NN			3NN			5NN		
	Rec.	Prec.	Acc.	Rec.	Prec.	Acc.	Rec.	Prec.	Acc.
LZ4	75,64%	79,45%	77,56%	79,49%	84,00%	82,69%	78,21%	84,42%	83,97%
Zstd	61,54%	79,73%	78,21%	69,23%	82,67%	81,41%	60,26%	82,43%	80,77%
LZ77	73,08%	74,51%	66,03%	76,92%	88,24%	75,00%	76,92%	89,13%	73,08%
gzip	73,08%	72,73%	69,23%	71,79%	87,10%	79,49%	79,49%	88,68%	76,28%
bzip2	76,36%	84,51%	81,41%	80,77%	87,14%	83,33%	83,33%	91,04%	85,26%
Snappy	48,72%	79,17%	76,92%	57,69%	84,85%	79,49%	52,56%	88,57%	84,62%
LZMA	76,92%	74,42%	76,92%	78,21%	77,91%	80,77%	78,21%	83,78%	82,05%
Brotli	82,05%	67,42%	69,87%	85,90%	72,73%	75,64%	79,49%	74,36%	74,36%
Huffman	76,92%	69,51%	70,51%	82,08%	70,59%	72,44%	74,36%	76,92%	76,92%
SF	53,85%	71,19%	66,03%	64,10%	76,92%	72,44%	62,82%	81,67%	74,36%
fpzip	48,72%	71,70%	64,74%	73,08%	67,06%	68,59%	73,08%	69,51%	70,51%
Rice	100%	50,00%	50,00%	100%	48,97%	48,08%	0,00%	50,00%	50,00%
pySmaz	98,72%	58,51%	60,26%	91,03%	56,98%	57,69%	51,28%	63,29%	63,46%
RLE	70,51%	50,00%	50,00%	62,82%	50,00%	50,00%	64,10%	0,00%	50,00%

Fonte: Elaborado pelo autor (2024).

O LZ77, que apresentou indícios de overfitting na análise anterior, demonstra uma estabilidade reduzida no conjunto combinado. Atingindo uma Acc. máxima de 75,00% em 3NN. Embora esse desempenho seja maior que de outros algoritmos, os resultados em um conjunto de dados maior sugere que o LZ77 pode ser menos confiável do que inicialmente indicado.

Algoritmos como Rice, pySmaz e RLE continuam a apresentar um desempenho insatisfatório em todos os valores de k, confirmando sua inadequação para esta tarefa específica. O RLE, em particular, não consegue identificar nenhum caso de Alzheimer (Prec. = 0%) para 5NN, destacando sua ineficácia para esta aplicação.

Os resultados da avaliação LOSO no conjunto combinado fornecem descobertas valiosas sobre as capacidades de generalização dos diversos algoritmos de compressão. O forte desempenho do bzip2, LZ4 e Snappy em diferentes valores de k sugere que esses algoritmos são particularmente adequados

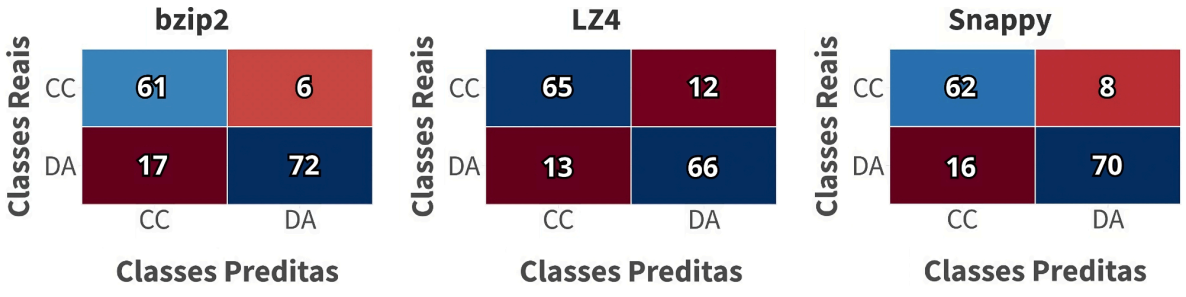
para capturar as nuances linguísticas que distinguem pacientes com Alzheimer de controles.

A estabilidade reduzida de alguns algoritmos (como o LZ77) neste conjunto de dados maior sublinha a importância de métodos de avaliação abrangentes. Também destaca os potenciais benefícios de utilizar conjuntos de dados maiores e combinados para uma avaliação mais robusta dos algoritmos em aplicações clínicas.

Além disso, o desempenho consistentemente ruim de certos algoritmos (Rice, pySmaz, RLE) em diferentes métodos de avaliação reforça a conclusão de que nem todos os algoritmos de compressão são igualmente adequados para este tipo de análise linguística.

A análise dos resultados com a aplicação de algoritmos de compressão na classificação de pacientes com Alzheimer e controles, conforme ilustrado nas matrizes de confusão, revela informações relevantes sobre o desempenho dos três melhores algoritmos: bzip2, LZ4 e Snappy. As matrizes de confusão, representadas na Figura 8, detalham o número de VN, FP, FN e VP gerados por cada algoritmo após a validação cruzada LOSO aplicada ao conjunto combinado de treino e teste. Nesta representação, o VN está localizado no quadrante superior esquerdo, o FP no quadrante superior direito, o FN no quadrante inferior esquerdo e o VP no quadrante inferior direito.

Figura 8 – Matrizes de Confusão para os Algoritmos bzip2, LZ4 e Snappy.



Fonte: Elaborado pelo autor (2024).

A análise da matriz de confusão revela importantes aspectos de cada algoritmo. O bzip2 demonstra boa capacidade de identificação de VP, mas a quantidade expressiva de FN é preocupante, pois indica que vários pacientes com Alzheimer foram incorretamente classificados como controles. O LZ4, embora tenha a maior quantidade de VN, apresenta o maior número de FP, sugerindo que alguns

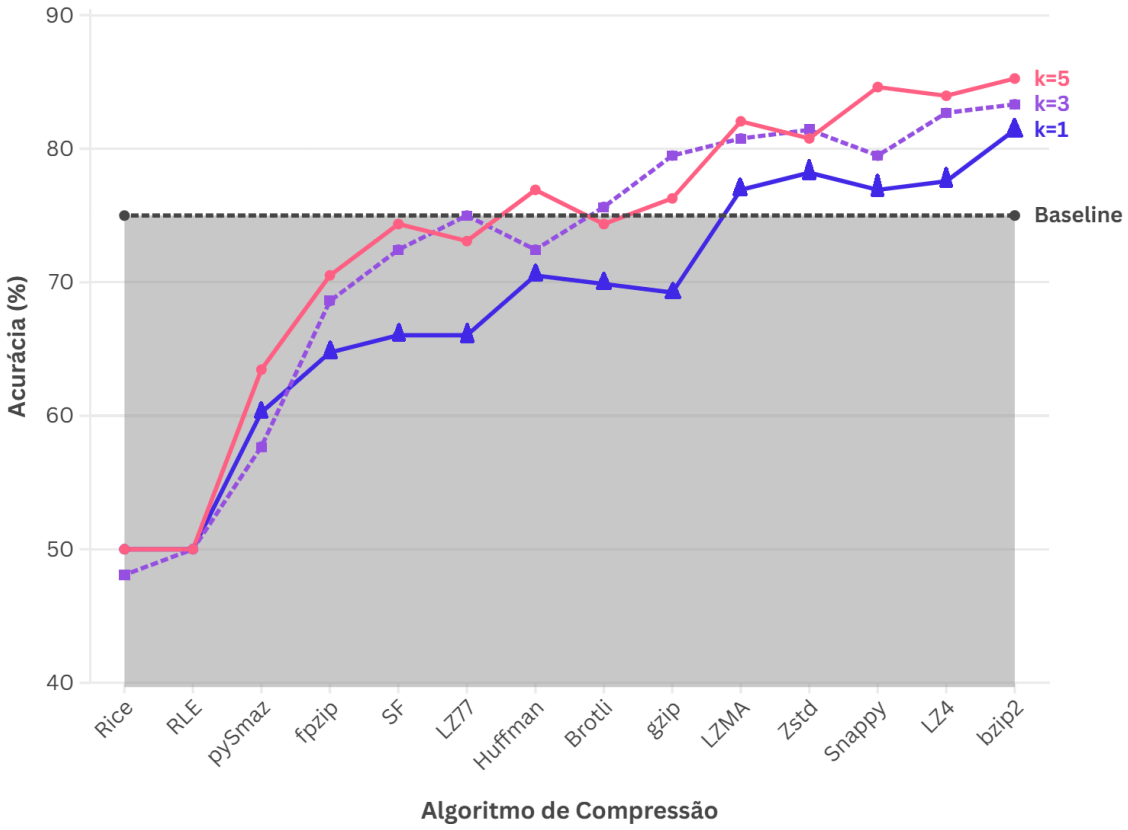
indivíduos saudáveis foram erroneamente identificados como portadores de Alzheimer. No entanto, o número relativamente baixo de FN indica que o LZ4 é eficaz em minimizar erros de não diagnóstico, o que é crucial na prática clínica. Por fim, o Snappy se destaca por apresentar um número inferior de FN em comparação ao bzip2, demonstrando maior Acc. na detecção de pacientes com Alzheimer, além de um número moderado de FP, que contribui para um equilíbrio mais satisfatório entre as classificações corretas e incorretas.

Uma consideração central na análise desses resultados é a minimização de FN, dado que erros desse tipo, ao classificar pacientes com Alzheimer como controles saudáveis, são extremamente prejudiciais por impedir que esses pacientes recebam o tratamento necessário. Em contextos clínicos, evitar tais erros é prioritário, uma vez que o diagnóstico precoce e o tratamento oportuno podem melhorar significativamente a qualidade de vida e retardar a progressão da doença. Como evidenciado nas matrizes de confusão, o algoritmo LZ4 se destaca na minimização de falsos negativos em relação aos outros métodos, enquanto o bzip2 apresenta uma taxa preocupante desses erros.

De forma geral, os três algoritmos demonstram bom desempenho, mas a escolha do mais adequado depende do equilíbrio entre a redução de falsos positivos e falsos negativos. O algoritmo LZ4, por exemplo, pode ser mais indicado em cenários onde a prioridade é evitar a falta de diagnóstico correto para pacientes com Alzheimer, apesar do maior número de falsos positivos. Por outro lado, o algoritmo Snappy oferece um equilíbrio mais favorável entre falsos positivos e negativos, tornando-se uma opção viável para maximizar a Prec. global do sistema de classificação.

Na Figura 9, observa-se a comparação do desempenho dos algoritmos de compressão no procedimento LOSO, aplicado ao conjunto combinado de treinamento e teste, considerando os valores de k = 1, 3 e 5 no algoritmo kNN. O eixo x está ordenado de forma crescente em relação à acurácia, com os algoritmos de pior desempenho à esquerda e os de melhor desempenho à direita.

Figura 9 – Comparação do Desempenho dos Algoritmos de Compressão em Classificação de Dados com k-Nearest Neighbors.



Fonte: Elaborado pelo autor (2024).

A análise do gráfico revela tendências importantes. Em geral, para a maioria dos algoritmos, observa-se uma melhora na Acc. conforme o valor de k aumenta, indicando que considerar um número maior de vizinhos próximos tende a resultar em classificações mais robustas. O algoritmo bzip2 destaca-se pelo melhor desempenho geral, com sua linha correspondente a 5NN atingindo o ponto mais alto do gráfico, corroborando os resultados anteriores e confirmando sua eficácia na captura de padrões linguísticos relevantes para a detecção de Alzheimer. O baseline para a tarefa de classificação foi de 75% de Acc. (Luz et al., 2021).

Os algoritmos LZ4 e Snappy apresentam curvas bastante próximas às de bzip2, especialmente para 3NN e 5NN, sugerindo que são alternativas viáveis e quase tão eficazes quanto o bzip2 para esta tarefa. Nota-se também uma distinção clara entre os algoritmos de melhor desempenho (bzip2, LZ4, Snappy, Zstd e LZMA) e aqueles com desempenho inferior (Rice, RLE, pySmaz). Algoritmos como Rice,

RLE e pySmaz mantêm desempenho abaixo da linha de base para todos os valores de k, confirmando sua inadequação para esta tarefa específica.

Em alguns casos, como nos algoritmos Zstd e gzip, observa-se um ponto de inflexão em que o aumento de k de 3 para 5 não resulta em melhora significativa e, em alguns casos, leva até a uma leve diminuição na Acc. Isso sugere que, para esses algoritmos, 3NN pode ser um valor ótimo, equilibrando a complexidade do modelo com a capacidade de generalização.

A consistência desses resultados com análises anteriores fortalece a confiabilidade das conclusões sobre a eficácia dos diferentes algoritmos de compressão. Ademais, a visualização gráfica facilita a identificação de padrões e tendências que podem não ser imediatamente evidentes em tabelas de dados.

Comparando nossos resultados no conjunto de teste com os obtidos na competição, o algoritmo LZ4 se destaca, apresentando um desempenho comparável aos resultados de Chen (University of Michigan) e Jiahong (Baidu USA), que alcançaram Acc. de 89,58%, ocupando o primeiro e o segundo lugares, respectivamente. O gzip também demonstra um desempenho sólido, enquanto bzip2 e Snappy alcançam 81,25%, superando outros concorrentes. O LZ77 está em linha com alguns participantes e supera o baseline de 75%. Em contrapartida, algoritmos como RLE (50%) e Rice (58,33%) apresentam desempenho inferior ao pior colocado, ficando abaixo de Higuchi (The University of Tokyo), que obteve 60,42%.

Tarefa de Regressão: Avaliação no Conjunto de Treinamento

Assim como na tarefa de classificação, a tarefa de regressão também foi realizada por meio da avaliação LOSO no conjunto de treino, utilizando o algoritmo kNN com diferentes valores de k (1, 3 e 5). O objetivo da regressão foi prever o MMSE dos pacientes, uma medida importante da gravidade da demência.

Os resultados desta análise estão na Tabela 4, incluindo as métricas de MAE, RMSE e o coeficiente de correlação r entre os valores verdadeiros e preditos para cada algoritmo de compressão testado.

Tabela 4 – Resultados da avaliação LOSO no conjunto de treino para diferentes algoritmos de compressão e valores de k na tarefa de regressão.

Compressor	1NN			3NN			5NN		
	MAE	**RMSE**	**r**	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**	**r**
LZ77	3,91	5,30	0,73	3,34	4,59	0,78	3,34	4,41	0,82
Zstd	3,79	5,52	0,70	3,50	4,73	0,77	3,54	4,69	0,78
LZ4	3,93	5,69	0,67	3,38	4,73	0,77	3,47	4,73	0,79
Snappy	4,37	5,89	0,64	3,41	4,77	0,76	3,55	4,70	0,78
bzip2	3,94	5,59	0,67	3,70	4,86	0,74	3,87	5,00	0,73
gzip	4,30	6,06	0,63	3,62	4,96	0,74	3,81	4,96	0,76
LZMA	4,74	6,53	0,55	3,75	5,12	0,71	3,92	5,05	0,74
Brotli	5,37	7,22	0,42	3,70	4,84	0,75	4,05	5,08	0,74
Huffman	5,60	7,84	0,39	5,39	6,87	0,39	5,63	6,98	0,33
fpzip	6,79	8,71	0,24	5,71	7,23	0,27	5,64	6,85	0,33
SF	6,37	8,71	0,25	5,74	7,02	0,34	5,81	7,08	0,28
Rice	6,62	9,21	0,16	6,39	8,30	0,05	6,56	8,10	-0,04
pySmaz	6,89	9,51	0,12	6,92	8,88	-0,03	6,90	8,55	-0,09
RLE	6,98	10,01	-0,09	6,79	9,78	0,08	6,76	9,74	-0,20

Fonte: Elaborado pelo autor (2024).

