REPORT OF THE PROJECT

The objective of this project is to evaluate and develop convolutional neural networks for the classification of images from the Rock–Paper–Scissors game. Convolutional neural networks are a specialized class of artificial neural networks designed to process data with a grid-like topology, such as images. CNNs make use of convolutional layers, where filters (also called kernels) slide across the input and automatically learn to detect relevant visual features. At the lower levels, these features may include simple patterns such as edges and corners, while deeper layers capture more complex structures such as textures, shapes, and object parts.

This hierarchical feature extraction makes CNNs particularly effective in computer vision tasks, as they are able to reduce the need for manual feature engineering. Moreover, CNNs exhibit translation invariance, meaning that they can recognize patterns regardless of their position in the image, and they tend to generalize well to unseen data when properly trained and regularized.

In the context of this project, the Rock–Paper–Scissors dataset offers a controlled but challenging classification problem, since the hand gestures share similar visual characteristics and may vary in orientation, lighting, and background conditions. The goal is therefore to design and compare different CNN architectures, analyse their performance using metrics such as accuracy, precision, recall, and F1-score, and discuss their ability to generalize beyond the training data.

The workflow proceeds through the following main stages:

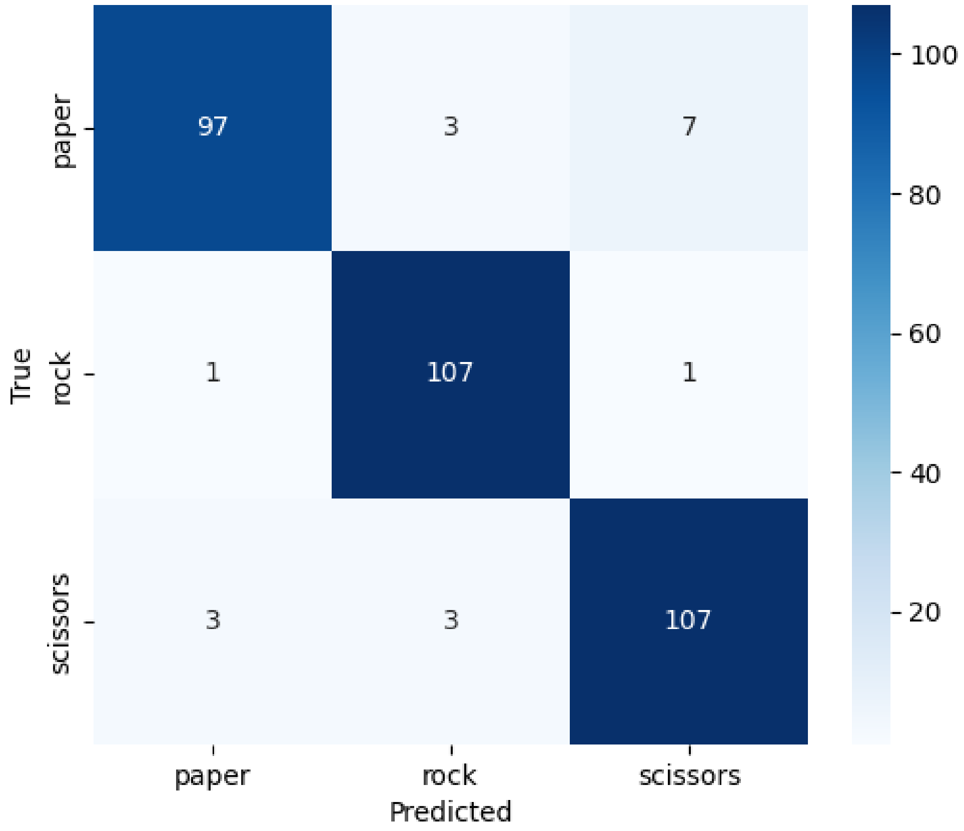
- The dataset preparation, where the dataset is loaded and organized into separate classes. The data are divided into training, validation and test sets to ensure that the models are evaluated on unseen images.
- The image preprocessing and augmentation, where all images are resized to a fixed resolution and pixel values are normalized to improve training stability. Data augmentation techniques such as rotation, flipping and zooming are applied to artificially increase dataset variability and help the models to generalize better.
- The creation of three distinct convolutional neural network which consist of: convolutional and pooling layers to extract spatial features from the images, fully connected layers to map extracted features to class probabilities and dropout layers to reduce overfitting.
- The training configuration, where the accuracy is chosen as the main evaluation metric during training.
- The cross-validation and hyperparametric tuning, to ensure robust result the k-fold cross-validation is applied, at the same time, a grid search is used to

automatically test combinations of key hyperparameters, such as learning rate, dropout rate, and batch size. This systematic approach identifies the most effective configuration for each model.

