university of groningen / faculty of science and engineering / artificial intelligence

Uncertainty Disentanglement in Face Age Estimation

Bachelor's Project Thesis

Jakub Ondrejka, s4360184, j.ondrejka@student.rug.nl,
Supervisors: Prof Dr M.A. Valdenegro Toro

Abstract: Face age estimation has been a popular problem in computer vision for quite some time, where the methods employed changed over time. Recognizing the distinction between apparent age and real age, apparent age estimation was chosen as a task for our model to perform. A recent development in machine learning is uncertainty disentanglement, crucial for further advancements in various tasks. However, its applications to face age estimation, particularly apparent age estimation, have yet to be explored. In order to study the performance of uncertainty disentanglement in apparent age estimation, we have employed DenseNet121 for feature extraction and implemented three methods for uncertainty estimation: Monte Carlo Dropout, Monte Carlo DropConnect, and Ensembles. Our results show that all three methods are capable of estimating apparent age, but struggle with providing appropriate uncertainties. All three methods struggle the most with the aleatoric uncertainty. It is worth mentioning that the Ensembles performed the best out of the three uncertainty disentanglement methods.

1 Introduction

Face age estimation represents a fascinating and complex field within computer vision, where the objective is to accurately determine a person's age based on their facial features. This field has quite a few applications in human-computer interaction, surveillance monitoring, and video content analysis, such as preventing underage accessing alcohol.

The fact that even humans have a hard time estimating real age, speaks of the challenge machines must overcome in automated age estimation. The inherent complexities come from environmental variabilities and significant intra-cohort age variance(Albert et al., 2007). This is when a group of people are born in the same generation, but age differently based on environmental factors such as smoking, drinking or stress.

Therefore, we have to clarify the difference between real and apparent age. Real age is the biological age of the person, whereas 'apparent' age is the age perceived by other people solely based on appearance only. This difference in tasks also has different possible applications. For example, real age estimation is used for age restriction(Angulu et al., 2018). On the other hand apparent age estimation can be used in applications where the human perspective is the crucial part, such as testing the effects of anti-ageing products or spotting faster rates of ageing caused by poor lifestyle(Goodyear et al., 2023).

One of the most significant problems in face age estimation is the already mentioned intra-cohort age variance. This complexity poses a formidable challenge to developing accurate models. Genetics, lifestyle choices, and health conditions contribute to a wide range of ageing expressions on the human face, making it difficult to standardize and model. Moreover, environmental influences and personal habits, such as exposure to sunlight and smoking, can accelerate or decelerate the visible signs of ageing (Morita, 2007; Kennedy et al., 2003), introducing additional layers of complexity. This calls for a sophisticated algorithm capable of capturing and interpreting the variability.

Age estimation based on facial features has different possible approaches. In the first part of the task, it is necessary to extract features from the pictures. These features can be either hand crafted, usually representing biological factors with the most impor-

tant factors including identity, gender and ethnicity(Guo et al., 2009; Bekhouche et al., 2016). Or the features can be automatically extracted from the pixels using techniques like Convolutional neural network (CNN)(Levi & Hassner, 2015; Han, 2020; K.-H. Liu et al., 2019). CNN models have become a more prevalent method because they lead to a fully learned end-to-end system that can estimate age from image pixels directly, without the need for humans to indicate and label features for the model.

The second part of the task is using the features and turning them into output. This can be divided into two categories: a classification task where a face is assigned to one of several specific age categories(N. Liu et al., 2020; Kwon & da Vitoria Lobo, 1999; Lanitis et al., 2004), or a regression task to predict a more precise age (N. Liu et al., 2020; Guo et al., 2009). In definition, doing classification may appear simpler, but results in an age range. This is sufficient enough for demographics analysis, and commercial user management. But even though regression is a bit harder it will result in an exact number, not just an age range. This is necessary for more sensitive applications, such as early spotting of environmental effects on health.

1.1 Uncertainty

Commonly, machine learning only gives output signifying the preferred answer of the model(Guo et al., 2009; Lanitis et al., 2004). This is satisfactory for a lot of applications, but with some other tasks, for example, tasks that deal with the risk of human life or have legal consequences, the black box output of the model is no longer satisfactory. Recently a solution has emerged: uncertainty quantification(Gal et al., 2016). With uncertainty quantification, the model also gives confidence in its output.

Two primary forms of uncertainties exist, aleatoric (data-related) and epistemic (model-related)(Gal et al., 2016). Although typically fused into a singular predictive uncertainty, disentangling these uncertainties provides clarity on their individual contributions. Specifically, aleatoric uncertainty highlights the inherent unpredictability of the data, suggesting that age prediction would be inherently unpredictable. In contrast, epistemic uncertainty underscores constraints of the model itself, suggesting that the model can be improved, by better architecture or with more learning opportunities.

There are quite a few uncertainty disentanglement methods available, and it is unclear which one is the best performing in which situation. We will look at three uncertainty disentanglement methods: Monte Carlo Dropout, Monte Carlo DropConnect, and Ensembles(Valdenegro-Toro & Mori, 2022). They will be explained in more detail in section 2.

1.2 Scope of This Thesis

Against this backdrop, our research is seeking to explore and expand the frontiers of age estimation technology. The central inquiry of our investigation is: "Can a neural network learn to recognize apparent human age and predict appropriate uncertainties?".

