Multi-label Image Classification on CelebA Dataset: A comparative study on CNNs, ResNet and Transformers

Giulia Bettuzzi

Università degli Studi di Verona

Master's Degree in Artificial Intelligence [LM-18]

Deep Learning Project

September 2025

Contents

Chapter 1

Motivation and Rationale

One of the central problems in Deep Learning is classification. Multi-label classification has become an interesting challenge in Deep Learning. The difference with multi-class classification is that a single sample can belong to more than one class.

An example of this problem is the recognition of facial attribute patterns in faces' images an it can be done with CelebA Dataset.

The Dataset is used as a benchmark in several studies to solve this problem. It contains over 200,000 images of celebrities labeled with 40 binary attributes. These attributes range from facial features to more complex traits.

This report aims to explore and compare multiple Deep Learning models applied to the prediction of facial attributes in the CelebA Dataset.

The main objective is to analyze the effectiveness of different classification approaches, using several models with different architectures and hyperparameter configurations.

The models included in the analysis are: Convolutional Neural Network, Residual Network, and Transformers. The differences between them are evaluated in terms of performance and scalability metrics.

Chapter 2

State of Art

In this report Deep Learning models (such as CNN, ResNet, ViT) are used to classify CelebA Dataset facial attributes, distinguishing hair color characteristics: brown, black, and blonde.
The introduction of advanced techniques such as Batch Normalization in Convolutional Neural Networks has improved stability and performance, while the adoption of Vision Transformers has offered new opportunities for image analysis.
These approaches are compared based on standard metrics: accuracy, precision, recall and F1-Score.
Advantages and limitations are highlighted based on hyperparameters and computational complexity.

Chapter 3

Objectives

The aim is to investigate CelebA Dataset and train Deep Learning models. By comparing the results obtained, conclusions can be drawn. The performance and precision of each are evaluated in relation to the others. The path is as follows:

1. Analyse CelebA Dataset
2. Train different Deep Learning models
3. Compare the models
4. Draw conclusions.

Chapter 4

Methodology

To get a complete overview of the problem and of the techniques, the dataset, the libraries and the methods used during the training and testing phase are analyzed.

4.1 Dataset

CelebFaces Attributes Dataset (CelebA) is a large-scale face attributes dataset with 202,599 celebrity images.
The dataset includes, in addition to the images:

- 10,177 identities
- 5 landmark locations
- 40 binary attributes annotations per image.

The images within the Dataset cover a range of locations and feature complex backgrounds, allowing for the diversity, quantity, and richness of the annotations to be highlighted.
CelebA Dataset:

→ Training samples 162,770 images

→ Validation samples 19,867 images

→ Test samples 19,962 images.

Three attributes, relating to hair color, are selected:

- Black_Hair: black hair
- Blond_Hair: blond hair
- Brown_Hair: brown hair.

The choice of attributes was based on the following factors: firstly, because it allows dealing with a Multi-label Classification problem, as a person can have mixed characteristics.
Secondly, the selected characteristics are visually distinguishable and, finally, it is possible to evaluate the model's ability to recognize specific details.

4.2 Data Pre-processing

In order to optimize the model, data augmentation is applied to the train images.
To ensure stable and efficient training, after pre-processing the pixel values of the images are scaled between [0, 1], then normalized. During the process data are flattened (ViT model is fed with a 1D input).

4.3 Analytical Tools

In order to decide if the model is performing well, the listed metrics are evaluated:

- Accuracy
- Precision
- Recall
- F1-Score
- Confusion matrix.

While accuracy measures the goodness of the classifier, precision, recall and F1-Score rely on the ability of the model to classify correctly the positive samples.

Indeed, accuracy deals with correctness of prediction, precision and recall with minimizing false positives and negatives respectively, and F1-Score balances precision and recall.
The confusion matrix is used to valuate the performance of the classification models. The matrix shows the number of times a class is correctly or not predicted by the model.

4.4 Computational Tools

The classifier is built using Google Colab. A cloud-based free software that allows the user to write Python code, test and share it.
Slow performance in large datasets is highlighted.
The main libraries exploited are:

- pandas and numpy: for performing matheatical operations on data
- torchvision [2]: from which the dataset used for the analysis is retrieved
- matplotlib: for the creation of graphs
- seaborn: for statistical distributions
- skikit-learn: open-source library for Machine and Deep Learning models.

4.5 Classification Model

As seen above, the report deals with a Multi-attribute Classification problem: Multi-label Classification rather than a Multi-class Classification problem, as each attribute represents an independent binary class and a subject may have mixed or ambiguous hair color characteristics.
Using the CelebA dataset, focusing on the recognition of three attributes related to hair color: Black_Hair, Blond_Hair, and Brown_Hair, the difficulties in classifying attributes due to the following factors emerge:

- Illumination variability: color perception is influenced by lighting conditions
- Color overlap: distinction between black and dark brown, or brown and dark blonde
- Multiple attributes

- Angular variability: different viewpoints and poses of the subject.

The classification task implemented can be formally defined as:

- **Input**: image $I \in \mathbb{R}^{H \times W \times D}$ where $H = 224$, $W = 224$, $D = 3$
- **Output**: prediction vector $y \in \{0, 1\}^3$

For the analysis, six model variants were developed and compared, each of them representing a family of neural architectures:

- Simple Convolutional Neural Network (CNN)
- Deep Convolutional Neural Network (CNN) with Batch Normalization
- Residual Network (ResNet)
- Vision Transformer (ViT).

