Project Report – Milestone 2 (DermaMNIST)

CS-233: Introduction to Machine learning Project

Mohamed Yebari-Omar Bensaid-Idriss Benjelloun

INTRODUCTION:

In this project, we tackle a multiclass classification task using a subset of the DermaMNIST dataset, composed of RGB images of skin lesions from 7 diagnostic categories. Our goal is to train models that can accurately classify each image into one of these categories. We implemented and compared two deep learning architectures: a Multi-Layer Perceptron (MLP) and a Convolutional Neural Network (CNN).

METHODOLOGY:

Data Preparation

We split the training data into a training and validation set (80/20), preserving class distribution via stratification. Images were normalized using per-channel mean and standard deviation. For MLPs, the images were flattened into 1D vectors. For CNNs, we preserved spatial structure and used the (N, C, H, W) format.

Model Architectures

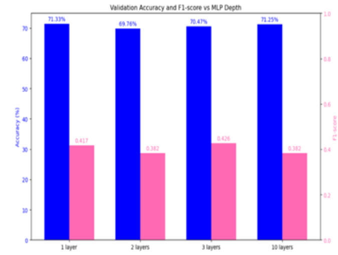
MLP: A three-layer feedforward network with ReLU activations and optional dropout. Input is a flattened 28×28×3 vector.
CNN: A two-layer convolutional architecture with ReLU and max-pooling, followed by fully connected layers.
All models were trained using the Adam optimizer with categorical cross-entropy loss.

EXPERIMENTS/RESULTS :

MLP:

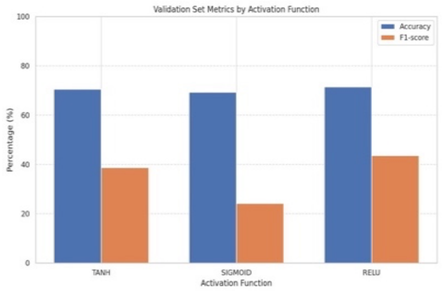
As expected, increasing the depth generally improves training performance: the train accuracy and F1-score increase from 92.7% to over 96% as we go from 1 to 3 layers.
This reflects a reflects a higher capacity of the model to fit the training data.

However, validation metrics tell a different story. The best validation F1-score (0.4256) is obtained with 3 hidden layers. Beyond this point (i.e., 10 layers), we observe a drop in generalization performance. While training accuracy remains high, the validation accuracy stagnates and the F1-score drops to 0.3818. This indicates overfitting: deeper networks memorize the training data but fail to improve predictions on unseen data.

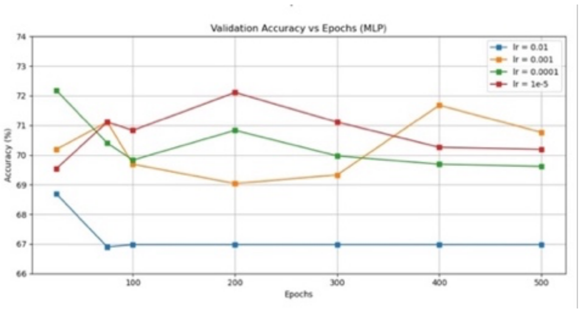


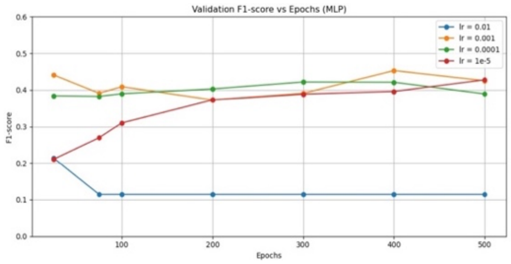
This analysis suggests that for our dataset, an MLP with 3 hidden layers achieves the best trade-off between expressivity and generalization.

Increasing depth further yields diminishing returns and may hurt performance due to overfitting or optimization challenges.



We compared the impact of three activation functions (ReLU, Tanh, and Sigmoid) on the performance of the MLP on the validation set.The results clearly show that ReLU offers the best performance, with an accuracy of 71.4% and an F1-score of 43.6%.
Tanh comes close behind, while Sigmoid yields the lowest scores, particularly in terms of the F1-score. These results confirm that ReLU is the most suitable function for our task, notably due to its speed and its ability to mitigate the vanishing gradient problem.



