Advanced Approaches to Age Recognition

LSE candidate Numbers - 26990, 35644, 28622, 22533

Abstract

Accurately estimating age from facial images is vital for various applications but presents challenges due to environmental factors and individual variability. In this study, we review age recognition methods, focusing on transfer learning using ImageNet with VGG-19 and ResNet-based architectures. Using a classification approach, we preprocessed 42,000 images from IMDb-WIKI data into five equally spaced age categories and explored the efficacy of applying transfer learning techniques in improving accuracy. Additionally, we investigated the models' interpretability through GradCAM to analyse the differences in the age prediction mechanisms of the two models. We find that the VGG-19-based model has more concentrated activation areas in the final convolutional layer, indicating a tendency to focus on specific features. Our findings demonstrated that the ensemble learning approach enhances age recognition performance through combining predictions from both the VGG-19 and ResNet architectures, achieving an overall accuracy of 70.27%.

1 Introduction

Age recognition from face image data is a fundamental task with diverse applications across domains such as biometrics, surveillance, human-computer interaction, and demographic analysis. Nevertheless, accurately estimating age from facial images presents significant challenges. In "in the wild" scenarios, where images are captured in unconstrained environments akin to real-world situations, diverse variations in pose, illumination, expression, and occlusion are exhibited, making them particularly important for the applicability of age recognition models. However, these variations introduce complexities that traditional age recognition methods struggle to address. Moreover, the variability in the ageing process across individuals, influenced by genetic factors, lifestyle choices, and environmental factors, further compounds the difficulty of accurately predicting age from facial images.

To address these challenges, our paper provides a review of state-of-the-art methods for age recognition that have demonstrated promising results in recent literature, each offering unique strategies to tackle this challenge. These approaches take advantage of transfer learning and pre-trained convolutional neural network architectures such as VGG-19 and ResNet152V2. In our project, we specifically aimed to evaluate the effect of transfer learning on these models. By leveraging transfer learning, these architectures can utilise knowledge gained from large-scale datasets and adapt this to the task of age recognition.

Although there are different ways to approach the task of age prediction, these largely fall into classification or regression-based methods. In this paper we adopt a classification approach, preprocessing the data by excluding ages above 99 and categorising ages into five equally spaced classes to ensure a balanced and practical age representation for model training and evaluation. This approach involves using the softmax probabilities outputted for each age category by the model and selecting the category with the highest predicted probability. This allows for a balance between simplicity, detail, and precision.

To train and validate the effectiveness of these approaches, we used a subset of 42,000 images (due to computational restrictions) from the IMDb-WIKI dataset. This dataset has a rich collection of over 500,000 face images annotated with gender and age labels. In our exploration, we sought to understand how leveraging pre-trained architectures (on ImageNet) could enhance the accuracy and robustness of age recognition systems.

Building upon this exploration, we delve into the interpretability of the models using GradCAM [11], allowing us to analyse the performance and insights provided by our models in different scenarios. Our aim is to deepen our understanding of the mechanisms driving age prediction.

Finally, despite computational restrictions, our analysis revealed interesting results. Notably, the ensemble model outperformed both individual architectures, achieving an overall accuracy of 70.27%, surpassing VGG-19's accuracy of 67.18% and ResNet's accuracy of 64.44%. Additionally, the ensemble model showcased competitive one-off accuracy at 98.08%, indicating its effectiveness in near-miss classifications. Furthermore, examination of confusion matrices revealed insights into the classification patterns of both VGG-19 and ResNet-based models, highlighting areas of improvement and indicating the potential for further refinement in the age classification process.

2 Related Work

Age estimation from facial images has garnered significant attention in research due to its broad applications across various domains. Previous studies have explored diverse methodologies, ranging from traditional machine learning approaches to more recent advancements in deep learning techniques.

Rothe et al.'s (2018) [10] paper stands out as a pivotal advancement within the realm of age estimation. First, their introduction of the IMDB-WIKI dataset, the most extensive collection available for age estimation, serves as a fundamental resource for both training and evaluating age estimation models. Moreover, their innovative approach circumvents the challenges posed by facial landmarks in unconstrained settings, advocating for a deep learning methodology utilising a CNN with the VGG-16 architecture pre-trained on ImageNet. This CNN can directly infer age from facial images without relying on explicit landmark localisation. By leveraging deep expectation, their methodology aims to capture both the genuine and perceived ages of individuals in images. Another notable aspect is their approach of redefining the age prediction task as a classification problem, discretising age values. This strategic adaptation adeptly tackles the instability inherent in training a CNN directly for regression, owing to the presence of substantially large gradients which makes it difficult for the network to converge. This not only stabilises the training process but also enhances the reliability of predictions, marking their methodology as a significant advancement in the field of age estimation. However, this approach is computationally intensive, which limits our ability to adopt the similar approach.

Similarly, Sheoran et al. (2021) [12] delve into age estimation from facial images, offering a comprehensive comparison between custom CNN architectures and pre-trained CNNs used as feature extractors. Their study underscores the superiority of pre-trained models over custom CNN architectures and classical machine learning algorithms for age and gender prediction tasks. Particularly noteworthy is their exploration of transfer learning, which exhibits substantial improvements in handling challenges associated with small dataset sizes and dataset imbalance. The study demonstrates that even simple linear regression models trained on features extracted from pre-trained models outperform custom CNN models trained from scratch in age estimation tasks. Among the pre-trained models evaluated, SENet50 pre-trained on VGGFace2 (denoted as SENet50_f) emerges as the top performer, providing the most effective feature extraction for both age estimation and gender classification tasks. Leveraging the UTKFace dataset, their research showcases the diversity of images in terms of expression, illumination, pose, resolution, and occlusion. By leveraging deep CNNs and transfer learning, particularly through pre-trained models as feature extractors, they highlight the effectiveness of combining custom CNN architectures with transfer learning for age and gender prediction tasks. Their findings suggest avenues for further exploration on larger and more balanced datasets to enhance model generalisability and mitigate biases.