Os resultados da regressão mostram variações significativas no desempenho dos algoritmos de compressão. O LZ77 destacou-se com os melhores resultados para 5NN, apresentando uma melhoria em relação ao baseline de RMSE de 5,2 reportado por Luz et al. (2021). O Zstd também demonstrou bom desempenho, indicando que ambos são eficazes na captura de características nas

transcrições para a previsão do MMSE. Por outro lado, RLE e pySmaz mostraram desempenhos inferiores, indicando sua inadequação para esta tarefa.

A forte correlação (r > 0,7) observada para os melhores algoritmos indica que o método de compressão, combinado com o kNN, é capaz de capturar eficientemente a relação entre as características textuais das transcrições e os scores do MMSE. Os resultados da regressão corroboram os achados da tarefa de classificação, destacando a eficácia dos algoritmos de compressão, particularmente o LZ77 e o Zstd, na análise de transcrições para a avaliação da demência.

Tarefa de Regressão: Avaliação no Conjunto de Teste (Competição)

Após a otimização do parâmetro k no conjunto de treino através do procedimento LOSO, procedemos à avaliação do desempenho desses modelos no conjunto de teste para a tarefa de regressão. Os resultados obtidos nesta análise de holdout, utilizando o valor de k otimizado no treino, estão na Tabela 5.

Tabela 5 – Resultados do *holdout* no conjunto de teste para diferentes algoritmos de compressão na tarefa de regressão.

Compressor	MAE	RMSE	r
LZ77 (k=5)	3,41	5,11	0,62
Zstd (k=5)	3,75	5,39	0,53
LZ4 (k=3)	3,41	5,18	0,60
Snappy (k=5)	3,73	4,98	0,62
bzip2 (k=3)	4,39	6,40	0,47
gzip (k=3)	4,20	6,05	0,49
LZMA (k=5)	3,77	5,27	0,55
Brotli (k=3)	4,40	6,13	0,43
Huffman (k=3)	4,78	6,00	0,38
fpzip (k=5)	4,37	5,49	0,45
SF (k=3)	4,44	5,74	0,40
Rice (k=5)	4,68	5,46	0,46
pySmaz (k=5)	5,75	7,13	-0,02
RLE (k=5)	5,64	8,17	0,00

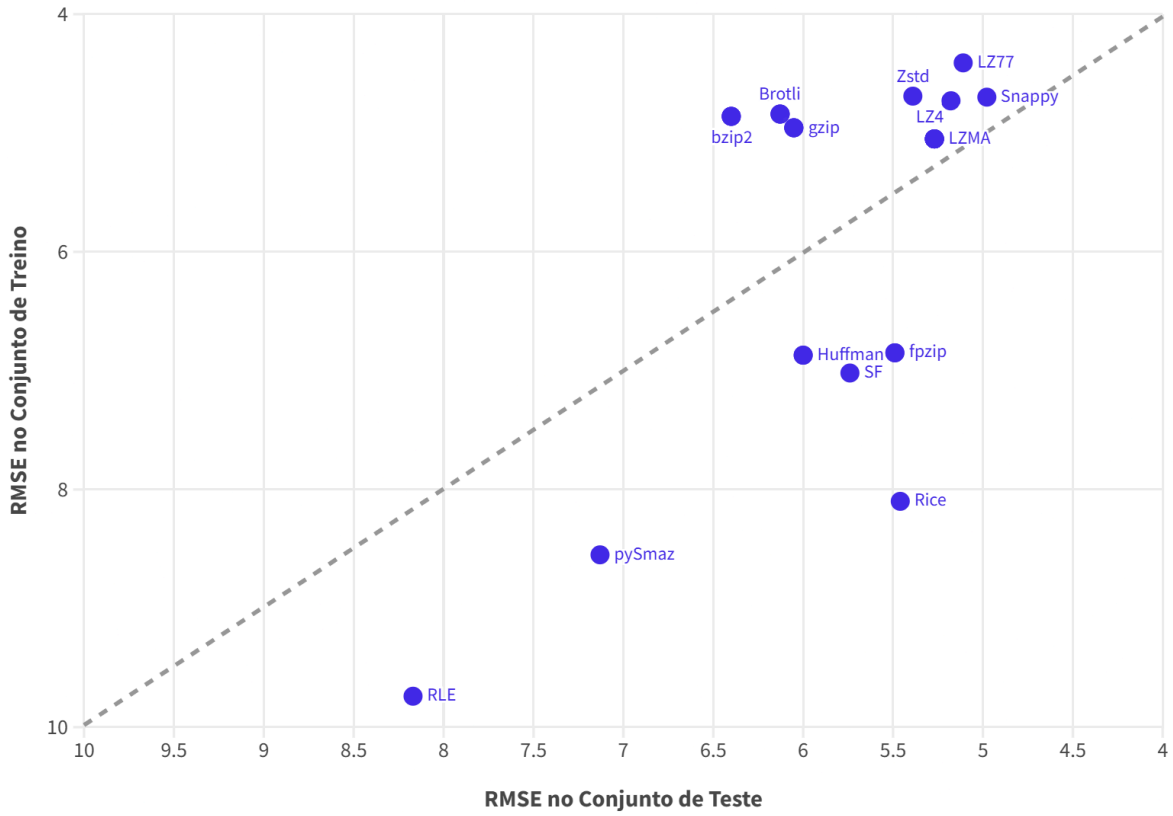
Fonte: Elaborado pelo autor (2024).

Os resultados no conjunto de teste para a tarefa de regressão revelam padrões interessantes e diferenças notáveis em relação ao desempenho observado no conjunto de treino. O algoritmo LZ77, que se destacou anteriormente, manteve um desempenho robusto no teste, apresentando um MAE de 3,41, um RMSE de 5,11 e um r de 0,62. Embora tenha ocorrido uma leve diminuição em comparação com os resultados do treino, o LZ77 ainda demonstra eficácia na previsão do MMSE, aproximando-se do baseline de RMSE de 5,2 reportado por Luz et al. (2021). Ao comparar os resultados com os da competição, o LZ77 apresenta um desempenho inferior ao do primeiro colocado, Koo (3,7472), e do segundo colocado, Chen (4,2902). O algoritmo Snappy destacou-se com um RMSE de 4,98, enquanto o LZ4 apresentou um RMSE de 5,18, ambos situando-se próximos ao top 50% da competição. Em contrapartida, algoritmos como RLE (8,17) e pySmaz (7,13) demonstraram desempenhos inferiores. Esses resultados indicam que a combinação de métodos de compressão com kNN pode ser otimizada para melhorar ainda mais a previsão do MMSE.

A correlação moderada a forte ($r > 0,6$) observada para os melhores algoritmos no conjunto de teste confirma que o método de compressão, combinado com o kNN, é capaz de capturar eficientemente a relação entre as características textuais das transcrições e os scores do MMSE, mesmo em dados não utilizados no treinamento.

A análise comparativa entre o desempenho dos modelos de regressão nos conjuntos de treino e teste oferece informações valiosas sobre a robustez e capacidade de generalização dos diferentes algoritmos de compressão utilizados. A Figura 10 representa um gráfico de dispersão onde o eixo x representa o RMSE no conjunto de teste e o eixo y o RMSE no conjunto de treino. A linha diagonal tracejada indica o ponto de equivalência entre os desempenhos nos dois conjuntos, facilitando a identificação de casos de overfitting e underfitting. Analisando o gráfico, podemos observar algoritmos com boa generalização como o LZ4, LZMA, Snappy e LZ77 apresentam resultados próximos à linha diagonal, indicando um desempenho consistente entre os conjuntos de treino e teste. Esses algoritmos demonstram boa capacidade de generalização.

Figura 10 – Comparação do RMSE entre conjuntos de treino e teste para diferentes algoritmos de compressão na tarefa de regressão.



Fonte: Elaborado pelo autor (2024).

Algoritmos como bzip2, Brotli e gzip mostram sinais de overfitting, com RMSE menor no conjunto de treino em comparação ao teste. O bzip, por exemplo, apresenta a maior discrepância, com um aumento significativo no RMSE ao passar do treino para o teste. Algoritmos como RLE, pySmaz e Rice exibem alto RMSE tanto no treino quanto no teste, indicando que essas abordagens de compressão não são adequadas para capturar os padrões linguísticos necessários para a tarefa de regressão do MMSE. Os algoritmos Huffman, SF e fpzip mostram um desempenho ligeiramente melhor no conjunto de teste em relação ao treino. Esse comportamento incomum pode sugerir que esses algoritmos se beneficiam de características particulares dos dados de teste ou apresentam uma capacidade única de generalização.

Esses resultados corroboram e expandem as observações feitas anteriormente sobre o desempenho individual dos algoritmos nos conjuntos de treino e teste. O bzip2, que havia se destacado no conjunto de treino, mostra uma queda

significativa de performance no teste, indicando uma tendência ao overfitting. Por outro lado, algoritmos como Snappy e LZMA demonstram maior consistência, sugerindo uma melhor capacidade de generalização.

É importante notar que, mesmo para os algoritmos com melhor desempenho, há um aumento no RMSE ao passar do conjunto de treino para o teste. Isso é esperado e indica que os modelos enfrentam desafios ao lidar com dados não vistos, ressaltando a importância de considerar não apenas o desempenho no treino, mas também a capacidade de generalização ao selecionar o algoritmo mais adequado para a tarefa de regressão do MMSE.

Tarefa de Regressão: Avaliação Geral

A análise foi ampliada para incluir uma avaliação LOSO nos conjuntos de dados combinados de treinamento e teste, proporcionando uma avaliação mais abrangente do desempenho dos algoritmos de compressão na tarefa de regressão. A Tabela 6 ilustra esses resultados para diferentes valores de k (1, 3 e 5) no algoritmo k-Nearest Neighbors. Com base nesses resultados, observa-se uma melhoria consistente no desempenho à medida que o valor de k aumenta para a maioria dos algoritmos. Isso sugere que considerar um número maior de vizinhos próximos geralmente leva a previsões mais precisas do MMSE.

Os resultados do LOSO no conjunto combinado são geralmente consistentes com as observações feitas nos conjuntos de treino e teste separadamente. No entanto, alguns algoritmos, como o LZ77, demonstram um desempenho mais robusto no conjunto combinado, sugerindo uma boa capacidade de generalização. O desempenho do LZ77 no conjunto combinado (RMSE = 4,70 para 5NN) é superior ao baseline de RMSE de 5,2 reportado por Luz et al. (2021), indicando uma melhoria significativa na Prec. da previsão do MMSE.

Tabela 6 – Resultados da avaliação LOSO no conjunto combinado (treinamento + teste) para diferentes algoritmos de compressão e valores de k na tarefa de regressão.

Compressor	1NN			3NN			5NN		
	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r
LZ77	4,19	5,86	0,62	3,71	5,14	0,68	3,51	4,70	0,74
Zstd	5,81	7,97	0,31	5,22	7,01	0,35	4,99	6,54	0,40
LZ4	4,19	6,17	0,58	3,63	5,20	0,67	3,47	4,81	0,72
Snappy	4,41	6,26	0,55	4,03	5,51	0,62	3,69	4,97	0,70
bzip2	4,52	6,64	0,53	4,14	5,92	0,57	3,87	5,32	0,65
gzip	4,46	6,42	0,54	3,73	5,21	0,66	3,82	5,23	0,66
LZMA	4,58	6,71	0,52	3,89	5,66	0,60	3,78	5,29	0,65
Brotli	5,74	7,68	0,35	4,97	6,55	0,46	4,71	6,09	0,50
Huffman	5,36	7,54	0,39	4,87	6,58	0,44	4,82	6,35	0,46
fpzip	5,74	7,80	0,35	5,38	7,16	0,33	5,27	6,90	0,32
SF	5,81	7,97	0,31	5,22	7,01	0,35	4,99	6,54	0,40
Rice	6,03	8,83	0,19	5,51	7,34	0,28	5,56	7,16	0,26
pySmaz	5,59	8,13	0,29	6,06	8,10	0,11	6,45	8,22	0,01
RLE	6,64	9,55	-0,08	6,35	9,21	0,07	6,45	9,32	0,07

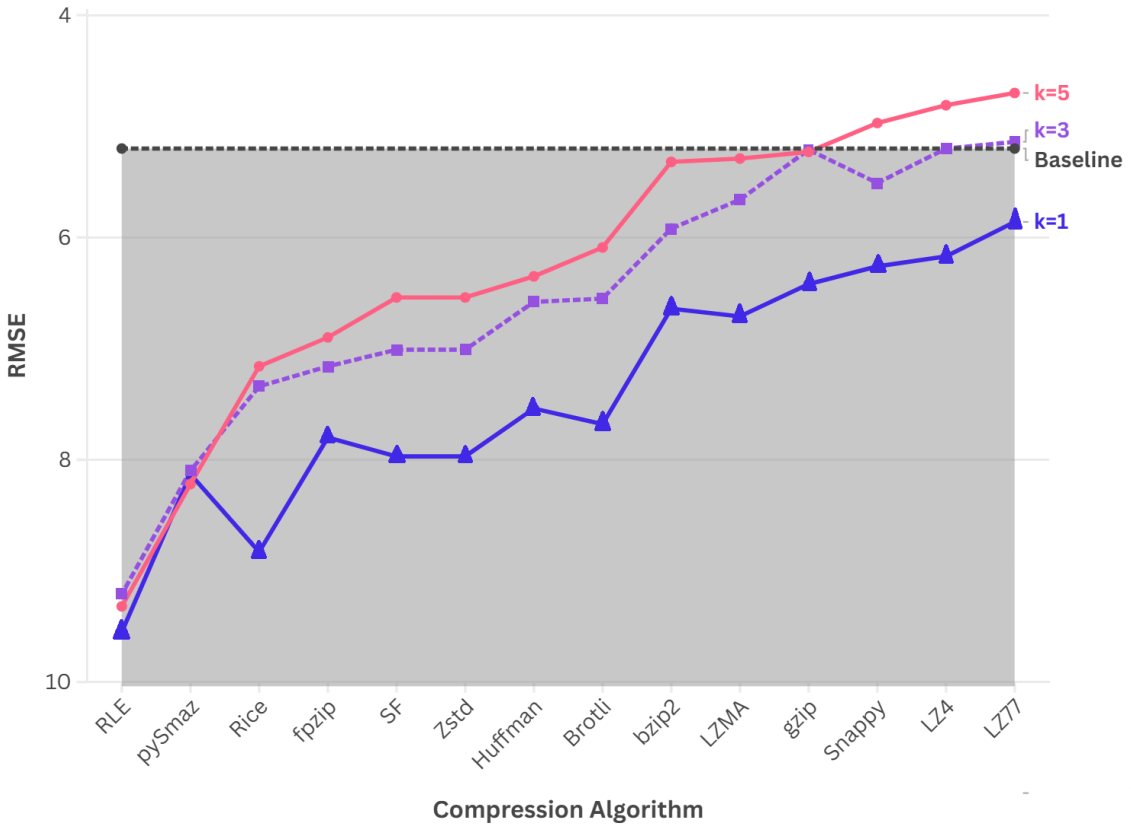
Fonte: Elaborado pelo autor (2024).

A forte correlação ($r \geq 0,7$) observada para os melhores algoritmos (LZ77, LZ4 e Snappy) no conjunto combinado confirma que o método de compressão, combinado com o kNN, é capaz de capturar eficientemente a relação entre as características textuais das transcrições e os scores do MMSE em um conjunto de dados mais abrangente.

