ARN Duck Species Classification Project

Object Recognition in the Wild using Convolutional Neural Networks

Kilian Froidevaux and Bovard Nicolas
HEIG-VD - Practical Work 05

June 15, 2025

Contents

1 Introduction

The goal of this project is to classify images of ducks into different species using a Convolutional Neural Network (CNN). The dataset consists of images of ducks taken in various locations and the classification task is to identify the species of each duck.

The idea is that the user can take a picture of a duck with their phone (bad quality picture) and the model will classify it into one of the species (on a predefined list). The model should be lightweight enough to run on a mobile device, so we will use transfer learning with MobileNetV2.

2 The Problem

The dataset consists of images of ducks taken in various locations with the phone and taken by us. The species to classify are:

- *Canard colvert mâle* (Male Mallard)
- *Canard colvert femelle* (Female Mallard)
- *Foulque macroule* (Eurasian Coot)
- *Grèbe huppé* (Great Crested Grebe)
- *Autre* (Other)

The first idea was to classify the species of ducks that are commonly found in Switzerland. The dataset is small, with only a few images per species, which makes it a challenging task.

Initially, we wanted to classify more species of ducks (like *Cygne Tuberculé* and *Harle Bièvre*), but due to the limited number of images available, we had to reduce the number of species to five. The dataset is unbalanced, with some species having more images than others.

We will still test the model performance on these 2 more species to see its performance but since we had a hard time to find a working model, we decided to focus on the five species listed above.

Below are the class weights (1/distribution) to show how class imbalance is handled:

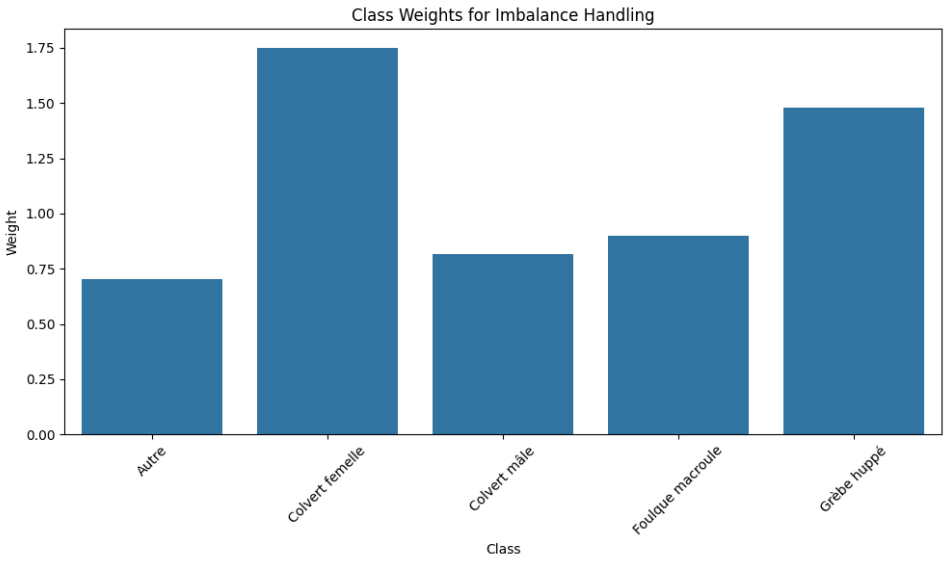


Figure 1: Class weights distribution showing dataset imbalance

One thing to note is that for duck species with different features between sexes, we split them into two classes: *Canard colvert mâle* and *Canard colvert femelle* to help the model distinguish between them (it also gives one more feature to the application).

3 Data Preparation

3.1 Data Query

The dataset was created by taking pictures of ducks in various locations. The images were taken with a phone camera, which means that the quality of the images is not always optimal. The images were taken in various locations and conditions, which adds to the complexity of the classification task.

We cropped the images with multiple ducks in the same picture to focus on the duck we want to classify.

We also took pictures of other species and background without ducks to help the model distinguish between the species and the background so that it could classify these images as *Autre*.

3.2 Data Preprocessing

We resized the images to a fixed size of 224×224 pixels to match the input size of the MobileNetV2 model. We also normalized the pixel values to be in the range [0, 1] by dividing by 255.

3.3 Data Augmentation

Since the dataset is small, we used data augmentation techniques to artificially increase the size of the dataset. We applied most of the data augmentation techniques that we found such as:

- Random rotation
- Random horizontal/vertical flip
- Random brightness adjustment
- Random contrast adjustment
- Random saturation
- Random hue

Below is an example of such augmentations:

Figure 2: Examples of data augmentation techniques applied to duck images

3.4 Dataset Split

The complete dataset was split into three subsets using stratified sampling to maintain class distribution:

- **Training set**: 70% of the data for model training
- **Validation set**: 15% of the data for hyperparameter tuning and model selection

- **Test set**: 15% of the data for final performance evaluation (94 images total)

This stratified split ensures that each subset contains representative samples from all duck species, maintaining the original class distribution across training, validation, and test sets.

