

# Technologies for Multimodal Data Representation and Archives

## Project Report

### 1. Introduction

In this project, we explicitly chose the image classification domain, focusing on the automatic recognition of letters engraved on ancient Byzantine seals. This task poses a challenging real-world scenario due to the historical nature of the data and the visual degradation of the artifacts.

The goal of this work is to design, train, and evaluate machine learning models capable of correctly classifying individual letters extracted from images of Byzantine seals. Unlike modern character recognition tasks, this dataset is affected by noise, erosion, low contrast, and significant stylistic variability, making the image classification problem particularly complex and suitable for experimentation with deep learning techniques.

### 2. Dataset and Problem Description

The selected task is an image classification problem formulated as an Optical Character Recognition (OCR) task on historical artifacts. Specifically, the objective is to recognize individual characters cropped from images of ancient Byzantine seals. These seals are small lead artifacts bearing inscriptions in Byzantine Greek, historically used to authenticate official documents during medieval times.

The dataset used in this project originates from the BHAI (Byzantine Historical Artificial Intelligence) project, which released a collection of more than 2,000 character crops extracted from hundreds of Byzantine seals dating from different historical periods. Unlike conventional OCR datasets, the images present unique challenges due to the three-dimensional nature of the engraved characters, surface erosion, irregular lighting conditions, and significant stylistic variability. Additionally, characters appear on different materials such as lead, silver, bronze, and gold, introducing further visual heterogeneity.



*Figure 1: Example character crops from the dataset, highlighting visual variability, surface degradation, and material-dependent appearance that make the classification task particularly challenging.*

Due to a severe class imbalance in the original dataset, this project focuses on a subset of the 24 most frequent character classes, including letters and symbols. The resulting dataset consists of approximately 1,610 images used for training and validation, and 430 images reserved for testing. Each image is labeled with a class identifier corresponding to a specific character, and filenames follow the format `<class_id>_<index>.jpg`, with label mappings provided in a companion file.

The distribution of samples across classes remains highly uneven, with a small number of characters dominating the dataset while others are represented by significantly fewer samples. This imbalance poses an additional challenge for supervised learning and must be taken into account during model training and evaluation.

The character crops vary considerably in spatial resolution, ranging from approximately  $60 \times 60$  pixels to over  $180 \times 180$  pixels. As a result, image resizing is required to ensure a consistent input size for neural network models. Given the limited size of the dataset, a cautious data-splitting strategy is adopted, allocating only a small portion of the training data for validation purposes, while preserving a fixed test set for final evaluation. Overall, the dataset represents a challenging benchmark for historical OCR tasks.

### 3. Data Preprocessing and Augmentation

The Byzantine seal dataset presents several challenges, including heterogeneous image sizes, limited data availability, and strong class imbalance. To address these issues, a dedicated preprocessing and augmentation pipeline was designed to ensure consistent input representation and improve model generalization.

The original dataset was provided in a flat directory structure, with class identifiers encoded in the image filenames. To enable the use of Keras data-loading utilities, the dataset was reorganized into class-specific subfolders. A minor filename inconsistency was manually corrected during this process.

Since character crops vary in size from approximately  $60 \times 60$  to over  $180 \times 180$  pixels, all images were resized to a fixed resolution of  $112 \times 112$  pixels. This choice balances computational efficiency with the preservation of relevant visual details. Pixel values were normalized to the  $[0, 1]$  range using a rescaling layer embedded in the model architectures.

The training data were split into 90% training and 10% validation subsets, following dataset recommendations and considering the limited number of samples. A separate test set was kept isolated and used exclusively for final evaluation to avoid data leakage.

To reduce overfitting and increase robustness, data augmentation was applied during training for convolutional models. The adopted transformations include random rotations, zooming and contrast adjustments. These operations simulate variations caused by seal orientation, material properties, surface wear, and illumination conditions, while preserving the semantic meaning of the characters.



#### 4. Model Architectures

To address the Byzantine seal character classification task, several neural network architectures were explored following a progressive and experimental approach. Starting from a simple baseline, increasingly more expressive models were introduced in order to evaluate the impact of spatial feature learning, regularization, and transfer learning on classification performance.

