Universitat Pompeu Fabra

FACIAL AGE CLASSIFICATION WITH CNN

Deep Learning

Astrid Alins Moragón, Miquel Bisbe Armengol i
Emma Climent

2025

Content:

1. INTRODUCTION

Facial image analysis plays a critical role in modern computer vision, enabling machines to extract high-level information from human faces. Among the many facial attributes that can be analyzed, age estimation stands out as a particularly challenging yet impactful task, with applications in security, health monitoring, and personalized services.

This project addresses the task of predicting a person's age from facial images, formulating it as a **classification problem** with 100 discrete age classes (0–99 years). Since age involves many fine distinctions and complex visual data, it can be hard to predict correctly, even for human criteria. Developing accurate models requires not only large and balanced datasets but also architectures capable of extracting both general and age-specific features from facial imagery.

To tackle this, we propose a **transfer learning approach** in which a convolutional neural network (CNN), originally trained for facial emotion recognition (FER) (a 7-class classification problem) is adapted and fine-tuned for age classification with 100 classes. Our main objectives are:

- To repurpose a pre-trained CNN by modifying and optimizing the model
- To evaluate its performance using valid metrics methods
- To analyze the impact of class imbalance and data augmentation strategies on model generalization.

This work aims to demonstrate how knowledge learned from one facial analysis task can be effectively transferred to another, highlighting the practical advantages of transfer learning in domains with limited labeled data.

For this purpose, we use the UTKFace dataset, which provides over 23,000 facial images labeled with age, gender, and ethnicity. We focus solely on the age attribute. The base CNN model we adapted was originally developed for emotion classification and shared publicly on Kaggle by user @mohamedchahed. We used his implementation as the foundation of our work, modifying its architecture and retraining it to perform multi-class age classification through transfer learning techniques.

2. STATE OF THE ART

Facial Emotion Recognition (FER) has been extensively studied, with CNN-based models achieving over 75% accuracy on benchmarks like FER2013 [1]. Given the shared visual structure across facial tasks, transfer learning has become a standard strategy. CNNs trained for FER can be adapted to age prediction, leveraging the general features already learned. As shown by Yosinski et al.], early convolutional layers tend to be transferable across visual domains, while deeper layers require fine-tuning for specific targets. This principle guides our project, where a FER-trained CNN is repurposed for age estimation.

Unlike FER's categorical nature, age prediction is more complex: it involves fine-grained and continuous changes over a wide range, making it data-hungry and sensitive to class imbalance. To tackle this, the literature presents three main paradigms:

- **Classification-based** models treat each age as a separate class. Niu et al. (2016) used Label Distribution Learning (LDL) to soften hard labels and improved performance on Morph II [5]. However, these methods ignore the ordinal relation between classes.

- **Regression-based** approaches predict continuous age values using losses like Mean Squared Error. Rothe et al. (2015) proposed the DEX model using a VGG-16 backbone, achieving competitive results by framing age prediction as a regression problem [6]. These models offer precision but can be unstable with noisy data.
- **Ordinal regression** methods aim to capture the ordered nature of age. CORAL, for example, models age as a sequence of binary decisions and has shown superior performance, especially under class imbalance and in-the-wild settings [5].

Despite significant progress, challenges remain—particularly under real-world variability (lighting, occlusion, demographics) and uneven age distributions in datasets like UTKFace. These issues call for models that are not only accurate but also robust and adaptable to such inconsistencies.

3. METHODOLOGY

3.1 DATA ANALYSIS

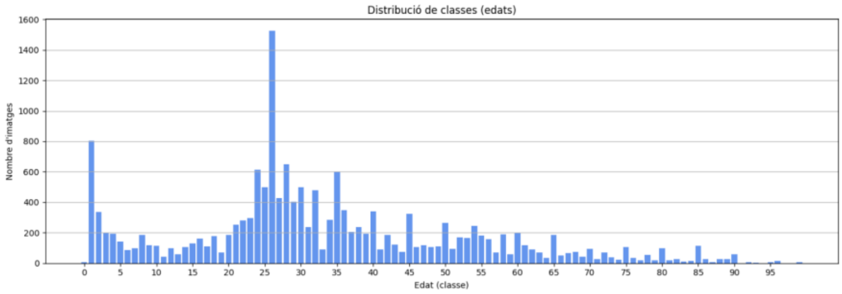
This study builds upon a CNN originally trained on the **FER2013 dataset**, a widely used benchmark in FER. It contains 35,887 grayscale face images (48×48 pixels), each labeled with one of seven emotion categories. The dataset's low resolution, varied expressions, and crowdsourced labels introduce noise and subjectivity, making automated recognition particularly challenging.

However, since our goal is age prediction, we transitioned to the **UTKFace** dataset, which includes explicit age annotations and is therefore better suited to our task. UTKFace contains 23,708 face images annotated with age, gender, and ethnicity. In our case, we focused exclusively on the age attribute. As an initial preprocessing step, we **removed all samples with age > 100**, since these classes were sparsely populated and introduced noise without significant benefit to model generalization.

To ensure compatibility with the model architecture initially designed for the FER2013 dataset, all images from UTKFace were resized to **48×48 pixels** and converted to **grayscale**, aligning them with the input format expected by the network. Additionally, we applied **one-hot encoding** to the age labels. Finally, the dataset was **split into training (70%)**, **validation (15%)**, and **test (15%)** subsets. This division allowed us to fine-tune hyperparameters, monitor for overfitting, and evaluate model generalization in a robust and consistent manner.