Several other studies have leveraged transfer learning for age estimation. Fang et al. (2019) [4] introduced a multi-stage learning approach using a pre-trained VGG-19 model, incorporating a saliency detection network and a modified VGG-19 model for age and gender estimation. Guo et al. (2019) [5] employed pre-trained architectures (VGG16, VGG-19, ResNet50, InceptionV3, and Xception), emphasising a regression-based approach with a fine-tuning process. Antipov et al. (2017) [1] utilised pre-trained models for age and gender estimation, and highlighted the effectiveness of the modified VGG16 architecture. Similarly, Uddin et al. (2021) [13] utilised pre-trained models (VGG16, ResNet50, SENet50) with K-fold cross-validation for age group classification, addressing overfitting and data disparity issues. Dagher and Barbara (2021) [3] explored facial age estimation using pre-trained CNNs and transfer learning, determining the optimal number of age classes and the ideal age range between classes for accurate estimation. Finally, Nam et al. (2020) [9] improved age estimation accuracy on low-quality images using a Conditional Generative Adversarial Network (CGAN) model, while Liu et al. (2020) [8] introduced a novel approach based on data augmentation and a lightweight CNN, achieving competitive results.

3 Methodology

In this section, we detail the data preprocessing and the two approaches used in our study: the VGG and ResNet architectures. These architectures, pre-trained on ImageNet, have demonstrated promising results in recent literature, serving as the foundation for our investigation. Each approach is fine-tuned to address the task at hand.

3.1 Preprocessing

We used the extensive IMDB-WIKI dataset, boasting over 500,000 labelled samples from 20,284 individuals aged between 1 and 126 years old, standing as one of the largest publicly available face image datasets. This collection merges samples sourced from people listed on IMDB and Wikipedia, predominantly within the age range of 20 to 50 years old, albeit with fewer images representing individuals aged 20 and below (see Appendix B Figure 12).

To ensure the integrity and quality of the data, we employed dlib's CNN-based face recognition model [7] for extracting facial regions. This model, designed to efficiently detect and recognise faces in images using CNNs, streamlines the facial recognition task while eliminating the need for training a model from scratch.

Our preprocessing pipeline began by standardising facial images to a dimension of 128x128 pixels with a 25% margin, aiming to adequately capture facial features while minimising background noise. Despite time and computational constraints on Google Colab's GPUs, we efficiently processed 42,000 well-aligned facial images.

Subsequently, to obtain a well-balanced dataset and mitigate potential biases, we undertook further data processing steps. Initially, to focus our analysis on ages within a reasonable range, we excluded entries with ages below 0 and above 99. While this may exclude some valid data points, such as individuals who lived beyond 99 years, visual inspection revealed that these extreme ages often represented anomalies or erroneous entries with mislabelled data. Thus, this approach helped mitigate potential noise in the data without significant loss of valid information.

Next, to standardise the image quality across the dataset, we performed pixel value clipping, constraining the pixel values to lie within the range of 0 to 255. This step helped mitigate potential inconsistencies in image brightness or contrast that could arise from variations in the original pixel values. Furthermore, we resized all images to a uniform dimension of 224x224 pixels. This standardisation ensured compatibility with the CNN architectures that were utilised later, which require input dimensions of 224 x 224 pixels.

The following step was to sort the integer value ages to five different categories, each with a class width of 20 years:

- Category 0: Ages 1 to 19 (inclusive)
- Category 1: Ages 20 to 39 (inclusive)
- Category 2: Ages 40 to 59 (inclusive)
- Category 3: Ages 60 to 79 (inclusive)
- Category 4: Ages 80 to 99 (inclusive)

From our experimentation with a variety of classes and class widths, we conclude that this strikes an optimal balance between granularity and practicality in age representation for model training and evaluation.

Afterwards, acknowledging the gender imbalance in the dataset, which skewed towards males (Appendix B Figure 13), and considering that aging characteristics may differ between genders, we proactively rebalanced the dataset to ensure a 50/50 split in genders in the final dataset. This process guarantees a balanced representation of gender within each age category, minimising bias during model training.

We then implemented down-sampling to reduce the influence of over-represented age classes on model performance (Figure 1). By calculating the average number of samples per age class and targeting 60% of this value as the maximum sample size, we achieve a more balanced distribution across age categories (Figure 2).

Following this down-sampling step, data augmentation techniques were applied to enrich the dataset and introduce variability, crucial for robust model training. Leveraging augmentation parameters such as rotation, shifting, shearing, zooming, and flipping, we generated diverse representations of each image, enhancing the model's ability to generalise. Moreover, conditional augmentation further customised transformation intensities based on age categories, with more aggressive augmentation applied to underrepresented categories such as categories 0 (ages 0-19) and 4 (ages 80-99). This approach accounted for inherent differences in age groups, optimising the augmentation process for improved model performance.

After augmentation, we normalised pixel values ranging from 0 to 1 facilitating uniform data representation and aiding in convergence during model training. Finally, the dataset was split into training, validation, and test sets with a 80%/10%/10% split, maintaining stratification by age class to preserve category proportions across partitions.

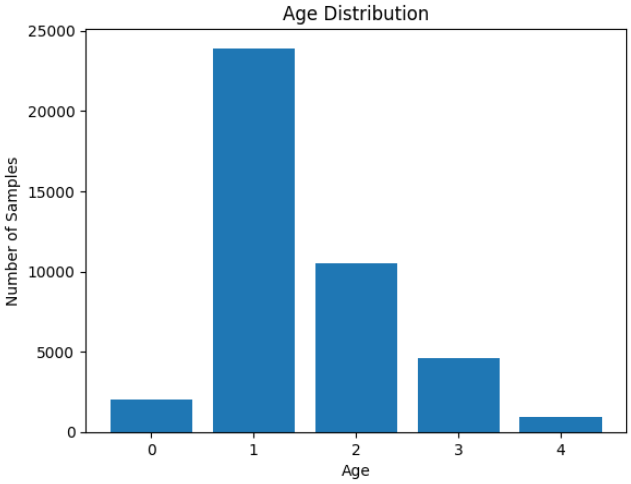

Figure 1: Initial age distribution through the 5 categories.