Na Figura 11 é apresentado o desempenho dos diferentes algoritmos de compressão na tarefa de regressão para prever o MMSE dos pacientes. O gráfico mostra o RMSE para cada algoritmo, com linhas distintas representando os diferentes valores de k (1, 3 e 5) utilizados no kNN. O eixo x está ordenado de forma que os algoritmos com melhor desempenho (menores valores de RMSE) estão posicionados à direita, enquanto os de pior desempenho (maiores valores de RMSE) estão à esquerda. Analisando o gráfico, observa-se uma tendência geral de melhoria no desempenho (diminuição do RMSE) à medida que o valor de k aumenta. Isso

sugere que considerar um número maior de vizinhos próximos geralmente leva a predições mais precisas do MMSE. O algoritmo LZ77 se destaca como o mais eficaz, apresentando o menor RMSE para todos os valores de k, especialmente para 5NN. Isso corrobora os resultados anteriores e confirma a capacidade do LZ77 em capturar padrões linguísticos relevantes para a previsão do MMSE.

Figura 11 – Comparação do RMSE na avaliação LOSO no conjunto combinado para diferentes algoritmos de compressão na tarefa de regressão.



Fonte: Elaborado pelo autor (2024).

LZ4 e Snappy apresentam curvas muito próximas às do LZ77, especialmente para 3NN e 5NN, indicando que são alternativas interessantes e quase tão eficazes quanto o LZ77 para esta tarefa de regressão. Há uma clara distinção entre os algoritmos de melhor desempenho (LZ77, LZ4, Snappy) e aqueles com desempenho inferior (Rice, RLE, pySmaz). Os algoritmos de pior desempenho mantém RMSE elevado para todos os valores de k, confirmando sua inadequação para esta tarefa específica. Em alguns casos, como no algoritmo gzip, observa-se que o aumento de k de 3 para 5 não resulta em melhoria no RMSE. Isso sugere que, para esse algoritmo, 3NN pode ser um valor ótimo, equilibrando a complexidade do

modelo com a capacidade de generalização. O desempenho dos melhores algoritmos (LZ77, LZ4, Snappy) com 5NN supera o baseline de RMSE de 5,2 reportado por Luz et al. (2021), indicando uma melhoria substancial na previsão do MMSE.

Os resultados do LOSO no conjunto combinado são geralmente consistentes com as observações feitas nos conjuntos de treino e teste separadamente, reforçando a confiabilidade das conclusões sobre a eficácia dos diferentes algoritmos de compressão. A consistência no desempenho superior dos algoritmos LZ77, LZ4 e Snappy através de diferentes valores de k e em comparação com a tarefa de classificação sugere que estes são particularmente eficazes na captura de características linguísticas relevantes para a avaliação da gravidade da demência.

É interessante notar que, enquanto na tarefa de classificação o bzip2 se destacou como o melhor algoritmo, na tarefa de regressão o LZ77 assume essa posição. Isso pode indicar que diferentes aspectos das transcrições são mais relevantes para a classificação binária em comparação com a predição de valores contínuos da gravidade da demência.

4.2.3. Análise Estatística

No segundo bloco experimental, foi conduzida uma análise estatística com o objetivo de investigar as diferenças de desempenho entre distintos algoritmos de compressão. Para tal, foram utilizados dados de 14 algoritmos, cada um com três amostras correspondentes às configurações de 1NN, 3NN e 5NN. Os valores de Acc. e RMSE foram obtidos por meio da validação LOSO aplicada ao conjunto combinado de treino e teste.

A primeira etapa da análise consistiu na aplicação do teste de normalidade de Shapiro-Wilk, que revelou que os dados de Acc. para o algoritmo RLE e o RMSE para o algoritmo gzip não apresentavam distribuição normal. Em decorrência dessa não-normalidade, optou-se por aplicar o teste de Friedman, que resultou em um p-valor < 0,01 para a classificação e um p-valor < 0,01 para a regressão. Assim, a hipótese nula foi rejeitada, aceitando-se a hipótese alternativa, que sugere a existência de diferenças significativas entre os algoritmos de compressão tanto para a tarefa de classificação quanto para a de regressão.

Para identificar quais grupos de algoritmos apresentaram diferenças significativas, foi realizado o teste post-hoc de Nemenyi, que se destina a comparar pares de grupos de dados. Os resultados deste teste revelaram diferenças significativas entre os seguintes pares de algoritmos para a tarefa de classificação: LZ77 e pySmaz, além de LZ77 e RLE. Para a tarefa de regressão, as diferenças significativas foram observadas entre LZ77 e pySmaz, LZ77 e RLE, e LZ4 e RLE.

Esses resultados indicam que o algoritmo LZ77 apresenta um desempenho superior em comparação com os algoritmos pySmaz e RLE, tanto na tarefa de classificação quanto na de regressão. Isso sugere que, em termos de Acc., o LZ77 é mais eficaz para as amostras testadas, o que pode ser um fator decisivo na escolha de algoritmos para aplicações que exigem alta Prec.

4.3 BLOCO EXPERIMENTAL 3: ANÁLISE BASEADA EM *EMBEDDINGS* NLP

Na análise de dados complexos, como sinais de voz ou neuroimagem, a representação eficiente da informação é crucial. Os *embeddings* emergem como uma técnica fundamental neste contexto, permitindo a transformação de dados de alta dimensionalidade em representações vetoriais mais compactas e matematicamente tratáveis, preservando as relações semânticas relevantes entre os dados. Esta técnica é particularmente valiosa no processamento de dados biomédicos, onde a dimensionalidade elevada pode comprometer a eficiência computacional e a capacidade de generalização dos modelos.

Considerando esta fundamentação, no terceiro bloco experimental, foi realizada uma análise inovadora baseada em embeddings de texto, utilizando as transcrições de áudios de pacientes com e sem Alzheimer do segundo bloco. Este bloco experimental é particularmente relevante porque explora embeddings nunca antes utilizados para o nosso propósito específico de análise da DA. O uso de embeddings de texto, especificamente desenvolvidos para capturar a semântica e o contexto, permite identificar nuances linguísticas que os métodos tradicionais de compressão não capturam. Esta abordagem é especialmente significativa no contexto das doenças neurodegenerativas, uma vez que a linguagem é um dos primeiros domínios cognitivos afetados pela DA.

A implementação desta abordagem baseada em embeddings visa estabelecer um comparativo entre a eficácia de técnicas de representação de texto mais avançadas em relação aos métodos de compressão utilizados no segundo bloco experimental. Essa comparação não apenas valida a utilidade dos embeddings no nosso contexto, mas também fornece uma visão sobre a capacidade de diferentes representações textuais na avaliação da DA, permitindo uma análise mais abrangente e precisa das alterações linguísticas associadas à doença.

4.3.1 Método

O pré-processamento das transcrições seguiu o mesmo procedimento do bloco anterior, com a eliminação das falas dos entrevistadores e a remoção de símbolos não relevantes para a análise textual.

A geração dos *embeddings* foi realizada com base nos sete melhores modelos selecionados segundo o Massive Text Embedding Benchmark (MTEB) Leaderboard (MUENNIGHOFF et al., 2022), utilizando o critério de Classification Average em doze conjuntos de dados. Os modelos escolhidos, em ordem de desempenho, são: NV-Embed-v2 (NV2), SFR-Embedding-2_R (SFR2), bge-en-icl (BGE), stella_en_1.5B_v5 (STELLA), gte-Qwen2-7B-instruct (GTE), jina-embeddings-v3 (JINA) e Linq-Embed-Mistral (LEM). Adicionalmente, foram testados três *embeddings* populares para a classificação de textos de pacientes com e sem Alzheimer: bert-base-multilingual-uncased-sentiment (BERT-S), ernie-2.0-large-en (ERNIE) e roberta-base-sentiment-latest (RoBERTa-S). Todos os *embeddings* foram obtidos por meio da série de bibliotecas e ferramentas da Hugging Face.

Para as tarefas de classificação e regressão, foram empregados os mesmos algoritmos utilizados no primeiro bloco experimental (Cardus, Maciel e Zalewski, 2024), com algumas modificações. O LDA foi substituído pelo *Logistic Regression* (LR), dado que os *embeddings* normalmente não seguem a distribuição normal presumida pelo LDA. O DT foi substituído pelo Naive Bayes (NB) devido ao desempenho insatisfatório do DT e à popularidade do NB no uso com *embeddings* (Kadam et al., 2018; Cichosz, 2023). Além disso, o algoritmo 1NN foi incluído nesta fase.

Assim como no segundo bloco experimental, foram realizados três procedimentos de avaliação, incluindo o LOSO no conjunto de treino, o *holdout* no conjunto de teste e o LOSO no conjunto combinado de treino e teste. Estas avaliações foram aplicadas tanto para a tarefa de classificação (distinção entre pacientes com Alzheimer e controles) quanto para a tarefa de regressão (previsão do MMSE).

Para a tarefa de classificação, manteve-se a Acc. como métrica principal, junto com Rec. e Prec., permitindo uma comparação direta com os resultados dos blocos experimentais anteriores. Na tarefa de regressão, foram utilizadas as métricas de MAE, RMSE e o coeficiente de correlação r entre os valores verdadeiros e preditos.

4.3.2 Resultados e Discussão

Tarefa de Classificação: Avaliação no Conjunto de Treinamento

Os resultados do procedimento de LOSO no conjunto de treinamento para os diferentes *embeddings* e algoritmos de classificação são apresentados na Tabela 7. Nesta tabela são apresentadas as métricas de Rec., Prec. e Acc. para cada combinação de *embedding* e algoritmo de classificação. Analisando os resultados, observa-se uma variação significativa no desempenho entre os diferentes *embeddings* e algoritmos de classificação. De modo geral, os *embeddings* mais recentes e especializados, como GTE, STELLA e LEM, demonstraram um desempenho superior em comparação com *embeddings* mais tradicionais como BERT-S, ERNIE e RoBERTa-S.

Estes resultados sugerem que a escolha do *embedding* e do algoritmo de classificação tem um impacto significativo na capacidade de discriminar entre pacientes com Alzheimer e controles saudáveis. Os *embeddings* mais recentes e especializados, particularmente quando combinados com algoritmos como NN, LR, RF e SVM parecem capturar características mais relevantes nas transcrições, resultando em um melhor desempenho de classificação.

Em comparação com os resultados do segundo bloco experimental baseado em algoritmos de compressão, os *embeddings* de texto demonstraram um potencial promissor, com alguns modelos igualando a Acc. máxima de 88,89% obtida pelo algoritmo LZ77 com 5NN. Isso indica que as representações de texto baseadas em *embeddings* podem oferecer vantagens na captura de nuances linguísticas relevantes para a detecção da DA.

Tabela 7 – Resultados do procedimento LOSO no conjunto de treinamento para diferentes *embeddings* e algoritmos de classificação.

Embedding	kNN (k=1)			NB			SVM			RF			LR			NN		
	Rec.	Prec.	Acc.	Rec.	Prec.	Acc.	Rec.	Prec.	Acc.	Rec.	Prec.	Acc.	Rec.	Prec.	Acc.	Rec.	Prec.	Acc.
GTE	77,55%	77,55%	75,00%	82,50%	89,58%	88,89%	88,00%	97,73%	88,89%	88,00%	89,58%	85,19%	87,50%	93,75%	88,89%	92,00%	84,91%	84,26%
STELLA	75,00%	80,49%	73,15%	81,25%	85,42%	81,48%	88,00%	93,48%	87,04%	90,00%	83,64%	84,26%	85,00%	86,54%	85,19%	80,00%	90,20%	87,96%
LEM	78,17%	80,85%	76,85%	79,63%	79,63%	79,63%	76,58%	89,36%	84,26%	85,58%	88,46%	87,04%	84,79%	84,31%	82,41%	75,93%	87,23%	82,41%
NV2	79,58%	80,49%	73,15%	81,87%	85,42%	81,48%	87,25%	93,48%	87,04%	88,84%	83,64%	84,26%	84,75%	77,78%	77,78%	85,00%	90,20%	87,96%
JINA	77,82%	79,41%	68,52%	82,91%	88,64%	81,48%	80,00%	95,00%	83,33%	77,78%	85,71%	82,41%	85,51%	88,64%	81,48%	83,51%	84,91%	84,26%
BERT-S	72,85%	73,58%	73,15%	62,58%	60,00%	63,89%	83,52%	80,77%	79,63%	82,75%	85,11%	80,56%	84,83%	84,31%	82,41%	84,00%	85,71%	82,41%
SFR2	73,08%	66,67%	66,67%	70,25%	78,00%	75,93%	77,87%	89,36%	84,26%	77,35%	77,36%	76,85%	79,49%	66,67%	65,74%	79,63%	87,76%	84,26%
ERNIE	69,57%	70,45%	66,67%	72,50%	75,00%	74,07%	68,52%	67,92%	67,59%	67,57%	78,43%	76,85%	80,00%	82,35%	80,56%	81,48%	80,00%	80,56%
RoBERTa-S	62,58%	62,50%	62,96%	68,00%	75,00%	72,22%	77,79%	79,55%	74,07%	76,52%	68,42%	69,44%	79,25%	77,36%	76,85%	76,57%	79,25%	78,70%
BGE	64,62%	65,31%	63,89%	62,56%	66,67%	60,19%	61,74%	68,42%	56,48%	74,78%	72,55%	71,30%	78,57%	77,78%	77,78%	72,25%	73,47%	71,30%

Fonte: Elaborado pelo autor (2024).

Tarefa de Classificação: Avaliação no Conjunto de Teste (Competição)

Após a realização do procedimento de LOSO no conjunto de treinamento, procedeu-se à avaliação do desempenho dos modelos no conjunto de teste através do procedimento de *holdout*. Os resultados obtidos são apresentados na Tabela 8, destacando as métricas de Rec., Prec. e Acc. para cada combinação de *embedding* e algoritmo de classificação. Analisando os resultados apresentados na tabela, observa-se uma variação significativa no desempenho entre os diferentes *embeddings* e algoritmos de classificação no conjunto de teste, revelando padrões importantes e algumas divergências em relação ao desempenho observado no conjunto de treinamento.

O *embedding* LEM destacou-se com uma Acc. de 87,50% ao ser combinado com os algoritmos LR e NB, evidenciando bom equilíbrio na detecção de casos de Alzheimer. O ERNIE também obteve Acc. de 87,50% com o algoritmo NB, demonstrando uma boa generalização. O GTE manteve desempenho sólido no teste com o SVM, alcançando Acc. equivalente e evidenciando sua robustez.

Entre os algoritmos de classificação, o LR e o RF se destacaram no conjunto de teste, proporcionando os melhores resultados para vários *embeddings*. Isso sugere que esses algoritmos são particularmente eficazes na captura de padrões relevantes para a discriminação entre pacientes com Alzheimer e controles saudáveis, mesmo em dados não utilizados no treinamento.

Os *embeddings* mais tradicionais, como BERT-S e RoBERTa-S, apresentaram desempenhos variados no conjunto de teste. O BERT-S, por exemplo, alcançou sua melhor performance com o algoritmo LR, com uma Acc. de 79,17% e uma Prec. notável de 88,89%. Já o RoBERTa-S obteve seus melhores resultados com o RF, atingindo 81,25% de Acc.. É interessante notar que o BGE, que havia apresentado resultados mais baixos no conjunto de treinamento, mostrou uma melhoria significativa no conjunto de teste, especialmente quando combinado com o 1NN, alcançando uma Acc. de 77,08% com um bom equilíbrio entre Rec. e Prec.

Tabela 8 – Resultados do holdout no conjunto de teste para diferentes *embeddings* e algoritmos de classificação.