- Model evaluation, where the standard performance metrics are computed to provide a balanced view of classification quality.
- Visualization of results, where confusion matrix, loss/accuracy curves and sample prediction are used to analyse model behaviour.
- Model comparison, all three models are summarized in a table including test performance and cross-validation.

Now that I have presented how the code works and what I used it for, I will show the experimental results obtained from it for all the model.

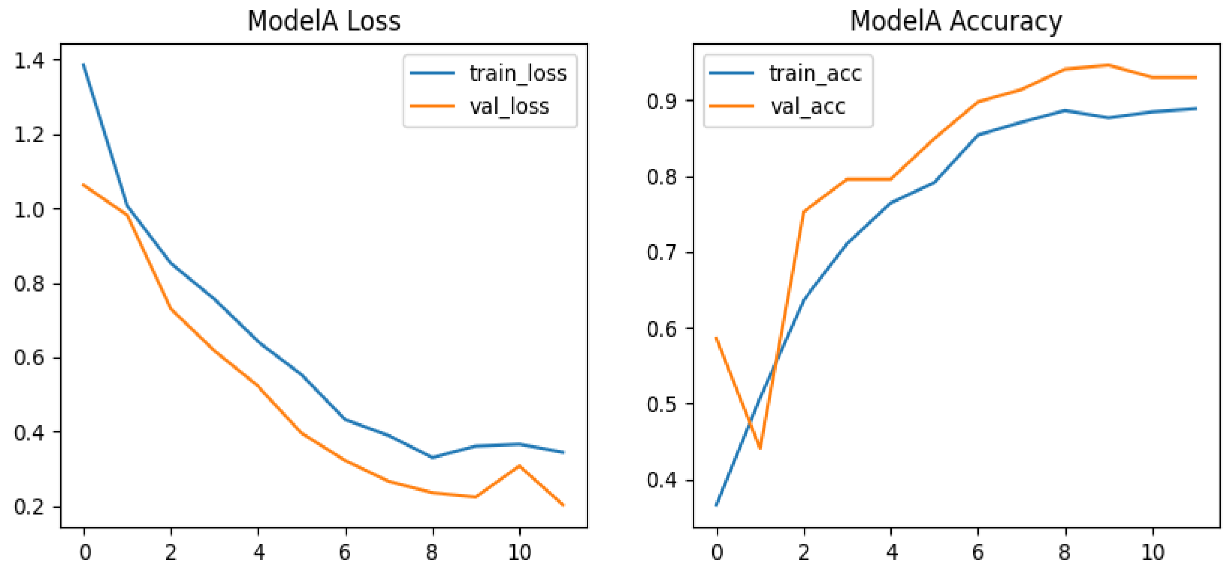
MODEL A



The confusion matrix of Model A on the test set shows that the model performs very well in classifying the Rock-Paper-Scissors images, specifically:

- **Paper**: 97 images were correctly classified as paper, while 3 were misclassified as rock and 7 as scissors.
- **Rock**: 107 out of 109 images were correctly classified, with only 1 misclassified as paper and 1 as scissors. This indicates that rock is the most reliably recognized gesture.
- **Scissors**: 107 images were correctly classified, with 3 incorrectly predicted as paper and 3 as rock.

Overall, the matrix highlights that most errors occur between the paper and scissors classes, which the model tends to confuse in some cases. Despite these minor misclassifications, Model A demonstrates high classification accuracy and balanced performance across all three classes.

