# Analyzing Performance Degradation of Facial Recognition in Convolutional Neural Networks: A Facial Feature Study

Priyadarshini Saha

Tanisha Dhami

Department of Systems Design Engineering

SYDE 552/ BIOL 487

University of Waterloo

Professor Micheal Furlong & Professor Terry Stewart

April 23, 2024

# Letter of Submission

April 23, 2024

Evaluators

SYDE 552

University of Waterloo

Waterloo, Ontario

N2L 3G1

Dear Evaluators,

The enclosed report, entitled "Analyzing Performance Degradation of Facial Recognition in Convolutional Neural Networks : A Facial Feature Study" was prepared during the Spring 2024 term.

We discussed the subject of the report at the brainstorming stage with our professors, Terry Stewart and Michael Furlong. We would like to express our gratitude to them for their invaluable assistance in selecting our topic and for their unwavering support and encouragement throughout the work term as we tackled the challenges outlined in this report. Additionally, we extend our appreciation to Trevor Yu, the TA of SYDE 552, whose tutorials and guidance were instrumental in helping us navigate the complexities of our project.

This report was authored entirely by us and has not received any previous academic credit at this or any other institution.

Sincerely,

Priyadarshini Saha, Tanisha Dhami.

# Abstract

Rationale

This study explores convolutional neural network architecture modifications, aiming to mimic facial recognition deficits using a neural network.

Objective

The objective was to understand how specific artificial impairments within a convolutional neural network model could mimic neurological dysfunction. Our hypothesis was that by selectively disabling neurons or certain neural network layers within the VGGFace 16 convolutional neural network, akin to neurological impairments such as prosopagnosia (commonly known as face blindness), the model would exhibit compromised facial feature recognition.

Methods

Utilizing a pre-trained VGGFace 16 model, we first established a baseline performance for face detection by measuring accuracy, precision, recall, F1-scores and confusion matrices. Then, the model was modified to simulate neurological deficits through methods such as feature ablation (remove of critical data inputs), selective dropout (randomly ignoring certain neurons), noise injection (adding random data into inputs), and deactivating specific neurons associated with the recognition of eyebrows, eyes, nose, and mouth.

Results

Results indicated that such modification of the model produced similar results to running the original model on modified datasets without certain facial features. The metrics such as accuracy, precision, F1-score varied greatly among the different modified models and datasets, but still aligned with expectations.

Conclusions

Degrading the performance of the models led to the predicted outcome. This research highlights the potential of CNNs to model and understand neurological deficits in the brain for facial recognition, suggesting directions for future studies to improve diagnostics models and therapeutic strategies.

# Table of Contents


**1.0 Introduction** **5**
- 1.1 Purpose and Problem Statement 5
- 1.2 Literature Review 5
- 1.3 General Approach and Objectives 6
- 1.4 Model Interests 7

**2.0 Methods** **8**
- 2.1 Model Description and Baselines 8
- 2.2 Modifications to the Model 9
- 2.3 Assessment of Performance 11

**3.0 Experimental Design and Hypotheses Testing** **12**
- 3.1 Datasets 12
- 3.2 Metrics 12
- 3.3 Models 13
- 3.4 Procedure 14
- 3.5 GRAD-CAM Activation Maps 14

**4.0 Results and Analysis** **16**
- 4.1 Original model 16
- 4.2 Original Dataset 16
- 4.3 Model Analysis 17
- 4.4 Neuron Intervention - Critical Layers 18
- 4.5 Masked brows dataset - Feature Abrasion 19

**5.0 Discussion** **20**
- 5.1 Important Findings 20
- 5.2 Unexpected Observations 21
- 5.3 Summary of Results 22
- 5.4 Interpretation of results 23
- 5.5 Future Directions 24

**6.0 References** **25**

**7.0 Appendix** **26**

# List of Figures and Tables

**Figures List**


Figure 1 VGG-16 Model Architecture…… 8
Figure 2 Baseline Accuracy and Loss …… …… 9
Figure 3 Heatmap Later Layer …… 15
Figure 4 Heatmaps Middle Layer …… 15
Figure 5 Heatmap Early Layer …… 15
Figure 6 Original Model Performance …… 16
Figure 7 Overview of Results …… 21
Figure 8 Original Model Line Graph …… 24


**Tables List**


Table 1 Baseline Performance Metrics …… 8
Table 2 Performance of Original Model …… 16
Table 3 Accuracies on original dataset…… 17
Table 4 Accuracies on augmented datasets…… 18
Table 5 Summary of Findings…… 20
Table 6 Summary of Results …… 22
Table 7 Full Dataset Summary …… 27

# 1.0 Introduction

The research investigates whether modifications to a pre-trained convolutional neural network (CNN) VGGFace model 16 through feature ablation, selective dropout, noise injection, and neuron deactivation degrades CNN's performance on face detection as we remove key facial features such as eyes, eyebrows, nose and mouth, thereby simulating neurological deficits similar to those observed in conditions like prosopagnosia (face blindness).

## 1.1 Purpose and Problem Statement

This project dives deep into understanding facial recognition issues using convolutional neural networks (CNNs), which are computer systems inspired by how the brain works. Our main aim is to replicate conditions like face blindness (prosopagnosia) or similar problems that affect how people recognize faces.

We carefully study how CNNs work on a dataset of faces and tweak them to mimic real-life problems people might face in recognizing faces. We use techniques like dropout layers, noise injection, and adjusting how much the network learns from different parts of the face, like the mouth, nose and eyes. These changes are meant to decrease the network's accuracy when it is run on the dataset, bringing it close to the accuracy when run on augmented data without certain facial features.

Our project is about combining learning from different fields like cognitive neuroscience and computer vision. We're thankful for the help we've received from our mentors and classmates, who've guided us through this project.

Through various methods of modifying the model, we've managed to impair the recognition capacity. We believe our work not only helps us understand more about how our brains work but could also lead to better technology to help people with face recognition difficulties.

In short, our project is a step forward in understanding how computers and brains interact when it comes to recognizing faces. We hope it sparks more research and ideas to help people with similar challenges.

## 1.2 Literature Review

Existing studies leverage CNNs for facial recognition tasks, showing high accuracy in identifying distinct facial features. However, there is limited research on deliberately impairing these models to mimic neurological conditions that affect facial recognition capabilities. This project aims to fill this gap by modifying CNN behavior to study such deficits.

#### 1.2.1 Convolutional Neural Networks (CNNs) for Face Detection:

We adopted a CNN-based approach for face detection, a methodology widely recognized for its efficacy in extracting intricate features from images. CNNs operate by passing the input image through multiple layers of convolutions to gradually extract features at different levels of abstraction. The architecture we employed, VGGFace with VGG16, has been pre-trained on extensive datasets, enabling it to detect faces reliably. This methodology draws inspiration from the hierarchical processing of visual information in the human brain, where different areas specialize in detecting various facial features.