##### Baseline Model: Fully Connected Network (LeNet-300)

As an initial benchmark, a fully connected neural network inspired by the LeNet-300 architecture was implemented. The model consists of a flattening layer followed by two dense layers with 300 and 100 neurons respectively, and a final softmax output layer. Input images were resized to  $112 \times 112$  pixels and normalized within the model.

The training curves confirm this limitation. Over  $\sim 150$  epochs, accuracy increases only modestly and saturates at around **0.14–0.16** ( $\approx 14\text{--}16\%$ ) for both training and validation. This is higher than random guessing ( $\approx 4\%$  for 24 classes), indicating the model learns some signal, but overall performance remains low. Validation accuracy is also highly unstable, fluctuating strongly from epoch to epoch, which is consistent with a small validation set and a model that does not learn stable, generalizable features.

The loss curves show a rapid drop from a very high initial training loss in the first epoch, followed by a slow improvement. Training loss gradually decreases, while validation loss stays roughly flat around  $\sim 3.0\text{--}3.3$  and even trends slightly upward near the end. Rather than clear “classic” overfitting (very high train accuracy with collapsing validation accuracy), the dominant issue here is **limited representational power**: the fully connected baseline cannot reliably capture the spatial patterns of the characters, resulting in a low accuracy ceiling and noisy validation behavior. This baseline supports the conclusion that **spatial feature extraction (convolutions)** is essential for the Byzantine seal character classification task.

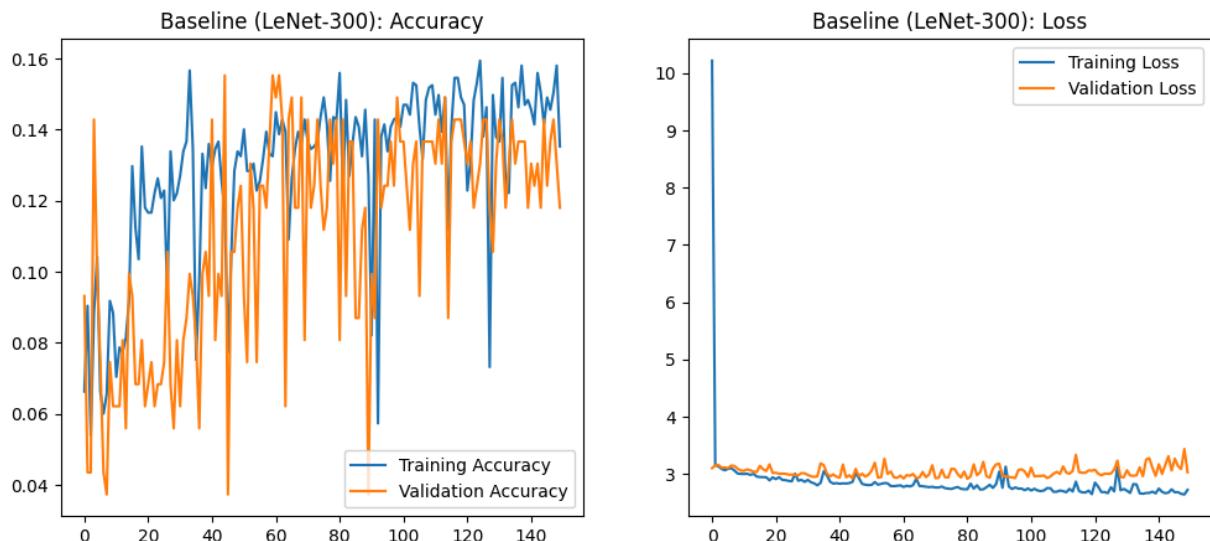


Figure 2: Training and validation performance of the baseline model (LeNet-300).

