Age Prediction with Image and Numerical Features- Machine Learning – Final report

Benjamin Philipose

ECEGR4750 – Introduction to Machine Learning

Fall 2023

12/8/2023

Introduction

This project aims to develop a model capable of estimating a person's age using either facial images, numerical input features, or a combination of both. I will be classifying it as a regression task due to age being a continuous target variable. Our approach utilizes the UTKFace dataset, which comprises of eight primary features: filename (face images), gender, race, number of haircuts in life, TikTok usage, remembering disco era, skincare habits, and maximum annual earnings. We will evaluate and compare the performance of three distinct models: a basic Linear Regression Model, a Convolutional Neural Network (CNN), and a Multi-Modal Neural Network that processes both image and numerical inputs.

Bias Data Analysis

When conducting data analysis for an age prediction model, both the categorical and numerical features were checked to identify any potential biases. Biases in this context refer to any disproportional representation or systematic skew within the data that could potentially lead to a model misinterpreting or overfitting to specific patterns. This means that biases can adversely impact the model's ability to generalize and perform equally across diverse data samples. Upon a close examination of the data distributions, illustrated in Figures 1 and 2, it became evident that the features "Race," "TikTok," and "Disco" showed the most significant potential for bias.

Race Bias

The "Race" feature was dominated by the 'white' category (Figure 1), constituting nearly half of the dataset. This uneven distribution risks a model that is predominantly trained on and, consequently, predicts better for white individuals ages, while probably underperforming for other racial groups. The imbalance is especially pronounced for the 'Asian' and 'Other' categories, which encompass a broad array of ethnicities that are not as well-represented, therefore ‘race’ was removed.

Disco Bias

For the "Disco" feature, it was noticed that this could cause a strong cultural and generational bias, with a majority indicating no memory of disco (Figure 1). This suggested a model predisposition towards younger age groups, potentially leading to inaccuracies in age prediction for older individuals, who may have experienced the disco era. Considering that disco was very popular within American culture, and assuming that the age prediction model would have no cultural bias, and would be aimed for international application, "Disco" was deemed unsuitable for a global age prediction model and subsequently removed.

Max Annual Earnings Bias

Similarly, "Max Annual Earnings" displayed an extreme bias with an overwhelming majority of the data points indicating low earnings (Figure 2). This feature was problematic for two reasons: first, low earnings could disproportionately represent individuals from impoverished backgrounds or areas. Secondly, annual earnings are not a reliable age indicator due to economic differences across geographical regions, such as varying cost of living. To avoid these socioeconomic biases and ensure the model's focus on more direct age-related features, "Max Annual Earnings" was also excluded from the dataset.

TikTok Bias

The "TikTok" feature exhibited a significant skew as well, with most individuals not using TikTok (Figure 1). However, given the platform's global reach and relevance across various age groups, it was decided to retain this feature. To attempt to reduce any associated bias, stratified sampling was utilized during the train-test-validation split, ensuring that each subset of data was representative of the overall usage distribution of TikTok. This approach aimed to preserve the model's learning process by maintaining a balanced representation of TikTok users and non-users across all data splits.

In summary, the removal of "Disco," "Race," and "Max Annual Earnings" was a deliberate choice to minimize cultural, racial, and socioeconomic biases, respectively. By doing so, it would help refine the model's focus on features that provide a globally relevant and minimized biased estimation of age. For the "TikTok" feature, where the majority of the dataset's individuals are non-users, stratified sampling was implemented to ensure balanced model training and validation, which help to contribute to a goal of an unbiased model's performance in real-world applications globally.

