Identificação de Esforço Cognitivo com Auxílio de Dispositivos Vestíveis e Inteligência Artificial

Mateus Antonio Franceschina[1], **Felipe André Zeiser**[1,2,3], **Cristiano André da Costa**[2], **Adriana Vial Roehe** [3], **Mateus Henrique Zeiser**[4]

[1]Universidade Comunitária da Região de Chapecó
Chapecó, SC – Brasil

[2]Software Innovation Laboratory - SOFTWARELAB
Applied Computing Graduate Program
Universidade do Vale do Rio dos Sinos (UNISINOS)
São Leopoldo, Brasil

[3]Departamento de Patologia e Medicina Legal
Universidade Federal de Ciências da Saúde de Porto Alegre
Porto Alegre, Brasil

[4]Instituto de Matemática e Estatística,
Universidade de São Paulo,
São Paulo, SP, Brasil

Abstract. *Cognitive effort can be detected through psychophysiological signals such as PPG, EDA, and EEG. This work explores two approaches: feature engineering with SVM, KNN, and GBDT, and end-to-end learning with CNN, FCN, LSTM, and ResNet. Using data from three volunteers, we evaluated model generalization. Results show that commercial wearables like the Samsung Galaxy Watch 4 can match clinical devices such as the Empatica E4 in cognitive effort detection (73% accuracy, AUC 0.698 vs. 74.3% and 0.696). These findings support the use of low-cost devices for mental state monitoring and emotion detection.*

Resumo. *O esforço cognitivo pode ser identificado por meio de sinais psicofisiológicos como PPG, EDA e EEG. Este trabalho aplica duas abordagens: engenharia de características com SVM, KNN e GBDT, e aprendizado de ponta a ponta com CNN, FCN, LSTM e ResNet. Os modelos foram avaliados com dados de três voluntários para verificar sua capacidade de generalização. Os resultados mostram que dispositivos vestíveis comerciais, como o Samsung Galaxy Watch 4, podem obter desempenho semelhante ao de equipamentos clínicos, como o Empatica E4 (acurácia de 73% e AUC de 0,698 vs. 74,3% e 0,696). Esses achados reforçam o potencial de dispositivos acessíveis no monitoramento mental e na detecção de emoções.*

1. Introdução

Também referenciado como esforço ou carga mental, o esforço cognitivo é definido como uma construção multidimensional, a qual representa a carga de uma tarefa imposta ao

sistema cognitivo de um indivíduo [Longo et al. 2022]. Atualmente, existem quatro formas principais para estimar o esforço cognitivo, sendo elas: Avaliação subjetiva; Performance da tarefa; Comportamento; e Métricas psicofisiológicas [Paas et al. 2003]. No entanto, conforme relatado por [Fleming et al. 2023], grande parte da pesquisa sobre estes métodos de mensuração do esforço cognitivo, até o presente momento, dedicou-se a formas que dependem da realização de alguma tarefa específica ou profissionais qualificados, algo inviável para realização em escala ou de modo automático por qualquer indivíduo. Como solução para este problema, e devido a maior disponibilidade computacional, métricas psicológicas ganham força no uso em detecção de esforço cognitivo. Estas métricas consistem em mudanças inconscientes no indivíduo, como movimento dos olhos, dilatação da pupila, resposta galvânica da pele, medidas de eletromiografia e eletroencefalografia, variação de batimentos cardíacos entre outros para identificação do esforço cognitivo [Cinaz 2013].