#### 1.2.2 Model Architecture and Training:

Our face detection model architecture comprises layers such as flattening, dense, dropout, and output layers. These components are integral to the model's ability to learn and generalize from the data. The utilization of rectified linear unit (ReLU) activation functions and dropout regularization aids in preventing overfitting, ensuring the model's robustness. Through extensive training on labeled datasets, the model learns to recognize images containing faces

#### 1.2.3 Evaluation Metrics:

To assess the performance of our model, the metric we have employed is accuracy, and we also measured the time taken by the model per step. These metrics provide insights into the model's ability to correctly identify faces while also looking at how time effective it was. The resulting accuracy values obtained indicate the effectiveness of our approach in affecting the model’s performance.

#### 1.2.4 Facial Features Studied:

We have decided to augment the dataset for each of the major facial features present in the front view of the face, including eyes, nose, and mouth. We also decided to incorporate eyebrows since masking eyebrows lead to the most significant deficits in face recognition performance using deep convolutional neural networks (DCNNs), even more so than masking eyes, mouth, or nose, corroborating previous human studies on the critical role of eyebrows for this task (Müller et al., 2024).

### 1.3 General Approach and Objectives

Existing studies leverage CNNs for facial recognition tasks, showing high accuracy in identifying distinct facial features. However, there is limited research on deliberately impairing these models to mimic neurological conditions that affect facial recognition capabilities and study the model performance on face

recognition as key facial features are removed. This project aims to fill this gap by modifying an existing pre-trained CNN for face recognition to simulate neurological defects.

### 1.3.1 Approach

The primary approach involved utilizing the pretrained VGGFace 16 model and evaluating its performance on the LFW (Labeled Faces in the Wild) dataset. Subsequently, the dataset is augmented to create subsets with specific facial features, including eyes, nose, mouth, and eyebrows, blacked out. The four subsets with masks over facial features (eyes, eyebrows, nose, and mouth) are generated using computer vision techniques employed by Python dlib and OpenCV libraries that detect features using facial landmarks and predictors. The original (unchanged) VGGFace 16 model is then applied to these augmented datasets to establish a baseline performance. Additionally, the original model is then deliberately modified in five different ways to degrade its performance on face recognition. Finally, the modified models are evaluated on the five datasets, including the original and generated masked datasets, to compare their performance against the original model's performance as facial features are removed or masked.

### 1.3.2 Objectives

The primary objective of this study is to achieve comparable performance between the original model on augmented datasets and the modified model on the original dataset. This alignment in performance would indicate that the modified model behaves as if the images lack certain facial features, thereby simulating neurological conditions affecting facial recognition capabilities.

## 1.4 Model Interests

The VGGFace 16 model, celebrated for its precision in facial recognition tasks, serves as a foundational tool for understanding complex neural mechanisms. Its adaptability and performance in varying conditions make it an excellent candidate for experimental modification to study disorders like prosopagnosia. The main concern in this study is facial features, so we focus on a dataset with front-view images, and for such images, the accuracy can reach 100% (Nakada, 2017).

Delving into such a well-established model allows us to explore beyond typical usage, potentially uncovering novel neural behaviours and contributing to a deeper understanding of cognitive impairments in facial recognition. These insights could guide the development of more nuanced diagnostic tools and therapeutic strategies.

## 2.0 Methods

### 2.1 Model Description and Baselines

We are utilizing the VGGFace 16 model, available through a specific implementation in the Keras library (source: Keras-VGGFace GitHub). Our evaluation benchmarks the model's baseline performance in terms of average accuracy, precision, recall, and F1 score across five distinct datasets that we prepare from the LFW dataset. This dataset is a mix of 10,000 training images. The newly created datasets are normal facial images, and images with eyes, eyebrows, nose, or mouth masks. This comparative analysis provides a comprehensive understanding of how the model's face detection capabilities are affected by systematic feature obstructions.

Model Description:

The model VGG16 is designed for image processing tasks like face detection. It comprises multiple sequential layers, including convolutional layers that extract features from the input images, max-pooling layers that reduce the spatial size of the representation, and fully-connected layers that process the features to perform classification. The network ends with a softmax layer that outputs probability distributions for different classes. This structured architecture, as can be seen in Figure 1, allows the CNN to learn complex hierarchies of features at various levels of abstraction.

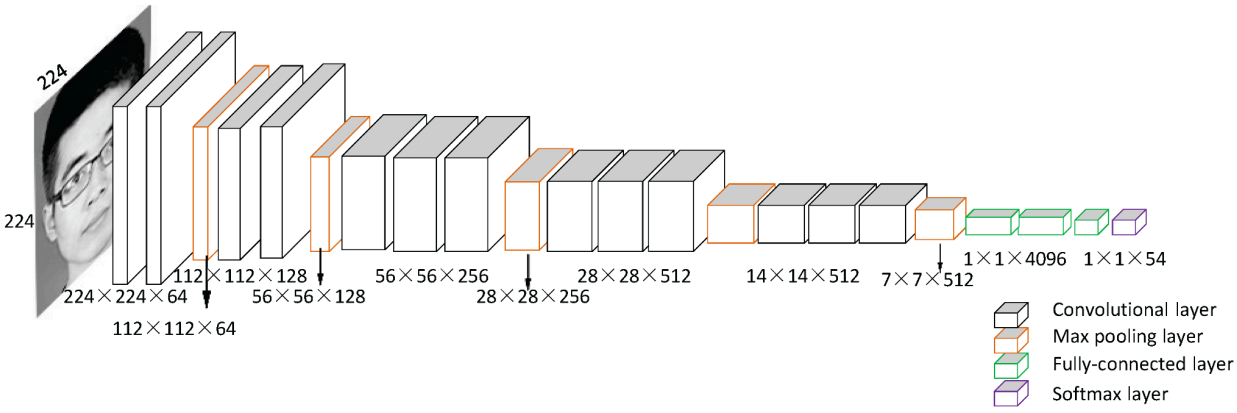



Figure 1: VGG-16 Model Architecture (Pei et al., 2019, Figure 1)

| Original Model | Normal Data |
|---|---|
| Average Accuracy | 72.98% |
| Precision | 0.995040061 |
| Recall | 0.9905051272 |
| F1 Score | 0.9927674153 |

Table 1: Baseline performance metrics

Table 1 shows the baseline metrics obtained by running the original model on the original dataset. We will utilize these metrics as a reference point for comparing our modified model and datasets.