4.5.1 Simple Convolutional Neural Network

Simple CNNs are fundamental Deep Learning architectures for computer vision tasks that can be used to solve Multi-label Classification problems through hierarchical feature extraction.
By tuning the **learning rate**, **dropout rate**, and **weight decay** parameters, the balance between model capacity and generalization can be optimally managed to prevent underfitting and overfitting.
The configuration tested are:

- **Simple CNN**
  - learning rate: 0.001
  - dropout rate: 0.5
  - weight decay: 0.0001
- **Simple CNN**
  - learning rate: 0.0001
  - dropout rate: 0.3
  - weight decay: 0.00001

4.5.2 Deep Convolutional Neural Network

Deep CNNs are enhanced convolutional architectures that incorporate batch normalization and deeper layer structures for representing complex features in classification tasks.
By optimizing the **learning rate** for stable convergence, **dropout rate** for regularization, and **weight decay** for parameter penalty, the model achieves better generalization while managing the increased complexity of deeper networks.
The configuration tested is:

- learning rate: 0.001
- dropout rate: 0.5
- weight decay: 0.0001

4.5.3 Residual Network

ResNets are residual learning frameworks designed for deep neural networks that solve the vanishing gradient problem through skip connections in classification tasks.
By configuring the **learning rate** for gradient optimization, **dropout rate** for overfitting prevention, and **weight decay** for regularization strength, the residual connections enable training of significantly deeper networks while maintaining gradient flow.
The configuration tested are:

- **ResNet**
  - learning rate: 0.001
  - dropout rate: 0.5
  - weight decay: 0.0001
- **ResNet**
  - learning rate: 0.0001
  - dropout rate: 0.3
  - weight decay: 0.00001

4.5.4 Vision Transformer

Vision Transformers are attention-based architectures that adapt the transformer mechanism from NLP to computer vision tasks, treating images as sequences of patches for multi-label classification. By setting the **learning rate** for stable attention weight updates, **dropout rate** for preventing overfitting in attention mechanisms, and **weight decay** for parameter regularization, the model captures global dependencies without convolutional inductive biases.

The configuration tested is:

- learning rate: 0.0001
- dropout rate: 0.1
- weight decay: 0.0001

Chapter 5

Results

5.1 Confusion Matrix

The confusion matrix helps evaluate the performance of the classification by comparing predicted values with actual values. It allows to count how often the algorithm predictions are correct or wrong.

The Simple CNN model [5.1] predicts blonde only when certain. This attribute can be considered the easier to classify and it is also the simplest to distinguish visually.
It tends to classify too many hairs as brown, probably due to the chromatic overlap (between black