Neste cenário, o presente trabalho propõem um sistema capaz de utilizar dados de sensores instalados em dispositivos vestíveis de qualidade comercial, responsáveis pela coleta de métricas psicofisiológicas, para a inferência do estado mental do indivíduo em esforço cognitivo ou não. Para realizar essa identificação, são utilizadas tecnologias de inteligência artificial, tanto com algoritmos tradicionais de classificação quanto com modelos de aprendizado profundo, os quais permitem formar um vínculo entre medidas fisiológicas do usuário e o esforço cognitivo, para classificar se no presente momento o sujeito executa alguma tarefa mentalmente exigente. Desta forma, as principais contribuição do trabalho são:

- Desenvolvimento de modelos para detecção de esforço cognitvo usando dispositivos vestíveis comerciais.
- Proposição de um framework aplicável a outras áreas, como classificação de estados emocionais e processamento de sinais psicofisiológicos.

O artigo está dividido em seis seções principais. Na Seção 2, são discutidos os trabalhos relacionados. A Seção 3 descreve os materiais e métodos. Na Seção 4, são apresentados os resultados e a discussão. Finalmente, a Seção 5 destaca as principais conclusões e orientações para trabalhos futuros.

2. Trabalhos Relacionados

Historicamente, a identificação do esforço cognitivo baseava-se em avaliações subjetivas. Com os avanços tecnológicos, o uso de sinais psicofisiológicos — como frequência cardíaca, condutividade da pele e EEG — passou a ser explorado.

Em [Shu et al. 2020], 25 voluntários assistiram a vídeos curtos com diferentes emoções, enquanto sinais de PPG eram coletados por um bracelete. Usaram normalização e algoritmos tradicionais (com LOO), destacando-se os baseados em árvores.

[Borisov et al. 2021] propuseram um sistema vencedor da competição CogLoad@UbiComp 2020, usando PPG de bracelete para classificar esforço cognitivo binário em 23 voluntários. Aplicaram extração de estatísticas simples e usaram GBDT em uma validação *k-fold*.

Já [Ding et al. 2020] utilizaram dispositivos de alta precisão para capturar EDA, EMG e ECG de 18 voluntários realizando tarefas mentais, aplicando filtros, transformadas wavelet e validação cruzada com diversos algoritmos.

Embora tragam contribuições importantes, os estudos possuem limitações: ausência de foco direto no esforço cognitivo ([Shu et al. 2020]), tamanho reduzido e baixa taxa de amostragem dos dados ([Borisov et al. 2021]), e uso de sensores clínicos caros e inviáveis para aplicações cotidianas ([Ding et al. 2020]). Esses fatores comprometem a generalização, escalabilidade e aplicabilidade dos modelos propostos.

3. Materiais e Métodos

O desenvolvimento do modelo de inteligência artificial para detecção de esforço cognitivo envolve quatro etapas principais, apresentadas na Figura 1. Inicialmente, define-se o conjunto de dados, incluindo os dispositivos e sensores utilizados, garantindo compatibilidade com dispositivos vestíveis amplamente disponíveis. Em seguida, os sinais coletados são processados em dois fluxos distintos: um que aplica engenharia de características para extração de informações relevantes e redução de ruídos, e outro que utiliza aprendizado de máquina de ponta a ponta, empregando redes neurais diretamente nos sinais brutos. Na fase de treinamento, o primeiro fluxo utiliza algoritmos tradicionais *support vector machine* (SVM), *k-nearest neighbors* (KNN) e *gradient boosting decision tree* (GBDT), enquanto o segundo adota redes neurais do tipo *convolutional neural network* (CNN), *long-Short term memory* (LSTM), Residual Network (ResNet) e *fully convolutional network* (FCN), selecionadas por sua eficácia na classificação de estados mentais. Ambos os fluxos passam por validação *leave-one-subject-out cross validation* (LOSOCV) e são avaliados com as métricas de desempenho F1-score, precisão, sensibilidade, acurácia e Área sob a curva (AUC). O diagrama na Figura 1 demonstra as etapas executas em sequência.