As we can observe from the data, the accuracy of the model stayed almost the same over the different epochs. As expected, the loss declined to 0 within the first epoch. Figure 2 plots the accuracy and loss over 10 epochs of training the original model on normal dataset:

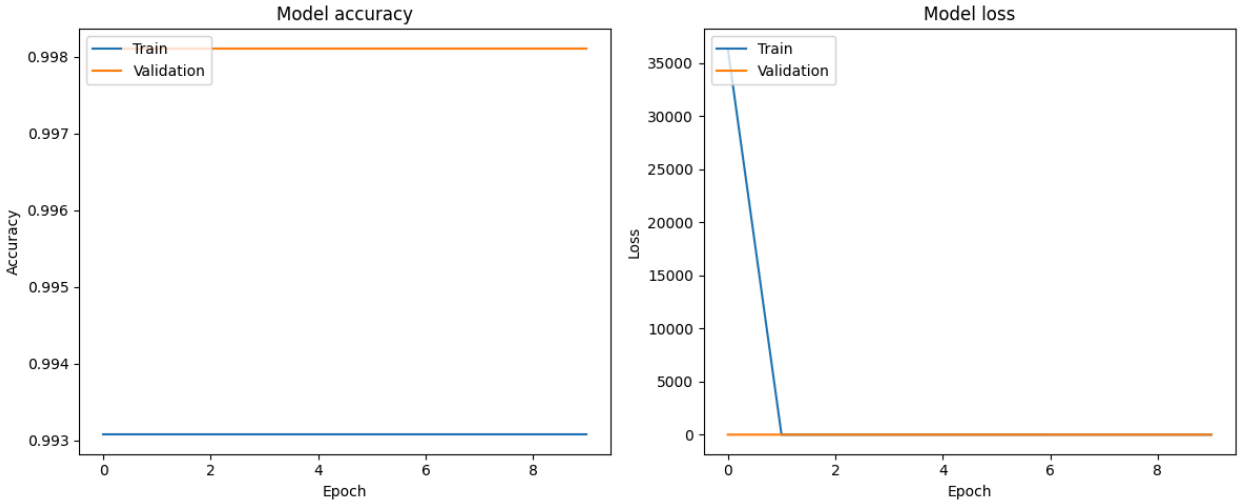



Figure 2: Baseline accuracy and loss

## 2.2 Modifications to the Model

We use four approaches to simulate neurological deficits in the pre-trained CNN.

### 2.2.1 Feature Ablation

In the VGGFace 16 CNN, feature ablation involves systematically removing or 'ablating' specific features, which in this context refers to activations of certain layers. This process allows us to study the impact on the network's performance, similar to simulating a neurological deficit caused by brain damage or dysfunction in specific neural regions. For example, by zeroing out activations in layers responsible for recognizing facial features like eyes or mouth, we can observe how the absence of these critical detections affects the network's ability to recognize faces, similar to how certain brain injuries can impair visual recognition abilities.

### 2.2.2 Noise Injection

The intervention involves injecting Gaussian noise into specific layers of the cloned VGGFace model. By doing so, the noise simulates random disturbances that can occur during neural processing, akin to sensory noise or synaptic dysfunction in a biological nervous system. This intervention aims to assess the

model's robustness and its ability to maintain performance when faced with the kind of unpredictable information that a neurological system might encounter due to deficits or external interference.

#### 2.2.3 Selective Dropout

This intervention involves selectively applying dropout - “selectively chooses the best neurons" to drop out (Barrow, Eastwood, & Jayne, 2016) - to the VGGFace 16 model, aimed at specific layers chosen for their relevance to the facial recognition task using GRAD-CAM results. It involves randomly deactivating a subset of neurons during training, mimicking neurological deficits observed in brain injuries or diseases. By simulating partial loss of neural function, this approach allows us to evaluate the network's dependency on specific neuron groups for facial feature recognition and its ability to adapt to reduced neural capacity. This not only helps prevent overfitting but also provides insights into the model's redundancy and fault tolerance.

#### 2.2.4 Neuron Intervention

This technique involves selectively activating or inhibiting specific neurons within a model's layer by adjusting their outputs. First, critical neurons for a particular feature are selected within a network layer. Then, their activity is modulated by adding or subtracting a fixed value from their outputs. In our case, we're setting these neuron outputs to zero, effectively disabling them. Since neurons transmit information through their outputs, setting them to zero means they no longer contribute to subsequent calculations in the network. This simulates the effect of non-functional or 'lesioned' neurons in a biological brain, akin to creating a virtual lesion in the neural network.

The manipulations described—feature ablation, noise injection, selective dropout, and neuron intervention—serve to simulate various aspects of neurological deficits that can contribute to face blindness, or prosopagnosia, in the brain. Each technique targets the neural network in a way that mimics different potential disruptions in brain function that might lead to difficulties in recognizing faces.

### 2.3 Assessment of Performance

We have used the following metrics to assess the performance of the model:

1. Accuracy of a model is the percentage of correct predictions made by a model. This is the most significant metric in the experiments, with the average accuracy over five runs being reported. This measure is the main measure we are looking at when concluding if changes to the model have led to a degradation of performance.

All others are reported for the best run of the model:

2. Precision measures the accuracy of positive predictions (i.e., the proportion of predicted positives that are actually true positives).Further, since all images are faces, this should always be 100%.
3. Recall measures the ability of the classifier to find all the positive samples (i.e., how well the model can detect actual positives from the dataset).
4. F1 Score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall.

## 3.0 Experimental Design and Hypotheses Testing

**Hypothesis**

- Null Hypothesis (H0): Interventions such as feature ablation, noise injection, selective dropout, and neuron intervention will not significantly change the facial recognition accuracy of the model compared to the baseline.
- Alternate Hypothesis (H1): Interventions such as feature ablation, noise injection, selective dropout, and neuron intervention will significantly improve or degrade the facial recognition accuracy of the model compared to the baseline.

**Experimental Design**

### 3.1 Datasets

1. Original Dataset: Using the LFW (Labeled Faces in the Wild) dataset, which is a standard facial recognition dataset, with typical preprocessing steps like normalization and resizing to fit as input to VGGFace 16 model.
2. Modified Dataset: Using the original dataset, several augmented datasets are created by systematically removing specific facial features like the eyebrows, eyes, nose, and mouth by masking them out with a black box. To achieve this, we leverage Python dlib and OpenCV's computer vision techniques to detect facial features (eyebrows, eyes, nose and mouth) and then use a convex hull to draw a bounding box and set the pixels to black for masking. This creates 4 masked datasets: masked_eyebrows, masked_eyes, masked_nose, and masked_mouth.
3. Deficit Training Dataset: This dataset is created by using training samples from the masked datasets - masked eyebrows, masked eyes, masked nose, and masked mouth. We randomly select samples from the four modified dataset to create this new dataset which is a mix of 10,000 images with facial features masked. This dataset is then used to train the treatment models (modified models).
4. Labels: Since we are doing face detection through facial feature recognition. The samples will have label 0 for no face detected if face outline or any facial feature among eyebrows, eyes, nose and mouth are missing. The samples will have a label 1 if a face along with all its facial features: eyes, eyebrows, nose and mouth are detected.