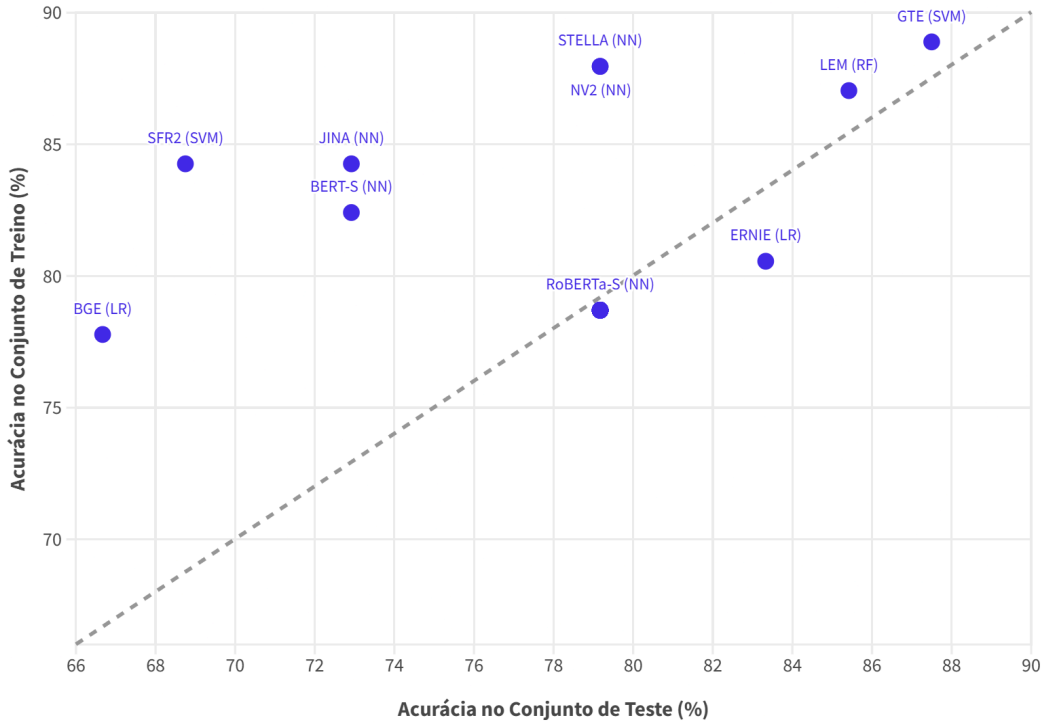
Embedding	kNN (k=1)			NB			SVM			RF			LR			NN		
	Rec.	Prec.	Acc.	Rec.	Prec.	Acc.	Rec.	Prec.	Acc.	Rec.	Prec.	Acc.	Rec.	Prec.	Acc.	Rec.	Prec.	Acc.
GTE	69,33%	72,22%	66,67%	84,72%	86,36%	83,33%	89,20%	90,91%	87,50%	79,41%	77,78%	81,25%	87,94%	90,48%	85,42%	77,67%	78,26%	77,08%
STELLA	79,17%	64,71%	60,42%	79,17%	81,82%	79,17%	78,50%	90,00%	83,33%	78,50%	79,17%	79,17%	78,50%	80,00%	81,25%	78,50%	79,17%	79,17%
LEM	62,50%	88,24%	77,08%	84,72%	90,91%	87,50%	79,17%	90,48%	85,42%	83,33%	86,96%	85,42%	85,94%	90,91%	87,50%	79,17%	86,36%	83,33%
NV2	79,79%	86,67%	72,92%	79,89%	80,95%	77,08%	82,08%	85,00%	79,17%	83,33%	86,96%	85,42%	86,67%	90,00%	83,33%	82,08%	85,00%	79,17%
JINA	81,25%	81,25%	70,83%	77,50%	79,17%	79,17%	75,00%	85,71%	81,25%	79,17%	82,61%	81,25%	82,50%	86,36%	83,33%	78,50%	76,19%	72,92%
BERT-S	68,89%	71,43%	68,75%	62,50%	52,63%	54,17%	62,50%	83,33%	75,00%	52,63%	76,47%	68,75%	86,67%	88,89%	79,17%	77,50%	86,67%	72,92%
SFR2	68,89%	73,68%	68,75%	78,50%	84,21%	77,08%	78,50%	90,91%	68,75%	83,33%	86,96%	85,42%	82,50%	88,89%	79,17%	72,50%	80,00%	75,00%
ERNIE	71,15%	73,68%	68,75%	87,50%	87,50%	87,50%	79,41%	77,78%	81,25%	84,71%	84,00%	85,42%	83,33%	83,33%	83,33%	89,20%	90,91%	87,50%
RoBERTa-S	76,00%	70,00%	75,00%	76,00%	78,95%	72,92%	75,00%	69,23%	70,83%	78,50%	80,00%	81,25%	79,17%	76,00%	77,08%	78,50%	79,17%	79,17%
BGE	68,18%	78,26%	77,08%	56,25%	46,67%	47,92%	77,08%	55,56%	52,08%	46,67%	73,91%	72,92%	55,56%	68,18%	66,67%	56,25%	68,18%	66,67%

Fonte: Elaborado pelo autor (2024).

Comparando estes resultados com os obtidos no segundo bloco experimental baseado em algoritmos de compressão, observa-se que os *embeddings* de texto mantiveram sua vantagem no conjunto de teste. O melhor resultado obtido com *embeddings* (LEM+NB, ERNIE+NB, GTE+SVM, ERNIE+NN e LEM+LR com 87,50% de Acc.) superou o melhor resultado dos algoritmos de compressão (LZ4 com 5NN, que atingiu 85,42% de Acc.). Isso reforça o potencial das representações de texto baseadas em *embeddings* na captura de nuances linguísticas relevantes para a detecção da DA.

A Figura 12 apresenta uma comparação entre os resultados obtidos no procedimento LOSO no conjunto de treino e os resultados do *holdout* no conjunto de teste, mostrando apenas o melhor classificador para cada *embedding*. Esse gráfico de dispersão oferece uma análise visual da capacidade de generalização dos diferentes *embeddings* testados, comparando a Acc. no conjunto de treino (eixo y) com a Acc. no conjunto de teste (eixo x). A linha diagonal tracejada, que indica o ponto de equivalência entre os desempenhos nos dois conjuntos, facilita a identificação de casos de *overfitting* e *underfitting*.

Figura 12 – Comparação da Acc. entre os conjuntos de treino e de teste para diferentes *embeddings* na tarefa de classificação.



Fonte: Elaborado pelo autor (2024).

Observa-se que a maioria dos *embeddings* apresenta um desempenho superior no conjunto de treino em comparação ao conjunto de teste. No entanto, a magnitude dessa diferença varia consideravelmente entre os *embeddings*. O *embedding* GTE (SVM) demonstra o melhor desempenho geral, com a maior Acc. tanto no conjunto de treino (89,89%) quanto no conjunto de teste (87,50%). Apesar de apresentar uma pequena queda no desempenho do teste, o GTE mantém uma bom resultado, indicando uma boa capacidade de generalização. Os *embeddings* LEM (RF), STELLA (NN) e NV2 (NN) se destacam pelo desempenho, com o LEM mostrando boa consistência entre treino e teste. SFR2 (SVM) e JINA (NN) apresentam desempenho intermediário com sinais de overfitting, enquanto BERT-S (NN) e ERNIE (LR) têm resultados inferiores. O BGE (LR) mostra o pior desempenho, indicando inadequação para a classificação de DA.

Em comparação com os resultados obtidos no segundo bloco experimental utilizando algoritmos de compressão, os *embeddings* de texto demonstram, em geral, uma maior variação entre os desempenhos de treino e teste. Isso sugere que os compressores podem oferecer uma representação mais generalizável das características textuais relevantes para a detecção da DA.

Tarefa de Classificação: Avaliação Geral

A Tabela 9 apresenta os resultados do procedimento LOSO aplicado ao conjunto combinado de treinamento e teste, para diferentes *embeddings* e algoritmos de classificação. Esta análise proporciona uma visão mais abrangente do desempenho dos modelos, ao considerar todos os dados disponíveis. Ao analisar os resultados, observa-se a manutenção de padrões consistentes e algumas mudanças em relação às análises anteriores. O *embedding* LEM manteve um desempenho robusto, alcançando uma Acc. de 85,90% com o algoritmo NN, o que demonstra sua capacidade de generalização em um conjunto de dados ampliado. O GTE alcançou uma Acc. de 86,54% com SVM, enquanto STELLA atingiu a maior Acc. global de 87,18% com NN e NV2 obteve 84,62% com LR. A LR se destacou, superando o baseline na maioria dos *embeddings*, exceto SFR2. Em contrapartida, o 1NN apresentou desempenho inferior, não superando o baseline em nenhum dos casos.

Tabela 9 – Resultados do procedimento LOSO no conjunto combinado (treinamento + teste) para diferentes *embeddings* e algoritmos de classificação.

Embedding	kNN (k=1)			NB			SVM			RF			LR			NN		
	Rec.	Prec.	Acc.	Rec.	Prec.	Acc.	Rec.	Prec.	Acc.	Rec.	Prec.	Acc.	Rec.	Prec.	Acc.	Rec.	Prec.	Acc.
GTE	70,26%	75,36%	72,44%	87,25%	85,14%	83,33%	85.18%	88,00%	86,54%	84,58%	83,33%	83,33%	83,33%	87,67%	85,26%	82,05%	82,05%	82,05%
STELLA	76,24%	80,70%	72,44%	80,27%	82,86%	79,49%	85,64%	89,71%	84,62%	80,77%	87,14%	83,33%	84,62%	84,62%	84,62%	87,18%	87,18%	87,18%
LEM	75,25%	80,95%	75,00%	82,05%	82,05%	82,05%	87,25%	88,73%	85,26%	78,21%	84,42%	83,97%	85,18%	85,53%	84,62%	86,25%	87,84%	85,90%
NV2	73,34%	75,36%	72,44%	48,72%	79,17%	76,92%	80,74%	82,61%	78,85%	80,77%	87,14%	83,33%	84,62%	84,62%	84,62%	82,05%	82,05%	82,05%
JINA	76,24%	78,85%	69,23%	79,25%	81,69%	78,85%	86,57%	88,06%	82,69%	79,49%	84,00%	82,69%	80,12%	86,76%	82,05%	72,14%	72,00%	71,15%
BERT-S	64,24%	63,41%	64,10%	60.17%	59,09%	62,82%	80,74%	82,61%	78,85%	80,77%	81,94%	79,49%	78,85%	78,85%	78,85%	76,46%	76,25%	76,92%
SFR2	69,24%	70,42%	68,59%	78,26%	79,17%	76,92%	42,19%	24,53%	32,69%	70,78%	80,00%	78,85%	70,17&	70,27%	69,23%	78,14%	81,33%	80,13%
ERNIE	71,12%	73,13%	69,87%	80,25%	80,26%	79,49%	81,24%	80,72%	82,69%	79,08%	79,22%	78,85%	83,92%	84,42%	83,97%	82,78%	83,12%	82,69%
RoBERTa-S	64,24%	64,86%	64,10%	73,73%	75,00%	71,79%	76,54%	77,78%	75,64%	73,46%	73,42%	73,72%	80,12%	80,52%	80,13%	76,54%	77,50%	78,21%
BGE	67,76%	69,86%	68,59%	60,34%	64,44%	58,33%	65,18%	68,00%	61,54%	74,18%	73,75%	74,36%	74,36%	76,92%	76,92%	77,26%	77,22%	77,56%

Fonte: Elaborado pelo autor (2024).

Comparando os resultados atuais com os obtidos anteriormente, verifica-se uma ligeira queda nas Acc. máximas em comparação com as obtidas no conjunto de treinamento isolado, o que era esperado, uma vez que o conjunto combinado inclui dados de teste, tornando a tarefa de classificação mais desafiadora.

Observa-se também um melhor equilíbrio entre a Rec. e Prec. para a maioria das combinações de *embedding* e algoritmo. Este equilíbrio reflete uma melhoria na capacidade dos modelos em identificar corretamente tanto os casos de Alzheimer quanto os controles saudáveis. O destaque é para a combinação STELLA com NN, que alcançou uma Acc. equilibrada de 87,18%, demonstrando um bom equilíbrio entre sensibilidade e especificidade. *Embeddings* mais tradicionais, como BERT-S e RoBERTa-S, apresentaram desempenho inferior em comparação com os *embeddings* mais recentes e especializados, o que reforça a importância de utilizar representações textuais mais avançadas e adaptadas para tarefas específicas, como a detecção da DA.

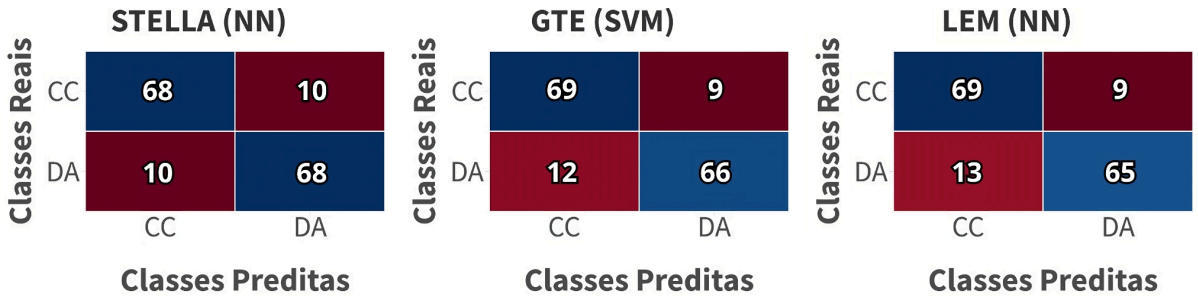
Ao comparar estes resultados com os obtidos no segundo bloco experimental, que utilizou algoritmos de compressão, nota-se que os melhores resultados obtidos com a combinação STELLA+NN (87,18% de Acc.) são ligeiramente superiores aos resultados dos algoritmos de compressão, como o bzip2 com 5NN, que alcançou 85,26% de Acc.

Os *embeddings* proporcionam uma maior diversidade de combinações com bom desempenho, oferecendo mais opções para a seleção de modelos. A variabilidade no desempenho entre diferentes *embeddings* e algoritmos é maior do que a observada entre os algoritmos de compressão, sugerindo que a escolha adequada da combinação *embedding*-algoritmo é fundamental para otimizar os resultados.

A análise dos resultados obtidos com a aplicação de *embeddings* na classificação de pacientes com Alzheimer e controles, conforme ilustrado nas matrizes de confusão, revela informações importantes sobre o desempenho dos três melhores algoritmos: STELLA (NN), GTE (SVM) e LEM (NN). As matrizes de confusão, representadas na Figura 13, detalham o número de VN, FP, FN e VP

gerados por cada algoritmo após a validação cruzada LOSO aplicada ao conjunto combinado de treino e teste.

Figura 13 – Matrizes de Confusão para os *embeddings* STELLA (NN), GTE (SVM) e LEM (NN).



Fonte: Elaborado pelo autor (2024).

