DESENVOLVIMENTO DE UMA INTERFACE DE FALA SILENCIOSA UTILIZANDO DEEP LEARNING, EMG E PROCESSAMENTO DE SINAIS

Mateus de Aquino Batista, Mario Oliveira Lima, Wagner dos Santos Clementino de Jesus.

Universidade do Vale do Paraíba, Faculdade de Ciências da Saúde, Avenida Shishima Hifumi, 2911, Urbanova - 12244-000 - São José dos Campos-SP, Brasil, mateusaqb@gmail.com, mol@univap.br, wagner@univap.br.

Resumo
O presente estudo trata-se de um projeto de instrumentação de uma interface de fala silenciosa de baixo custo, compreendendo o desenvolvimento de uma rede neural capaz de classificar estímulos musculares. O *software* é compatível com diversas placas modulares de eletromiografia (EMG) para capturar e armazenar os dados de atividade muscular, que podem ser processados e treinados por uma rede neural convolucional, a fim de identificar as palavras subvocalizadas ou demais estímulos musculares, tudo integrado ao *software*. A rede neural foi treinada e validada com arquivos de dados de EMG de domínio público, divididos em 90% para treinamento e 10% para validação, o valor da acurácia final do treinamento da rede neural foi de 96.99% enquanto da validação alcançou 96.58%. Com resultados positivos, será possível expandir o vocabulário da interface, aumentando sua utilidade no cotidiano de pacientes com dificuldades de comunicação. Em conclusão, o trabalho busca desenvolver uma solução acessível e eficiente para auxiliar na comunicação de pessoas com dificuldades de fala, usando tecnologias de eletromiografia e redes neurais.
Palavras-chave: Interfaces Cérebro-Computador. Processamento de Sinais Assistido por Computador. Redes Neurais de Computação.

Curso: Biomedicina.

Introdução

De acordo com Darley et al. (1969), a disartria é um distúrbio motor caracterizado pela perda da capacidade de articular as palavras de forma regular devido a danos no sistema nervoso central ou periférico. A fala disártrica é caracterizada pela interrupção da fonação e/ou pela precisão articulatória reduzida, tornando-a difícil de compreender para os ouvintes humanos, como afirmado por Deng et al. (2009). Em casos mais graves de disartria, pode ser necessário o uso de dispositivos externos de comunicação, como mencionado por Huang (2021). Uma abordagem emergente para comunicação utiliza dispositivos capazes de interpretar e processar os movimentos faciais em texto, estabelecendo uma interface entre a fala silenciosa (subvocalização) com o computador. Essa técnica tem sido explorada por pesquisadores, conforme mencionado por Deng et al. (2009), Galego (2016) e Kapur et al. (2018).

A subvocalização refere-se à articulação inaudível com os órgãos responsáveis pela fala e desempenha um papel no processamento cognitivo e retenção de informações, reduzindo a carga cognitiva, como mencionado por Rayner et al. (2011). Já a eletromiografia (EMG) é uma técnica que permite registrar estas atividades elétricas musculares, abrindo caminho para várias aplicações clínicas e biomédicas, conforme destacado por Valderrama et al. (2021). Ao contrário das interfaces cérebro-máquina tradicionais, a eletromiografia não acessa informações ou pensamentos privados, e a entrada é voluntária por parte do usuário (Kapur et al., 2018). Neste estudo, a proposta é estender a interface de forma não invasiva, obtendo e classificando dados de EMG de forma customizável.

O objetivo principal desta pesquisa é identificar qualquer estímulo muscular e classificá-los em um conjunto predefinido de palavras através de uma rede neural convolucional. As palavras identificadas podem ser utilizadas para escrever em um computador sem a necessidade de um teclado, reproduzir áudio por meio de um sintetizador *Text-To-Speech* automaticamente ou até mesmo controlar dispositivos conectados à internet.