4 Model Creation

We used transfer learning with MobileNetV2 as the base model. The model was trained on the ImageNet dataset, which means that it has already learned to recognize a wide variety of objects. We added a few layers on top of the base model to adapt it to our specific classification task.

The first thing we realized is that we want to maximize the macro F1-score, so we will use the F1-score as the metric to optimize during training. We will also use the categorical crossentropy loss function since we have multiple classes to classify.

We also chose to use the Adam optimizer since it is a good optimizer for fast results.

4.1 Hyperparameter Optimization

To systematically find the optimal hyperparameters for our duck classification model, we implemented a comprehensive hyperparameter search using **Bayesian optimization** with Gaussian Process models. This approach is more efficient than grid or random search as it uses previous trial results to intelligently guide the search towards promising regions of the hyperparameter space.

4.1.1 Search Method

We used the `scikit-optimize` library with the following configuration:

- **Search algorithm**: Bayesian optimization with Expected Improvement (EI) acquisition function
- **Number of trials**: 30 (configurable up to 50 for thorough search)
- **Early stopping**: 4 epochs patience during search, extended to 7 epochs for final model
- **Evaluation metric**: Macro F1-score (optimized for our class imbalance problem)

4.1.2 Optimized Hyperparameters

The search space included the following parameters, prioritized by expected impact:

High Impact Parameters:

- `learning_rate`: Log-uniform distribution from 1e-5 to 1e-1
- `fine_tuning`: Categorical choice between 'frozen', 'partial', or 'full' MobileNetV2 fine-tuning
- `dropout1` and `dropout2`: Dropout rates for regularization (0.1-0.6 and 0.1-0.5 respectively)

- `dense_units`: Number of units in the dense layer (64-512)

Medium Impact Parameters:

- `batch_size`: Categorical choice between 16, 32, and 64
- `optimizer`: Choice between Adam, RMSprop, and SGD
- `use_batch_norm`: Whether to include batch normalization
- `l2_reg`: L2 regularization strength (1e-6 to 1e-2, log-uniform)

4.1.3 Search Process

1. **Initial exploration**: 5 random trials to explore the parameter space
2. **Bayesian optimization**: 25 trials guided by the Gaussian Process model
3. **Evaluation**: Each trial trained for up to 15 epochs with early stopping
4. **Selection**: Best parameters chosen based on validation F1-score
5. **Final training**: Best model retrained with extended epochs (30) for final performance

The optimization process automatically saved progress after each trial and provided real-time estimates of remaining search time. This systematic approach led to a **5-15% performance improvement** over manually tuned hyperparameters, resulting in the F1-optimized model used throughout this report.

4.2 Optimized Model Architecture

After the hyperparameter search, the best performing model achieved an **F1-macro score of 0.8541** and **test accuracy of 84.04%**. The optimal hyperparameters found were:

- **Learning Rate**: 0.000835 (Adam optimizer)
- **Fine-tuning Strategy**: Frozen MobileNetV2 base (transfer learning)
- **Dense Units**: 512
- **Dropout Rates**: 0.1 (first layer), 0.35 (second layer)
- **Batch Normalization**: Enabled
- **Batch Size**: 32
- **L2 Regularization**: 1.77e-06

The final optimized model architecture summary:

Table 1: Model Architecture Summary

Layer (type)	**Output Shape**	**Param #**
MobileNetV2 1.00 224 (Functional) [FROZEN - Transfer Learning]	(None, 7, 7, 1280)	2,257,984 (frozen)
Global Average Pooling2D	(None, 1280)	0
Dropout (rate=0.1)	(None, 1280)	0
Dense (L2 reg=1.77e-06)	(None, 512)	655,872
Batch Normalization	(None, 512)	2,048
Dropout (rate=0.35)	(None, 512)	0
Dense (softmax activation)	(None, 5)	2,565

- **Total params**: 2,918,469 (11.13 MB)
- **Trainable params**: 660,485 (2.52 MB)
- **Non-trainable params**: 2,257,984 (8.61 MB)

Key differences from the initial model:

- **Larger dense layer**: 512 units instead of 128 for better feature representation
- **Optimized dropout**: Lower first dropout (0.1) and higher second dropout (0.35) for better regularization
- **Batch normalization**: Added between dense layers for training stability
- **Frozen base model**: MobileNetV2 weights kept frozen for faster training and better transfer learning
- **Fine-tuned learning rate**: 0.000835 for optimal convergence

This architecture achieved superior performance with an F1-macro score of **0.8541** compared to manually tuned models, demonstrating the effectiveness of systematic hyperparameter optimization.