## **Convolutional Neural Network: LeNet-5**

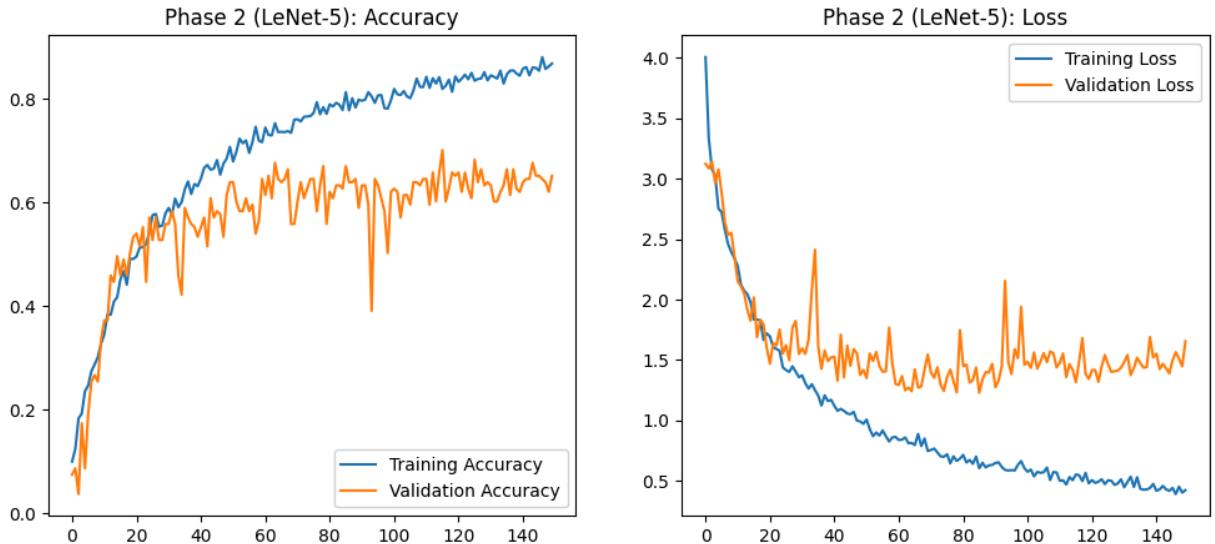
To incorporate spatial learning, a convolutional neural network inspired by the LeNet-5 architecture was introduced. The model consists of two convolutional blocks, each composed of a convolutional layer followed by average pooling and batch normalization. The convolutional feature extractor is followed by fully connected layers for classification.

To improve robustness and mitigate overfitting, several regularization strategies were applied. Data augmentation layers were used at the input level to introduce invariance to small rotations, scale changes, and contrast variations. In addition, a dropout layer was inserted before the final classification stage to reduce co-adaptation of neurons.

The training curves demonstrate a clear improvement over the fully connected baseline. Training accuracy increases steadily throughout training, eventually reaching approximately **85–88%**, indicating that the model is able to fit the training data very well. Validation accuracy rises rapidly during the early epochs and then stabilizes around **60–65%**, with moderate fluctuations caused by the relatively small validation set.

The loss curves further highlight this behavior. Training loss decreases smoothly and continuously, while validation loss drops sharply during the initial phase and then plateaus around **1.4–1.6**, exhibiting occasional spikes but no sustained divergence. The growing gap between training and validation performance indicates the onset of overfitting; however, the validation accuracy remains stable rather than collapsing, suggesting that the learned representations generalize reasonably well.

Overall, these results confirm that incorporating convolutional layers and spatial inductive bias is crucial for this task. Despite its relatively simple architecture, the LeNet-5 model effectively learns hierarchical features such as edges, strokes, and character shapes, achieving a strong balance between model capacity and generalization on this small and specialized dataset.



