DermaMNIST CNN Multi-class classification

Team members: Samantha Chan, Jessica Ho, Seon Min Kim, Wenli Xie, Belinda Chen

Project Overview: Skin Lesion Classification with DermaMNIST

Objective: Develop a CNN model to classify dermatoscopic images into seven skin lesion categories using the DermaMNIST dataset.

Goal: Support dermatological diagnostics by enabling robust and reproducible image-based disease classification.

Dataset:

- Derived from the HAM10000 dataset via MedMNIST.
- 10,015 RGB images resized to 28×28 pixels.
- Pre-split into training (7,007), validation, and test (2,005) sets.

Data Handling:

- Images normalized from [0, 255] → [-1, 1] for stable training.
- Efficient loading via PyTorch DataLoader.

Project Overview: Skin Lesion Classification with DermaMNIST

Model:

- Custom CNN with convolutional + pooling layers for hierarchical feature extraction.
- Fully connected layers for multi-class classification across seven categories.

Training:

- Optimizer: SGD | Loss: Cross-Entropy | Epochs: 10 | Batch Size: 100 | LR: 0.01
- Metrics logged per epoch: Accuracy, AUC

Evaluation:

- Performance evaluated on the test set.
- Results analyzed using confusion matrices and per-class metrics.

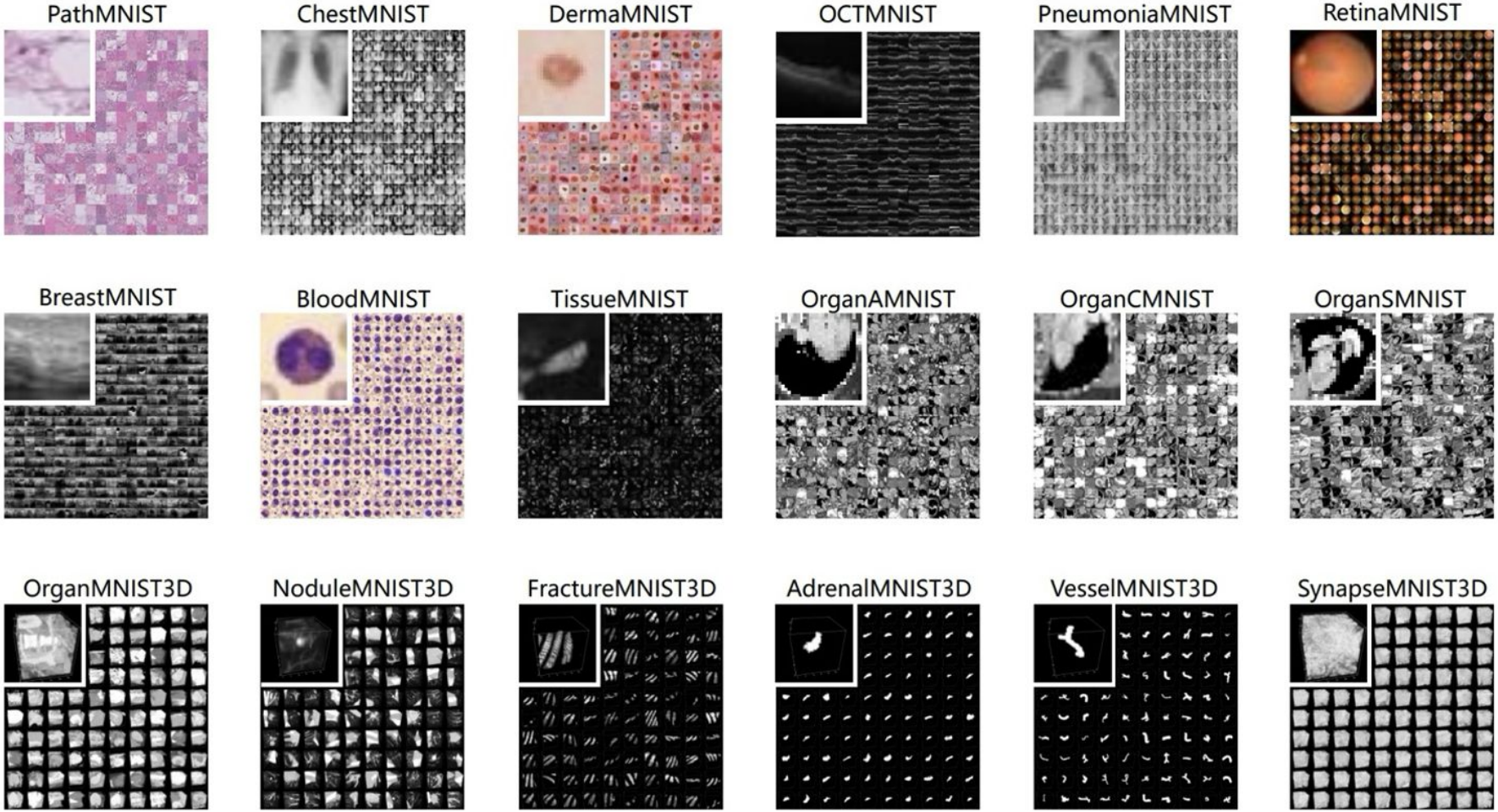
Expected Outcomes:

- Establish baseline classification performance on DermaMNIST.
- Identify key challenges (e.g., class imbalance, visual similarity).
- Propose next steps such as data augmentation or deeper models.

Description

MedMNIST is a comprehensive collection of biomedical image datasets designed to facilitate machine learning research and education. It is comprised of 12 datasets for 2D and 6 datasets for 3D, encompassing a variety of imaging modalities such as X-rays, CT scans, ultrasounds, and dermatoscopic images. Each dataset is preprocessed into uniform sizes, 28×28 for 2D images and 28×28×28 for 3D volumes, making them lightweight and accessible for rapid prototyping and algorithm benchmarking. MedMNIST supports diverse classification tasks, including binary, multi-class, ordinal regression, and multi-label classification, with dataset sizes ranging from 100 to 100,000 samples. Additionally, MedMNIST+ offers higher-resolution versions (64×64, 128×128, and 224×224 for 2D; 64×64×64 for 3D) to support the development of more complex models.

DermaMNIST, a subset of MedMNIST, focuses on dermatological image classification. It is derived from the HAM10000 dataset, which contains 10,015 dermatoscopic images of pigmented skin lesions. These images are categorized into seven classes, including melanoma, basal cell carcinoma, and benign keratosis, among others. For consistency with the MedMNIST framework, the original images 600×450 pixels are resized to 28×28 pixels. The dataset is split into training, validation, and test sets in a 7:1:2 ratio. With its compact size and standardized format, DermaMNIST is particularly suitable for algorithm prototyping, and lightweight model evaluation in dermatological imaging.


PathMNIST
ChestMNIST
DermaMNIST
OCTMNIST
PneumoniaMNIST
RetinaMNIST
BreastMNIST
BloodMNIST
TissueMNIST
OrganAMNIST
OrganCMNIST
OrganSMNIST
OrganMNIST3D
NoduleMNIST3D
FractureMNIST3D
AdrenalMNIST3D
VesselMNIST3D
SynapseMNIST3D

Dependencies

Python version: 3.8+

Dataset: medmnist

To run the code, please ensure the following packages are installed:

- Torch
- Torchvision
- Scikit-learn
- Numpy
- Pandas
- Matplotlib
- Medmnist (dermamnist)

Run the code in the Jupyter Notebook

Methods - Data Pre-processing

1. Split the dataset into training and testing sets
    - Training: 7,007
    - Testing: 2,005
2. Convert images to tensors
    - Converting pixel values from [0, 255] to [0, 1]
3. Normalize tensors by rescaling from [0,1] to [-1, 1] to center image around 0.
    - Make gradients more stable
    - Avoid bias due to undcentered inputs
    - Helps layers operate more efficiently

