

# AI-based Analysis and Classification of EGG signals

**Abstract**—Electrogastrography (EGG) is an examination method for investigating the myoelectrical activity of a stomach. EGG is widely used in research and diagnosis of GI diseases such as nausea, vomiting, functional dyspepsia, gastroparesis, motion sickness, etc. In our study, we demonstrate the usage of machine learning algorithms to predict the functional states of the stomach. Our dataset consists of two states, fasting and postprandial. We utilize feature engineering techniques to extract time domain features, frequency domain features, and fractal features. We then use the extracted features to classify the signals using SVM and Random Forest algorithms. We obtained 89% test accuracy with SVM, 93% test accuracy with Random Forest, and 95% test accuracy with Neural Networks.

**Index Terms**—Electrogastrography, Butterworth filters, Index-based discrete cosine transform (IBDCT), linear minimum mean squared error (LMMSE), Support Vector Machine Classifier (SVM), Random Forest Classifier (RF), Neural Networks(NN),

trogastrogram (EGG) dataset.[DOI: 10.5281/zenodo.3730617] Throughout the entire phase of the research, We used this dataset for segmentation, analysis, and classification. The dataset contains 40 records of 20 healthy individuals (8 Females and 12 Males), and each record consists of 3 channels' recordings. For each subject, EGG was recorded from three locations before (fasting state) and after (postprandial state) a commercial oatmeal (274 kcal). Two 20 minutes recordings (files) are obtained for each subject - fasting and postprandial with a sampling rate of 2 Hz. Overall, each record had 2400 samples for each channel. The method for EGG recording and pre-processing together with subjects' data can be found in Popović et al. 2019 [7].

## B. Data Cleaning

We started by unzipping the data and saving it in a folder as text files then saving the path and label for each record in a data frame for easier access to the records in the future. We wrote a function to separate the records channels into three separate arrays for better analysis, segmentation, and feature engineering results.

1) *Segmentation*: Two levels of segmentation were implemented. The first level segments the records' channels into 3 separate arrays. The second level segments each channel into one-minute intervals. The interval length was chosen based on a suitable CPM value to capture the signal properties which is 3 CPM, which translates to a 1-minute interval [8].

2) *Noise Removal*: We experimented with three denoising methods, the IIR-Butterworth filter, index-based discrete cosine transform (IBDCT), and linear minimum mean squared error estimator (LMMSE). However, LMMSE didn't perform well on our signal. All methods filter for frequencies in the range [0.03,0.25] with sampling frequency  $F_s = 2\text{Hz}$  according to the dataset creators' recommendations [8].

a) *IIR filter - ButterWorth filter*:: We built a band-pass sos filter for the predefined frequency range then we applied the sos filter to the signal using Scipy `sosfiltfilt` function.

b) *Index based cosine transform (IBDCT)*: : We used the following steps to implement the IBDCT method for noise removal.

- Convert signal to frequency space with discrete cosine transform DCT
- Get index (k) of frequencies to keep from the low frequency (Klow) and high frequency (Khigh) thresholds [0.03,0.25] in our dataset.
- Block the frequencies outside the range of Klow to Khigh.
- Convert the signal back using IDCT.

## I. INTRODUCTION

Electrogastrogram (EGG) stands for gastric electrical activity which can be measured non-invasively by the placement of surface electrodes over the stomach [1]. EGG belongs to a wider group of electrophysiological signals. The main components of the EGG signal are the slow-wave and spike potentials [2]. The slow wave originates in the proximal body and propagates normally to the distal antrum with a regular rhythm of approximately 3 cycles per minute (CPM) in healthy humans. The slow wave triggers the onset of spike potentials which in return elicit muscle contractions [3]. EGG is widely used in research and diagnosis of GI diseases such as nausea, vomiting, functional dyspepsia, gastroparesis, motion sickness, etc[2]. Recently, EGG signals are being used in automotive vehicles for gastric motility monitoring [4,5] and in VR applications to detect user sickness [5].