To understand the distribution of classes, we plotted the number of samples per age class (Figure 1). This visualization revealed a clear imbalance: some age groups had a disproportionately high number of samples, while others were severely underrepresented. For example, the most frequent age classes were age 26 with 2,197 images and age 1 with 1,123 images. In contrast, some classes had extremely limited representation, such as age 99 with only 9 images and age 93 with 5 images. Such imbalance poses a risk of biasing the model toward the most frequent classes and degrading its ability to generalize.

Figure 1: Initial UTKFace Distribution



Although we initially explored the dataset and noted the age distribution (see our intuitions in the Data Analysis section), it was not until after our first model execution that we fully understood the extent of the imbalance. The confusion matrix from this initial run (Figure A1) revealed that the model was predicting almost every sample as age 26, confirming a severe skew in the dataset. This prompted us to take corrective action by applying **undersampling**, specifically reducing by 500 samples the number of samples for age 26.

In addition to undersampling, we applied **data augmentation** techniques such as random horizontal flips, rotations, and slight zoom variations during training.

These strategies are expected to improve the model's learning and reduce errors caused by uneven class distribution. However, the actual effectiveness of these methods will be analyzed and discussed in detail in the **Experiments and Discussion** section, where we will evaluate their impact on model performance.

The boxplot (Figure A2) reveals a median age of X years with a heavily right-skewed distribution concentrated in younger demographics. The interquartile range spans at X years, with 75% of samples being under 40 years old, while numerous outliers extend toward 100 years representing a much smaller older population.

3.2 MODEL AND OPTIMIZATION

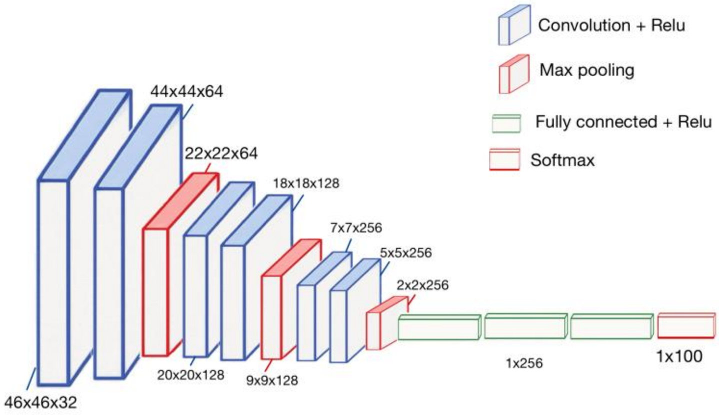


Figure 2: Resulting optimized model architecture

The final model architecture analyzed in this study is the optimized version of our CNN, designed for age classification. Following Zeiler and Fergus (2014) [7], we froze the earlier layers responsible for extracting low-level, general features. Specifically, the entire first convolutional block and the first two layers of the second block were kept frozen, preserving their pretrained weights to maintain stable feature extraction and reduce overfitting.

The remaining layers, including the deeper convolutional blocks and the fully connected (Dense) layers, were fine-tuned. This approach enabled the model to refine task-specific features for age prediction while retaining general visual representations from the frozen layers.

The CNN consists of three convolutional blocks followed by a classifier with three FC layers. Each convolutional block contains 2 Conv2D layers with ReLU activation, followed by BatchNormalization, MaxPooling2D, and Dropout. The convolutional kernels are fixed at 3×3, which ensures local spatial feature extraction. Spatial dimensions reduce progressively from 46×46 in the first block to 2×2 in the last, while filter depth increases from 32 to 256. The classifier head includes three Dense layers with 512, 512, and 256 units respectively, each followed by BatchNormalization and Dropout with a 0.5 rate. The final layer outputs 100 logits, passed through a **softmax** function to convert raw scores into class probabilities, enabling the model to estimate prediction confidence for the multi-class task (Goodfellow, Bengio, & Courville, 2016)[8]. The model has around 1.8 million parameters, where the first dense layer accounts for more than 500,000 parameters, enhancing the model's ability to learn detailed age-related features. Approximately 93,000 parameters in early convolutional blocks remain frozen to preserve pretrained features and reduce overfitting.

We used categorical **cross-entropy** as the loss function, which is standard for multi-class classification problems. Initially, for training with the original pretrained weights from Mohammed's model, we used a learning rate of 0.0001. This value was chosen because it is not too small, allowing the model to learn at a reasonable speed without unnecessarily slowing down training. When fine-tuning deeper parts of the network where the learning became more specific, we lowered the learning rate to 1e-5 to ensure stable convergence. These learning rate values have proven to work best with our model and dataset.

4. EXPERIMENTS

The experimental phase was designed to evaluate the impact of **undersampling**, **layer freezing, architectural modifications**, **evaluation metrics, age group classification and data augmentation** on model performance.

Since we did not have access to pre-trained weights for the original CNN, we **trained the baseline model for 50 epochs** to obtain initial weights (replicating the approach of Mohammed). His original FER-trained model reaches a validation accuracy of around **0.65**, which we also observed during our replication. These weights were then **imported** and used as the starting point for further fine-tuning.

4.1 FREEZING

We initially performed an exploration of layer freezing strategies, progressively unfreezing layers from the most specific (deepest) to the most general (earliest convolutional blocks). Our initial attempt involved unfreezing only the last convolutional block. However, we quickly realized that this approach was insufficient (Figure A3). We therefore continued unfreezing layers until only the first convolutional block and the first two layers of the second block remained frozen.

In the end, we ended up fine-tuning most of the network but taking advantage of the original CNN structure (more extended in the Discussion section). During this phase, we noticed that the model's overall accuracy remained modest, and the confusion matrix lacked a strong diagonal structure (Figure A4), indicating difficulties in making consistent predictions.