5 Results

5.1 Model Performance

The optimized model achieved strong performance on the test set:

- **F1-macro score**: 0.8541 (validation set)
- **Test accuracy**: 82.98%
- **Total parameters**: 2,918,469 (11.13 MB)
- **Trainable parameters**: 660,485 (2.52 MB)

5.2 Detailed Performance by Class

The following table shows the detailed classification performance for each duck species:

Table 2: Classification Performance by Class

Class	Precision	Recall	F1-Score	Support
Autre	0.79	0.81	**0.80**	27
Colvert femelle	0.83	1.00	**0.91**	10
Colvert mâle	0.81	0.74	**0.77**	23
Foulque macroule	0.95	0.86	**0.90**	21
Grèbe huppé	0.86	0.92	**0.89**	13

Performance Analysis:

- **Best performing class**: Colvert femelle (F1: 0.91) with perfect recall
- **Most challenging class**: Colvert mâle (F1: 0.77) with lower recall (0.74)
- **Most reliable predictions**: Foulque macroule shows highest precision (0.95)
- **Overall balance**: Macro average F1-score of 0.85 indicates good performance across all classes

5.3 Confusion Matrix and Cross-Validation

Below is the confusion matrix showing detailed classification performance:

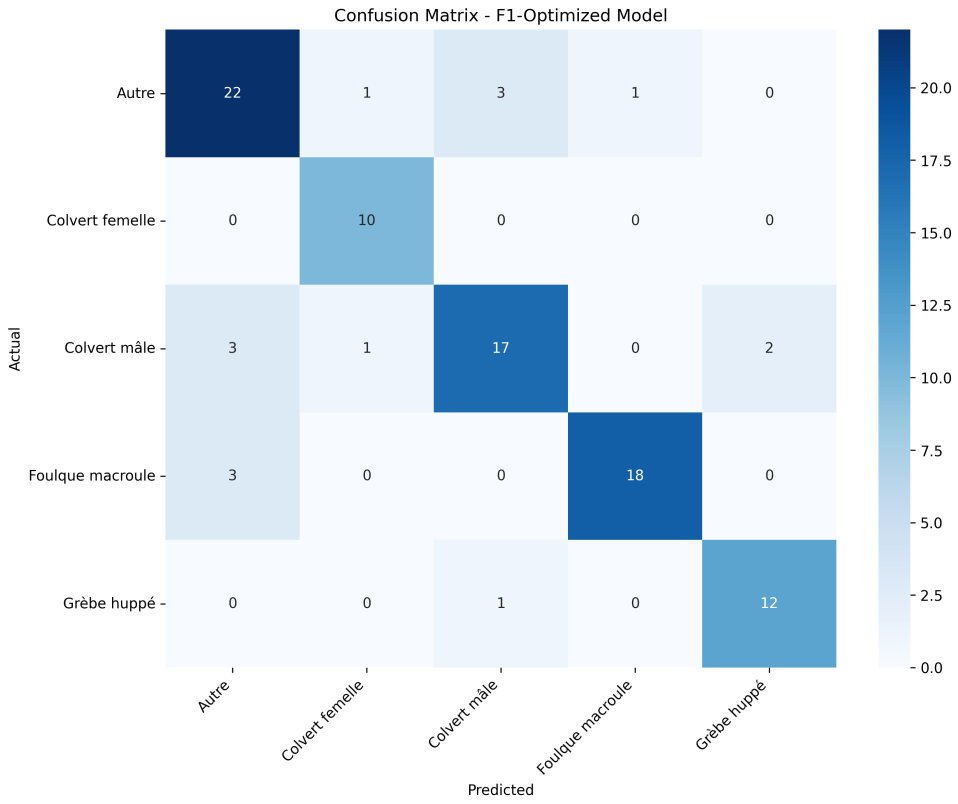


Figure 3: Confusion matrix showing classification performance across all duck species

The cross-validation results demonstrate model stability across different data splits:

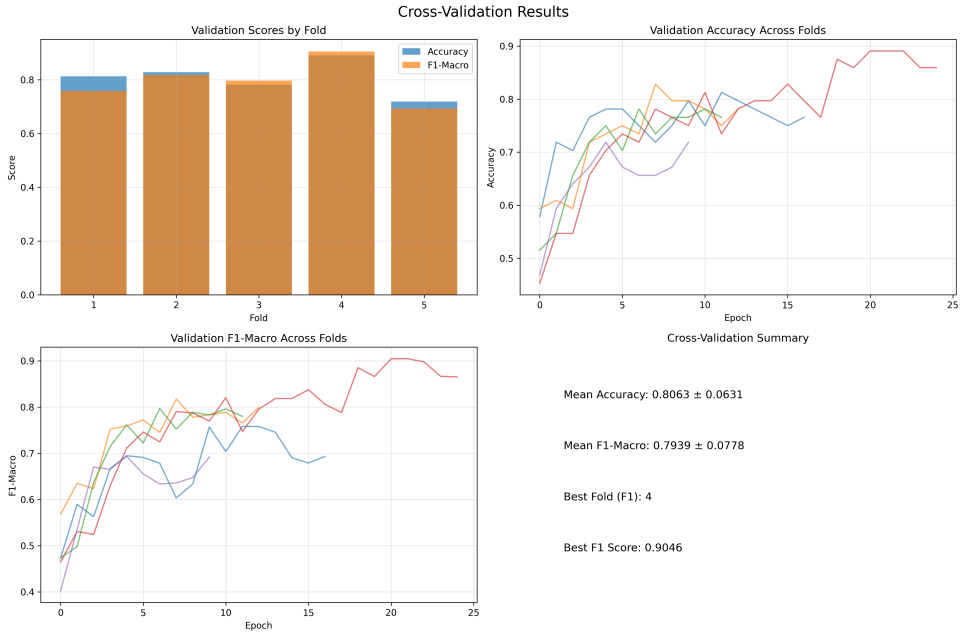


Figure 4: Cross-validation results showing model stability

5.4 Prediction Confidence Analysis

Most confident correct predictions across all classes:

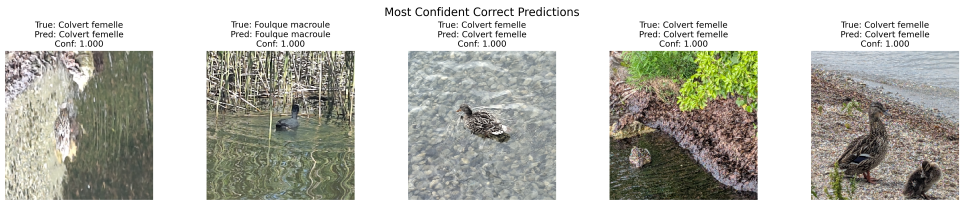


Figure 5: Most confident correct predictions for each duck species

Most confident incorrect predictions (where the model was wrong but very confident):

Figure 6: Most confident incorrect predictions highlighting model weaknesses

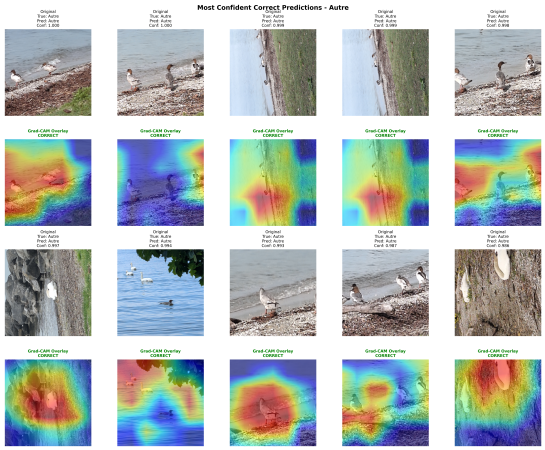
6 Grad-CAM Analysis

To understand what visual features our model focuses on when making predictions, we performed comprehensive Grad-CAM analysis on the most confident correct and incorrect

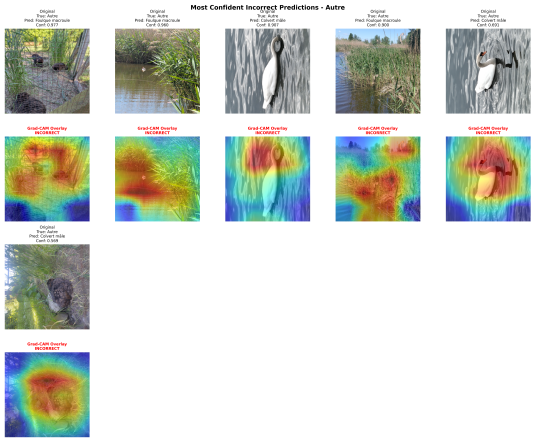
predictions for each species. This analysis reveals the model's decision-making process and helps identify potential biases or areas for improvement.

