Digital Biomarker Discovery for Eye Disorders using EEG Data

Prepared by Kui Hong Lim (MSc of Medical Informatics)
Email: kuihong.lim@students.fhwn.ch | Github
Date : 08 January 2024
Module : Digital Biomarker
University of Applied Sciences and Arts Northwestern Switzerland

---

Objective

The objective of this project is to develop a digital biomarker using EEG data that can accurately identify individuals with eye disorders characterized by the need for long and stronger blinks. Students will analyze EEG signals obtained from four electrodes to find a biomarker that can be used to diagnose this eye disorder efficiently.

Project Tasks

Index	Task
1	Data download and Preprocessing
2	Feature Extraction
3	Biomarker Selection and Visualization
4	Model Development, Validation and Evaluation

1. Data download and preprocessing

1.1 Data cleaning and understanding

Reading the EEG data from the provided four electrodes by running the file collected during the long blink and short blink session with equal length (510 points pro session). Conduct the preliminary data processing and visualization steps to gain some insight from the EEG recording, started with the long blink data:

- Split and clean the data
- Select the first 510 points for visualization

The initial visualization of the data reveals that the amplitudes of the four signals range from 400 Hz to 1200 Hz. The first, third, and fourth electrode signals exhibit a similar wave pattern with powerline

noise, while the second electrode shows less perturbation with powerline noise.

1.2 Load data, and parsing it from string to appropriate data type. Then, select the first 510 points for visualization.

1.3 Data format compatibility

In this step, the MNE-Python package is used. MNE is an open source tool for exploring, visualizing, and analyzing human neurophysiological data, such as EEG. A .fif file needs to be generated from csv to fit the raw data format used in this package for analysis.

1.4 Scaling the value

A healthy human EEG displays a certain pattern of activity correlate with how awake a person is. The range of frequencies observed are between 1 and 30 Hz, with amplitudes vary between 20 and 100 μV". However, the raw signals of the four electrodes provided exhibit different magnitude range: ("500 - 1100" or "400 - 1200").

The objective here is to standardize the magnitude of the variations in the raw signale from the four electrodes into a consistent range between 20 and 100 μV. Standardizing the range helps to achieve consistent and comparable magnitudes across the electrodes. Additionally, scaled data of similar amplitudes can lead to better convergence and efficient training for machine learning.

In order to rescale the raw signal values to the desired range between 20 and 100 μV, min-max scaling is applied. Transposing the DataFrame is done to ensure that each row corresponds to a specific time point, which aligns with the typical structure of time-series data.

1.5 Artifact detection

MNE-Python includes tools for automated detection of certain artifacts such as blinks; one can always visually inspect the data to identify and annotate artifacts as well. Before looking at artifacts, SSP projectors is set aside in a separate variable and then remove them from raw object using del_proj() method so that the data can be inspected in its original, raw state. Signal-space projection (SSP) is a technique for removing noise from EEG signals by projecting the signal onto a lower-dimensional subspace.

1.6 Apply low-frequency drifts

Low-frequency drifts are most readlily detected by visual inspection using the basic plot() method, it is helpful to plot a relatively long time span and to disable channel-wise DC shift correction. In this context, a 2 seconds is plotted. It is observed that there are approximately 12 pulses occuring every

0.25 seconds. This pattern results in around 50 pulses per second, corresponding to a frequency of 50 Hz. Then the power spectral density is applied to unveil the distribution of power across various frequencies in the EEG signal.

The Power Spectral Density (PSD) examines how the power of the signal is distributed across different frequencies. The PSD analysis reveals an increase in noise magnitude from 10 Hz to 30 Hz, with a notable peak at 50 Hz, likely indicating power line noise (commonly at 50 Hz in the EU). Additionally, other low-power frequencies are detected around 56 Hz and between 72 to 80 Hz. Given this observation, it is planned to apply a notch filter later, which is commonly use to eliminate specific inteferences frequences such as power line noise. The goal is to mitigate pertubations caused by unwanted frequency in the data.

1.7 Filtering EEG data

Before initiating the filter process, the preprocessed raw data is converted into a DataFrame.