4.2 MODIFYING ARCHITECTURE (ADDING 1 FC)

To improve the model's expressiveness, we experimented with adding FC layers to the classifier head. Introducing a single FC layer with 512 units led to a marked improvement in both accuracy and the structure of the confusion matrix, which showed a much clearer diagonal (Figure A5). Before this modification, the model achieved an initial accuracy of 0.24 on the training set and 0.27 on the validation set. After this addition, these values increased to 0.36 and 0.54, respectively, highlighting a substantial gain, particularly in generalization. This configuration yielded particularly promising results.

When modifying the model architecture, we initially considered adding more convolutional layers. However, given that the final feature maps were already reduced to 2×2 (after three MaxPooling layers), adding convolutionals would have required careful adjustments to padding and filter sizes.

We opted for a simpler and more effective solution: expanding the fully connected layers. This approach provided a better balance between model capacity, implementation effort, and interpretability, making it the most practical choice for enhancing performance in our age prediction task.

4.3 MODIFYING ARCHITECTURE (ADDING 2 FC)

We also tested adding one and two more FC layers, but both configurations resulted in overfitting (Figure A6). This was evident from the widening gap between training and validation metrics, as well as a deterioration in generalization performance.

Even though we implemented dropout with a rate of 0.5 to mitigate overfitting, it was not sufficient to compensate for the increased model complexity. With limited data, the model can still memorize training examples rather than learn generalizable patterns, leading to reduced validation performance. Consequently, we retained the version with a single added FC layer as the most balanced configuration.

4.4 METRICS AND TOLERANCE ACCURACY

During our analysis of different freezing strategies and classifier head configurations, we realized that traditional accuracy and confusion matrix alone were not sufficiently informative for the fine-grained task of age prediction. Small deviations, such as predicting age 24 instead of 25, were counted as fully incorrect, even though such errors are semantically minor. To better evaluate the model’s performance, we incorporated additional metrics, including precision, recall, F1-score and support. Additionally, due to class imbalance in the dataset, we computed all evaluation metrics using **weighted averages**, ensuring that underrepresented age classes were fairly accounted for.

We also computed the Mean Absolute Error (MAE), which provided a more nuanced view by measuring the average absolute difference between predicted and true ages. At different stages of model development, we observed MAE values between 2.09 and 2.65 years. Indicating that the model’s predictions typically deviated by less than 3 years from the actual age.

To further refine our evaluation, we introduced a custom **tolerance-based accuracy metric**, which considers a prediction correct if it falls within a range of ±*k* years from the ground truth. Using a tolerance of ±3 years, the model achieved a much more realistic and meaningful accuracy of **76.97%**, compared to the traditional accuracy of **57%**. Similarly, the weighted F1-score also improved. The significance of adopting tolerance-based metrics is further elaborated in the **Discussion** section detailed analysis of the model's performance under different evaluation criteria, both with and without tolerance, is presented in the **Results** section.

4.5 AGE RANGES

Given the strong performance observed when applying tolerance-based accuracy to the 100-class configuration, we were motivated to test whether training the entire model under this evaluation perspective could lead to further improvements. As part of this exploration, we considered simplifying the classification task by grouping individual ages into broader intervals. This was based on the hypothesis that reducing the number of output classes might help the model generalize better (passing from a 100-class to a 11-class).

To define meaningful age ranges, we reviewed best practices in the literature and drew inspiration from the DEX (Deep EXpectation of apparent age from a single image) paper by Rothe et al.[6], which used the LAP dataset to establish age brackets grounded in real-world perception of age (Figure A7).

However, results showed that this approach did not yield better performance. The confusion matrix for the age-range model was less diagonal (Figure A8), and the model exhibited difficulties in accurately distinguishing between adjacent ranges. These outcomes suggest that the original fine-grained 100-class setup, especially when combined with a tolerance-based evaluation, offered superior precision and generalization. (A detailed discussion is provided in the **Discussion** section.)

4.6 DATA AUGMENTATION

As our final experiment, we explored **data augmentation** in an effort to address the persistent issue of data imbalance, particularly the **very low number of samples in certain age classes**. To mitigate this, we applied augmentation only to the training set, using a combination of transformations intended to increase the variability and robustness of the input data.

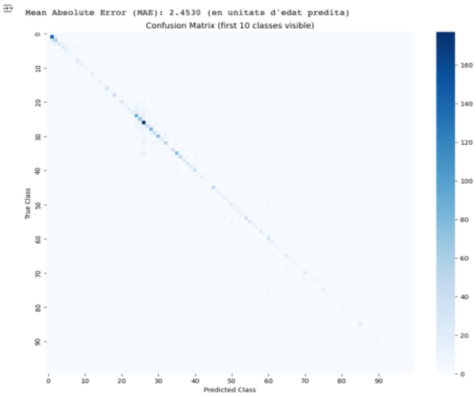
Specifically, we used **random horizontal flips**, **random rotations (±10 degrees)**, and **random affine transformations** with slight translations (up to 10% in each direction), in addition to resizing and converting images to grayscale. The results of this augmentation attempt are discussed further in the **Discussion section**, where we analyze why this strategy ultimately **did not yield the improvements we had hoped for**.