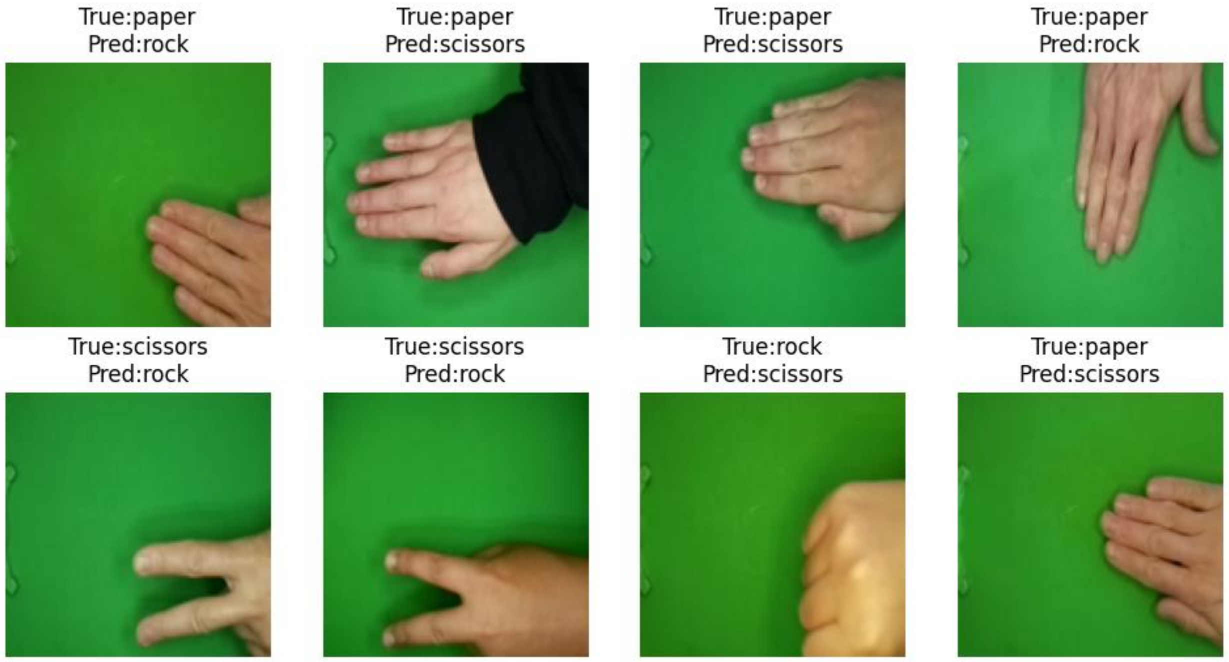


The plots of loss and accuracy provide insights into the training dynamics:

- **Loss Curve**: Both the training and validation losses steadily decrease over the epochs, with no signs of divergence. This indicates that the model is learning effectively and is not overfitting to the training data. The validation loss is even slightly lower than the training loss in the later epochs, which suggests good generalization to unseen data.
- **Accuracy Curve**: The training accuracy increases consistently, while the validation accuracy improves rapidly after the first epochs and stabilizes above

