# Improving MNIST Accuracy through Feature Engineering

**Aina Shafqat**
2021-EE-058
Department of Electrical Engineering
University of Engineering and Technology
2021ee58@student.uet.edu.pk

**Ayesha Ahmad**
2021-EE-052
Department of Electrical Engineering
University of Engineering and Technology
2021ee52@student.uet.edu.pk

## 1 Background

The domain of our project is feature engineering of images, with a primary goal of improving the accuracy of the machine learning model, to climb up in the kaggle competition rankings.

### 1.1 Importance of Research

The accuracy of the machine learning model is vital in any application. A low accuracy can result in huge losses of money and credibility. An example of this is Zillow's erroneous property valuations, which resulted in losses exceeding $500 million stemming from their flawed algorithms. This underscores the significance of algorithm accuracy in real-world applications and the machine learning industry.

### 1.2 Aim

Our project will use feature engineering to improve the accuracy of the Support Vector Machine (SVM) classifier on the MNIST dataset to compete with the Berkley students in the Kaggle competition.

## 2 Data Source

The dataset to be used in this paper is MNIST which is provided in the homework 1 (h.w1). It is contained in the file "mnist data.mat" and comprises the following: • Number of samples: The dataset consists of 60,000 labeled digit images specifically curated for training purposes. • Number of features: Each digit image is represented as a grayscale image with dimensions of 28 x 28 pixels. • Test set: An additional set of 10,000 digit images designated for testing purposes All the images are accompanied by one of 10 possible labels, corresponding to the digits 0 through 9.

## 3 Relevance to this class

This project will implement concepts taught in class, such as hyper-parameter tuning, Support Vector Machine (SVM) classifiers, and feature engineering. Through careful design and selection of relevant features, our goal is to raise model accuracy.

### 3.1 Feature Engineering

Feature engineering involves extracting, refining, and selecting features to boost model accuracy.

Preprint. Under review.

Initially, diverse datasets and case studies are explored to identify distinctive features tailored to specific problem-solving needs. Through extraction and construction methods, features are shaped to contain essential information.

Subsequently, feature selection techniques aid in identifying the most influential features, facilitating model development.

Finally, evaluating models using these selected features enables performance assessment and identification of areas for improvement.

Proficiency in these techniques enhances model accuracy and enables efficient resolution of real-world challenges by ensuring that models are trained on the most relevant and informative features, thereby maximizing predictive performance.

### 3.2 Hyper-parameters

Hyper-parameters play a crucial role in training machine learning models for optimal performance. In our project, we'll explore various hyper-parameter optimization techniques to ensure that our SVM models are well trained. By adjusting hyper-parameters such as the regularization parameter and kernel parameters, we aim to find the optimal settings that maximize model accuracy. Additionally, hyper-parameter optimization helps us combat overfitting and underfitting, ensuring that our models generalize well to unseen data. By systematically training hyper-parameters, we can enhance the robustness and effectiveness of our SVM classifiers, ultimately improving our performance in the competition.

### 3.3 SVM

SVMs encounter challenges with non-linearly separable data and sensitivity to outliers, impacting classification accuracy. Techniques like soft-margin SVMs and outlier detection methods address these issues for strong performance. To optimize SVM models, we'll employ advanced hyperparameter optimization and explore various kernel functions tailored to our feature-engineered dataset.

This highlights the importance of feature engineering and how our project applies class concepts to real-world situations. Overall, integrating these techniques is expected to enhance model performance.

### 3.4 Project work Composition

The components of the project and the expected time division is shown in Table 1.

Table 1: Work Composition

| Divisions | % Time |
|---|---|
| Literature survey | 30% |
| Coding | 50% |
| Theory | 20% |