Secondly, to comprehensively assess the efficiency of each uncertainty estimation method in performing our face age estimation task, we plan to implement all three approaches: Monte Carlo Dropout, Monte Carlo DropConnect, and Ensembles. By analyzing the precision of uncertainty estimates and their impact on the overall performance and reliability of age predictions, we aim to identify the most effective method for capturing predictive uncertainty in the context of face age estimation.

By focusing on apparent age, we acknowledge the subjective nature of perceiving ageing. Moreover, by aiming to predict uncertainties, we are committing to a level of transparency and reliability that goes beyond traditional models.

This thesis contributes in three key ways. Firstly, we propose a model capable of both estimating apparent age and disentangling the aleatoric and epistemic uncertainties associated with the estimation. Secondly, this approach leverages the availability of standard deviation associated with apparent age to estimate the epistemic uncertainty of the aleatoric uncertainty prediction as well (details in section 2). Lastly, a comparison between DropOut, DropConnect and Ensembles, about their disentangling quality in the apparent age estimation task.

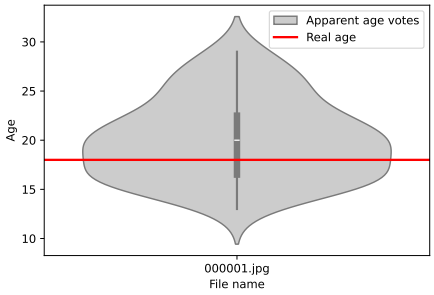


Figure 2.1: This plot shows the distribution of apparent age votes for face "000001.jpg". The red line indicates real age.

2 Methods

2.1 Dataset

For our purpose, we will be utilizing the APPA-REAL database(Agustsson et al., 2017), a pioneering dataset that offers both real and apparent age annotations for each image. This dataset emerges in response to the scarcity of publicly available databases that include face images annotated with apparent age labels contrary to the extensive availability of datasets with only "real" age data.

The APPA-REAL database comprises 7,591 images(split into 4113 train, 1500 valid and 1978 test images), enriched with nearly 300,000 votes on apparent age, resulting in an average of approximately 38 votes per image(example of voting in the figure 2.1). This extensive collection of votes per image ensures a high degree of reliability in the apparent age data, evidenced by a standard error of the mean of just 0.3. Each image in the database is labelled with both the real age and apparent age, the latter derived from the mean of the raw votes post-outlier removal. The dataset spans a diverse age range, making the dataset more representative of the real world.

Analysis of the APPA-REAL database reveals a strong correlation between real and apparent ages(figure 2.2), though with individual variances that can exceed up to 20 years. This variance is indicative of the subjective nature of age perception, influenced by factors such as lifestyle, genetics, and environmental conditions. Recognizing such variance as natural in real-world scenarios, and since we want our model to be able to reflect the uncertainty of the real world, it justifies our choice of the APPA-REAL dataset.

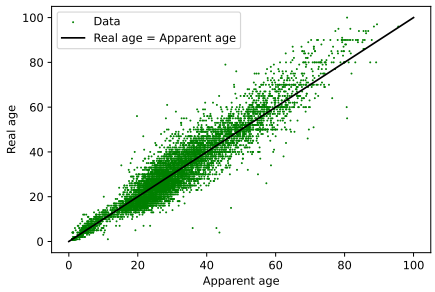


Figure 2.2: Scatterplot showing the relationship of Apparent and Real age.

Given the diverse applications of age estimation, the APPA-REAL dataset encompasses a wide range of image conditions without controlling for background and lighting. This variability reflects real-world scenarios, where facial images may be captured in uncontrolled environments, thus presenting both a challenge and an opportunity for enhancing the robustness of our age estimation model.

2.2 Preprocessing Steps

To prepare the dataset for training, we used the fact that the APPA-REAL dataset already provides images that are cropped, centered, and aligned, focusing exclusively on the face. To further augment our dataset and improve the model's generalizability to different facial orientations and lighting conditions, we applied runtime data augmentation techniques. These included rotations within +/- 10 degrees, horizontal and vertical shifts of up to 10% of the image dimensions, shear transformations, zoom adjustments ranging from 80% to 120%, horizontal flips, brightness variations between 80% and 120%, and normalizing of pixels (RGB values) from [0,255] to [0,1].

2.3 Feature Extraction

For the task of extracting meaningful features from the images, our model incorporates DenseNet121(Huang et al., 2017), a deep CNN known for its efficiency and effectiveness in handling complex image recognition tasks. DenseNet121 stands out for its unique architecture

that facilitates feature reuse, making it particularly suitable for the nuanced task of face age estimation.

The architecture comprises of 121 layers, including convolutional layers, pooling layers, and densely connected blocks that collectively contribute to the network's deep learning capabilities. The model employs concatenation of feature maps from all preceding layers at each layer, thereby enriching the feature set with minimal increase in parameters, making it both efficient and powerful.

We have decided to use DenseNet implemented by Keras(Chollet et al., 2015), because of its convenience. In addition, the library we are using for Uncertainty estimation(introduced in the next subsection) is based on Keras, which supports our decision.

2.4 Uncertainty Methods

We implement three distinct methods for uncertainty estimation: Monte Carlo Dropout, Monte Carlo DropConnect, and Ensembles(Valdenegro-Toro & Mori, 2022).

