CS-233 Project (2025) - Milestone 2 Report

Introduction

This report details the implementation and evaluation of deep learning methods for skin lesion classification using the DermaMNIST dataset[1]. Written by Alessandro Annibale Cioffi (358488), David Gorgiev (362628), and Sacha Heizmann (363265), this work applies Multi-Layer Perceptrons (hereby referred to as MLPs) and Convolutional Neural Networks (likewise referred to as CNNs) to classify dermoscopic images across seven diagnostic categories.

The DermaMNIST dataset, omitting the validation samples, contains 9,012 images (28×28 RGB) with severe class imbalance ranging from 103 samples (Dermatofibroma) to 6,034 samples (Melanocytic nevus). Our objective was to develop architectures capable of handling this multi-class classification while addressing class imbalance through cleverly chosen weighting strategies.

Methodology

For MLP as well as for CNN, the images were flattened to 2,352-dimensional vectors with normalization to [0,1]. The spatial structure was preserved, but transposed to PyTorch's (N,C,H,W) format. Both used stratified 80/20 train-validation splits.

Class weighting proved to be problematic. Standard balanced weighting failed catastrophically due to extreme class ratios, causing training instability (loss stuck at ~1.94). Square root weighting reduced ratios to 8:1, improving F1-score from 0.16 to 0.31. Final cubic root weighting provided optimal stability while maintaining class balance correction.

MLP and CNN layers and neuron counts were determined through experimentation, as well as optimal hyperparameters. Alternative activation functions were explored and successfully employed, and dropout helped achieve optimal results. (See results section for details).

For MLP, the architecture initially consisted of a singular layer of 64, then 128 neurons, quickly yielding sub-par accuracy and F1-scores (~50% to 60% and ~0.16 to 0.31 respectively). A second layer was added, slightly improving performance but increasing execution times. After numerous parameter tests, the MLP's architecture settled on a two hidden layer network (2,352 → 384 → 384 → 7), leveraging GELU activations instead of the standard ReLU, and using 0.2 dropout. Learning rates were suboptimal at the e-3 and e-5 range, yielding poor accuracy; the optimal choice for this parameter turned out to be 1e-4.

Fine-tuning revealed 320 epochs as optimal (0.4912 F1-score) vs. 325 epochs (0.4731 F1-score) and 330 epochs (0.4903 F1-score). The 384-neuron architecture provided sufficient capacity without excessive overfitting.

CNN's architecture consists of four convolutional layers with batch normalization and ReLU activation, followed by max-pooling and dropout to reduce overfitting. The feature maps are flattened and passed through two fully connected layers (128 → 7). The input size is fixed at 28x28x3, and the number of channels increases progressively (16 → 32 → 64 → 128).

Results

MLP Performance

Configuration	Train Acc	Train F1	Test Acc	Test F1	Epochs
320 epochs (0.2 dropout)	80.26%	0.6733	70.82%	0.4912	320
325 epochs (0.2 dropout)	81.49%	0.6565	72.67%	0.4731	325
330 epochs (0.2 dropout)	81.00%	0.6848	71.07%	0.4903	330

The results shown are for predictions on the test set after hyperparameter tuning on the validation set. Multiple interesting configurations were retained to highlight run-to-run variance, as no seed had been used for these measures (at that point - see CNNs for seeded results). Batch size was 64 for all runs.

Other results, not shown, hovered around the 50% to 60% accuracy range, with F1-scores ranging from 0.16 to 0.31 (as mentioned in the Methodology section). A combination of insufficient epochs (50 to 200), suboptimal learning rate and batch size selection (namely 128 and 32) were the cause, combined with early class weighting troubles.

[1] https://medmnist.com/, last accessed on 30.05.2025.

CNN Performance

Config	Train Acc	Train F1	Val Acc	Val F1	Epochs
4 CL	99.47%	0.9946	73.94%	0.5251	150
2 CL	75.33%	0.7941	63.49%	0.5040	100

Config	Train Acc	Train F1	Test Acc	Test F1	Epochs
4 CL	97.76%	0.9772	74.11%	0.5799	150

For practicality, we set a seed at the top of the main file to evict randomness and result fluctuation between runs. We initially experimented with two convolutional layers trained for 100 epochs. The training scores and validation set F1-score are good but the resulting validation accuracy was unsatisfactory. Adding two additional convolutional layers led to a significant improvement in validation accuracy and a modest increase in F1-score. However, the growing gap between training and validation performance revealed signs of overfitting. To address this, we introduced dropout regularization in the four-layer model, which helped slightly but did not lead to a substantial improvement in generalization. Importantly, the test set results confirm the generalization capacity of the final CNN model. Despite concerns of overfitting during validation, the final test results remain strong. This suggests that the use of dropout and batch normalization, although not eliminating overfitting, helped mitigate it sufficiently.
Regarding training speed, we observed that both the MLP and CNN were significantly slower than all the algorithms implemented in Milestone 1. Obtaining results over approximately 100 epochs (typically 200 to 300) took around 20 to 30 minutes, as we did not have access to a dedicated GPU to perform the computations.
As expected, increasing the number of layers and training iterations further slowed down the process. Regardless, the CNN required less iterations to achieve better results than the MLP.

Discussion

MLP

In spite of longer execution times, the MLP achieved 70.82% test accuracy and 0.4912 F1-score, demonstrating that fully connected networks can classify dermoscopic images despite spatial information loss.
The 9.44% train-test accuracy gap indicates controlled overfitting with good generalization.
Our class weighting evolution represents a critical methodological contribution for extreme imbalance scenarios. During optimization, we observed that configurations with higher test accuracy (e.g. 72.67% at 325 epochs) often yielded lower F1-scores (0.4731), highlighting the accuracy-F1 trade-off in imbalanced datasets. We prioritized F1-score optimization given the severe class imbalance, leading to our final 320-epoch configuration. The choice of GELU activation over ReLU also proved critical to boost the accuracy beyond 70%.

CNN

Compared to MLP and earlier CNN attempts, the last four-layer CNN model showed the strongest performance overall, though signs of overfitting remained. Despite the use of dropout and batch normalization, the performance gap indicates that the CNN may still rely too heavily on dominant class patterns. Nevertheless, it significantly outperformed the MLP in both raw accuracy and F1-score, suggesting that spatial features are valuable for this classification task. We can suppose that CNN would perform even better with higher resolution images, since the 28x28 images of DermaMNIST are pretty low resolution.
Interestingly, the CNN achieved its optimal performance with approximately half the number of epochs required by the MLP, indicating that it may also be more efficient at learning from this dataset despite its greater architectural complexity.

Usually we say that for neural networks, the more layers the better. Our results show that simply adding more layers can indeed lead to better results, but need adjustments to avoid overfitting.

Conclusion

Our architectures were successfully used to classify the DermaMNIST dataset with an accuracy of 72% to 74%, and F1-scores in the 0.49 to 0.57 range. While the MLP proved to be useful for the task, it was clearly outperformed by the CNN, namely in terms of F1-score and epoch count. This is unsurprising, given how well CNNs work with grid-like data, and images in particular. We therefore recommend the usage of a CNN for this task.