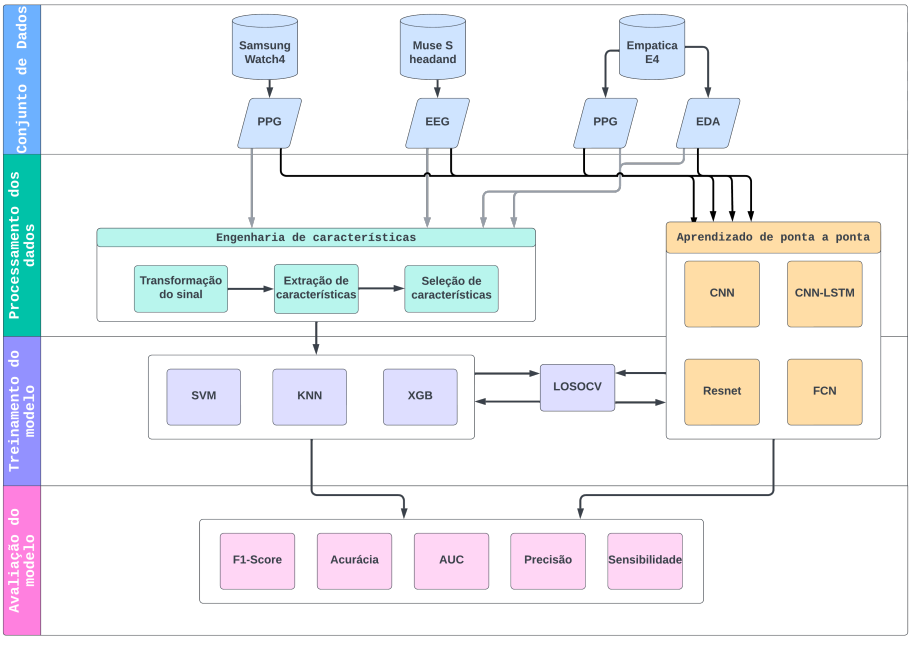


Figura 1. Visão da estrutura elaborada

3.1. Dataset

O conjunto de dados CogWear, desenvolvido por [Grzeszczyk et al. 2023], visa avaliar a viabilidade de sensores vestíveis comerciais na detecção de esforço cognitivo. Ele é

composto por dois grupos: um piloto, utilizado para testes iniciais de coleta com 11 voluntários durante o teste de *Stroop*, e um principal, onde 13 participantes realizaram o mesmo teste antes de responder questionários gamificados para analisar o impacto da gamificação no esforço cognitivo. Os dados foram coletados simultaneamente por três dispositivos: *Empatica E4* (clínico, para controle), *Muse S EEG headband* e *Samsung Galaxy Watch4*. Tendo em vista que o objetivo é elaborar um modelo capaz de detectar esforço cognitivo em dados comuns, os quais podem facilmente ser obtidos através de algum dispositivo popular, o foco deste trabalho foi no aparelho *Galaxy Watch4*, por tratar-se do cenário mais comum de dispositivo vestível disponível. Mas todos os aparelhos serão utilizados a fim de comparação.

3.2. Indicadores coletados

Os sensores em cada dispositivo captaram sinais psicofisiológicos diferentes, mas que podem ser utilizados para derivar as mesmas métricas a fim de comparação ou indicadores diferentes. O *Empatica E4* mediu valores de volume de pulso sanguíneo (BVP), do qual podem ser derivados valores para variabilidade de variabilidade do ritmo cardíaco (HRV) e intervalo entre batimentos (IBI). Também possui sensor para temperatura da pele (TEMP), um para atividade elétrica cutânea (EDA) e por ultimo aceleração (ACC) em três eixos, mas este último não foi disponibilizado para uso no conjunto de dados. A taxa de coleta para BVP foi de 64 Hz através de um sensor fotopletismografia (PPG), 4 Hz para o EDA e TEMP também com 4 Hz, no entanto, a informação de temperatura não foi utilizada. O dispositivo *Galaxy Watch4* da Samsung coletou sinal de PPG verde, semelhante ao dispositivo da *Empatica*, a uma taxa de 25 Hz. Finalmente, o sinal coletado pelo *Muse S EEG headband* foi um eletroencefalograma (EEG) através de eletrodos localizados nas regiões pré-frontal e temporal, com uma taxa de coleta de 256 Hz e possuem contato seco com a pele. Ademais, este aparelho também possui uma informação de giroscópio, que poderia ser utilizada para um processo avançado de remoção de artefatos de movimento, mas não caberá no escopo deste trabalho.

