Rock, Paper and Scissors Image Classification

Alessandro Bottoni

February 12, 2026

1 Introduction and Dataset

The following project aims at developing a Convolutional Neural Network capable of recognizing rock, paper and scissors hand gestures.
It consists of the development of three different models, with increasing complexity and in the comparison of their performance.
The dataset is composed of 2189 pictures in a .png format and divided in three classes:
1. Rock, with 726 pictures.
2. Paper, with 712 pictures.
3. Scissors, with 750 pictures.

2 Preprocessing and Data Augmentation

To prepare the dataset for training, a series of preprocessing steps and data augmentation techniques were applied. First, all images were resized to a uniform size of 256×256 pixels to ensure consistent input dimensions for the CNN. Next, pixel values were normalized to the range by dividing by 255, which supports faster convergence during training. To increase training-data diversity and improve model generalization, several augmentation techniques were used. These included random horizontal flips to reflect natural variability in hand gestures and random rotations within ±15° to account for different hand orientations. In addition, random color jittering was applied to vary brightness, contrast, and saturation, improving robustness to changing lighting conditions.

Altough it is a best practice to use cross validation to validate models, in order to avoid overfitting on the validation set, for the first model a simple train/validation/test split was used. This choice was made to speed up the training process, which is already quite long with the current hardware. The dataset was split into 70% for training, 15% for validation and 15% for testing. The split was performed separately for each class to ensure that the distribution of classes is consistent across all sets. As for the other two models a cross validation approach has been used, to better evaluate the performance of the models and avoid overfitting on the validation set.

3 Simple CNN Model and Training

3.1 Model Architecture: Simple CNN

The first model designed is a very simple Convolutional Neural Network (CNN) architecture. It consists of a single convolutional layer followed by a ReLU activation function, a pooling layer, a flatten layer and a fully connected one. The convolutional layer takes as input 3 channels (RGB color images), the only fixed parameter, while the number of output channels and the kernel size are hyperparameters that can be tuned in the following way:

- Number of output channels: 16, 32, 64
- Kernel size: 3, 4, 7

The padding is automatically calculated as the kernel size divided by 2, to preserve the spatial dimensions of the input. The purpose of this layer is to extract spatial features from the input images, such as edges and textures. The ReLU activation function introduces non-linearity to the model, allowing it to learn more complex patterns, while the pooling layer reduces the spatial dimensions of the feature maps, which helps to reduce the number of parameters and computational cost. The kernel size of the pooling layer is tunable, and can be set to either 2 or 4, where 2 preserves more details, and 4 is faster but loses information. Finally, the flatten layer converts the 2D feature maps into a 1D vector, which is then passed to the fully connected layer for classification.

3.2 Model Training Methodology

3.2.1 Hyperparameter Search Space and Strategy

The project employs Ray Tune, a distributed hyperparameter optimization framework, to systematically explore the hyperparameter space. Ray Tune provides efficient resource management, parallel trial execution, and sophisticated scheduling algorithms for early stopping.

3.2.2 Architecture Hyperparameters (Grid Search)

Architecture-related hyperparameters are explored using exhaustive grid search to ensure comprehensive evaluation of model capacity:

- **Output Channels** $\in \{8, 16, 32\}$: Controls the number of filters in the convolutional layer, directly affecting model capacity. Lower values reduce parameters and computational cost, while higher values increase representational power.
- **Convolutional Kernel Size** $\in \{3, 4, 7\}$: Determines the receptive field of each filter. Smaller kernels capture fine-grained local features, while larger kernels capture broader spatial context. The sizes were chosen to represent small (3×3), medium (4×4), and large (7×7) receptive fields.

Grid search generates $3 \times 3 = 9$ unique architecture combinations.

3.2.3 Training Hyperparameters (Random Search)

Training-related hyperparameters are sampled randomly to balance exploration efficiency with computational cost:

- **Pooling Kernel Size** $\in \{2, 4\}$: Sampled using `tune.choice`. Controls the downsampling factor. A kernel size of 2 is standard practice, preserving more spatial detail, while a kernel size of 4 performs aggressive downsampling, trading spatial resolution for faster computation.
- **Batch Size** $\in \{16, 32, 64\}$: Sampled uniformly with equal probability. Smaller batch sizes provide noisier gradient estimates but often generalize better, while larger batch sizes offer more stable training and faster computation per epoch.