*Figure 3: Training and validation performance of the LeNet-5 model.*

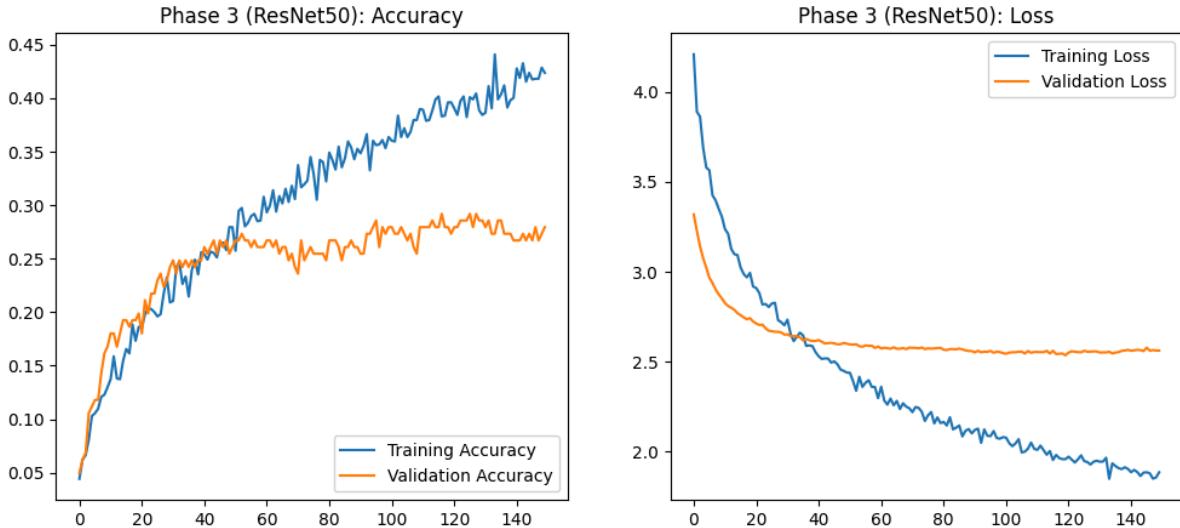
### Transfer Learning with ResNet-50

To explore the upper performance bound, a deep convolutional architecture based on ResNet-50 was evaluated using transfer learning. The model was initialized with weights pretrained on ImageNet, and the convolutional backbone was kept frozen to preserve the learned representations. A lightweight classification head composed of global average pooling, dropout, and a dense softmax layer was added on top.

The training curves reveal that the frozen ResNet-50 model is able to fit the training data to a limited extent, with training accuracy gradually increasing to approximately **40%**. In contrast, validation accuracy improves only during the early epochs and then quickly saturates around **27–29%**, showing minimal gains thereafter. This persistent gap between training and validation accuracy indicates limited generalization rather than effective transfer.

The loss curves further support this interpretation. Training loss decreases steadily throughout training, while validation loss drops rapidly at the beginning and then plateaus around **2.6**, exhibiting only marginal improvement despite continued optimization. This behavior suggests that the classification head adapts to the frozen features, but the features themselves are not well suited to the target task.

Overall, these results demonstrate a pronounced domain mismatch between ImageNet (natural, colored photographs) and the Byzantine seal dataset (monochromatic, embossed character crops). Although ResNet-50 provides rich and expressive representations for natural images, its frozen features fail to capture the fine-grained stroke patterns and shape variations required for character recognition in this domain. As a result, transfer learning with a fully frozen backbone yields substantially worse performance than the simpler, task-specific LeNet-5 architecture.



*Figure 4: Training and validation performance of the ResNet-50 model with frozen backbone.*

### Fine-Tuning the ResNet Model

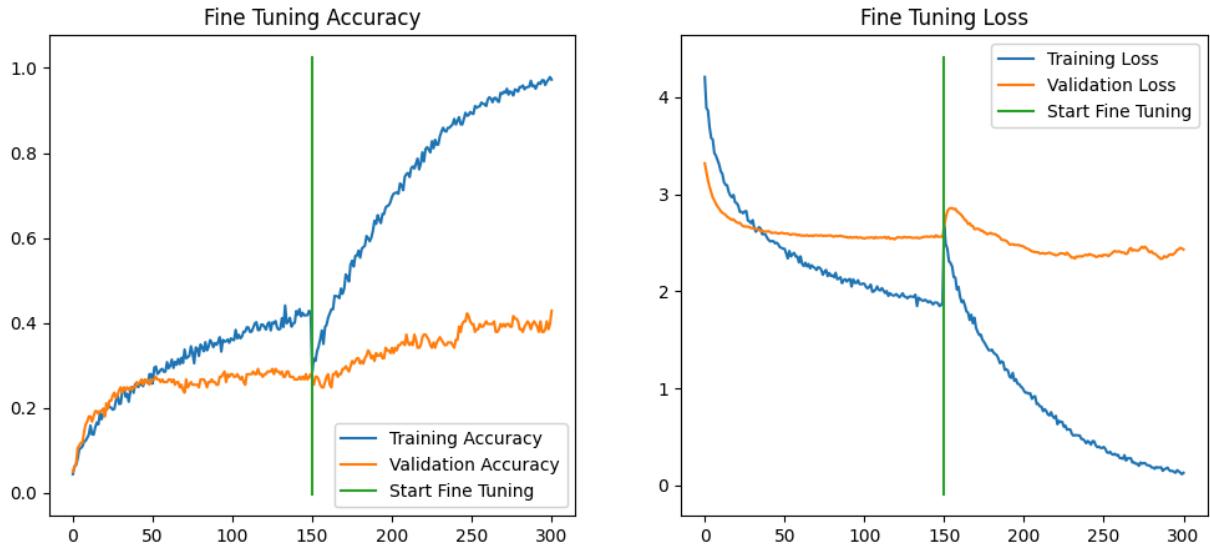
To reduce the domain gap observed in the frozen transfer learning setup, a fine-tuning strategy was applied by unfreezing the upper layers of the ResNet-50 backbone while keeping the lower layers fixed. Fine-tuning was performed using a substantially reduced learning rate to allow the model to adapt to the target domain without destroying low-level edge and texture detectors.