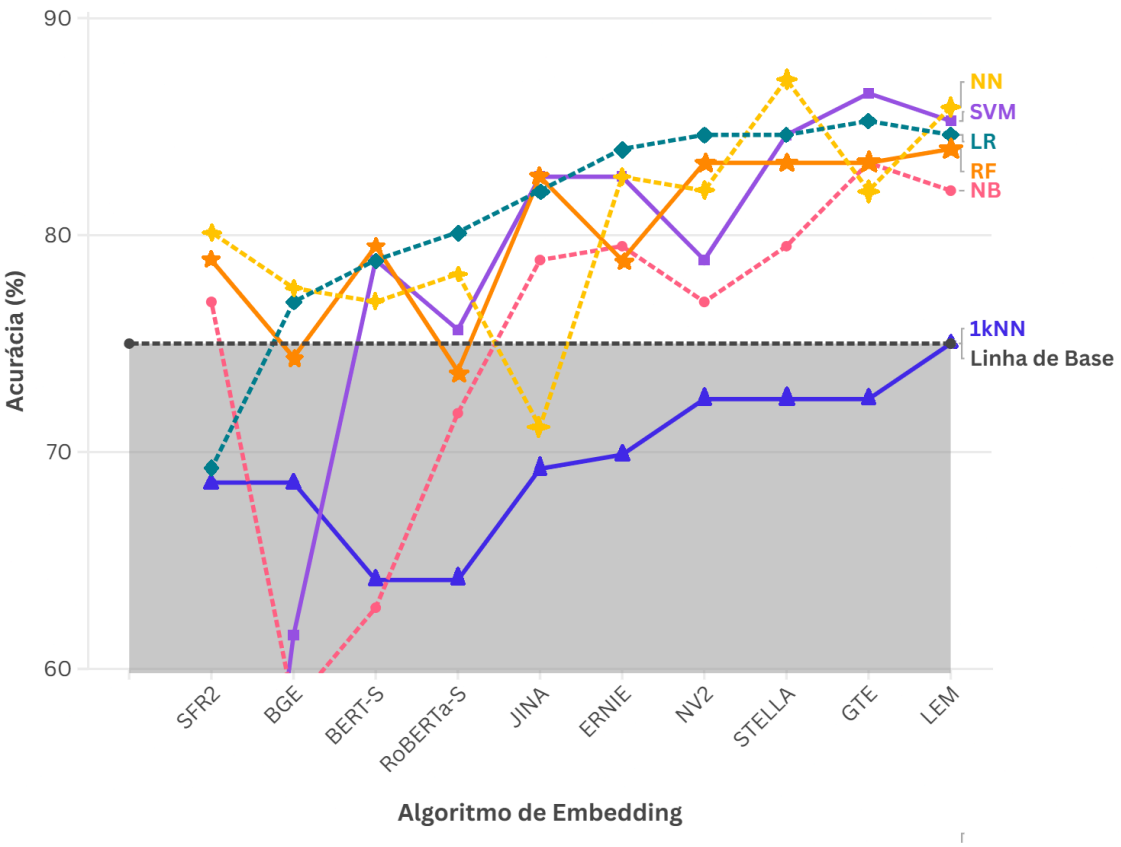
O *embedding* STELLA com NN se destaca por apresentar o menor número de FN (10) entre os três modelos. Comparando estes resultados com os obtidos no segundo bloco experimental utilizando algoritmos de compressão, observa-se uma melhoria geral no desempenho. Por exemplo, o melhor modelo baseado em compressão (LZ4) apresentou 65 VN, 12 FP, 13 FN e 66 VP, enquanto o STELLA (NN) alcança 68 VN, 10 FP, 10 FN e 68 VP, demonstrando uma redução nos erros de classificação tanto para controles quanto para pacientes com Alzheimer. O STELLA (NN) oferece o melhor equilíbrio entre a minimização de falsos positivos e falsos negativos, tornando-se uma opção atraente para maximizar a Prec. global do sistema de classificação. O GTE (SVM) e o LEM (NN) também apresentam resultados sólidos, com uma leve vantagem na identificação de controles, mas um pequeno aumento nos falsos negativos em comparação com o STELLA.

Na Figura 14, é apresentado um gráfico comparativo do desempenho dos diferentes embeddings e algoritmos de classificação no procedimento LOSO aplicado ao conjunto combinado de dados (treinamento + teste). O eixo x está ordenado de forma que os algoritmos com pior desempenho (menores valores de acurácia) estão posicionados à esquerda, enquanto os de melhor desempenho (maiores valores de acurácia) estão à direita. A análise do gráfico revela tendências significativas e variações importantes no desempenho dos diferentes embeddings e algoritmos de classificação. O baseline para a tarefa de classificação foi estabelecido em 75% de acurácia, representado pela linha tracejada no gráfico.

Em geral, observa-se que a maioria das combinações de *embeddings* e algoritmos superou o baseline, indicando a eficácia geral da abordagem baseada em *embeddings* para a detecção de Alzheimer a partir de transcrições de fala. No entanto, há uma variabilidade considerável no desempenho entre diferentes *embeddings* e algoritmos.

Os *embeddings* mais recentes e especializados, como STELLA, GTE e LEM, demonstraram um desempenho superior em comparação com *embeddings* mais tradicionais como BERT-S e RoBERTa-S. O *embedding* STELLA, em particular, alcançou o pico de desempenho com o algoritmo NN, atingindo o melhor resultado global observado. Entre os algoritmos de classificação, o NN, SVM e LR se destacaram, proporcionando os melhores resultados para vários *embeddings*. O RF também demonstrou um bom desempenho geral, embora ligeiramente inferior aos três mencionados anteriormente. O NB, apesar de superar o baseline na maioria dos casos, geralmente apresentou desempenho inferior aos outros algoritmos.

Figura 14 – Desempenho dos *embeddings* na tarefa de classificação para diferentes algoritmos.



Fonte: Elaborado pelo autor (2024).

É interessante notar o desempenho consistentemente baixo do algoritmo 1NN para todos os *embeddings*, raramente superando o baseline. Isso contrasta fortemente com os resultados obtidos no segundo bloco experimental com algoritmos de compressão, onde o kNN com valores maiores de k (3 e 5) apresentou bom desempenho. Essa discrepância sugere que, no contexto de *embeddings*, considerar apenas o vizinho mais próximo não é suficiente para capturar a complexidade das relações entre as amostras no espaço de alta dimensionalidade dos *embeddings*.

Os *embeddings* tradicionais, como BERT-S e RoBERTa-S, geralmente apresentaram desempenho inferior em comparação com os *embeddings* mais recentes e especializados. No entanto, é notável que mesmo esses *embeddings*, quando combinados com algoritmos adequados como NN e SVM, ainda foram capazes de superar significativamente o baseline.

O *embedding* SFR2 apresentou o desempenho mais baixo entre todos os testados, com Acc. abaixo do baseline para a metade dos algoritmos. Isso sugere que este *embedding* pode não ser adequado para a tarefa específica de classificação de Alzheimer a partir de transcrições de fala.

A variabilidade no desempenho entre diferentes *embeddings* e algoritmos é maior do que a observada entre os algoritmos de compressão, sugerindo que a escolha adequada da combinação *embedding*-algoritmo é fundamental para otimizar os resultados. Esta variabilidade também oferece mais opções para a seleção de modelos, permitindo um equilíbrio entre desempenho e outras considerações práticas, como tempo de processamento ou interpretabilidade do modelo.

Tarefa de Regressão: Avaliação no Conjunto de Treinamento

Na análise baseada em *embeddings* de texto, foram avaliados diferentes algoritmos de regressão aplicados aos *embeddings* gerados pelos modelos selecionados. Na Tabela 10, são apresentados os resultados obtidos no procedimento LOSO no conjunto de treino para cada combinação de *embedding* e algoritmo de regressão.

Tabela 10 – Resultados do Procedimento LOSO no Conjunto de Treinamento para Diferentes Combinações de Embeddings e Algoritmos de Regressão.

Embedding	kNN (k=1)			SVM			LR			NN			RF			NB		
	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r
LEM	5,08	7,59	0,47	5,51	7,23	0,54	5,23	6,61	0,49	4,86	6,17	0,54	4,99	6,23	0,51	4,66	5,95	0,56
NV2	6,34	8,92	0,23	5,89	7,67	0,49	5,23	6,70	0,50	4,35	5,79	0,60	4,79	5,97	0,59	4,45	5,81	0,59
GTE	5,41	8,07	0,37	5,52	7,30	0,57	5,75	7,18	0,43	4,92	6,36	0,53	5,00	6,17	0,53	4,48	5,93	0,56
SFR2	6,57	8,96	0,31	6,09	8,04	0,35	4,88	6,52	0,54	5,01	6,10	0,56	5,16	6,32	0,48	4,42	5,78	0,59
JINA	6,43	9,04	0,27	5,45	7,21	0,54	6,93	8,69	0,37	4,69	6,39	0,53	4,63	5,79	0,62	4,48	6,03	0,55
BGE	6,65	9,23	0,20	6,12	8,00	0,05	5,57	7,22	0,07	5,05	6,48	0,54	5,12	6,26	0,49	4,89	6,12	0,52
BERT-S	6,10	8,66	0,31	5,50	7,26	0,42	6,21	7,73	0,42	5,51	7,56	0,38	5,01	6,20	0,51	4,96	6,10	0,53
ERNIE	6,27	8,92	0,26	6,05	8,01	0,32	5,19	6,69	0,53	6,46	8,27	-0,13	5,09	6,25	0,51	4,39	5,70	0,61
STELLA	5,99	8,56	0,27	5,49	7,28	0,59	6,69	8,07	0,41	7,00	8,44	0,39	4,77	6,04	0,55	4,36	5,74	0,60
RoBERTa-S	7,28	9,75	0,22	5,97	7,76	0,33	6,63	8,48	0,34	6,05	7,96	0,14	5,36	6,41	0,46	4,85	6,04	0,54

Fonte: Elaborado pelo autor (2024).

Pela análise dos resultados, é possível observar uma variação considerável no desempenho dos diferentes *embeddings* e algoritmos de classificação. É importante notar que nenhum dos modelos conseguiu superar o baseline de RMSE de 5,2 relatado por Luz et al. (2021), ficando todos com RMSE maior que 5,2. Entre os *embeddings* avaliados, o LEM apresentou um desempenho relativamente melhor em comparação com os outros, especialmente quando combinado com o algoritmo NB. Para esta combinação, obteve-se um MAE de 4,66, um RMSE de 5,95 e um r de 0,56. O *embedding* NV2 também demonstrou um desempenho competitivo, particularmente quando usado com o algoritmo NN, alcançando um MAE de 4,35, RMSE de 5,79 e r de 0,60.

É interessante notar que o desempenho dos algoritmos variou significativamente dependendo do *embedding* utilizado. Por exemplo, o 1NN e o SVM apresentaram resultados consistentemente inferiores em comparação com outros algoritmos, independentemente do *embedding*. O algoritmo NB mostrou um desempenho relativamente estável entre os diferentes *embeddings*, com RMSE variando entre 5,70 e 6,12. Isso sugere que o NB pode ser menos sensível às variações nas representações de texto geradas pelos diferentes modelos de *embedding*. Entre os *embeddings* populares testados o BERT-S apresentou resultados intermediários, com seu melhor desempenho alcançado com o algoritmo RF (MAE de 4,96, RMSE de 6,10 e r de 0,53).

Embora nenhum dos modelos tenha superado o *baseline*, alguns se aproximaram mais do que outros. Isso sugere que, com ajustes adicionais ou a exploração de outras técnicas de *embedding* ou classificação, pode ser possível melhorar esses resultados. Esses resultados indicam que, apesar do potencial dos *embeddings* de texto na análise de transcrições para a avaliação da demência, ainda há desafios significativos a serem superados para alcançar um desempenho superior ao dos métodos convencionais.

Tarefa de Regressão: Avaliação no Conjunto de Teste (Competição)

Após a avaliação dos modelos no conjunto de treino através do procedimento LOSO, prosseguimos com a análise do desempenho desses modelos no conjunto de teste, utilizando o procedimento de *holdout*. Na Tabela 11 são apresentados os resultados obtidos nesta análise para cada combinação de

embedding e algoritmo de regressão. Como podemos observar, os resultados da avaliação *holdout* no conjunto de teste revelam padrões interessantes e algumas diferenças em relação ao desempenho observado no conjunto de treino. O *embedding* GTE demonstrou um desempenho robusto no conjunto de teste, especialmente quando combinado com o algoritmo RF. Esta combinação alcançou o menor RMSE de 4,77, um MAE de 3,89 e um r de 0,66. É notável que este resultado supera o baseline de RMSE de 5,2 relatado por Luz et al. (2021), indicando uma melhoria na previsão do MMSE.

O algoritmo NB apresentou um desempenho consistentemente bom entre os diferentes *embeddings* no conjunto de teste, com RMSE variando entre 4,90 e 5,66. Isso reforça a observação anterior de que o NB pode ser menos sensível às variações nas representações de texto geradas pelos diferentes modelos de *embeddings.*

A correlação moderada a forte (r > 0,6) observada para as melhores combinações de embeddings e algoritmos no conjunto de teste confirma que esta abordagem é capaz de capturar eficientemente a relação entre as características textuais das transcrições e os scores do MMSE, mesmo em dados não utilizados no treinamento.

Estes resultados sugerem que a utilização de *embeddings* modernos, como o GTE e o NV2, combinados com algoritmos de aprendizado de máquina adequados, pode oferecer melhorias significativas na previsão do MMSE em pacientes com Alzheimer. No entanto, a variabilidade nos resultados entre diferentes *embeddings* e algoritmos ressalta a importância de uma seleção cuidadosa e possivelmente uma abordagem de *ensemble* para otimizar o desempenho em aplicações práticas.

Tabela 11 – Resultados do Procedimento de Holdout no Conjunto de Teste para Diferentes Combinações de *embeddings* e Algoritmos de Regressão.

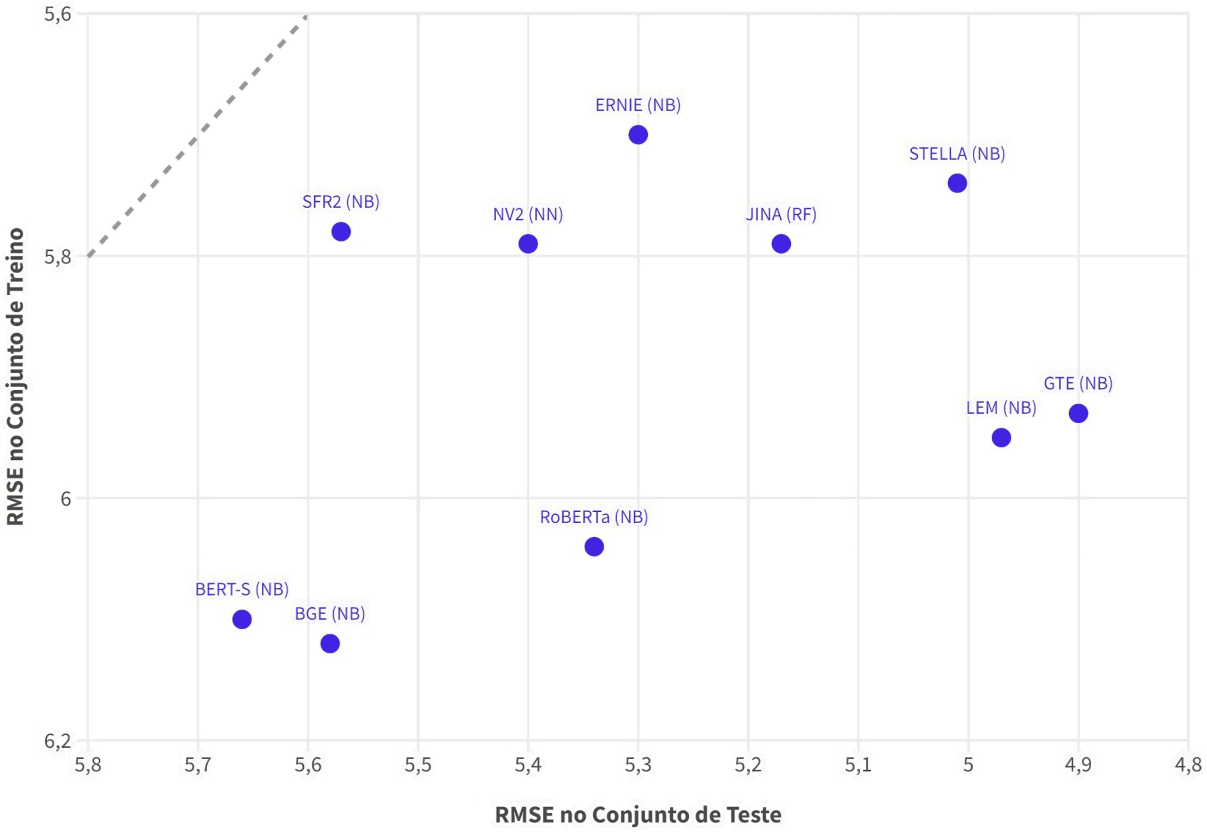
Embedding	kNN (k=1)			SVM			LR			NN			RF			NB		
	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r
LEM	5,15	7,36	0,42	4,17	5,63	0,67	4,28	5,78	0,56	3,96	5,32	0,59	4,12	5,20	0,54	3,76	4,97	0,60
NV2	4,06	5,76	0,56	4,48	6,09	0,58	4,82	6,61	0,51	3,73	5,40	0,56	3,86	5,01	0,58	3,57	5,13	0,58
GTE	4,21	6,46	0,52	4,16	5,73	0,67	4,52	6,10	0,57	4,36	5,74	0,57	3,89	4,77	0,66	3,57	4,90	0,61
SFR2	4,60	6,52	0,45	4,80	6,54	0,45	5,05	6,92	0,34	4,15	5,29	0,51	4,37	5,41	0,50	3,93	5,57	0,47
JINA	4,94	7,20	0,35	3,98	5,56	0,64	6,63	8,82	0,16	4,14	6,02	0,45	4,07	5,17	0,54	3,95	5,53	0,48
BGE	4,88	7,25	0,55	4,92	6,54	0,07	4,94	6,51	0,34	5,64	7,20	0,31	4,85	5,98	0,31	4,43	5,58	0,45
BERT-S	6,04	8,17	0,35	4,85	6,04	0,31	5,95	7,94	0,22	5,33	6,71	0,14	4,63	5,86	0,37	4,45	5,66	0,41
ERNIE	5,17	7,02	0,25	4,64	6,38	0,64	5,17	6,55	0,41	5,21	6,71	-0,13	4,73	5,63	0,49	4,52	5,30	0,56
STELLA	5,88	8,23	0,16	4,28	5,79	0,59	5,48	6,99	0,54	5,93	7,41	0,48	4,02	5,01	0,63	3,85	5,01	0,60
RoBERTa-S	6,19	8,54	0,22	4,75	6,21	0,32	5,80	7,65	0,43	4,71	6,61	0,17	4,53	5,38	0,53	4,20	5,34	0,53

Fonte: Elaborado pelo autor (2024).