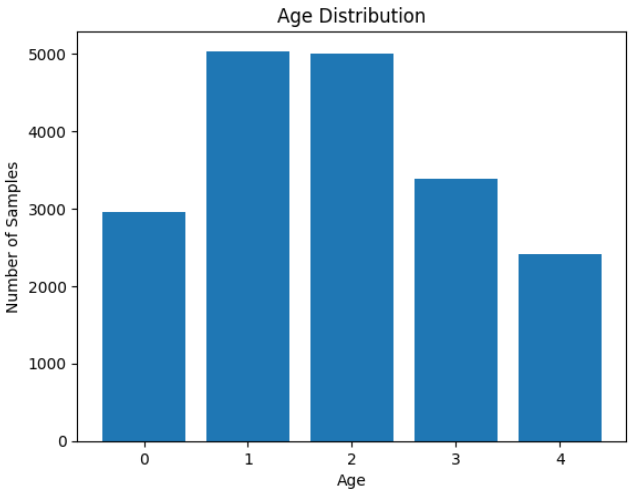

Figure 2: Age distribution through the 5 categories after rebalancing.

This comprehensive preprocessing pipeline prepares the dataset for subsequent model development and evaluation, laying a solid foundation for accurate age prediction in diverse scenarios.

3.2 The Architectures

In this section, we detail the methodology adopted for age classification using Convolutional Neural Networks (CNNs), leveraging the robust capabilities of pre-trained models to enhance prediction accuracy. Our approach incorporates two distinct CNN architectures as base models — VGG-19 and ResNet152V2. These architectures are not only widely utilised due to their proven effectiveness in various image recognition tasks, but also offer a unique opportunity to compare the performance implications of different model depths. These base models are each supplemented with a custom-designed architecture based on a saliency-to-CNN structure. These additional layers on top of the base models focus on extracting salient aspects relevant to age differentiation, later funnelling these through additional convolutional layers to finally classify inputs into one of the predefined age categories. To further improve prediction accuracy, we then explore ensembling methods to combine the outputs from the two models.

VGG-19-Based Model The first architecture employs VGG-19 as its base model, a 19-layer deep CNN developed by the University of Oxford's Visual Geometry Group. VGG-19 uses small convolution filters and processes RGB images with dimensions of 224x224 pixels. Initially, it calculates the average RGB values across all training set images, which are then fed into the convolutional network. The network consistently uses filters of sizes 3x3 or 1x1 during the convolution steps. The model has a total of 143.7 million parameters. Around this foundational base, we develop a customised model that integrates a saliency detection mechanism with additional convolutional neural network layers. The saliency network is structured with an encoder-decoder architecture to generate a saliency map from the input images, emphasising features critical for age differentiation. This network begins with consecutive convolutional and max pooling layers to downsample the input, followed by a bottleneck layer, and concludes with upsampling layers merged with the corresponding earlier layers to reconstruct the feature space, outputting a saliency map. This map is then expanded across three channels and fed into a pre-trained VGG-19 model, where the top eight layers have been fine-tuned for this task. The output from the VGG-19 base model is processed through one more convolutional layer followed by batch normalisation and several dense layers, which include dropout and L2 regularization to prevent overfitting. This setup leads to a softmax layer that classifies the predicted age into five distinct categories.

ResNet-Based Model Conversely, our second architecture utilises the ResNet152V2 model as our base model, which is characterised by its deep residual learning framework that aids in training deeper networks without performance degradation. Developed by Microsoft Research, the ResNet family introduces this residual learning concept of enabling the gradient to flow directly through the skip connections across layers without undergoing transformation by deep layer weights, thus helping to solve the vanishing gradient problem that is commonly observed with deeper models. The ResNet152 and ResNet152V2 models are the deepest models in the family at 152

layers, contrasting to the 50 layers of the ResNet50 base model, thus allowing for the learning of more complex features. Furthermore, the V2 variants of the ResNet-based models apply batch normalisation and ReLU activation before the convolution operation within each residual block (pre-activation), which maintains the mean and variance of the input distribution throughout the depth of the network and promotes a more stable training process. The combination of increased model depth and improved architecture of the residual blocks allows the model to more effectively maintain feature integrity across deeper network layers, ensuring robust age predictions. Similarly to the VGG-19-Based model, we utilise the saliency network to generate a saliency map before feeding into the ResNet152V2 base model, and then pass the output from the base model through the same convolutional and dense layers with dropout and L2 regularisation before generating predictions from the final softmax layer.

3.3 Training Methods

The training process for both models employs data augmentation techniques and leverages callbacks such as EarlyStopping and ModelCheckpoint for optimisation.

VGG-19-Based Model Central to the training process is the Adam optimiser with an initial learning rate of 0.0001. This learning rate is adaptively reduced upon encountering a plateau in validation loss, a strategy facilitated by the ReduceLROnPlateau callback, which decreases the learning rate by a factor of 10 if no improvement is seen in validation loss for two consecutive epochs, with a floor set at 0.000001. Model training is conducted over 25 epochs with a batch size of 32, utilising a data generator for augmented input batches to enhance generalisability and prevent overfitting. Overfitting is further mitigated by applying dropout layers within the dense segments of our network and employing L2 regularisation in key neural layers. Model performance is closely monitored through validation data comprising separate sets of images and labels. As implied above, we utilise the ModelCheckpoint callback to save the model weights at the epoch where validation loss is minimised, thereby ensuring recovery of the most effective model state post-training. The EarlyStopping callback is also employed, halting training if the validation loss does not improve over five epochs. It restores weights from the best-performing epoch to prevent the loss of training progress due to potential overtraining.