Monte Carlo Dropout involves applying dropout not just during training but also during inference, creating multiple stochastic forward passes. By treating dropout as a Bayesian approximation, it provides a distribution of outputs from which uncertainty can be quantified.

Monte Carlo DropConnect involves randomly dropping weights in the network rather than activations. This variation offers a different perspective on model uncertainty by assessing the impact of altering the network's connections.

Ensembles involve training several models independently and then aggregating their predictions, the variance among the outputs serves as a measure of uncertainty. Consequently, Ensembles differ from the Monte Carlo methods by relying on diversity across multiple trained models rather than stochasticity in a single model's forward passes.

For all three methods, we are utilizing the Keras-uncertainty library(Valdenegro, 2024).

2.5 Uncertainty Disentanglement

As mentioned before there are two types of uncertainty, aleatoric and epistemic. These are derived from the set of equations and are ultimately defined as $E_i[\sigma_i^2(x)]$ and $\text{Var}_i[\mu_i(x)]$ respectively, where $\mu_i(x)$ is mean output and $\sigma_i^2(x)$ is predictive variance with $i \in [1, M]$ being an index for different samples or ensembles.

In our case, the dataset we use also includes labels for the Std of the apparent age, which can be interpreted as aleatoric uncertainty. Therefore the model will be two-headed with one head outputting the mean apparent age and the other head outputting the aleatoic uncertainty of apparent age. This way we can learn the aleatoric uncertainty from the data instead of approximating it.

For the epistemic uncertainty we only need to calculate the Variance of the output. Since we have two outputs, we will have to calculate epistemic uncertainty for both: Epistemic uncertainty of mean apparent age - $\text{Var}_i[\mu_i(\text{apparent age})]$ and Epistemic uncertainty of std apparent age - $\text{Var}_i[\mu_i(\text{STD of apparent age})]$

2.6 Hyper-parameters and Model Configuration

The model itself differs slightly based on the uncertainty quantification method used. All three models start with DenseNet followed by 2 dense layers. After that, we fork the model, making it into two-headed. Each fork has 1 dense layer followed by a softplus output layer. For Monte Carlo Dropout and DropConnect, these dense layers are replaced by dropout and dropconnect layers which allow for the specific properties of the method.

For the ensemble approach, we trained five distinct models, 150 epochs each, capitalizing on diverse model initializations in order to enhance the robustness and accuracy of our age estimation. In contrast, for the Monte Carlo Dropout and DropConnect, we opted for a longer training duration of 200 epochs to ensure convergence, with dropout and dropconnect rates set at 0.2 and 0.05, respectively.

2.7 Model Outputs and Proposed Evaluation

Our model has four outputs(shown in the table 2.1), two of which are learned through labelled data: Mean Apparent age(μ_{AppAge}) and the Aleatoric Uncertainty of Apparent age(σ_{AppAge}). And two outputs which are calculated; these represent the epistemic uncertainties: Epistemic Uncer-

tainty of Mean Apparent age(E_μ) and Epistemic Uncertainty of Std of Apparent age(E_σ). These two types of outputs need different evaluations.

For the labelled data, we have decided to use a simple Root-mean-square-error(RMSE) as it provides a straightforward analysis of how the model performed with that output. On the other hand, RMSE might be insufficient, because it summarizes the whole model performance in one number, while it could be performing differently on different age ranges. Therefore we will also look at the plot of predictive output versus label, to see how the trend of error looks throughout different age categories.

For the other two outputs representing the epistemic uncertainties, we will calculate the calibration curve and its corresponding error to see if the model is overconfident, underconfident or balanced. The calibration curve will also be used to look at how the uncertainty evolves with accuracy. Furthermore, we will look at the Epistimic uncertainty versus absolute error plot, to again see how well the model predicts the uncertainty and the trend of the Epistimic uncertainty in regards to the errors the model makes.

Lastly, we will look case by case at specific pictures, representing the best and worst at the respective outputs and uncertainty estimation methods. This way we can discuss why the model performs better or worse on which types of data.

3 Results

3.1 Descriptive Results

For μ_{AppAge}, the ground truth average is 33.09 years, with predictions closely aligning: DopConnect at 28.51 years, DopOut at 29.73 years, and Ensembles at 32.25 years. Although all the methods are closely aligned, with the Ensembles method yielding the closest average to the truth, the range of predicted ages (min: 1.72 to 1.79 years, max: 75.00 to 89.95 years) across methods shows a smaller range than the ground truth (min: 0.95, max: 95.57 years), indicating a potential limitation in capturing the full spectrum of ages accurately.

In terms of σ_{AppAge}, the ground truth's average is 4.84. The methods' predicted averages for σ_{AppAge} are somewhat lower (DopConnect: 1.86, DopOut: 1.92, Ensembles: 2.05), suggesting a conservative estimation of uncertainty compared to human judgment variability. The ranges of predicted σ_{AppAge} (min: 0.08 to 0.72 years, max: 2.79 to 3.00 years) across methods show a smaller range than the ground truth (min: 0.00, max: 13.59 years), indicating that all three methods have limitations in estimating the higher side of the spectrum of the aleatoric uncertainty.

For E_μ Ensembles exhibit the broadest range of epistemic uncertainty (min: 0.29, max: 21.20) compared to DropOut (min: 0.90, max: 12.35) and DropConnect (min: 0.22, max: 13.65). This suggests that Ensembles potentially offer a more nuanced understanding of prediction confidence across age estimates. For E_σ it is the same, with Ensembles exhibiting the broadest range (min: 0.03, max: 0.44), compared to DopOut () and DopConnect (min: 0.01, max: 0.42).