### 3.2 Metrics

- Primary Metric: Average accuracy (over 5 runs), measured as a percentage of correctly categorized face detection for evaluating the performance of each model.

- Secondary Metrics: Precision, recall, and f1-score for evaluating the performance of the model. These metrics are reported over the last run of the model (not the best or worst) as the last iteration will be the weights of the model that will be used for predictions.
- Decision Metrics: Implement the GRAD-CAM technique to visualize which parts of the images are most significant for predictions at each convolutional neural network layer to help interpret and decide which layers contain neurons that are the most active or sensitive towards facial features.

### 3.3 Models

**3.3.1 Original Model:** Pretrained VGGFace 16 model, with added flattened and dense layer for categorical predictions. This model is trained on the original unmodified dataset, and tested on the original unmodified dataset and masked datasets to establish a baseline performance.

#### 3.3.2 Treatment Models Creation

- Feature Ablation Model (critical layers): Introduce feature ablations at the critical layers as determined by the GRAD-CAM activation maps (see Jupyter Notebook). It is trained on the deficit dataset, and tested on original and masked datasets.
- Noise Injection Model (critical layers): Introduce a high Gaussian noise of 0.8 at the critical layers as determined by the GRAD-CAM activation maps (see Jupyter Notebook). It is trained on the deficit dataset, and tested on original and masked datasets.
- Selective Dropout Model (critical layers): Introducing random selective dropout of neurons with a rate of 0.8 at the critical layers as determined by the GRAD-CAM activation maps (see Jupyter Notebook). It is trained on the deficit dataset, and tested on original and masked datasets.
- Combines Model (critical layers): Introduce feature ablation, noise injection with Gaussian noise of 0.8, and selective dropout of 0.8 at the critical layers of the GRAD-CAM activation maps to see what happens when we incorporate all the modifications. It is trained on the deficit dataset, and tested on original and masked datasets.
- Neuron Intervention Models (critical layers): We find the neurons that are the most active at a specific critical layer for each facial feature (eyes, eyebrows, nose, and mouth) and then deactivate those neurons by setting it to 0 to essentially stop those neurons from contributing to the calculations. In total, we create 12 models in this setup - one for each layer (conv3_2, conv4_2, and conv5_3) and for 4 facial features (eyebrows, eyes, nose and mouth).

### 3.4 Procedure

1. Split the datasets into training, validation and test sets
2. Training: Each model (control and treatment models) is trained on its respective datasets (original and deficit datasets) as appropriate. We use the validation set to fine-tune hyperparameters if necessary.
3. Testing: All models, except the neuron intervention models, are tested on the original dataset and the masked datasets. The neuron intervention models are only tested on the original dataset as we want to compare how specifically deactivating neurons used to recognize particular facial features (essentially removing the facial feature) affects performance.

### 3.5 GRAD-CAM Activation Maps

We use activation heat maps, particularly through the GRAD-CAM (Gradient-weighted Class Activation Mapping) method to visualize which parts of an image influence neural network decisions, especially in CNNs. This visualization helps us determine which layer of the model is most responsive to specific features of the input image. We call these layers critical layers and these are the layers where we induce modifications.

The purpose of Grad-CAM provide insight into the decision-making of CNNs by highlighting areas of the input image that are important for predictions. This method is particularly useful for understanding which layers respond to specific features, such as facial parts (eyes, nose, mouth) in a facial recognition system.

We apply the Grad-CAM technique to generate heat maps for the undistorted faces and each masked facial feature on early layers (con1_1 to conv3_3), middle layers (conv4_1 to conv4_3) and later layers (conv5_1 to conv5_3) of the VGGFace 16 model. To determine the critical layers, we compare heat maps across different layers for the same image to see which layers are most responsive to specific features. We usually compare the normal image heat maps with the masked feature heatmaps over the same set of layers. The layer in which the normal image heat map has bright colors for a specific feature but dark colors in the masked feature heat map is the layer that is the most responsive to that specific masked feature. Thus, by applying Grad-CAM to both original and masked images, we can observe changes in layer activations, revealing how each layer processes the presence or absence of specific features.