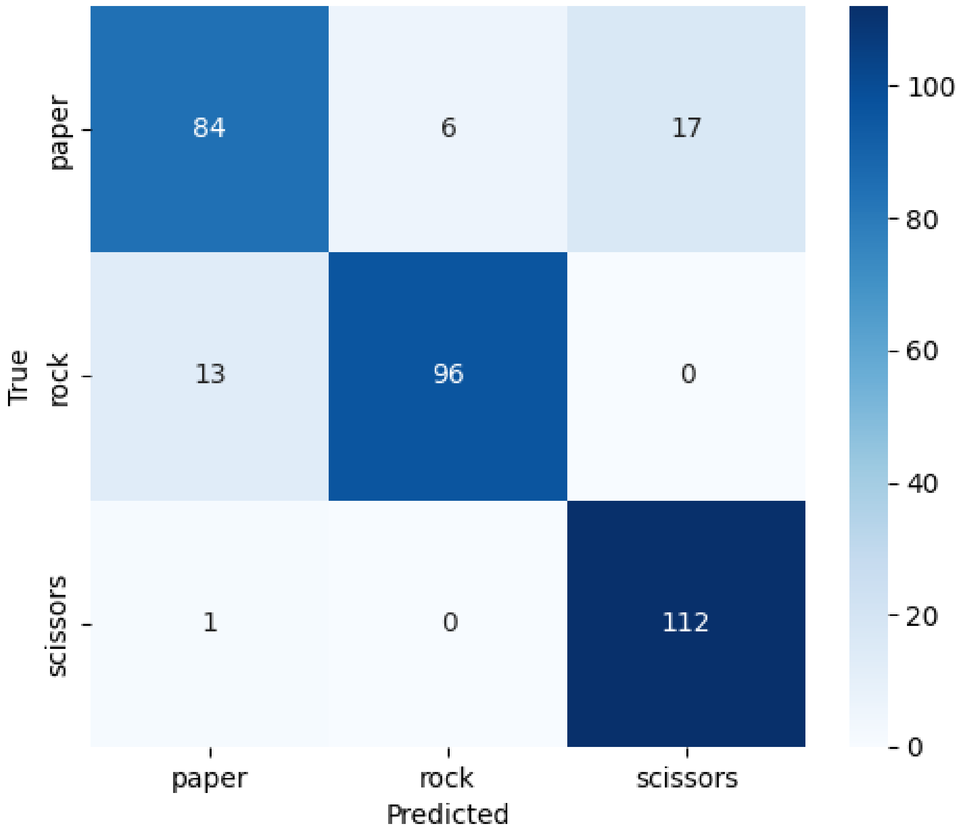
90%. The close alignment between the two curves demonstrates that the model maintains a balanced learning process without significant overfitting or underfitting.

Overall, these curves confirm that Model A converges well during training and achieves strong performance on the validation set.



The figure illustrates examples where the CNN misclassified hand gestures. In particular paper was confused with scissors or rock, this may be due to the similar shape of a flat hand compared to certain orientations of the rock gesture, or to ambiguities introduced by the angle of the hand or partial occlusion- Some scissors gestures were predicted as rock. This typically happens when the two extended fingers are not clearly visible, or when the image quality and background cause the model to focus on the palm area rather than the finer details of the fingers. Variability in hand positioning and lighting conditions may lead to reduced feature distinctiveness. The green background helps to some extent, but shadows or overlapping parts of the hand might confuse the model. These errors highlight that although the model performs well overall there are still cases where subtle visual similarities lead to misclassification.

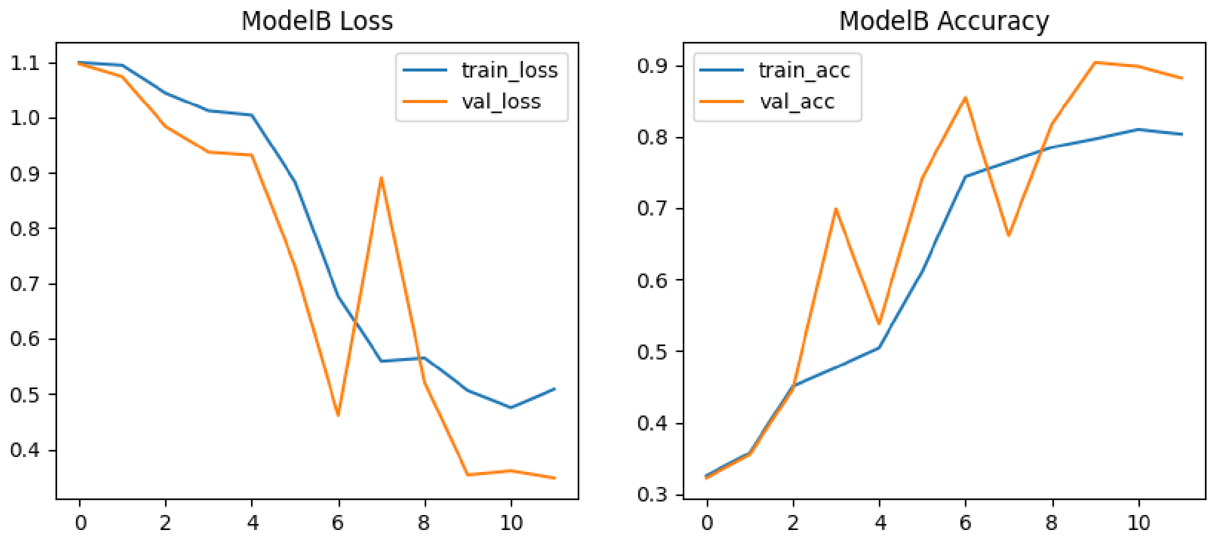
MODEL B



The confusion matrix of Model B gave also good result in classification, specifically:

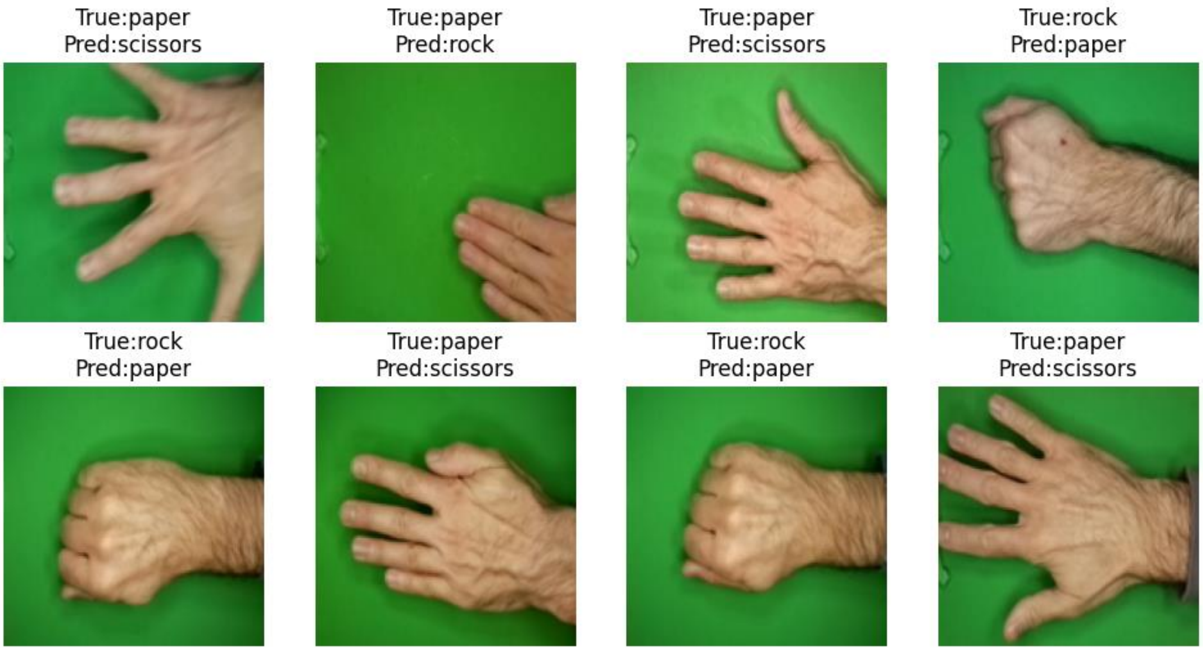
- **Scissors**: 112 out of 113 samples were correctly classified, showing that Model B is highly reliable at recognizing the scissors gesture. Just one case of misclassification so scissors is the most robustly recognized class.
- **Rock**: 96 out of 109 samples were correctly classified, which is a strong performance but leaves room for improvement. In all this cases, rock were misclassified in paper, this could happen when the hand is not full clenched making the shape looking similar to a flat hand.
- **Paper**: 84 out of 107 samples were correctly classified, which is notably lower compared to scissors. In particular 17 cases of paper misclassified in scissors, this suggest that the model sometimes fails to distinguish between a flat hand and two extended fingers, especially when fingers are not clearly separated. We got also 6 cases where paper was misclassified in rock, a small number, probably due to the similarities in hand orientation.

Compared to Model A, Model B shows **a** stronger specialization in scissors recognition, but struggles slightly more with paper and rock separability.



