**Facial Attribute Analysis Using Deepface and CNN**

A PROJECT THESIS

Submitted by

## Venkata Harsha Vardhan Beeraka

19MIS1203

in partial fulfillment for the award of the degree of

Master of Technology

in

Software Engineering (5 Year Integrated Programme)



*DeepFace Revisited: Model Performance and Application Study*

Venkata Harsha Vardhan B.   
*School of Computer Science Engineering (SCOPE)*

*Vellore Institute of Technology, Chennai, India*

*Abstract*—"DeepFace Revisited" In this study, we examine the performance of Facebook's 'deepface' model, launched in 2014, and compare it with a custom-built Convolutional Neural Network (CNN) model. The goal is to assess the accuracy of 'Deepface' in facial recognition tasks and to evaluate its capabilities against a newly designed CNN model. The analysis focuses on both models' accuracy, efficiency, and adaptability in various facial recognition scenarios.

Keywords— Facial Recognition, DeepFace Analysis, Accuracy Comparison, Image Processing

# Introduction

Facial recognition technology has evolved significantly, notably since Facebook's 'deepface' model emerged in 2014. Initially, facial recognition relied on basic geometric models, offering limited accuracy. A paradigm shift occurred with Turk and Pentland's introduction of 'eigenfaces' in 1991[1], steering the field towards statistical approaches. Deep learning, particularly through Convolutional Neural Networks (CNNs), has since revolutionized this domain. DeepFace achieved a groundbreaking 97.35% accuracy y, resembling human-level performance, by training a nine-layer neural network with four million images.

However, facial recognition still confronts challenges such as varying lighting conditions and diverse facial expressions. Continuously expanding datasets aim to encompass this vast range of expressions and scenarios [2]. Additionally, privacy concerns[3], ethical implications, and potential biases in these technologies add layers of complexity to their development and deployment. The evolution and challenges of facial recognition technologies underline the importance of ongoing research and debate regarding their efficacy and ethical use.

Simultaneously, there's a growing emphasis on addressing the ethical and privacy concerns associated with facial recognition. Efforts are underway to establish transparent and equitable guidelines for the use of this technology, ensuring it respects individual privacy and minimizes biases. This involves scrutinizing the data used for training these models to prevent any inadvertent perpetuation of racial or gender biases. Moreover, the potential applications of facial recognition technology are expanding. Beyond security and identification, it's being integrated into fields like marketing, where it can offer personalized experiences, and healthcare, for patient monitoring and diagnosis. This broadening of scope necessitates a multidisciplinary approach, combining insights from technology, ethics, and social sciences to harness its benefits while safeguarding against its risks.

As facial recognition technology continues to evolve, its impact on society and individual rights remains a critical area of study and discussion. The balance between technological advancement and ethical responsibility will likely define the trajectory of its future development and integration into everyday life.