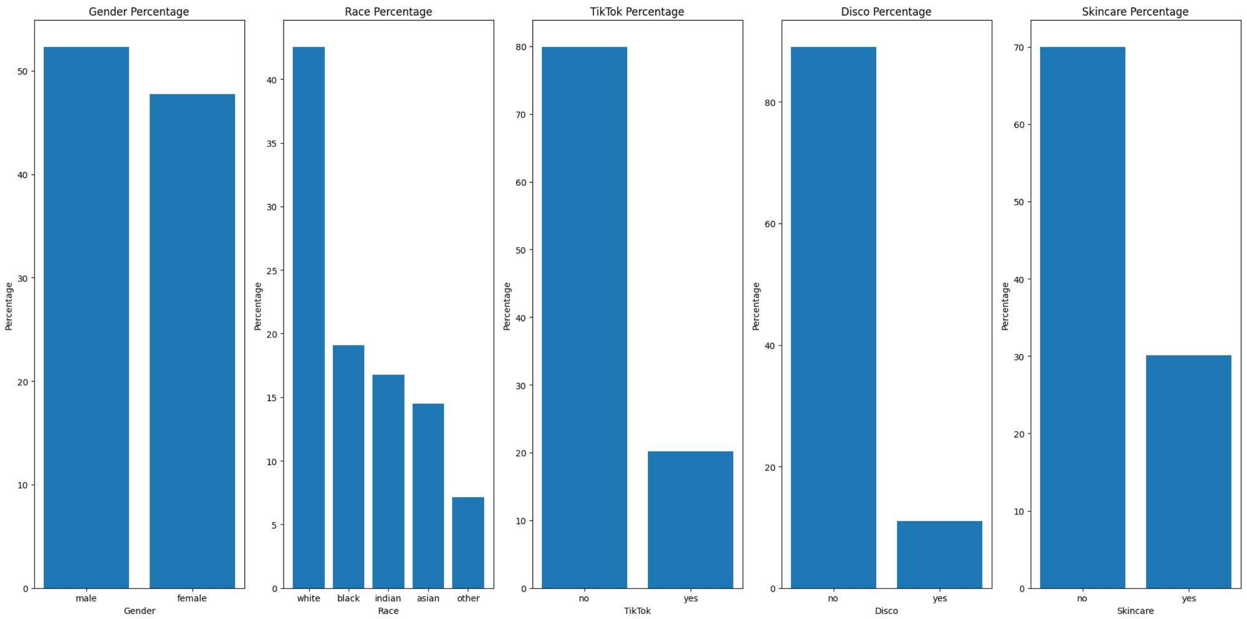

Figure 1 - Categorical Features Data Distribution

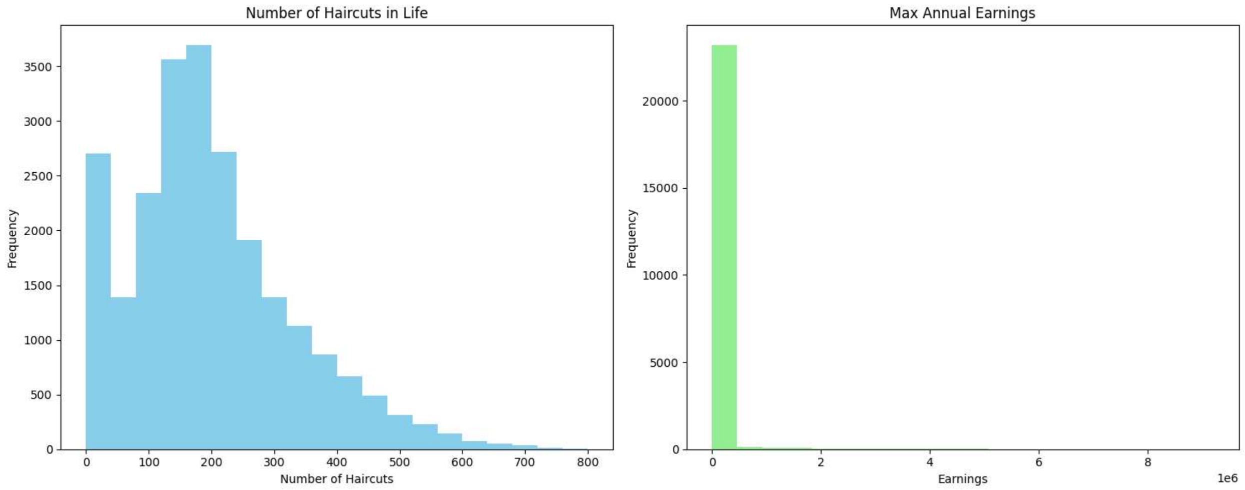

Figure 2 - Numerical Features Data Distribution

Dataset Processing

For the Dataset Processing stage, these are the features that were selected for this step: ‘Number of Haircuts Lifetime’, ‘TikTok’, ‘Skincare’, ‘gender’, Images.

Image Data Preprocessing

For image data preprocessing in our project, the images were processed in batches as they were streamed from the disk. Each image, once loaded into the data loader, was first resized to a uniform dimension of 64x64 pixels. This resizing ensured consistent input sizes for the model. Next, the images were converted to grayscale, which simplified the data input by reducing the requirement to just one layer for recording pixel brightness values, as opposed to three layers needed for RGB values. After conversion, the grayscale pixel brightness values, ranging from 0 to 255, were normalized to maintain consistent scaling across all images, and standardize image data for the model. The image data was not preprocessed more than

Continuous Numerical Features

In the continuous numerical features category, specifically for 'Number of Haircuts Lifetime', we applied standardization only in the linear regression model, utilizing the Standard Scaler. This method normalized the data into a uniform distribution, ensuring no single data point disproportionately influenced the model. After further testing, this standardization was not implemented in the multimodal neural network data, as it was noted that non-standardized haircut data improved age prediction accuracy. The potential reasoning for this is that non-normalized feature is a high correlated feature to age, and thus by not standardizing the data, it starts the model with greater initial weights, which initializes the model to have high emphasis on that feature to predict age. Note this difference in data processing between the linear regression model and the multimodal neural network was the only difference between the two model’s numerical data preparation.

Discrete Numerical Features

In the data processing phase, Gender, TikTok, and Skincare were categorized as categorical features. TikTok and Skincare, each with two classes - 'yes' and 'no' - were represented using binary encoding (0 for 'no', 1 for 'yes'). This approach avoided the need for one-hot encoding for these binary features. However, Gender, a potential multi-class categorical feature, was processed using one-hot encoding. Although initially including only male and female categories, the processing was designed without assuming exclusively biological genders, allowing the possibility of incorporating a broader range of gender classifications in the future.

Age Regression V.S. Age Classification

The data preprocessing steps in our study were designed with the objective of approaching the Age Prediction task as a regression problem. The advantage of a regression approach is its ability to provide more precise age predictions. However, this precision comes at the cost of increased sensitivity to variations in the input age data, which can potentially lead to inaccuracies. On the other hand, the classification approach simplifies the problem by categorizing ages into ranges. While this can allow for a wider range of error, it sacrifices the level of detail and precision achievable with a regression model.

Models

The 3 models that are tested for the age prediction task are Linear Regression, Convolutional Neural Networks, and lastly Multi-Modal Neural Network.

Linear Regression Model:

Model Architecture

The model predicts age based on a linear relationship between the inputs (which are only numerical input features, such as ‘Number of Haircuts Lifetime’) and the target age. The core idea behind this model is to establish a linear relationship between the input feature and age, as shown in Figure 3. Unlike more complex models, the Linear Regression Model doesn't involve sophisticated layers or activations. It directly maps the input to the output, treating age prediction as a straightforward regression problem.