Filtering is an important process of removing unwanted frequencies such as noise and artifacts from a signal and preserve the desired frequencies to extract relevant information.

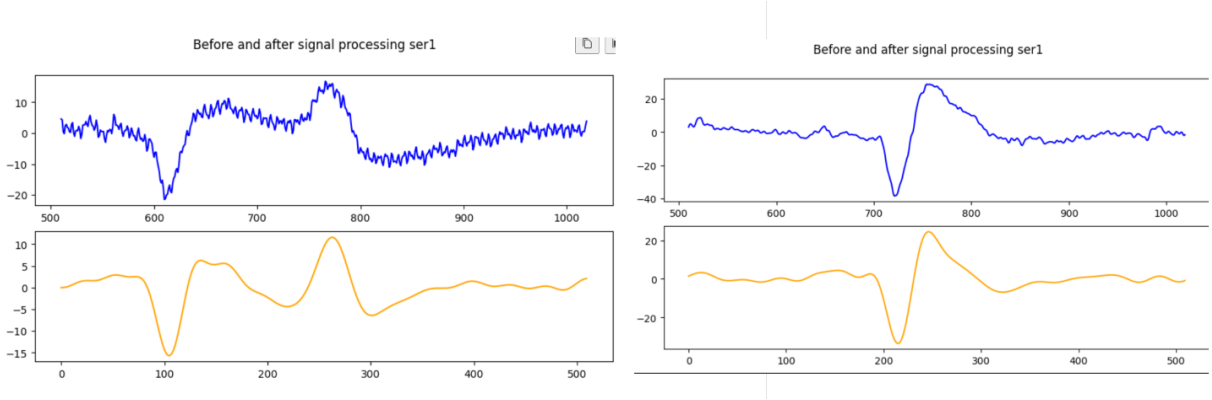
A window of data is segmented for further processing. The preprocessing employs third-order filters, including a median filter for noise removal and retention of important features. Subsequent processing involved additional filtering steps using both low-pass and high-pass filters. Note that for both the high-pass and low-pass filters, the power does not drop of sharply right at the cutoff frequency, rather, there is a roll-off, indicating a gradual decrease in power over a range of frequencies. The chosen Butterworth filter, with a fifth-order design, is employed, configuring cutoff frequencies at 0.5 Hertz for the low-pass filter and 50 Hertz for the high-pass filter. The Butterworth filter is known for maintaining a flat frequency response in the passband. To address power interference at 50 Hertz, a notch filter is then applied.

In the final step, a finite impulse response (FIR) filter was designed using the "signal.firwin" function from the SciPy library. FIR filter is widely used for signal processing tasks such as EEG signal processing. This digital filter was configured with 430 taps (coefficients). The cutoff frequencies for this FIR filter were set in the range of 0.01 to 0.06. The lower cutoff frequency of 0.01 Hz is chosen to preserve very low-frequency components in the EEG signal. This includes elements like baseline shifts or gradual trends, which can be important features. The upper cutoff frequency of 0.06 Hz indicates that the filter permits relatively higher-frequency components within the low-frequency range to pass. This selective allowance is designed to capture specific frequency components while mitigating both lower-frequency drifts and potential higher-frequency noise. Essentially, FIR filter is used to isolate targeted frequency components within the EEG signal, striking a balance between retaining essential low-frequency information and minimizing unwanted noise. The chosen cutoff frequencies

and the overall design aim to enhance the quality and interpretability of the EEG data for subsequent analyses.

The application of FIRWIN demonstrates a notable improvement in the filtered EEG signal for blinks, resulting in a smoother and cleaner representation compared to the preprocessed state. This enhanced clarity enhances the interpretability and quality of the EEG data for subsequent analyses. The images are placed side by side for comparison.