Machine learning (ML) is one of the main applications of artificial intelligence, where computers learn automatically without explicit programming [6]. ML algorithms automatically find patterns in data and use them to make predictions. Support vector Machines (SVM), K-Nearest Neighbor (KNN), and Logistic Regression (LG) are among the most used machine learning algorithms in biomedical engineering [6]. The aim of our research is to classify EGG signals into functional states using ML algorithms. The performance parameters considered in this research are accuracy, F1 score, precision, and recall.

## II. MATERIALS AND METHODS

### A. Data Acquisition

In our search for an open source dataset for EGG, We were only able to locate one, the zenodo three-channel surface elec-

$$k = \frac{L \times f}{F_s \cdot 2} \quad (1)$$

$$f = \frac{k \times F_s}{2L} \quad (2)$$

c) *Linear minimum mean squared error estimator (LMMSE)*:: The motion artifacts present in the EGG are typically manifested as short bursts (i.e., a few seconds or less) of high-amplitude activity (i.e., on the order of mV). Instead of deleting windows of recordings with motion artifacts, we can identify and suppress artifacts with LMMSE (linear minimum mean squared error estimator).[9] The objective is to obtain a clean signal  $e$  from the original signal  $y$  by identifying the noise  $x'$  in each segment and removing it.

$$y = x + e \quad (3)$$

$$e = y - x' \quad (4)$$

$$x' = E[y] + \frac{\max\{0, \text{Var}(y) - \sigma_e^2\}}{\max\{\text{Var}(y), \sigma_e^2\}}(y - E[y]) \quad (5)$$

### III. METHODOLOGY

#### A. Feature Extraction

Feature extraction is the process of building numerical features from raw data that capture and describe the information and relations contained in the raw data. Feature extraction eliminates information redundancy and reduces the data input to machine learning and deep learning algorithms, thus improving their performance.

##### 1) Time domain features::

- Mean
- Mode
- Median
- Standard deviation
- Skewness: A distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed.
- Kurtosis: a measure that describes the shape of a distribution's tails in relation to its overall shape. It has 3 types
  - Mesokurtic: The distribution follows a normal distribution.
  - Leptokurtic: The distribution has a sharp "skinny" peak. The "skinniness" of a leptokurtic distribution is a consequence of the outliers, which stretch the horizontal axis of the histogram graph, making the bulk of the data appear in a narrow ("skinny") vertical range.
  - Platykurtic: distributions with extreme values less than that of the normal distribution.
- RMS: Root means square of the signal.
- Line Length: Total vertical length of the signal.
- Entropy MVN: Entropy for the determinant of the multivariate normal distribution [10].

- Approximate Entropy: a technique used to quantify the amount of regularity and the unpredictability of fluctuations over time-series data [11].
- Permutation Entropy: provides a quantification measure of the complexity of a dynamic system by capturing the order relations between values of a time series and extracting a probability distribution of the ordinal patterns.
- SVD Entropy: Indicator for the number of vectors needed for an accurate explanation of the dataset.
- Lempel-Ziv (LZ) Complexity: A tool to quantify the uncertainty contained in time series data. It measures how diverse the patterns are present in a particular signal.
- Hjorth Parameters: Hjorth Mobility and Hjorth Complexity were used [12].

##### 2) Frequency based features:

- Dominant Frequency (DF) in CPM: DF is the frequency with the maximum amplitude.
- Dominant Power (DP): Maximum magnitude of the power spectrum density(PSD). Power spectrum density measures the power content of a signal against frequency.
- Percentage of PSD that has a higher value than DP/4
- Spectral Entropy: Calculates the Shannon entropy for the normalized PSD of the signal.
- Bandpower: Bandpower is the average power in the band of frequencies. We used the band power between 3-7 CPM, 7-11 CPM, and 11-15 CPM.
- Crest Factor of PSD
- Median Frequency: Frequency where the area under PSD is 50% of total power.
- Mean Power Frequency (MPF).
- MFCCs: Mel-frequency cepstrum (MFCCs) are a small set of features (usually 10-20 features) that describe the overall shape of the spectral envelope of the signal. We call coefficients of the mel-scale (MFCCS) as the mel-scale is a scale that maps the actual frequency to the frequency that human ears will perceive. The mel scale can be obtained as follows: first, we obtain FFT of a signal then we map the powers of the spectrum obtained from the FFT to mel-scale from the equation below. After that we take the logs of the powers at each mel frequency and apply discrete cosine transform (DCT) of mel logs. The MFCCs are the amplitudes of the resulting spectrum. The top 5 MFCCs were used in our research.