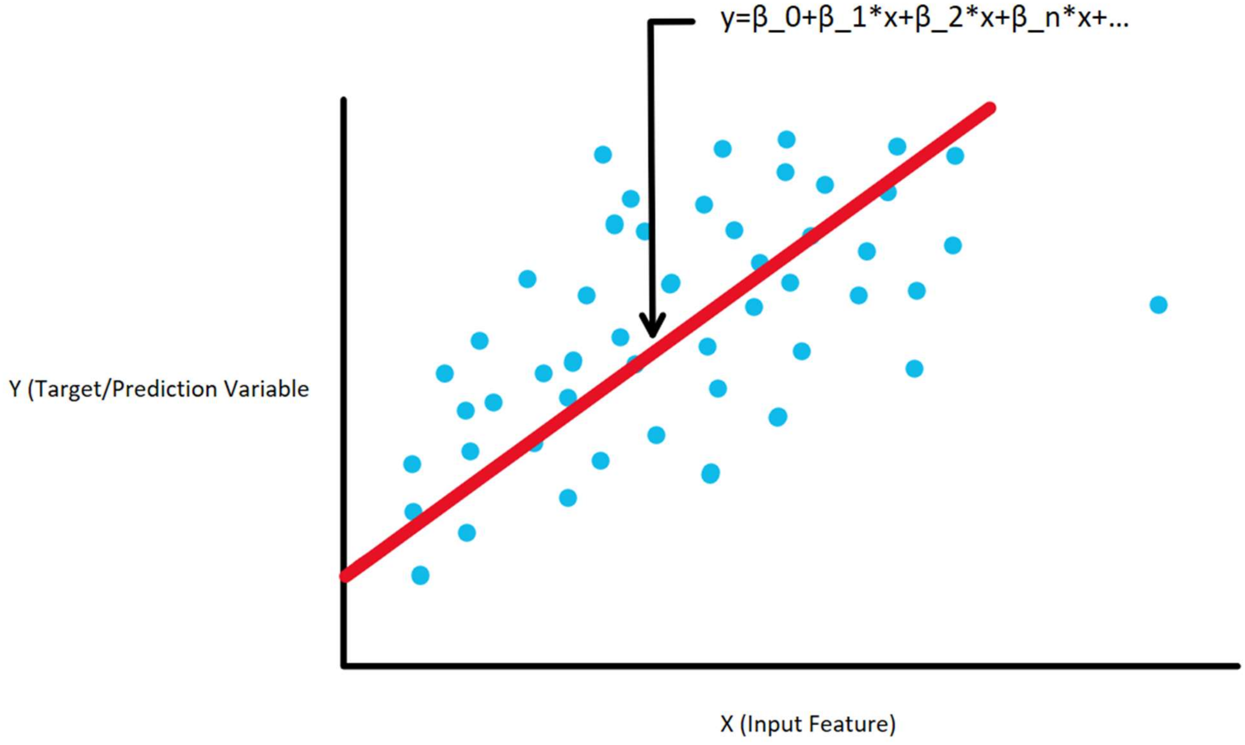

Figure 3 - Linear Regression Model Analysis

Hyperparameter Tuning

For Linear Regression, when performing hyperparameter tuning, the two most important features when training the model were epochs and learning rate. Learning rate dictates how much the model weights are updated, and epochs represent how many times the model goes through the dataset completely. Ensuring that if there is a low learning rate of 1e-5, then there needs to be enough epochs for the training and validation loss to flatten out. For my hyperparameters, I noticed that the smaller the learning rate, the better the loss lowered, so I set my learning rate to 1e-4, and set epochs to be 100, as it gave plenty of epochs for the training and validation curves to flatten out (Figure 4), and because training this simple model was degrees quicker than training next two models. For the Linear Regression model, the main challenge was figuring out when to stop training the model, even if the model didn't start to overfit at all with separation of the validation and training loss lines. Due to the model's quick loss drop at around 30 epochs, it was difficult to choose whether to continue and train the model in hopes that the loss would continue to drop even more down the line or stop at 100 epochs. This was tested a few times, once with 10,000 epochs, and another with 10. The 10,000 epochs didn't show any signs of a lower loss drop, so there was not much reason to continue to train, as the model had found its best regression line formula. 10 epochs on the other hand were too little epochs, and it didn't allow the model to have enough chances to re-tweak the linear regression equation and allow it to let the training and validation loss plateau.

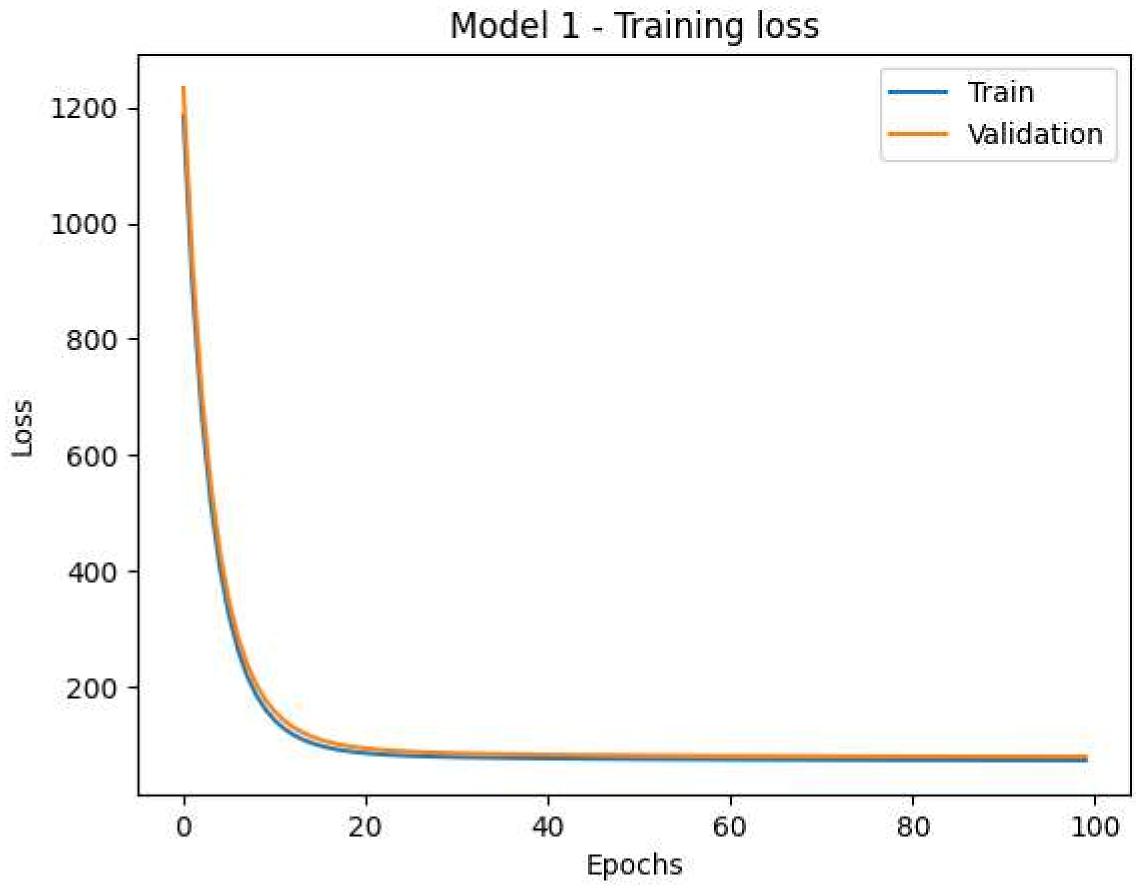

Figure 4 - Linear Regression Final Train/Validation Loss Graph

CNN (Convolutional Neural Network):