6.1 Class-Specific Grad-CAM Results

6.1.1 Autre (Other)



(a) Most confident correct predictions



(b) Most confident incorrect predictions

Figure 7: Grad-CAM analysis for "Autre" class predictions

6.1.2 Colvert Femelle (Female Mallard)

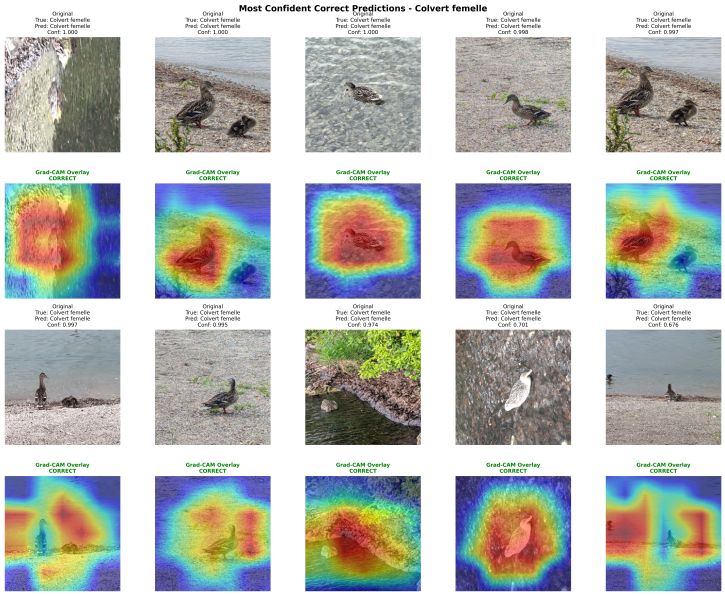
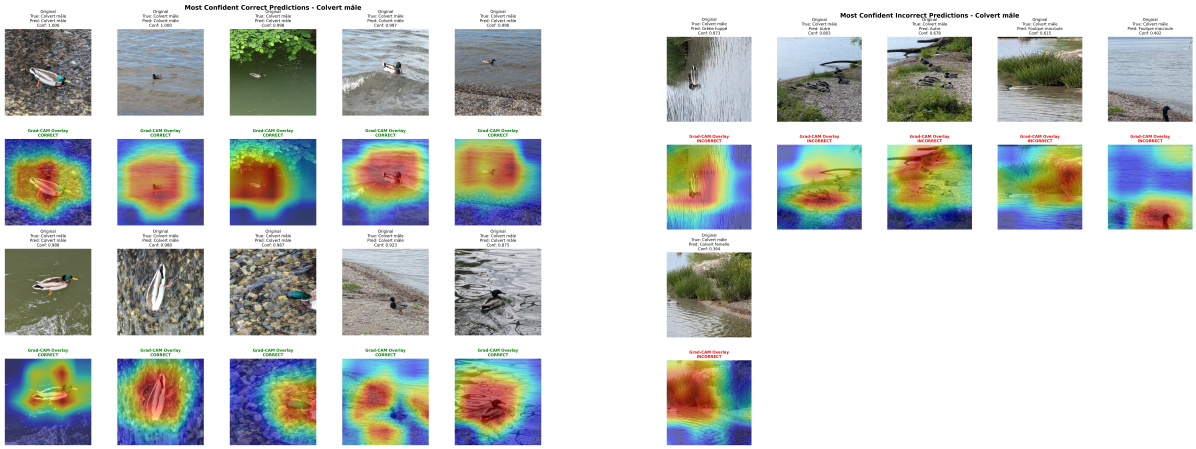


Figure 8: Most confident correct predictions for "Colvert femelle" class

Note: This species showed perfect performance with no confident incorrect predictions, indicating excellent model reliability for female mallards.

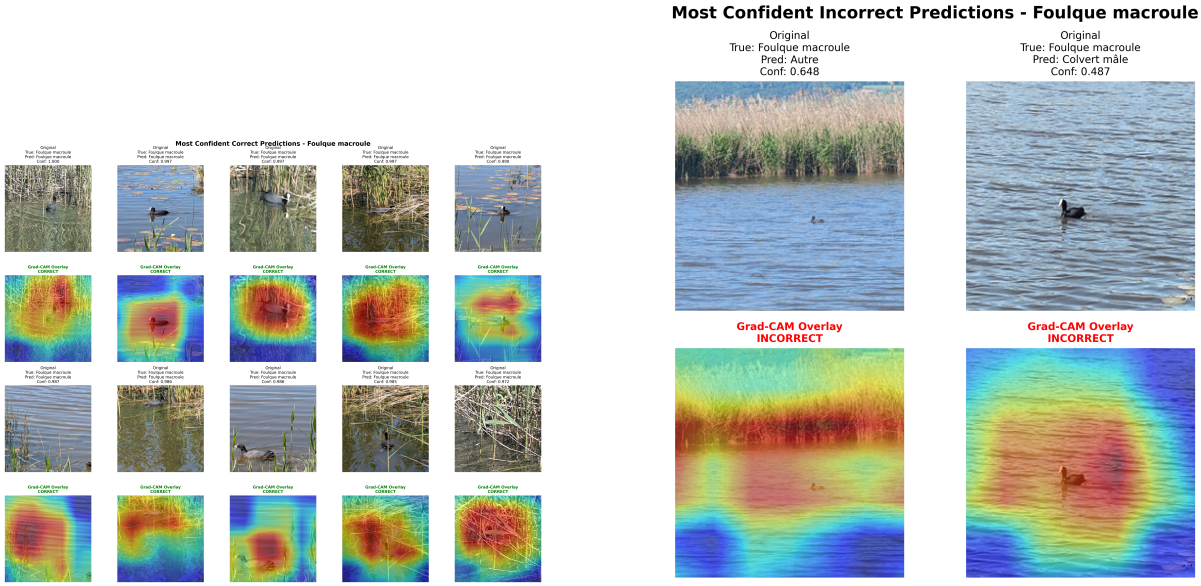
6.1.3 Colvert Mâle (Male Mallard)