3.2.4 Training Hyperparameters (Log-Uniform)

- **Learning Rate**: Sampled from a log-uniform distribution. Log-uniform sampling ensures equal representation across orders of magnitude, as learning rates often require logarithmic exploration. Mathematically:

With `num_samples=10`, each of the 9 architecture combinations is evaluated with 10 different random samples of training hyperparameters, yielding a maximum of $9 \times 10 = 90$ potential trials.

3.2.5 Ray Tune Configuration and Training Loop

Before describing the tuning workflow, it is useful to summarize the training setup at a high level. The model is trained using stochastic gradient descent with momentum, optimized against a cross-entropy loss for a three-class classification task. Each trial runs for a limited number of epochs (up to 10), and the batch size is treated as a tunable parameter, so the amount of data processed per update can change across trials. Data are fed through separate training and validation loaders: the training loader shuffles samples to improve generalization, while the validation loader keeps a fixed order to make evaluation stable; a small number of worker processes is used to overlap data loading with computation.

In this project, hyperparameter tuning is organized as a collection of independent *trials*. Each trial corresponds to one concrete set of hyperparameters and runs a short training job, producing comparable validation metrics that Ray Tune can track across experiments. The overall goal is to explore many configurations efficiently while keeping the execution stable and reproducible.

To keep the machine responsive, the tuning process is run under a controlled resource budget. Instead of letting every trial consume as many resources as it wants, each trial is assigned a fixed amount of compute (a certain number of CPU cores). This explicit resource assignment determines how many trials can run at the same time: if each trial needs more CPU, fewer trials can run in parallel; if each trial needs less, more trials can run concurrently. Ray Tune supports this kind of per-trial resource allocation so that parallelism is predictable and does not overwhelm the system. [7]

To avoid wasting time on bad configurations, the search uses the ASHA scheduler (Asynchronous Successive Halving). ASHA implements an early-stopping policy that continuously compares trials as they progress: at predefined milestones in training, a trial's performance is evaluated against other trials that have reached the same point. Trials that are clearly underperforming are stopped early, and the freed resources are immediately reused to start new trials or continue more promising ones. Because decisions are made asynchronously, trials do not need to wait for each other, which improves resource utilization and speeds up the overall search. [7]

Checkpointing is used to preserve useful model states during tuning. Rather than saving many snapshots, the configuration can be set up to retain only the most relevant checkpoint per trial (typically the best one according to the chosen validation metric). This makes it easier to recover the best-performing model after tuning and reduces disk usage, which is important when running many trials. [7]

Inside each trial, the training loop follows a standard pattern. Each epoch is split into a training phase (where model parameters are updated) and a validation phase (where performance is measured without updating the model). After each epoch, the trial reports key metrics (such as validation loss, validation accuracy, and training loss) back to Ray Tune so that the scheduler can make early-stopping decisions and so that results are logged in a consistent format. [7]

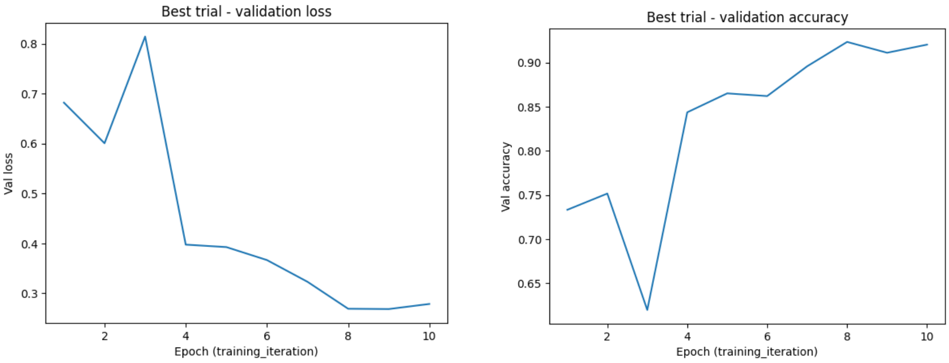
Finally, the entire experiment is orchestrated by the Tune *Tuner*. The Tuner combines the search space definition with the optimization objective (which metric to minimize or maximize) and the scheduler policy. It generates trial configurations, queues them, runs them when resources are available, and collects the results into a single structured output that you can inspect to select the best configuration and its corresponding checkpoint. [7]

