Analysis of HMAX Model for Object Recognition Task And Evaluation of a Confidence Metric

Authors
Zahra Maleki rosamaleki1382@gmail.com
Shervin Mehrtash ishervinmehrtash@gmail.com

Abstract

Object recognition is a fundamental task in both biological and artificial vision systems. The HMAX model, inspired by the hierarchical processing of the human visual cortex, has been widely used to extract features at multiple levels, from simple edge detection (S1) to complex shape representation (C2). In this study, we investigate the effectiveness of the HMAX model for an object recognition task and evaluate a confidence metric associated with classification performance. To achieve this, we first conducted an experiment in which two human subjects classified images into animal and non-animal categories. We recorded their responses, reaction times, and confidence levels to assess human performance in the task. Subsequently, we used the HMAX model to extract hierarchical features from the same dataset and trained both Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) classifiers using C2-level features. The trained models were then evaluated in terms of accuracy and confidence scores to compare their performance with human participants. Our findings provide insights into the alignment between biological and computational object recognition mechanisms, highlighting the role of confidence as a key metric in evaluating model reliability.

1 Introduction

Visual neuroscience focuses on understanding how the brain processes visual information, enabling us to perceive the world. This field explores the structure and function of the visual system, including the eye, retina, and brain regions responsible for vision. Visual input begins at the retina, where light is converted into electrical signals. These signals travel through the optic nerve to the lateral geniculate nucleus (LGN) and then to the primary visual cortex (V1), where basic features like edges, orientation, and motion are processed.
Advanced processing occurs in higher cortical areas like V4 and the inferotemporal cortex, which help recognize complex patterns, colors, and objects. Visual neuroscience bridges the gap between biology and perception, explaining how neural pathways transform raw sensory data into meaningful visual experiences [10].

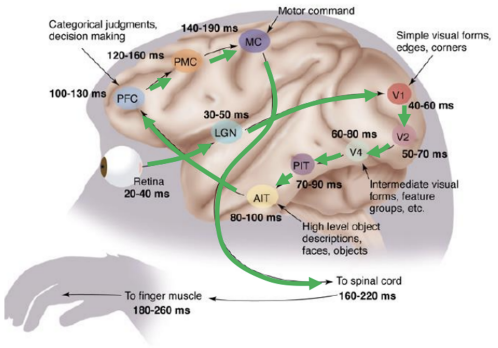


Figure 1: Ventral and Dorsal Visual Pathways

As you can see in Figure 1 The dorsal and ventral pathways are two distinct visual processing streams in the brain, originating from the primary visual cortex (V1) and serving different roles in perception and action.

- The dorsal pathway, also known as the ”Where” or ”How” pathway, extends from the V1 through the parietal lobe and is responsible for processing spatial information, motion, and guiding actions. It determines an object’s location, movement, and how to interact with it, such as catching a ball by calculating its trajectory in real-time. Damage to the dorsal stream can lead to deficits like difficulty in perceiving motion or spatial neglect [2].

- The ventral pathway, also known as the ”What” pathway, projects from the V1 to the inferior temporal cortex and specializes in object recognition and form perception. It identifies ”what” an object is by analyzing features like color, shape, and texture. This pathway is essential for recognizing faces, objects, and visual details. Damage to the ventral stream can cause conditions like visual agnosia, where individuals cannot recognize objects despite intact vision [2].

Together, these pathways work in parallel to integrate what we see with where it is and how to act upon it, enabling seamless interaction with our environment.

1.1 Object Recognition

Object recognition is the ability of the visual system to identify and differentiate objects, a process fundamental to interacting with the environment. This begins with the perception of simple visual features such as edges, lines, and orientations in the primary visual cortex (V1). As visual information advances hierarchically,

areas like V2 and V4 integrate these basic features into intermediate forms, such as shapes and contours. Eventually, the anterior and posterior inferior temporal (AIT and PIT) cortices process this information to identify whole objects, faces, or complex stimuli. The system achieves invariance, meaning it can recognize objects despite changes in size, lighting, angle, or position. Such robustness is achieved through hierarchical pooling of features across neural layers, where increasing complexity allows for accurate recognition under varying visual conditions [10].

Neuroscientific models of object recognition are inspired by both biological systems and computational approaches. For instance, feedforward models describe how visual signals move sequentially through neural layers, with each stage building upon the previous. Studies such as those by Riesenhuber and Poggio (1999) [6] demonstrate that neurons progressively encode more complex features while maintaining selectivity for specific objects. Additionally, the visual system relies on mechanisms such as context effects and priming, where prior knowledge or surrounding stimuli influence recognition speed and accuracy.

1.2 Decision Making

Decision-making is the process of evaluating alternatives and selecting an action. We tend to search for information related to the problem at hand, estimate the probabilities of different alternatives, and attach meanings and values to anticipated outcomes. Therefore, decisions are a choice among courses of action. People who must make too many decisions too quickly have to trade off the speed of decision-making against the accuracy of decision outcomes. In addition, the time to make an accurate decision is related to the amount of uncertainty in the decision. Naturally, the more uncertainty we have, the longer it takes us to search for the information, estimate probabilities of different alternatives, and attach values to outcomes[1].

1.3 Confidence

Confidence is the belief in the accuracy or correctness of a decision, action, or outcome. When we make choices, confidence reflects how sure we are that our decision is the right one. For example, if you answer a question and feel very sure it's correct, your confidence level is high. Confidence can be influenced by experience, the information available, and how often we've been correct in similar situations before. In the brain, confidence often involves areas like the prefrontal cortex, which processes reasoning and decision-making, and it helps guide actions by assessing how likely we are to succeed[3].

1.4 Goal of This Research

This research is concentrated on understanding the brain's visual system. In this study, we design a psychophysics task, collect behavioral data, and analyze the results. Additionally, we will analyze the HMAX model, a computational model inspired by the visual cortex. We will also explore how this model extracts visual features and, after conducting analysis, we will compare its results with human behavioral data. Moreover, we will calculate confidence levels using this model.

2 HMAX Model

The HMAX model (Hierarchical Model and X) is a biologically inspired computational model of object recognition, designed to replicate the hierarchical feedforward processing observed in the primate visual cortex. It was initially developed by Riesenhuber and Poggio and has since been refined in numerous studies [7], [8].

2.1 Overview of the Model

The HMAX model mirrors the functionality of the ventral visual pathway, specifically the primary visual cortex (V1) to the inferotemporal cortex (IT). It uses a hierarchical approach where features are processed in alternating layers of simple (S) and complex (C) units. The alternating structure allows the model to achieve a tradeoff between selectivity (identifying specific features) and invariance (robustness to transformations like position, scale, and rotation).

- **S Layers:** Simple units perform feature extraction using a Gaussian tuning operation. They detect specific patterns such as edges, orientations, and contours.
- **C Layers:** Complex units pool information from S units via a max operation, enabling position and scale invariance.

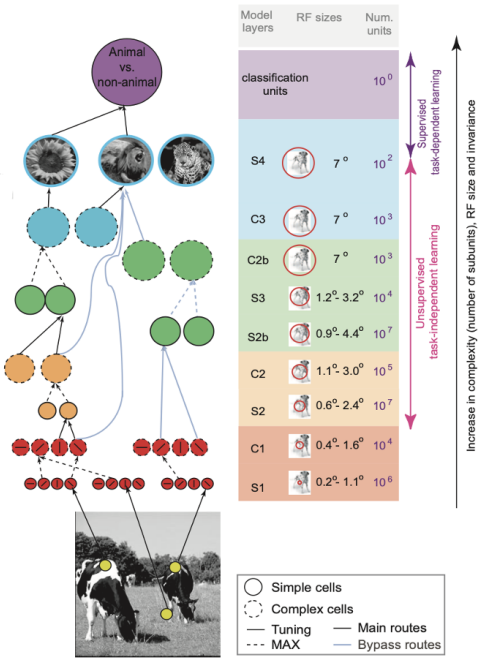


Figure 2: An illustration of the HMAX model structure, showing the hierarchical flow of information from simple to complex representations.