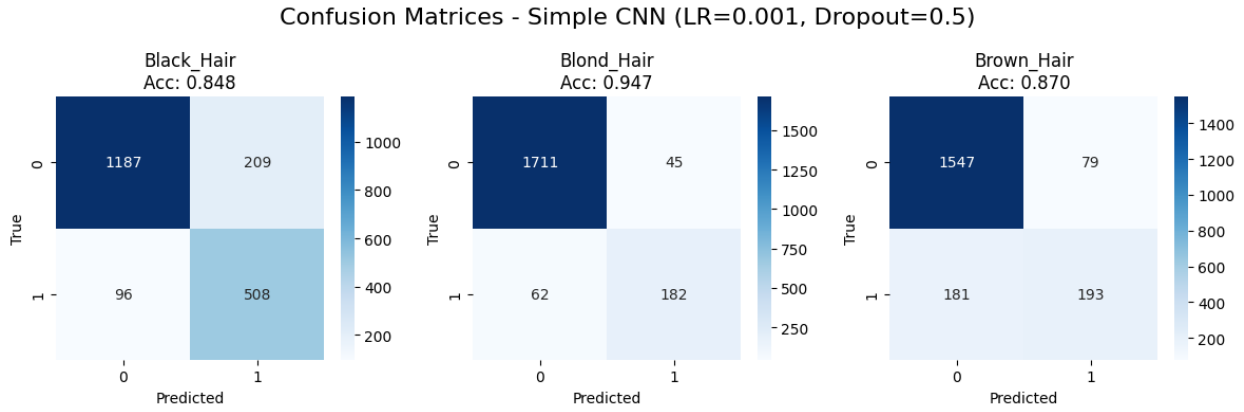

Figure 5.1: Simple CNN

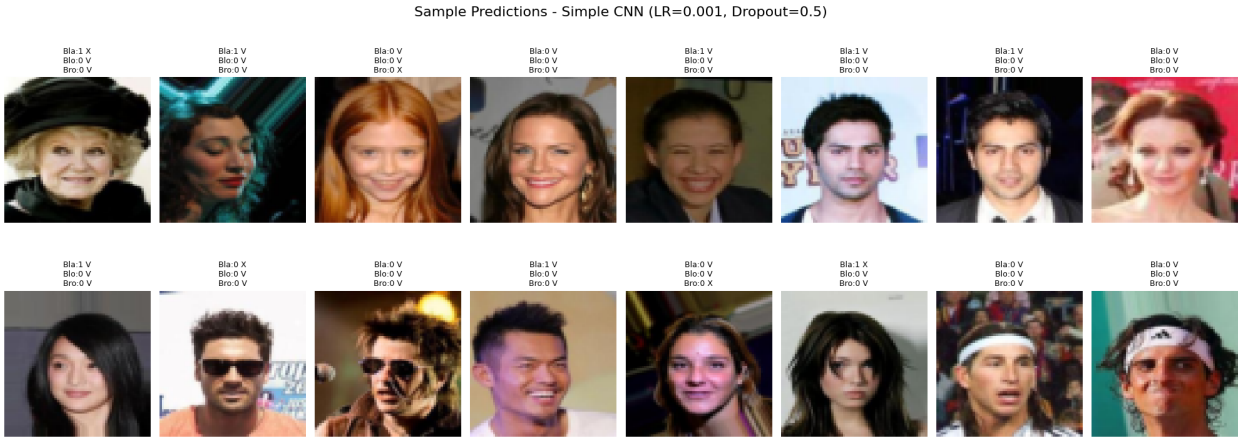

Figure 5.2: Prediction vs Ground Truth Simple CNN

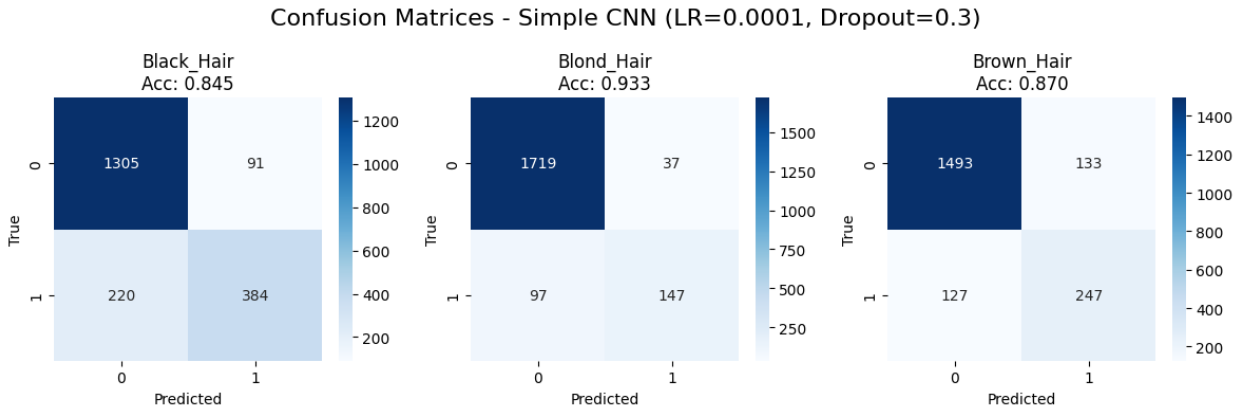

Figure 5.3: Simple CNN

and dark brown and brown and dark blonde), the greater variability of browns, and the possibility of multicolored hair.

For black, there is good separability with an intermediate balance between precision and recall (balanced errors - false positives and false negatives).

It can be seen that when comparing the two Simple CNN models, there is almost the same performance on brown hair. In the second case [5.3] all metrics get poorer and the model has not achieved optimal convergence.

In this case, it predicts black only when certain, there is an overestimate of black hair, as can be seen in the classification of black hair itself, where there are many false positives but few false negatives: the model predicts the positive class more aggressively.