Based on just descriptive results, we can estimate that Ensembles are our best method, because of the higher ranges in all four outputs, Ensembles are capable of a broader spectrum of values than the other two methods. This is not a definitive answer as the broader spectrum might be a negative thing as well. Nevertheless, looking at the μ_{AppAge} and σ_{AppAge}, Ensembles are closer to the range the ground truth shows.

3.2 Mean Apparent Age and Aleatoric Uncertainty of Apparent Age

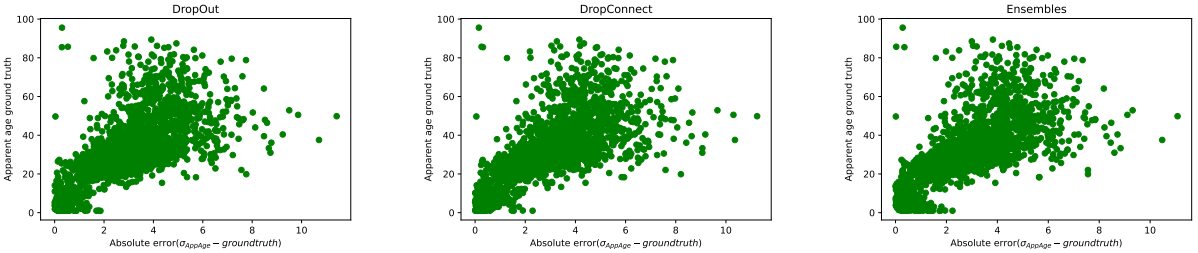
Ensembles exhibited superior performance with the lowest RMSE of μ_{AppAge} (table 3.1), indicating more accurate age predictions compared to Monte Carlo Dropout and Monte Carlo DropConnect. The baseline method had a significantly higher RMSE than all the methods, suggesting that the model was able to learn the apparent age, with Ensembles performing the best.

The RMSE values of σ_{AppAge} were closely matched across the three methods, with Ensembles slightly outperforming both Monte Carlo Dropout and Monte Carlo DropConnect. The baseline has a significantly lower RMSE of σ_{AppAge} than any of the methods, suggesting the model was not able to learn the aleatoric uncertainty from the data.

Furthermore, we can analyse scatter plots for σ_{AppAge} predictions versus the ground truth(figures

Name	Aleatoric Uncertainty	Epistemic Uncertainty	Ground truth	Notation
Mean Apparent age	No	No	Apparent age label	μ_{AppAge}
Aleatoric Uncertainty of Apparent age	Yes	No	Std of Apparent age label	σ_{AppAge}
Epistemic Uncertainty of Mean Apparent age	No	Yes	NA	E_{μ}
Epistemic Uncertainty of Std of Apparent age	No	Yes	NA	E_{σ}

Table 2.1: Caption



(a) DropOut - Pearson coefficient of 0.674

(b) DropConnect - Pearson coefficient of 0.673

(c) Ensembles - Pearson coefficient of 0.685

Figure 3.1: Scatter plot illustrating the apparent age versus absolute error of σ_{AppAge}, using the DropOut(a), DropConnect(b) and Ensembles(c). The data points suggest a trend where the error increases with the apparent age of the individuals.

RMSE:	μ_{AppAge}	σ_{AppAge}
dropout	12.272	3.471
dropconnect	12.968	3.499
ensembles	11.101	3.389
baseline	17.674	2.096

Table 3.1: Root-mean-square error for both labelled outputs(mean age and σ_{AppAge}), comparing the three uncertainty estimation methods and baseline

3.1a, 3.1b and 3.1c). We can deduce interesting information about the performance of our three methods: with increasing apparent age, the error in predicting aleatoric uncertainty increases. We have done a Pearson correlation test, yielding high positive correlations of 0.673 up to 0.685 This might suggest that the model has an easier time estimating younger faces.

3.3 Epistemic Uncertainty of Mean Apparent Age

Looking at the regressor calibration curve for E_{μ} (figure 3.2) we can see the balance between accuracy and confidence of the model with the three different methods. All three methods are overconfident, meaning they give low uncertainty values for how inaccurate they are. Ensembles are the best of the three options with a calibration error of 0.157. DropConnect is second best with the calibration error of 0.237. DropOut comes third with the calibration error of 0.298, almost double the value of Ensembles.

Looking at scatter plots of E_{μ} over the absolute error of predictive μ_{AppAge} and ground truth, we would want to see an increase in Epistemic Uncertainty as the absolute error increases. Ensembles, as indicated in the figure 3.2, did best, this is also supported by the figure 3.3c where we can see an

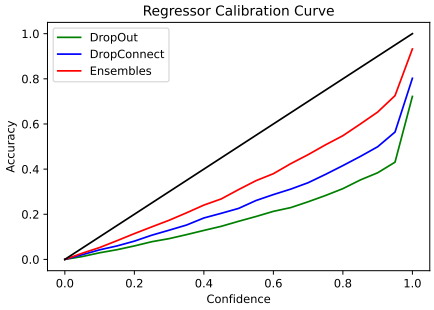


Figure 3.2: RCC of the E_μ for the three uncertainty estimation methods.The green line represents DropOut, the blue line represents DropConnect and the red line is Ensembles. The black line represents a perfectly calibrated model.

increasing trend for the Epistemic Uncertainty with the increasing absolute error. On the other hand, we have DropOut which did worse on RCC(figure 3.2), and we can also see that in figure 3.3a the Epistemic Uncertainty is not increasing nearly as fast enough as in Ensembles. Looking at the Dropconnect in figure 3.3c, we see a middle ground between Ensembles and Dropout, just like in figure 3.2. We also calculated the Pearson correlation coefficients, and we got a medium positive correlation for Ensembles (0.304), and low or negligible correlations for DropOut(0.133) and DropConnect(0.025). The Pearson correlation coefficients suggest only Enseblems is able to estimate somewhat appropriate E_μ.