Building upon our methodology, we extended the VGG-19-based model's training to an eight-category age classification scheme to investigate the model's adaptability to finer age distinctions. This phase maintained the original optimiser settings and regularisation strategies, ensuring methodological consistency while evaluating the model's capability to distinguish more nuanced age-related features across a broader spectrum of age categories.

ResNet152V2-Based Model We utilise a similar training process for the ResNet-Based model, initially freezing the pretrained weights for the base ResNet152V2 model learned from the ImageNet database, and training the unfrozen weights for 20 epochs with a batch size of 32. We also use the Adam optimiser for the training process with the EarlyStopping, ModelCheckpoint, and ReduceLROnPlateau callbacks to optimise convergence and prevent overtraining. To fine-tune the model, we adopt a gradual unfreezing approach for the weights in the ResNet base model. Following the initial training cycle, we unfreeze the weights of the final 10 layers in the base model, decrease the learning rate by a factor of 10 compared to our initial training cycle, and train the entire model again for 10 epochs. We sequentially proceed to unfreeze and train the preceding 10 layers, continuing this pattern through the network until model performance fails to improve for 2 sequential training cycles. Through our experimentation process, we find that fine-tuning fails to produce improvements in validation accuracy after the 4th fine-tuning training cycle. We also consider fine-tuning the base model through unfreezing the entire model and training all of the weights in the base model with a low learning rate - however, this fails to yield as significant of an accuracy improvement as the gradual unfreezing approach.

Ensemble Further we construct an ensemble model, combining the outputs from the two primary neural networks (VGG-19-based and ResNet-based models) through various integration techniques. Initially, a simple averaging method was applied to merge the predictions. Subsequently, to refine this integration, a more sophisticated approach was adopted involving a neural network meta-model. The meta-model was designed as a sequential neural network consisting of multiple dense layers with activations. It starts with a 512-unit dense layer and progresses through decreasing layer sizes down to a final 5-unit dense layer using a softmax activation suitable for multi-class classification. The model is trained using the combined predictions from both the VGG-19-based and ResNet-based models, aiming to learn the underlying relationships between these model predictions and the actual age labels. Through this process, the model seeks to capture and exploit any complementary strengths and patterns present in the outputs of the individual models. To further enhance our model's optimisation, a random search was conducted to fine-tune the hyperparameters of a similar neural network structure. The search varied the number of dense units, the number of layers, dropout rates, and learning rates to identify the optimal configuration. The best model parameters were determined based on validation accuracy over a series of 20 trials.

4 Numerical Results

The primary goal of the numerical evaluations presented in this paper is to assess the effectiveness of the proposed age classification models, specifically comparing the performance of individual models based on VGG-19 and ResNet architectures against an ensemble approach. We focus on overall accuracy and class-specific accuracies, alongside one-off accuracy to measure near-miss classification performance. The baseline for comparison includes the individual model performances, where the VGG-19-based and ResNet-based models provide benchmarks for assessing the enhancement offered by the ensemble method. The ensemble model combines predictions from both base models using a meta-model approach, aiming to leverage complementary strengths. Numerical results are detailed in Table 1, showcasing the accuracies achieved by each model setup.

The principal accuracy metric we use is the categorical accuracy, reflecting the proportion of correct predictions over the total number of cases evaluated. This metric offers a straightforward interpretation of the model's predictive accuracy across the designated age classes. Complementarily, the loss function employed during

Table 1: Classification accuracies and one-off accuracies for VGG-19, ResNet, and Ensemble models

Metric	VGG-19	ResNet	Ensemble
Overall Accuracy	67.18%	64.44%	70.27%
One-Off Accuracy	97.52%	95.98%	98.08%
Class 0	60.14%	61.66%	69.76%
Class 1	73.16%	67.20%	74.16%
Class 2	68.00%	62.70%	69.70%
Class 3	59.32%	54.59%	56.36%
Class 4	72.67%	79.50%	83.44%

training and evaluation phases is categorical cross-entropy. This function measures the dissimilarity between the predicted probability distribution over age categories and the true distribution, where the true labels are one-hot encoded. A lower cross-entropy value indicates that the model's predicted probabilities are converging towards the actual labels, thus signifying better performance.

We can observe from Table 1 that the VGG-19-based model outperforms the ResNet-based model, with a overall prediction / one-off accuracy of 67.18% / 97.52% to the ResNet-based model's scores of 64.44% / 95.98%. This difference in performance can be attributed to the challenges associated with fine-tuning the ResNet-based model, which has a significantly deeper architecture with over 150 layers. In our experiments, fine-tuning beyond the last 40 layers did not yield improvements, suggesting that many of the deeper layers retained their initial training focused on general image recognition from the ImageNet dataset, which might not align well with the specific nuances of age recognition. This indicates that for our specific application and given computational constraints, a shallower model like VGG-19, which has only 19 layers, may be more effective as it can be more comprehensively fine-tuned to adapt to the task of age recognition, thus providing better generalisation from the training to the application domain.

The ensemble approach of combining the predictions from the two models achieves a superior overall accuracy compared to the two individual models with 70.27% accuracy and 98.08% one-off accuracy. By integrating the unique strengths of each model, the ensemble approach mitigates their individual weaknesses, resulting in a more robust and accurate age classification system.

We next turn to the confusion matrices for the classifications made by the VGG-19-based model and ResNet-based model for a more thorough breakdown of model performance by category. Firstly, Figure 3 displays the results for the ResNet-based model, and Figure 4 displays the results for the VGG-19-based model. The cell values in the left matrix represent the absolute number of images in each cell, and the cell values in the right matrix represent the percentage of predictions for each true class. As reflected by the one-off accuracy results in Table 1, the two sets of confusion matrices show that for the 3757 observations in the test set, the majority of inaccurate predictions are one-off the true category - with the remaining cell values (located in the bottom left and top right corners) representing less than 5% of the total observations in each category.