5. RESULTS

5.1 RESULTS 100 outputs

Performance Analysis

Without applying tolerance, the model's performance was moderate. The **weighted accuracy** reached **57%**, with a **weighted precision** of **0.59**, **weighted recall** of **0.55**, and a **weighted F1-score** of **0.56**, suggesting a decent performance that accounts for the class imbalance.



```
Mean Absolute Error (MAE): 2.4530 (en unitats d'edat predita)
```

Figure 3: Conf matrix (strong diagonal)

However, when we examined the **macro-average** metrics, where each class is treated equally regardless of its frequency, the results were significantly lower: **precision** was **0.56**, **recall dropped** to **0.45**, and **F1-score** was only **0.47**. These values reveal that the model struggled specially with underrepresented classes, retrieving less than half of the true labels correctly and failing to generalize across all classes uniformly.

Issues Detected: Due to the low support in some classes (with only 1, 2, or 4 samples), several labels (such as **0**, **13**, **64**, and **71**), obtained an **F1-score of 0.00**, meaning the model never predicted them correctly. This clearly exposes the challenge of extreme class imbalance in fine-grained classification settings.

Accuracy Tolerance Evaluation

To better reflect realistic prediction quality, we introduced a **tolerance-based evaluation** (explained in the experiments section): The results were as follows:

- **±1 year** → Accuracy: **64.38%**
- **±2 years** → Accuracy: **71.06%**
- **±3 years** → Accuracy: **76.97%**

While not always predicting the exact age, our model tends to make close estimations. The relatively high accuracy under tolerance, especially at ±2 or ±3 years, supports the model's ability to extract meaningful age-related features from the data.

Performance with ±3-Year Tolerance

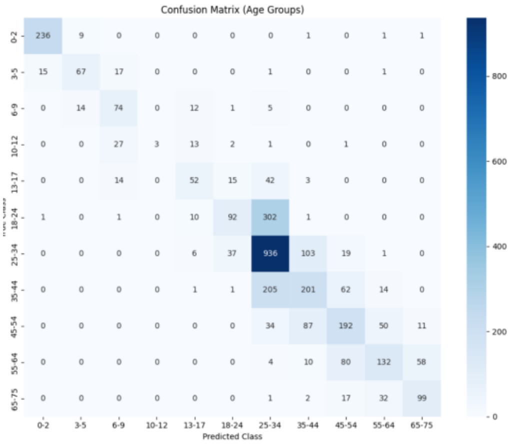
When evaluating with a ±3-year tolerance, the model's overall performance improved substantially. It reached a **tolerant accuracy of 76.97%**, with a **weighted precision** of **0.78**, **weighted recall** of **0.77**, and a **weighted F1-score** of **0.77**, demonstrating robust behavior even in the presence of class imbalance. The **MAE** also decreased to **2.09 years**, indicating that most predictions were very close to the true age.

While the overall results discussed above are promising, we would like to emphasize that a more granular analysis reveals important limitations. Specifically, we evaluated the **F1-score for each individual age class** and observed **very low values for many of them**. The detailed breakdown of F1-scores (Figure A9) and causes behind these inconsistencies is provided in the **Discussion section**.

5.2 RESULTS 11 outputs

Performance Analysis

The model achieves an **overall accuracy of 60.76%** on the validation set, which is **insufficient** for a multiclass classification task with 11 age categories. The **weighted f1-score** of **0.5843** confirms that the model performs **poorly**, with a limited ability to balance false positives and false negatives. The **macro-average f1-score** (0.5452) and **macro recall** (0.5376) further indicate that the model **struggles to perform consistently across all classes**. Although the **validation**



loss (1.05) is slightly lower than the training loss (1.16), suggesting no overfitting, it also implies that learning has stabilized.

Looking at performance by age group, there is a **clear variability**. The model performs best on the **0–2 years** group, with excellent metrics: **precision of 0.9365**, **recall of 0.9516**, and **f1-score of 0.9440**. This suggests highly accurate and reliable classification for this group. The **25–34 years** group also stands out, with a strong **f1-score of 0.7110** and **recall of 0.8494**, showing good detection capacity.

In contrast, performance drops significantly in other groups. The **10–12 years** category shows the **lowest performance**, with an **f1-score of just 0.1200** due to a very low **recall (0.0638)**, despite having a **precision of 1.0000**. This indicates that the model **almost never correctly identifies samples from this class**. Similarly, the **18–24** and **35–44 years** groups show **poor performance**, with f1-scores of **0.3315** and **0.4507**, respectively.

Other groups such as **55–64** and **65–75 years** achieve **moderate results**, with f1-scores close to **0.6**. Meanwhile, the **3–5**, **6–9**, and **45–54 years** groups show **acceptable performance**, with f1-scores ranging from **0.5 to 0.7**.

In summary, the model shows **very uneven performance depending on the age group**, with **strong results only in a few categories**, and **overall metrics that indicate a need for substantial improvement**. Distribution is shown in Figure A7.

6. DISCUSSION

6.1 Architecture and Transfer Learning Effectiveness

Our starting hypothesis was that a CNN pre-trained on FER could be effectively repurposed for facial age classification. Initially, we assumed only the final layers would require fine-tuning, however, our results showed that meaningful performance improvements only emerged after unfreezing most of the convolutional layers. This confirms that, unlike FER, prediction with 100 classes requires deeper, task-specific feature adaptation.

Although the model architecture remained largely intact, nearly all layers had to be retrained to capture the fine details of age-related features. **We ended up having to retrain 1.569.448 parameters, which represent approximately 94% of the existing ones.**

6.2 Model Depth and Fully Connected Layers

Increasing the depth of the model by adding a FC layer significantly improved the performance (3rd experiment). This change enhanced both accuracy and the diagonality of the confusion matrix, indicating that the model was better at capturing meaningful patterns. This improvement suggests that deeper classifiers help capture complex, non-linear patterns between facial features and age.