2.2 How the Model Works

The HMAX model processes images through the following stages:

1. **S1 Layer (Input Stage):** The input image is analyzed by S1 units, which detect low-level features such as edges at specific orientations and scales, similar to V1 simple cells.
2. **C1 Layer (Complex Representation):** Outputs from S1 units are pooled to form position- and scale-invariant representations, resembling the behavior of V1 complex cells.
3. **S2 Layer (Intermediate Complexity):** At this stage, S2 units detect combinations of features such as corners and contours by pooling responses from C1 units.

4. **C2 Layer:** C2 units further pool S2 outputs across positions and scales, producing a dictionary of robust, invariant features.

5. **Higher Layers (S3 and Beyond):** The final stages (e.g., S3) integrate increasingly complex features, akin to visual processing in higher cortical areas like V4 and IT.

2.3 Biological and Computational Insights

- **Rapid Categorization:** It explains the brain's ability to quickly classify objects, even in cluttered or degraded scenes.
- **Robustness:** The model handles transformations such as changes in position, scale, and rotation effectively.
- **Feature Learning:** Using unsupervised learning, the model adapts to the statistics of natural images, tuning its feature detectors for better performance.

3 Dataset

3.1 Dataset Description

The dataset consists of gray-scale 256 × 256 pixel images, divided into 600 animal stimuli and 600 non-animal stimuli. Animal stimuli are split into four categories: head, near-body, medium-body, and far-body. In the category of non-animal stimuli, there are 300 images of natural scenes and 300 images of artificial scenes, which are divided into four groups according to the mean distance from the camera: head ($\leq$ 1 m), close-body (5–20 m), medium-body (50–100 m), and far-body ($\geq$ 100 m).
The set of stimuli is divided into two separate training and test sets, with the same number of exemplars in each of the four mentioned groups, as presented in Figure 3 The dataset is divided into two subsets. Each subset contains 600 images with an equal number of images in each category. We used one of the subsets as a training set and the other one as the test set.

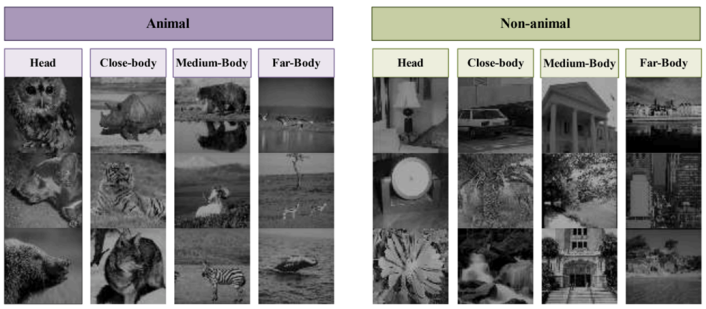


Figure 3: A Few Examples of The Dataset

4 Behavioral Task

4.1 Task Paradigm

This experiment consists of 1200 trials, divided into 10 blocks, with each block containing 120 trials. Subjects should be allowed to rest between blocks. In the first 5 blocks, training images will be used, while the remaining blocks will use test images. Each trial consists of a sequence of events in which either an animal or a non-animal image is randomly presented. Figure 4 shows the task paradigm:

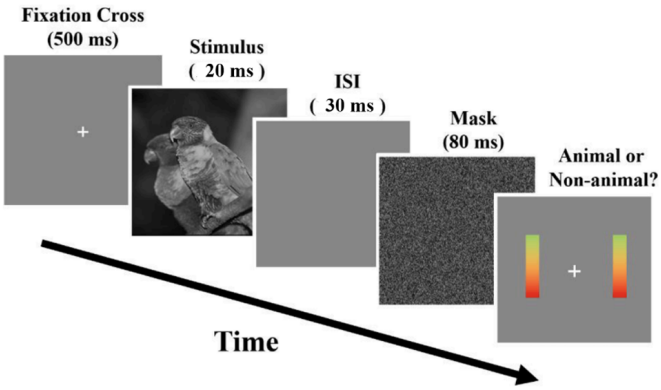


Figure 4: Behavioral Task Paradigm

The sequence for each trial is as follows:

1. Fixation Cross: A plus sign (+) will appear in the center of the screen for 500 ms.
2. Stimulus Presentation: After the fixation cross disappears, a stimulus (either an animal or non-animal image) will be shown for 20 ms.
3. Gray Screen (ISI): The screen will then display a gray color for 30 ms.
4. Mask: A masked version of the image will be shown for 80 ms.
5. Decision Screen: Subjects are then asked to make a decision regarding whether the presented image was an "Animal" or a "Non-Animal." The screen displays two color bars as columns: if their choice is "Animal," they should select the right column, and for "Non-Animal," they should select the left column. The height of the bar indicates their confidence level, with a higher green portion representing greater confidence and a higher red portion representing lower confidence. Confidence height is between 0 and 1.

It should be noted that in order to evaluate the performance of subjects in challenging scenarios, we have recorded data in two other situations as well, where images were noisy and rotated.
The following figure represents a subject during data acquisition:

Figure 5: A data recording session where the subject is instructed to select whether the shown image is "animal" or "non-animal"

4.2 Subject Results

As stated earlier, we have extracted the performance, reaction time, and confidence of each subject individually. In order to find the performance, we have collected the correct and incorrect answers and then found the correctness as a measure to represent the performance. In terms of reaction time (RT), we have considered the time between the demonstration of the image and the response of the subject to be the RT. Also, the confidence is evaluated based on the location where the subject has chosen on the confidence bars shown on the monitor.

The following tables represent the performance, RT, and confidence for each subject (and each category of images individually) and the mean of both subjects (using original, noisy, and rotated images):

Subject 1:

	All	Head-View	Near-Body	Medium-Body	Far-Body
Original	0.93	0.93	0.95	0.95	0.89
Rotated	0.87	0.95	0.86	0.87	0.80
Noisy	0.89	0.92	0.90	0.91	0.82

(a) Performance

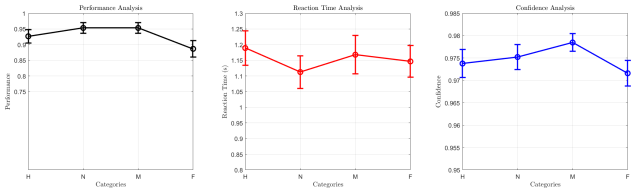
	All	Head-View	Near-Body	Medium-Body	Far-Body
Original	1.16	1.19	1.11	1.17	1.15
Rotated	1.24	1.20	1.31	1.17	1.28
Noisy	1.19	1.15	1.13	1.02	1.45

(b) Reaction Time

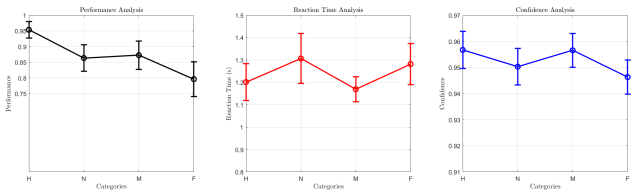
	All	Head-View	Near-Body	Medium-Body	Far-Body
Original	0.98	0.97	0.98	0.98	0.97
Rotated	0.96	0.96	0.95	0.96	0.95
Noisy	0.95	0.96	0.96	095	0.94