DOI: https://dx.doi.org/10.18066/inic0371.23

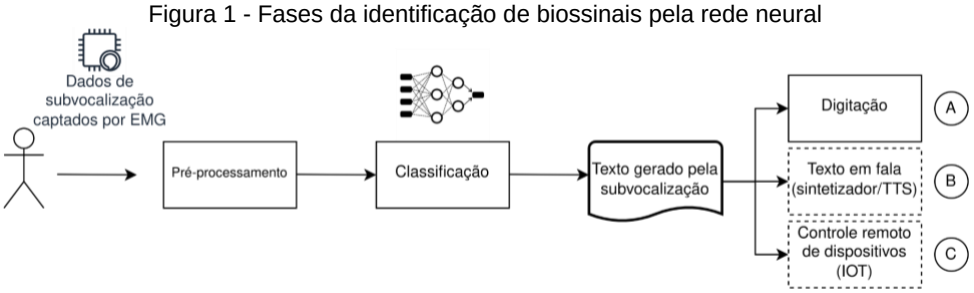
Metodologia

Antes do desenvolvimento da aplicação (*software*), foi realizada uma prova de conceito (PoC) da pesquisa em um IPYNB (*IPython Notebook*, plataforma de computação para documentação e codificação em tempo real de códigos em Python). Depois foi realizada a montagem do dispositivo de coleta (*hardware*) para testar a conexão com o aplicativo desenvolvido.

Toda a lógica da instrumentação até o processamento dos dados foi realizada na linguagem de programação Python para maior agilidade e fácil reprodutibilidade do experimento, utilizando a biblioteca Brainflow, oferecendo suporte para 16 tipos de placas de biossensores diferentes conforme recomendado pela documentação da OpenBCI. A OpenBCI é conhecida por seu desenvolvimento de ferramentas de neurociência e biossensores de código aberto, desta forma, foi utilizada a placa Cyton de 8 canais da OpenBCI para o teste da conexão de ponta a ponta.

Após o treinamento da rede neural, os dados podem ser pré-processados e classificados em tempo real, conforme a Figura 1. Neste estudo foi priorizada a solução A para a prova de conceito, porém, com o resultado da palavra já identificada é possível realizar integrações com qualquer biblioteca de sintetização de texto ou controle de dispositivos externos.

Figura 1 - Fases da identificação de biossinais pela rede neural

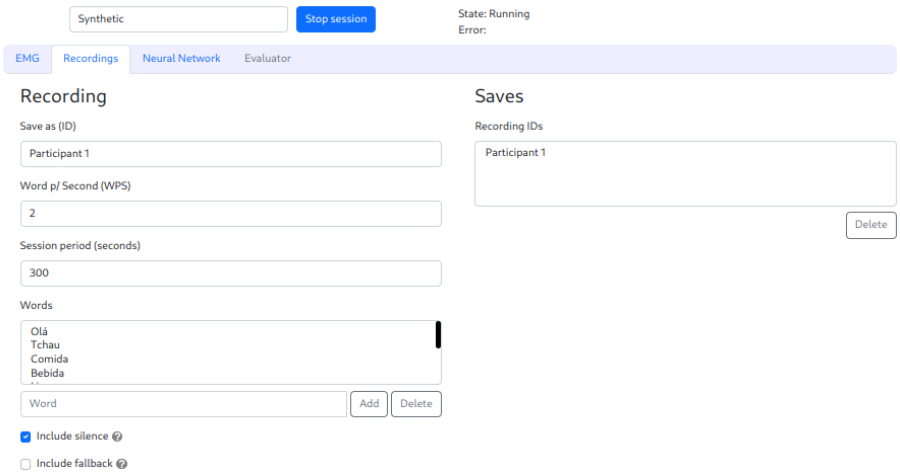


Fonte: O Autor (2022).

O pré-processamento, faz-se necessário para reduzir os ruídos coletados pelo EMG, como evidenciado no estudo de Deng et al. (2009), os dados brutos devem ser inicialmente filtrados, para remover a interferência da rede, o deslocamento da corrente contínua, que muda lentamente devido a diferenças de potencial naturais na interface eletrodo/pele, bem como qualquer ruído de alta frequência. Para isso, foi utilizado o filtro Butterworth, no qual os registros de sinais são filtrados na faixa de frequência de 2 a 45Hz.

A tela do *software* desenvolvido pode ser observada abaixo na Figura 2 na aba de Gravação.

Figura 2 - Aba de gravação do *software* desenvolvido



Fonte: O Autor (2023)

DOI: https://dx.doi.org/10.18066/inic0371.23

Além da linguagem Python no *backend*, a aplicação também utiliza as linguagens HTML (*HyperText Markup Language*), Javascript e CSS (*Cascading Style Sheets*) com Bootstrap no *frontend* para a criação da interface, a comunicação entre as duas pontas é feita via *websocket* com a biblioteca Eel para maior velocidade na apresentação dos resultados vindos da placa de eletromiografia, e também pode ser testada com dados sintéticos que são obtidos através de um gerador de ondas aleatórias, sem a necessidade de um participante para testes.