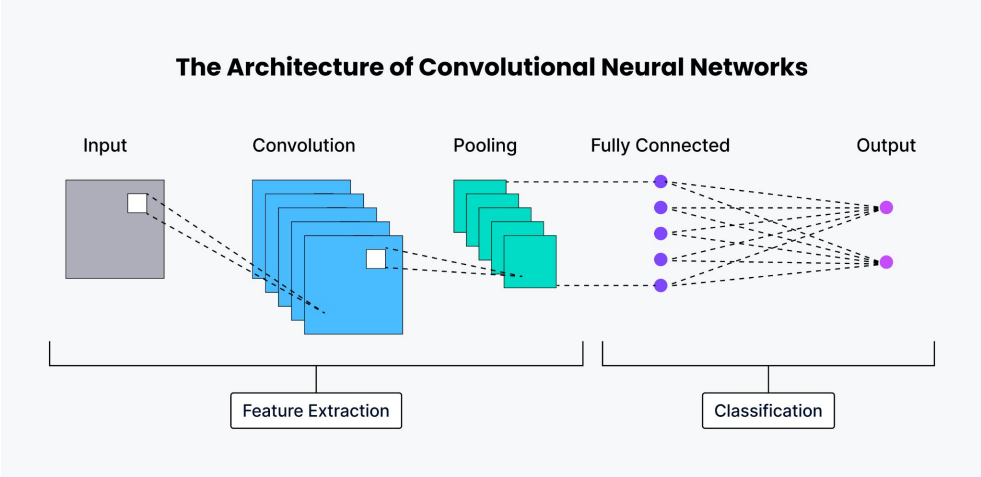
Methods - Convolutional Neural Network (CNN)

Why a Convolutional Neural Network?

- Hierarchical feature learning that enables extraction of complex image features
- Computationally efficient
- Automatic feature extraction; does not need manual feature engineering

What is a CNN?

- Type of feed-forward neural network that learns features via filter optimization
- Consists of input layer, hidden layers, and output layer



Methods - Our CNN Model

Input Layer: 3-channel RGB dermatoscopic image (28 pixels x 28 pixels)

Hidden Layer	Operation Steps in Layer	Output Channels
Conv1	3x3 Convolution + Batch Normalization + ReLu Activation Function	16
Conv2	3x3 Convolution + Batch Normalization + ReLu Activation Function + Max Pooling Layer	16
Conv3	3x3 Convolution + Batch Normalization ReLu Activation Function	64
Conv4	3x3 Convolution + Batch Normalization + ReLu Activation Function + Max Pooling Layer	64

Output Layer: Vector of probabilities across the 7 classes

Methods - Model Training

Model Training Hyperparameters

- **Number of Epochs:** 10 epochs
- **Optimizer:** SGD
- **Batch Size:** 100 samples per batch
- **Learning Rate:** 0.01
- **Loss Function:** Cross-Entropy Loss

Training loop steps:

1. Input image batch and labels
2. Predict probabilities for each class using CNN
3. Compute loss
4. Backpropagate to calculate gradients
5. Update model weights
6. Reset gradients

Methods - Model Evaluation

1. Set model to evaluation mode to disable dropout and batch norm updates
2. Obtain model outputs of test images from trained CNN
3. Compare predicted probability vs. binary true label for each class
4. Average AUCs across all valid classes for final AUC metric