3.3 Simple CNN (Ray Tune)'s Results

Best hyperparameters: Ray Tune selected the following configuration that turned out to the best by minimum validation loss across all trials::

- $out\_channels = 32$
- $conv\_kernel\_size = 7$
- $pool\_kernel\_size = 4$
- $lr = 7.6444576 \times 10^{-4}$
- $batch\_size = 16$

Performance: For the selected trial, the final validation loss was 0.2784 and the final validation accuracy was 0.9202. While on the test set, over $n = 331$ images, the model achieved test loss 0.2737 and test accuracy 0.9154. It is important to point out that due to the lack of cross-validation, these results are highly dependent on the specific train/validation/test split used, and may not be fully representative of the model's generalization performance.



(a) Model 1: validation loss across epochs (best Ray Tune trial).

(b) Model 1: validation accuracy across epochs (best Ray Tune trial).

Figure 1: Learning curves for the best Model 1 configuration.

Confusion matrix: Figure 1 reports the confusion matrix on the test set (rows = true class, columns = predicted class).

Table 1: Model 1 confusion matrix counts on the test set (rows: true class; columns: predicted class).

	Pred: Paper	Pred: Rock	Pred: Scissors
True: Paper	98	1	9
True: Rock	2	95	13
True: Scissors	1	2	110

Conclusion (Model 1). Despite its minimal architecture, Model 1 provides a strong baseline for the task: the best Ray Tune configuration achieves 0.920 validation accuracy and 0.915 test accuracy (test loss 0.274 on $n = 331$ images). The learning curves indicate rapid convergence within the 10 training epochs used during tuning, with validation loss decreasing and accuracy stabilizing near its final plateau, suggesting stable optimization. The confusion matrix shows that most predictions fall on the diagonal (98, 95, 110 correct per class), while the remaining errors are concentrated between visually similar gestures (notably some Rock and Paper samples predicted as Scissors). Overall, this computationally efficient model seems to already generalize well, making it a solid reference point for the upcoming comparison with deeper architectures and cross-validation.

4 Medium CNN Model and Training

4.1 Training procedure

In the second experiment we moved from the baseline (Model 1) to a moderately deeper CNN architecture, while keeping the same overall input pipeline and classification goal (3 classes: rock, paper, scissors).
Differently from Model 1, here the model assessment and model selection were carried out using **K-fold cross-validation**, in order to obtain a more reliable estimate of the generalization performance and to reduce dependence on a single train/validation split.The preprocessing and data augmentation pipeline is the same as in the model 1, with the exception of the train/validation split, which is now performed within each fold of the cross-validation procedure.

4.1.1 Model Architecture

Model 2 increases capacity by stacking two convolutional blocks and introducing explicit regularization (Batch Normalization and Dropout).The network is composed of a `features` module (convolutions, normalization, ReLU and pooling) and a `classifier` module (fully-connected layers).

Feature extractor:

- **Block 1:** `Conv2d(3` → `16, kernel=7, stride=1, padding=3)` → `BatchNorm2d(16)` → `ReLU` → `MaxPool2d(2)`.
- **Block 2:** `Conv2d(16` → `32, kernel=3, padding=1)` → `BatchNorm2d(32)` → `ReLU` → `MaxPool2d(2)`.

Classifier: After the convolutional blocks, the feature maps are flattened and passed to:

- `LazyLinear(128)` → `ReLU` → `Dropout(0.5)` → `Linear(128, 3)`.