Upon inspecting the two filtered EEG signals for both long and short blink data, it becomes apparent that the short blink cycle (image on the right) is shorter, particularly noticeable in the first, second, and fourth electrodes. Conversely, the long blink cycle (image on the left) is longer. Additionally, more fluctuation is observed in the third electrode for both long and short blinks, suggesting the potential influence of cognitive activity being recorded in this electrode. An example of one of the images from the four electrodes is displayed below:



1.8 Data preprocessing and filtering on short blink data

The same steps to process and filter long blink data are applied on the short blink data. In contrast to the long blink data, the short blink data exhibits reduced power line noise. However, it is affected by artifacts and noises within the amplitude range of 600 Hz to 900 Hz. Many small fluctuations are noticed before the rise or descent to the peak at 60 seconds (an indication of eye blink). The Power Spectral Density (PSD) analysis indicates a gradual increase in noise from 10 Hz towards the end of the segment, featuring a distinct peak at around 56 Hz.

2. Feature Extraction

2.1 Binarization

The dataset, comprising 25,500 rows, is segmented into batches, with each batch corresponding to a blink epoch of 2 seconds. Each epoch consists of 510 rows, resulting in a total of 50 blinks (including

both long and short blinks). This segmentation leads to 50 chunks, each containing 510 data points. Subsequently, a new column is added to classify the filtered dataset into two classes, denoted as 0 and 1. To facilitate time series processing, a function is created to split the data into 100 batches, each containing 4 rows of 510. The batched data are then transposed, treating each array as a unit model for processing.

2.2 Statistical Extraction of Features

Subsequently, statistical feature extraction is employed to identify important signal attributes such as amplitude variation. These features offer valuable insights into the signal and contribute to the system's accuracy. While these statistical features provide valuable information, studies suggest that relying solely on them may not yield a comprehensive diagnosis as they may fall short in capturing the full complexity of brain signals. The statistical features used are:

1. Mean

Mean represents the average value of the EEG signal across the selected time domain.Changes in the mean may indicate shifts in the baseline or overall amplitude of the signal.

2. Variance

Variance provides information about the spread or dispersion of amplitude values in the signal. Specifically, variance quantifies how much individual data points in the signal deviate from the mean (average) value. A higher variance indicates greater variability in the amplitude values, while a lower variance suggests that the values are more tightly clustered around the mean.

3. Energy

Energy is calculated to find the spreading out of the data values around the mean.Energy provides information about the intensity or power of the signal. Peaks in energy may correspond to periods of increased neural activity or certain events captured by the EEG.

4. Standard Deviation

Standard deviation measures the amount of variability or dispersion of the EEG signal values around the mean.A higher standard deviation suggests greater variability in signal amplitudes. It can be indicative of the signal's overall volatility or the presence of irregularities.