# Methodology

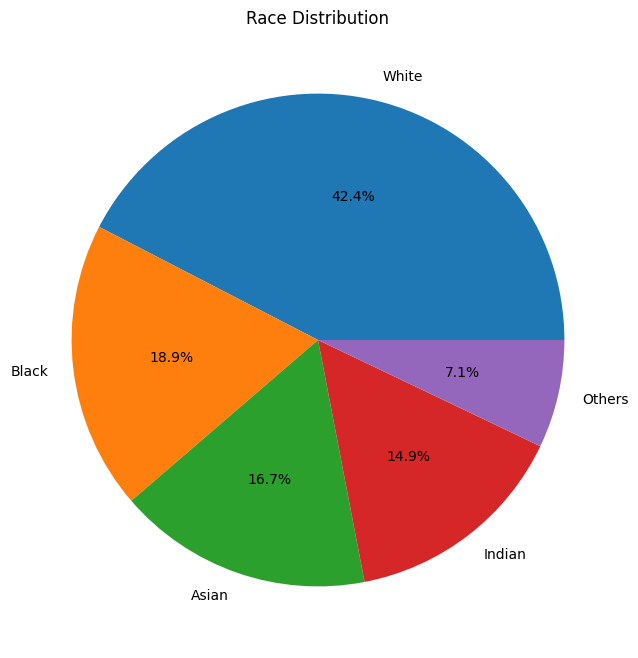
## The Preface

The deepface model was introduced in 2014 when it was tested on the dataset, (LFW)[4] Labelled Faces in the Wild, it achieved an accuracy of 97.35% on 13,233[5] images of 5749 different identities. Over the years, the size of the dataset has grown from 13,233 to 24,107 images in total. We’re now trying to analyze if the same 97.35% accuracy of the older dataset is valid or not.

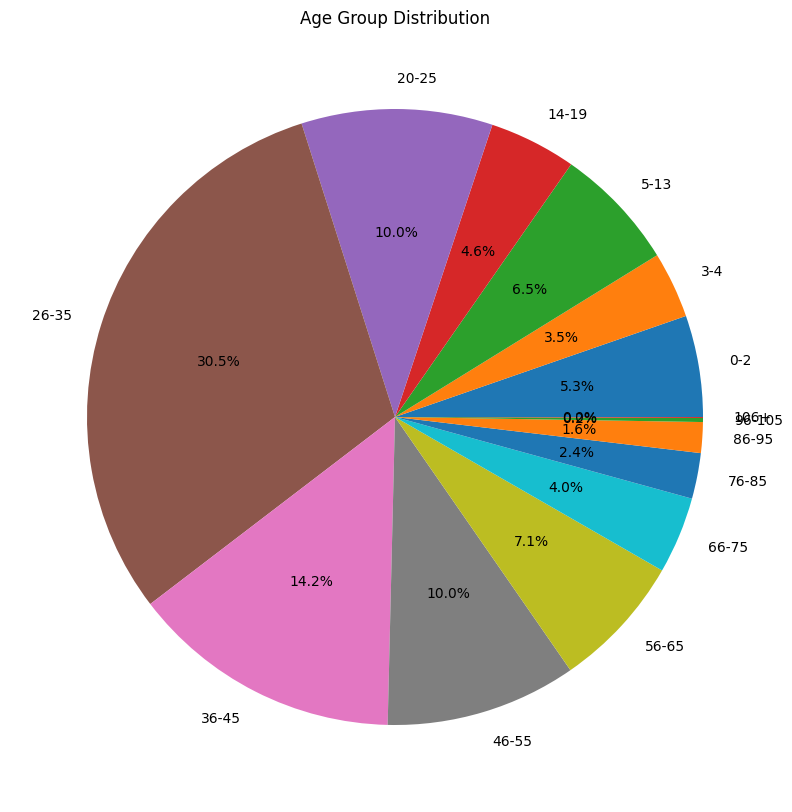
## Dataset Description

The dataset used is *Labelled Faces in the Wild.* The total images that are present to be analyzed is 24,103 images. Out of those images 52.2% of the images are *male* images, and the corresponding 47.8% is of *females*. The images corresponding to the race, *white, 42.4%* being the dominant race, followed by *black, 18.9%, Asian, 16.4%, Indian, 14.9%* and finally, *others (Hispanic, Latino, Middle Eastern), 7.1%.*

*The following graphs illustrates the presented information:*



*Fig.1.*



*Fig. 2.*

## The Process of Evaluating the Models

We are using the Labelled Faces in the Wild(LFW) dataset to evaluate the two models. The ‘deepface’ model as a model in itself to test is not attainable, since the file corresponding to the weights used for the model couldn’t be obtained. However, the model is also listed as a library. The library wraps different face models- "VGG-Face", "Facenet", "Facenet512","OpenFace","DeepFace","DeepID"   "ArcFac-e",   "Dlib",   "SFace"[6][7][8][9]. But, when the library is run without any specified model, it defaults to ‘DeepFace’ and, it should work as intended.And, for the custom-built models for predicting age, gender and race, we develop them using CNN, and modify them to suit our needs.