(c) Confidence

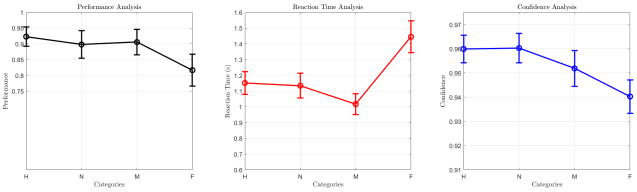
Figure 6: Evaluation of results for Subject 1



(a) Original Images



(b) Rotated Images



(c) Noisy Images

Figure 7: Visual Representation of results for Subject 1

Subject 2:

	All	Head-View	Near-Body	Medium-Body	Far-Body
Original	0.86	0.90	0.89	0.87	0.78
Rotated	0.81	0.84	0.86	0.71	0.83
Noisy	0.82	0.82	0.81	0.88	0.78

(a) Performance

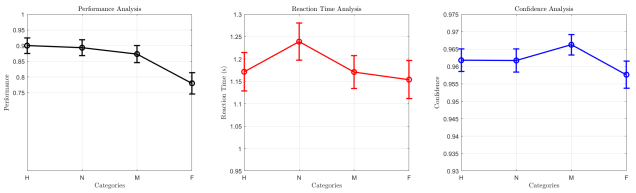
	All	Head-View	Near-Body	Medium-Body	Far-Body
Original	1.18	1.17	1.24	1.17	1.15
Rotated	1.17	1.24	1.14	1.15	1.15
Noisy	1.19	1.10	1.16	1.24	1.24

(b) Reaction Time

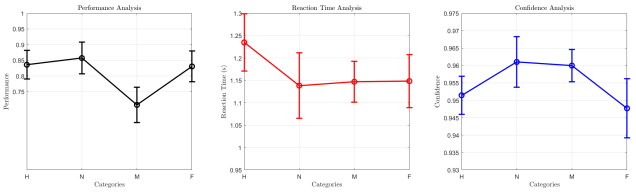
	All	Head-View	Near-Body	Medium-Body	Far-Body
Original	0.96	0.96	0.96	0.97	0.96
Rotated	0.96	0.95	0.96	0.96	0.95
Noisy	0.97	0.97	0.96	0.97	0.96

(c) Confidence

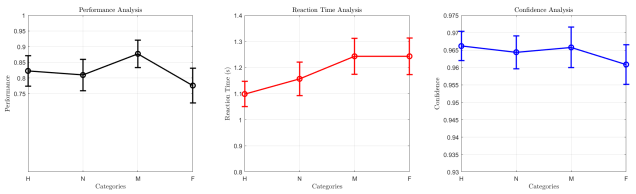
Figure 8: Evaluation of results for Subject 2



(a) Original Images



(b) Rotated Images



(c) Noisy Images

Figure 9: Visual Representation of results for Subject 2

The results indicate a correlation among the extracted data. Specifically, higher performance is associated with lower reaction times and higher confidence levels. Additionally, performance tends to be better in both the Head View and Near-Body View, suggesting that subjects can more easily recognize animals in these views compared to the Middle View or Far View. Moreover, when analyzing noisy and rotated images, performance decreases in these more challenging conditions, but there remains a notable ability to distinguish between animal and non-animal figures.

It is important to note that the correlation between the three extracted parameters (performance, reaction time, and confidence) is not always consistent. This discrepancy may arise because a subject's choices could be influenced by intuition or other mental states,

which were not accounted for in the analysis. Consequently, performance is identified as the most reliable parameter for accurate comparisons across different views (Head, Near-Body, etc.) and image qualities (normal, noisy, or rotated).

Average of Subjects:
The following tables represent the average results of both subjects:

	All	Head-View	Near-Body	Medium-Body	Far-Body
Original	0.90	0.92	0.92	0.91	0.84
Rotated	0.84	0.90	0.86	0.79	0.82
Noisy	0.86	0.87	0.86	0.90	0.80

(a) Performance

	All	Head-View	Near-Body	Medium-Body	Far-Body
Original	1.17	1.18	1.17	1.17	1.15
Rotated	1.21	1.22	1.23	1.16	1.22
Noisy	1.19	1.13	1.15	1.13	1.36

(b) Reaction Time

	All	Head-View	Near-Body	Medium-Body	Far-Body
Original	0.97	0.97	0.97	0.98	0.97
Rotated	0.96	0.96	0.96	0.96	0.95
Noisy	0.96	0.97	0.96	0.96	0.95

(c) Confidence

Figure 10: Evaluation of Average Results for Both Subjects

The analysis of average results confirms the previous findings, indicating that performance is slightly lower for rotated and noisy images. Meanwhile, the Head View and Near-Body View yield higher performance and shorter reaction times. As mentioned earlier, inconsistencies were observed in the reaction time and confidence data, which can be attributed to the subjects' unclear mental state or varying performance during the data recording session.

4.3 Questions

4.3.1 Why the stimulus presentation period is very short?

The stimulus presentation period is very short (20 ms) to test participants' ability to recognize the image quickly and unconsciously, tapping into their automatic visual processing capabilities. This brief exposure ensures that the task relies more on implicit memory and rapid perceptual processes, rather than deliberate, conscious recognition. By shortening the presentation time, the task reduces the possibility of subjects consciously analyzing the image, making it more of an implicit recognition task.

4.3.2 What is the role of ISI in the paradigm?

The ISI (30 ms gray screen) serves multiple purposes:

- **Distraction and Reset:** It provides a brief break between the stimulus and the mask, allowing the visual system to reset and reducing the possibility of lingering afterimages from the initial stimulus. This helps to maintain the accuracy of the mask and subsequent decision-making.
- **Segmentation of Stimuli:** By separating the stimulus presentation and the mask, the ISI creates a clear boundary between the two, making it easier for the brain to process each component (stimulus and mask) independently.
- **Attention Shift:** It acts as a cue for participants to prepare for the next phase of the task, encouraging the brain to shift attention away from the stimulus and focus on the mask or decision phase.

4.3.3 What is the role of Mask in the paradigm?

The mask (displayed for 80 ms) plays a crucial role in interfering with conscious recognition of the stimulus. Since the mask is a scrambled version of the original image, it effectively blurs or disrupts the visual information, preventing participants from fully processing the stimulus. This ensures that the recognition task primarily tests the initial, automatic processing of the image before the mask is shown. The mask is used to limit the duration of conscious processing, so participants are forced to rely on their implicit visual memory rather than detailed analysis.

4.3.4 What were your challenges during implementing this task and collecting data from subjects? How did you overcome them?

The primary challenge in this task was to make sure that the subject made a reasonable decision when considering the image as animal or non-animal. In order to overcome this challenge and make sure that this incorrect decision-making chance does not affect the entire process, 5 blocks of data were considered as train datasets and were used to make the subject familiar with the nature of the task. This initial encounter with the task paradigm significantly increases the decision-making ability of the subjects.

4.3.5 What was the effect of challenging scenarios (Noise/Rotation) on accuracy, response time, and confidence?

As mentioned earlier, the challenging scenarios that included the noisy and rotated images resulted in lower performance (accuracy) for all subjects and across all image views. The expected effect of these challenging scenarios on the response time and confidence is to make these parameters higher and lower, respectively. Although this behavior is observed in some cases, in other cases there exists a violation which is a result of the subject's mental state during the data recording session. It simply means that subjects may choose an answer with uncertainty but make their response too fast and with a high reported confidence. In such cases, only the accuracy of the answers is the reliable parameter for further analysis.

5 Model Training and Evaluations

5.1 Training the Model

After running the MATLAB script `demoRelease.m`, the output consists of C2 features, which are high-level representations of the images processed by the HMAX model. The script first loads images from the specified dataset directories, including training and testing sets for animal and non-animal categories. It then applies Gabor filters to extract C1 features, capturing edge and texture information. These C1 features are further processed to compute C2 features, which serve as robust descriptors for classification tasks. The extracted C2 features are saved in a `.mat` file, allowing them to be used for training classifiers such as Support Vector Machines (SVM) and Multi-Layer Perceptron (MLP).