One key pattern observed in the results for both models, is that they particularly struggle to accurately distinguish between categories 1 (ages 20-39) and 2 (ages 40-59). For example, for true category 1 in the ResNet-based model, there is a significant misclassification rate where 24.75% of the instances are incorrectly predicted as category 2, compared to a much lower rate of 7.36% being misclassified as category 0. Conversely, for true category 2, the model more frequently underestimates the age, with 23.4% of instances incorrectly classified as category 1, which is substantially higher than the 11.2% misclassified as category 3.

The confusion between age categories 1 (20-39) and 2 (40-59) could be largely due to subtle ageing signs that are less distinct between young and middle-aged adults compared to the clearer distinctions with younger (0-19) and senior (60-79, 80-99) groups. Our models, trained to predict biological rather than apparent age, face further challenges from our image sources—Wikipedia and IMDb. These platforms often feature individuals from industries such as fashion and media who may appear younger due to cosmetic treatments and professional photography. This can obscure the age distinctions between the 20-39 and 40-59 categories, leading to frequent underestimations or overestimations of age. As Figure 5 shows, even human observers may find it difficult to accurately categorise these images with high confidence. However, this inability to distinguish between categories 1 and 2 could also be due to the significant underrepresentation of some categories in the original dataset, which we attempted to mitigate by oversampling these categories. Despite employing extensive data augmentation, these categories may still suffer from a lack of diversity in the training data.

The wider confusion matrix also reflects the VGG-19-based and ResNet-based model's ability to be able to clearly differentiate between the 20-59 age range (Categories 1 & 2), and the 0-19 / 60-99 age ranges (Categories 0, 3 & 4). The model demonstrates a high degree of precision in avoiding false predictions for categories 0 (children and teenagers), 3 (early seniors), and 4 (advanced seniors), with all incorrect predictions for these categories being below 16%. This suggests that the model is particularly adept at recognising the distinct age-related features that are more pronounced in the youngest and oldest age groups. However, again, this may be due to underrepresentation of categories 0, 3, and 4 in the original dataset, leading to the model being more frequently exposed to younger categories.

An interesting pattern emerges when analysing the rate of true positive predictions across the categories. For both models, the accuracy decreases from category 1 (ages 20-39) to category 3 (ages 60-79) but then unexpectedly increases for category 4 (ages 80-99), achieving scores of 79.50% and 72.67% for the ResNet-based and VGG-19-based model respectively. Under usual circumstances, one might expect a gradual decline in prediction accuracy as age increases, given the potential overlap in ageing features between the middle categories. The high accuracy for the oldest age group could be attributed to several factors. Older individuals often exhibit more distinctive ageing features such as wrinkles, skin sagging, and grey hair, which may be easier for the model to identify compared to the subtler signs of ageing in younger adults. Additionally, the oldest age group might exhibit less variability in facial features related to ageing compared to younger groups, where lifestyle and genetic factors can cause a broader range of appearances at the same age.

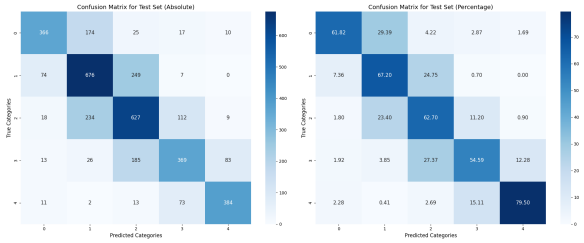

Figure 3: Test set confusion matrices for the ResNet-based model, for five categories

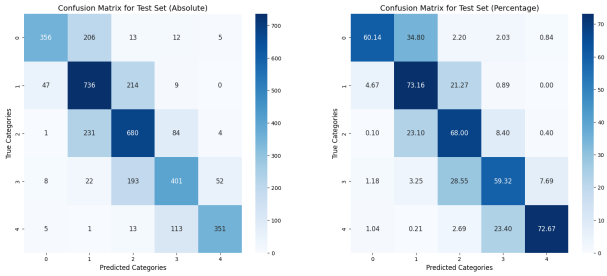

Figure 4: Test set confusion matrices for the VGG-19-based model. Absolute numbers (left) and relative numbers (right)

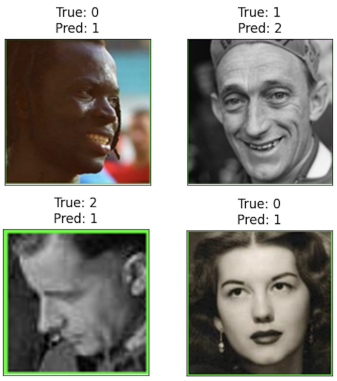

Figure 5: Selection challenging samples misclassified by the VGG-19-based model

Another potential challenge is the broad age range contained within each category. For instance, individuals aged 40 and 59 are grouped together, despite likely significant visible ageing effects across this span. This is the reason, why the VGG-19-based model was additionally trained on a dataset divided into eight categories. The performance outcomes were a test accuracy of 45.2% and a one-off accuracy of 87.1%. Comparing this model's performance with others is difficult due to the increased complexity from predicting eight categories instead of five, leading to an approximate 20% decrease in test accuracy. Nevertheless, it is noteworthy that the one-off accuracy diminishes by only about 10%. This smaller impact on one-off accuracy implies potential benefits in increasing the number of categories, as errors are relatively less critical with more categories. However, due to constraints in computational resources and time, further exploration of this hypothesis was not feasible. It remains an intriguing question whether the declines in accuracy would diminish with a continued increase in the number of categories. The confusion matrix in Figure 14 contains further details on model performance with eight categories.