A Figura 15 representa uma visualização da comparação do desempenho de cada combinação dos conjuntos de treinamento e de teste. A linha diagonal tracejada indica o ponto de equivalência entre os desempenhos nos dois conjuntos, facilitando a identificação de casos de overfitting e underfitting. Analisando o gráfico, podemos observar que todos os pontos se encontram abaixo da linha diagonal, indicando que todos os modelos apresentaram um desempenho melhor no conjunto de teste do que no conjunto de treino. Isso é um comportamento inesperado, pois os modelos tendem a se ajustar melhor aos dados que foram utilizados durante o treinamento.

O *embedding* GTE combinado com o algoritmo NB se destaca como o melhor, apresentando o menor RMSE no conjunto de teste. O LEM e STELLA, também com NB, apresentam desempenhos competitivos, mas ligeiramente inferiores. Por outro lado, os *embeddings* ERNIE e NV2 mostram resultados similares, mas são os menos destacados em relação aos demais, indicando uma performance menos robusta.

Figura 15 – Comparação do RMSE entre conjuntos de treino e teste para diferentes *embeddings* e algoritmos na tarefa de regressão.



Fonte: Elaborado pelo autor (2024).

A predominância do algoritmo NB entre as melhores performances sugere que este método pode ser particularmente adequado para lidar com as representações de texto geradas pelos *embeddings* na tarefa de previsão do MMSE. No entanto, a variabilidade observada entre diferentes *embeddings* ressalta a importância de uma seleção cuidadosa da combinação *embedding*-algoritmo para otimizar o desempenho em aplicações práticas.

Tarefa de Regressão: Avaliação Geral

Após a análise dos resultados nos conjuntos de treinamento e teste de forma isolada, realizamos uma avaliação mais abrangente utilizando o procedimento LOSO no conjunto combinado de treinamento e teste. A Tabela 12 ilustra os resultados desta análise para cada combinação de *embedding* e algoritmo de regressão.

Observam-se padrões interessantes e algumas discrepâncias em relação ao desempenho observado nos conjuntos isolados. Os resultados obtidos no conjunto combinado indicam uma tendência de valores de RMSE superiores em comparação com os resultados anteriores, o que é esperado devido ao aumento na variabilidade dos dados.

O *embedding* STELLA destacou-se por demonstrar um bom desempenho, apresentando o menor RMSE (5,15) quando combinado com o algoritmo RF. O NB emergiu como o algoritmo mais consistente entre os diferentes *embeddings*, frequentemente produzindo os menores valores de RMSE e MAE. Tal resultado corrobora as observações anteriores sobre a estabilidade do NB em relação às variações nas representações de texto.

Diferentemente dos resultados anteriores, poucas combinações de *embedding* e algoritmos foram capazes de superar o baseline de RMSE de 5,2 relatado por Luz et al. (2021). Isso sugere que a tarefa de regressão se torna mais desafiadora quando aplicada a um conjunto de dados mais amplo e diversificado.

Tabela 12 – Resultados do procedimento LOSO no conjunto combinado (treinamento + teste) para diferentes *embeddings* e algoritmos na tarefa de regressão.

Embedding	kNN (k=1)			SVM			LR			NN			RF			NB		
	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r
LEM	5,12	7,19	0,49	4,41	5,78	0,56	4,90	6,39	0,60	4,82	6,10	0,52	4,55	5,24	0,57	4,21	5,50	0,59
NV2	5,12	7,31	0,37	4,04	5,59	0,58	5,00	6,88	0,52	5,03	6,66	0,51	4,34	5,43	0,59	4,00	5,50	0,59
GTE	4,75	7,29	0,45	4,68	6,00	0,55	4,89	6,62	0,63	5,34	6,64	0,50	4,45	5,46	0,59	3,96	5,42	0,58
SFR2	5,97	8,48	0,29	4,96	6,02	0,53	5,69	7,60	0,36	5,11	6,77	0,47	4,62	5,80	0,55	4,28	5,73	0,55
JINA	5,73	8,40	0,27	4,18	6,00	0,54	4,71	6,42	0,60	7,45	9,69	0,26	4,30	5,62	0,58	4,01	5,49	0,49
BGE	5,99	7,91	0,27	5,12	6,40	0,54	5,65	7,53	0,15	5,03	6,85	0,21	4,77	5,69	0,55	4,49	5,68	0,56
BERT-S	6,71	8,82	0,23	5,32	7,32	0,34	5,30	7,06	0,42	6,30	7,97	0,34	4,74	6,01	0,49	4,72	5,90	0,51
ERNIE	5,46	7,88	0,31	5,20	6,72	0,21	4,74	6,42	0,61	5,95	7,94	0,22	4,90	5,85	0,50	4,75	5,79	0,54
STELLA	4,11	5,39	0,60	6,47	7,93	0,43	5,50	7,31	0,58	5,90	8,51	0,22	4,51	5,15	0,64	4,20	5,40	0,60
RoBERTa-S	6,74	9,16	0,21	5,38	7,30	0,14	5,30	6,82	0,34	6,22	8,10	0,39	4,98	5,79	0,50	4,80	5,90	0,50

Fonte: Elaborado pelo autor (2024).

As correlações de Pearson observadas foram geralmente moderadas, variando entre 0,5 e 0,6 para as melhores combinações. Este fato indica que, embora exista uma relação consistente entre as características textuais capturadas pelos *embeddings* e os scores do MMSE, tal relação é menos forte do que a observada nos conjuntos isolados.

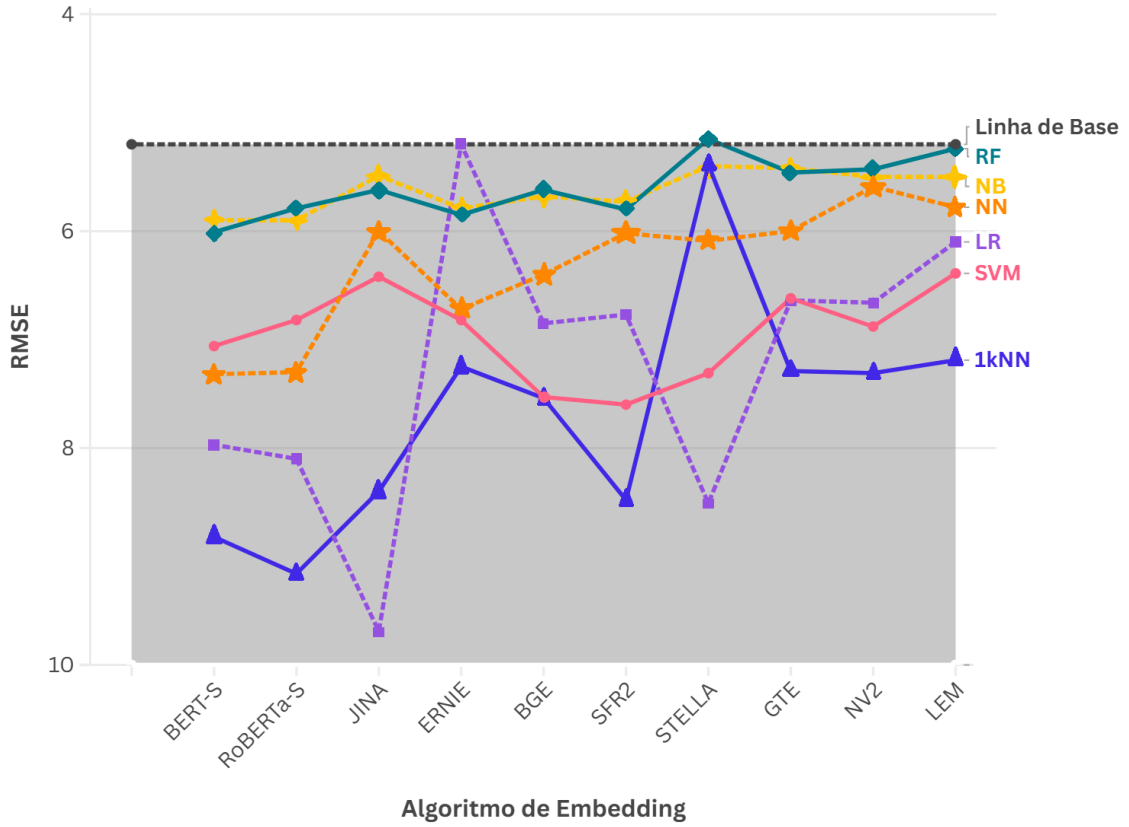
Entre os *embeddings* populares testados, o ERNIE apresentou um desempenho médio, com seu melhor resultado (RMSE de 5,79) alcançado em combinação com o algoritmo NB. Observou-se uma variabilidade significativa no desempenho dos diferentes algoritmos para cada *embedding*. Por exemplo, para o *embedding* LEM, o RMSE variou de 5,24 (RF) a 7,19 (1NN), ressaltando a importância da escolha apropriada do algoritmo de regressão.

O algoritmo 1NN consistentemente apresentou os piores resultados para todos os *embeddings*, sugerindo que abordagens mais simples de kNN podem não ser adequadas para capturar a complexidade da relação entre as características textuais e o MMSE em um conjunto de dados mais amplo.

Embora o desempenho geral tenha sido ligeiramente inferior ao observado nos conjuntos isolados, a consistência de determinados *embeddings* (como LEM, NV2 e GTE) e a eficácia do algoritmo NB reforçam a validade da abordagem baseada em *embeddings* para a previsão do MMSE em pacientes com Alzheimer.

Na Figura 16, observa-se uma comparação visual do desempenho dos diferentes algoritmos de regressão aplicados aos diversos embeddings no conjunto combinado de treinamento e teste. O eixo x está ordenado de forma que os algoritmos com pior desempenho (maiores valores de RMSE) estão posicionados à esquerda, enquanto os de melhor desempenho (menores valores de RMSE) estão à direita. O gráfico mostra o RMSE para cada combinação de embedding e algoritmo de regressão, com linhas coloridas representando os diferentes algoritmos utilizados. A linha pontilhada horizontal representa a linha de base de RMSE de 5,2, conforme relatado por Luz et al. (2021), servindo como referência para avaliar a eficácia dos modelos testados.

Figura 16 – Comparação do Desempenho dos Algoritmos de Regressão Aplicados aos *embeddings* na Previsão do MMSE.



Fonte: Elaborado pelo autor (2024).

Ao analisar o gráfico, observam-se várias tendências significativas. A maioria das combinações de *embedding*-algoritmo apresenta RMSE abaixo da linha de base, indicando que superar o benchmark estabelecido é um desafio considerável. Há uma notável variação no desempenho entre os diferentes *embeddings*; por exemplo, o LEM e o NV2 demonstram um desempenho mais consistente entre os algoritmos, enquanto outros, como JINA e ERNIE, exibem maior variabilidade.

Os algoritmos RF e NB destacam-se como os mais consistentes, frequentemente apresentando os menores valores de RMSE para a maioria dos *embeddings*. Isso corrobora nossas observações anteriores sobre a estabilidade do NB em relação às variações nas representações de texto. O *embedding* STELLA, quando combinado com o algoritmo RF, apresenta o melhor desempenho geral, sendo o único ponto no gráfico que se encontra abaixo da linha de base. Essa combinação específica sugere uma eficácia particular na captura de características linguísticas relevantes para a previsão do MMSE.

O algoritmo 1NN consistentemente apresenta os piores resultados para todos os *embeddings*, confirmando nossas observações anteriores sobre sua inadequação para esta tarefa específica. Entre os *embeddings* populares testados (BERT-S, ERNIE, RoBERTa-S), o ERNIE parece ter o melhor desempenho geral, especialmente quando combinado com NB ou RF.

Os *embeddings* LEM, NV2 e GTE demonstram uma performance relativamente boa e consistente entre os diferentes algoritmos, reforçando sua eficácia na captura de características linguísticas relevantes para a avaliação da gravidade da demência.

Ao comparar estes resultados com os obtidos no segundo bloco experimental, que utilizou algoritmos de compressão, observamos algumas diferenças notáveis. No segundo bloco, algoritmos de compressão (LZ77, LZ4, Snappy) com 5NN superaram a linha de base. Em contraste, no terceiro bloco, apenas uma combinação conseguiu esse feito.

Além disso, no segundo bloco, o aumento do valor de k no kNN geralmente levava a melhores resultados. No terceiro bloco, não observamos uma tendência clara de melhoria para um algoritmo específico, com RF e NB apresentando um desempenho mais consistente. Os resultados com *embeddings* também demonstram maior variabilidade entre as diferentes combinações de *embedding*-algoritmo em comparação com os resultados dos algoritmos de compressão.

4.3.3 Análise Estatística

A mesma análise do bloco experimental 2 foi realizada para avaliar os diferentes *embeddings* do bloco 3. O teste de Friedman apresentou um p-valor < 0,01 para os classificadores e p-valor < 0,01 para os regressores. Para a tarefa de classificação, os embeddings LEM e RoBERTa-S, LEM e BERT-S, LEM e BGE, LEM e SFR2, GTE e RoBERTa-S, GTE e BERT-S, GTE e BGE, GTE e SFR2, e STELLA e BGE mostraram-se estatisticamente diferentes entre si. Da mesma forma, na tarefa de regressão, os embeddings LEM e BGE, e LEM e SFR2 também apresentaram diferenças significativas, conforme confirmado pelo teste de Nemenyi.

Esses resultados indicam que as diferentes abordagens de embeddings têm um impacto significativo no desempenho dos classificadores e regressores utilizados, sugerindo que a escolha do embedding pode influenciar diretamente a eficácia do modelo em tarefas de classificação e regressão. A superioridade de determinados *embeddings* sobre outros implica que a seleção cuidadosa do método de representação de dados é crucial para a otimização do desempenho em aplicações práticas.

Essas descobertas são relevantes para a comunidade científica e industrial, pois destacam a importância da investigação sobre *embeddings* em modelos de aprendizado de máquina. O entendimento das diferenças de desempenho entre *embeddings* pode guiar pesquisadores e profissionais na escolha de técnicas apropriadas, contribuindo para o avanço das aplicações de aprendizado de máquina em diversas áreas, como processamento de linguagem natural e análise de dados.