5.2 Classification

Support Vector Machine (SVM) is a supervised learning algorithm that finds an optimal hyperplane for separating classes in a high-

dimensional space. Given a dataset $\{(x_i, y_i)\}_{i=1}^{n}$ with labels $y_i \in \{-1, 1\}$, SVM solves the following optimization problem:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad s.t. \quad y_i(w^T x_i + b) \geq 1, \quad \forall i$$

where w is the weight vector and b is the bias.

The model achieved an accuracy of **76.5%** in the test set.

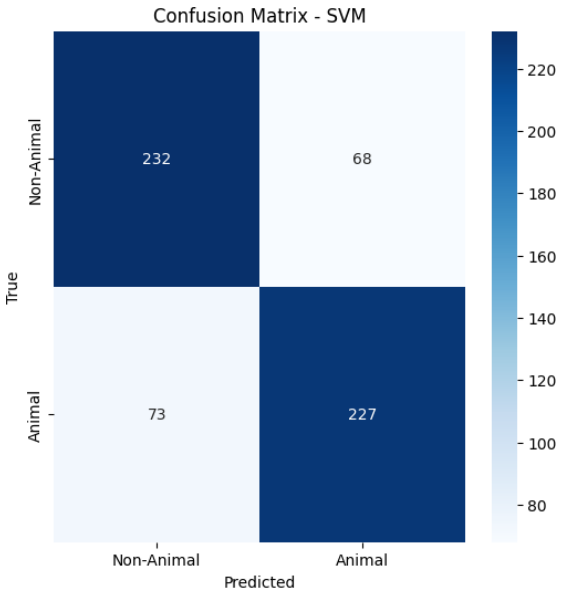


Figure 11: Confusion Matrix - SVM

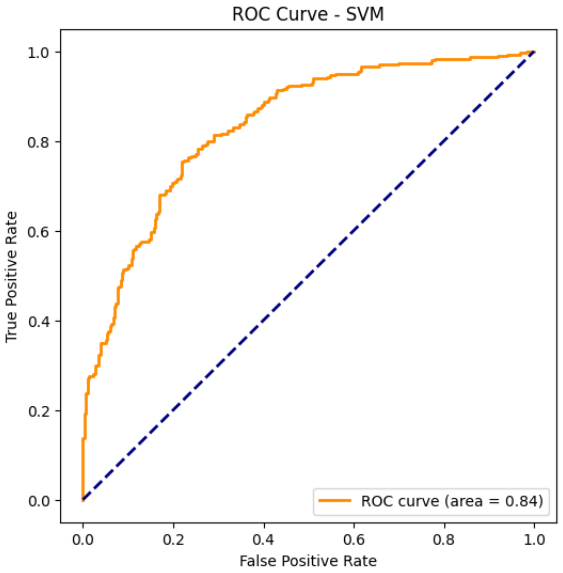


Figure 12: ROC Curve - SVM

Multi-Layer Perceptron (MLP) is a type of artificial neural network consisting of an input layer, one or more hidden layers, and an output layer. The forward pass in an MLP is given by:

$$h = \sigma(W_1 x + b_1), \quad y = \sigma(W_2 h + b_2)$$

where W_1, W_2 are weight matrices, b_1, b_2 are biases, σ is an activation function, and h represents the hidden layer activations.

The model achieved an accuracy of **61.7%** in the test set.

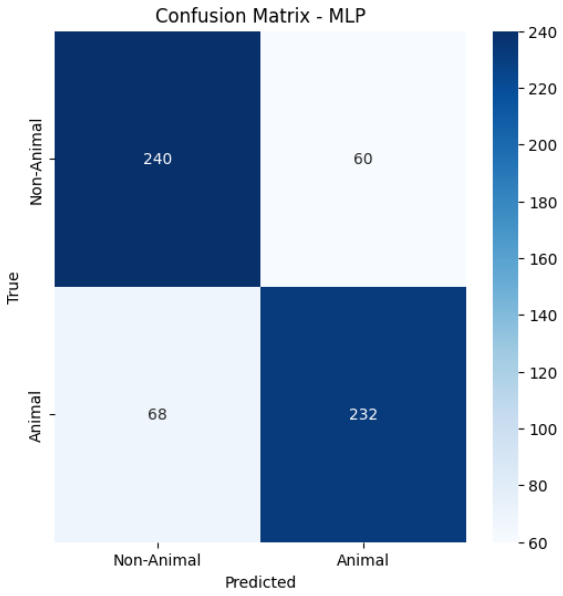


Figure 13: Confusion Matrix - MLP

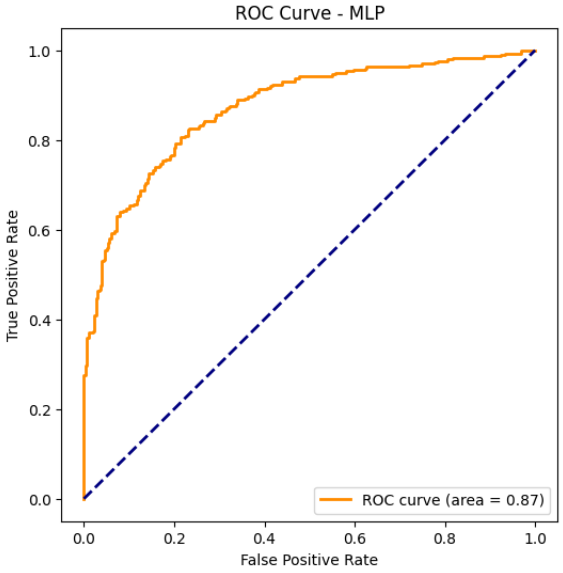


Figure 14: ROC Curve - MLP

Accuracy measures the proportion of correctly classified instances out of the total test samples, providing a general performance indicator of the classifier. Area Under the Curve (AUC) from the Receiver Operating Characteristic (ROC) curve quantifies the model's ability to distinguish between classes, where a higher AUC indicates better classification performance. While accuracy gives an overall success rate, AUC evaluates the model's sensitivity to different classification thresholds.

5.3 Task

5.3.1 Comparing Table

The comparison of two methods is given in Table 1.

Metric	SVM	MLP
Validation Accuracy	1.0000	1.0000
Test Accuracy	0.7650	0.7867
AUC	0.84	0.87

Table 1: Performance Metrics for SVM and MLP

5.3.2 Classifier Performance Comparison

Based on the results, **MLP performs better than SVM** on the given data. While both classifiers achieve perfect validation accuracy, **MLP has a higher test accuracy (0.7867 vs. 0.7650) and a better AUC score (0.87 vs. 0.84)**, indicating that MLP generalizes slightly better and has a stronger ability to distinguish between classes.

5.3.3 accuracy for all categories

The dataset consists of four categories: head, close-body, medium-body, and far-body. The head category includes close-up images focusing on the head of an object or animal. Medium-body images show the entire body from a moderate distance, while far-body images depict subjects from a greater distance.

The SVM classifier performed consistently across all categories, with the highest accuracy in the head category and slightly lower scores for medium-body and far-body In contrast, the MLP classifier performed well on head images but showed a significant drop in accuracy for distant categories, for far-body images. This suggests that SVM maintains better stability across all categories, while MLP struggles with recognizing subjects in distant images.

Category	SVM	MLP
Head	77.33	74.67
Close-body	70.67	54.67
Medium-body	66.67	48.00
Far-body	66.67	30.67

Table 2: Category-wise Accuracy

5.3.4 Category-wise Performance

The plot shows the category-wise comparison of SVM and MLP across three metrics: accuracy, reaction time, and confidence. SVM maintains a relatively stable performance across all categories, while MLP's accuracy declines significantly for distant categories. Reaction time remains similar across both classifiers, whereas SVM shows higher confidence scores overall.