5 Interpretability

Interpretability in deep learning refers to the ability to understand and explain how a deep neural network makes its predictions or decisions. Its causes include increasing trust in model predictions, identifying influential factors, and gaining insights that aid in developing or improving models. One method that is used for interpretability analysis is Gradient-weighted Class Activation Mapping (GradCAM)[6]. GradCAM visualises which regions of an image are crucial for a model's predictions by backpropagating the gradients from the target output to the convolutional layers. This creates a heatmap, pinpointing areas most influential to the decision-making process [11]. This is particularly useful in tasks such as image classification, object detection, and segmentation, where understanding the model's focus areas can provide valuable insights into its reasoning but also helps verify that the model's decisions align with human judgement [2].

Additionally, GradCAM can aid in identifying model biases or areas where the model may be making predictions based on spurious correlations within the data. This can guide improvements to the model architecture or the dataset to enhance the model's robustness and generalisation capabilities [11]. Moreover, GradCAM is a widely applicable technique that can be used with various convolutional neural network architectures without requiring significant modifications or retraining of the model, making it suitable for this study due to the limited resources available [11].

In this study, GradCAM analysis was performed using *viridis* colourmap, ranging from purple, corresponding to 0, to yellow, corresponding to 1. Figure 6 shows the GradCAM analysis of the final convolutional layer of the VGG-19-based model. The figure shows three correctly identified male and female images. For women (indices 150, 162, and 74), the heatmaps highlight the central features of the face, particularly around the eyes and nose, demonstrated by the bright yellow regions representing strong activation in these areas. For men (indices 159, 43, and 40), the activation seems more spread across the face, but still focuses significantly around the central facial features such as the nose and cheeks. However, some heatmaps also show activations extending into the forehead area, possibly indicating a focus on broader facial structures or differentiating features like beard shadows or facial expressions. The differences might indicate that the model has learned specific cues associated with gender.

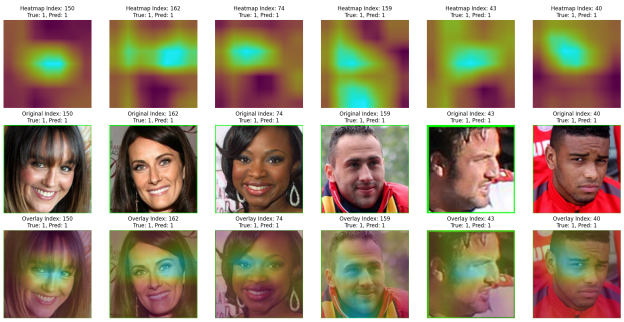

Figure 6: GradCAM analysis of VGG-19-based model - Correct Labels

Figure 7 presents the GradCAM analysis from the final convolutional layer of the ResNet-based model. This analysis, focused on the

same correctly identified male and female images, reveals a different pattern of activation. For women, the ResNet-based's heatmaps also concentrate around central facial features, however, the activation zones are more dispersed and less intense compared to the VGG-19-based model, suggesting a broader integration of facial regions into the decision-making process. For men, the ResNet-based model demonstrates an even more distributed activation across the face, including subtle activations in the peripheral facial regions. This diffusion could imply that the ResNet-based model utilises a more holistic approach to facial analysis, integrating a wider range of features for classification.

The differences in heatmap activations between the VGG-19-based and ResNet-based models underscore their inherent architectural distinctions. The VGG-19-based model's activations are localized and intense, reflecting its architectural bias towards capturing texture and specific details. Conversely, the ResNet-based model, with its residual learning approach, appears to capitalise on a broader contextual understanding of the image, which might explain its slightly lower test accuracy of 64% compared to the VGG-19-based model's 67%.

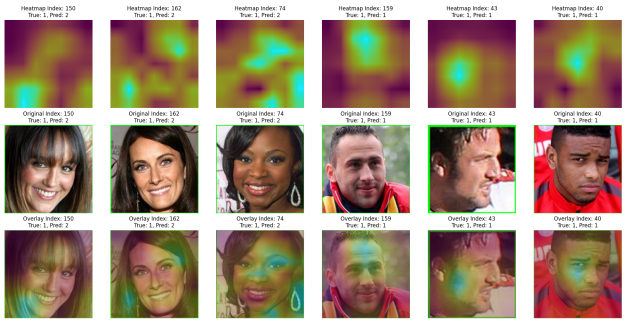

Figure 7: GradCAM analysis of ResNet-based model - Correct Labels

Figure 8 shows analysis of the same layer of the VGG-19-based model as previously. The figure shows falsely categorised images. The misclassified female images show more spread out activation compared to the correctly identified images in Figure 6. Furthermore, it is worth noting abnormalities in the misclassified images. Image 1 (index 1295) presents the model with a distorted face. The misalignment of the face is most likely the predominant reason for the larger activation surface. Image 3 (index 2197) presents obstruction in the form of glasses. Although the model focuses on the central facial area, the glasses seem to lead to a greater spread in activation. The fourth image (index 619) also introduces obstruction in the form of a bandanna. This model focuses on the obstruction rather than the facial features that could identify the age category. For the remaining images (indexes 1391, 2377, and 2406), the predicted age categories have manually been evaluated and agree with human predictions, or were categorised to be close to the edge of the corresponding predicted and true age categories. Therefore, these images likely represent outliers within their categories, either as they do not visually match their actual age, or because they are positioned near the threshold of adjacent age categories.