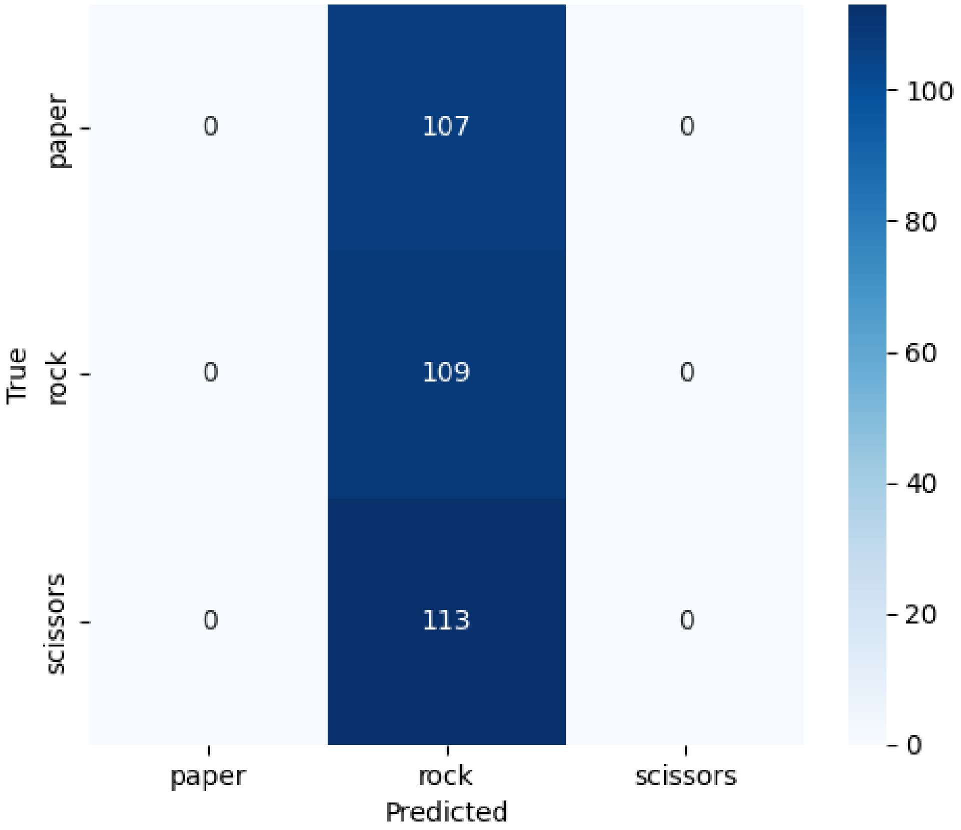
For the loss curves, at the beginning, both training and validation loss are relatively high, which is expected since the model starts untrained. Both losses decrease steadily across epochs, indicating that the model is learning meaningful features. The validation loss drops below the training loss around epoch 6 and remains lower until the end. This can happen due to regularization techniques (like data augmentation) that make the training set harder than the validation set. A small fluctuation in validation loss around epoch 7 suggests some instability, but it quickly recovers.

While for the accuracy curves we can say that: training accuracy starts low, then steadily increases to 0.80 by the final epochs. Validation accuracy shows a more rapid improvement, reaching 0.90 near the end of training. We can say that validation accuracy is consistently higher than training accuracy, so the model generalizes relatively well on unseen data and that data augmentation or dropout make the training task harder. The fact that validation accuracy stay high shows no sign of overfitting and the gap between training and validation accuracy indicates that the model is learning and has strong generalization.