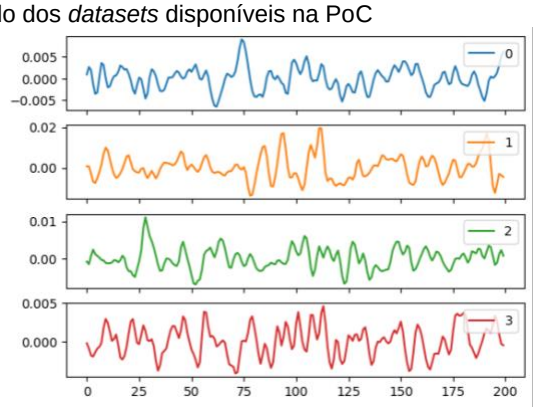
Toda a prova de conceito (PoC), o passo a passo para a montagem do equipamento, instalação do software, código-fonte, equipamentos e os arquivos de EMG coletados de participantes de domínio público para a prova de conceito foram armazenados em um repositório público no GitHub (https://github.com/MateusAquino/subvocalization-emg), uma plataforma online de hospedagem de código-fonte e arquivos com controle de versão para que o experimento possa ser replicado por outros pesquisadores, melhorado ou até mesmo apenas revalidado sem a necessidade da compra dos equipamentos ou submissão de participantes às coletas dos dados de eletromiografia, sendo uma abordagem positiva para a transparência e colaboração científica.

O *dataset* público utilizado para o treinamento e validação da rede neural possui um tamanho de 30.3 MB, no total 759.559 linhas com dados de 4 canais de biossinais de EMG de superfície, que, após o pré-processamento foi convertido em 3800 matrizes de gestos distintos. Este *dataset* foi disponibilizado pelos pesquisadores Toro-Ossaba et al., 2022. Este contém dados de 8 participantes com cinco gestos distintos para classificação: mão aberta, mão fechada, pinça lateral, sinalização e maloik. É possível conferir abaixo na Figura 3 a esquerda uma amostra do conteúdo do arquivo CSV após o processamento, e a direita os mesmos dados de um único gesto (mão aberta) dispostos em um gráfico com os 4 canais (eletrodos) criado com a biblioteca de plotagem de gráficos Matplotlib.

Figura 3 - Amostra do conteúdo dos *datasets* disponíveis na PoC

	WORD	Channel_1	Channel_2	Channel_3	Channel_4
0	Open hand	0.004898	0.004003	-0.004031	-0.001103
1	Open hand	0.003446	-0.004977	0.001012	-0.002247
2	Open hand	-0.003934	-0.009833	0.007323	-0.003097
3	Open hand	-0.002250	-0.007743	-0.004610	-0.000540
4	Open hand	-0.003268	-0.006243	0.002576	-0.003321
...	...	...	...	...	...
759555	Rock sign	0.034819	-0.008418	-0.006649	0.027904
759556	Rock sign	0.043744	-0.007028	-0.001524	0.050832
759557	Rock sign	0.055965	-0.002811	0.007052	0.015889
759558	Rock sign	0.054340	-0.014413	-0.030651	-0.001178
759559	Rock sign	0.049726	-0.020201	-0.052454	-0.040208

759560 rows × 5 columns



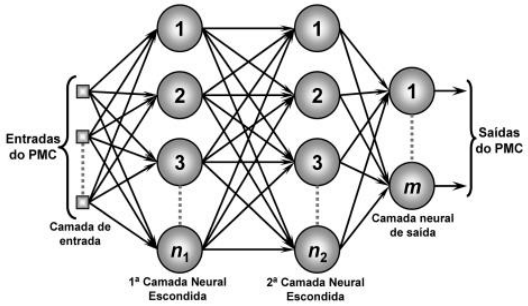
Fonte: O Autor (2023)

Para a criação da rede neural, cada camada convolucional sequente desenvolve filtros que são capazes de reconhecer características sucessivamente mais complexas nos dados normalizados de EMG. Aprendendo a taxa de disparo dos neurônios motores, o algoritmo é capaz de encontrar as frequências nas quais os recursos mais úteis estão presentes, facilitando a detecção de frequências e fonemas posteriores. A nível de código, foram utilizadas duas bibliotecas importantes no desenvolvimento: o Numpy, uma biblioteca para o processamento de grandes e multi-dimensionais arranjos e matrizes, e o Keras, uma biblioteca de *deep learning* modular capaz de implementar arquiteturas de redes convolucionais de forma ágil.