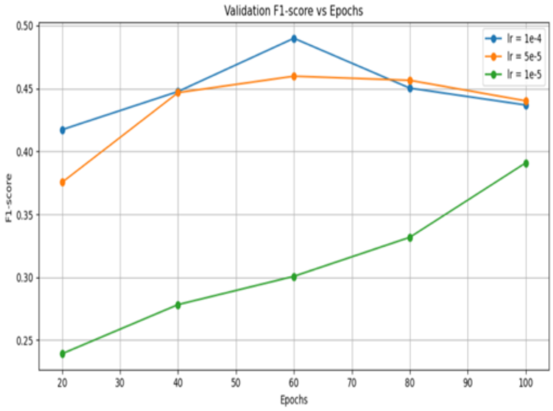


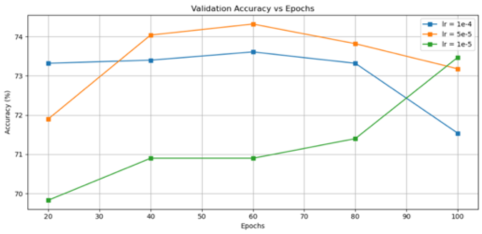
To optimize our MLP, we experimented with multiple combinations of learning rates (`0.01`, `0.001`, `0.0001`, `1e-5`) and training epochs (ranging from 25 to 500), with a fixed batch size of 32. Since the DermaMNIST dataset is imbalanced, we emphasized the **F1-score** as our primary evaluation metric. The results clearly show that `lr = 0.01` fails to converge, with F1-scores stagnating at 0.114. `lr = 1e-5` improves gradually but too slowly, requiring 500 epochs to reach 0.4277. The learning rate `0.0001` shows a smoother curve and performs well (max F1 = 0.4218 at 300 epochs), but the best results come from **`lr = 0.001`**, which peaks at **0.4530 F1-score at 400 epochs**. We therefore select **`lr = 0.001` with 400 epochs** as the optimal configuration, balancing convergence speed, stability, and high performance while mitigating overfitting.

CNN:

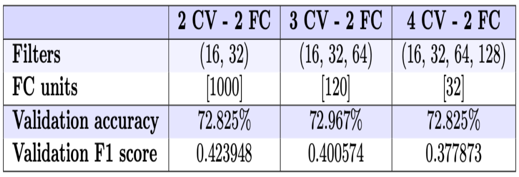
To optimize our CNN, we varied the learning rate (`1e-4`, `5e-5`, `1e-5`) and the number of epochs (20–100), with a fixed batch size of 32. Given the class imbalance in DermaMNIST, we focused on the F1-score.

Results show that while performance improves with more epochs, overfitting begins beyond 60. A rate of `1e-5` is too low, and `5e-5` remains stable but underperforms. The best result (F1-score = 0.4899) is achieved with `lr=1e-4` at 60 epochs, making it our optimal configuration.

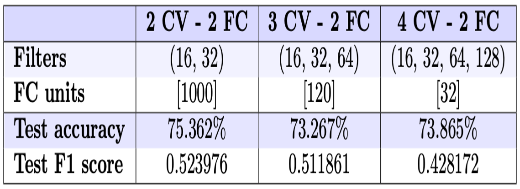




The performance of the three CNN architectures (2 CV-2 FC, 3 CV-2 FC, and 4 CV-2 FC) shows that increasing the network's complexity does not guarantee better performance. The 2 CV - 2 FC model, with only two convolutional layers (16 and 32 filters) and a fully connected layer of 1000 units, achieves the highest accuracy on the test set (75.36%) as well as the best F1-score (0.524). In comparison, the deeper architectures (3 CV and 4 CV) yield slightly lower performance, likely due to overfitting or a loss of information caused by excessive reduction in spatial dimensions. This observation is confirmed by the results on the validation set, where the scores remain very close between the three models, but the simplest model remains competitive or even better. This suggests that a simpler network is sufficient for this dataset, and that adding more layers may hinder generalization.



	2 CV - 2 FC	3 CV - 2 FC	4 CV - 2 FC
Filters	(16, 32)	(16, 32, 64)	(16, 32, 64, 128)
FC units	[1000]	[120]	[32]
Validation accuracy	72.825%	72.967%	72.825%
Validation F1 score	0.423948	0.400574	0.377873



	2 CV - 2 FC	3 CV - 2 FC	4 CV - 2 FC
Filters	(16, 32)	(16, 32, 64)	(16, 32, 64, 128)
FC units	[1000]	[120]	[32]
Test accuracy	75.362%	73.267%	73.865%
Test F1 score	0.523976	0.511861	0.428172

DISCUSSION/CONCLUSION:

Both MLP and CNN achieved competitive results on **DermaMNIST**. For the MLP, adding depth improved performance up to a point (3 layers), beyond which overfitting degraded generalization. CNNs showed that **simpler models generalize better** on this dataset.
Optimal configurations were:
MLP: 3 layers, lr = 0.001, epochs = 400
CNN: 2 CV, lr = 1e-4, epochs = 60
This highlights the importance of balancing model complexity and training dynamics when working with imbalanced medical image data.