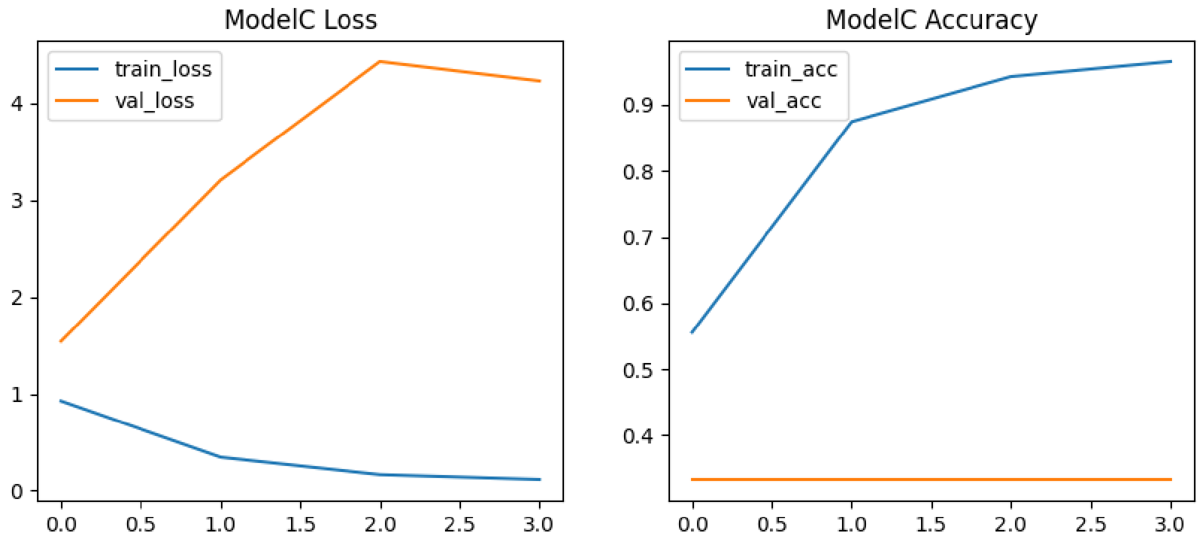


This figure displays several test images where the predicted label does not match the true class. For example a “Paper” gesture is sometimes classified as “Scissors” or “Rock” gesture is misclassified as “Paper”. Paper and Scissors both involve extended fingers, which may confuse the network, especially if the hand is partially closed or rotated, while rock and paper both cover most of the hand area, making it harder for the model to distinguish them in certain angles. Some other reasons could be that since all gestures share the same green background, the network may rely heavily on subtle hand features, making it sensitive to noise or lighting variations, or the fact that different individuals might perform the same gesture slightly differently could increase intra-class variability. We could say that the model focus strongly on finger position but struggles when fingers are only partially visible. These misclassifications highlight limitations in the model’s generalization ability.