So, now we build the necessary models. For age, we are applying the model to perform a  regression task, we deploy a convolutional layer with 32 layers and a 3x3 kernel. And, to reduce the dimensionality we use *MaxPooling*, to make the model less sensitive in the position of features for the given input image. Following that, we convert the 2D feature map into a 1D Vector input since the images are processed in a batch to be fed into the dense layer. We here set the *batch\_size* to 128, therefore there are 128 neurons present, and activating the ‘relu’ (Rectified Linear Unit) function, which helps the model learn and associate complex patterns in the images, before passing it on to the final Dense Layer, i.e., the output layer. The ‘adam’ optimizer is used to optimize and reduce loss for all the models, due to its flexibility, efficiency, and robustness while training large datasets.

And the model while running through the epochs, the model realises the differences between its predicted ages and the true ages(which are obtained from the file names of the dataset), and adjusts its weights accordingly through backpropagation, trying to reduce the loss function, which in this case is MSE. As for the hyperparameters for the loss function, every parameter is set  at its default values (i.e., η ,learning rate= 0.001, β1, forgetting factor for gradients= 0.9, β2, forgetting factor for second moments of gradients=0.999).

Similarly, for gender and race, we need them to perform a binary classification task and multi-class classification tasks respectively using ‘relu’ for handling the 1st dense layer, and for the output layer, we use ‘sigmoid’ and ‘softmax’ activation functions. For the loss functions, ‘binary\_crossentropy’(1) and ‘categorical\_crossentropy’ (2) are used.

1. Binary Cross-entropy[10]:



Here, N is the number of samples, yi is the true label (0 or 1), and pi is the predicted probability of the sample being in class.

1. Categorical Cross-entropy[11]:



In this, yo,c is a binary indicator of whether class label c is the correct classification for observation ‘o’ and p o,c is the predicted probability of observation ‘o’ being in class c.

# Methodology: From Data Preprocessing to Model Evaluation

Trying to analyse the images, includes creating the model, train, validate and test them too. We first start by importing libraries for data processing (like NumPy, pandas), image manipulation (OpenCV), deep learning (TensorFlow, Keras), and visualization (matplotlib, seaborn), reading an image from a given path, convert it to RGB from BGR format, resize the images to 224x224 pixels, and normalize pixel values. Then, we extract labels (like age, gender, and race) from the filename to get a benchmark and later compare the results from the trained models to determine the accuracy. Since we have a large dataset of images, approx. 12,400, we use *data generators* help us yield batches of images and their corresponding labels for model training, instead of splitting the dataset instead of *x\_train, x\_test.* For the data processing, we split the dataset into 55% for training the data, 20%, 25% for validating and testing the model respectively.

The performance metrics for measuring all the three attributes were: MAE’[Mean Absolute Error], indicates how far off the *predicted age* is the *original age,* on average. Gender and race use *accuracy* as their performance metric.

# Results And Discussion

Here are the detailed analysis and interpretation of results for both models.

## Custom-Built CNN Algorithm

The model used 5 layers(1 Conv2D layer, 1 -MaxPooling2D layer, 1 Flatten layer, 2 Dense layers )to train and predict the images. Now, let us analyze the graphs and the values obtained by running the model.