Model Architecture

The second model, Convolutional Neural Network (CNN), is tailored for image data processing and feature extraction. The model takes grayscale facial images, resized to 64x64 pixels, as inputs and leverages convolutional layers (Conv2d), batch normalization (BatchNorm2d), max pooling (MaxPool2d), and fully connected layers (linear) to predict a person's age based on facial features as shown in Figure 6. The convolutional layers apply filters to the inputs, shown in Figure 5. The activation functions are utilized typically after the convolutional to determine if neuron should be activated. The pooling layers reduce the dimensionality of the feature maps. The Fully connected layers (linear layers) process the features to make predictions. This model is developed to focus on capturing patterns and relationships within facial images, making it useful for age prediction tasks that involve image data. Note that I loosely modeled my CNN architecture after the study "Age Prediction using Image Dataset using Machine Learning" by Verma et al. (2019) [1], as they achieved good results on a similar age prediction task. Like their model, mine also uses three convolutional layers and two fully connected layers.

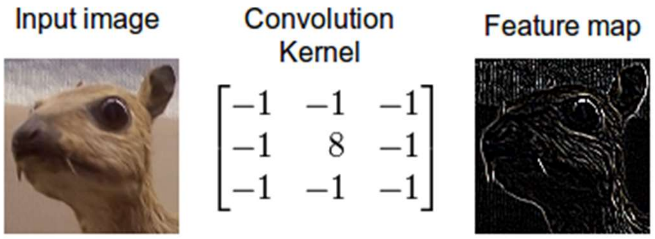

Figure 5 - Convolution Kernel Example [https://developer.nvidia.com/discover/convolution]

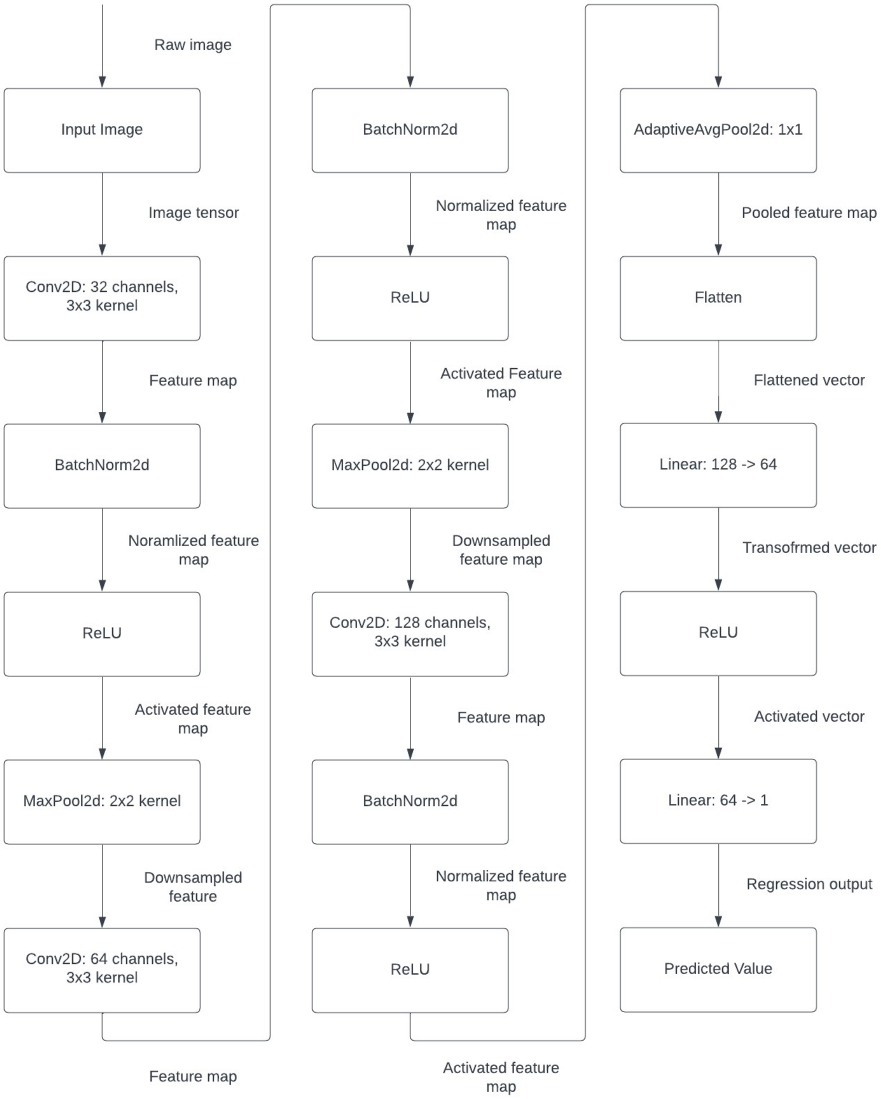

Figure 6 - Convolutional Neural Network architecture

Hyperparameter Tuning

The main hyperparameter I had tuned for the CNN model was the batch size, which dictates how many data points the model goes over before adjusting its weights accordingly. Batch size is important as it sets the number of training data samples processed before the model's weights are updated. A larger batch size can provide a more accurate gradient estimate, but it also requires more memory for storing the image data. This trade-off is evident in Figure 7. Despite a low learning rate of 1e-4 and a relatively high batch size of 100, there were notable spikes in loss at around 2.5, 7.5, and 19.5 epochs. These spikes were caused by the presence of outliers in certain batches, which influenced the weight updates negatively.

Initial hyperparameter testing resulted in significant spikes in the loss, attributed to a smaller batch size that made the model more sensitive to outliers. After several retraining attempts (4-7 retraining's), it became clear that a smaller batch size led to more chaotic kernel weight updates due to these outliers being the majority trained in some of the small batches. This testing made me realize that a balance needed to be set by having sufficiently large batch sizes to smooth over the outlier's impact, and not too large batches which would potentially exceed memory capacity. This balance between the two I found was important in optimizing the model's performance.