O modelo utilizado foi o sequencial da biblioteca Keras, que é mais apropriado para uma pilha simples de camadas onde cada camada tem exatamente um tensor de entrada e um tensor de saída. O modelo criado da rede neural convolucional terá as seguintes características de arquitetura dispostas na Figura 4. O Perceptron de Múltiplas Camadas (PMC) utilizado neste estudo possui as seguintes características: cada uma das entradas do PMC corresponde aos dados de EMG e são provenientes de sete canais da região facial. Essas entradas são unidimensionais, representando uma única onda de EMG. Além disso, o PMC possui múltiplas entradas, pois cada onda de EMG é composta por vários pontos, dependendo da frequência de leitura.

DOI: https://dx.doi.org/10.18066/inic0371.23

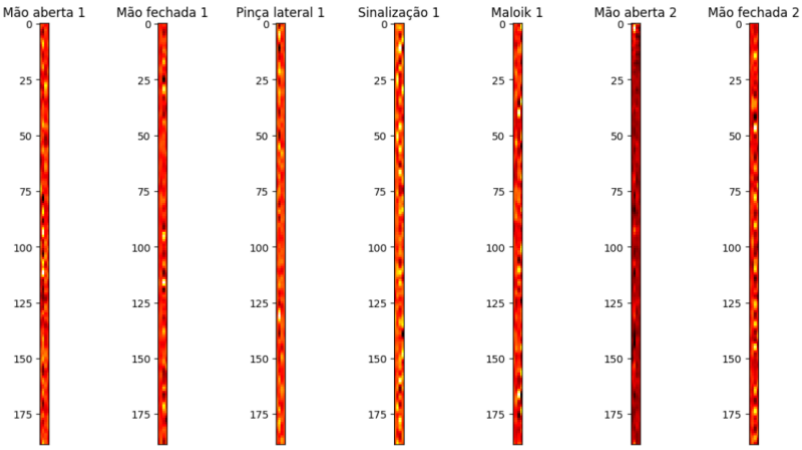
Figura 4 - Ilustração de rede Perceptron multicamadas



Fonte: Silva *et al*. (2010)

O PMC possui uma primeira camada para detectar recursos e várias camadas internas (*hidden layers*, ou camadas escondidas) para a extração de recursos de nível superior. Essas camadas internas são responsáveis por identificar recursos mais complexos nos dados de EMG. Por fim, as saídas do PMC são projetadas para reconhecer inicialmente apenas 5 palavras (ou gestos) diferentes. Portanto, a rede neural tem 5 saídas distintas, correspondendo a cada uma dessas palavras. Após o pré-processamento dos dados, cada um dos gestos realizados em um intervalo de tempo pode ser plotado em 4 canais (colunas dos eletrodos) em função do tempo de registro (189 linhas das samples), onde a mesma amplitude da onda exibida na Figura 3 é convertida na intensidade da cor, conforme a Figura 5.

Figura 5 - Resultado dos dados de EMG após o processamento separados por gestos



Fonte: O Autor (2023)

Cada imagem ilustra uma representação em mapa de calor correspondente a um único gesto. Consequentemente, considerando uma configuração de 4 eletrodos por 189 samples, a matriz de entrada para a PMC consiste em 756 elementos e resultando em um total de 5 saídas. Para a divisão dos dados. Quanto à proporção de dados utilizados destinados para o treinamento e a validação, foram treinados dois modelos utilizando proporções diferentes, um de 80/20 e outro de 90/10, considerando que, a maior parte das matrizes de dados são utilizadas para o treinamento, deixando o restante destinado apenas à validação.

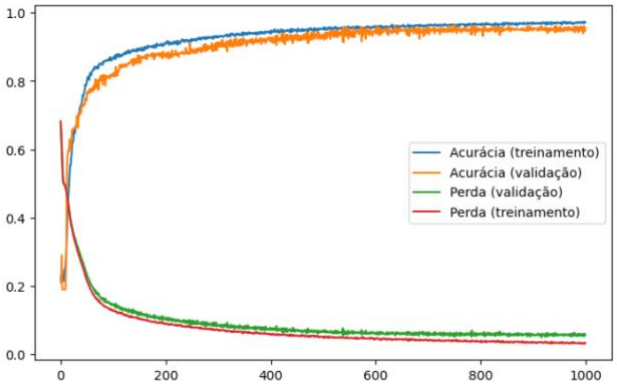
Resultados

Conforme experimentos previamente realizados por Kapur et al., 2018 e Chandrashekhar, 2021, na validação da rede neural, espera-se como uma taxa de acerto satisfatória o valor entre 90 a 95%, aproximadamente, no teste de precisão para a classificação e identificação das palavras (definidas em um vocabulário enxuto com 10 termos).