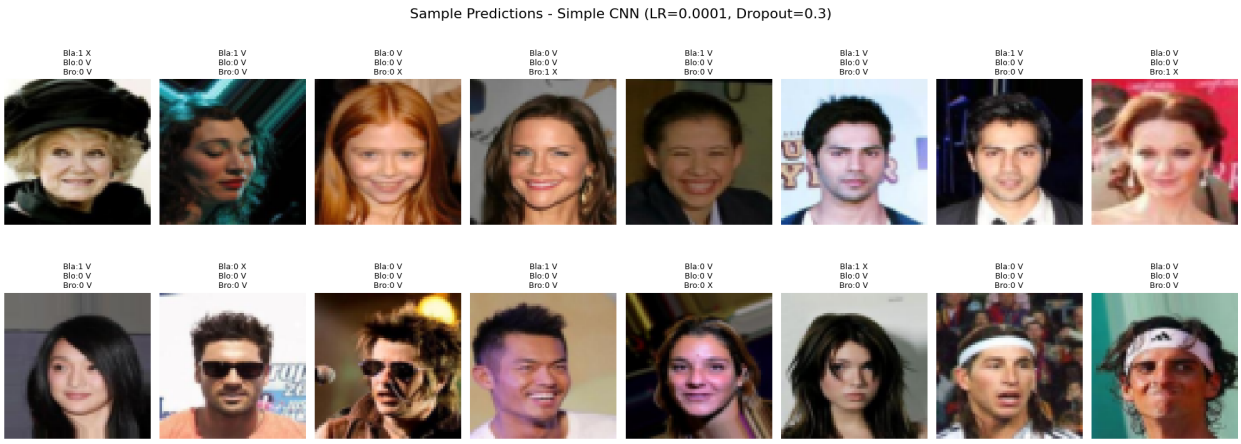

Figure 5.4: Prediction vs Ground Truth Simple CNN

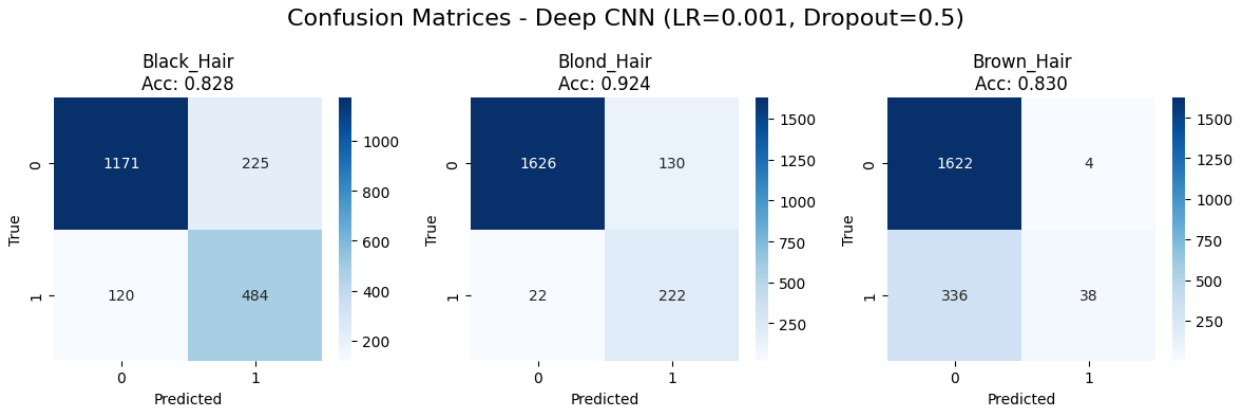

Figure 5.5: Deep CNN

It has lost accuracy in predicting blonde and it results that the first model is better in predicting blonde.

The problem of separability for brown is still present.

This demonstrates that in Deep Learning, undertraining can lead to outcomes worse than overfitting, especially in multi-label tasks where each attribute requires independent convergence.

It is important to observe that in Deep CNN [5.5] there is a decay in performance: more layers do not imply better performance.

The model presents a very particular pattern for brown hair, with an unusually low number of false positives, but a high number of false negatives. This indicates great prudence in predicting brown hair.

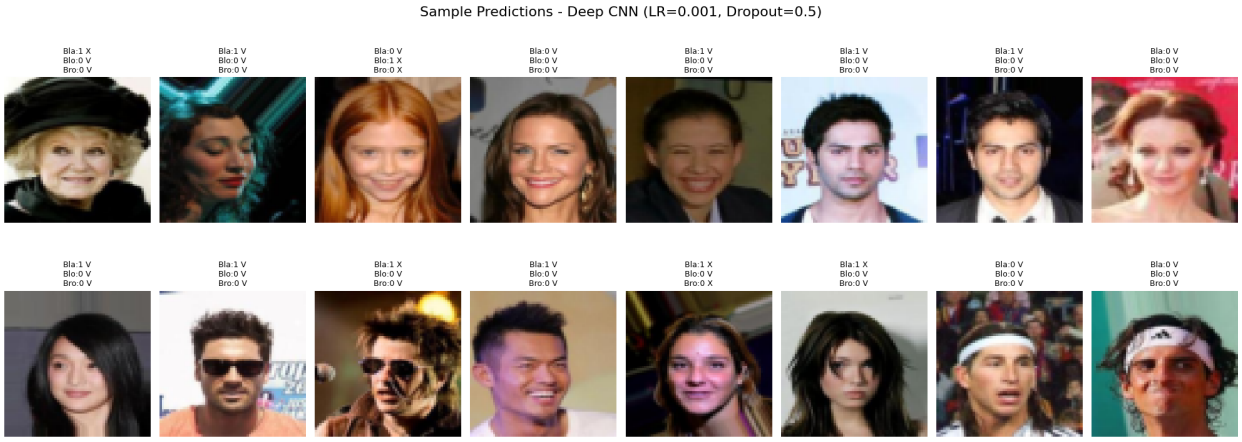

Figure 5.6: Prediction vs Ground Truth Deep CNN

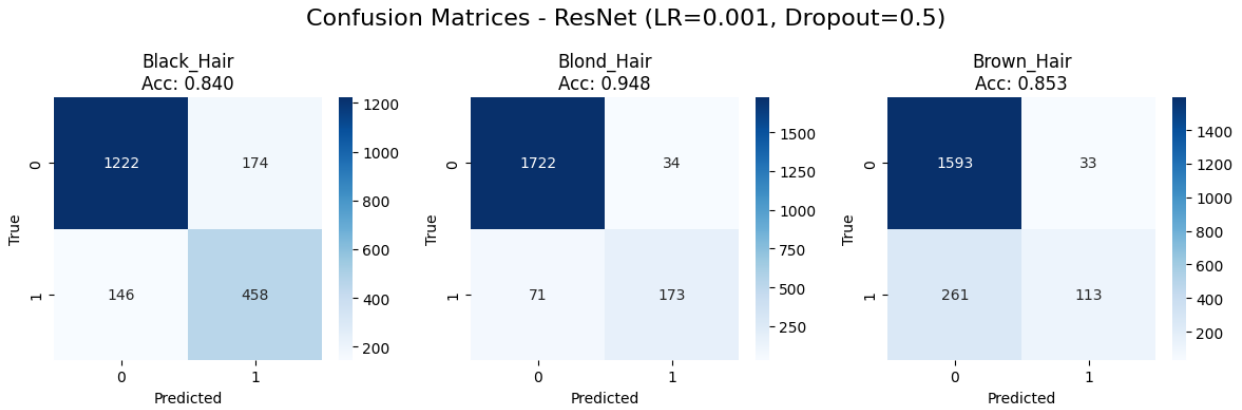

Figure 5.7: ResNet

The complexity of Deep CNN is not necessary for the task and there is a possible overfitting. Simple CNN performs better in all aspects for this task.