$$\text{mel}(f) = 1127 \ln\left(1 + \frac{f}{700}\right) \quad (6)$$

##### 3) Fractal Features:

- Petrosian fractal dimension.
- Katz Fractal dimension: Katz's method calculates the fractal dimension of a sample by calculating the sum and average of the Euclidean distances between the successive points of the sample (L and a, resp.) as well as the maximum distance between the first point and any other point of the sample (d). The fractal dimension of sample (D) then becomes

$$D = \frac{\log(\frac{L}{a})}{\log(\frac{d}{a})} = \frac{\log(n)}{\log(n) + \log(\frac{d}{L})} \quad (7)$$

Where n is L divided by a [13].

- Hurst Index: Quantifies the relative tendency of a signal to regress to the mean or to cluster in a direction [14].

All feature data were scaled to have zero mean and unit variance. Scikit-learn StandardScaler function was used to fit and transform training data then transform validation and testing data.

### B. Machine Learning Algorithms

Dataset was split into 80% training data and 20% testing data for ML algorithms.

1) *Support Vector Machine (SVM)*: SVMs are a non-parametric classification technique that showed great results in the medical diagnostics field. SVMs are suitable for binary classification tasks [15]. For multiclass classification tasks, SVMs divide classes into one versus all approach or compares each class to every other class in a one-to-one approach. SVM performs classification by creating a line or hyperplane between the data of the two classes. SVM use kernels to make classifications. A kernel can be a linear, polynomial, or radial basis function (RBF). In SVM there exists a tradeoff between complexity and correctly classifying data. This tradeoff is controlled by the hyperparameter C. A large value of C means more data is classified correctly. The SVM model hyperparameters were obtained using the GridSearch function in scikit-learn. The final hyperparameters are C=59 and Kernel = RBF.

2) *Random Forest (RF)*: A random forest is an estimator that fits various decision trees on sub-samples of the dataset, where each decision tree is trained on a different subset of the dataset than the rest of the decision trees and uses the averaging technique to perform final classification. Random forests are better than individual decision trees as they reduce over-fitting and improve predictive accuracy. The Random Forest classifier hyperparameters were obtained using the Optuna hyperparameter tuning study. The final hyperparameters are n\_estimators = 809, max\_depth = 39, min\_sample\_split = 4, min\_sample\_leaf = 2, and max\_features = 29. Where n\_estimators represent the number of decision trees the random forest uses, max\_depth represents the maximum depth of trees, min\_sample\_split represents the minimum number of samples required to split an internal node, min\_sample\_leaf represents the minimum number of samples required to be at a leaf node, and max\_features represents the number of features to consider when looking for the best split.

### C. Deep Learning Algorithms

For the Neural Network, the dataset was split into 80% training data, 10% validation data, and 10% testing data.

1) *Neural Networks*: Neural Networks are made from a series of algorithms that work to learn the underlying patterns of data for accurate classification or prediction. Our Network consists of eight dense layers. The first layer takes the input of shape (1,28) followed by six dense layers followed by a final layer for classification into one of the two classes: fasting or postprandial. All layers are followed by batch normalization and the first, second, and second to last layers are followed by a dropout rate of 0.2. The figures below illustrate the structure of the network in depth. The Neural Network uses a subset of the features mentioned above. The features for the neural network were chosen by obtaining the mutual information score for each feature and then choosing the top 28 features. The mutual information score measures the similarity between each feature and the labels. The Neural Network training time is 6.15 minutes for 300 epochs with batch\_size = 64.