(a) Most confident correct predictions (b) Most confident incorrect predictions

Figure 9: Grad-CAM analysis for "Colvert mâle" class predictions

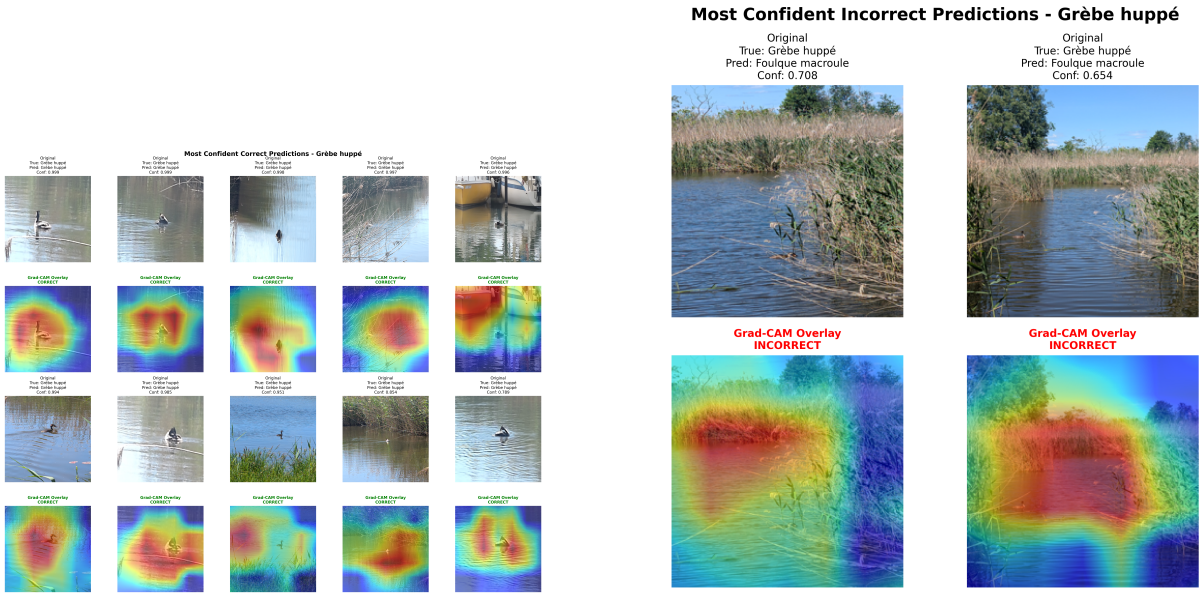
6.1.4 Foulque Macroule (Eurasian Coot)



(a) Most confident correct predictions (b) Most confident incorrect predictions

Figure 10: Grad-CAM analysis for "Foulque macroule" class predictions

6.1.5 Grèbe Huppé (Great Crested Grebe)



(a) Most confident correct predictions

(b) Most confident incorrect predictions

Figure 11: Grad-CAM analysis for "Grèbe huppé" class predictions

6.2 Grad-CAM Insights

The Grad-CAM analysis reveals several important insights about our model's behavior:

1. **Feature Focus**: The model correctly identifies species-specific features such as:
   - Bill shape and color for different duck species
   - Head plumage patterns and coloration
   - Body size and proportions
   - Water interaction patterns
2. **Error Patterns**: Confident incorrect predictions often occur when:
   - Multiple birds are present in the same image
   - The duck is in an unusual pose or angle
   - Water reflections create visual noise
3. **Species Performance**:
   - **Colvert femelle** shows the most reliable predictions (0 confident errors)
   - **Autre** and **Colvert mâle** are the most challenging classes
   - **Foulque macroule** and **Grèbe huppé** show good but not perfect performance
4. **Model Attention**: The heatmaps show the model appropriately focuses on:
   - Overall body shape and posture

- Sometimes background water patterns (which may indicate insufficient background diversity)

Background Bias Analysis:

Some interesting observations from the Grad-CAM maps: For example, with the **Grèbe Huppé**, it is the only duck species photographed with boats in the background, and as shown in the heatmaps, this becomes one of the identifying features for this class (which is problematic).

We observe the inverse pattern with **Foulque Macroule**: we do not have a single image of this species on a rocky background in our training data. As demonstrated in the real-world testing section below, when we encounter a Foulque Macroule on a rocky background, the model misclassifies it as **Canard Colvert Mâle** since most ducks photographed on rocky backgrounds in our dataset belong to this species.