The final layer outputs logits for the three-class classification task.

4.1.2 Training setup

The model is trained with cross-entropy loss (`CrossEntropyLoss`) and optimized using stochastic gradient descent with momentum (`SGD`, momentum = 0.9).We fix the main hyperparameters to the best configuration previously found while experimenting with Model 1 (learning rate $\approx 5.12 \times 10^{-4}$, batch size = 16), and we do not run a new tuning procedure for Model 2.
Each training run lasts 20 epochs.

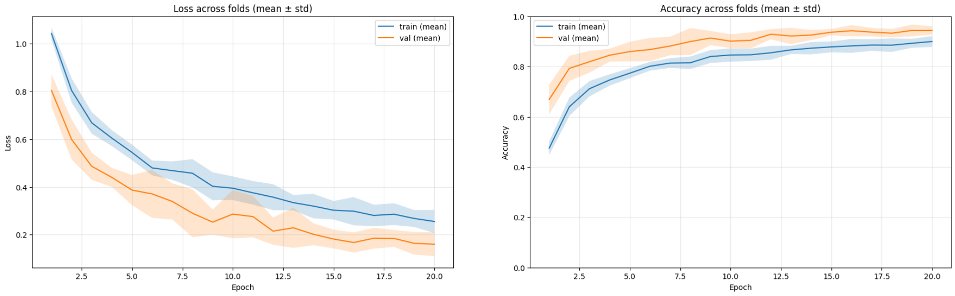
To improve reproducibility, we set a global seed (`SEED`=42) for Python, NumPy, and PyTorch, and we enable deterministic behavior where applicable.Training is executed on Apple Silicon `MPS`

4.1.3 K-fold cross-validation protocol

To perform cross-validation, we combine the available training and validation folders into a single dataset and then split it using `KFold` with $K = 10$, shuffling enabled and `random_state`=42.
For each fold, the model is trained *from scratch* on $K - 1$ folds and validated on the remaining fold, repeating the process until each fold has served as validation exactly once.
Within each fold, we build a shuffled training dataloader and a non-shuffled validation dataloader (batch size = 16, `num_workers`=2).
During training we log training and validation loss/accuracy at each epoch, and we track the best validation loss and best validation accuracy observed within each fold.
After completing the 10 folds, we retrain a final instance of the same architecture on the full (train+validation) dataset using the selected training configuration, and we save the resulting weights for the final test evaluation.

4.2 Results for Medium CNN with K-fold Cross-Validation

Cross-validation performance (K-fold): Across the 10 folds, the mean *best* validation accuracy was 0.9596 ± 0.0147 (mean ± std), while the mean *best* validation loss was 0.1294 ± 0.0352. Figure 2 reports the learning curves aggregated across folds (mean ± std), showing stable convergence and limited variance across splits.



(a) Model 2: loss across folds (mean ± std) over epochs.

(b) Model 2: accuracy across folds (mean ± std) over epochs.

Figure 2: Aggregated cross-validation learning curves for Model 2.

Final test performance: After cross-validation, a final instance of Model 2 was retrained on the full (train+validation) dataset and evaluated on the held-out test set. On the test set (classes: `paper`, `rock`, `scissors`, $n = 331$), the model achieved test loss 0.1294 and test accuracy 0.9668.

Confusion matrix: Figure 3 shows the confusion matrix on the test set (rows = true class, columns = predicted class), with the corresponding raw counts reported in Table 2.

Table 2: Model 2 confusion matrix counts on the test set (rows: true class; columns: predicted class).

	Pred: Paper	Pred: Rock	Pred: Scissors
True: Paper	100	1	7
True: Rock	0	109	1
True: Scissors	0	2	111

Precision/Recall/F1 (test set). Table 3 reports precision, recall, and F1-score per class on the test set (support shown in the last column). Overall, the macro-averaged F1-score was 0.9668 and the weighted-averaged F1-score was 0.9667.