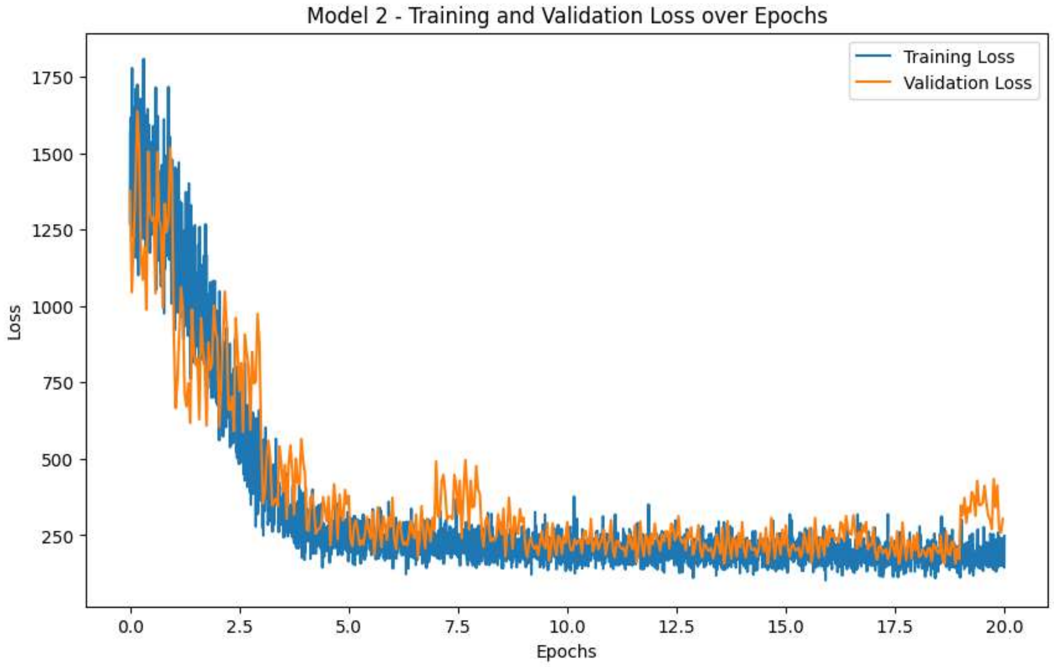

Figure 7 - CNN Final Train Validation Loss Graph

Multi Modal Neural Network:

Model Architecture

The Multi Modal Neural Network is a combined input model setup for an age prediction task that combines two different types of data: grayscale facial images and numerical features. This network architecture is particularly useful for age prediction as it utilizes a variety of information from multiple data modalities. This will allow the network to capture a wide range of relevant features for improved age prediction accuracy. Network Architecture shown in figure 8.

The neural network takes first the input grayscale facial images with dimensions of 64x64 pixels, which are processed by a series of convolutional layers, batch normalization, and max pooling layers and becomes a condensed representation of the facial features. These convolutional layers are responsible for extracting visual patterns and features from facial images, similar to CNN.

After the CNN layers are done with the image, the model then takes numerical features into account as part of the input data. These numerical features are merged with the image data after the convolutional layers processed image data, which the combined data is then passed through 2 more connected layers (linear layers). Note the network utilizes ReLU as the activation function. This implementation ensures that both types of information are considered during the prediction process. This integration enables the model to make accurate age predictions by considering a wider range of visual and numerical features.

