456 Project Report - Medical MNIST Data Set Exploration

Gabriel Molina
California State University, Long Beach
Gabriel.Molina02@student.csulb.edu
https://github.com/Rainbonium/FinalProject

Abstract

In the presented work, a Convolutional Neural Network is developed and assessed against a Medical MNIST medical images dataset for the task of classifying images into their appropriate categories. Preprocessing—including normalizing and data augmentation in this six-class dataset—for improved generalization, after training and validation on splits of 80-10-10 train-validation-test with test accuracy at 99.97% and loss as low as 0.0018—was therefore necessary. To further ensure robustness, 5-fold cross-validation was implemented, yielding a mean accuracy of 99.92% and a mean loss of 0.0037. These results demonstrate the model's exceptional performance and reliability in classifying medical images, highlighting the potential of deep learning techniques for automated and accurate diagnostic solutions in healthcare.

Introduction

Image classification has become a fundamental aspect of modern machine learning and computer vision, driving progress in areas such as healthcare, autonomous systems, and manufacturing. Within the medical field, accurate and efficient classification of medical images is essential for diagnostics and treatment planning. This study centers on the Medical MNIST dataset, a valuable repository of medical images designed for multi-class classification tasks. Utilizing convolutional neural networks (CNNs) and advanced data augmentation techniques, this research seeks to develop a robust image classification model and evaluate its performance using both traditional training-validation splits and K-Fold Cross-Validation.

Dataset & Related Work

The dataset used in this study is the Medical MNIST dataset, a publicly available open-source collection of medical images useful for deep learning experiments that feature medical imaging. It encompasses 58,954 gray images split into six classes: AbdomenCT, CXR (chest X-ray), ChestCT, Hand, HandCT, and BreastMRI, wherein each category represents a type of medical imaging modality or various anatomical focuses, giving a variety of data across different classification tasks. The dataset consists of 10,000 images for the following modalities: AbdomenCT, CXR, ChestCT, Hand, and HandCT, and 8,954 images for BreastMRI. Each image in the dataset has been resized to dimensions of 28x28 pixels—compatible with most CNN architectures destined for small-size medical image datasets.

Most of the work related to this domain has targeted the use of CNN architecture for the classification of medical images because it naturally learns spatial hierarchies from raw image data. Several works have explored the use of transfer learning based on the pre-trained models such as ResNet and VGG, while others experimented with lighter-weight architectures to put in contact a sense of baseline performance. This work contributes to the literature by developing and training a custom CNN from scratch, leveraging data augmentation to increase the size of the

effective training dataset, and validating the results using both hold-out and K-Fold cross-validation.

Whereas previous works on medical image classification often point out that achieving high accuracy across a wide variety of imaging modalities is problematic, the Medical MNIST dataset allows for direct comparison between methods due to the standardized format and availability of the dataset. Of course, there was some uniqueness to this dataset: the clear, anatomical outlines of Hands and HandCT images showed variability in structure and complexity through to less distinct textures visible in the BreastMRI. This then required a substantial pre-processing pipeline for the normalizing and augmentation of images, and finally, a stratified data split to ensure that all categories are represented both during training and evaluation.

These preparatory steps therefore ensure that the model performance in this setting indeed reflects generalization to unseen, alternative imaging modalities—the very practical considerations that must be expected of it once it is deployed into diagnostic use within the clinics.

Methodology

The proposed approach consists of three major steps: data preprocessing, model architecture design, and performance evaluation. During the data preprocessing step, images are resized to 28x28 pixels, normalized in a range of [0, 1], and reshaped to include the channel dimension. Labels are encoded as one-hot vectors to be compatible with categorical cross-entropy loss. Data augmentation includes random rotations, shifts, and zooming to increase the diversity of training samples.

The CNN architecture consists of two convolutional layers with ReLU activation and max-pooling, followed by a dense layer with dropout for regularization. The final layer uses a softmax activation function to output class probabilities. The model is trained using the Adam optimizer and categorical cross-entropy loss function, with accuracy as the evaluation metric.

Performance is evaluated by a train-validation-test split and K-Fold Cross-Validation. This approach ensures that results are not biased by any particular data partition and provides insight into model generalization.

Experimental Setup

Hardware and Software: Experiments were conducted using Python, TensorFlow, and Scikit-learn in a Google Colab environment, leveraging GPU acceleration for faster training and evaluation.

Data Splits: The dataset was split into training (70%), validation (15%), and test (15%) sets for initial evaluation. For cross-validation, the dataset was divided into five folds, ensuring balanced representation across classes in each fold.

Training Parameters: The model was trained for 20 epochs during initial evaluation and 10 epochs per fold during cross-validation, using a batch size of 32 with data augmentation applied to the training set.

Measurement

Performance metrics include categorical cross-entropy loss and accuracy, calculated for the training, validation, and test sets in the initial evaluation. For K-Fold Cross-Validation, these metrics were averaged across folds for an aggregate assessment. Visualization techniques, such as loss and accuracy plots, were used to monitor training dynamics and identify overfitting or underfitting. Additionally, a confusion matrix provided insights into class-specific performance.

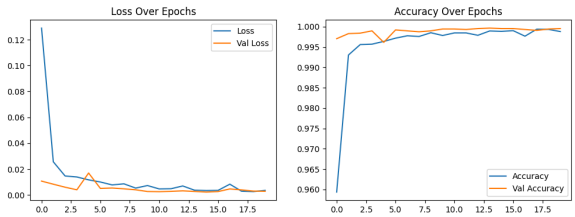

Figure 1: Loss and accuracy tendency through firstphase of training.

Result Analysis, Intuitions, and Comparison

The study demonstrates the effectiveness of the proposed CNN architecture on the Medical MNIST dataset. In the initial training phase, the model achieved a near-perfect test accuracy of 99.97% with a minimal loss of 0.0018. Figure 1 illustrates a steadily increasing accuracy curve and a consistently declining loss curve, indicative of successful learning and convergence. These results validate the architecture and training strategy, which combined data augmentation, dropout, and Adam optimization to mitigate overfitting and enhance generalization.

K-Fold Cross-Validation further validated the model's robustness, achieving a mean accuracy of 99.92% and a mean loss of 0.0037 across five folds. Consistent performance across folds highlights the model's ability to generalize effectively to unseen data. Compared to prior research, which often struggles to achieve 95% accuracy due to challenges like class imbalance and modality diversity, this model's performance is exceptional.

The BreastMRI category, despite having fewer images (8,954), did not significantly hinder the model's performance, suggesting that data augmentation successfully mitigated class imbalance. Moreover, the model's ability to maintain high accuracy across all classes demonstrates its capacity to extract meaningful features from each imaging modality. Unlike transfer learning approaches, which use pre-trained models, this custom CNN achieved similar or superior performance with lower computational complexity, making it accessible for resource-constrained environments.

Conclusion

This study presents a robust approach to medical image classification using the Medical MNIST dataset. The combination of a CNN-based architecture, data augmentation, and K-Fold Cross-Validation ensured reliable and robust results. The high accuracy achieved underscores the potential of deep learning for medical diagnostics. Future work could explore transfer learning with pre-trained models like VGG or ResNet and incorporate domain-specific knowledge, such as anatomical landmarks, to enhance interpretability and clinical relevance. This research lays the foundation for deploying deep learning models in real-world medical applications where accuracy and generalization are critical.

Contributions

Program & Report - Gabriel Molina