The training curves clearly illustrate the effect of fine-tuning. Before unfreezing the backbone (left of the green vertical line, around epoch 150), both training and validation accuracy increase slowly and remain relatively low. After fine-tuning is enabled, training accuracy rises sharply, eventually reaching almost **100%**, and training loss decreases rapidly. In contrast, validation accuracy improves only modestly, increasing from roughly **28–30%** to around **38–42%**, before stabilizing and slightly fluctuating.

The loss curves further emphasize this divergence. While training loss continues to decrease steadily after fine-tuning, validation loss shows only a small initial improvement and then plateaus around **2.5**, with a slight upward trend toward the end of training. This widening gap between training and validation performance indicates pronounced overfitting, as the model increasingly memorizes the training samples rather than learning features that generalize to unseen data.

Overall, fine-tuning provides a measurable but limited benefit over the frozen ResNet-50 model. Although unfreezing the upper layers allows the network to adapt some of its representations to the Byzantine seal domain, it does not fully bridge the domain mismatch between ImageNet (natural, color images) and monochromatic embossed characters. This experiment reinforces the conclusion that increased model complexity and transfer learning do not necessarily yield superior performance in data-scarce, highly specialized visual

domains, and that simpler, task-specific architectures such as LeNet-5 can generalize more effectively.



*Figure 5: Training and validation performance during ResNet-50 fine-tuning.*

### Model Comparison and Selection

Among all evaluated architectures, the LeNet-5-inspired convolutional network achieved the best balance between performance and generalization. Its moderate depth and task-specific design proved more effective than both the fully connected baseline and the large pretrained ResNet model.

These results suggest that, for historical OCR tasks with limited and specialized data, lightweight convolutional architectures combined with appropriate regularization outperform deep transfer learning approaches. Consequently, the LeNet-5 model was selected as the final architecture for evaluation and cross-validation.

## 5. Results and Evaluation

This section presents the quantitative and qualitative evaluation of the proposed models. Performance is assessed using standard classification metrics and training dynamics, with the goal of comparing different architectures and identifying their strengths and limitations for the Byzantine seal character recognition task.

### Evaluation Metrics

Model performance was evaluated using classification accuracy as the primary metric, complemented by precision, recall, macro-averaged F1-score, and confusion matrices. These metrics provide a comprehensive view of both overall performance and per-class behavior, which is particularly important given the strong class imbalance present in the dataset.

Final results are reported on a held-out test set, while training and validation curves are

used to analyze convergence behavior, stability, and overfitting during model development.

## Baseline Results

The fully connected baseline model inspired by the LeNet-300 architecture achieved limited performance, reaching approximately **14–16% validation accuracy**. While this result is significantly higher than random guessing ( $\approx 4\%$  for 24 classes), overall performance remained low and unstable. Training and validation accuracy curves closely track each other and exhibit strong fluctuations, indicating that the model struggles to learn robust and generalizable representations.

The loss curves further highlight this limitation. Training loss decreases steadily after an initial sharp drop, while validation loss shows only minor improvement and remains largely flat throughout training. Rather than exhibiting classical overfitting, the baseline model suffers primarily from **limited representational capacity**: flattening the input images removes crucial spatial information, preventing the network from effectively modeling character shapes and stroke patterns.

These results confirm that a purely dense architecture is insufficient for this task and that spatial feature extraction is essential for reliable character recognition.

## Performance of the LeNet-5 Model

The LeNet-5-inspired convolutional neural network achieved the best overall performance among all evaluated architectures. Validation accuracy increased rapidly during the early training phase and stabilized at approximately **60–65%**, while final test accuracy exceeded **60%**, representing a substantial improvement over the baseline model.