3.4 Epistemic Uncertainty of Std of Apparent Age

We can do the same analysis for E_σ as for E_μ. Starting with RCC in figure 3.4 we can see that all three methods are not calibrated well, they are all overconfident. Both DropConnect and DropOut have an almost vertical line and calibration error of 0.471 and 0.4767 respectively. This suggests that both methods give low values for E_σ for when they are correct and when they are not, no-discriminatively. On the other hand, Ensembles does a bit better than the other two, with the calibration error of 0.388. We can also look at the corresponding scatter plots of E_σ over the absolute error of predictive σ_{AppAge} and ground truth. We can see that both figures 3.5a and 3.5b give very low uncertainty values, even with increasing absolute error. We can see the same behaviour from Ensembles in figure 3.5c, with the difference that on average the uncertainty values are higher. Calculating the Pearson correlation coefficients we get negative correlations for DropOut(-0.276) and DropConnect(-0.331), suggesting that both methods are not estimating the E_σ correctly. For Ensembles, we get no correlation(-0.021), which again suggests the inability of Ensembles to estimate E_σ. Also, it is worth mentioning that no correlation is better than a negative correlation, making the Ennsebles best method.

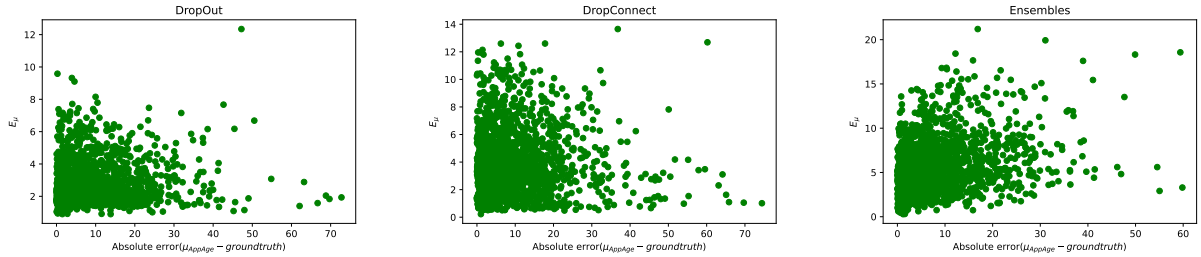
3.5 Example Analysis

In this section, we will go over a few examples from the test split, and show the results of the three uncertainty disentanglement methods. The examples that were selected show 2 images with higher apparent age(figures 3.6b and 3.6c) and 2 images with lower apparent age(figures 3.6a and 3.7a). Also figures 3.6a and 3.6b are examples of when the methods estimate μ_{AppAge} correctly and incorrectly, in that order. Furthermore, figures 3.6c and 3.7a are examples of well-estimated and poorly estimated Epistemic uncertainties, in that order.

Comparing the examples against each other we can notice the previously mentioned problem with higher errors in aleatoric uncertainty as the apparent age increases. Namely figures 3.6a and 3.7a have much closer prediction of σ_{AppAge} than figures 3.6b and 3.6c.

Figure 3.6a shows an example where all three of the methods perform well. We can see that DropOut and Ensembles both have predicted the μ_{AppAge} almost perfectly. DropConect is a bit off on the μ_{AppAge}, but not too far, as the E_μ makes up for the difference. The E_μ DropOut and Ensembles are too high for how precisely it got the answer. Next looking at the σ_{AppAge}, we see that all three methods give a lower value than the ground truth. When we take into consideration E_σ, the σ_{AppAge} prediction does not get any better as the E_σ values are too low. Here Ensembles perform best, giving the highest E_σ, whereas DropOut and DropConnect give negligible values.

Figure 3.6b shows an example where all three of the methods perform poorly. What we can notice



(a) DropOut - Pearson coefficient of 0.133

(b) DropConnect - Pearson coefficient of 0.025

(c) Ensembles - Pearson coefficient of 0.304

Figure 3.3: Scatter plot showing E_μ versus absolute error of the σ_{AppAge} and the corresponding ground truth, for the DropOut, DropConnect and Ensembles.