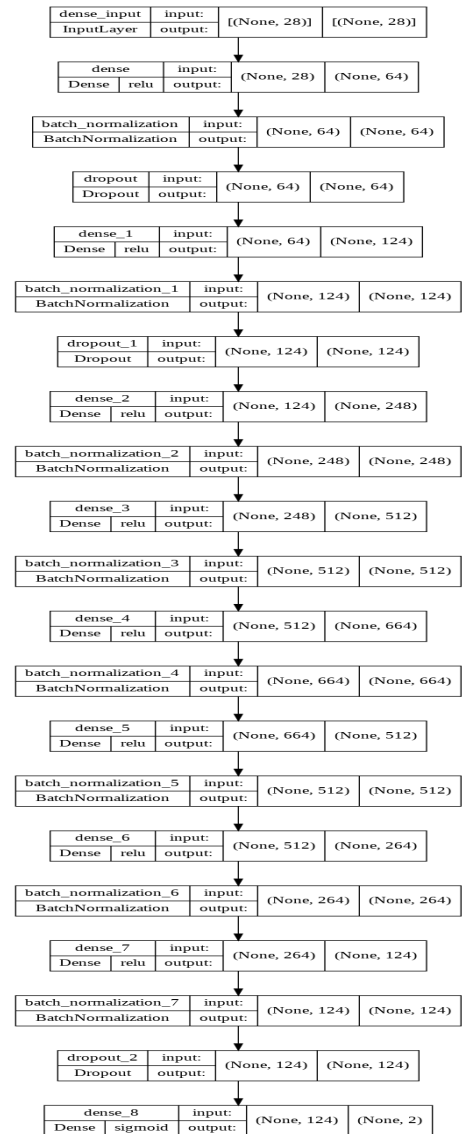


Fig. 1. Neural Network Model Structure



Fig. 2. Neural Network Model Pipeline

#### D. Classification Parameters

To analyze and compare the results of the algorithms used for classification we calculate the accuracy, precision, recall, and F1 score.

# Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Fig. 3. Confusion Matrix

1) *Accuracy*: Accuracy is the total number of correctly classified data points over the total number of predictions.

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \quad (8)$$

2) *Precision*: The number of correctly classified positive predictions over the total number of positive predictions made by the classifier.

$$Precision = \frac{T_p}{T_p + F_p} \quad (9)$$

3) *Recall*: The number of correctly classified positive predictions over the total number of positive data points in the dataset.

$$Recall = \frac{T_p}{T_p + F_n} \quad (10)$$

4) *F1 score*: The F1 score is the harmonic mean of precision and recall. It's a measure of the accuracy of an algorithm on data.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (11)$$

#### IV. RESULTS

In this study, the main objective was to classify electrogastrography (EGG) signals into two classes (fasting and postprandial). We implemented SVM and Random Forest machine learning algorithms and dense neural networks. We used time domain features such as mean, mode, standard deviation, and kurtosis; frequency domain features such as dominant frequency, dominant power, entropy, and MFCCs;

and fractal features such as Hurst fractal dimension and Katz fractal dimension. Our results show that the Neural Network achieved a maximum F1 score, precision, recall, and accuracy of 0.95, 0.95, 0.95, and 0.95 respectively.

Table I shows the classification results of the models on the binary classification problem of our dataset of labels fasting and postprandial.

TABLE I  
CLASSIFICATION METRICS OF MODELS

Model	Accuracy	Precision	Recall	F1 score
SVM	0.89	0.88	0.89	0.88
RF	0.93	0.93	0.92	0.93
NN	0.95	0.95	0.95	0.95