Figure 9 extends the analysis of misclassification to the ResNet-based model. Unlike the VGG-19-based model, the ResNet-based

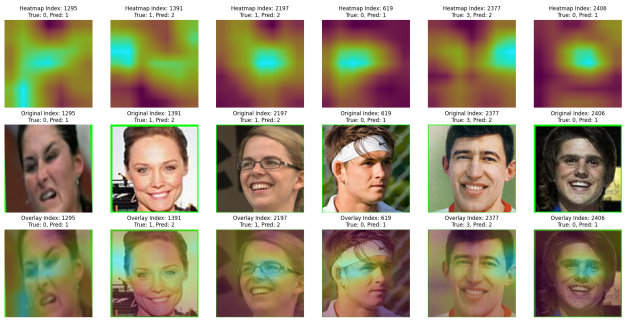

Figure 8: GradCAM analysis of VGG-19 based model - False Labels

model's heatmaps appear to have a wider and more uniform distribution across the face. This could suggest that the ResNet-based model, even in cases of misclassification, attempts to integrate a broader range of facial features into its decision-making process.

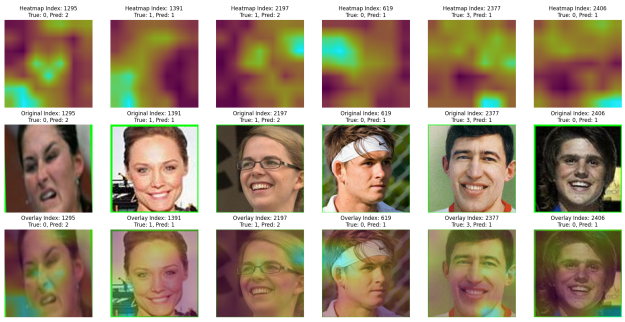

Figure 9: GradCAM analysis of ResNet based model - False Labels

Particularly, the ResNet-based model exhibits a notable activation in the peripheral regions of the face, such as the hairline and jaw, which might indicate an attempt to use more contextual information from the face's outline. However, similar to the VGG-19-based model's analysis, obstructions and facial distortions impact the model's accuracy. For example, in the misclassified images with obstructions such as glasses (index 2197) or facial accessories (index 619), the ResNet-based model, too, seems to misinterpret these features as relevant to the age classification. The spread of the heatmap across non-facial areas in these instances points to potential confusion in the model about which features are most predictive of age.

Additionally, for the images evaluated to be on the boundary of age categories or those presenting non-typical age features (indexes 1391, 2377, and 2406), the ResNet-based model's heatmaps are strongly diffused, suggesting a lack of strong defining features that align with the expected age characteristics. This diffuse activation could be attributed to the model's residual learning structure, which may incorporate broader image context but still fails to pinpoint the crucial age-discriminating features in challenging cases.

Figure 10 shows the GradCAM analysis of black and white images of the VGG-19-based model as before. It demonstrates that the heatmaps continue to focus on the central features of the faces. However, the heatmaps also show broader and more diffused areas of activation compared to the earlier colour images. This could

indicate a reliance on texture, shape, and intensity gradients rather than colour, which may be less effective. Overall, the fraction of misclassified black and white images was higher than coloured images, indicating potential challenges the model faces with grayscale inputs.

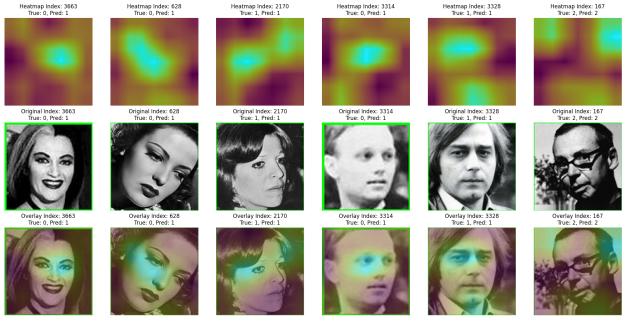

Figure 10: GradCAM analysis of VGG-19 based model - Grayscale Images

Continuing from the analysis of the VGG-19 model, Figure 11 presents the GradCAM analysis for the ResNet-based model using black and white images. Again, the heatmaps for ResNet-based model are typically more dispersed across the entire face. This pattern could indicate that the ResNet architecture attempts to extract a broader spectrum of structural and textural information from the grayscale images.

Figure 11: GradCAM analysis of ResNet based model - Grayscale Images

However, the results also reveal challenges similar to those faced by the VGG-19-based model when processing black and white images. The broader activations might reflect an uncertainty or a compensatory mechanism where the model struggles to identify decisive features that are usually highlighted by colour differences in coloured images. The misclassification rates for black and white images are comparably high, which underscores potential weaknesses in the model's capability to interpret grayscale inputs effectively. This issue is further complicated by the predominance of colour images in the training set, which does not adequately prepare the model for grayscale image recognition.

For practical applications, if the use of black and white images is anticipated, enhancing the training regimen with grayscale augmentation or employing image processing techniques to better define features independent of colour is advisable. This approach would help improve feature detection in grayscale and ensure more robust performance across diverse imaging conditions. Moreover, given that many of the black and white images are older, this raises the question about the relevance of training the model extensively on historical images, unless they are expected to be part of the model's operational environment.

In summary, both the VGG-19-based and ResNet-based models show nuanced and context-dependent heatmap activations, illustrating their architectural tendencies and operational distinctions. Correctly categorised images highlight each model's ability to focus on decisive facial features, while misclassifications and grayscale analyses reveal the challenges posed by obstructions, atypical features, and non-colour-based inputs. This study underscores the importance of tailored model training that anticipates actual deployment scenarios, suggesting that the inclusion of diverse image conditions, such as grayscale, may significantly enhance model robustness. Additionally, the varied activation patterns observed between the VGG-19-based and ResNet-based model across the same datasets emphasises the need for careful model selection based on specific application requirements to optimise both accuracy and interpretability.

6 Conclusion

Overall, our paper offers a comprehensive overview of age recognition using CNNs, focusing on the utilisation of pre-trained models to enhance prediction accuracy. We introduced and evaluated two distinct CNN architectures—VGG-19 and ResNet152V2—augmented with custom-designed saliency-based layers. Additionally, we explored ensemble methods to combine the outputs from these models, aiming for improved performance.