### Results of Grad-CAM

## Later Layers:

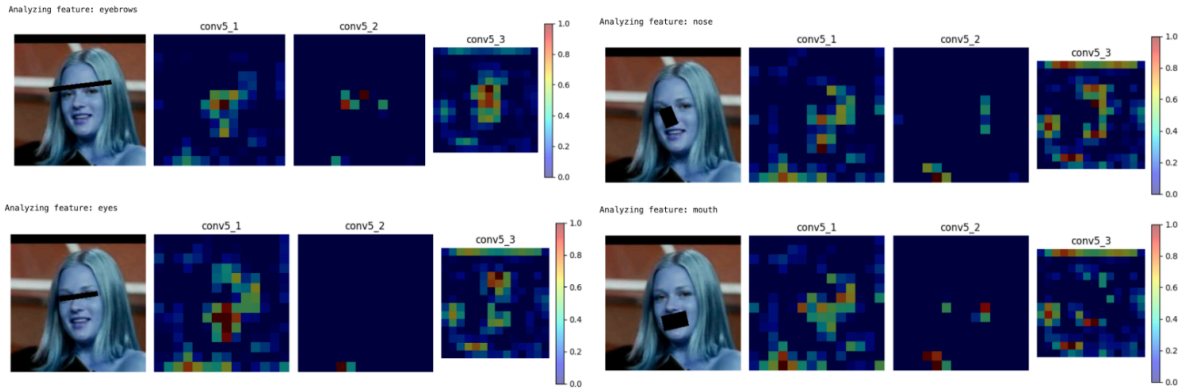



Figure 3: Heatmap Later Layers

## Middle Layers

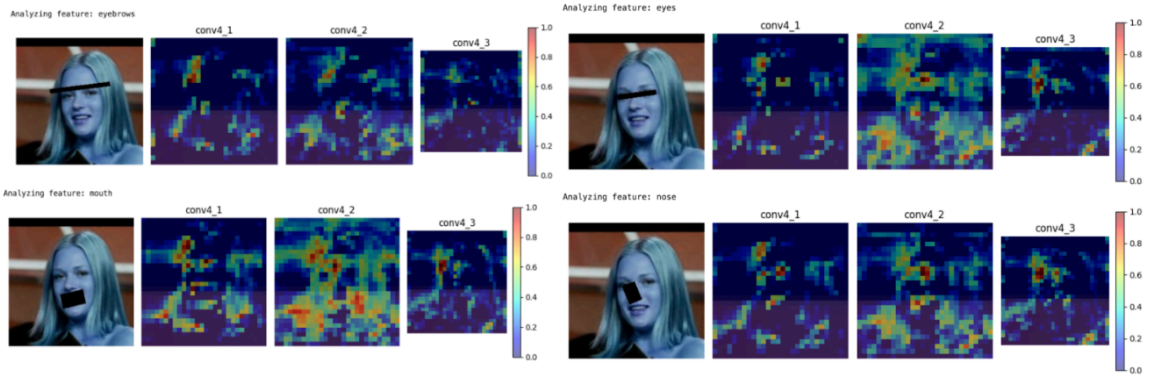



Figure 4: Heatmap Middle Layers

## Early Layers

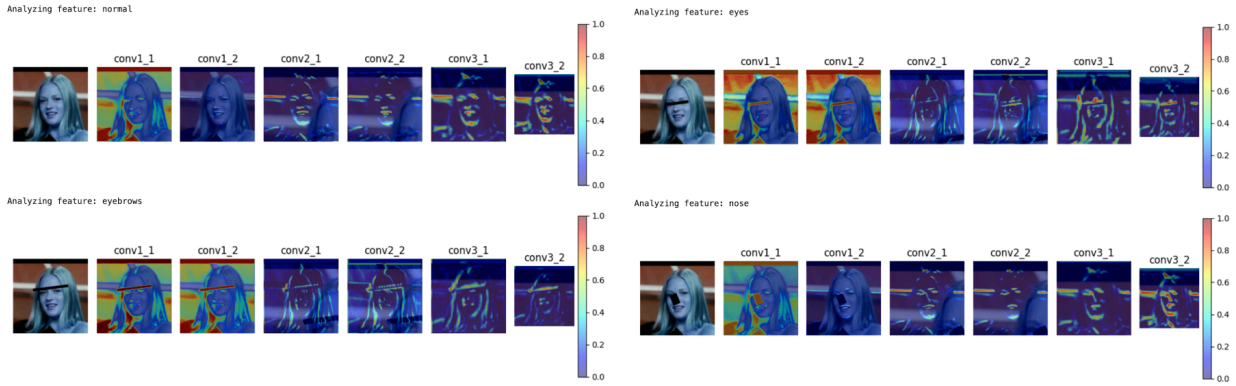



Figure 5: Heatmap Early Layers

Based on this analysis, we have determined that the critical layers are **conv3_2, conv4_2, and conv5_3.**

## 4.0 Results and Analysis

The modified models were evaluated based on accuracy. Additionally, model performance was observed using time taken for each step. Finally, we were able to run the modified model on the original dataset and achieve the reduced accuracy, which showed that the model was mimicking a facial recognition defect.

### 4.1 Original model

| Original Model | Normal Data | Masked Eyebrows | Masked Eyes | Masked Nose | Masked Mouth |
|---|---|---|---|---|---|
| Average Accuracy | 72.98% | 61.59% | 54.25% | 68.96% | 50.28% |
| Precision | 0.995040061 | 0.9952531646 | 0.9936541512 | 0.9945705824 | 0.9952114924 |
| Recall | 0.9905051272 | 0.2389817629 | 0.7136346373 | 0.7652867452 | 0.9952114924 |
| F1 Score | 0.9927674153 | 0.3854166667 | 0.8306808134 | 0.8649924877 | 0.9706168515 |

Table 2: Performance of original model

We include a line chart to visualize the different metrics for each model on the original dataset.

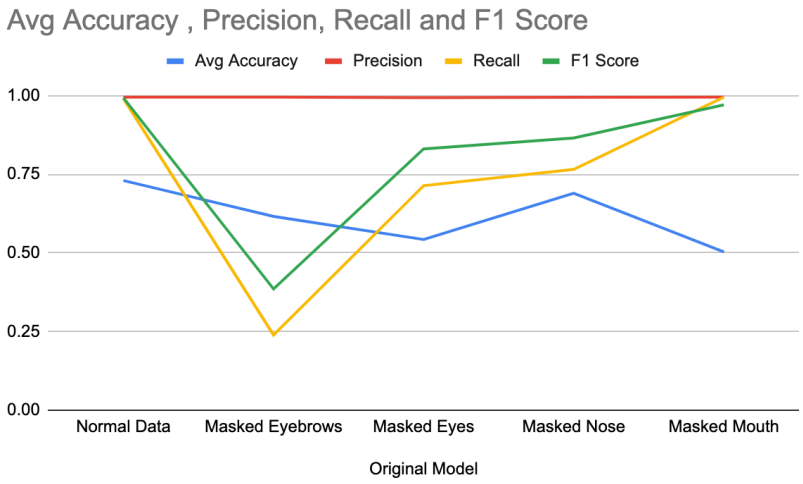



Figure 6: Original Model Performance

### 4.2 Original Dataset

The original model outperforms the modified ones, aligning with our intended objective. Each of these modifications reduced the ability of the model to detect faces. The original model consistently demonstrated superior performance compared to the modified versions across all evaluated metrics, in line with our primary objective of degrading model performance to simulate facial recognition deficiencies. Specifically, the feature abrasion, noise injection, and dropout modifications led to

significant reductions in accuracy, with the average accuracy dropping to 26.79%, 37.96%, and 46.94%, respectively, compared to the original model's accuracy of 72.98%. This suggests that each modification successfully hindered the model's ability to accurately detect faces. (See Table 2).

Interestingly, we observed varying degrees of impact among the different modifications, with feature abrasion resulting in the most drastic decrease in accuracy. These findings underscore the importance of understanding the nuanced effects of different modifications on model performance.

### 4.3 Model Analysis

| **Model** | **Original** | **Feature Abrasion** | **Noise injection** | **Dropout** | **Combined Model** |
|---|---|---|---|---|---|
| No masking | 72.98% | 26.79% | 37.96% | 46.94% | 22.67% |
| Eyebrows masked | 61.59% | 68.15% | 31.24% | 36.69% | 39.44% |
| Eye masked | 54.25% | 85.13% | 72.80% | 60.24% | 33.28% |
| Nose masked | 68.96% | 51.3% | 73.55% | 65.61% | 28.56% |
| Mouth masked | 50.28% | 38.33% | 55.27% | 40.11% | 21% |

Table 3: Average accuracies on original and masked datasets

To analyze the data and discuss the performance changes of the VGGFace 16 model, we will compare the average accuracy for each model across the different datasets in comparison to the original model:

1. Feature Ablation Model: This model saw a significant drop of 46.19% in average accuracy with original data, suggesting that ablation of the identified CAM critical layers substantially impacted the model’s ability to detect faces. This goes to show that neurons in the critical layers are responsible for detecting facial features and ablating them significantly reduces performance.
2. Noise Injection Model: This noise injection at critical layers caused a significant drop of 35.02% in average accuracy on normal data. This result is hypothesized that injecting gaussian noise at critical layers will effectively confuse the model and will degrade performance.
3. Dropout Model: The dropout model experienced a moderate decrease of 26.04% in average accuracy compared to the original model. This goes to show that the random dropout simulated at

the critical layers effectively dropped values of neurons responsible for detecting features for facial recognition.