Training accuracy continued to increase throughout training, eventually reaching around **85–88%**, resulting in a moderate but controlled gap between training and validation performance. The loss curves show smooth convergence: training loss decreases steadily, while validation loss drops sharply early on and then stabilizes around a constant level, with occasional fluctuations due to the limited size of the validation set.

Confusion matrix analysis reveals that frequently represented and visually distinctive classes (e.g., Omega, Tau, Alpha, Rho) are recognized with high precision and recall. In contrast, misclassifications are more common among visually similar characters and underrepresented classes, reflecting both intrinsic ambiguities in the glyph shapes and the effects of class imbalance.

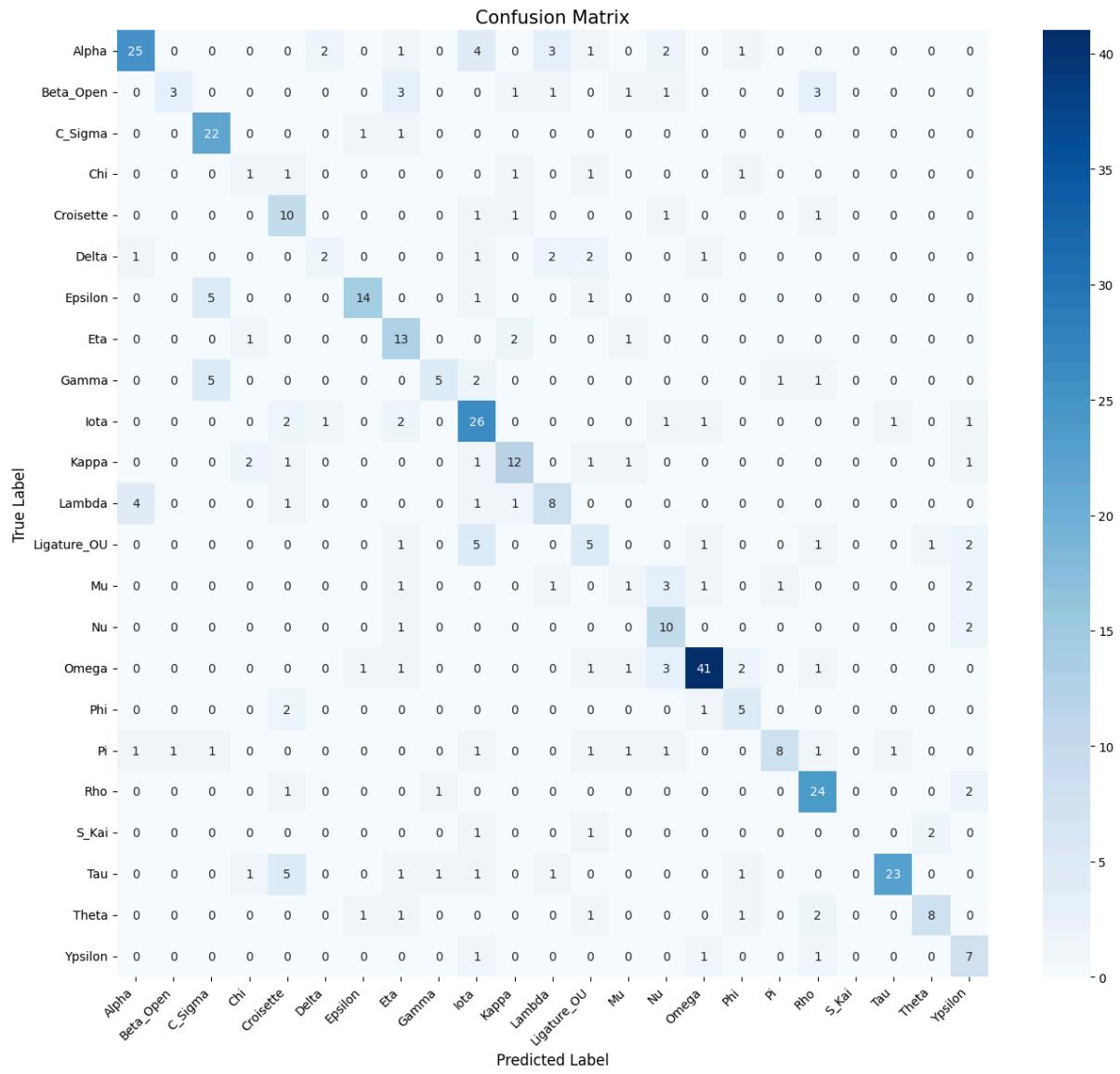


Figure 6: Confusion matrix of the LeNet-5 model on the test set.

