# Contribution Information

## *VAMPIRE*: vectorized automated ML pre-processing and post-processing framework for edge applications

This paper aims to present a novel Machine Learning framework which consists of multiple pre-processing and post-processing algorithms, which are applied in an edge computing environment and demand lower processing requirements and allow for improvement in prediction accuracy. The pre-processing algorithms extract features from biomedical time-series by generating CSV format datasets in an automated manner and consist of two feature extraction arrangements that allow the creation of the datasets regardless of the quality of the medical annotation. Also, the post-processing algorithm that has been conceived is targeted towards hardware platforms with limited processing capabilities, such as the Raspberry Pi, where all Machine Learning forecasts are performed in this paper. This new post-processing block, named Activation Engine, aims to extract binary classification thresholds for 2 Softmax outputs, to apply on the test set based on training set accuracy. Due to the limited processing power of IoT nodes, automated Machine Learning techniques such as meta-learning and ensemble learning are not feasible, since they are time-consuming, require sufficient resources, and are additionally difficult to implement. Also, forecasts through a commercial software were also implemented to test the novel pre-processing algorithm, to test other setups and ensure robustness. Furthermore, already pre-processed datasets were tested using the Activation Engine for the same reciprocal purpose.

In [1], authors aim to convert time-series into words which are used as features. However, concerning the application field, this method might not be consistent in terms of performance due to possible information loss. In [2], [3], wavelet transform is used to extract featured before performing ML forecasts using classical algorithms. This technique requires tuning and manual operation to achieve accurate results. In fact, the pre-processing algorithms described in this paper is completely automated due to the introduction of a novel V.Hz waveform which varies over time and is extracted from the biomedical time-series. Additionally, the three biomedical time-series are diverse enough to prove the robustness of the framework.

Authors in [4], present Auto-Sklearn, which is an automated version of and built over the Sklean Python Library. It implements a hybrid automated Machine Learning setup with a meta-learning configuration that has an ensemble learning block at the output. This technique can achieve improved and effortless forecast accuracy; however, it remains substantially heavy in terms resource and processing requirements and is naturally time-consuming. Contrarily, the post-processing method provided in this paper (Activation Engine) is light on resource requirements and achieves accurate results without relying heavily on hyperparameter tuning.

Concerning the experimental results, in case of the ECG dataset, when 5 classes are considered, the best performance was achieved using SVM where the accuracies for the 5 labels are: 99.57%, 94.61%, 98.8%, 96.88%, and 89.11%. Regarding the EEG dataset forecasts were performed using Rulex and the Activation Engine, where the best accuracy achieved for Rulex was using LLM where the accuracies for the labels are: 85.09% and 85.99%. As for the Activation Engine, the results are: 87.58% and 83.07%. The PPG dataset was applied using an MLP with an Activation Engine at the output, where the overall binary accuracy for two patients combined is 83.59%. Furthermore, the pre-processed datasets achieved higher accuracies than those reported their original publications, and by a wide margin, where for the Activity Detection dataset, a near perfect testing accuracy was achieved using the Activation Engine.

## Declarations

Not Applicable.


## Funding

This research received no external funding.


## Conflicts of interest

The authors declare no conflict of interest.

## Ethics approval

Not Applicable.

## Consent to participate

Not Applicable.

## Consent for publication

Not Applicable.

## Availability of data and material

This is the link for the dataset used for ECG arrythmia classification:

https://www.kaggle.com/mondejar/mitbih-database

This is the link for the dataset used for EEG seizure classification:

https://www.kaggle.com/knight079/epilepsyseizure

This is the link for the dataset used for PPG gaming classification:

https://www.kaggle.com/canaria/5-gamers

This is the link for the dataset used for Human activity classification using smartphones:

https://www.kaggle.com/uciml/human-activity-recognition-with-smartphones

The code for the radar classification dataset is not publicly available.

## Code availability

The source code used in this paper cannot be accessed online.

## Authors' contributions

Ali Walid Daher developed and implemented every algorithm in the framework, and also wrote the main draft of the paper. All authors discussed the presented results and revised the current version of the manuscript.

## References


[1] J. Wang, P. Liu, M. F. H. She, S. Nahavandi, and A. Kouzani, "Bag-of-words representation for biomedical time series classification," *Biomed. Signal Process. Control*, vol. 8, no. 6, pp. 634–644, Nov. 2013, doi: 10.1016/j.bspc.2013.06.004.

[2] T. Kalayci and O. Ozdamar, "Wavelet preprocessing for automated neural network detection of EEG spikes," *IEEE Eng. Med. Biol. Mag.*, vol. 14, no. 2, pp. 160–166, Mar. 1995, doi: 10.1109/51.376754.

[3] "Face Gender Image Classification Using Various Wavelet Transform and Support Vector Machine with various Kernels - ProQuest."

https://www.proquest.com/openview/6fbb981e646162367f8854f32e27fbe8/1?pq-origsite=gscholar&cbl=55228 (accessed Jul. 22, 2021).

[4] “Auto-sklearn: Efficient and Robust Automated Machine Learning | SpringerLink.” https://link.springer.com/chapter/10.1007/978-3-030-05318-5_6 (accessed Jun. 28, 2021).