3.3. Pré-processamento

Nesta etapa, os dados passarão por uma análise exploratória, que visa buscar um melhor entendimento das informações a fim de tomar decisões mais assertivas sobre as transformações que serão imposta aos dados. De forma análoga ao descrito por [Tukey et al. 1977], o próprio conjunto de dados foi utilizado para responder perguntas relevantes no processo de identificação de quais características contribuem para observação de esforço cognitivo com auxílio de gráficos para identificar padrões. Ademais, análises quantitativas também podem ser empregadas em momentos que os recursos visuais não forneçam uma tendência clara [Buja et al. 2009].

O pré-processamento pode ser dividido em três fases, a primeira delas busca reduzir a presença de artefatos causados por movimento ou interferências e ruídos nos sinais de sensores que possam ter sido causados por outros dispositivos ou o próprio mecanismo interno do aparelho. Para este objetivo, foram utilizadas técnicas de filtros passa-faixa e transformadas *wavelet*, uma vez que tanto trabalhos relacionados costumam utilizá-las, quanto por estas demonstrarem eficácia comprovada no tratamento de sinais [Joseph et al. 2014].

A partir do sinal limpo, com quantidade de ruídos reduzida, o processo de extração de características, o qual busca inferir o maior número possível de informações a partir do sinal psicofisiológico coletado. Entre as características buscadas, estão aquelas relatadas na Tabela 1, quanto valores psicofisiológicos como HRV, IBI, ritmo cardíaco (HR) e BVP.

Tabela 1. Características por domínio

Domínio	Característica
Tempo	Máximo, mínimo, média, mediana, moda, desvio padrão, Hjorth (atividade, mobilidade, complexidade), entropia, *High-Order Crossing* (HOC), comprimento de onda (WL)
Frequência	Análise de faixas espectro de densidade de energia (PSD)
Tempo-frequência	Eficiência de energia recursiva, energia de transformadas *wavelet* discretas, espectrograma, espectro de Hilbert-Huang

Especificamente, todos os sinais foram divididos em janelas de 30 segundos, equilibrando a eficiência computacional com a capacidade de capturar informações relevantes. Estudos indicam que janelas de 90 segundos poderiam fornecer melhor predição do esforço cognitivo [Ferreira et al. 2014], mas a duração reduzida das coletas inviabilizou esse tamanho, levando à escolha da janela de 30 segundos como compromisso entre qualidade e disponibilidade de dados.

O pré-processamento do sinal de EDA foi realizado com a biblioteca NeuroKit2. O sinal bruto passou por filtragem e foi convertido em onze colunas representando diferentes componentes da resposta galvânica da pele. As estatísticas mínima, máxima, média, mediana e desvio padrão foram extraídas e normalizadas entre zero e um. Para EEG, também foi utilizada a NeuroKit2, que gerou duas colunas: dissimilaridade global e campo de potência global. Com base nos dados brutos de 20 canais, foram calculadas estatísticas mínima, máxima, média, mediana e desvio padrão, além da normalização entre zero e um.

