**Research goals**

We will present our research goals, their treatment and their evolution through the past couple of weeks in this section. Dividing the research goals into 4 logical blocks, we focused on these following major goals:

1. Feature Engineering.
2. Application of learning algorithms on the extracted features.
3. Analysis and optimization of the learning models.
4. Creating a coherent end-to-end application for final submission and the competition.

Our target is to accomplish these goals and push the classification accuracy to as high as possible with the given experimental setup.

***Feature Engineering:***

As mentioned in the sections above, feature engineering was added to our set of research goals after we realized that we would have to extract the features from the raw images since no preprocessed datasets were available. Further, feature engineering translated into the following tasks:

1. In-depth knowledge of the domain (the Where?)
2. The features that we should extract based on the domain knowledge (the What?)
3. How to extract these features (the How?)

In an attempt to answers these questions above, we turned to the literature in the field of handwriting recognition, signature forensics & image processing and in due time we started our search for an image processing platform that we could use for the feature extraction algorithms that we would use. While navigating through the domain at hand, a major portion of our focus shifted towards extracting the features from the raw image data.

We decided to use Intel’s OpenCV library for image processing. The features that represent the individual idiosyncrasies of the signatures can be divided into related groups:

1. Features based on the whole image – Macro Features.
2. Features based on the character level changes – Micro Features.
3. Features based on the DTW approach – Dynamic Time Warping.

We decided to go ahead with using the Macro Features and the Micro Features initially. However, due to our incomplete understanding of the extraction process of the Micro Features and the limited amount of time we had on our hands, we researched solely on the Macro Features. The Macro Features gave us information about the image on the whole and generated 11 features.

These features were: entropy of gray levels, gray level threshold, number of black pixels, number of interior contours, number of exterior contours, number of vertical slope components, number of horizontal slope components, number of positive slope components, number of negative slope components, slant and height. We aimed our research towards implementing these extraction algorithms using the OpenCV library from the given set of scanned signature images. A lot of background research went into analyzing how images are represented using Chain Codes and various other techniques like thresholding using the ‘Otsu’ algorithm, Sobel operators for edge detection etc.

***Application of learning algorithms on the features:***

After some deliberation, we realized that using the extracted features directly would not help us solve the proposed problem. Thus we needed to transform our feature-set into a new representative space which would make it compatible to the learning algorithms

We faced the following challenges posed by this goal:

1. How to use the features to build our learning model.
2. What feature transformation would be a helpful representation?
3. Which learning algorithms can be applied?

We transformed the original feature space into a distance space by taking ‘Inter-class’ distances between the known forgeries of the given original signature samples of a given writer. Similarly, we took the ‘Intra-class’ distances of the original samples of a given writer. This allowed us to see how forgeries are distant from the originals in this new space.

The learning algorithms we are applying to the final dataset include Neural Networks, Decision Trees & SVM and SVM with Adaptive Boosting. We have built some preliminary models using these algorithms and will present the results we have obtained till now in the later sections.

***Analysis and optimization of the learning models***

After running our preliminary tests on the extracted dataset, we are focusing on the alternatives available to us at this point:

1. Extract more features from other feature categories mentioned previously.
2. Exploit the existing feature set by improving and tweaking the learning algorithms.

Since we had decided to use only the Macro features for constructing the dataset our aim is to determine if adding more features and more feature types will be helpful to improve the performance of the models we have obtained and the other option is to solely focus on the dataset we have obtained. There are pros and cons for both the approaches here, adding new features will be time consuming and might not guarantee significant improvements. It will also make us revisit and add more feature extraction algorithms on the OpenCV platform but it might empower us with more highly correlated features. On the other hand, trying to optimize the learning algorithms may prove unworthy since some important features might be missing but at the same time tweaking the algorithms and combining them might give us good results.

***Creating a coherent end-to-end application for final submission and the competition***

Currently, our application is fragmented into various modules. The major module sits on the OpenCV platform while another component written in Python used this OpenCV module to extract the image features and then generate the datasets. The third part is the set of learning algorithms for which are currently using Weka. However, our aim to provide a single point of entry (a set of original and forgery samples) and a single point of exit which would be a classification – deciding whether the sample in question is an original or a forgery.