5 ANÁLISE COMPARATIVA DOS RESULTADOS

5.1 ANÁLISE INTEGRADA DA TAREFA DE CLASSIFICAÇÃO

A análise integrada dos três blocos experimentais realizados no presente estudo, todos efetuados no conjunto de teste (*holdout*), fornece uma visão abrangente sobre o desempenho dos diferentes métodos empregados na tarefa de predição da DA. Cada bloco abordou uma estratégia distinta, desde a aplicação de seletores de características acústicas até a utilização de *embeddings* para representar as características textuais das transcrições.

No primeiro bloco experimental, foram extraídas características acústicas das amostras de fala do conjunto ADReSS 2020, utilizando três conjuntos de atributos amplamente utilizados: eGeMAPS, ComParE e emobase. Após a extração, técnicas de seleção de características como SUC, ERC, SFM e SSC foram aplicadas, retendo os 10%, 20% e 30% melhores atributos. Diversos algoritmos de aprendizado de máquina foram testados, e o melhor desempenho foi obtido pela combinação de NN e a técnica SFM, com uma Acc. de 79,17% utilizando o conjunto emobase e 20% dos atributos selecionados. Embora os algoritmos NN e RF tenham apresentado os melhores resultados em vários cenários, com destaque também para RF em ComParE (77,08%), alguns algoritmos, como LDA e DT, apresentaram desempenho mais modesto, sugerindo sensibilidade à seleção de características e à complexidade dos dados.

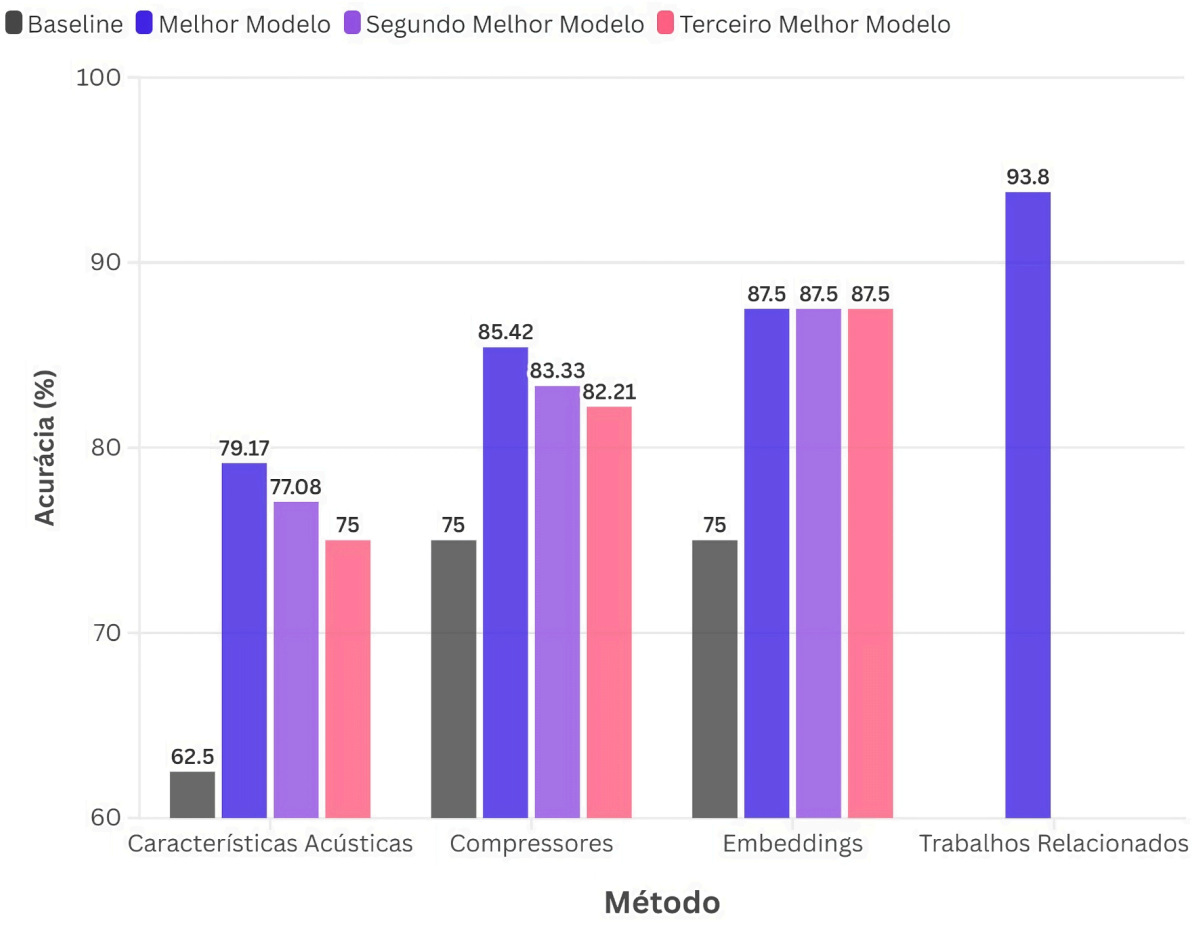
No segundo bloco experimental, foi realizada uma análise utilizando diversos algoritmos de compressão na classificação de pacientes com Alzheimer e controles. O algoritmo LZ4 destacou-se como o melhor modelo, apresentando uma Acc. de 85,42% com o classificador 5NN, evidenciando sua eficácia em identificar casos de Alzheimer. O gzip obteve a segunda maior Acc., com 83,33%, demonstrando um equilíbrio notável entre as classes. Em seguida, o algoritmo bzip2 e Snappy apresentaram uma Acc. de 81,25%, reforçando sua competitividade na tarefa de classificação. Esses resultados indicaram que os algoritmos LZ4, gzip, bzip2 e Snappy foram particularmente eficazes para a classificação de Alzheimer, superando outros métodos.

No terceiro bloco experimental, foi realizada uma avaliação dos modelos de classificação aplicados ao conjunto de teste, utilizando o procedimento de holdout. Entre os diversos *embeddings* e algoritmos testados, destacaram-se os três melhores desempenhos. O *embedding* LEM, quando combinado com os classificadores de Regressão Logística (LR) e Naive Bayes (NB), apresentou uma Acc. notável de 87,50%, demonstrando eficácia na detecção de casos de Alzheimer. De maneira similar, o *embedding* ERNIE também alcançou uma Acc. de 87,50% ao ser utilizado com o classificador NB, evidenciando sua capacidade de generalização em relação aos dados não vistos. Por fim, o *embedding* GTE, quando combinado com o classificador de Máquinas de Vetores de Suporte (SVM), obteve uma Acc. sólida, igualmente de 87,50%, confirmando sua robustez na identificação de pacientes com Alzheimer em comparação aos controles. Esses resultados ressaltam a eficácia dos *embeddings* em captar nuances linguísticas essenciais para o diagnóstico da Doença de Alzheimer, superando o desempenho dos algoritmos de compressão previamente avaliados.

A Figura 17 ilustra de forma comparativa os resultados obtidos nos três blocos experimentais, evidenciando uma progressão na Acc. dos métodos empregados. Notavelmente, todos os melhores modelos superaram seus respectivos *baselines*, com os *embeddings* apresentando a menor variabilidade entre modelos e as maiores Acc. globais, todas acima de 85%. Os três modelos apresentados no gráfico correspondem aos três melhores desempenhos para cada abordagem. Esta análise visual corrobora a eficácia crescente das abordagens utilizadas, desde métodos acústicos até representações textuais mais sofisticadas, na detecção da DA.

Observa-se que as características acústicas e os compressores apresentaram maior variabilidade em comparação com os *embeddings*. A cada bloco experimental, é possível notar uma melhoria em relação ao anterior, refletindo uma evolução na capacidade de detecção. Os *embeddings*, por sua vez, demonstraram o melhor desempenho geral entre as abordagens analisadas.

Figura 17 – Comparação dos Resultados dos Blocos Experimentais: Métodos Acústicos, Algoritmos de Compressão e *embeddings* na Detecção da DA.



Fonte: Elaborado pelo autor (2024).

Na revisão bibliográfica, o melhor resultado foi uma Acc. de 93,8%, obtido por Wang et al. (2022b), que utilizaram *embeddings* BERT e RoBERTa combinados com uma SVM como classificador, aplicando fusão de modelos em texto transcrito via reconhecimento automático de fala (ASR). De forma semelhante, Martinc et al. (2021) também alcançaram uma Acc. de 93,8%, empregando *embeddings* como *Bag-of-n-grams* e utilizando um *ensemble* com *k-means* e *Random Forest*, otimizados por recursos de representação ativa de dados (ADR). Observa-se que os modelos de melhor desempenho apresentam maior complexidade em comparação com as nossas abordagens.

5.2 ANÁLISE INTEGRADA DA TAREFA DE REGRESSÃO

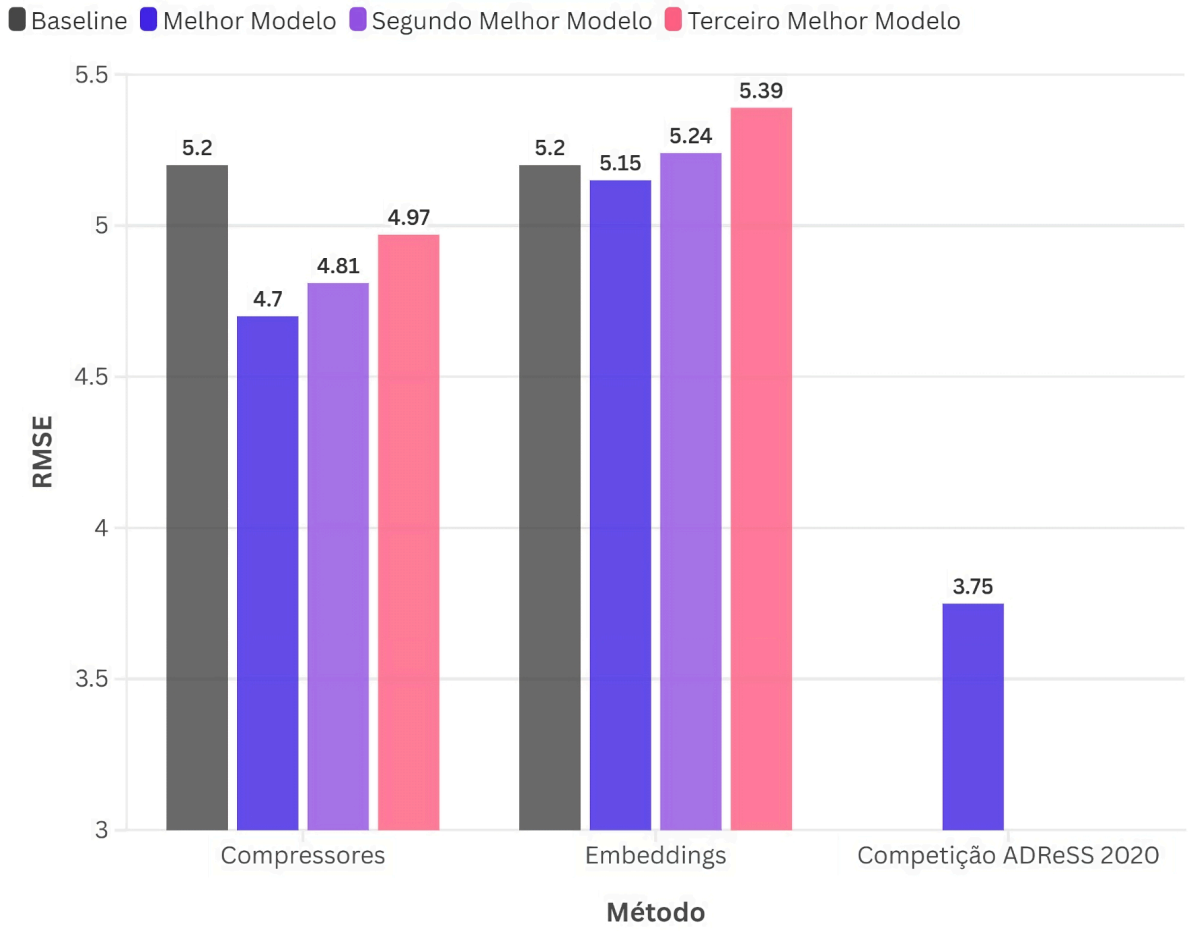
A análise integrada da tarefa de regressão foi realizada com o intuito de avaliar o desempenho dos diferentes algoritmos de compressão e *embeddings* na previsão do MMSE para pacientes com DA. As avaliações foram conduzidas utilizando o procedimento de LOSO, abrangendo um conjunto de dados mais completo que combina as amostras de treinamento e teste. Essa abordagem visa fornecer uma compreensão mais aprofundada da eficácia dos modelos implementados.

No segundo bloco experimental, os resultados evidenciaram que o algoritmo LZ77 apresentou o melhor desempenho, alcançando um RMSE de 4,70 ao empregar o classificador 5NN, o que indica uma notável eficácia na captura de padrões relevantes para a previsão do MMSE. Em seguida, os algoritmos LZ4 e Snappy demonstraram desempenhos sólidos, com RMSEs de 4,81 e 4,97, respectivamente, também utilizando o classificador 5NN. Esses resultados corroboram a eficácia dos métodos de compressão selecionados na tarefa de regressão. Em contraste, algoritmos como gzip, Rice, pySmaz e RLE exibiram desempenhos inferiores, com RMSEs elevados, sugerindo sua inadequação para a previsão do MMSE nesta análise.

No terceiro bloco experimental, os resultados revelaram que o *embedding* STELLA, em combinação com o algoritmo RF, apresentaram o melhor desempenho, alcançando um RMSE de 5,15, o que indica uma eficácia significativa na previsão do MMSE. Em segundo lugar, o *embedding* LEM, também utilizando RF, destacou-se com um RMSE de 5,24, corroborando sua relevância nas análises. Embora o algoritmo 1kNN tenha demonstrado uma consistência inferior, frequentemente gerando os maiores valores de RMSE entre as diversas combinações, obteve um desempenho notável ao utilizar o *embedding* STELLA, com um resultado registrado em 5,39. Esses achados ressaltam a importância da escolha do *embedding* e do algoritmo de regressão na tarefa de previsão do MMSE, evidenciando que o desempenho pode variar significativamente conforme as combinações utilizadas.

A Figura 18 ilustra, de forma comparativa, as diferenças de desempenho entre os métodos empregados na predição do MMSE. O gráfico apresenta o RMSE para cada método, onde valores mais baixos indicam melhor performance.

Figura 18 – Comparação dos Resultados dos Blocos Experimentais: Métodos de Compressão e *embeddings* em Algoritmos de Regressão



Fonte: Elaborado pelo autor (2024).

Os melhores modelos do segundo bloco experimental superaram notavelmente a linha de base, evidenciando uma tendência de melhoria no RMSE à medida que a complexidade dos métodos aumenta. No entanto, no terceiro bloco, apenas o modelo com melhor desempenho conseguiu ultrapassar o *baseline*, destacando a dificuldade em superar essa referência com nossa abordagem. O melhor modelo da Competição ADReSS 2020 destacou-se, apresentando um RMSE de 3,75 na tarefa de regressão do MMSE, alcançado por Koo et al. (2020), que utilizou uma abordagem multimodal combinando características acústicas e textuais extraídas de redes pré-treinadas com uma arquitetura modificada de rede neural convolucional recorrente (CRNN). Esse resultado reforça a importância de abordagens mais avançadas e integradas na predição do MMSE para a detecção da Doença de Alzheimer.