DOI: https://dx.doi.org/10.18066/inic0371.23

Após o treinamento da rede neural com o *dataset* público de EMG, composto por 5 termos distintos, as taxas de acurácia podem ser observadas na Figura 6. Após 1000 gerações, os resultados alcançaram valores de acurácia de 96.99% e 96.58% durante o treinamento, com perdas de 3.39% e 3.49% no cenário de divisão de dados 80/20 (treinamento/validação). Por outro lado, ao adotar a proporção de 90/10, a acurácia de validação tende a diminuir, já que a rede neural foi treinada com menos dados. Nessa configuração, as acurácias atingiram 96.97% no treinamento e 93.29% na validação, com perdas de 3.31% e 6.37%, respectivamente.

Figura 6 - Valores de acurácia e perda da rede neural de acordo com o histórico do treinamento e validação ao longo das gerações.



Fonte: O Autor (2023)

Discussão

A abordagem proposta destaca-se em termos de acessibilidade e usabilidade. Apesar de requerer uma placa de coleta de EMG, é uma alternativa viável e de baixo custo para a democratização ao acesso da comunicação assistida. Além disso, é possível utilizar a ferramenta sem exigir um conhecimento avançado prévio em programação, pois todo o fluxo pode ser controlado pela interface gráfica, seja para a gravação e treinamento da rede neural ou para sua aplicação prática. Isso amplia seu alcance para um público mais amplo, incluindo aqueles sem experiência em programação. Por fim, o software é de código aberto, disponível para a comunidade científica e desenvolvedores, promovendo a transparência e incentivando colaborações e melhorias contínuas.

É relevante destacar que este estudo alcançou resultados superiores em comparação com tentativas de treinamento anteriores realizadas, utilizando o mesmo *dataset*, a mesma quantidade de gerações e proporção de validação 80/20. Os resultados anteriores atingiram na fase de treinamento uma acurácia de 93% e uma taxa de perda de 16%, com uma validação de 92% e perda de 22% (Toro-Ossaba et al., 2022). Essa diferença substancial pode ser atribuída, em parte, à implementação de um processo abrangente de pré-processamento de dados, incluindo a remoção de ruídos e artefatos, juntamente com a diferente abordagem na modelagem da rede neural.

Considerando que as palavra ou os gestos foram identificados, o software exibe na tela o termo identificado, desta forma, é possível estender a aplicação, tanto pelo *backend* via Python quanto pelo *frontend* via Javascript para definir ações personalizadas para cada termo, por exemplo sintetizar fala, realizar cálculos, controlar um teclado virtual ou até mesmo controlar um dispositivo externo como uma cadeira de rodas elétrica.

Os resultados obtidos com o conjunto de dados utilizado para o treinamento e validação da rede neural são promissores, com taxas de acerto significativas. No entanto, reconhece-se a necessidade de expandir e aprimorar o conjunto de dados. Após a aprovação do comitê de ética e pesquisa (CAAE 65587722.5.0000.5503), planeja-se a coleta de dados de EMG diretamente de participantes humanos. Isso permitirá um treinamento mais abrangente e a adaptação da rede neural para uma maior variedade de sinais, garantindo uma solução ainda mais precisa e eficaz, visto que, a pesquisa atual não incorporou testes diretos em humanos, limitando-se ao uso de dados públicos, a próxima etapa evolutiva do estudo implicará a substituição dos gestos disponíveis nas bases por dados de eletromiografia faciais provenientes da coleta direta em participantes.

DOI: https://dx.doi.org/10.18066/inic0371.23

Uma possível evolução interessante para a pesquisa envolve a identificação de fonemas em vez de palavras completas. Isso ampliaria significativamente as possibilidades, permitindo que o software reconhecesse e interpretasse qualquer sequência de fonemas pronunciados silenciosamente, tornando-o aplicável a qualquer palavra do dicionário. Embora a identificação de fonemas apresenta desafios adicionais, como uma análise mais detalhada dos sinais de EMG, seu sucesso poderia tornar o software mais acessível e adaptável a diferentes idiomas e dialetos, ampliando sua utilidade e flexibilidade.