O pré-processamento do PPG exigiu etapas adicionais devido à variabilidade na qualidade do sinal, especialmente do Samsung Galaxy Watch4 e Empatica E4. Inicialmente, realizou-se a winsorização dos 1% extremos dos dados para remover leituras atípicas, seguida da aplicação de um filtro passa-faixa Butterworth com limites de 0.1Hz a 9Hz para minimizar distorções. Em seguida, os picos do sinal foram identificados com o SciPy, utilizando ajustes para melhorar a precisão na detecção. Após essa etapa, foram extraídas características do sinal com o uso da biblioteca HeartPy, incluindo intervalos RR, análise no domínio da frequência e filtragem de anomalias. Dessa forma, foram geradas as métricas batimentos por minuto (BPM), IBI e PSD. Por fim, todas as colunas resultantes foram normalizadas e submetidas a estatísticas descritivas para garantir a padronização dos dados e viabilizar sua utilização nos modelos de aprendizado de máquina.

Com o conjunto de características disponível, técnicas para selecionar as características mais relevantes e independentes entre si serão utilizadas. A redução do número de características permitirá maior agilidade no treinamento de diferentes modelos, minimizando a perca de informações, uma vez que apenas características com maior grau de redundância ou baixa correlação com esforço cognitivo serão removidas. Este pro-

cesso de seleção de características apenas foi realizado através de métodos embutidos nos algoritmos de classificação, caso aplicável.

3.4. Treinamento

Neste trabalho, foram explorados dois paradigmas distintos para a detecção de esforço cognitivo a partir de dados psicofisiológicos. O primeiro paradigma envolveu a engenharia de características, onde os sinais coletados foram processados para remover ruídos e interferências, derivando assim informações representativas do estado físico dos participantes. Esses dados foram então utilizados como entrada para três algoritmos tradicionais de aprendizado de máquina. O segundo paradigma aplicou aprendizado profundo por meio de quatro modelos distintos, os quais operaram diretamente sobre os dados brutos, sem pré-processamento para limpeza ou extração de características.

Os algoritmos de aprendizado de máquina tradicional foram treinados utilizando os dados processados conforme descrito na Seção 3.3. Foram empregados três modelos: Support Vector Machine (SVM) e K-Nearest Neighbors (KNN) e o (GBDT). Enquanto para o aprendizado profundo, quatro arquiteturas voltadas para a classificação de dados temporais foram utilizadas, todas previamente adotadas em estudos relacionados à detecção de esforço cognitivo. O primeiro, é o time-CNN (CNN) [Ismail Fawaz et al. 2019], este modelo é composto por duas camadas convolucionais seguidas de camadas de pooling médio, com filtros em múltiplos de seis e 12, e tamanhos de filtro múltiplos de sete. As camadas de pooling possuem tamanho três. A saída é combinada em uma camada densa com ativação sigmoide. O segundo modelo utilizado foi a FCN [Ismail Fawaz et al. 2019]. Esse modelo mantém o tamanho da série temporal inalterado, devido à ausência de camadas de pooling. Ele é composto por três blocos de convolução, normalização em lote e ativação ReLU, com filtros configurados em múltiplos de (128, 8), (256, 5) e (128, 3). A última camada antes da saída é uma Global Average Pooling (GAP), conectada a uma camada densa com ativação sigmoide.

Enquanto isso, o terceiro modelo foi o ResNet [Ismail Fawaz et al. 2019]. Originalmente, essa arquitetura possuía três blocos convolucionais, mas a quantidade foi tratada como hiperparâmetro. Um diferencial desse modelo é a conexão residual, que permite que a entrada de um bloco seja somada diretamente à sua saída, reduzindo o risco de dissipação do gradiente em redes profundas. Cada bloco contém três camadas convolucionais, normalização em lote e ativação ReLU, com filtros fixos em 64 e tamanhos múltiplos de oito, cinco e três. O atalho residual é composto por uma convolução de tamanho um e normalização em lote. Após o último bloco, a saída passa por uma camada GAP antes da densa final com ativação sigmoide.