For better result in Deep CNN, it is possible to try to reduce the learning rate and increasing the dropout rate, and finally consider early stopping.

Comparing the two ResNet models, the first one [5.7] is better than the second [5.9] for all the classes: higher learning rate and dropout are more efficient. The second model has a more aggressive behavior in positive class prediction indeed there are more false positives and less false negatives for both black and brown hair.

The blonde benefits from the conservative (high precision, low recall) configuration, bringing im-

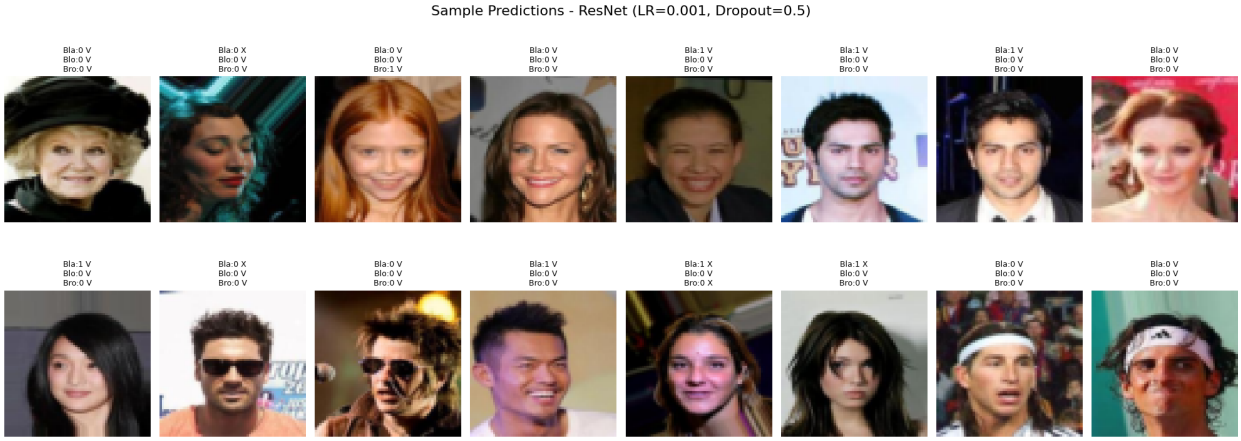

Figure 5.8: Prediction vs Ground Truth ResNet

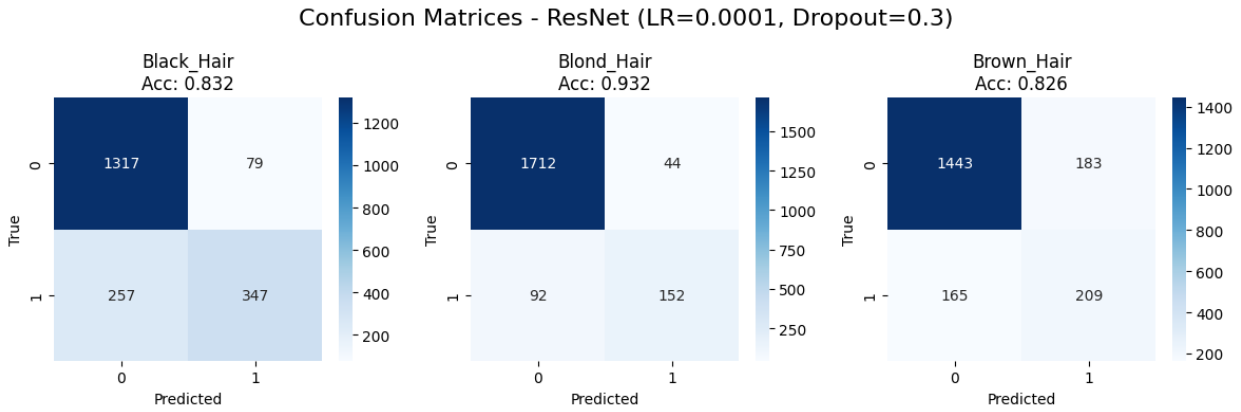