Conclusão

O trabalho busca desenvolver uma solução acessível e eficiente para auxiliar na comunicação de pessoas com dificuldades de fala. Com resultados positivos, essa interface tem o potencial de expandir o vocabulário disponível e aumentar sua utilidade no cotidiano de pacientes com dificuldades de comunicação.

Através do uso da abordagem de rede neural convolucional, o software desenvolvido neste trabalho demonstrou uma promissora aplicação na classificação de sinais de EMG, representando um avanço significativo na comunicação assistida. No entanto, para otimizar ainda mais a precisão e o desempenho da interface, torna-se imperativo o acesso e treinamento a mais dados, bem como a ampliação da quantidade de termos predefinidos para a rede neural.

Além disso, considerando a perspectiva futura de identificação de fonemas em vez de palavras completas, à medida que a pesquisa evolui, a aplicação prática dessa abordagem pode se estender não apenas à comunicação de palavras, mas também à construção livre de frases e expressões, tornando-a mais inclusiva e adaptável a diversos contextos linguísticos e necessidades individuais.

Referências

CHANDRASHEKHAR, V. The classification of EMG signals using machine learning for the construction of a silent speech interface. **The Young Researcher**, 5 (1), 266-283. 2021. ISSN: 2560-9815 (Print) 2560-9823 (Online).

DARLEY, F.; ARONSON, AE.; BROWN, JR. Differential diagnostic patterns of dysarthria. **J Speech Hear Res.** 1969; 12: 246-69.

DENG, Yunbin. *et al*. Disordered speech recognition using acoustic and sEMG signals. Proceedings of the Annual Conference of the International Speech Communication Association, 2009. **INTERSPEECH**. 644-647.

GALEGO, J. S. **Aquisição e processamento de biosinais de eletromiografia de superfície e eletroencelografia para caracterização de comandos verbais ou intenção de fala mediante seu processamento matemático em pacientes com disartria**. Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal do Rio Grande do Sul, 2016.

HUANG, J. **Disartria**. Manual MSD Versão Saúde para a Família. 2021. Disponível em: <https://www.msdmanuals.com/pt-br/casa/distúrbios-cerebrais,-da-medula-espinal-e-dos-nervos/disfunção-cerebral/disartria>. Acesso em: 15 nov. 2022.

KAPUR. A.; KAPUR S.; MAES P. ALTEREGO: A Personalized Wearable Silent Speech Interface. **23rd International Conference on Intelligent User Interfaces (IUI 2018)**. Tóquio, Japão, p 43-53, mai. 2018. DOI: 10.1145/3172944.3172977.

RAYNER, K.; POLLATSEK, A.; ASHBY, J.; CLIFTON Jr., C. **Psychology of Reading**. 2nd ed. New York: Psychology Press., 9 nov. 2011. DOI: 10.4324/9780203155158.

SILVA, IN.; SPATTI, DH.; FLAUZINO, RA. **Redes neurais artificiais para engenharia e ciências aplicadas**. 2. ed. São Paulo: Artliber Editora, 10-0725. 2010.

TORO-OSSABA, A.; JARAMILLO-TIGREROS, J.; TEJADA, J.C.; PEÑA, A.; LÓPEZ-GONZÁLEZ, A.; CASTANHO, R.A. **EMG hand gesture dataset (2.1)**. 2022 [Data set]. Zenodo. DOI: 10.5281/zenodo.7668251.

TORO-OSSABA, A.; JARAMILLO-TIGREROS, J.; TEJADA, J.C.; PEÑA, A.; LÓPEZ-GONZÁLEZ, A.; CASTANHO, R.A. LSTM Recurrent Neural Network for Hand Gesture Recognition Using EMG Signals. **Appl. Sci**. 2022, 12, 9700. https://doi.org/10.3390/app12199700

DOI: https://dx.doi.org/10.18066/inic0371.23

VALDERRAMA, M.; GONZÁLEZ, E.; GARCÍA F. Development of a low-cost surface EMG acquisition system device for wearable applications. **IEEE 2nd International Congress of Biomedical Engineering and Bioengineering**. CI-IB&BI, pp. 1-4, 2021. DOI: 10.1109/CI-IBBI54220.2021.9626100.

DOI: https://dx.doi.org/10.18066/inic0371.23