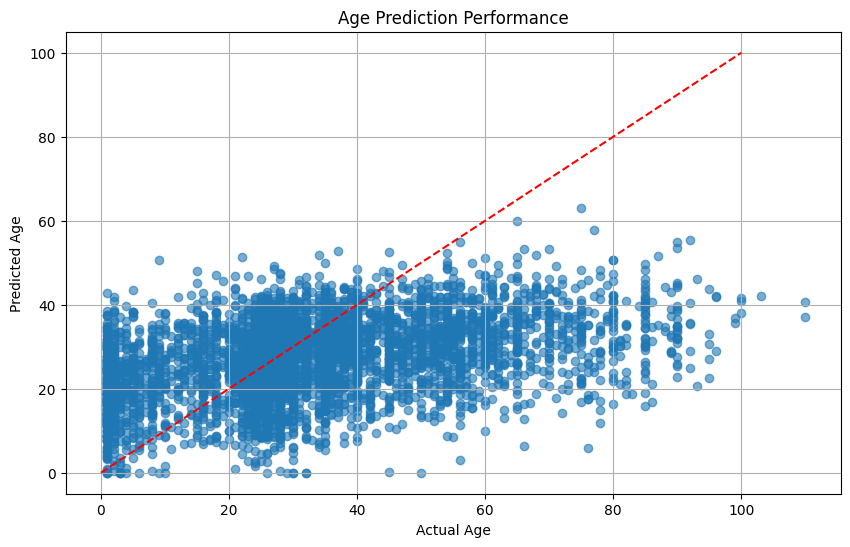


Fig. 3.

Analysing Fig. 3, we obtain the following observations:

The scatter plot describes the performance of the model in the age prediction. The predictions are plotted against the actual data obtained. We can observe concentration of data points along the diagonal, indicating many accurate predictions. The difference between predicted and actual ages increases as the actual age increases, indicated by the vertical spread of data points. The red dashed line represents the perfect prediction accuracy, and here, we can learn that the model is better at predicting younger ages than older ages. The model tends to underestimate the age for older individuals and overestimate it for younger individuals. MAE achieved: 14.18902297991213

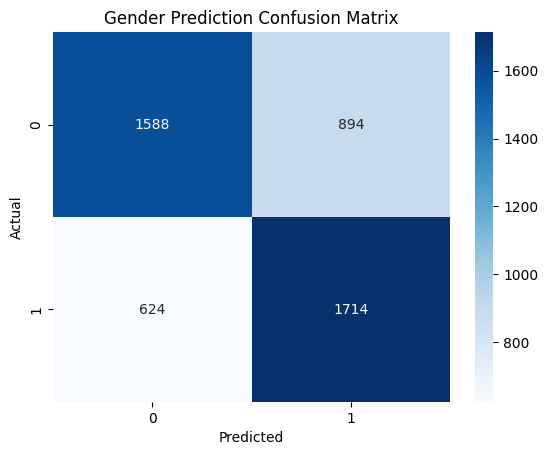


Fig. 4

Analysing Fig. 4, we observe following observations:

The confusion matrix indicates a binary classification model for gender prediction, with two classes: 0(male) and 1(female). True Positives: The model correctly predicted 'Male' for 1,588 entities. True Negatives: The model correctly predicted 'Female' for 1,714 entities. False Positives: The model incorrectly predicted 894 entities as 'Female' when they were ‘Male’. False Negatives: The model incorrectly predicted 624 entities as 'Male' when they were 'Female'. The has a higher accuracy in predicting 'Female' compared to 'Male', as seen by a larger number of correct predictions and fewer false positives for 'Male'. Accuracy for Gender Prediction: 0.6850622406639004

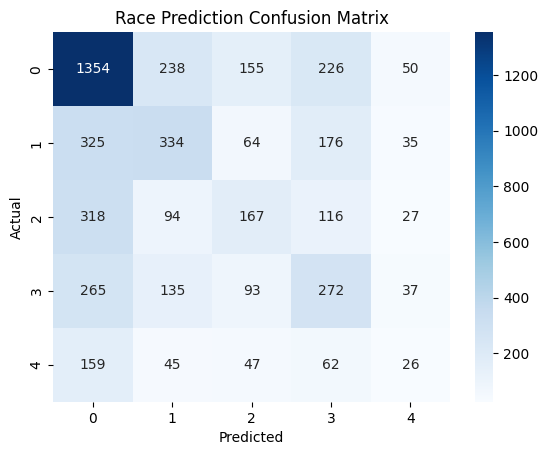


Fig. 5.

Analysing Fig. 5, one can deduce these observations:

The confusion matrix for race prediction shows a multi-class classification with five categories, indicated by the numbers 0 through 4 on both axes. Class 0 has the highest number of correct predictions (1354), but also significant misclassifications with other classes, especially with class 3 (226) and class 1 (238).