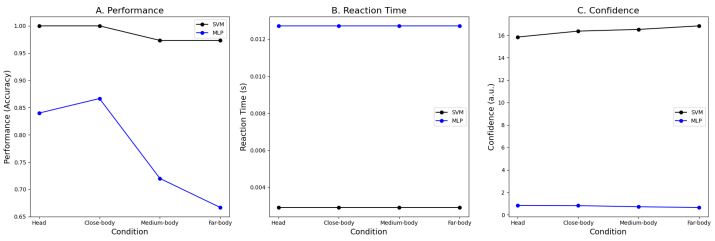


Figure 15: Comparison of performance, reaction time, and confidence.

5.3.5 Effect of Hyper-parameter Changes

For SVM, the linear and RBF kernels achieve higher accuracy, while the polynomial kernel performs worse, suggesting that simpler kernels generalize better. For MLP, increasing the number of layers or neurons does not significantly improve accuracy but increases reaction time, leading to higher computational costs. SVM maintains higher confidence across configurations, while MLP's confidence remains lower and stable.

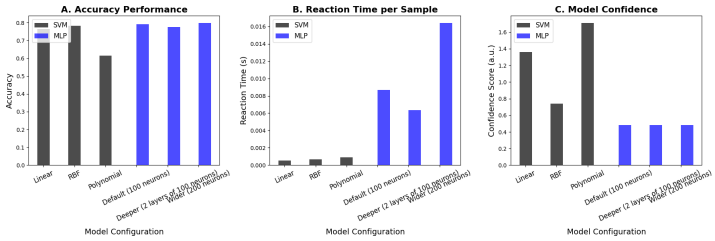


Figure 16: Impact of hyper-parameter changes.

5.3.6 Interpretation of the ROC Curve

The ROC curves for SVM and MLP illustrate the trade-off between the true positive rate and false positive rate at different classification thresholds. The diagonal line represents random guessing, while curves above this line indicate better classification performance. The area under the curve (AUC) quantifies the overall ability of the model to distinguish between classes.

The SVM model achieves an AUC of 0.84, indicating good classification performance. However, the MLP model performs slightly better with an AUC of 0.87, suggesting that MLP has a stronger ability to separate positive and negative classes. The higher the AUC, the better the model is at distinguishing between classes.

5.3.7 Advantages and Disadvantages of Classifiers

Support Vector Machines (SVM) are highly effective for binary classification tasks, especially when the data is well-separated. They work well with smaller datasets and are less prone to overfitting due to the use of regularization techniques. SVMs are also robust to high-dimensional spaces and can handle both linear and non-linear classification using kernel functions. However, they can be computationally expensive, especially with large datasets, and require careful tuning of hyperparameters such as the kernel type and regularization parameter. Additionally, SVMs struggle with noisy data and overlapping class distributions, making them less suitable for complex real-world problems.

Multi-Layer Perceptrons (MLP) offer greater flexibility and can model complex relationships in data by leveraging multiple hidden layers and non-linear activation functions. They excel in handling large datasets and can automatically learn important features, making them ideal for deep learning applications. However, MLPs require extensive computational resources, large amounts of labeled data, and careful tuning of hyperparameters such as the number of layers, neurons, and learning rate. They are also prone to overfitting, especially when the model is too complex or trained on small datasets, requiring techniques like dropout and batch normalization to improve generalization.

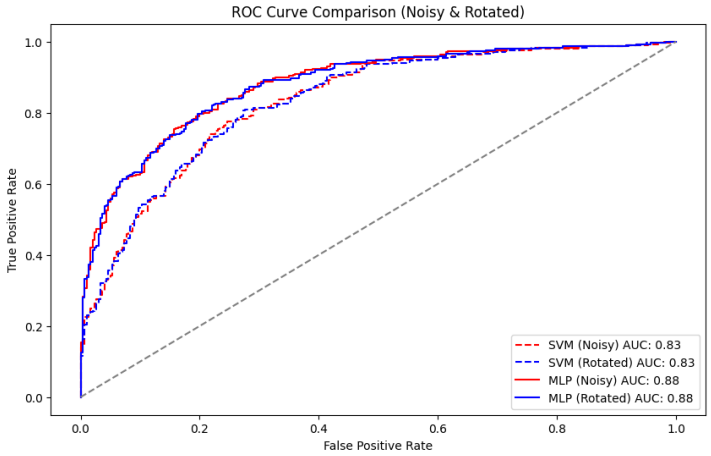


Figure 17: ROC curve for Noisy and Rotated dataset

5.4 robustness of the HMAX model

5.4.1 Evaluating the robustness

To evaluate the robustness of the classifiers, two modified test sets were created by adding Gaussian noise and applying random rotations to the images.

Metric	SVM N	SVM R	MLP N	MLP R
Acc	0.7617	0.7533	0.7933	0.7950
AUC	0.83	0.83	0.88	0.88

Table 3: Comparison on Modified Test Sets (N:Noisy,R:Rotated

The results show that both SVM and MLP maintained stable performance, with MLP demonstrating slightly better adaptability to distortions. The AUC values remained consistent across all test conditions, indicating that both classifiers retained their ability to distinguish between classes despite the transformations. However, minor variations in accuracy suggest that noise and rotations affect classification performance differently, with MLP showing greater resilience.

5.4.2 comparing the classifier

MLP appears to be more robust to noise and rotation compared to SVM. While both classifiers maintained similar AUC values across all test conditions, MLP showed more stable accuracy, indicating better adaptability to distortions. This can be attributed to the model's ability to learn complex patterns through its multiple layers, making it less sensitive to small perturbations in the input data. In contrast, SVM relies heavily on margin-based separation, which can be affected by noise and geometric transformations. Therefore, MLP demonstrates greater resilience to variations in image quality.

5.4.3 Improving the HMAX Model

To enhance the robustness of the HMAX model against noise and rotation, several modifications can be considered. One approach is augmenting the training data with artificially generated noisy and rotated images, allowing the model to learn invariant features. Additionally, incorporating deeper hierarchical processing with more complex feature extraction layers could improve its ability to capture structural patterns that remain stable under transformations. Another potential improvement is integrating normalization techniques or adaptive thresholding in feature selection to reduce sensitivity to noise. Finally, hybridizing HMAX with deep learning methods, such as convolutional neural networks (CNNs), could enhance feature representation and adaptability to distortions.

5.5 Effect of Dimension Reduction

5.5.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms a dataset into a new coordinate system where the most significant variations are captured along the principal components.

First, the dataset is standardized by subtracting the mean and scaling to unit variance. Then, the covariance matrix is computed:

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T \tag{1}$$

where x_i is the i-th data point, μ is the mean vector, and n is the total number of samples.

Next, the eigenvalues and eigenvectors of the covariance matrix are determined by solving:

$$Cv = \lambda v \tag{2}$$

where λ represents the eigenvalues, and v represents the eigenvectors, which form the principal components. The eigenvalues are sorted in descending order, and the top k components that capture the most variance are selected. The data is then projected onto these selected components:

$$Z = XW \tag{3}$$

where W is the matrix of selected eigenvectors. This transformation reduces dimensionality while preserving essential information, making PCA useful for visualization and improving computational efficiency in machine learning tasks.

5.5.2 Task and Analysis

PCA was applied to reduce the dimensionality of the C2 features while retaining 95% of the variance. This reduced the feature count from 1000 to 178. Both SVM and MLP classifiers were then trained and tested using the reduced dataset.