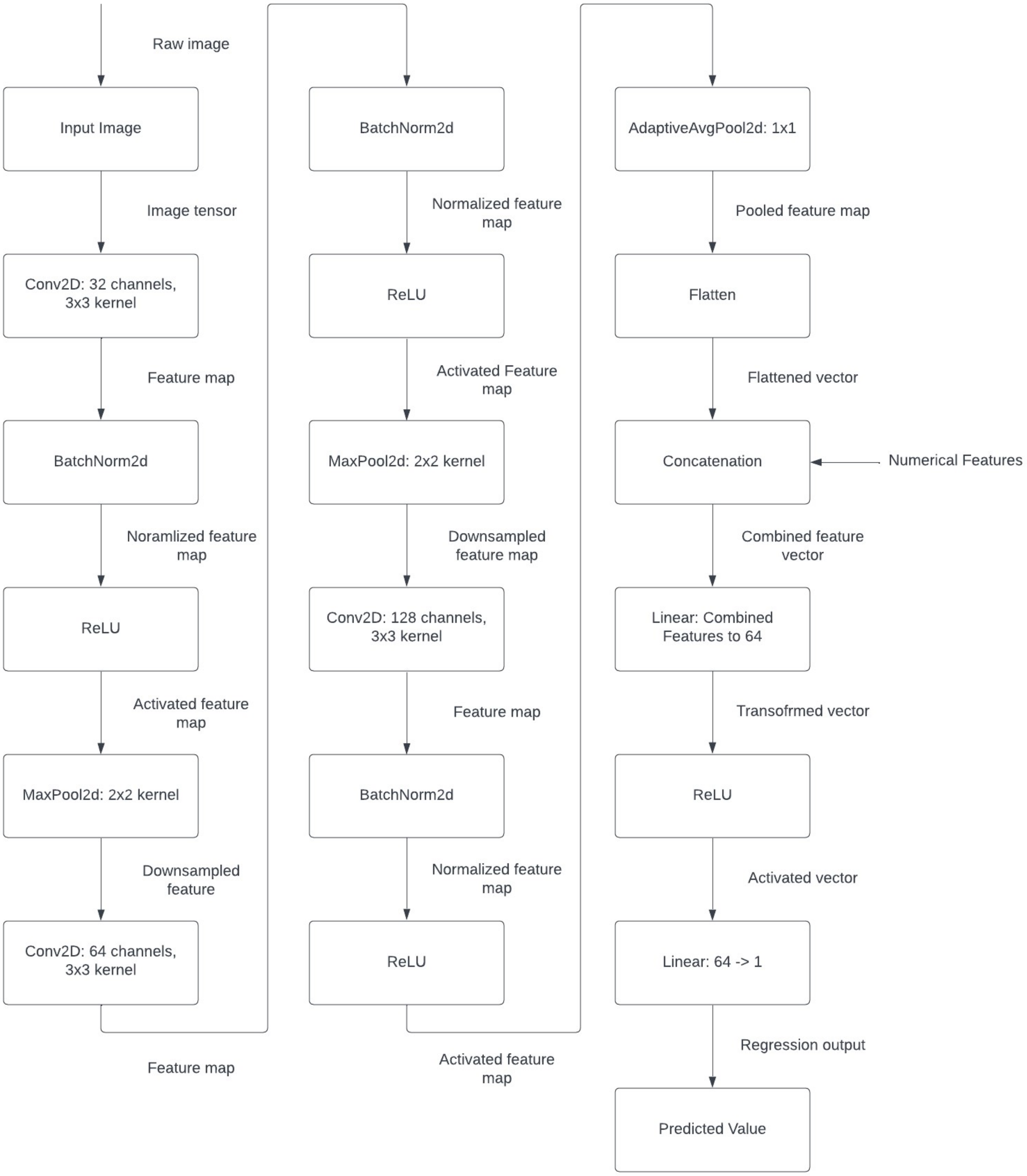

Figure 8 - Multi Modal Neural Network Block Diagram

Hyperparameter Tuning

When performing hyperparameter tuning for this model, the most important hyperparameter to tune was learning rate and batch size. Learning rate dictates how much the model can update its weights during training, while batch size dictates how many data points are observed before updating model's weights. When initially training the multimodal model, there were random spikes in the training data. When performing hyperparameter tuning, it was found out that I had a high learning rate and a low batch size, this was causing sudden spikes, because the model would occasionally get a batch filled with outlier data, and this would cause a drastic change in the weights, which thus causes training loss and occasionally validation loss to spike during training. This outlier occurrence was found to be more prominent with the introduction of numerical features in the network. After figuring the relationship between learning rate and batch size, I was able to strike a balance between the two with a learning rate of 1e-4 and a batch size of 200, as seen in training/validation loss in Figure 9.

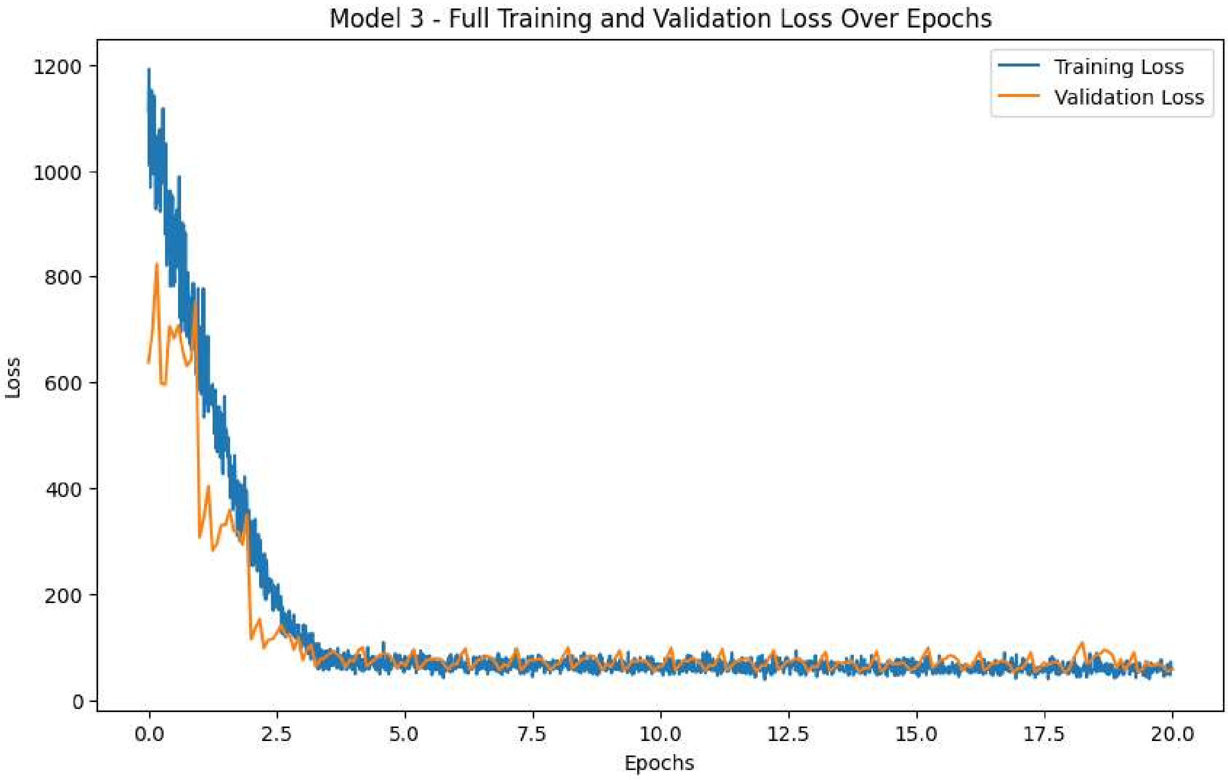

Figure 9 - Multi Modal Neural Network Final Train Validation Loss Graph

Results

Model 1- Linear Regression- Strengths and Weaknesses

The Linear Regression model (Model 1) showed great relative performance for younger ages, but its accuracy declined for individuals older than 20 years, as depicted in Figure 11. This decline in prediction accuracy is shown for ages above 20. Additionally, Model 1 had a consistent tendency to overpredict age. This is evident in Figure 12, where the histogram for Model 1 shows a sharper peak and a narrower base compared to Model 2. Model 1 is observed to have a bias towards over predicting age, with the most common residuals being around 8 years above the actual target age.

Furthermore, Model 1 showed a lack of confidence in predicting ages below 5 years. As seen in Figure 10, the model rarely predicted ages lower than approximately 5-7 years. This trend is further supported by the comparison of predicted and actual values in Figure 11, which shows a bias in the model that prevents it from predicting ages in the lower range (5-7 years old).

In terms of Mean Squared Error (MSE), Model 1 performed well, achieving a score almost ¼ that of the baseline, making it the second-best among the three models. This indicates that Model 1 is less prone to large errors. The Root Mean Squared Error (RMSE) backs this observation, showing that Model 1's predictions are generally consistent and do not often result in significant errors.