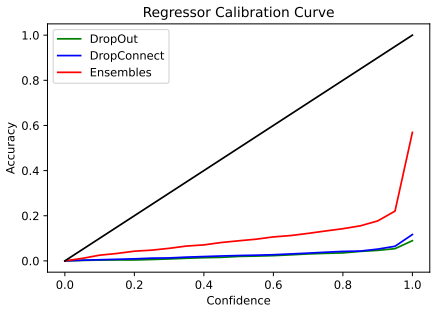
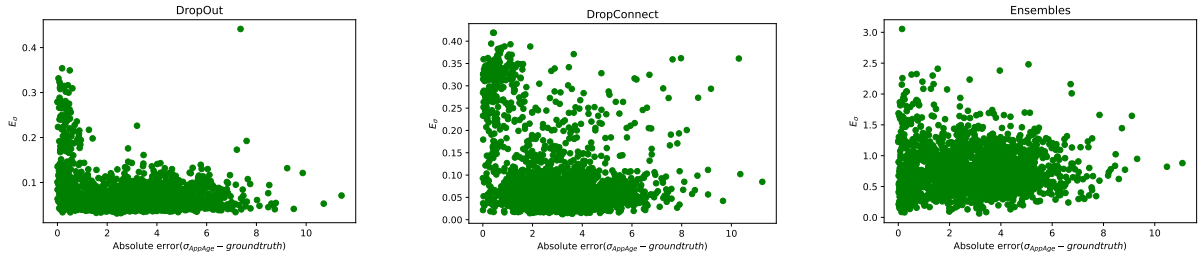


Figure 3.4: RCC of the E_σ for the three uncertainty estimation methods. The green line represents DropOut, the blue line represents DropConnect and the red line is Ensembles. The black line represents a perfectly calibrated model.

right away is that all three methods have an error of 20+ years in μ_{AppAge}. Secondly, all three methods give the wrong value for the σ_{AppAge}. This example seems to be one of the hardest for humans to estimate, indicated by the large Std of Apparent age ground truth, but all three methods have failed to give appropriate uncertainties. The already mentioned σ_{AppAge} and both of the Epistemic uncertainties are undervalued. On the other hand, it is necessary to mention that even though the values are low for the uncertainties, Ensembles gives the highest ones compared to the other two.

Looking at figure 3.6c, all three of our methods have failed to estimate μ_{AppAge} and σ_{AppAge} correctly. But what seems to be interesting in this example is the Epistemic uncertainties. For both E_μ and E_σ the Ensembles perform well. We can see that for both μ_{AppAge} and the σ_{AppAge}, the ground truth is within the corresponding Epistemic Uncertainty error bars. This is only the case for the Ensembles, both DropOut and DropConnect estimate low values of Epistemic Uncertainties. This example suggests that even in case of a bad prediction, Ensembles can provide higher values of uncertainty, compared to the other two methods. The example provided in figure 3.7a, is somewhat of the opposite of figure 3.6c. Firstly, all three methods perform well on predicting μ_{AppAge} and the σ_{AppAge}, except for Ensembles, which perform slightly worse on the μ_{AppAge} estimation. What is interesting, is how the methods perform on the Epistemic Uncertainty part. We can see that both DropOut and DropConncet give the usual small values of Epistemic Uncertainties, which seems to be alright in this case, as they got the μ_{AppAge} and σ_{AppAge} estimation correctly. On the other hand, Ensembles give their usual higher values of Epistemic Uncertainties, which is counterproductive. Since Ensembles did make a decent μ_{AppAge} and σ_{AppAge} estimation, the high values of Epistemic Uncertainties are uncalled for. This example suggests that the tendencies of the methods(low versus high Epistemic Uncertainty for DropOut, DropConncet versus Ensembles) are sometimes making the methods look like they perform better(DropOut, DropConncet) or worse than they usually do(Ensembles).

Looking at figure 3.7b, we can see that all three of our methods perform badly on all aspects of the task. Predicted μ_{AppAge} is off by 10 up to 30 years, and σ_{AppAge} is also predicted with too low values. Looking at the Epistemic uncertainties, we see the methods do not reflect the fact that the model is

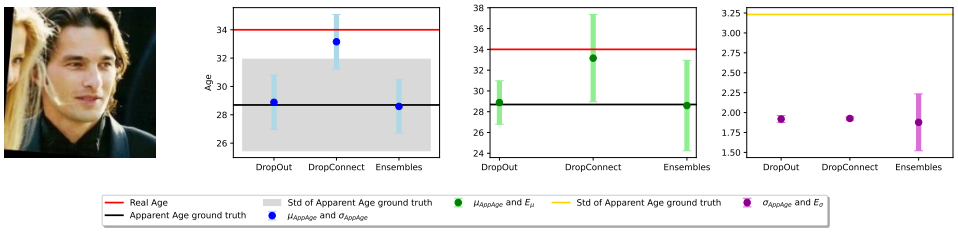


(a) DropOut - Pearson coefficient of -0.276

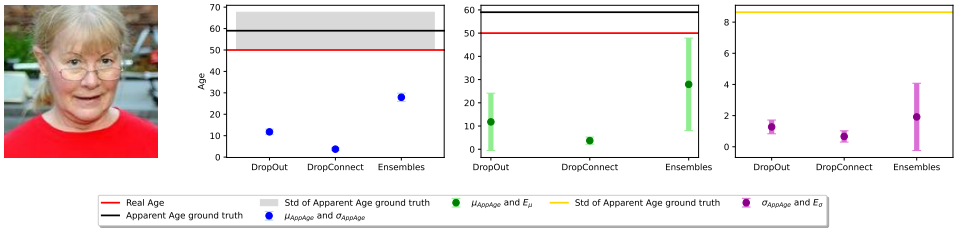
(b) DropConnect - Pearson coefficient of -0.331