The overfitting likely resulted from the sharp increase in trainable parameters caused by adding extra dense layers**, adding complexity but no new spatial information**. This becomes especially problematic with a limited dataset, as the model tends to memorize rather than generalize. As

Goodfellow, Bengio, and Courville (2016) [8] explain, 'deeper networks can learn more complex patterns, but they also require stronger regularization to avoid overfitting'.

6.3 Class Imbalance: Undersampling and Data augmentation

To address class imbalance, we applied manual undersampling. However, even after removing approximately 500 samples, the dataset remained heavily biased toward age 26, limiting the impact of regularization. This suggests that a more aggressive undersampling strategy might have been necessary to better balance the data and prevent the model from overfitting to this dominant class.

In parallel, we experimented with various **data augmentation techniques**, aiming to synthetically increase variability within underrepresented classes. Unfortunately, **no significant performance gains** were observed. This outcome is expected, as **augmenting rare samples does not introduce fundamentally new information**, but rather small perturbations of already limited data. Such transformations are insufficient to compensate for the lack of real diversity in the training set.

This limitation is clearly reflected in **Figure A11**, which shows the **confusion matrix** (after applying data augmentation) Which lacks a strong diagonal. Moreover no age class stands out with consistently correct predictions, highlighting that **the model fails to confidently classify any specific age**, even after augmentation. This reinforces the conclusion that data augmentation alone is **not an effective solution** in the context of extreme class imbalance (5 samples in some classes) and limited sample diversity.

6.4 Evaluation Metrics and Tolerance Accuracy

Given the fine-grained nature of age classification, standard accuracy metrics often fail to reflect the model's true performance. Predicting 24 instead of 25 is a minor, semantically acceptable error, yet is treated as entirely incorrect. This rigid evaluation misrepresents model reliability and motivated our use of **tolerance-based accuracy**.To address the strong class imbalance, we used **weighted metrics** to ensure a fairer evaluation across the age range.

Although we achieved a reasonably good tolerance accuracy with our best model configuration, we focused on the F1 score for each individual class to better understand how well the model represents each age group (seen in Figure A9). We observed very high F1 scores for classes with abundant data, indicating strong performance where the model had sufficient examples to learn from. In contrast, classes with fewer samples showed much lower F1 scores, sometimes close to zero, reflecting poor predictive ability. This discrepancy occurs because the **model struggles to learn meaningful patterns for underrepresented ages** due to limited training data despite overall tolerance accuracy appearing acceptable.

6.5 Range-Based Classification (4.5 age ranges)

Empirical results showed a decline in both accuracy and F1-score compared to the fine-grained model. A key issue was semantic ambiguity near class boundaries, where ages such as 24 and 25, though visually similar, fall into different intervals. As a result, the model was penalized for predictions that were close in absolute terms but assigned to neighboring bins, reducing its ability to generalize effectively.

Moreover, grouping ages into broader ranges did not solve the class imbalance problem. For instance, the age groups 18–24 and 35–44 were often misclassified as 25–34, the range with the most samples. We believe this occurs because facial features in these neighboring ranges are quite similar, making it difficult for the model to distinguish them. In contrast, age 1 was predicted accurately due to the clearly distinct facial characteristics of infants.

These results show that while **simplifying the task** with age ranges can help reduce complexity, it also **limits the model's ability to capture subtle, year-to-year facial changes** and **worsens performance on less represented groups.**

Future improvements**:**

- **Train on a less biased dataset for better generalization:** We could involve training on more balanced or demographically diverse datasets, such as Morph II [2]+ or APPA-REAL [3], to improve robustness and fairness, particularly in underrepresented age groups.
- **Modify CNN architecture: adding convolutional layers with padding:** Introducing additional convolutional layers (possible if we added padding) could preserve spatial dimensions and allow the model to extract deeper hierarchical features. This may enhance age-related pattern recognition, especially in the mid and deep layers.
- **Analyze model performance across ethnicities:** UTKFace includes ethnicity annotations, which were not used in our current experiments. Conducting a subgroup analysis could reveal whether the model performs equitably across ethnic groups or if biases are present.
- **Exploring Alternative Loss Functions:** As discussed in the presentation, we could benefit from testing different loss functions designed to handle class imbalance better, potentially improving the model's performance on underrepresented ages and making predictions more balanced and reliable.
- **Compare classification vs. regression approaches:** As discussed in the presentation, reformulating the problem as a regression task could better reflect the continuous nature of age, so we could implement it. That could create a more precise error analysis and offer valuable insights when comparing loss behaviors.

7.CONCLUSION

This project set out to repurpose a CNN pre-trained for facial emotion recognition and adapt it for age classification across 100 discrete classes. We found that the number of frozen layers in transfer learning had a key impact on balancing general visual knowledge with age-specific feature learning. Additionally, the structure of the fully connected layers proved crucial to optimize the model's adaptability and classification accuracy.

To fairly assess performance, we relied on multiple, task-specific evaluation metrics beyond standard accuracy. These metrics revealed that while the model performed well on frequently occurring ages, it struggled on underrepresented ones, highlighting the importance of considering class imbalance in both model training and evaluation.

Despite attempts to mitigate this imbalance through data augmentation, improvements were limited due to insufficient base diversity. Future work should explore regression-based reformulations, alternative loss functions, and more balanced datasets to boost robustness and fairness. Ultimately, this project demonstrated the potential of transfer learning across facial analysis tasks and the need for careful, context-aware evaluation in real-world applications.