Figure 5.9: ResNet

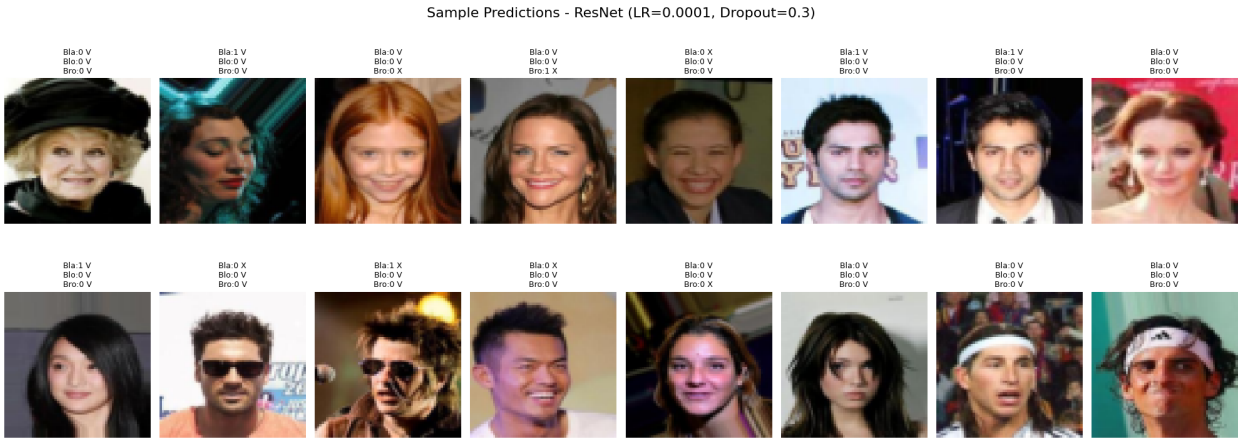

Figure 5.10: Prediction vs Ground Truth ResNet

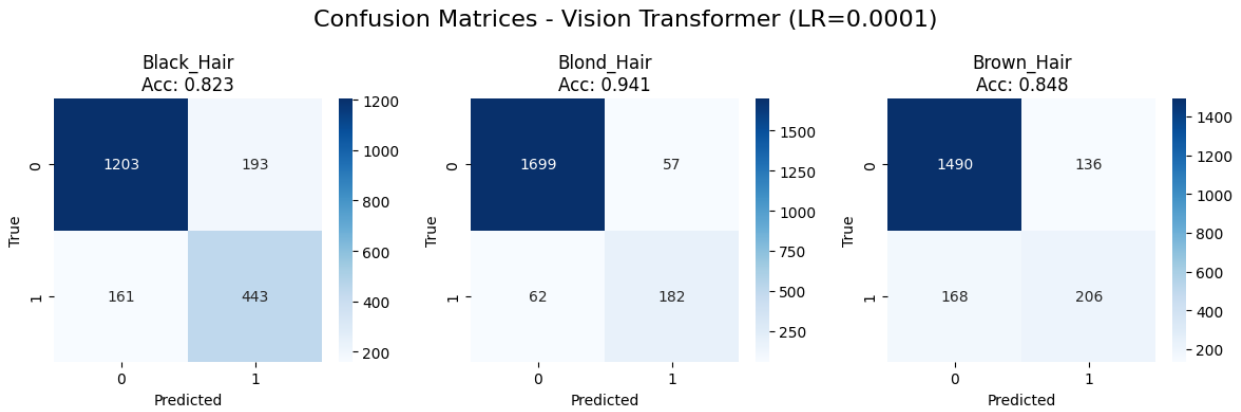

Figure 5.11: ViT

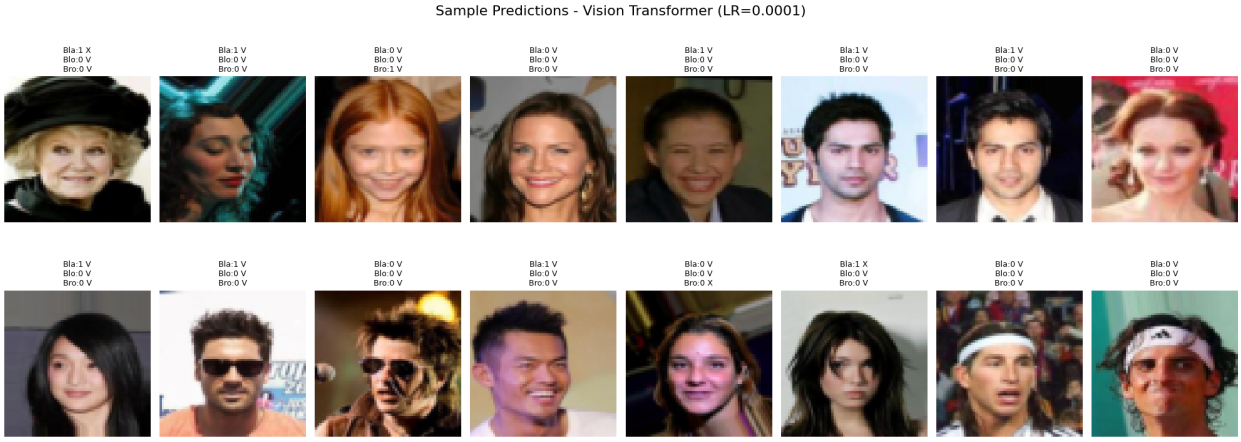

Figure 5.12: Prediction vs Ground Truth ViT

provements in both recall and F1-Score, and brown remains a difficult attribute to identify.

The Vision Transformer [5.11] represents a case of architectural oversizing: it uses a sophisticated attention mechanism for the task of hair color recognition, resulting in the worst computational efficiency among all the models tested and failing to fully exploit its attention capabilities due to insufficient training data.