O último modelo foi o CNN-LSTM [Kanjo et al. 2019]. Esse modelo combina redes convolucionais com redes recorrentes do tipo LSTM. Foram realizadas adaptações devido a restrições computacionais, reduzindo a quantidade de filtros e aumentando seu tamanho. O modelo incorpora mecanismos de controle de memória, permitindo capturar relações de longo prazo nos dados. A arquitetura do CNN-LSTM inclui duas camadas convolucionais, com filtros em múltiplos de quatro e oito, seguidas por pooling máximo de tamanho dois. Diferentemente dos demais modelos, os canais individuais são concatenados antes de passar por uma camada densa e um bloco LSTM, cuja quantidade de células de memória foi tratada como hiperparâmetro. A saída final é processada por uma camada densa com ativação sigmoide.

Os dados foram divididos em um conjunto de treinamento, composto por 10 voluntários, e um conjunto de testes, com os 3 voluntários restantes, que foi utilizado exclusivamente para avaliação final. No treinamento, aplicou-se a técnica LOSOCV, onde o modelo foi validado em um voluntário enquanto era treinado nos outros nove, repetindo o processo dez vezes. Esse método visa maximizar a capacidade de generalização dos modelos para novos indivíduos.

Por fim, para cada conjunto de modelo e sensor foi realizada a busca de hiperparâmetros, que são variáveis de cada algoritmo prévias ao processo de aprendizado e que podem influenciar a performance do modelo. Para este cenário, foi utilizado um algoritmo *tree-structured Parzen Estimator* (TPE) de busca Bayesiana.

4. Resultados e Discussão

Todos os modelos foram treinados a partir dos dados de 10 voluntários e avaliados nos dados de outros três, com métricas para os melhores modelos agrupados por sensor disponíveis na Tabela 2. As métricas exibidas são uma média da performance de cada modelo entre todos os participantes de teste.

Pode-se observar que, em média, os modelos tradicionais tiveram melhor desempenho na classificação correta de valores, independentemente de qual era o estado cognitivo, enquanto os modelos de aprendizado profundo apresentaram um viés a favor de esforço cognitivo, por isso tendo, em média, um valor de F1 mais elevado. Isto ocorre pelo fato de existir mais informação disponível para esforço cognitivo, de modo que um modelo o qual sempre resulte em uma classificação positiva para uma janela, tenha um valor de precisão mediano e sensibilidade igual a um, resultando em uma boa métrica para F1.

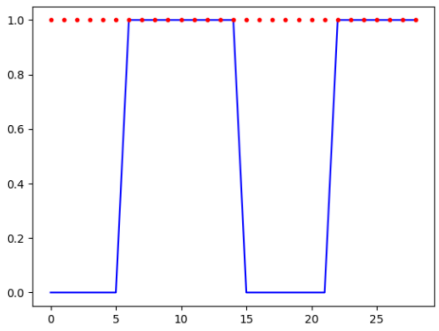
Para exemplificar a magnitude de variações entre voluntários, na Figura 2a é exibido a classificação para o modelo FCN de melhor performance para o sinal EEG ao testar o primeiro voluntário do conjunto de testes. Nesta imagem, a linha azul representa os valores reais do conjunto de dados, onde zero equivale ao estado normal e um a esforço cognitivo, enquanto os pontos vermelhos representam o estado inferido pelo modelo.

Enquanto que na Figura 2b, é exibido o teste do mesmo modelo mas para o segundo participante do grupo de testes. Embora que na primeira imagem o modelo tenha classificado todas as janelas como esforço cognitivo, para o segundo participante a classificação foi perfeita. Este exemplo repete-se para outros modelos e para outros sinais.