6 CONSIDERAÇÕES FINAIS

Este estudo propôs uma abordagem multidimensional para a detecção e predição da DA, combinando análises acústicas e textuais da fala espontânea com métodos inovadores, como algoritmos de compressão e *embeddings* de última geração. A aplicação dessas técnicas demonstrou grande potencial no diagnóstico precoce, oferecendo uma perspectiva mais completa e eficiente para o monitoramento da doença

Os resultados obtidos, com Acc. superiores a 85%, indicam que tanto os métodos baseados em compressão quanto os baseados em *embeddings* são comparáveis em eficácia. A robustez dos *embeddings*, como STELLA e GTE, se destacou por capturar melhor as nuances textuais associadas à DA. Por outro lado, os algoritmos de compressão, como bzip2, também foram eficientes na diferenciação entre pacientes com DA e controles, mostrando que métodos computacionalmente eficientes, como o uso de NCD, são alternativas promissoras. Essa eficácia comparável sugere que é possível adotar soluções acessíveis e escaláveis, como implementações em plataformas móveis ou online, possibilitando triagens em larga escala e monitoramento remoto da progressão da DA.

Adicionalmente, a análise comparativa dos métodos demonstrou uma progressão na Acc. ao longo dos experimentos, com os *embeddings* apresentando a menor variabilidade e as maiores Acc. Apesar disso, observou-se que a ausência de ajustes específicos nos *embeddings* pré-treinados pode ter limitado o desempenho máximo desses modelos, especialmente na predição do MMSE. Além disso, a análise estatística mostrou que é possível otimizar a eficiência dos modelos sem comprometer significativamente o desempenho, o que pode levar a soluções mais interpretáveis e adaptáveis.

6.1 IMPLICAÇÕES PARA A DETECÇÃO E PREDIÇÃO DA DOENÇA DE ALZHEIMER

O uso de aprendizado de máquina na análise de dados de fala espontânea tem o potencial de transformar a triagem da DA, sobretudo em contextos com acesso limitado a especialistas. Essa metodologia pode ser adaptada para dispositivos móveis, viabilizando monitoramento remoto e acessível, além de contribuir para uma triagem inicial em larga escala. A possibilidade de personalizar os diagnósticos para diferentes perfis de pacientes amplia as opções de análise, enquanto a eficácia comprovada de métodos alternativos, como compressão e *embeddings*, sugere que abordagens multimodais podem enriquecer o diagnóstico da DA, permitindo uma análise mais completa dos dados.

6.2 LIMITAÇÕES DO ESTUDO

Apesar dos resultados promissores, o estudo apresenta algumas limitações importantes. A primeira está relacionada ao conjunto de dados do Desafio ADReSS 2020, que embora balanceado demograficamente, não representa uma ampla diversidade de condições cognitivas e origens étnico-culturais. Essa limitação restringe a generalização dos resultados para populações mais amplas e heterogêneas. Além disso, o ruído de fundo e as sobreposições de vozes presentes em algumas amostras acústicas comprometeram a qualidade dos dados e, consequentemente, a Acc. dos modelos baseados em características sonoras.

Outra limitação envolve a falta de ajustes específicos nos *embeddings* utilizados. A aplicação de *embeddings* pré-treinados sem personalização para a tarefa específica de predição de Alzheimer pode ter limitado o desempenho máximo desses modelos, especialmente durante a tarefa de regressão. A inclusão de técnicas de ajuste fino para capturar as nuances linguísticas associadas à progressão da demência pode ser uma abordagem mais eficiente.

6.3 PROPOSTAS PARA PESQUISAS FUTURAS

Pesquisas futuras podem explorar a utilização de técnicas de *ensemble*, combinando diferentes modelos para aumentar ainda mais a Acc. e a robustez das predições. A personalização de *embeddings* para tarefas específicas, como a detecção da DA, pode ser uma área promissora, assim como o uso de arquiteturas de aprendizado profundo, que poderiam capturar de forma mais precisa os padrões linguísticos e acústicos associados à demência.

Além disso, é crucial ampliar a base de dados com mais diversidade de amostras, tanto em termos demográficos quanto em relação a diferentes estágios de progressão da DA. A inclusão de dados de outros idiomas e culturas também pode ajudar a avaliar a generalização dos modelos propostos. Seria interessante comparar o desempenho das técnicas apresentadas em outros conjuntos de dados, como o ADReSS 2021 e o Pitt Corpus. A inclusão desses datasets, pode fornecer informações valiosas sobre a generalização dos modelos e sua aplicação em cenários mais amplos.

Finalmente, uma análise mais aprofundada sobre a integração de dados multimodais (combinação de fala, texto, imagens cerebrais e outros biomarcadores) poderia oferecer uma visão mais holística da progressão da doença, melhorando a Acc. e possibilitando a criação de sistemas mais completos e personalizados de monitoramento da DA.

Esses avanços têm o potencial de revolucionar a maneira como a DA é diagnosticada e monitorada, oferecendo ferramentas mais acessíveis, precisas e não invasivas, capazes de contribuir de forma significativa para o manejo da doença em uma escala global.

REFERÊNCIAS

ABUKURI, Daniel Naawenkangua. Novel Biomarkers for Alzheimer's Disease: Plasma Neurofilament Light and Cerebrospinal Fluid. **International Journal of Alzheimer's Disease**, v. 2024, n. 1, p. 6668159, 2024.

AHANGER, Ab Basit et al. Alzhinet: an explainable self-attention based classification model to detect Alzheimer from 3D volumetric MRI data. **International Journal of System Assurance Engineering and Management**, p. 1-10, 2024.

ALPAYDIN, Ethem. Introduction to machine learning. **MIT press**, 2020.

ALZHEIMER'S DISEASE INTERNATIONAL. World Alzheimer Report 2024: Global changes in attitudes to dementia. London, England: **Alzheimer's Disease International**, 2024.

BERTOLA, Laiss et al. Prevalence of dementia and cognitive impairment no dementia in a large and diverse nationally representative sample: the ELSI-Brazil study. **The Journals of Gerontology**: Series A, v. 78, n. 6, p. 1060-1068, 2023.

BLUMENFELD, Jessica et al. Cell type-specific roles of APOE4 in Alzheimer disease. **Nature Reviews Neuroscience**, v. 25, n. 2, p. 91-110, 2024.

CARDUS, J. C. R.; MACIEL, J. N.; ZALEWSKI, W. Predição da demência por meio da fala usando algoritmos de seleção de atributos e aprendizado de máquina. In: **Congresso de Engenharias e Ciências Aplicadas das Três Fronteiras (MEC3F)**, 5., 2024, Foz do Iguaçu: UNILA, 2024. p. 1-10.

Cichosz. Bag of words and embedding text representation methods for medical article classification. **International Journal of Applied Mathematics and Computer Science**, 33(4), 603-621. 2023.

COHEN, Andrew R.; VITÁNYI, Paul MB. Normalized compression distance of multisets with applications. **IEEE transactions on pattern analysis and machine intelligence**, v. 37, n. 8, p. 1602-1614, 2014

DEVI, Gayatri. A how-to guide for a precision medicine approach to the diagnosis and treatment of Alzheimer's disease. **Frontiers in Aging Neuroscience**, v. 15, p. 1213968, 2023.

DONG, Mengjin et al. Comparison of Focal Hippocampal T2-Weighted MRI and Whole-Brain T1-Weighted MRI for Detection of Longitudinal Atrophy Using Deep Learning-Based and Conventional Approaches. **Alzheimer's & Dementia**, v. 19, p. e083228, 2023.

EL NAQA, Issam; MURPHY, Martin J. What is machine learning?. **Springer International Publishing**, 2015

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**. 2010. p. 1459-1462.

FERRETTI, Ceres et al. An assessment of direct and indirect costs of dementia in Brazil. **PLoS One**, v. 13, n. 3, p. e0193209, 2018.

F. Eyben et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," in **IEEE Transactions on Affective Computing**, vol. 7, no. 2, pp. 190-202, 1 April-June 2016, doi: 10.1109/TAFFC.2015.2457417.

GAUBERT, Fanny; BORG, Céline; CHAINAY, Hanna. Decision-Making in Alzheimer's Disease: The Role of Working Memory and Executive Functions in the Iowa Gambling Task and in Tasks Inspired by Everyday Situations. **Journal of Alzheimer's Disease**, v. 90, n. 4, p. 1793-1815, 2022.

HARDY, John A.; HIGGINS, Gerald A. Alzheimer's disease: the amyloid cascade hypothesis. **Science**, v. 256, n. 5054, p. 184-185, 1992.

IVANOVA, Olga; MARTÍNEZ-NICOLÁS, Israel; GARCÍA MEILÁN, Juan José. Speech changes in old age: Methodological considerations for speech-based discrimination of healthy ageing and Alzheimer's disease. **International Journal of Language & Communication Disorders**, v. 58, n. 1, p. 12-30, 2023.

JACK JR, Clifford R. et al. NIA-AA research framework: toward a biological definition of Alzheimer's disease. **Alzheimer's & dementia**, v. 14, n. 4, p. 535-562, 2018.

JAMIESON, Maggie et al. Carers: the navigators of the maze of care for people with dementia—a qualitative study. **Dementia**, v. 15, n. 5, p. 1112-1123, 2016.

JIANG, Zhiying et al. “Low-resource” text classification: A parameter-free classification method with compressors. In: **Findings of the Association for Computational Linguistics: ACL 2023**. 2023. p. 6810-6828.

KADAM, GALA, GEHLOT, KURUP, GHAG. Word embedding based multinomial naive bayes algorithm for spam filtering. **In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-5)**. IEEE. 2018, August.

KADHIM, Thair A. et al. A Review of Alzheimer’s Disease and Emerging Patient Support Systems. **2023 20th International Multi-Conference on Systems, Signals & Devices (SSD)**. IEEE, p. 379-386, 2023.

KALTSA, Maria et al. Eliciting evidence on linguistic perspectives for Subjective Cognitive Impairment, Mild Cognitive Impairment and Alzheimer’s Disease: A cross‐sectional study. **Alzheimer's & Dementia**, v. 19, p. e076053, 2023.

KOO, Junghyun et al. Exploiting multi-modal features from pre-trained networks for Alzheimer's dementia recognition. **arXiv preprint arXiv**:2009.04070, 2020.

LUZ, Saturnino et al. Alzheimer's dementia recognition through spontaneous speech. **Frontiers in computer science**, v. 3, p. 780169, 2021.

MARTINC, Matej et al. Temporal integration of text transcripts and acoustic features for Alzheimer's diagnosis based on spontaneous speech. **Frontiers in Aging Neuroscience**, v. 13, p. 642647, 2021.

MARTINC, Matej; POLLAK, Senja. Tackling the ADReSS Challenge: A Multimodal Approach to the Automated Recognition of Alzheimer's Dementia. In: **Interspeech**. 2020. p. 2157-2161.

MCKHANN, Guy M. et al. The diagnosis of dementia due to Alzheimer’s disease: Recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer's disease. **Alzheimer's & dementia**, v. 7, n. 3, p. 263-269, 2011.

MELO, S. C.; CHAMPS, A. P. S.; GOULART, R. F.; MALTA, D. C.; PASSOS, V. M. A. Dementias in Brazil: increasing burden in the 2000-2016 period. Estimates from the Global Burden of Disease Study 2016. **Arquivos de Neuropsiquiatria**, v. 78, n. 12, p. 762-771, dez. 2020.

MONTINE, Thomas J. et al. National Institute on Aging–Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease: a practical approach. **Acta neuropathologica**, v. 123, p. 1-11, 2012.

MORRONE, Christopher Daniel et al. Proteostasis failure exacerbates neuronal circuit dysfunction and sleep impairments in Alzheimer's disease. **Molecular Neurodegeneration**, v. 18, n. 1, p. 27, 2023.

MUENNIGHOFF, Niklas; TAZI, Nouamane; MAGNE, Loïc; REIMERS, Nils. MTEB: Massive Text Embedding Benchmark. **arXiv preprint arXiv**:2210.07316, 2022.

NICHOLS, Emma et al. Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. **The Lancet Neurology**, v. 18, n. 1, p. 88-106, 2019.

NICHOLS, Emma et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. **The Lancet Public Health**, v. 7, n. 2, p. e105-e125, 2022.

PAN, Yilin et al. Using the Outputs of Different Automatic Speech Recognition Paradigms for Acoustic-and BERT-Based Alzheimer's Dementia Detection Through Spontaneous Speech. In: **Interspeech**. 2021. p. 3810-3814.

PETERS-FOUNSHTEIN, Gregory et al. Lost in space (s): Multimodal neuroimaging of disorientation along the Alzheimer's disease continuum. **Human Brain Mapping**, v. 45, n. 4, p. e26623, 2024.

PEZZOLI, Stefania et al. A multimodal neuroimaging and neuropsychological study of visual hallucinations in Alzheimer's disease. **Journal of Alzheimer's Disease**, v. 89, n. 1, p. 133-149, 2022.

POLIN, Clément et al. Repetitive Behaviors in Alzheimer's Disease: A Systematic Review and Meta-Analysis. **Journal of Alzheimer's Disease**, n. Preprint, p. 1-15, 2023.

REHAN, SOLISCH, BLAZHKOVA, SUSŁOW, SZWED, SZCZĘSNA. Advancing Alzheimer's Diagnosis: The Role of AI-A Review. **Journal of Education, Health and Sport**, 77, 57093-57093. 2025.

RIZZI, Liara; AVENTURATO, Ítalo Karmann; BALTHAZAR, Marcio LF. Neuroimaging research on dementia in brazil in the last decade: scientometric analysis, challenges, and peculiarities. **Frontiers in Neurology**, v. 12, p. 640525, 2021.

ROHANIAN, Morteza; HOUGH, Julian; PURVER, Matthew. Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs. **arXiv preprint arXiv**:2106.15684, 2021.

SARAWGI, Utkarsh et al. Multimodal inductive transfer learning for detection of Alzheimer's dementia and its severity. **arXiv preprint** arXiv:2009.00700, 2020.

SCHEFFELS, Jannik F. et al. The influence of age, gender and education on neuropsychological test scores: Updated clinical norms for five widely used cognitive assessments. **Journal of Clinical Medicine**, v. 12, n. 16, p. 5170, 2023.

SCHULLER, Björn et al. Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge. **Computer Speech & Language**, v. 53, p. 156-180, 2019.

SÖRENSEN, Silvia; CONWELL, Yeates. Issues in dementia caregiving: effects on mental and physical health, intervention strategies, and research needs. **The American Journal of Geriatric Psychiatry**, v. 19, n. 6, p. 491-496, 2011.

SYED, Zafi Sherhan et al. Tackling the ADRESSO Challenge 2021: The MUET-RMIT System for Alzheimer's Dementia Recognition from Spontaneous Speech. In: **Interspeech**. 2021. p. 3815-3819.

THERRIAULT, Joseph et al. Biomarker-based staging of Alzheimer disease: rationale and clinical applications. **Nature Reviews Neurology**, v. 20, n. 4, p. 232-244, 2024.

VERMA, N. Revolutionizing Dementia Management: Insights and Progress from the Asian Pacific Region. **Acta Scientific Medical Sciences**, v. 8, n. 2, p. 63-72, fev. 2024.

WANG, Tianzi et al. Conformer Based Elderly Speech Recognition System for Alzheimer’s Disease Detection. **Proc. Interspeech**, 2022a (Grenoble: ISCA), 4825–4829.

WANG, Yi et al. Exploring linguistic feature and model combination for speech recognition based automatic ad detection. **arXiv preprint arXiv**:2206.13758, 2022b.

WEBERS, Alessandra; HENEKA, Michael T.; GLEESON, Paul A. The role of innate immune responses and neuroinflammation in amyloid accumulation and progression of Alzheimer's disease. **Immunology and cell biology**, v. 98, n. 1, p. 28-41, 2020.

YE, Zi et al. Development of the cuhk elderly speech recognition system for neurocognitive disorder detection using the dementiabank corpus. In: **ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. IEEE, 2021. p. 6433-6437.

YUAN, Jiahong et al. Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease. In: **Interspeech**. 2020. p. 2162-6.