Overall, the ensemble model achieved an accuracy of 70.27% and outperformed the individual VGG-19-based and ResNet-based models in overall accuracy, class-specific accuracies, and one-off accuracy metrics. Additionally, the analysis of the confusion matrices provided key insights into specific areas of model performance and misclassifications.

Furthermore, our investigation into interpretability using Grad-CAM, revealed insightful differences between the VGG-19-based and ResNet-based models in their decision-making processes. While both showed strengths in identifying crucial facial features, misclassification instances and grayscale images highlighted some challenges. These findings emphasise the necessity of tailored model training to anticipate real-world scenarios, suggesting the inclusion of diverse image conditions like grayscale for improved robustness.

In addressing prevalent issues, such as data scarcity and quality concerns, future research could focus on refined data collection strategies and leverage advancements in architectures like GANs for improved data diversity. Proactively tackling these challenges and integrating interpretability techniques hold promise for enhancing the robustness and accuracy of age recognition systems across various domains.

References

[1] G. Antipov, M. Baccouche, S.-A. Berrani, and J.-L. Dugelay. Effective training of convolutional neural networks for face-based gender and age prediction. *arXiv preprint arXiv:1706.08476*, 2017.

[2] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, 10 2017.

[3] I Dagher and D Barbara. Facial age estimation using pre-trained cnn and transfer learning. *Multimedia*, 2021.

[4] J. Fang, Y. Yuan, X. Lu, and Y. Feng. Multi-stage learning for gender and age prediction. *arXiv preprint arXiv:1908.08579*, 2019.
[5] X. Guo, S. Li, J. Yu, J. Zhang, J. Ma, L. Ma, W. Liu, and H. Ling. Pfld: A practical facial landmark detector. *arXiv preprint arXiv:2019*, 2019.
[6] Yi han Sheu. Illuminating the black box: Interpreting deep neural network models for psychiatric research, 10 2020.
[7] Davis King. Dlib: A toolkit for making real world machine learning and data analysis applications in c++, 2024.
[8] X. Liu, Y. Zou, H. Kuang, and X. Ma. Face image age estimation based on data augmentation and lightweight convolutional neural network. *Journal of Electronic Imaging*, 29(1):013013, 2020.
[9] S.H. Nam, Y.H. Kim, N.Q. Truong, J. Choi, and K.R. Park. Age estimation by super-resolution reconstruction based on adversarial networks. *IEEE Access*, 8:29891–29903, 2020.
[10] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2):144–157, 2018.
[11] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *arXiv preprint arXiv:1610.02391v4*, December 2019.
[12] Vikas Sheoran, Shreyansh Joshi, and Tanisha R Bhayani. Age and gender prediction using deep cnns and transfer learning. In *Computer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part II*, volume 5, pages 293–304. Springer Singapore, 2021.
[13] S.S. Uddin, S. Morshed, M.I. Prottoy, and A.A. Rahman. Age estimation from facial images using transfer learning and k-fold cross-validation. *arXiv preprint arXiv:2101.07216*, 2021.

A Statement about individual contributions

- 28622: ResNet-based model architecture design, training, and respective write up
- 26990: VGG-19 Model architecture design, training, and respective write up
- 35644: Interpretability analysis and respective write up
- 22533: Preprocessing and remaining write up

B Figures

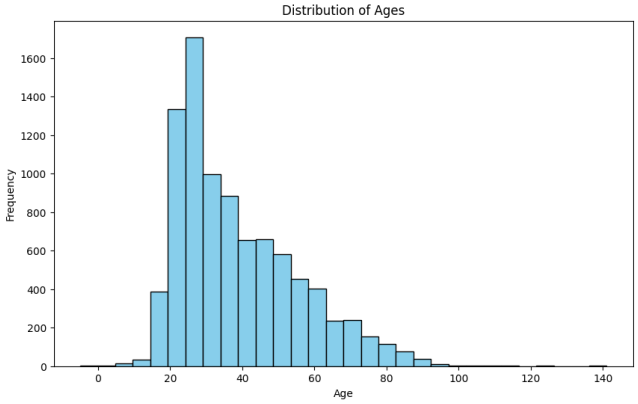

Figure 12: Age distribution (in a random 9,000 image sample).

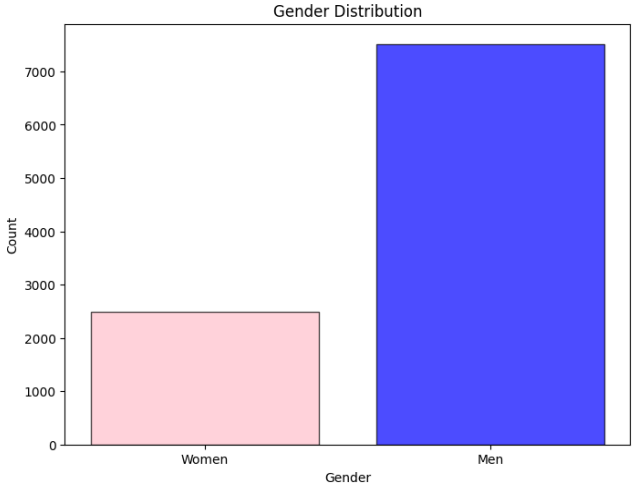

Figure 13: Gender distribution (in a random 9,000 image sample).

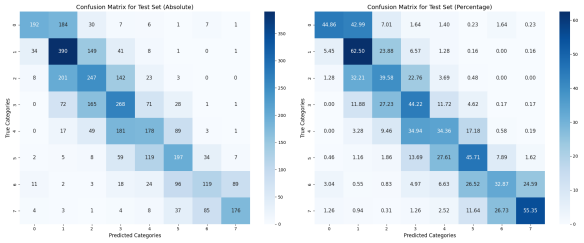

Figure 14: Test set confusion matrices for the VGG-19-based model trained on eight categories. Absolute numbers (left) and relative numbers (right)