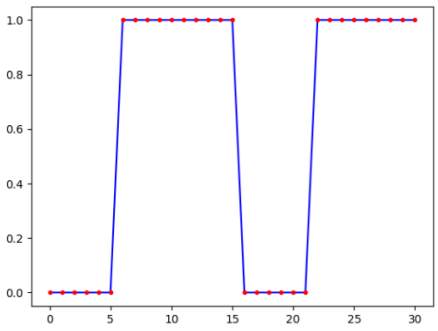
Outro ponto importante é a comparação da performance dos sinais coletados por dispositivos comercias em relação ao dispositivo clínico. O melhor valor absoluto observado foi obtido pelo sinal EDA coletado pelo *Empatica E4*, enquanto que o valor de PPG coletado pelo mesmo dispositivo apresentou performance inferior aos dados vindos do *Samsung Galaxy Watch4*. Contudo, a seleção de características para PPG foi limitada e baseada pelos dados vindos do dispositivo da Samsung, devido a baixa qualidade do sinal, o que pode sugerir um viés nos dados obtidos no processo de engenharia de características. Mais um fator para corroborar esta tese do viés parte do fato de que a performance dos modelos em aprendizado de ponta a ponta, onde não houve intervenção nos dados coletados, foi melhor para o dispositivo de qualidade clínica, da Empatica.

Tabela 2. Resumo dos resultados

	Modelo	Acurácia	Precisão	Sensibilidade	F1	AUC
Samsung PPG	CNN	0.619	0.626	0.929	0.748	0.534
	FCN	0.682	0.671	0.966	0.789	0.608
	LSTM	0.608	0.608	1.0	0.756	0.5
	RESNET	0.714	0.686	1.0	0.812	0.639
	KNN	0.695	0.718	0.808	0.756	0.666
	SVM	0.653	0.643	0.95	0.767	0.579
	XGB	**0.73**	**0.754**	**0.844**	**0.791**	**0.698**
Muse EEG	CNN	0.595	0.589	0.667	0.609	0.603
	FCN	**0.667**	**0.658**	**1.0**	**0.768**	**0.681**
	LSTM	0.412	0.629	0.309	0.308	0.464
	RESNET	0.533	0.53	0.979	0.682	0.515
	KNN	0.602	0.601	0.754	0.662	0.603
	SVM	0.524	0.524	1.0	0.684	0.5
	XGB	0.571	0.558	0.922	0.688	0.56
Empatica PPG	CNN	0.55	0.64	0.6	0.618	0.536
	FCN	0.696	0.705	0.912	0.792	0.637
	LSTM	0.548	0.597	0.808	0.686	0.473
	RESNET	0.642	0.7	0.826	0.746	0.594
	KNN	0.636	0.686	0.746	0.714	0.602
	SVM	**0.709**	**0.74**	**0.842**	**0.78**	**0.661**
	XGB	0.614	0.614	1.0	0.761	0.5
Empatica EDA	CNN	0.602	0.609	0.983	0.752	0.492
	FCN	0.613	0.613	1.0	0.76	0.5
	LSTM	0.537	0.575	0.842	0.677	0.449
	RESNET	0.613	0.613	1.0	0.76	0.5
	KNN	0.688	0.733	0.799	0.758	0.658
	SVM	0.711	0.711	0.917	0.8	0.644
	XGB	**0.743**	**0.75**	**0.885**	**0.812**	**0.696**



(a) Resultado FCN EEG para primeiro participante de teste



(b) Resultado FCN EEG para segundo participante de teste

Figura 2. Comparação dos resultados FCN EEG para dois participantes de teste

5. Conclusão

Neste trabalho, discutiu-se a viabilidade do uso de inteligência artificial aplicada a dados de dispositivos vestíveis populares para detectar janelas de esforço cognitivo. A adoção dessa abordagem possibilita a criação de um sistema contínuo e independente de monitoramento, eliminando a necessidade de equipamentos clínicos de alto custo, o que favorece a democratização do acesso a esse tipo de tecnologia. Além das implicações para a área da saúde, a possibilidade de monitoramento em tempo real de estados psicológicos abre novas perspectivas para pesquisas sobre a efetividade de medicamentos para transtornos de atenção, bem como para o acompanhamento do desempenho cognitivo no ambiente de trabalho. Esse tipo de tecnologia pode oferecer insights relevantes sobre a relação entre carga cognitiva e transtornos mentais, incluindo depressão, ansiedade e síndrome de burnout.