Overall, these results demonstrate that lightweight convolutional architectures combined with data augmentation and regularization are well suited for historical OCR tasks with limited and specialized data.

### Transfer Learning and Fine-Tuning Result

The ResNet-50 model pretrained on ImageNet performed poorly when used as a frozen feature extractor, achieving validation accuracy of approximately **25–30%**. Despite stable optimization and steadily decreasing training loss, the model failed to learn sufficiently discriminative representations for Byzantine seal characters, indicating a strong domain mismatch between natural images and embossed historical artifacts.

Fine-tuning the upper layers of the ResNet-50 backbone led to a substantial increase in training accuracy; however, the improvement in validation performance was limited.

Validation accuracy increased modestly, from approximately **28–30%** to around **40%**, before plateauing. The growing divergence between training and validation curves clearly indicates overfitting. These results demonstrate that increased model complexity and transfer learning do not necessarily translate into superior generalization, especially in data-scarce

and domain-specific settings.

### Cross-Validation Results

To further assess model robustness, stratified 5-fold cross-validation was performed using the LeNet-5 architecture. The model achieved a mean accuracy of **0.575** with a standard deviation of **0.035**, indicating stable performance across different data splits.

The macro-averaged F1-score was lower than overall accuracy, reflecting the difficulty of correctly classifying minority classes with limited training samples. This result emphasizes the challenges posed by class imbalance and highlights potential directions for future improvement.

### Discussion

The experimental results clearly show that model design must be aligned with data characteristics. While deep pretrained networks excel in large-scale natural image recognition tasks, they are not necessarily effective for specialized historical OCR problems. In contrast, the LeNet-5 architecture strikes a favorable balance between capacity and generalization, making it the most suitable model for this dataset.

The achieved performance demonstrates that meaningful character recognition is possible despite severe data limitations and visual degradation, and that carefully designed lightweight models can outperform significantly more complex architectures in such scenarios.

## 6. Conclusions

In this project, an image classification approach was applied to the recognition of individual characters cropped from images of ancient Byzantine seals. Several neural network architectures were evaluated, ranging from a fully connected baseline to convolutional models and deep transfer learning approaches. Experimental results show that a lightweight LeNet-5-inspired convolutional network achieves the best performance, reaching approximately 63% accuracy on the test set. This result significantly exceeds random guessing and demonstrates that meaningful visual patterns can be learned despite severe data limitations and degradation of historical artifacts.

The evaluation highlights that model complexity alone does not guarantee improved performance. While transfer learning with ResNet-50 is highly effective in many computer vision tasks, it proved unsuitable in this context due to strong domain mismatch and limited training data. In contrast, a shallow architecture specifically tailored to the scale and characteristics of the dataset generalizes better and exhibits reduced overfitting.

When compared to the results reported by Rageau et al. in “Character recognition in Byzantine seals with deep neural networks”, where character classification accuracy above 90% is reported under controlled experimental conditions, the performance obtained in this work is substantially lower. This discrepancy is primarily explained by fundamental methodological and data-related differences, including dataset curation, class selection, and evaluation protocols. Rageau et al. employ a two-stage pipeline in which characters are

first localized using ground-truth or high-quality bounding boxes and then classified using a deep network trained on carefully curated and augmented character crops. Moreover, their evaluation focuses on well-represented classes and benefits from expert-level annotations and controlled cross-validation at the seal level.

In contrast, the present work addresses a more constrained setting, relying on a smaller subset of character images with pronounced class imbalance, higher visual noise, and less controlled variability in cropping quality. As a result, the classification task is inherently more challenging, and the lower accuracy reflects these constraints rather than a failure of the modeling approach.

Overall, the results confirm that Byzantine seal character recognition remains a difficult problem under realistic data scarcity conditions. Nevertheless, the achieved performance demonstrates that even simple convolutional models can provide valuable support for historical OCR tasks, offering a solid baseline for further research in the digital analysis of cultural heritage artifacts.