5. Kurtosis

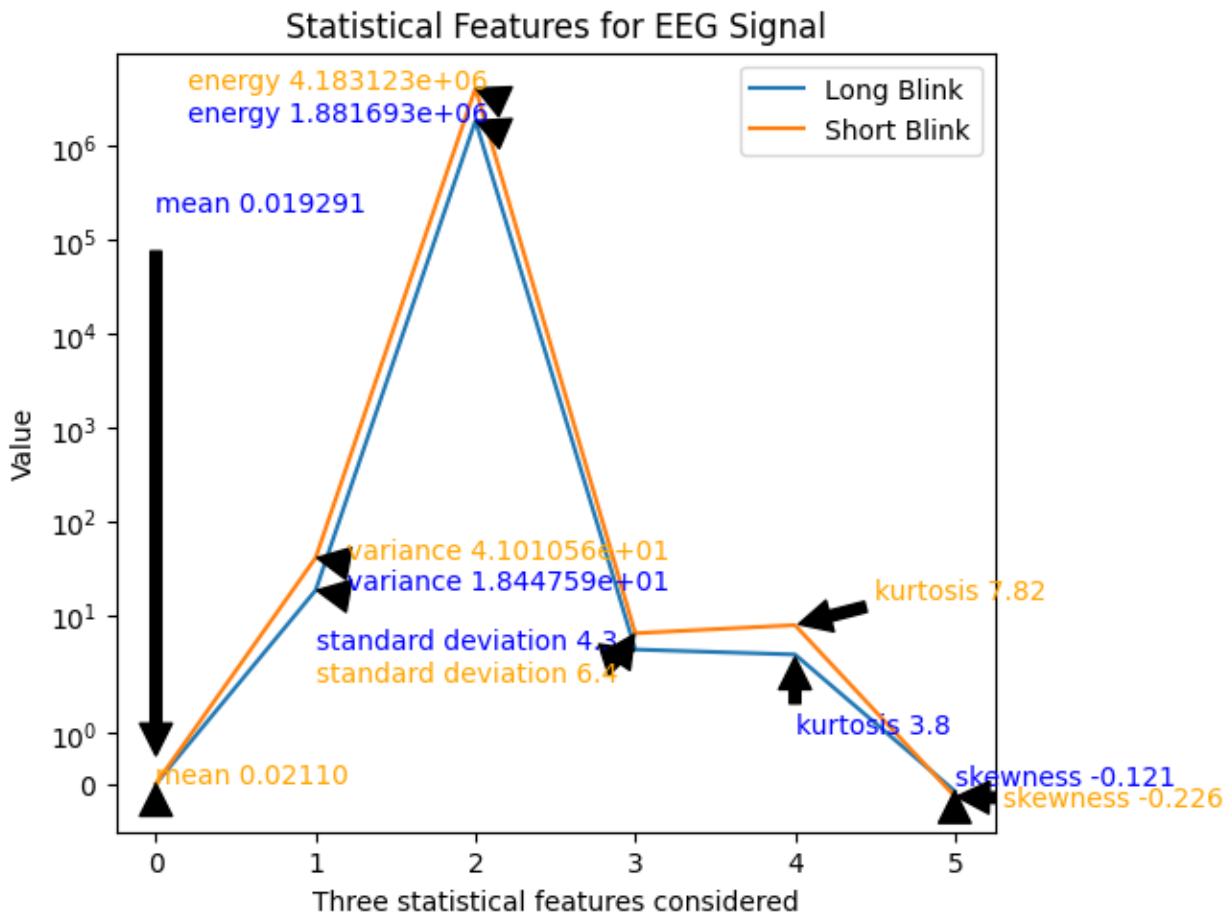
In a probability distribution, it measures the outliers present.

6. Skewness

Skewness describes asymmetry from the normal distribution in a set of statistical data, as data becomes more symmetrical as its value approaches zero. Normally distributed data, by definition has little skewness and on other hand positively skewed or right sided skewed data has positive and negatively skewed or left sided skewed has negative value.

Statistical features	**Long blink**	**Short blink**
Mean	0.019291	0.02110
Variance	1.844759e+01	4.101056e+01
Energy	1.881693e+06	4.183123e+06
Standard Deviation	4.3	6.4
Kurtosis	3.78	7.82
Skewness	-0.0121	-0.0226

It is observed that the mean and skewness do not show much variation between the two classes. On the other hand, variance, energy, standard deviation, and kurtosis show some good variation between the two classes. Therefore, these parameters are considered for analysis.