Apesar dos avanços apresentados, algumas limitações devem ser destacadas. O conjunto de dados utilizado foi relativamente pequeno, com poucos participantes e um período curto de coleta, o que pode restringir a capacidade de generalização dos modelos. Além disso, a ausência de informações adicionais sobre os voluntários, como faixa etária e condições individuais, pode influenciar nos indicadores psicofisiológicos analisados. Outra limitação refere-se ao uso de apenas um sensor, enquanto dispositivos vestíveis mais modernos frequentemente incorporam múltiplos sensores, como PPG e EDA, que poderiam fornecer uma análise mais robusta. Além disso, a qualidade dos sensores comerciais implica uma relação sinal-ruído inferior, exigindo estratégias avançadas de pré-processamento. Por fim, o treinamento dos modelos foi realizado em um cenário controlado, sem exposição a dados obtidos em um ambiente real, o que pode impactar a aplicabilidade prática da solução.

Como trabalhos futuros, propõe-se a expansão da base de dados, a inclusão de múltiplos sensores para aprimorar a precisão dos modelos e a realização de coletas em contextos reais, permitindo uma avaliação mais abrangente da robustez e aplicabilidade das soluções propostas.

Agradecimentos

Os autores gostariam de agradecer à CAPES (C.F. 001) e ao CNPq (nº 309537/2020-7).

Referências

Borisov, V., Kasneci, E., and Kasneci, G. (2021). Robust cognitive load detection from wrist-band sensors. *Computers in Human Behavior Reports*, 4.

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F., and Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4361–4383.

Cinaz, B. (2013). *Monitoring of cognitive load and cognitive performance using wearable sensing*.

Ding, Y., Cao, Y., Duffy, V. G., Wang, Y., and Zhang, X. (2020). Measurement and identification of mental workload during simulated computer tasks with multimodal methods and machine learning. *Ergonomics*, 63.

Ferreira, E., Ferreira, D., Kim, S., Siirtola, P., Röning, J., Forlizzi, J. F., and Dey, A. K. (2014). Assessing real-time cognitive load based on psycho-physiological measures for younger and older adults. In *2014 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB)*, pages 39–48. IEEE.

Fleming, H., Robinson, O. J., and Roiser, J. P. (2023). Measuring cognitive effort without difficulty. *Cognitive, Affective, & Behavioral Neuroscience*, 23(2):290–305.

Grzeszczyk, M. K., Blanco, R., Adamczyk, P., Kus, M., Marek, S., Prkecikowski, R., and Lisowska, A. (2023). Cogwear: Can we detect cognitive effort with consumer-grade wearables?

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., and Muller, P.-A. (2019). Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963.

Joseph, G., Joseph, A., Titus, G., Thomas, R. M., and Jose, D. (2014). Photoplethysmogram (ppg) signal analysis and wavelet de-noising. In *2014 Annual International Conference on Emerging Research Areas: Magnetics, Machines and Drives (AICERA/iCMMD)*, pages 1–5.

Kanjo, E., Younis, E. M., and Ang, C. S. (2019). Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Information Fusion*, 49.

Longo, L., Wickens, C. D., Hancock, G., and Hancock, P. A. (2022). Human mental workload: A survey and a novel inclusive definition. *Frontiers in psychology*, 13:883321.

Paas, F., Tuovinen, J. E., Tabbers, H., and Gerven, P. W. V. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38.

Shu, L., Yu, Y., Chen, W., Hua, H., Li, Q., Jin, J., and Xu, X. (2020). Wearable emotion recognition using heart rate data from a smart bracelet. *Sensors (Switzerland)*, 20.

Tukey, J. W. et al. (1977). *Exploratory data analysis*, volume 2. Springer.