Results: Class Distribution

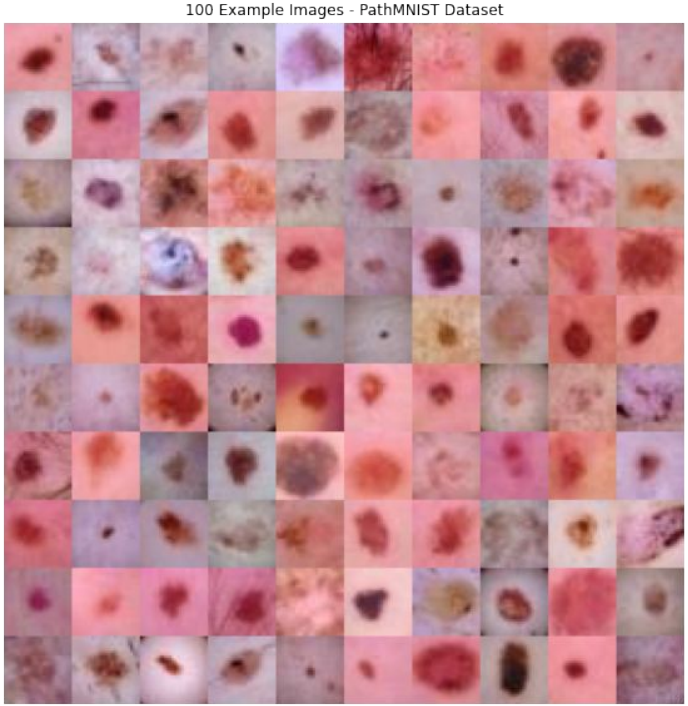
The dataset has large class imbalances

- Melanocytic nevi accounts for ⅔ of the data
- Remaining classes are underrepresented, particularly dermatofibroma & vascular lesions

Impact on learning:

- Model is biased toward majority class predictions
- During training, the loss function is dominated by the majority class, making it harder for the model to learn features from minority classes

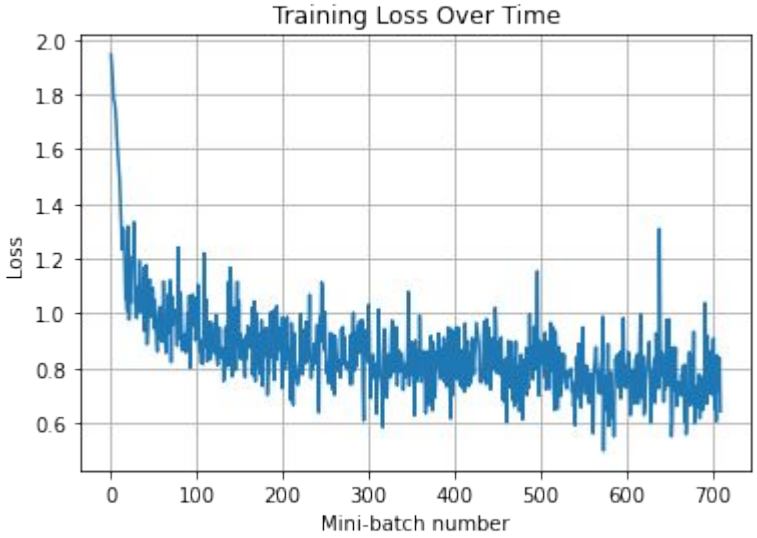
	Class Name	Prevalence (%)
0	actinic keratoses and intraepithelial carcinoma	3.25
1	basal cell carcinoma	5.12
2	benign keratosis-like lesions	10.97
3	dermatofibroma	1.14
4	melanoma	11.12
5	melanocytic nevi	66.98
6	vascular lesions	1.41



Subset of skin lesion imaging data used to train model on multi-class classification

Key Findings

- Training loss trend show clear improvement over time (downward slope) → an indication that the model is successfully learning
- Model is able to discriminate fairly well between classes (**AUC: 0.838**) w/ an overall accuracy 0f 0.7
- As expected from the data imbalance, the model is biased → Analysis reveals that the model performs well on Melanocytic Nevi classes, but fails on underrepresented classes
    - **Melanocytic Nevi (67% of data):** High precision (0.74), very high recall (0.99)
    - **Melanoma & Other Minority Classes:** Extremely low recall (Melanoma recall = 0.04), F1-scores of 0.00 for classes like actinic keratoses, dermatofibroma



	precision	recall	f1-score	support
actinic keratoses and intraepithelial carcinoma	0.00	0.00	0.00	66
basal cell carcinoma	0.26	0.35	0.30	103
benign keratosis-like lesions	0.60	0.16	0.25	220
dermatofibroma	0.00	0.00	0.00	23
melanoma	0.38	0.04	0.07	223
melanocytic nevi	0.74	0.99	0.84	1341
vascular lesions	0.00	0.00	0.00	29
accuracy			0.70	2005
macro avg	0.28	0.22	0.21	2005
weighted avg	0.62	0.70	0.62	2005