Class 1 has a balanced distribution of misclassifications with other classes, suggesting there are overall lesser obvious features learned for this category. Class 2 has lower correct predictions (167) and is confused with class 0 (318) and class 3 (116). Class 3 has a considerable number of correct predictions (272), with notable confusion with class 0 (265) and class 1 (176). Class 4 shows the least number of correct predictions (62), which could indicate potential difficulty for the model to distinguish this category or fewer entities in the training data.Accuracy achieved: 0.4466804979253112

## DeepFace

DeepFace is a sophisticated facial recognition system created by Facebook that employs a nine-layer neural network and uses over 120 million parameters to fine-tune the results.

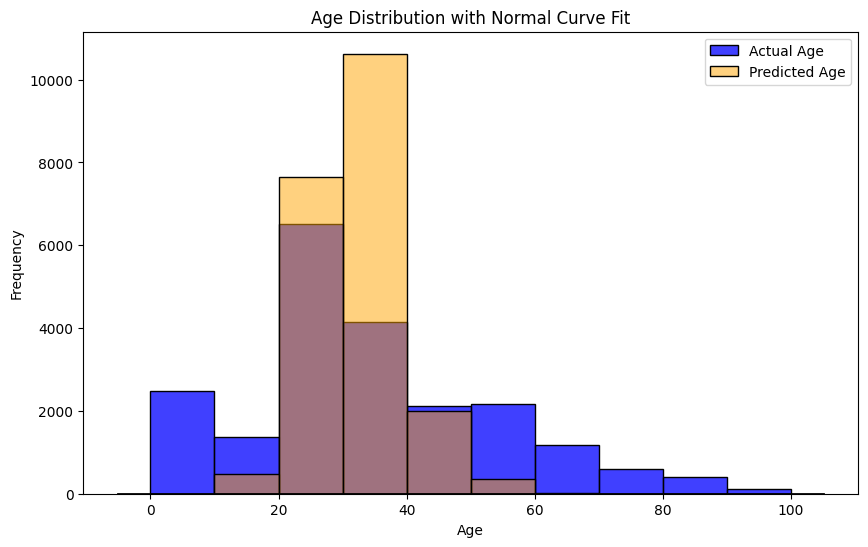


Fig. 6.

Analysing Fig. 6, one can obtain the following observations:

The histogram overlaid with a normal curve fit. This indicates frequency distribution of actual and predicted ages. There's a prominent peak in the distribution at the 20-30 age range, suggesting a higher concentration of both actual and predicted ages in this group. The actual age distribution is slightly right skewed. The predicted age distribution closely follows the actual age trend. There is a clear overestimation in the predicted ages for the 0-10 age range. The distributions diverge as age increases, especially after the 60-age mark, marking potential inaccuracies in age prediction for older demographics. MAE achieved: 12.32

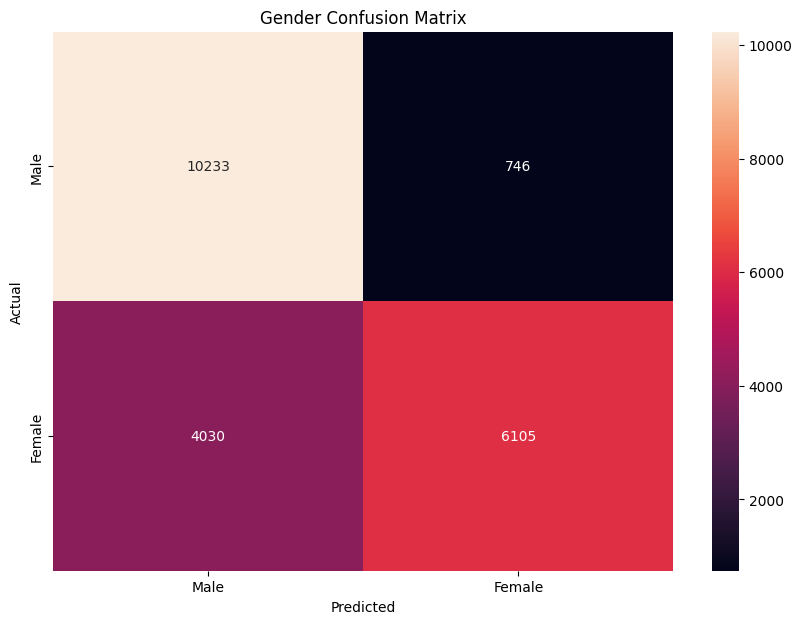


Fig. 7.