Figure 18 presents the ROC curves for both classifiers after PCA, showing that the AUC values remain largely unchanged compared to the original dataset. Figure 19 displays the confusion matrices, indicating a slight reduction in classification accuracy.

Table 5.5.2 compares the accuracy and AUC of SVM and MLP before and after applying PCA. The accuracy of both classifiers slightly decreased, reflecting the trade-off between dimensionality reduction and classification performance. However, the AUC values remained consistent, suggesting that PCA preserves the overall decision-making capability of the classifiers.

Metric	SVM (PCA)	MLP (PCA)
Accuracy	0.7600	0.7750
AUC	0.84	0.87

Table 4: Comparison of Classifiers Before and After PCA

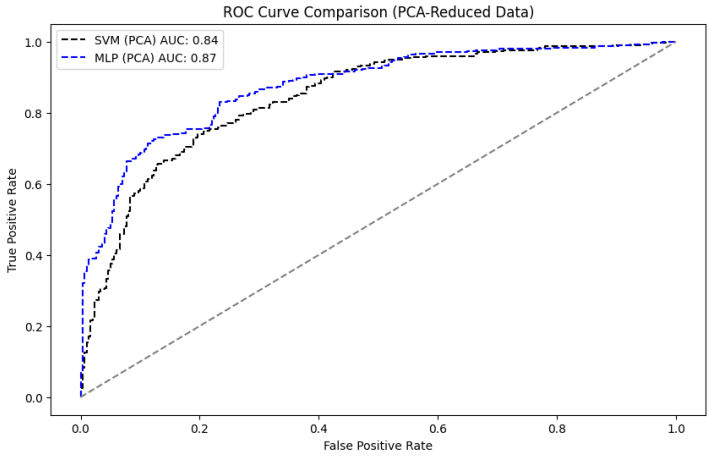


Figure 18: ROC Curve Comparison After PCA

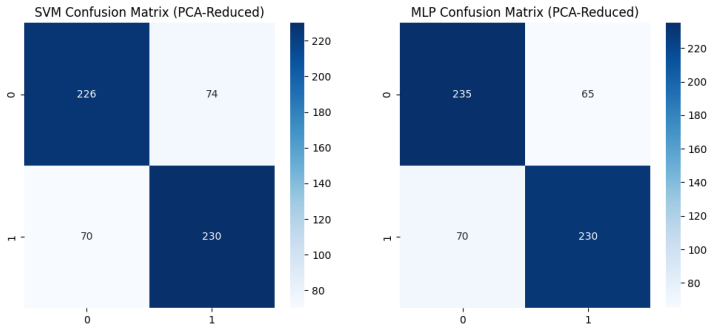


Figure 19: Confusion Matrices for SVM and MLP After PCA

5.5.3 Questions

1. PCA reduces the dimensionality of the dataset while preserving most of the variance, leading to a more compact representation of features. For both SVM and MLP classifiers, the accuracy slightly decreases after PCA, indicating a minor loss of information. However, the AUC values remain unchanged, suggesting that the classifiers still retain their ability to distinguish between classes. This trade-off shows that while PCA can improve computational efficiency and reduce redundancy, it may slightly degrade classification accuracy due to the loss of some feature details.

2. Applying PCA improves computational efficiency by reducing the number of features, leading to faster training and inference times for both SVM and MLP classifiers. With fewer dimensions, SVM benefits from reduced complexity in finding the optimal decision boundary, while MLP experiences lower computational overhead in training and backpropagation. This makes the models more efficient, especially for large datasets, without significantly compromising classification performance.

3. Using PCA for dimensionality reduction in this context has both advantages and disadvantages.

   Advantages: PCA reduces the number of features while retaining most of the variance, leading to improved computational efficiency and faster training times for both SVM and MLP classifiers. It helps eliminate redundant or less informative features, potentially reducing overfitting. Additionally, by transforming data into a new coordinate system, PCA can improve model generalization on unseen data.

   Disadvantages: Despite maintaining most variance, PCA can lead to a slight loss of discriminative information, which may result in reduced classification accuracy, as seen in the performance drop after applying PCA. Moreover, the transformed features lack interpretability since PCA generates new feature representations that may not have direct real-world meaning. Lastly, selecting the optimal number of principal components requires careful tuning to balance between performance and dimensionality reduction.

6 Exploring Confidence

6.1 Brief Explanation

When using classifiers like Support Vector Machines (SVM) and Multi-Layer Perceptrons (MLP) for object recognition, obtaining probability estimates instead of just binary decisions is crucial for understanding model confidence and improving decision-making. Analyzing confidence scores is important because it provides insight into model uncertainty, allowing for better handling of ambiguous cases. It also enables threshold adjustments for applications where precision or recall is more critical. Furthermore, probability-based outputs facilitate integration with downstream decision systems, such as Bayesian frameworks or ensemble methods, leading to more robust and interpretable predictions.

6.2 Calculating Absolute Difference of the Outputs

In order to find the confidence metric for each sample, we use the following equation:

$$Confidence = |P_{animal} - P_{non-animal}|$$

By finding this metric for each sample of the test inputs (600 test images including animal and non-animal images with different views), we can analyze the confidence of the models.

6.2.1 Confidence Metric Across All Samples in SVM

Analysis of Confidence Score Distribution:
In the SVM model, the confidence metrics across different samples are distributed as follows:

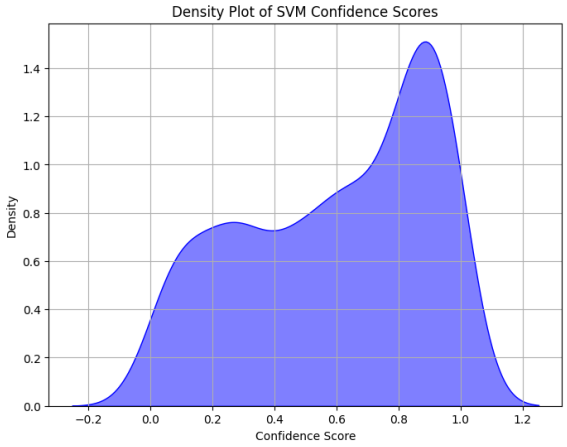


Figure 20: Density Plot of SVM Confidence Scores

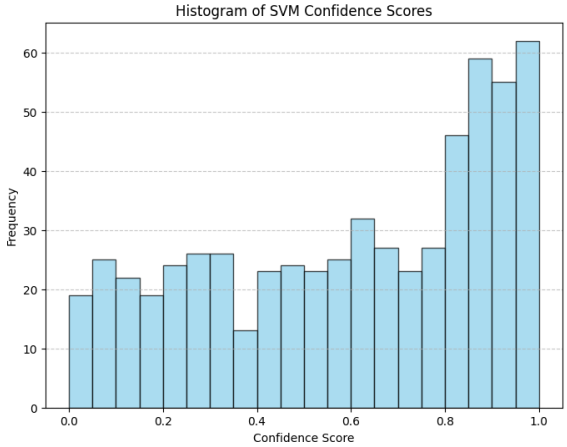


Figure 21: Histogram of SVM Confidence Scores

The elicited results suggest that the confidence metric in SVM model is centered around values higher than 0.6 and is less populated around lower values, meaning that the confidence is relatively high and the outcomes of this model are reliable.

Mean Confidence Score:
The Mean confidence metric for all test samples would be:

$$Mean \simeq 0.60$$

Confidence Score Across Different Categories:
However, the confidence metric can be found for different categories of images individually. The result for individual categories is a bit lower than the accumulative confidence score for the entire dataset of test images, as in this section, the test images from different categories are given to the model individually.