The Mean Absolute Error (MAE) for Model 1 suggests an average age prediction discrepancy of about 6 years, which is half of what was observed in the baseline. This demonstrates the model's precision and ranks it second among the three models for this metric. The R-Squared metric, which indicates the model's ability to capture variation in age, shows that Linear Regression performed second best among the models and significantly surpassed the baseline, with an 81% higher score. This suggests that Model 1, despite its limitations in predicting certain age ranges, generally understands and captures the age variation in the dataset effectively.

Model 2- Convolutional Neural Network – Strengths and Weaknesses

The performance of the CNN model (Model 2) was the worst among the three models we tested. As shown in Figure 10, Model 2 showed a major lack of confidence in predicting ages below 20. This pattern could be due to the model being predominantly trained on faces of individuals aged 20 and above, resulting in less accuracy for younger age groups. Moreover, there was a gradual decline in performance when predicting ages over 60, suggesting a generally limited range of accurate age prediction.

This pattern of inaccuracy is further shown in Figure 11, Model 2, where a significant frequency of overpredictions for ages 20 and younger is observed. The model appears prone to errors, often overestimating ages by an average of 10 years. This tendency to overpredict is explicitly evident in the younger age group.

When evaluated using Mean Squared Error (MSE), Model 2's performance was 3rd best, falling just 50 points short of the baseline and 260 points behind Model 1. This high MSE score shows a tendency towards larger errors in predictions. The Root Mean Squared Error (RMSE) confirms this, with the model scoring only 1 point better than the baseline and 10 points worse than Model 1. This shows that the model's predictions frequently involve significant errors and lack consistency in accuracy (Table 1).

The Mean Absolute Error (MAE) further underscores the model's limited performance, with an average age prediction difference of approximately 14 years, which is close to the baseline. This result is likely impacted by the model's poor performance in predicting ages under 20. Model 2's R-Squared metric reflects the amount of variation in age that the model can understand/deal with, in which it was the lowest among the three models. It scored only slightly better than the baseline and substantially lower than the next best model, Model 1. This low R-Squared value indicates a limited ability of Model 2 to capture and understand the variability in age, especially in the younger and older age groups (Table 1).

Model 3- Multi Modal Neural Network – Strengths and Weaknesses

The Multimodal Neural Network (Model 3) displayed a combination of strengths and weaknesses from both Model 1 and Model 2. Like the other models, Model 3 struggled with maintaining high prediction accuracy for older ages, as shown in Figure 11, Model 3. However, the magnitude of this inaccuracy was less severe compared to Models 1 and 2. A notable strength of Model 3 is its broader range of confident predictions, including more accurate guesses for very young ages, approaching close to 0, as demonstrated in Figures 10 and 11 for Model 3.

Regarding common prediction errors, Model 3 often overestimated ages, but typically by a smaller margin of about 5-6 years, as illustrated in Figure 12, Model 3. This suggests a more balanced prediction error compared to the other models.

In terms of Mean Squared Error (MSE), Model 3 outperformed the others, achieving a score that is approximately 1/7th of the baseline, the best among the three models. This low MSE score indicates that Model 3 is least prone to large errors. The Root Mean Squared Error (RMSE) supports this observation, with Model 3 showing the most consistent predictions and the fewest significant errors. With an RMSE score of 7.5, it ranks 1 point better than the second-best model, as detailed in Table 1.

The Mean Absolute Error (MAE) of Model 3 is approximately 5.5 years, which is lower than the baseline and is evident of high precision in age prediction. This performance places Model 3 as the best among the three models in terms of MAE. Additionally, Model 3 performed very well in the R-Squared metric, scoring 85% higher than the baseline, the highest among all models. This high R-Squared value reflects Model 3's effective utilization of multimodal features, enabling it to understand and capture the age variation in the dataset more effectively than the other models, as summarized in Table 1.

Table 1 - Numerical Model Evaluation Metric Table

	Linear Regression (Model 1)	Convolutional Neural Network (Model 2)	Multi Modal Neural Network (Model 3)	Baseline
Mean Squared Error	72.659	339.983	56.391	382.160
Root Mean Squared Error	8.524	18.439	7.509	19.550
Mean Absolute Error	6.362	14.842	5.517	14.975

R-Squared	0.810	0.110	0.852	0.0

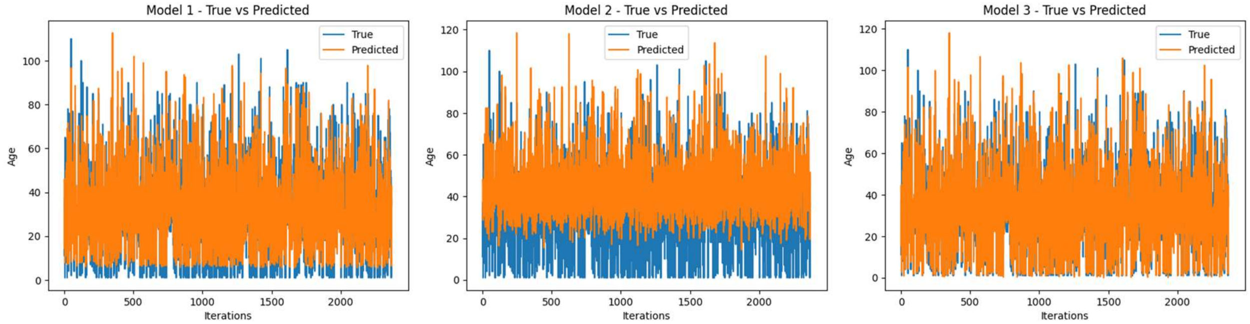

Figure 10 - True versus predicted per iterations graph comparison, Model 1 - Linear Regression, Model 2 - CNN, Model 3 - Multi Modal Network