REFERENCES

1. Goodfellow, I., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Bengio, Y. (2013). *Challenges in representation learning: A report on three machine learning contests*. In International Conference on Neural Information Processing (pp. 117–124). Springer. https://arxiv.org/abs/1307.0414

2. Ricanek, K., & Tesafaye, T. (2006). MORPH: A longitudinal image database of normal adult age-progression [Dataset]. Face Aging Group, University of North Carolina Wilmington
3. Agustsson, E., Timofte, R., Van Gool, L., & Esteban, D. (2017). APPA-REAL: Apparent and real age estimation dataset [Dataset]. Computer Vision Lab, ETH Zurich.
4. Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In Advances in Neural Information Processing Systems (NeurIPS) (pp. 3320–3328). https://arxiv.org/abs/1411.1792

5. Cao, B., Mirjalili, V., & Raschka, S. (2020). Rank-consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140, 469–475.
6. Rothe, R., Timofte, R., & Van Gool, L. (2015). DEX: Deep EXpectation of apparent age from a single image. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 10–15). https://doi.org/10.1109/ICCVW.2015.41

7. Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In *European Conference on Computer Vision* (ECCV) (pp. 818–833). Springer. https://doi.org/10.1007/978-3-319-10590-1_53

8. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. (See Chapter 6: Deep Feedforward Networks, Section 6.2.4 Softmax Units)

APENIDX:

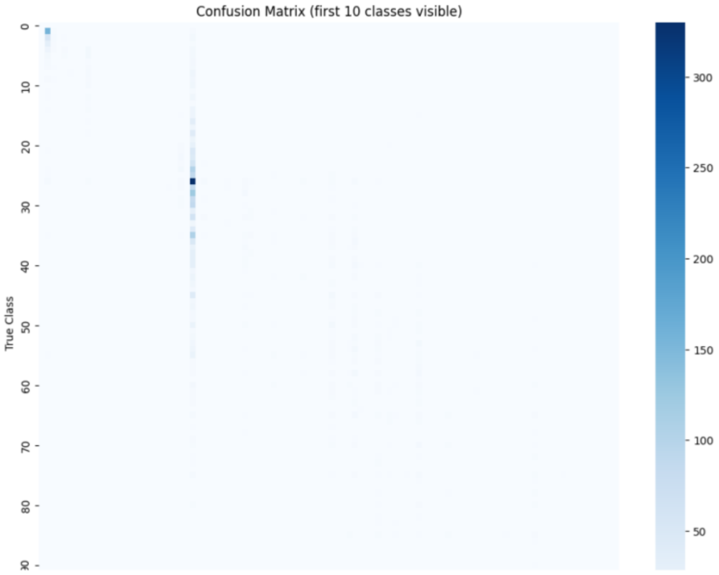


Figure A1: confusion matrix age 26 (without undersampling), data analysis

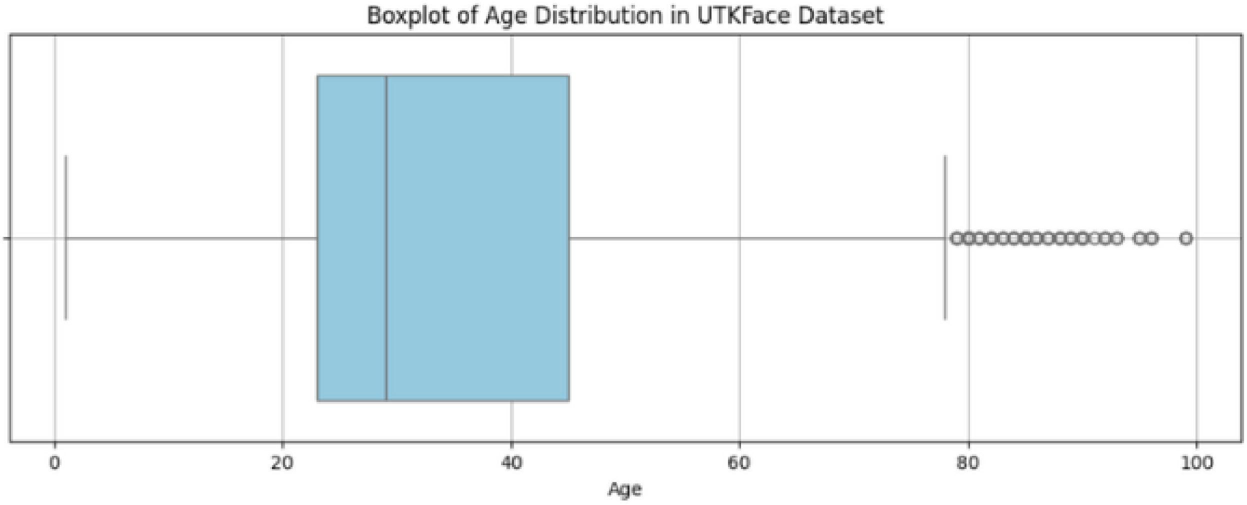


Figure A2: Boxplot of initial age Distribution

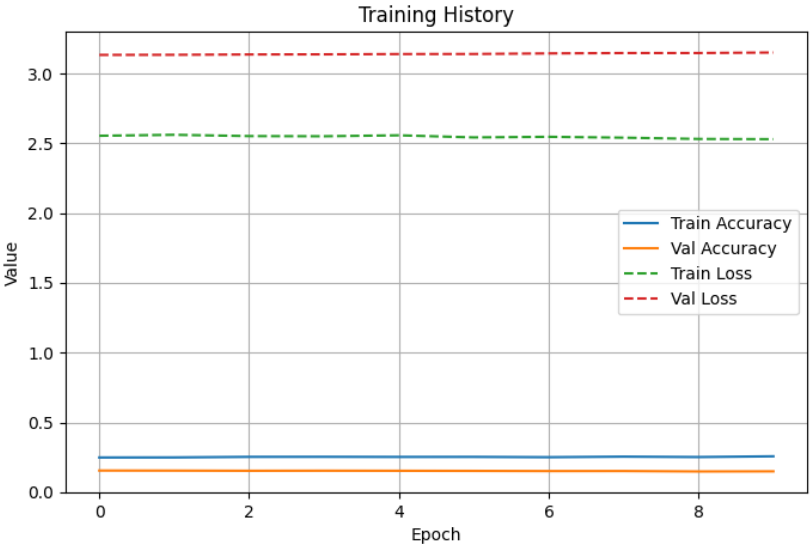


Figure A3: unfreezing the last convolutional block

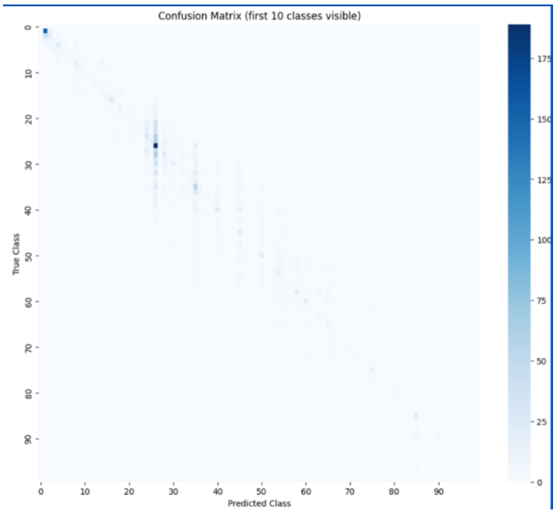
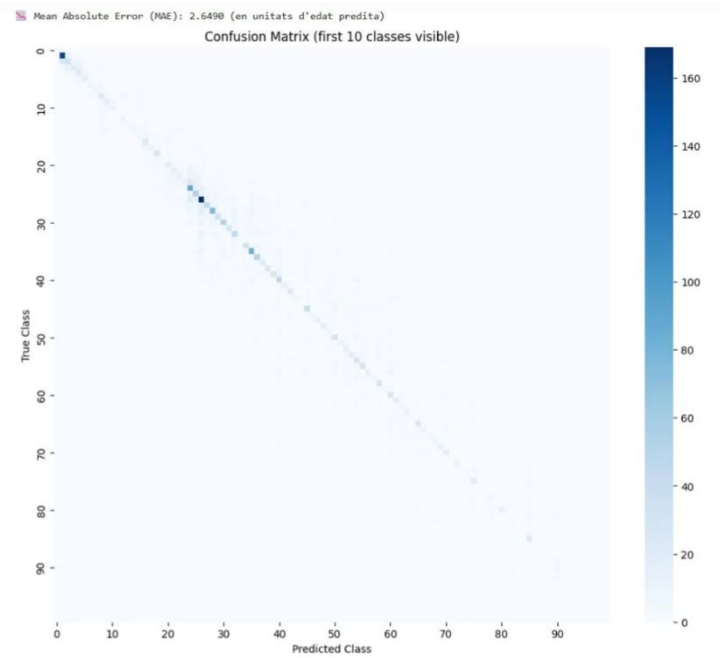


Figure A4: confusion matrix lacked a strong diagonal structure, second experiment



```
Mean Absolute Error (MAE): 2.6490 (en unitats d'edat predita)
```