4. Combined Model: Combining feature ablation, noise injection, and selective dropout led to a lower accuracy drop of 50.31% on normal data and masked feature data, suggesting the simultaneous application of different modifications can compound their effects and lead to a greater loss of facial recognition capability.

The varying increase and decrease of average accuracy on the masked eyes, masked eyebrows, masked nose, and masked mouth are expected as the model was trained using the deficit dataset and the varying changes may be due to unbalanced data.

### 4.4 Neuron Intervention - Critical Layers

In the neuron intervention, we find the neurons in each critical layer that is responsible for detecting a particular facial feature and then turning it off to see how it affects average accuracy. Based on the average accuracies in Table 4, we can determine which convolutional layer is the most sensitive to each facial feature by identifying the layer with the lowest accuracy for that feature. Sensitivity refers to the impact on the detection of the feature after turning off neurons in that particular layer. A higher sensitivity has a larger, resulting in lower accuracy.

| **Critical Layers** | **Eyebrows** | **Eyes** | **Nose** | **Mouth** |
|---|---|---|---|---|
| conv5_3 | 63.13% | 56.94% | 52.26% | 70.83% |
| conv4_2 | 66.55% | 58.56% | 36.16% | 70.83% |
| conv3_2 | 56.94% | 57.49% | 50.49% | 46.66% |

Table 4: Average Accuracies for on Original Dataset for Neuron Intervention Models

Here are the layers most sensitive to each feature:

- Eyebrows: The most sensitive layer for detecting eyebrows is **conv3_2** with an accuracy of 56.94%. This layer's low accuracy for eyebrows indicates its role in capturing essential textures and edges, which are crucial for eyebrow recognition, and deactivating neurons here likely strips away critical details needed to identify them.
- Eyes: The most sensitive layer for detecting eyes is **conv5_3** with an accuracy of 56.94%. This layer's sensitivity to the eyes suggests it's responsible for integrating complex visual patterns that

define the eyes, and impairing this layer disrupts the model's ability to discern these intricate features.

- Nose: The most sensitive layer for detecting the nose is **conv4_2** with an accuracy of 36.16%. This layer shows the highest sensitivity to the nose, implying it handles the spatial and shading cues necessary for nose identification, with neuron deactivation here leading to significant recognition loss.
- Mouth: The most sensitive layer for detecting mouth is **conv3_2** with an accuracy of 46.66%. The sensitivity to the mouth in this layer implies it detects the critical contours and contrasts defining the mouth's structure, and its lower accuracy reveals the importance of these mid-level features in mouth detection.

### 4.5 Masked brows dataset - Feature Abrasion

The masked brows dataset was expected to have the lowest accuracy according to the discussion in Section 2.4.4. For most models, we do see a decrease in average accuracy, yet we do not see eyebrow masked dataset having the lowest accuracies for all models. As can be seen in Table 3, feature abrasion even has a higher accuracy with the masked eyes dataset compared to the original model on the original dataset. At first, this may seem unexpected or unlikely, yet it is a positive outcome in this scenario. The model has a higher accuracy for a dataset with eyes blacked out means that the model could not have used the eye to detect faces. In other words, the feature abrasion done to the original model took away its ability to use eyes in recognition of faces.

Also to be noted, the eyebrows masked dataset had a slightly higher accuracy compared to other datasets in the original model, but was one of the lower ones for other models. Further investigation is warranted to understand the underlying factors contributing to this unexpected finding.

## 5.0 Discussion

In conclusion, this project provides valuable insights into the simulation of facial recognition deficiencies using convolutional neural networks. The implemented methodologies offer avenues for further research in understanding and addressing neurological deficits in facial recognition systems. By simulating aspects of conditions such as prosopagnosia, this work contributes to the development of more robust and inclusive facial recognition technologies.

We have used different modifications of the original VGG-16 models, such as by introduction of dropout layers, noise injection, feature ablation, and neuron intervention to see how we can mimic these deficiencies. We found that these manipulations did indeed hinder the performance of the original dataset with faces. Yet, when we ran the original model on modified datasets with certain features blacked out, it did not perform very well. Most importantly, these results align pretty closely, which means the ability of the model to use these particular features in face recognition was successfully diminished.

### 5.1 Important Findings

The main comparison we draw is between the accuracy of the original model on the original dataset and augmented models on the original dataset.

| | **Average accuracy** | **Precision** | **Recall** | **F1 Score** |
|---|---|---|---|---|
| Original model, original dataset | 72.98% | 0.995040061 | 0.9905051272 | 0.9927674153 |
| Feature Ablation | 26.79% | 0.9944303797 | 0.7459172047 | 0.8524305556 |
| Noise Injection | 37.96% | 0.994123408423114 | 0.385491834409418 | 0.5555555555555 |
| Dropout | 46.94% | 0.993961352657004 | 0.312571211545765 | 0.475585091014157 |
| Neuron Manipulation | 63.13% | 0.994832041343669 | 0.438663121914166 | 0.608856088560885 |
| Original model, augmented datasets | 58.77% | 0.9946723477 | 0.6782786594 | 0.7629267048 |
| Modified models, original dataset | 55.20% | 0.9946137282 | 0.5292008748 | 0.6649632053 |

Table 5: Summary of findings

Table 5 summarizes the results achieved. Figure 7 helps in visualizing this by taking the average of all metrics achieved by all modified models with the original dataset. By plotting it as a line chart, we can clearly see that the average results for modified models on the original dataset is very similar to the modified dataset on the original model.

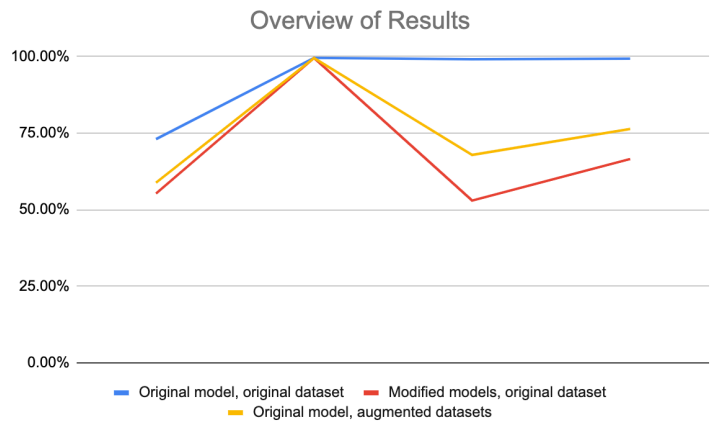



Figure 7: Overview of Results

## 5.2 Unexpected Observations

The results observed from the various models and their runs are largely in line with what we expected, yet there are a few things that are contrary to literature mentioned or the hypothesis.