Key Findings (cont.)

This confusion matrix provides further insight into model performance:

- **Melanocytic nevi** were predicted with high accuracy (1,321 correct) - the model excels here (as previously established)
- **Melanoma** shows moderate accuracy, but is often misclassified as benign lesions, which is a major clinical risk
- **Benign keratosis-like lesions** are frequently confused with melanocytic nevi, suggesting overlap in features
- **Actinic keratoses** are rarely correctly predicted and mostly misclassified
- **Dermatofibroma** and **vascular lesions** are largely misclassified likely due to low sample representation

Overall, the model favors common classes and underperforms on rarer or visually similar categories.

Next Steps: Improving model fairness & performance

- Mitigate class imbalance
  - Apply class-weighted loss to penalize errors on underrepresented classes more heavily during training
  - Oversample minority classes or augment data to balance training set w/o reducing total data volume (→ this risks information loss)
- Evaluation strategy
  - Tune prediction thresholds to increase sensitivity on rare classes
- Enhance model training
  - Experiment with lower learning rates or potentially gradient clipping to stabilize learning on rare classes

Decisions and Tradeoffs: Model Architecture

Why CNNs?

- Spatial locality (good at detecting edges/textures), translation invariance, parameters efficiency

Why not…

Fully connected neural networks	Loses spatial information; requires much higher parameter count for image input
RNNs/LSTMs	Not suitable for image data, only for 2D sequences
Transformers	Overkill for 28x28 image; requires very large dataset and compute

Decisions and Tradeoffs: Model Architecture

Decision	Benefit	Trade-off
4 convolutional layers with ReLU and BatchNorm	Helps gradient flow and training stability; quick run time	Manual tuning of architecture, less generalizable
MaxPooling after some layers	Reduces spatial size and computation	Risk of losing spatial information
Fully connected layer: 64×4×4 → 128 → 128 → num_classes	Appropriate capacity for this dataset	Hardcoded spatial assumptions; no global pooling

Decisions and Tradeoffs: Training Pipeline

Decision	Benefit	Trade-off
Optimizer: SDG with learning rate 0.01	Simple and well-understood	Slower convergence compared to other optimizers (Adam, etc.)
10 epochs for training	Reasonable “default” for this size data	May be too few epochs for convergence
Batch size of 100	Efficient memory usage, faster runtime	Might not completely reduce gradient noise beneficial for generalization

Reference

1. Chao Chen, Nor Ashidi Mat Isa, Xin Liu, A review of convolutional neural network based methods for medical image classification, Computers in Biology and Medicine, Volume 185, 2025, 109507, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2024.109507.

2. Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, & Bingbing Ni. (2024). [MedMNIST+] 18x Standardized Datasets for 2D and 3D Biomedical Image Classification with Multiple Size Options: 28 (MNIST-Like), 64, 128, and 224 (3.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10519652

3. Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, Bingbing Ni. Yang, Jiancheng, et al. "MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification." Scientific Data, 2023.

4. Sarvamangala DR, Kulkarni RV. Convolutional neural networks in medical image understanding: a survey. Evol Intell. 2022;15(1):1-22. doi: 10.1007/s12065-020-00540-3. Epub 2021 Jan 3. PMID: 33425040; PMCID: PMC7778711.

5. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., & Ni, B. (2023). MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data, 10(1), 41. https://doi.org/10.1038/s41597-022-01721-8