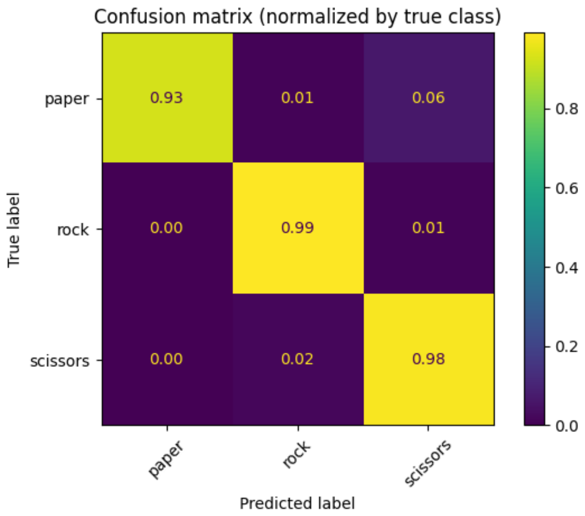


Figure 3: Model 2: confusion matrix on the test set.

Table 3: Model 2 classification report on the test set.

Class	Precision	Recall	F1-score	Support
Paper	1.0000	0.9259	0.9615	108
Rock	0.9732	0.9909	0.9820	110
Scissors	0.9328	0.9823	0.9569	113
Accuracy		0.9668		331
Macro avg	0.9687	0.9664	0.9668	331
Weighted avg	0.9681	0.9668	0.9667	331

Conclusion (Model 2). Model 2 achieves strong and consistent performance under cross-validation, with mean best validation accuracy 0.9596 and relatively low fold-to-fold variability. After retraining on the full development set, the final model generalizes well to the held-out test set, reaching 0.9668 accuracy with balanced precision/recall across the three classes. Most residual errors are limited and class-specific (e.g., a small number of `paper` samples predicted as `scissors`), as shown by the confusion matrix and per-class metrics.

5 Residual CNN Model and Training

5.1 Residual CNN and training procedure

In the third experiment we further increased model capacity by adopting a residual CNN architecture, i.e., a network built from residual blocks with skip connections, while keeping the same three-class classification objective (rock, paper, scissors). As for Model 2, model assessment is performed using **K-fold cross-validation**, which reduces dependence on a single train/validation split and provides a more stable estimate of generalization performance. The preprocessing and data augmentation pipeline is the same as in the model 1, with the exception of the train/validation split, which is now performed within each fold of the cross-validation procedure.

5.1.1 Residual blocks (BasicBlock)

Model 3 is based on residual learning, where the output of a transformation $F(\cdot)$ is added to the original input through a shortcut connection. In its simplest form, a residual block can be expressed as

$$y = F(x) + x,$$

which helps the optimization of deeper networks by allowing gradients to flow through identity paths.

In our implementation, each residual block (`BasicBlock`) contains two 3×3 convolutional layers, each followed by batch normalization, with a ReLU non-linearity after the first convolution and again after the residual addition. To introduce regularization, the block can optionally apply a spatial dropout (`Dropout2d`) after the first activation (with a drop probability set per stage). When the number of channels changes or spatial downsampling is required (i.e., `stride` $\neq 1$), the shortcut branch uses a projection consisting of a 1×1 convolution with matching stride followed by batch normalization, so that tensor dimensions align before the summation.

5.1.2 Architecture (ResCNN)

The network is organized into: (i) an initial convolutional stem, (ii) three residual stages, and (iii) a classification head.

Stem: The stem is composed of `Conv2d(3` → `16, kernel=7, stride=1, padding=3)` → `BatchNorm2d(16)` → `ReLU` → `MaxPool2d(2)`.

Residual stages.

- **Stage 1 (16 channels, no downsampling):** `BasicBlock(16` → `16, stride=1, drop=0.0)` ×2.
- **Stage 2 (32 channels, downsampling):** `BasicBlock(16` → `32, stride=2, drop=0.1)` followed by `BasicBlock(32` → `32, stride=1, drop=0.1)`.
- **Stage 3 (64 channels, downsampling):** `BasicBlock(32` → `64, stride=2, drop=0.2)` followed by `BasicBlock(64` → `64, stride=1, drop=0.2)`.