Analysing Fig. 7, we can deduce the following observations:

The confusion matrix indicates the performance of a gender classification model with two categories: Male and Female. True Positives: The model correctly predicted 'Male' for 10,233 instances. True Negatives: The model correctly predicted 'Female' for 6,105 instances. False Positives: The model incorrectly predicted 746 instances as 'Female' when they were ‘Male’. False Negatives: The model incorrectly predicted 4,030 instances as 'Male' when they were 'Female'. The model shows a higher accuracy in predicting 'Male' compared to 'Female', as seen by a larger number of correct predictions and fewer false positives for 'Male'. Accuracy achieved: 0.773823756

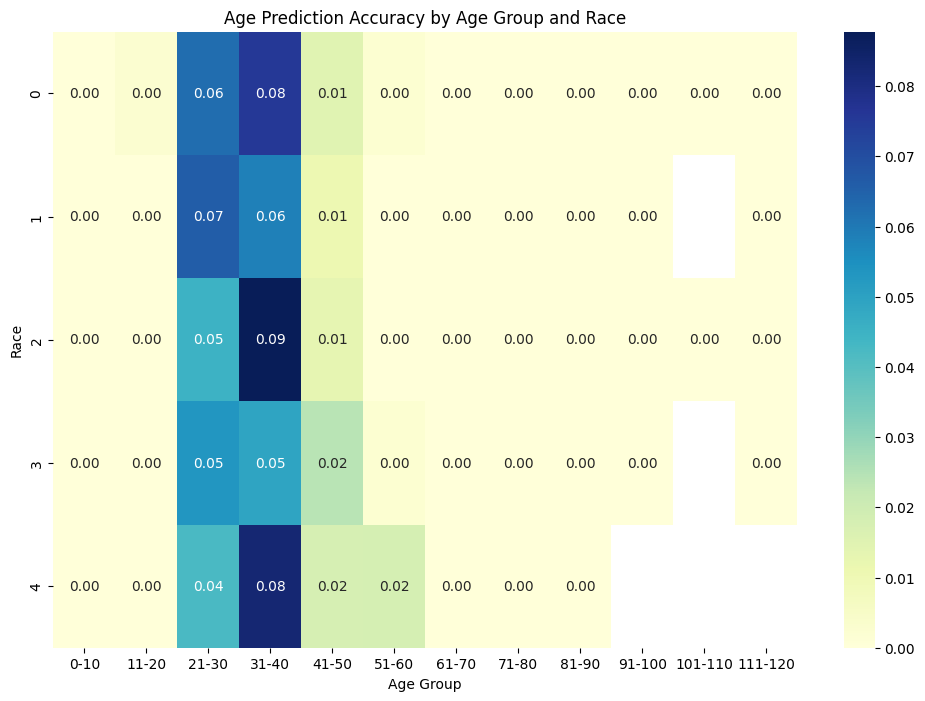


Fig. 8.

Analysing Fig. 8, one can obtain the following observations:

The age prediction accuracy varies significantly across different racial categories and age groups. The most accurate predictions occur in the age groups of 21-30 and 31-40, across all racial categories presented. Racial category 2 in the age group 31-40 shows the highest accuracy, with a score of 0.09. Racial categories 0 and 4 have lower accuracy scores, particularly in the age groups 41-50 and 51-60.

There is a general trend of declining accuracy as age increases, with very little or no accuracy in the oldest age groups (71-120). Accuracy achieved: 0.61599642681

# Conclusion

After analysing both the models, and measuring their various performance metrics, these were the main takeaways when the DeepFace failed to recognise any face.