### 5.2.1 Eyebrows masked - Original model

In line with the findings mentioned in Section 4.2.4, we expected that eyebrows would have the most detrimental effect on the accuracy of the models, “eyebrows lead to the most significant deficits in face recognition performance” (Müller et al., 2024). However, contrary to our expectations, our analysis revealed unexpected findings. While this was true when compared to other modified models, as can be seen in Table 5 of this paper, we found that the original model performed the worst when the mouth was masked, achieving only 50.28% accuracy, and had the best performance with the eyebrows masked datasets compared to other models. Further, it performed better

This discrepancy suggests that critical facial features and their influence on recognition may be more nuanced than previously thought. It raises questions about the complexity of facial features, model

specificity, data variability, interplay of features, and random variation. These could explain the unexpected result seen.

### 5.2.2 Consistently high precision values

As can be seen in Table 6, the precision values were consistently very high for all models and all dataset combinations. Precision is defined as the number of true positives over the total number of positives predicted. In our datasets, all images are faces, therefore all predictions made as positive will be correct. Therefore, all precision values are rightly 100%.

## 5.3 Summary of Results

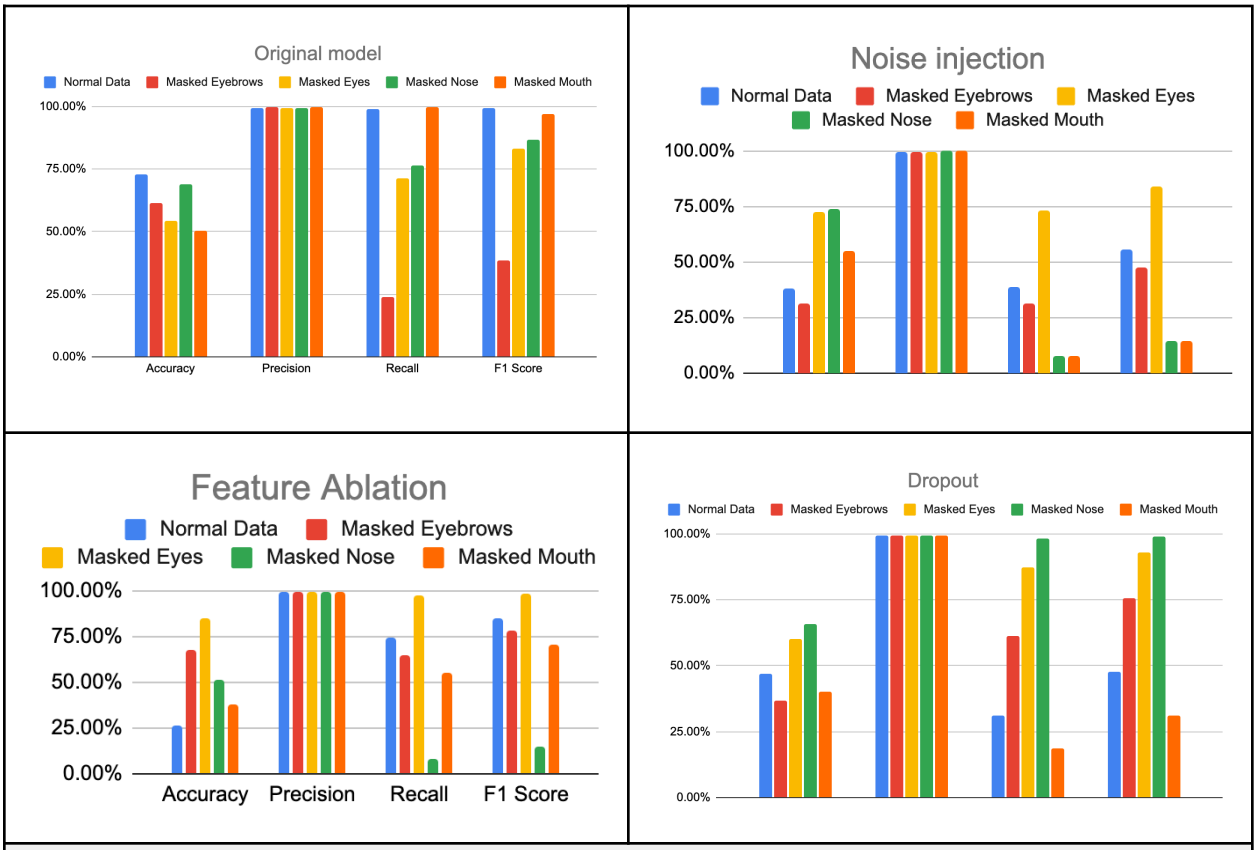



Table 6: Summary of results

From our research, we can conclude the following:

1. The activation heat maps help us deduce the critical layers from the early layers to the later layers in the model which are involved in detecting facial features and recognizing faces. These critical layers are found to be conv3_2, conv4_2, and conv5_3. In the altered models, we induce

modifications such as feature ablation, Gaussian noise, and selective dropout. We find that the average accuracy significantly decreases when run on the original dataset in comparison to the original model. This shows that the critical layers are important in facial detection and altering neurons in those layers will significantly degrade performance as expected.

2. The combined model with all the modifications - feature ablation, noise, and selective dropout - degraded the performance of the model by dropping accuracies to the range of 21-34%. This shows that altering the neurons in the critical layers simultaneously compounds the effects to simulate a strong and impactful neurological deficit.
3. Through the neuron intervention model, we found that certain critical layers are more sensitive to certain facial features than others. For eyebrows and mouth, early layers such as conv3_2 contain the most sensitive neurons for detecting these features as they are responsible for edges and lines. The middle layer conv4_2 contains neurons most responsible for detecting the nose as it is responsible for contours and contrasts. Finally, the later layer conv5_3 is the most sensitive to eyes as it is responsible for integrating complex visual patterns. From this we can discern that different layers in the CNN of VGGFace16 are responsible for detecting certain facial features, which all then come together during pooling.

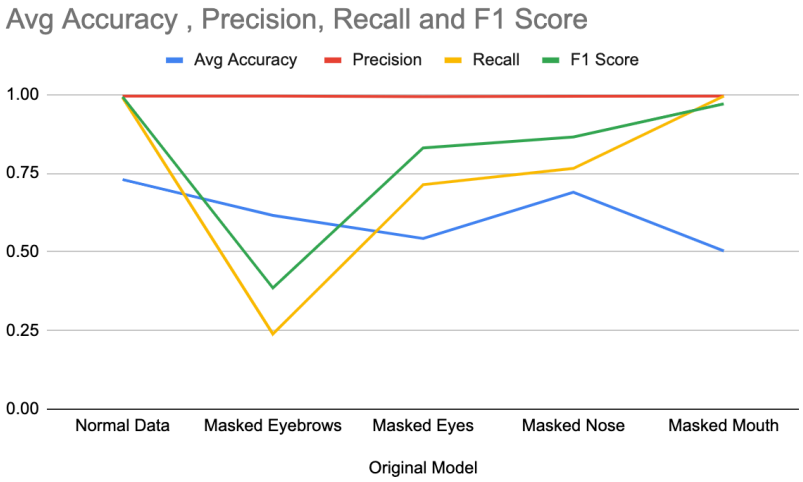
## 5.4 Interpretation of results

### 5.4.1 Comparing datasets

In conclusion, the original model's performance decreased with varying degrees for each of the augmented datasets, as expected. The highest degradation in accuracy being caused by masked eyebrows, once again, a part of our hypothesis thanks to literature cited.