Classifier. After the residual stages, features are flattened and passed to `LazyLinear(128)` → `ReLU` → `Dropout(0.5)` → `Linear(128, 3)`. The output layer produces logits for the three target classes.

5.1.3 Training setup

Training uses a multi-class cross-entropy loss (`CrossEntropyLoss`) and stochastic gradient descent with momentum (`SGD`, momentum = 0.9). We keep the learning rate fixed to 5.115859×10^{-4} and use a batch size of 16, training for 20 epochs per fold. To improve reproducibility, we set a global seed (`SEED`=42) for Python, NumPy, and PyTorch, and enable deterministic behavior where applicable. Training runs on Apple Silicon `MPS`.

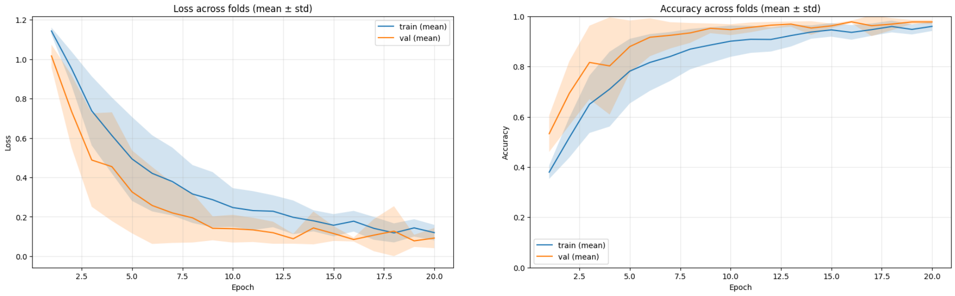
5.1.4 K-fold cross-validation protocol

Cross-validation is implemented by merging the existing training and validation folders into a single dataset (`ConcatDataset`) and then applying `KFold` with $K = 5$, shuffling enabled, and `random_state`=42. For each fold, the model is trained *from scratch* on $K - 1$ folds and validated on the remaining fold, so that each sample serves as validation exactly once. Within each fold, we build a shuffled training dataloader and a non-shuffled validation dataloader (batch size = 16, `num_workers`=0). During training we log train/validation loss and accuracy at each epoch and track the best validation metrics within each fold for later aggregation (reported in the Results section).

After completing cross-validation, we train a final instance of the same residual architecture on the full (train+validation) dataset using the same configuration and save the weights to disk for final test evaluation (reported separately).

5.2 Residual CNN with 5-fold Cross-Validation: Results

Cross-validation performance: Across the 5 folds, the mean *best* validation accuracy was 0.9849 ± 0.0036 (mean ± std), while the mean *best* validation loss was 0.0613 ± 0.0283. Figure 4 reports the learning curves aggregated across folds (mean ± std), showing strong convergence and limited variance across splits.



(a) Model 3: loss across folds (mean ± std) over epochs.

(b) Model 3: accuracy across folds (mean ± std) over epochs.

Figure 4: Aggregated cross-validation learning curves for Model 3.

Final test performance: After cross-validation, a final instance of Model 3 was retrained on the full (train+validation) dataset and evaluated on the held-out test set. On the test set (classes: `paper`, `rock`, `scissors`, $n = 331$), the model achieved test loss 0.0619 and test accuracy 0.9879.

Confusion matrix: Figure 5 shows the confusion matrix on the test set (rows = true class, columns = predicted class), with the corresponding raw counts reported in Table 4.

Table 4: Model 3 confusion matrix counts on the test set (rows: true class; columns: predicted class).