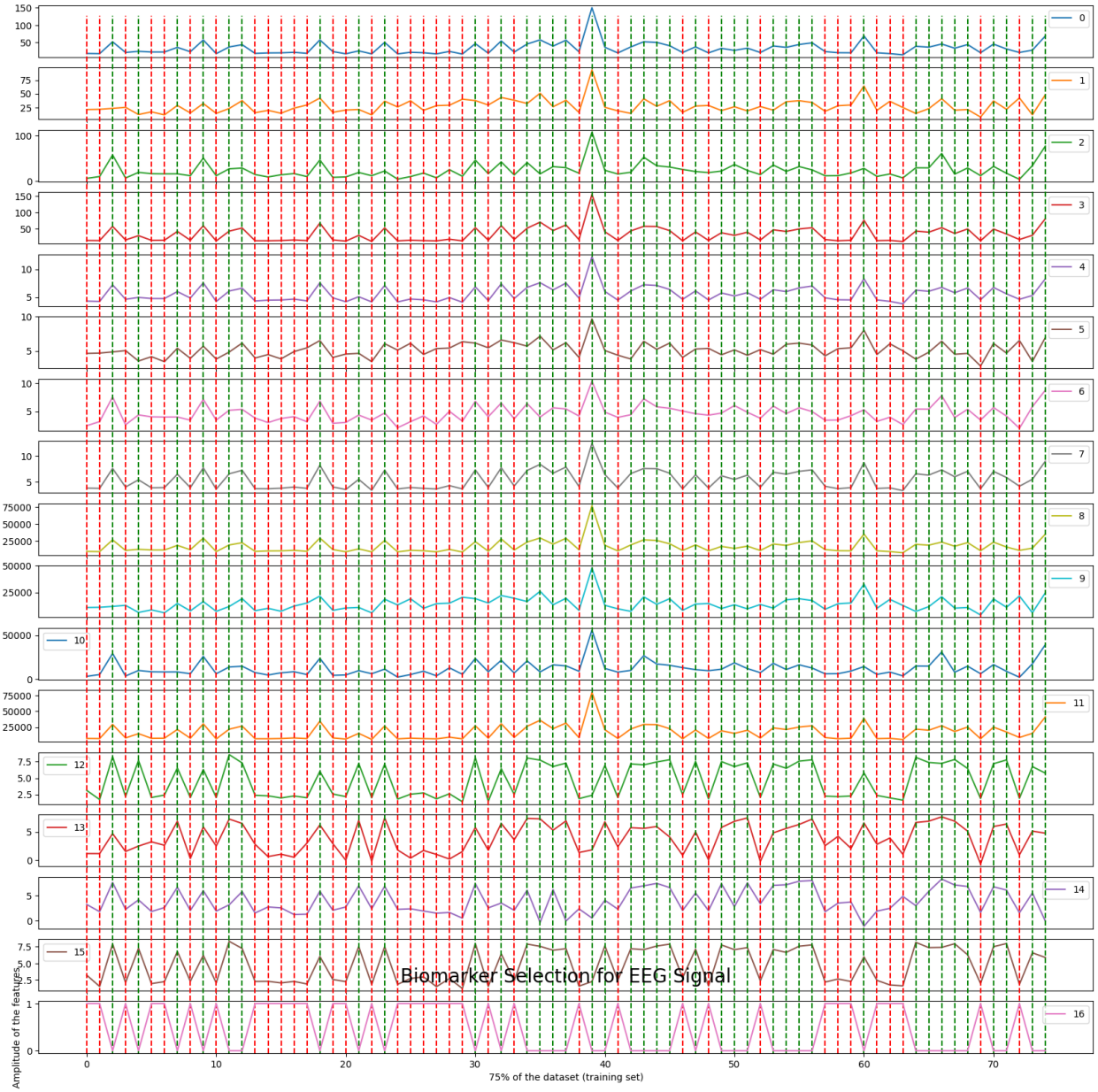
3. Biomarker Selection and Visualization

In the context of biomarker selection, the random forest algorithm, which employs multiple decision tree classifiers on various subsets of the dataset is applied. Averaging is used to enhance predictive accuracy and control overfitting.

To determine the optimal number of trees, five-fold cross-validation is applied. Interestingly, all estimators `n_trees in [50, 100, 200, 500]` achieved perfect accuracy on the training data. As a result, the retained features are applied with the smallest number of trees `n_estimators = 50`.

The selected biomarkers consist of 16 statistical features (4 features for each of the 4 electrodes). These features are extracted from 75% of the datasets, and two classes are considered. The visualized plot illustrates the sequences of statistical features from top to bottom as *Variance*, *Energy*, *Standard deviation*, and *Kurtosis*. Long blinks are annotated with a green dotted line, while short blinks are annotated with a red dotted line. The visual inspection indicates that the majority of long blinks (associated with eye disorder) appear flat, while short blinks (healthy state) display spikes. This distinction is particularly noticeable, especially in the *Kurtosis* feature.

```
Text(0, 0.5, 'Amplitude of the features')
```



4. Model Development, Validation and Evaluation

The first model architecture is a simple sequential neural network model for classification using `Keras library`. This model incorporates two `dense layers`, hidden layers with 20 neurons, and output layer with 2 classes **(long blink = 1, short blink = 0)**. The input layer is 8 (selected features). Activation function are specified for both hidden and output layers. Additionally, an optimizer, different batch sizes, and loss function are configured to quantify the difference between predicted

binary outcomes and actual binary labels during testing at 40 epochs.To fine-tune the model, three sets of hyperparameters (activation function, loss function, and optimizer) are employed. Following model training, the `classification_report()` function from the `sklearn library` is applied to generate the metrics commonly used to assess the quality of the model. From the report, **Precision**, **Recall**, **F1-Score** are specifically considered in the evaluation.