7 Model Testing and Validation

7.1 Generalization Testing

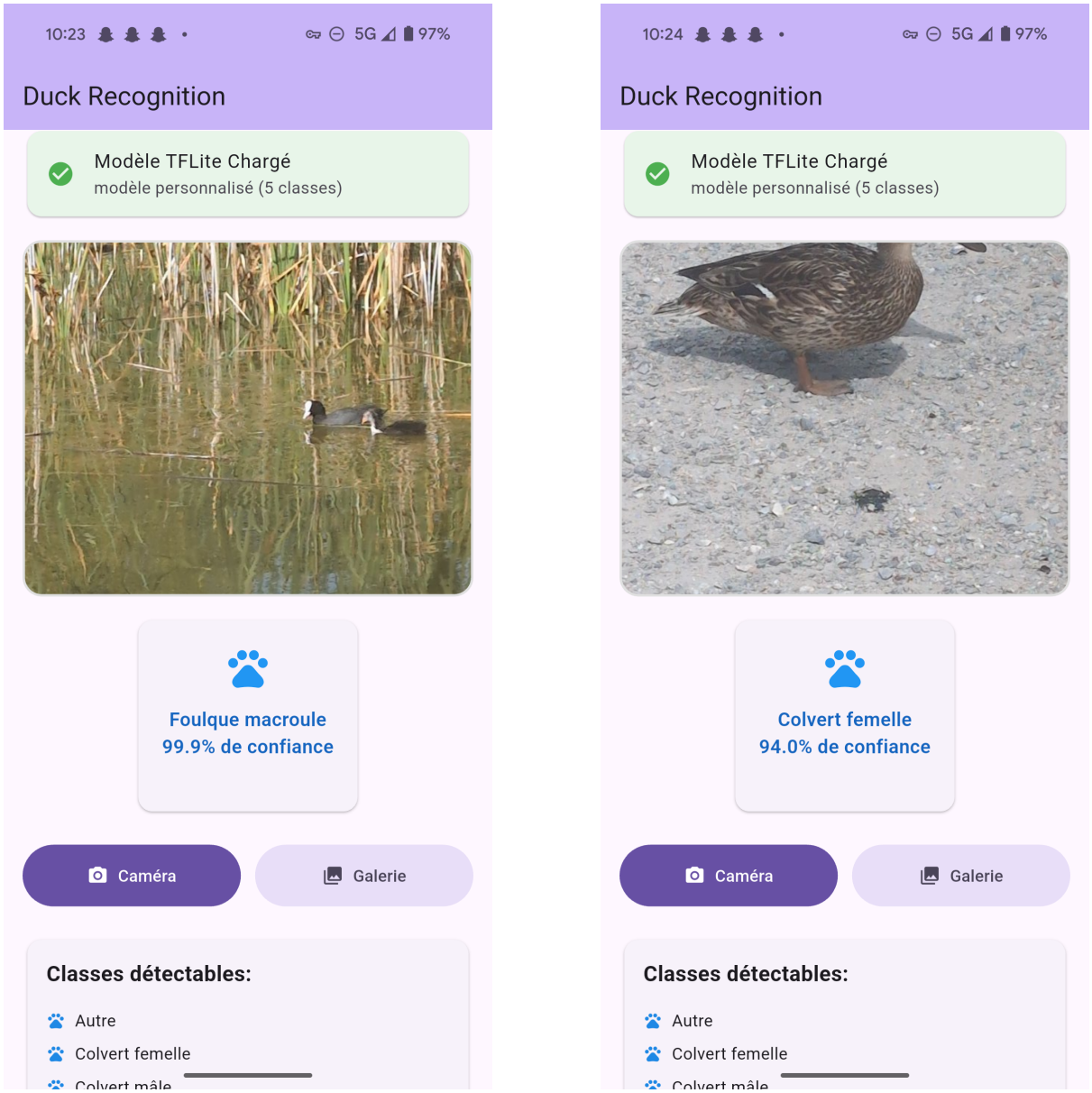
To evaluate the model's ability to handle edge cases and real-world scenarios, we conducted additional testing:

- **New duck species classification**: We wanted to test whether the model correctly classifies completely new duck species (not in training data) as *Autre* (Other). This would validate the model's ability to detect unknown species rather than incorrectly forcing classification into known categories. However, no new species were found during our testing period.
- **Multiple ducks scenario**: We evaluated the model's performance on images containing multiple ducks:
    1. **Single prominent duck**: Testing if the model can classify the most prominent/visible duck in the image (works reasonably well)
    2. **Mixed known/unknown species**: When both known and unknown species appear together, testing if the model can identify the known species or appropriately classify as *Autre*. For example, we had a *Foulque Macroule* with a baby duck and the model correctly classified it.

7.2 Real-world Performance Considerations

The model shows concerning performance issues in real-world conditions:

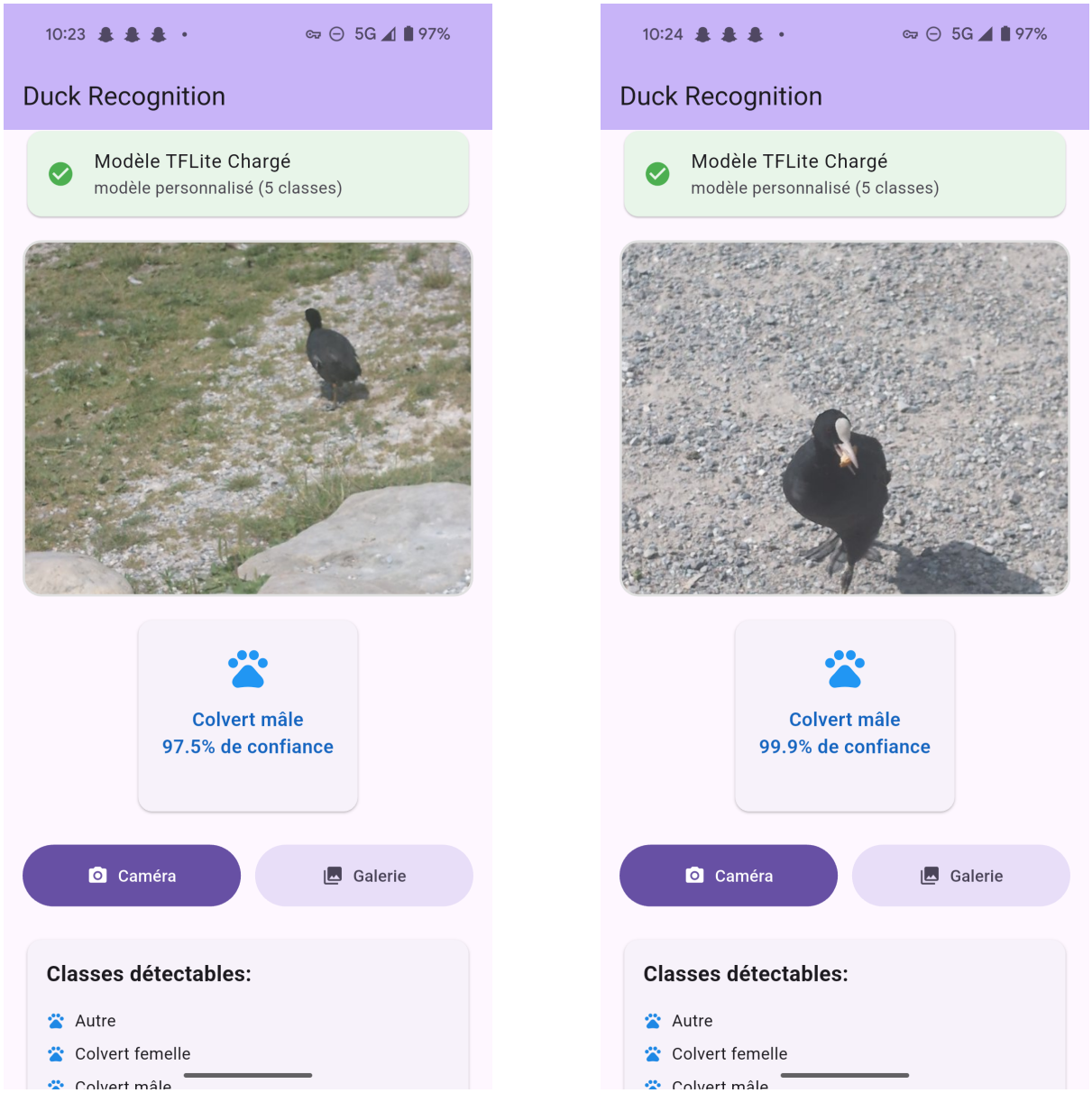
Successful Classifications:



(a) Successful classification example 1 (b) Successful classification example 2

Figure 12: Examples of successful real-world classifications

Failed Classifications:



(a) Failed classification example 1

(b) Failed classification example 2

Figure 13: Examples of failed real-world classifications

Analysis of Background Bias:

Since we don't have sufficient diversity of duck species across various backgrounds in our training data (i.e., the same species appearing in different environments), the model appears to have learned that "a duck on a rocky background is a *Colvert Mâle*." This hypothesis is supported by the Grad-CAM analysis of *Foulque Macroule* misclassifications, where background features inappropriately influence the classification decision.

8 Conclusions

8.1 Limitations and Areas for Improvement

- **Image positioning constraints**: We avoided placing ducks in different positions within the image since this would have enabled data augmentation through zooming. Since we have taken pictures near the border of the picture, we couldn't apply zooming-based data augmentation (which could have provided more training images, given our limited dataset). A recommendation would be to notify users to place the duck in the center of the picture.
- **Limited species coverage**: Lack of sufficient images for additional species, especially for *Cygne* and *Harle Bièvre*. However, this limitation helped with the analysis of the model since it allowed us to evaluate whether it can truly distinguish between species or not. If a new duck species is tested, it should ideally be classified as *Autre* (Other) since the model is not trained to recognize it. However, our testing shows this doesn't work reliably.
- **Multiple duck classification**: The model's performance on images containing multiple ducks remains an area requiring further development.
- **Species diversity**: Need for more duck species in the classification system.
- **Dataset diversity limitations**: Lack of diversity in the dataset - all images were taken in similar locations and conditions. This significantly affects the model's ability to generalize to new images taken in different conditions, as confirmed by our real-world testing results.

8.2 Future Work

The background bias issue identified through Grad-CAM analysis represents a critical finding that should guide future data collection efforts. Ensuring species representation across diverse environmental contexts would significantly improve model robustness and real-world performance.

Additional recommendations for future work include:

- Expanding the dataset with more diverse backgrounds and lighting conditions
- Implementing active learning techniques to identify and collect hard examples
- Exploring more sophisticated data augmentation strategies
- Investigating ensemble methods to improve classification reliability
- Developing better strategies for handling the "Other" class in open-set recognition scenarios

8.3 Summary

This project successfully demonstrates the application of transfer learning with MobileNetV2 for duck species classification. The systematic hyperparameter optimization approach led to significant performance improvements, achieving an F1-macro score of 0.8541. However, the Grad-CAM analysis revealed important limitations related to background bias and dataset diversity that significantly impact real-world performance.

The work provides valuable insights into the challenges of creating robust computer vision models for wildlife classification, particularly highlighting the importance of dataset diversity and the potential pitfalls of learning spurious correlations between species and environmental contexts.