**CHAPTER 3**

**METHODOLOGY**

**Research Design**

In this study, the researchers will adopt an experimental and quantitative method to achieve the objective of automatically detecting and removing artifacts from electrodermal activity (EDA) signals used for stress detection. The study will employ hybridized Long Short-Term Memory (LSTM) and Support Vector Machine (SVM) models and compare their performance to existing models, including binary classifiers such as SVM, k-Nearest Neighbors (KNN), Logistic Regression, Random Forest, and Naive Bayes. The goal is to determine the efficacy of the hybrid LSTM-SVM model in improving the overall performance in terms of accuracy, precision, recall, and F1-score, and the reliability of stress detection from EDA signals by effectively handling artifacts. To attain this objective, the researchers propose the application of hybridized LSTM-SVM to address the concern with the non-stationary characteristic of EDA data: it can change through the course of time.

*Variables?*

*Expected outcome?*

**Sources Of Data**

This study will utilize the Electrodermal Activity artifact correction BEnchmark (EDABE) data set as its primary source of data.EDABE dataset consist of electrodermal activity (EDA) recordings of hand and body motion artifacts. EDABE includes the signal with the following variables such as timestamp of the signal, raw data, cleandata, binary target, signal automatic, predArtifact,and postProcessedPredArtifacts.

**Research Instrument**

This study will use the Python as the primary programming language

*Programming language? Python*

*Dataset? EDABE Dataset by Llanes-Jurado et. al. (2023)*

*GUI?*

*Tools: Tensorflow, Scikit-learn, Google Colab*

**Procedures**

Data Collection: (reference)

* Will use publicly available dataset (EDABE)

Preprocessing:

* Load data into software (e.g., EEGLAB).
* Apply band-pass filter (e.g., 0.5-50 Hz).
* Resample data if necessary.

Artifact Detection:

* Perform ICA to decompose the EEG.
* Identify artifact components by visual inspection or automated criteria.

Artifact Removal:

* Remove identified ICA components.
* Alternatively, apply regression or adaptive filtering techniques.

Post-Processing:

* Reconstruct the clean EEG signal.
* Validate the removal process by checking the data quality and re-running analyses if necessary.

**System Architecture**

Raw EDA signal is uploaded to a web application

Raw EDA signal is split into 0.5 second segments

Segments are fed into validated LSTM-SVM model

Decoding which signals have been labeled as artifacts

Return labeled raw EDA signal

Gawin nalang picture/image ito

**Ethical Considerations**

In conducting this research on developing a hybridized LSTM-SVM model for artifact detection, we will ensure data privacy by using anonymized, consented datasets and adhering to all licensing agreements for data usage. The model's accuracy and limitations will be transparently reported, with considerations for minimizing false positives and negatives to mitigate adverse impacts on downstream applications. We will acknowledge the dual-use nature of the technology and propose safeguards against misuse. Performance comparisons with existing models will be conducted under identical conditions using standardized metrics to ensure fairness.

**Data Analysis (Procedure and Treatment)**

Llanes-Jurado et al. (2023) had essentially replicated the same methods used in the paper by Taylor et al. (2015) which employed the extraction of 62 hand crafted features in total

minimum, maximum, mean, median, standard

deviation and range.

These statistical features were also computed over

the first and second derivative of the segment.

Which in total would be 18 features for the first category of features

A low pass filtered version of the 0.5s segments of signals of 16hz (hertz) were then again used to compute the min, max, mean, median, std, and then differentiated once to again compute these statistical features and then finally differentiated for the second time to again compute these statistical features to achieve 18 features for this low pass filtered 16hz version of the 0.5s segments of the EDA signals

(40 of which was chosen using a backward selection feature based on a Support Vector Classifier) from a 0.5s segment of signals rather than 5s as it was in the latter paper, then fed to three different classifiers: Logistic Regression, Support Vector Classifier, and a Random Forest Classifier. Subsequently in order to select the best model for each classifier Llanes-Jurado et al. (2023) used hyperparameter tuning for each classifier to select the best model out of each classifier; 0.01, 0.1, 1, 10 and 100 for the C hyperparameter in the Logistic Regression classifier; 200, 400, and 600 & 10, 30 and 50 for the Estimators and Max Depth hyperparameters respectively for the Random Forest classifier; 1, 10, 100 and, 1000 & 0.001, 0.01, 0.1, and 1 for the C and Gamma hyperparameters respectively in the SVM model. The model with highest accuracy was selected among other fitted/trained models of each of the three kinds of classifiers, and finally

Moreover they also had reimplemented the same methodology used in the recent paper by Hossain et al. (2022) where instead of 5s segments as it was in the latter paper they used 0.5s segments of EDA signals. They engineered typical statistical features as with the paper by Taylor et al. (2015) such as mean, median, standard deviation, minimum, maximum, range, and the Shannon entropy from the raw signal itself, as well as the signals first and second order derivatives. Moreover they included as features the optimized coefficients of an autoregressive model (excluding however the bias/intercept coefficient). Finally they used two time frequency transformation methods to extract time frequency features in order to capture non-stationary characteristics (or variables of the data that do not change over time) from the signals. Namely these were wavelet transformation and variable frequency complex demodulation (VFCDM). The mean, standard deviation, median, and range were then computed from each level from the result of a three level wavelet decomposition using a Haar window, and finally VFCDM was applied to the 0.5s segment signals using 64hz, 48hz, 32hz, and 16hz frequencies to extract from these decompositions the last needed features: the standard deviation and mean. A total of 50 features were engineered and extracted from the raw EDA signal data, then reduced to 40 using a Random Forest classifier as a feature selection method to remove redundant features. The features before being fed as input for Support Vector Machine, Gradient Boosted Tree, Random Forest, and Logistic Regression classifiers were standard scaled and normalized using min-max.

In order to select also the best model for each classifier Llanes-Jurado et al. (2023) again used hyperparameter tuning for each classifier together with the use of 5-fold cross validation to select the best model out of each classifier; 0.01, 0.1, 1, 10 and 100 for the C hyperparameter in the Logistic Regression classifier; 200, 400, and 600, 0.01 and 0.1 & 3, 5 and 10 for the Estimators, the Learning Rate, and the Max Depth hyperparameters respectively for the Gradient Boosted classifier; 200, 400, and 600 & 10, 30 and 50 for the Estimators and the Max Depth hyperparameters respectively for the Random Forest classifier; 1, 10, 100 and, 1000 & 0.001, 0.01, 0.1, and 1 for the C and Gamma hyperparameters respectively in the SVM model. The model that had the highest accuracy out of each classifier category was defined as the best model.

This section proposes the use of a hybridized LSTM-SVM and BI-LSTM-SVM in the detection in the last 0.5 s of a 5 s signal segment. This model’s main purpose is to learn from the signal’s temporal evolution. The architecture of this model was inspired by the work of Antczak (2018) and Bento, Belo, and Gamboa (2020), who both used CNN and LSTM to extract features from a raw ECG signal.

We then use the extracted features by the LSTM/BI-LSTM as input to the SVM classifier

How to obtain first and second order derivatives of time series data:

<https://stackoverflow.com/questions/50766061/plot-a-derivative-of-a-time-series-with-a-smoothed-look-in-python>

Antczak, K. (2018). Deep recurrent neural networks for ECG signal denoising. CoRR,

abs/1807.11551. URL: http://arxiv.org/abs/1807.11551. arXiv:1807.11551.

Bento, N., Belo, D., & Gamboa, H. (2020). ECG biometrics using spectrograms and deep neural networks. http://dx.doi.org/10.18178/ijmlc.2020.10.2.929.