(c) Ensembles - Pearson coefficient of -0.021

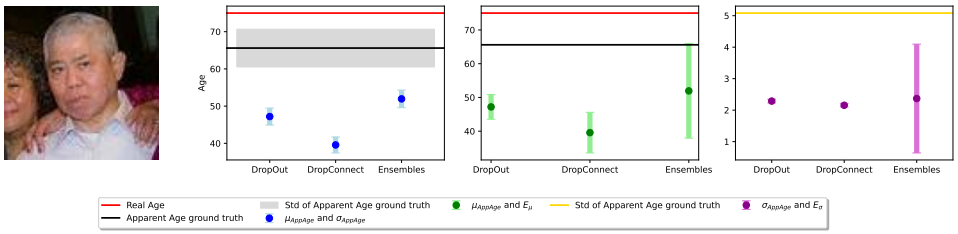
Figure 3.5: Scatter plot showing E_σ versus absolute error of the σ_{AppAge} and the corresponding ground truth, for the DropOut, DropConnect and Ensembles.



(a) Facial image 006450, is an example of input the model does well on.



(b) Facial image 006416, is an example of input the model struggles with.

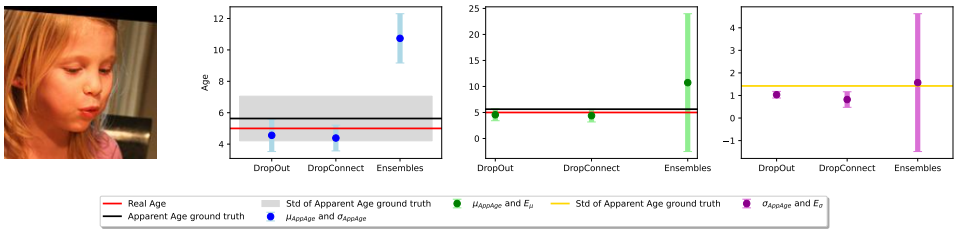


(c) Facial image 007185.

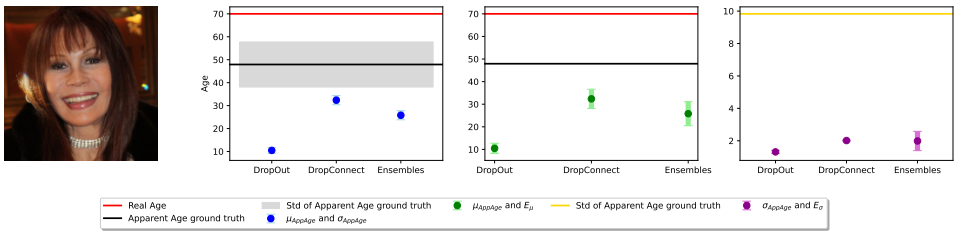
Figure 3.6: The three graphs show results from the three uncertainty disentanglement methods. The first graph compares the apparent age and Std of Apparent age ground truths to the predicted μ_{AppAge} and σ_{AppAge}. The second graph compares the apparent age ground truth to $\mu_{AppAge} + E_\mu$. The third graph compares the Std of apparent age ground truth to $\sigma_{AppAge} + E_\sigma$.

struggling with the μ_{AppAge} and σ_{AppAge}. This example once again shows the weakness of the model in estimating well faces of higher apparent age.

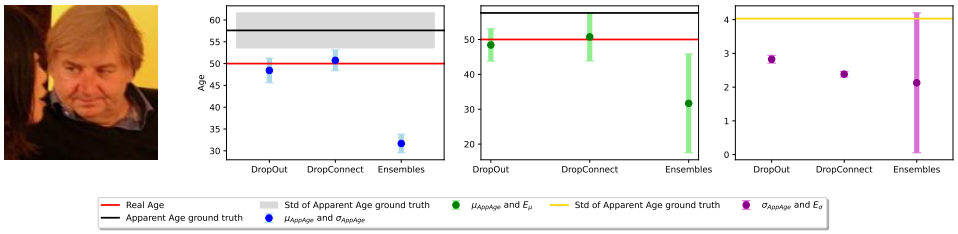
Lastly, figure 3.7c, shows a good example of Ensembles able to give higher values of Epistemic uncertainty in some of the cases when its predictions



(a) Facial image 007483.



(b) Facial image 007088.



(c) Facial image 006555.

Figure 3.7: The three graphs show results from the three uncertainty disentanglement methods. The first graph compares the apparent age and Std of Apparent age ground truths to the predicted μ_{AppAge} and σ_{AppAge}. The second graph compares the apparent age ground truth to $\mu_{AppAge} + E_{\mu}$. The third graph compares the Std of apparent age ground truth to $\sigma_{AppAge} + E_{\sigma}$.

of μ_{AppAge} and σ_{AppAge} are off. On the other hand, DropOut and DropConnect do both well on the μ_{AppAge} and E_{μ} prediction but once again fail the σ_{AppAge} and E_{σ}.

4 Discussion

In this thesis, we have defined two questions. Firstly, can neural networks learn to recognize apparent human age and predict appropriate uncertainties? We have looked into the apparent age estimation and uncertainty disentanglement with three methods: Monte Carlo Dropout, Monte Carlo DropConnect and Ensembles. This leads us to the second question, which of these three methods would perform the best? The results showed that all of the methods are capable of estimating apparent human age, but they all struggle with giving appropriate uncertainties. Namely, all three methods we not able to predict aleatoric uncertainty, and also the corresponding epistemic uncertainty. On the other hand, all three methods did estimate the epistemic uncertainty of the mean apparent age a bit better, but they were still unsatisfactory. Finally, the results also clearly showed Ensembles performing the best across all the evaluations, suggesting a potential pathway for future research to refine and build upon.