Figure A5: confusion matrix with strong diagonal structure, third experiment

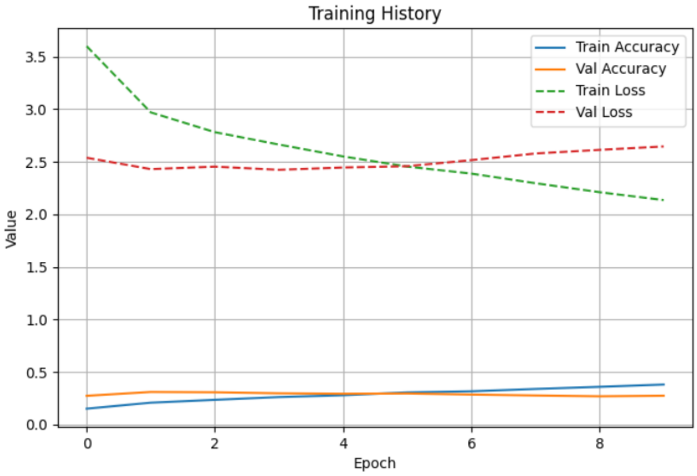


Figure A6: Results using CNN with 2 additional Fully Connected layers

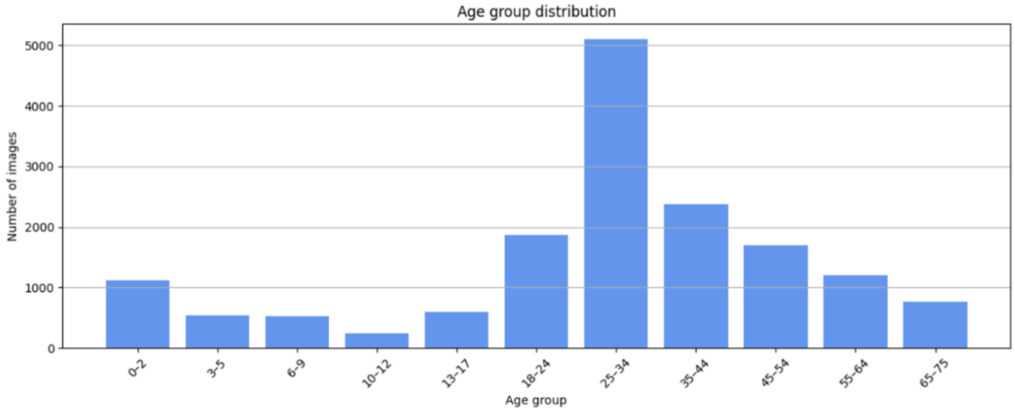
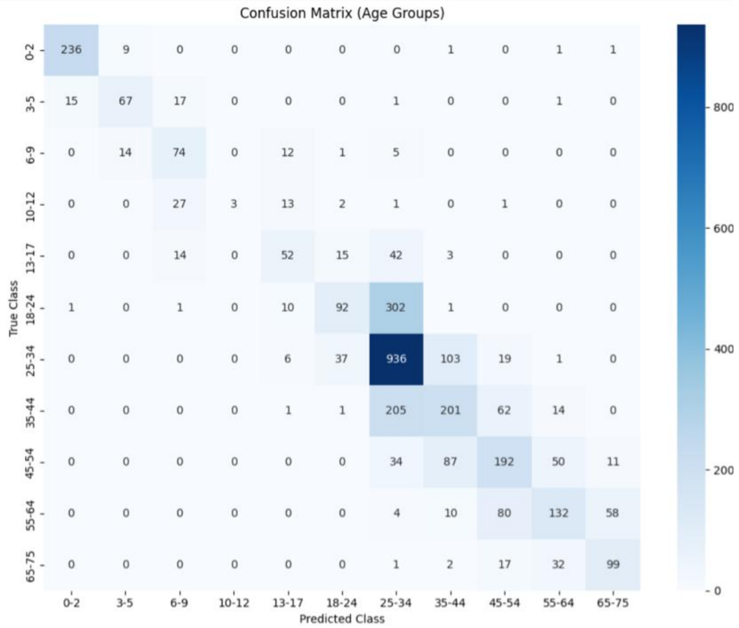


Figure A7: age ranges distribution using the paper by Rothe et al. Sixth experiment



	0-2	3-5	6-9	10-12	13-17	18-24	25-34	35-44	45-54	55-64	65-75
0-2	236	9	0	0	0	0	0	1	0	1	1
3-5	15	67	17	0	0	0	1	0	0	1	0
6-9	0	14	74	0	12	1	5	0	0	0	0
10-12	0	0	27	3	13	2	1	0	1	0	0
13-17	0	0	14	0	52	15	42	3	0	0	0
18-24	1	0	1	0	10	92	302	1	0	0	0
25-34	0	0	0	0	6	37	936	103	19	1	0
35-44	0	0	0	0	1	1	205	201	62	14	0
45-54	0	0	0	0	0	0	34	87	192	50	11
55-64	0	0	0	0	0	0	4	10	80	132	58
65-75	0	0	0	0	0	0	1	2	17	32	99

Figure A8: confusion matrix of the age ranges experiment. Sixth experiment

Figure A9: detailed breakdown of F1-scores:

Criteris per seleccionar edats importants-> del sense data augmentation

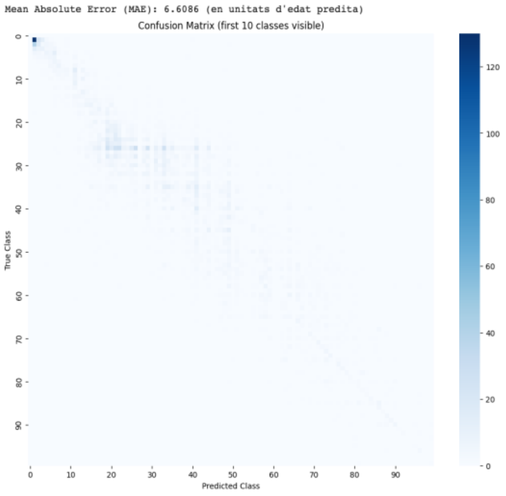
- **Suport ≥ 40**
- **F1-score ≥ 0.70**

(Classe)	F1-Score	Precisió	Exhaustivitat (Recall)	Suport
1	0.87	0.84	0.90	164
2	0.71	0.72	0.69	84
16	0.69	0.63	0.75	44
18	0.69	0.61	0.80	49
21	0.76	0.73	0.79	58
29	0.76	0.72	0.79	47
32	0.75	0.69	0.82	49
36	0.74	0.76	0.73	45
38	0.81	0.74	0.89	53

Figure A10: Taula amb edtas mes interessenats

RANKS	Precision	Recall	F1-score	Support
0-2	0.9365	0.9516	0.9440	248
3-5	0.7444	0.6634	0.7016	101

6-9	0.5564	0.6981	0.6192	106
10-12	1.0000	0.0638	0.1200	47
13-17	0.5532	0.4127	0.4727	126
18-24	0.6216	0.2260	0.3315	407
25-34	0.6114	0.8494	0.7110	1102
35-44	0.4926	0.4153	0.4507	484
45-54	0.5175	0.5134	0.5154	374
55-64	0.5714	0.4648	0.5126	284
65-75	0.5858	0.6556	0.6188	151



```
Mean Absolute Error (MAE): 6.6086 (en unitats d'edat predita)
```

Figure A11: cf 100 augm data