The following figure represents the confidence metric across different categories:

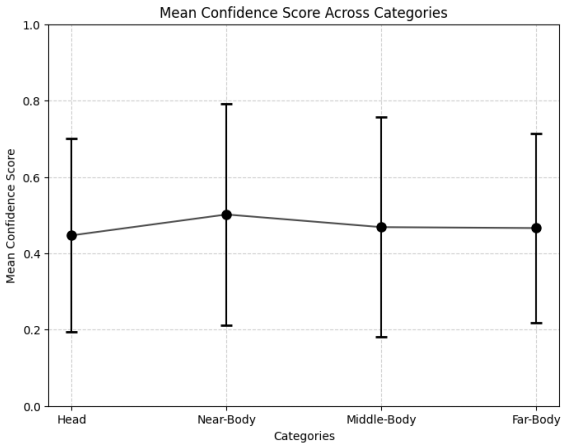


Figure 22: Confidence Metric Across Different Categories in SVM

The confidence score of the SVM model is higher for Head View and Near Body View images because these categories typically contain more distinct and well-defined features that the model can leverage for classification. In Head View, facial structures, contours, and other discriminative features are clearly visible, making it easier for the model to differentiate between objects or individuals. Similarly, in the Near Body View, important details such as clothing textures, body proportions, and defining characteristics are captured with higher resolution and clarity.

Applying PCA:
At this point, we applied the PCA to reduce the input vectors' dimension and then calculated the confidence metrics and plotted them for all categories again.
The following figure represents the confidence metric across different categories after applying PCA:

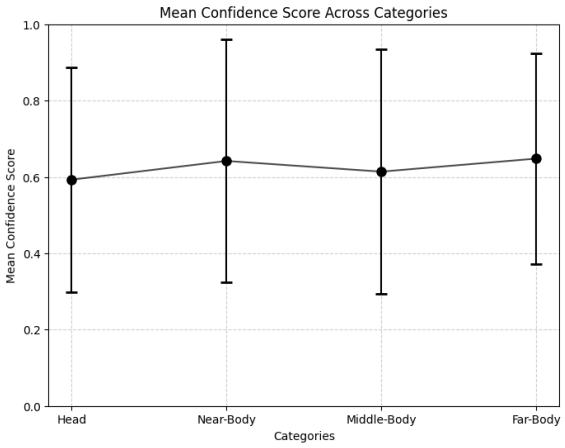


Figure 23: Confidence Metric Across Different Categories in SVM After Applying PCA

Applying Principal Component Analysis (PCA) can increase the confidence score of the SVM model in object recognition tasks because it enhances the signal-to-noise ratio by removing irrelevant or

redundant features and retaining the most discriminative information.

Results for Noisy and Rotated Datasets:
We went through the same process for noisy and rotated images as well and obtained the following outcomes:

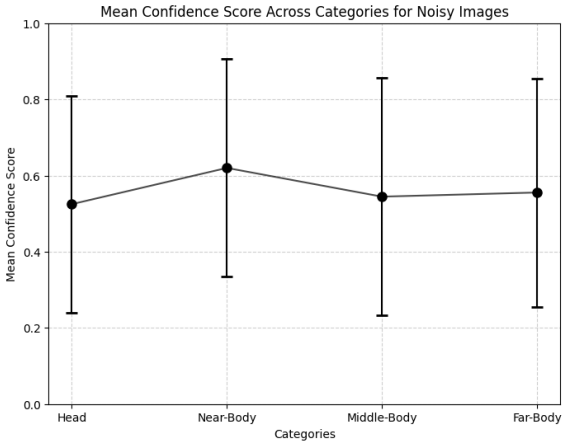


Figure 24: Confidence Metric Across Different Categories in SVM for Noisy Images

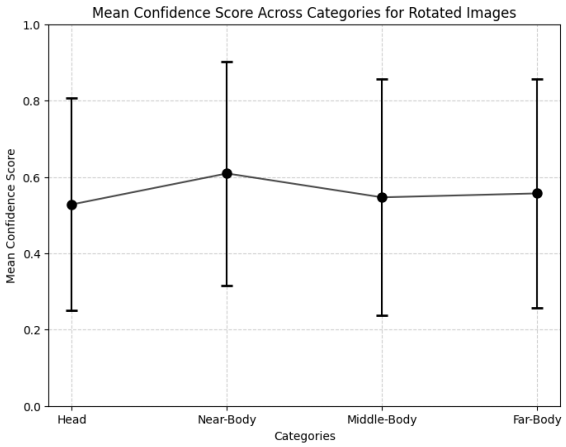


Figure 25: Confidence Metric Across Different Categories in SVM for Rotated Images

The confidence metric is lower in datasets with noisy or rotated images because these distortions negatively impact the quality of features that the SVM model relies on for classification.

6.2.2 Confidence Metric Across All Samples in MLP

Analysis of Confidence Score Distribution:
In the MLP model, the confidence metrics across different samples are distributed as follows:

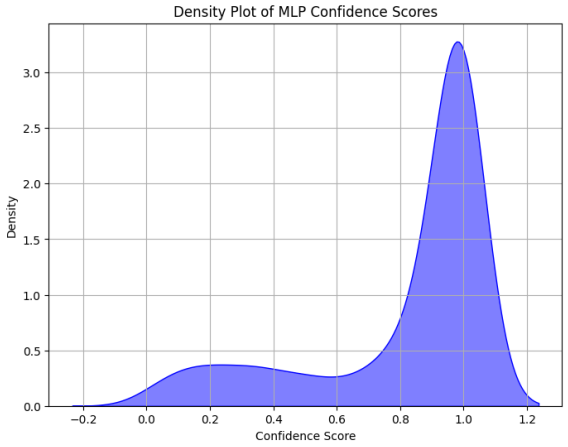


Figure 26: Density Plot of MLP Confidence Scores

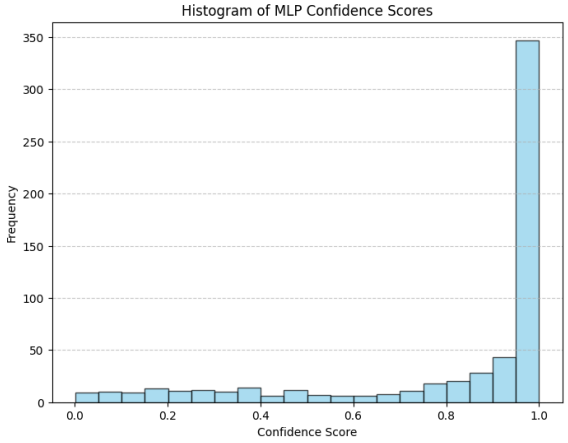


Figure 27: Histogram of MLP Confidence Scores

The elicited results suggest that the confidence metric in MLP model is relatively higher than the SVM model. Above outcomes highly suggest that this classifier demonstrated a better performance and the outcomes are more reliable.

Mean Confidence Score:
The Mean confidence metric for all test samples would be:

$$Mean \simeq 0.82$$

Confidence Score Across Different Categories:
The confidence metric can be found for different categories of images individually. For the MLP classifier, the obtained results are closer to the average of the entire dataset when analyzed without being divided into 4 categories.

The following figure represents the confidence metric across different categories:

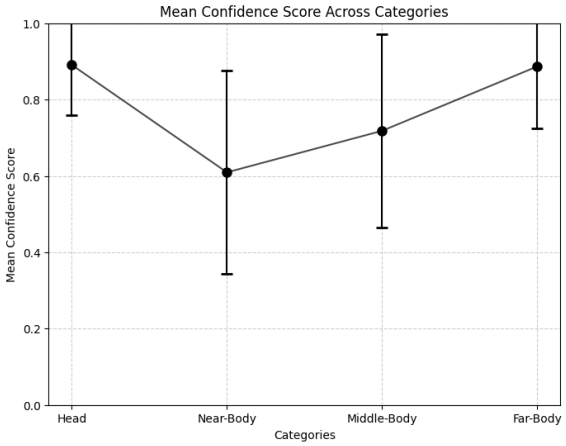


Figure 28: Confidence Metric Across Different Categories in MLP

The confidence score of the MLP model is higher in all different categories compared to the SVM model. However, this model shows a slightly different behavior, as the confidence score for middle and far body views is higher than the near view. This could be due to the characteristics of this classifier.