4.1 Limitations

The main limitation of our research is the model itself. The research was mostly focused on uncertainty disentanglement, and hence the basic age es-

timation has not been built to reach its full potential. The architecture of the model is relatively simple, compared to the newest competitive face age estimation models. Using a more sophisticated model would likely lead to better results, at least in the apparent age estimation part. However, it is unclear if this would also boost the performance of the uncertainty estimation parts.

Furthermore, we assume that the data is following Gaussian distribution, meaning the votes on each data point come together to make a normal distribution. And since all three of our methods' outputs follow the normal distribution, it might fail if not all of the data points follow the Gaussian distribution, which we did not control for.

That leads us to the third limitation, estimating aleatoric uncertainty is a difficult task, and there is more than one way of estimating it. We are using direct supervision as we hoped that would make it easier. This might not be the best way to estimate it.

4.2 Future Research

Given the limitations identified in our study, future research should focus on those first. Firstly, building a more sophisticated up-to-date model and then integrating the uncertainty disentanglement methods on top of it. Secondly, other distributions that might model the dataset better, could be explored, such as the Gaussian mixture model. In case the dataset has a different distribution, we would have to tweak the three methods to take this into consideration, and also follow the specific distribution. Lastly, trying other ways of estimating aleatoric uncertainty, to try to enhance the performance.

Furthermore, in our thesis, we only look at estimating apparent age. This could be expanded to real age as well. Although real age and apparent age are correlated with each other, making a multitask model estimating apparent age and real age as well, could lead to better overall performance.

References

Agustsson, E., Timofte, R., Escalera, S., Baro, X., Guyon, I., & Rothe, R. (2017). Apparent and real age estimation in still images with deep residual regressors on appa-real database. In *2017 12th ieee international conference on automatic face & gesture recognition (fg 2017)* (pp. 87–94).

Albert, A. M., Ricanek Jr, K., & Patterson, E. (2007). A review of the literature on the aging adult skull and face: Implications for forensic science research and applications. *Forensic science international*, *172*(1), 1–9.

Angulu, R., Tapamo, J. R., & Adewumi, A. O. (2018). Age estimation via face images: a survey. *EURASIP Journal on Image and Video Processing*, *2018*(1), 1–35.

Bekhouche, S. E., Ouafi, A., Taleb-Ahmed, A., Hadid, A., & Benlamoudi, A. (2016). Facial age estimation using bsif and lbp. *arXiv preprint arXiv:1601.01876*.

Chollet, F., et al. (2015). *Keras*. https://keras.io.

Gal, Y., et al. (2016). Uncertainty in deep learning.

Goodyear, K., Saffari, P. S., Esfandiari, M., Baugh, S., Rootman, D. B., & Karlin, J. N. (2023). Estimating apparent age using artificial intelligence: Quantifying the effect of blepharoplasty. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, *85*, 336–343.

Guo, G., Mu, G., Fu, Y., & Huang, T. S. (2009). Human age estimation using bio-inspired features. In *2009 ieee conference on computer vision and pattern recognition* (pp. 112–119).

Han, S. (2020). Age estimation from face images based on deep learning. In *2020 international conference on computing and data science (cds)* (p. 288-292). doi: 10.1109/CDS49703.2020.00063

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4700–4708).

Kennedy, C., Bastiaens, M. T., Willemze, R., Bavinck, J. N. B., Bajdik, C. D., & Westendorp, R. G. (2003). Effect of smoking and sun on the aging skin. *Journal of investigative dermatology*, *120*(4), 548–554.

Kwon, Y. H., & da Vitoria Lobo, N. (1999). Age classification from facial images. *Computer vision and image understanding*, *74*(1), 1–21.

Lanitis, A., Draganova, C., & Christodoulou, C. (2004). Comparing different classifiers for automatic age estimation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *34*(1), 621–628.

Levi, G., & Hassner, T. (2015). Age and gender classification using convolutional neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 34–42).

Liu, K.-H., Liu, H.-H., Pei, S.-C., Liu, T.-J., & Chang, C.-T. (2019). Age estimation on low quality face images. In *2019 ieee international conference on artificial intelligence circuits and systems (aicas)* (pp. 295–296).

Liu, N., Zhang, F., & Duan, F. (2020). Facial age estimation using a multi-task network combining classification and regression. *IEEE Access*, *8*, 92441–92451.

Morita, A. (2007). Tobacco smoke causes premature skin aging. *Journal of Dermatological Science*, *48*(3), 169-175. Retrieved from https://www.sciencedirect.com/science/article/pii/S0923181107003210 doi: https://doi.org/10.1016/j.jdermsci.2007.06.015

Valdenegro, M. (2024). *keras-uncertainty.* https://github.com/mvaldenegro/keras-uncertainty. GitHub.

Valdenegro-Toro, M., & Mori, D. S. (2022). A deeper look into aleatoric and epistemic uncertainty disentanglement. In *2022 ieee/cvf conference on computer vision and pattern recognition workshops (cvprw)* (pp. 1508–1516).