Upon analyzing the test dataset, which comprises 11 short blinks and 14 long blinks, set C (characterized by the activation function `Relu` in the hidden layer, `Softmax` in the last layer, `categorical_crossentropy` as loss function, `Adam` as optimizer and `batch size` as 20) emerges as the optimal choice among the three sets of hyperparameters. It demonstrates a relatively low loss score and excels in test accuracy, precision, recall, and F1-score.

Model Analysis

Model 1

Set	Activation Function *(Hidden layer)*	Activation Function *(Last layer)*	Loss Function	Optimizer	Batch Size
A	Relu	Sigmoid	binary_crossentropy	RMSProp	20
B	Relu	Softmax	categorical_crossentropy	RMSProp	20
C	Relu	Softmax	categorical_crossentropy	Adam	20

Model 1 result

Set	Loss Score	Test Accuracy
A	5	0.88
B	64	0.96
C	4	0.96

Classification report: Set A

Class	Precision	Recall	F1-score
0	0.83	0.91	0.87
1	0.92	0.86	0.89

Classification report: Set B

Class	Precision	Recall	F1-score
0	1.00	0.91	0.95
1	0.93	1.00	0.97

Classification report: Set C

Class	Precision	Recall	F1-score
0	1.00	0.91	0.95
1	0.93	1.00	0.97

Following that, the dataset is processed using the classical machine learning model, Support Vector Classifier (SVC), yielding amazing results:

SVC	Score
Classical SVC on the training dataset	1.00
Classical SVC on the test dataset	0.96

Classification report: SVC

Class	Precision	Recall	F1-score
0	0.92	1.00	0.96
1	1.00	0.93	0.96

When iterating with the same dataset, there is a noticeable decrease in test accuracy. Therefore, to assess the model's efficiency and performance more robustly, it is advisable to evaluate it on a different dataset.

5. Challenges and conclusion

Grasping the intricacies of EEG signals, encompassing both conceptual understanding and the practical application of mathematical concepts in programming, is a challenging and time-consuming endeavor. In many scholarly works, eye blinks are commonly treated as artifacts during

EEG signal processing. However, there is lack of literature regarding how to leverage eye blinks as potential biomarkers for visual disorders.

In conclusion, this project aimed to develop a digital biomarker using EEG data for the accurate identification of individuals with eye disorders characterized by distinct blink patterns. The initial phase involved comprehensive data preprocessing, including data cleaning, formatting, scaling, and artifact detection. To standardize the magnitude of EEG signal variations, min-max scaling was applied, ensuring a consistent range between 20 and 100 μV across the four electrodes. Subsequent artifact detection, low-frequency drift correction, and filtering steps, including third-order filters and a finite impulse response (FIR) filter, were implemented to enhance data quality and interpretability.

Feature extraction focused on statistical measures such as mean, variance, energy, standard deviation, kurtosis, and skewness. Random Forest, a robust algorithm for biomarker selection, was employed to identify important features. Visualization of the selected biomarkers revealed distinctive patterns between long and short blinks, particularly in the Kurtosis feature.

The model development phase utilized both a neural network and a classical machine learning approach, Support Vector Classifier (SVC). Evaluation metrics such as precision, recall, and F1-score were employed to assess model performance. Notably, set C of the neural network hyperparameters demonstrated optimal results. It is noteworthy that the Support Vector Classifier has also demonstrated superior performance. However, it is essential to note a decrease in test accuracy upon iteration with the same dataset, suggesting potential limitations or overfitting. To ensure robust evaluation, further testing on different datasets is recommended.

This project provides a comprehensive framework for leveraging EEG data as a potential biomarker for eye disorders. All references papers and codes can be found on Markdown.