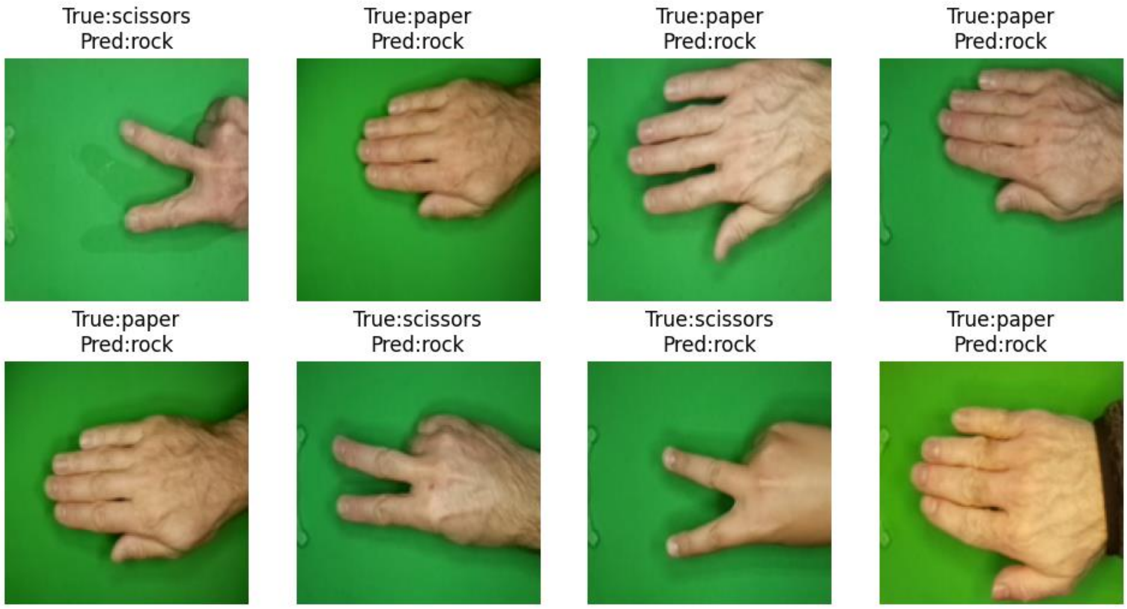
MODEL C



The confusion matrix of Model C shows a critical issue in the model’s behaviour: none of the classes were classified correctly. All predictions, regardless of the true label, were assigned to the rock class. The model is suffering from severe class imbalance in predictions, instead of learning to discriminate between the three classes, it defaulted to always predicting the majority or easiest class. This results show an high accuracy only for the rock class, but a total failure for paper and scissors.

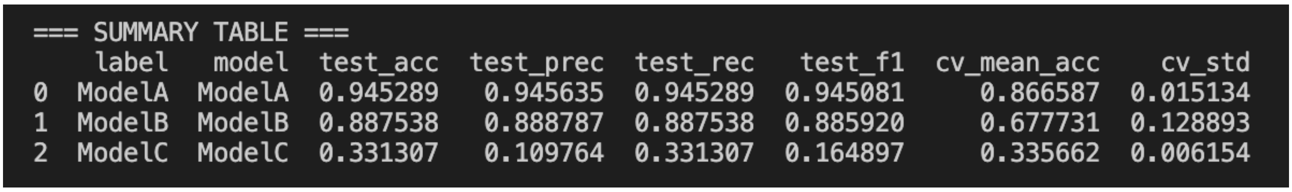


From the analysis of Training and Validation Curves we can say that the training loss decreases steadily, reaching very low values after just a few epochs, while the training accuracy increases quickly and reaches almost 100%. This suggests that the model is able to fit the training data extremely well. The validation loss behaves in the opposite way: it starts increasing rapidly, reaching values above 4. The validation accuracy remains constant across all epochs, which corresponds to random guessing among three classes, this indicates that the model completely fails to generalize to unseen data. This is a classic case of overfitting, where the model memorizes the training set perfectly but does not learn meaningful features for classification. At the same time this model shows signs of model collapse predicting always the same class.



In this case, all misclassifications are biased toward the “Rock” class. The network seems to have collapsed into predicting “Rock” almost exclusively. Unlike prediction of model B, where errors were diverse (Paper confused with Scissors or Rock), here the problem is systematic: the model fails to generalize beyond the Rock class. This indicates that the model is not balanced in how it recognizes the three gestures. This visualization reveals a serious limitation: the model is essentially unusable in its current form, since it lacks the ability to correctly classify Paper and Scissors.

Model comparison



```
=== SUMMARY TABLE ===
    label   model  test_acc  test_prec  test_rec   test_f1  cv_mean_acc    cv_std
0  ModelA  ModelA  0.945289   0.945635  0.945289  0.945081     0.866587  0.015134
1  ModelB  ModelB  0.887538   0.888787  0.887538  0.885920     0.677731  0.128893
2  ModelC  ModelC  0.331307   0.109764  0.331307  0.164897     0.335662  0.006154
```

Model A achieved the highest performance across all metrics, with a test accuracy of 94.5% and strong precision, recall, and F1 scores. Cross-validation results confirm its stability, with a mean accuracy of 0.867 and a low standard deviation, showing consistent generalization across folds.

Model B reached a good but lower accuracy (88.7%), with precision, recall, and F1 scores all around 0.88. Cross-validation mean accuracy is 0.678 with a higher standard deviation than model A, indicating less stable performance depending on the data splits. Still, it demonstrates a decent ability to generalize, though weaker than Model A.

Model C performs at chance level, with accuracy 0.33 and very poor precision and F1. Cross-validation confirms this collapse, with mean accuracy 0.336 and negligible variation. This indicates that Model C failed completely to learn meaningful features and collapsed into always predicting a single class.

In conclusion the model A result to be the best model, it has strong result across test set, cross-validation and confusion matrix. Model B could be an acceptable alternative being still a functional model despite the lower accuracy. While model C demonstrate to be unsuccessful, it failed to generalize with results equivalent to random guessing.