5.2 Comparisons

The models' performance :

Model	Attribute	Accuracy	Precision	Recall	F1-Score
Simple CNN (LR=0.001, Dropout=0.5)	Black_Hair	0.8475	0.7085	0.8411	0.7691
Simple CNN (LR=0.001, Dropout=0.5)	Blond_Hair	**0.9465**	0.8018	0.7459	**0.7728**
Simple CNN (LR=0.001, Dropout=0.5)	Brown_Hair	0.8700	0.7096	0.5160	0.5975
Simple CNN (LR=0.0001, Dropout=0.3)	Black_Hair	0.8445	0.8084	0.6358	0.7118
Simple CNN (LR=0.0001, Dropout=0.3)	Blond_Hair	0.9330	0.7989	0.6025	0.6869
Simple CNN (LR=0.0001, Dropout=0.3)	Brown_Hair	0.8700	0.6500	0.6604	0.6552
Deep CNN (LR=0.001, Dropout=0.5)	Black_Hair	0.8275	0.6827	0.8013	0.7372
Deep CNN (LR=0.001, Dropout=0.5)	Blond_Hair	0.9240	0.6307	**0.9098**	0.7450
Deep CNN (LR=0.001, Dropout=0.5)	Brown_Hair	0.8300	**0.9048**	0.1016	0.1827
ResNet (LR=0.001, Dropout=0.5)	Black_Hair	0.8400	0.7247	0.7583	0.7411
ResNet (LR=0.001, Dropout=0.5)	Blond_Hair	0.9447	0.8357	0.7090	0.7672
ResNet (LR=0.001, Dropout=0.5)	Brown_Hair	0.8530	0.7740	0.3021	0.4346
ResNet (LR=0.0001, Dropout=0.3)	Black_Hair	0.8320	0.8146	0.5745	0.6738
ResNet (LR=0.0001, Dropout=0.3)	Blond_Hair	0.9320	0.7755	0.6230	0.6909
ResNet (LR=0.0001, Dropout=0.3)	Brown_Hair	0.8260	0.5332	0.5588	0.5457
Vision Transformer (LR=0.0001)	Black_Hair	0.8230	0.6965	0.7334	0.7145
Vision Transformer (LR=0.0001)	Blond_Hair	0.9405	0.7615	0.7459	0.7036
Vision Transformer (LR=0.0001)	Brown_Hair	0.8480	0.6023	0.5508	0.5754

The average models' performance:

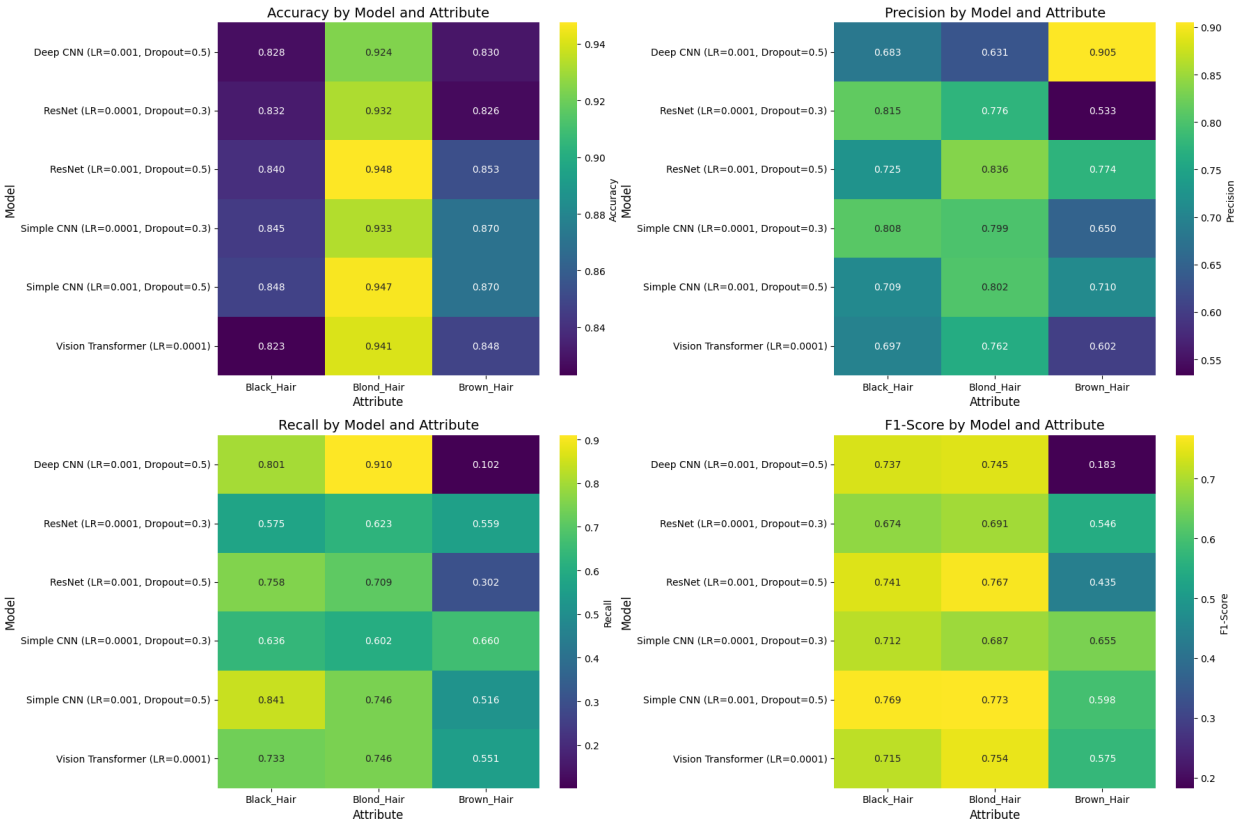

Figure 5.13: Comparative graphs

Model	Accuracy	Precision	Recall	F1-Score
Deep CNN (LR=0.001, Dropout=0.5)	0.8605	0.7394	0.6043	0.5550
ResNet (LR=0.0001, Dropout=0.3)	0.8633	0.7077	0.5854	0.6368
ResNet (LR=0.001, Dropout=0.5)	0.8802	**0.7781**	0.5898	0.6476
Simple CNN (LR=0.0001, Dropout=0.3)	0.8825	0.7524	0.6329	0.6846
Simple CNN (LR=0.001, Dropout=0.5)	**0.8880**	0.7399	**0.7010**	**0.7132**
Vision Transformer (LR=0.0001	0.8705	0.6868	0.6767	0.6812

By analyzing the performance matrices [5.13] for the classification of hair attributes, the following patterns emerge:

- Blonde hair is the easiest attribute to classify with accuracy, precision, and recall. All models perform similarly for this attribute.

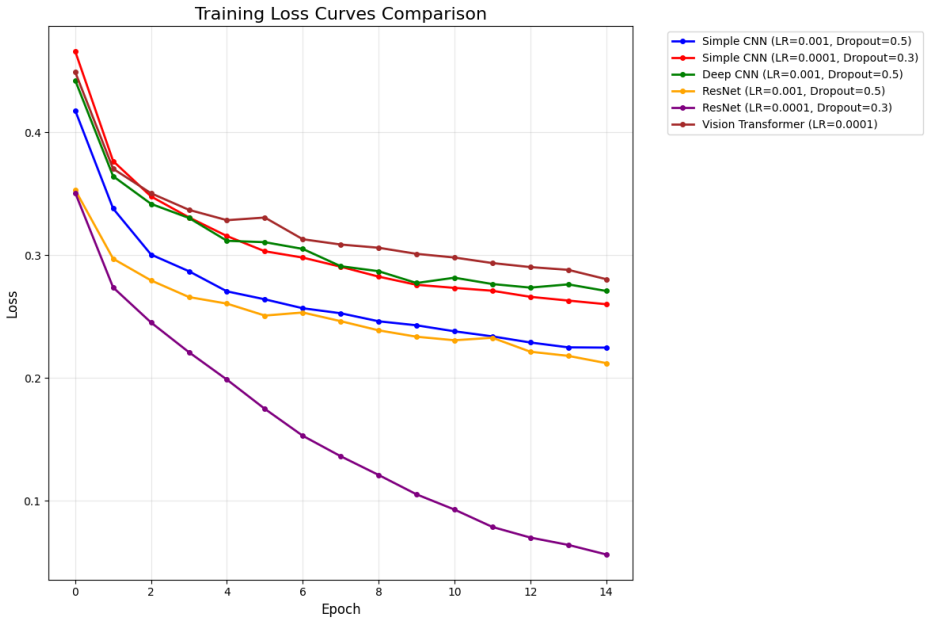

Figure 5.14: Training Loss curves comparison

- Black hair has moderate precision and low recall
- Brown hair has problematic recall, highlighting the difficulties in identifying this attribute.

In the training loss curve graph [5.14] Res Net (Dropout = 0.3) has a interesting behavior because it presents a rapid convergence, it has a descending curve. There is no evidence of overfitting or instability and it suggests an efficient learning.

ResNet (Dropout = 0.5) present a constant reduction but more gradual than ResNet (Dropout = 0.3). The first has final loss and a stable behavior without significant fluctuations.

On the other hand, Simple CNNs show some criticism like: due to the fact that it remains constant for many consecutive periods without significant improvements they have slower convergence and higher final loss. More strong learning rates could improve the models convergence.

Deep CNN highlight possible problems: it has a good initial convergence but at mid time it is almost constant and show a relatively high final loss. It suggests possible gradient flow problems or the need for different regularization parameters.

ViT results the worst performer also in this case with the highest final loss and slower convergence.

The performance discrepancy between training loss and final metrics is notable:

- ViT has the lowest loss but moderate performance on metrics

- Simple CNN has higher loss but competitive performance.

Chapter 6

Conclusions

Different approaches and hyperparameters can be used in order to explore the most of scenarios.
The best results for the image classification problem using CelebA Dataset is obtained with Simple CNN. The model offers the best performance - complexity ratio, achieving results that are, sometimes, better than the more sophisticated ones.
Blonde_Hair is the easiest attribute while Brown_Hair is the most problematic.
Even if ViT shows a fast convergence in the training phase, there is a big gap in real optimization and performance, maybe due to a possible overfitting.
To improve the performance on the more difficult attributes it is necessary to explore different strategies and specific loss function.

Bibliography

[1] https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

[2] https://docs.pytorch.org/vision/stable/index.html