DeepFace shows varying accuracy in different age groups, particularly less precise for very young [0-16] and very old [90-116] ages. The model's performance is hindered by foreground objects blocking the face. Accuracy in age and race recognition is significantly affected by lighting and image filters.The model only recognizes gender in binary terms, which may not cover all gender identities. High-quality images are required for accurate recognition, with a minimum dimension requirement of 150 pixels. The presence of glasses is negatively impacting the model's face recognition accuracy. Effective processing requires images to be larger than 150 pixels in width or height.Watermarks in images can disrupt the model's ability to recognize faces. The model's accuracy varies with portrait photos, especially affected by the depth of field.

TABLE I. Consolidated Results for two models

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithms | Attributes | | |
| Age (MAE) | Gender (Accuracy) | Race (Accuracy) |
| Custom Model | 14.667972297991 | 0.6850622406639004 | 0.4466804979253112 |
| DeepFace | 12.32 | 0.773823756 | 0.615996426 |

With the study's findings we also observe that age prediction accuracy by the model is influenced by both age and race, with a notable decrease in precision for age groups under 10 and over 70. The results underscore the challenges in achieving high accuracy across all demographics, particularly in certain racial categories. This suggests a need for diverse and balanced training datasets, as well as potentially refined model architectures to address underrepresented age groups.

Future research should focus on these areas to enhance the model's robustness and fairness in age prediction tasks.

##### Acknowledgements:

I extend my heartfelt gratitude to the department faculty for providing support, and my academic mentor, Dr Jahangeer Sidiq S. for their insightful guidance throughout this research, whose constructive feedback was instrumental in refining the paper. Finally, we thank our families and friends for their patience and encouragement, which were the bedrock of our perseverance and success in this endeavor.

##### References

1. Turk, M., & Pentland, A. (1991). Eigenfaces for Recognition. Journal of Cognitive Neuroscience
2. H. -B. Kim, N. Choi, H. -J. Kwon and H. Kim, "Surveillance System for Real-Time High-Precision Recognition of Criminal Faces From Wild Videos," in IEEE Access, vol. 11, pp. 56066-56082, 2023, doi: 10.1109/ACCESS.2023.3282451.
3. “UTKFace\_LFW Dataset.” *Utkface*, susanqq.github.io/UTKFace/
4. "FaceNet: A Unified Embedding for Face Recognition and Clustering" by Florian Schroff, Dmitry Kalenichenko, and James Philbin (2015):.
5. Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. Conference on Computer Vision and Pattern Recognition (CVPR).
6. M. Jha, A. Tiwari, M. Himansh and V. M. Manikandan, "Face Recognition: Recent Advancements and Research Challenges," 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2022, pp. 1-6, doi: 10.1109/ICCCNT54827.2022.9984308.
7. I. Masi, Y. Wu, T. Hassner and P. Natarajan, "Deep Face Recognition: A Survey," 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Parana, Brazil, 2018, pp. 471-478, doi: 10.1109/SIBGRAPI.2018.00
8. A. Firmansyah, T. F. Kusumasari and E. N. Alam, "Comparison of Face Recognition Accuracy of ArcFace, Facenet and Facenet512 Models on Deepface Framework," 2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE), Jakarta, Indonesia, 2023, pp. 535-539, doi: 10.1109/ICCoSITE57641.2023.10127799.
9. J. Deng, J. Guo, N. Xue and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 4685-4694, doi: 10.1109/CVPR.2019.00482.
10. Benjaminson, Emma. “Binary Cross-Entropy.” *Binary Cross-Entropy – Emma Benjaminson – Mechanical Engineering Graduate Student*, sassafras13.github.io/BiCE/
11. “Categorical Cross-Entropy Loss.” Understanding Categorical Cross-Entropy Loss, Binary Cross-Entropy Loss, Softmax Loss, Logistic Loss, Focal Loss and All Those Confusing Names, gombru.github.io/2018/05/23/cross\_entropy\_loss/