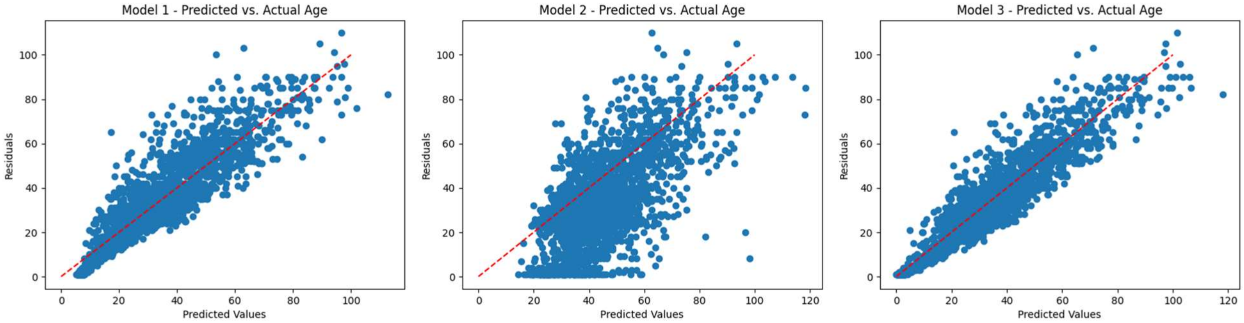

Figure 11 - True V.S. Actual Age scatter plot variant comparison Model 1 - Linear Regression, Model 2 - CNN, Model 3 - Multi Modal Network

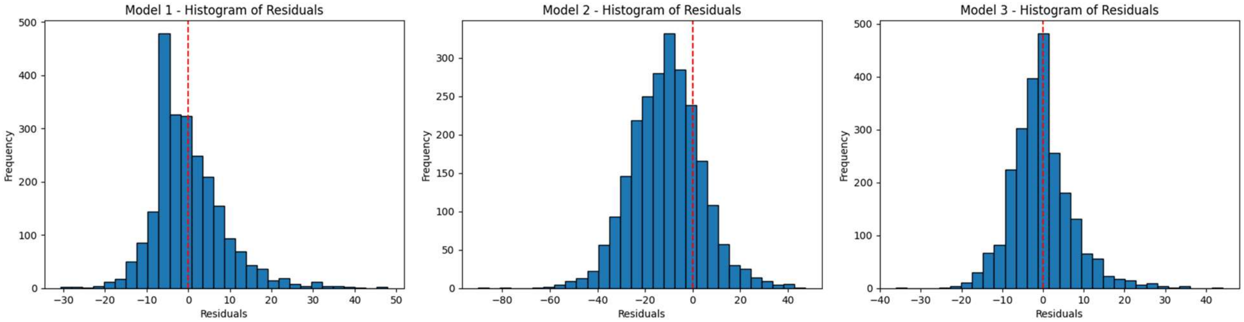

Figure 12 - Residuals Bar plot comparison Model 1 - Linear Regression, Model 2 - CNN, Model 3 - Multi Modal Network

Based on the comprehensive evaluation of the three models, Linear Regression (Model 1), CNN (Model 2), and Multimodal Neural Network (Model 3), for the age prediction task, I believe the best model to recommend it the Multimodal Neural Network (Model 3) as the primary model. This is due to several key observations during analysis. Firstly, Model 3 showed a great ability to accurately predict ages across a broad spectrum (Figure 10), including very young ages, a capability that was notably lacking in the other 2 models. This range of accurate predictions is crucial for a consistent and precise age prediction use.

Secondly, Model 3 demonstrated a more balanced error profile compared to Model 2 and 3. While it tended to overestimate ages, the margin of error was relatively smaller (approximately 5-6 years), suggesting a consistency in predictions that is vital for practical applications. (Figure 12)

Furthermore, the performance of Model 3 in the evaluation metrics is a major indicator of its superior performance compared to the other two models. It achieved the lowest Mean Squared Error (MSE) among the models, showing a reduced tendency for large errors. Its Root Mean Squared Error (RMSE) was also the lowest, proving the model's is consistently in making fewer significant prediction errors. On top of such, Model 3's Mean Absolute Error (MAE) was significantly lower than the baseline and the other models, proving its precision in age predictions. Lastly, the model's superior performance in the R-Squared metric, which was the highest among the models and greatly improved compared to the baseline, indicates its effectiveness in understanding the age variation in the dataset. Overall, the combination of Model 3's broad prediction range, balanced error profile, and outstanding performance across various key metrics makes it the most suitable choice for the age prediction task at hand.

Conclusion

In conclusion, this project has successfully explored and compared three different machine learning models (Linear Regression, Convolutional Neural Network, and Multimodal Neural Network) in the context of age prediction using the UTKFace dataset. Each model brought unique strengths and showed specific limitations, which gave valuable insights into the complexity of age prediction tasks.

The Linear Regression model showed good performance with younger ages but faced challenges in accurately predicting older ages. CNN, specialized in processing facial images, showed potential but was hindered by its inability to accurately predict ages, especially in the younger and older age groups. The Multimodal Neural Network, combining both image and numerical data, emerged as the most promising model. It showed the ability to capture a wider range of features which led to more balanced and accurate age predictions across different age groups, as evidenced by its superior performance in the tested model evaluation metrics (MSE, RMSE, MAE, and R-squared).

Overall, this project shows the potential benefits of performance that multimodal models can bring in addressing complex prediction tasks. More specifically, by utilizing numerical and image features, the model can catch nuances that may be missed if one type is used instead of the other.

As for future work that can further improve the age prediction models results. The first potential improvement includes eliminating sampling bias in the dataset, especially for factors such as race, and other features that don’t have equal representation of their data’s classes. Additionally, I believe that to capture the apparent complexity of the age prediction tasks, the neural networks implemented (CNN and Multimodal Neural Network) both should have more layers to allow for the increased ability to capture complex patterns in age predictions. The last improvement that could potentially improve results is incorporating more features into the dataset, as currently there are only 4 main features, and this makes the models susceptible to underfitting, while more features could potentially boost the models fitting to the data.

References

Figure 5 - Convolution. NVIDIA Developer. (n.d.). https://developer.nvidia.com/discover/convolution

[1] Age prediction using image dataset using machine learning - researchgate. (n.d.). https://www.researchgate.net/profile/Vijay-Kumar-319/publication/343162970_Age_Prediction_using_Image_Dataset_using_Machine_Learning/links/5f19aded299bf1720d5d3102/Age-Prediction-using-Image-Dataset-using-Machine-Learning.pdf