We include a line chart to visualize the different metrics for each model on the original dataset. Figure 8 clearly shows us that most metrics were the worst for eyebrow blacked out dataset, and masked mouth dataset had the least effect on all metrics (except accuracy). That being said, accuracy is the only metric that was repeated and averaged out over 5 runs, with all others only being recorded from the last run.

Thus these data do not mean that the eyebrow augmented dataset worsened the performance of the model




the most.

Figure 8: Original model line graph

### 5.4.1 Comparing Model Modifications

The various alterations made to the model consistently yielded notably lower average accuracy compared to the original model. This outcome is consistent with our initial hypothesis, as the modifications were implemented specifically to diminish the performance of the original model, mimicking deficits in facial recognition. The alignment of the results with our hypothesis suggests the effectiveness of our approach in inducing targeted impairments in model performance.

The performance of the different modified models varied across the datasets, with certain models demonstrating better performance on specific datasets while others performed better on different ones. Overall, there is no consistent trend in the performance of the models across all datasets, indicating the importance of considering the specific characteristics of each model and dataset when evaluating model performance. There could also be specific alignments with some models and certain datasets which could be leading to better performance.

### 5.4.2 Performance metric values

Since we are solving a face detection classification problem with all our datasets and models, the values of metrics achieved can be pretty misleading due to the nature of the task. The dataset consists solely of faces, with some features being blacked out and some images having difficult faces to detect. Thus, since the dataset might be skewed, the accuracy value can vary quite a bit. The classification models could be tested much better if we had an equal number of images belonging to each category - face, not face

(Mishra, 2018), yet that is not what we wanted in this study. So for the purpose of this study, the dataset was exactly as we'd like, but it is not ideal for the metrics to be of much significance.

### 5.5 Future Directions

- Further experimentation with different impairment techniques and model architectures to better simulate neurological deficits.
- Exploration of real-world applications, such as assistive technologies for individuals with facial recognition impairments.
- Collaboration with experts in neuroscience and psychology to refine the simulation methodologies and enhance the understanding of facial recognition deficiencies in humans.

## 6.0 References


Barrow, E., Eastwood, M., & Jayne, C. (2016). Selective Dropout for Deep Neural Networks. In A. Hirose, S. Ozawa, K. Doya, K. Ikeda, M. Lee, & D. Liu (Eds.), Neural Information Processing. ICONIP 2016. Lecture Notes in Computer Science (Vol. 9949). Springer. https://doi.org/10.1007/978-3-319-46675-0_57

Mishra, A. (2018, February 24). Metrics to Evaluate your Machine Learning Algorithm. Towards Data Science. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

Müller, S. M., Riesel, A., & Keil, A. (2024). Eyebrows are more important than eyes for deep neural face recognition. Frontiers in Computational Neuroscience, 18, 1209082. https://doi.org/10.3389/fncom.2024.1209082

Nakada, M. (2017). AcFR: Active Face Recognition Using Convolutional Neural Networks. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). DOI:10.1109/CVPRW.2017.11. Retrieved from ResearchGate: https://www.researchgate.net/publication/319284653_AcFR_Active_Face_Recognition_Using_Convolutional_Neural_Networks#pf3

OpenAI's ChatGPT, Response to "Format these references in APA format" ChatGPT, OpenAI, February 23, 2024. https://chatgpt.com

Pei, Z., Xu, H., Zhang, Y., Guo, M., & Yang, Y.-H. (2019). Face Recognition via Deep Learning Using Data Augmentation Based on Orthogonal Experiments. Electronics, 8(10), 1088. https://doi.org/10.3390/electronics8101088

## 7.0 Appendix

| | Normal Data | Masked Eyebrows | Masked Eyes | Masked Nose | Masked Mouth |
|---|---|---|---|---|---|
| **Original Model** | | | | | |
| Accuracy | 72.98% | 61.59% | 54.25% | 68.96% | 50.28% |
| Precision | 0.995040061 | 0.9952531646 | 0.9936541512 | 0.9945705824 | 0.9952114924 |
| Recall | 0.9905051272 | 0.2389817629 | 0.7136346373 | 0.7652867452 | 0.9952114924 |
| F1 Score | 0.9927674153 | 0.3854166667 | 0.8306808134 | 0.8649924877 | 0.9706168515 |
| | | | | | |
| **Feature Ablation** | | | | | |
| Accuracy | 26.79% | 68.15% | 85.13% | 51.30% | 38.33% |
| Precision | 0.9944303797 | 0.9953703704 | 0.9945987654 | 1 | 0.9931833674 |
| Recall | 0.7459172047 | 0.6532472465 | 0.9791112799 | 0.07823775161 | 0.5546250476 |
| F1 Score | 0.8524305556 | 0.788809906 | 0.9867942584 | 0.1451215217 | 0.7117733268 |
| | | | | | |
| **Noise Injection** | | | | | |
| Accuracy | 37.96% | 31.24% | 72.80% | 73.55% | 55.27% |
| Precision | 0.9941234084 | 0.9915254237 | 0.9948266943 | 1 | 1 |
| Recall | 0.3854918344 | 0.3111702128 | 0.7303456134 | 0.07823775161 | 0.07823775161 |
| F1 Score | 0.5555555556 | 0.4736842105 | 0.8423127464 | 0.1451215217 | 0.1451215217 |
| | | | | | |
| **Dropout** | | | | | |
| Accuracy | 46.94% | 36.69% | 60.24% | 65.61% | 40.11% |
| Precision | 0.9939613527 | 0.9938309685 | 0.9947984395 | 0.9950076805 | 0.9938650307 |
| Recall | 0.3125712115 | 0.6120820669 | 0.8716293202 | 0.9840486137 | 0.1850019033 |
| F1 Score | 0.475585091 | 0.7575828827 | 0.9291497976 | 0.9894978041 | 0.3119383825 |
| | | | | | |
| **Combined Model** | | | | | |
| Accuracy | 22.67% | 39.44% | 33.28% | 28.56% | 21% |
| Precision | 0.9966101695 | 0.9951876805 | 0.9931740614 | 0.9894459103 | 0.9890510949 |
| Recall | 0.2233194075 | 0.3928571429 | 0.3315609571 | 0.2848461831 | 0.2063189951 |
| F1 Score | 0.3648774434 | 0.5633342414 | 0.4971526196 | 0.4423473902 | 0.3414173228 |

**Table 7: Full Dataset Summary**