#### REFERENCES

- [1] N. B. Popović, N. Miljković and T. B. Šekara, "Electrogastrogram and electrocardiogram interference: Application of fractional order calculus and Savitzky-Golay filter for biosignals segregation," 2020 19th International Symposium INFOTEH-JAHORINA (INFOTEH), 2020, pp. 1-5, doi: 10.1109/INFOTEH48170.2020.9066278.
- [2] N. Jovanović, N. B. Popović and N. Miljković, "Empirical mode decomposition for automatic artifact elimination in electrogastrogram," 2021 20th International Symposium INFOTEH-JAHORINA (INFOTEH), 2021, pp. 1-6, doi: 10.1109/INFOTEH51037.2021.9400683
- [3] Chang FY. Electrogastrography: basic knowledge, recording, processing and its clinical applications. J Gastroenterol Hepatol. 2005 Apr;20(4):502-16. doi: 10.1111/j.1440-1746.2004.03751.x. PMID: 15836697.
- [4] Gruden T, Popović NB, Stojmenova K, Jakus G, Miljković N, Tomažič S, Sodnik J. Electrogastrography in Autonomous Vehicles—An Objective Method for Assessment of Motion Sickness in Simulated Driving Environments. Sensors. 2021; 21(2):550. https://doi.org/10.3390/s21020550
- [5] Popović NB, Miljković N, Stojmenova K, Jakus G, Prodanov M, Sodnik J. Lessons Learned: Gastric Motility Assessment During Driving Simulation. Sensors. 2019; 19(14):3175. https://doi.org/10.3390/s19143175
- [6] Park, C., Took, C.C. & Seong, J.K. Machine learning in biomedical engineering. Biomed. Eng. Lett. 8, 1–3 (2018). https://doi.org/10.1007/s13534-018-0058-3
- [7] Popović, Nenad B., Miljković, Nadica and Popović, Mirjana B. "Simple gastric motility assessment method with a single-channel electrogastrogram" Biomedical Engineering / Biomedizinische Technik, vol. 64, no. 2, 2019, pp. 177-185. https://doi.org/10.1515/bmt-2017-0218
- [8] Nanivadekar AC, Miller DM, Fulton S, Wong L, Ogren J, Chitnis G, McLaughlin B, Zhai S, Fisher LE, Yates BJ, Horn CC. Machine learning prediction of emesis and gastrointestinal state in ferrets. PLoS One. 2019 Oct 18;14(10):e0223279. doi: 10.1371/journal.pone.0223279. PMID: 31626659; PMCID: PMC6799899.
- [9] Gharibans AA, Smarr BL, Kunkel DC, Kriegsfeld LJ, Mousa HM, Coleman TP. Artifact Rejection Methodology Enables Continuous, Non-invasive Measurement of Gastric Myoelectric Activity in Ambulatory Subjects. Sci Rep. 2018 Mar 22;8(1):5019. doi: 10.1038/s41598-018-23302-9. PMID: 29568042; PMCID: PMC5864836.
- [10] Neeraj Misra, Harshinder Singh, Eugene Demchuk, Estimation of the entropy of a multivariate normal distribution, Journal of Multivariate Analysis, Volume 92, Issue 2, 2005, Pages 324-342, ISSN 0047-259X, https://doi.org/10.1016/j.jmva.2003.10.003
- [11] Delgado-Bonal A, Marshak A. Approximate Entropy and Sample Entropy: A Comprehensive Tutorial. Entropy (Basel). 2019 May 28;21(6):541. doi: 10.3390/e21060541. PMID: 33267255; PMCID: PMC7515030
- [12] Galvão F, Alarcão SM, Fonseca MJ. Predicting Exact Valence and Arousal Values from EEG. Sensors (Basel). 2021 May 14;21(10):3414. doi: 10.3390/s21103414. PMID: 34068895; PMCID: PMC8155937.

- [13] Chu Kiong Loo, Andrews Samraj, Gin Chong Lee, "Evaluation of Methods for Estimating Fractal Dimension in Motor Imagery-Based Brain Computer Interface", *Discrete Dynamics in Nature and Society*, vol. 2011, Article ID 724697, 8 pages, 2011. <https://doi.org/10.1155/2011/724697>
- [14] Alejandro A. Torres-García, Omar Mendoza-Montoya, Marta Molinas, Javier M. Antelis, Luis A. Moctezuma, Tonatiuh Hernández-Del-Toro, Chapter 4 - Pre-processing and feature extraction, Editor(s): Alejandro A. Torres-García, Carlos A. Reyes-García, Luis Villaseñor-Pineda, Omar Mendoza-Montoya, *Biosignal Processing and Classification Using Computational Learning and Intelligence*, Academic Press, 2022, Pages 59-91, ISBN 9780128201251, <https://doi.org/10.1016/B978-0-12-820125-1.00014-2>.
- [15] Auria, Laura and Moro, R. A., Support Vector Machines (SVM) as a Technique for Solvency Analysis (August 1, 2008). DIW Berlin Discussion Paper No. 811, Available at SSRN: <https://ssrn.com/abstract=1424949> or <http://dx.doi.org/10.2139/ssrn.1424949>