	Pred: Paper	Pred: Rock	Pred: Scissors
True: Paper	107	1	0
True: Rock	1	108	1
True: Scissors	0	1	112

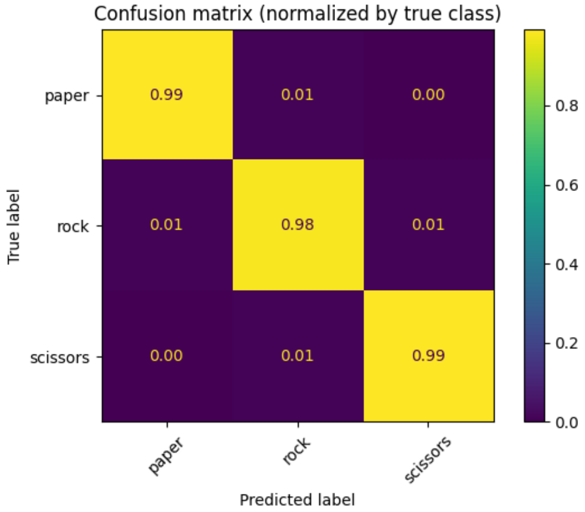


Figure 5: Model 3: confusion matrix on the test set.

Precision/Recall/F1: Table 5 reports precision, recall, and F1-score per class on the test set (support shown in the last column). Overall, the macro-averaged F1-score was 0.9879 and the weighted-averaged F1-score was 0.9879.

Table 5: Model 3 classification report on the test set.

Class	Precision	Recall	F1-score	Support
Paper	0.9907	0.9907	0.9907	108
Rock	0.9818	0.9818	0.9818	110
Scissors	0.9912	0.9912	0.9912	113
Accuracy		0.9879		331
Macro avg	0.9879	0.9879	0.9879	331
Weighted avg	0.9879	0.9879	0.9879	331

Conclusion for Model 3: Model 3 achieves excellent and stable performance under 5-fold cross-validation, with mean best validation accuracy 0.9849 and low fold-to-fold variability. After retraining on the full development set, the final model generalizes extremely well to the held-out test set, reaching 0.9879 accuracy with uniformly high precision/recall across all three classes. The confusion matrix confirms that errors are rare and mostly isolated (only a few samples per class are misclassified).

6 Conclusion

Across the three experiments, performance improves consistently as model capacity and architectural sophistication increase. Model 1 (a minimal single-convolution baseline tuned with Ray Tune) already achieves strong generalization, with 0.9154 test accuracy (test loss 0.2737 on $n = 331$). Model 2, which increases depth and introduces explicit regularization (Batch Normalization and Dropout) and is assessed via 10-fold cross-validation, provides a clear gain, reaching 0.9668 test accuracy (test loss 0.1294) with mean best validation accuracy 0.9596 ± 0.0147. Finally, Model 3 (residual CNN with skip connections) achieves the best and most stable results, with 0.9879 test accuracy (test loss 0.0619) and mean best validation accuracy 0.9849 ± 0.0036 under 5-fold cross-validation. Overall, Model 2 represents an effective accuracy–complexity trade-off, while Model 3 is the best final choice when maximizing accuracy and robustness is the primary objective.

7 Bibliography

References

[1] Understand "stride". https://medium.com/@bragadeeshs/stride-in-cnns-stepping-towards-efficient-image-processing-e58a34b02ff0

[2] Understand "padding". https://d2l.ai/chapter_convolutional-neural-networks/padding-and-strides.html

[3] Basic tutorials for inspiration: what is torch.nn?. https://docs.pytorch.org/tutorials/

[4] Basic tutorials for inspiration: training a Neural Network. https://docs.pytorch.org/tutorials/beginner/basics/buildmodel_tutorial.html

[5] Basic tutorials for inspiration: training an image classifier. https://docs.pytorch.org/tutorials/beginner/blitz/cifar10_tutorial.html

[6] Understand conv2d parameters. https://www.codegenes.net/blog/conv2d-parameter-object-input-pytorch/

[7] Hyperparameter fine-tuning with Ray. https://docs.pytorch.org/tutorials/beginner/hyperparameter_tuning_tutorial.html

[8] Stochastic Gradient Descent and Momentum. https://www.lunartech.ai/blog/mastering-stochastic-gradient-descent-the-backbone-of-deep-learning-optimization

[9] SkLearn for Cross Validation. https://discuss.pytorch.org/t/how-can-i-use-sklearn-kfold-with-imagefolder/36577

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.