Applying PCA:
We applied the PCA similar to the previous section to reduce the input vectors' dimension and then calculated the confidence metrics and plotted them for all categories again.
The following figure represents the confidence metric across different categories after applying PCA for MLP:

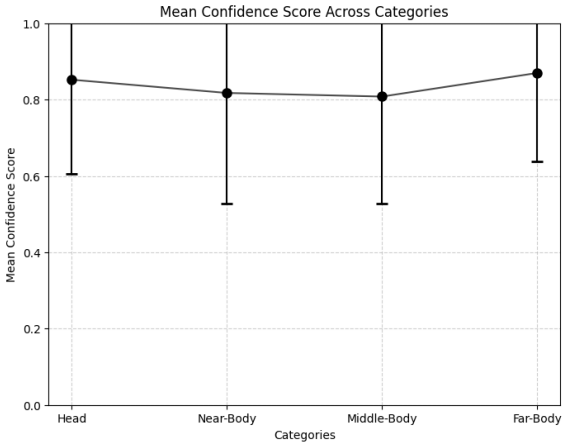


Figure 29: Confidence Metric Across Different Categories for MLP Model After Applying PCA

As shown in the figure above, applying Principal Component Analysis (PCA) increased the confidence score in almost all categories by eliminating unwanted features.

Results for Noisy and Rotated Datasets:
We applied the same methodology to noisy and rotated images, analyzing their impact on recognition performance and confidence metrics. The results obtained from these variations provide further insights and are as follows:

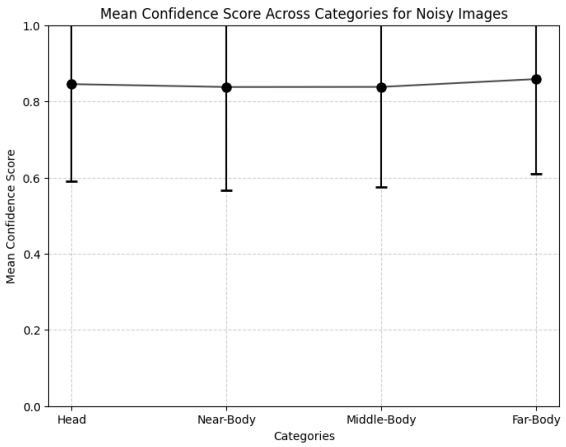


Figure 30: Confidence Metric Across Different Categories in MLP for Noisy Images

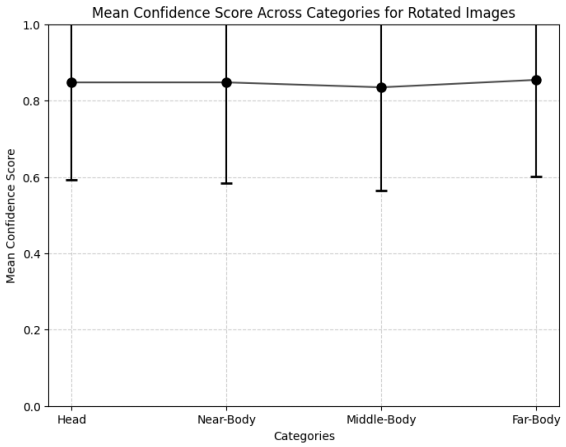


Figure 31: Confidence Metric Across Different Categories in MLP for Rotated Images

The figures above illustrate distinct effects of adding noise or rotating images. First, there is an increase in the confidence score, contrasting with the outcomes observed for the SVM model. Second, the confidence metrics appear more uniform across all categories. A detailed explanation of this behavior is provided in the following section.

6.3 Questions

6.3.1 Compare the confidence scores for the SVM and MLP classifiers and discuss the results. Which classifier provides higher confidence?

Based on the elicited results, the confidence scores are higher in MLP model. However, there are differences in different datasets (normal, noisy, and rotated images) and also across different categories. Generally, the improved confidence scores of the MLP model compared to SVM in this object recognition task can be attributed to several key factors:

- MLP Captures Non-Linear Relationships Better:
  SVM with a linear kernel primarily finds a linear decision boundary, which may not be optimal for object recognition tasks where feature distributions are often non-linearly separable. MLP, with its multiple layers and non-linear activation functions, can learn complex hierarchical representations, making it more robust to variations such as noise and rotation.

- Feature Extraction in HMax (C2 Level) Benefits MLP:
  The C2 features from the HMax model represent complex patterns and object structures. MLP's architecture allows it to better leverage these high-level features for classification compared to an SVM, which might struggle with high-dimensional feature representations.

6.3.2 How does confidence relate to classification accuracy?

While accuracy tells you how often your model is correct, confidence gives you insight into how certain the model is about each prediction.
High confidence typically correlates with high accuracy, but not always. For example, high-confidence predictions (e.g., 0.9 or higher probability) are more likely to be correct, but there can be cases where the model is overly confident in incorrect predictions, especially when the model is poorly calibrated or is likely to misclassify difficult cases.
Analyzing confidence can give you a deeper understanding of model performance, especially in uncertain predictions, beyond just the accuracy metric.

6.3.3 What is the effect of PCA on confidence level?

PCA reduces the dimensionality of the feature space by projecting the original data onto a set of orthogonal components (principal components) that capture the most significant variance in the data. This reduction often leads to a more compact and cleaner feature representation, eliminating less important features (often containing noise or irrelevant variations). As a result, the models (SVM, MLP, etc.) are trained on more discriminative features, which can lead to higher confidence scores as the model's decision boundaries become more distinct and reliable.
In this study, it is seen that the PCA has resulted in higher confidence scores in both models.

6.3.4 What is the effect of noise and rotation on confidence level?

The effect of noise and rotation on the confidence metric is complex and depends on the model's ability to handle such distortions. In this study, the MLP model exhibited an increase in confidence levels after adding noise and rotation, which might seem counterintuitive at first. This outcome can be attributed to the complex structure and multiple layers of the MLP, which allow it to learn robust, high-level features even in the presence of noise and rotation. MLPs are capable of capturing intricate patterns through non-linear activations, and their architecture may help them generalize better, even when faced with distorted data. The model might learn to adapt to these transformations in a way that increases its certainty about the classification decision. However, this result does not provide a straightforward answer, as the impact of noise and rotation on the confidence metric can vary across different datasets and model configurations. Therefore, while we observe this trend in the MLP, it's important to note that we cannot give a solid, universal answer to the question, as the effect of such distortions on the confidence metric is context-dependent.

References

Joshua I. Gold and Michael N. Shadlen. The neural basis of decision making. *Annual Review of Neuroscience*, 30:535–574, 2007.

Melvyn A. Goodale and A. David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992.

Piercesare Grimaldi, Hakwan Lau, and Michele Basso. There are things that we know that we know, and there are things that we do not know we do not know: Confidence in decision-making. *Neuroscience and Biobehavioral Reviews*, 55, April 2015.

M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Schölkopf. Support vector machines. *IEEE Intelligent Systems and Their Applications*, 13(4):18–28, 1998.

Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (PCA). *Computers Geosciences*, 19(3):303–342, 1993.

Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025, December 1999.

Thomas Serre, Aude Oliva, and Tomaso Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, 104(15):6424–6429, 2007.

Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–426, 2007.

Gurpreet Singh and Manoj Sachan. Multi-layer perceptron (MLP) neural network technique for offline handwritten Gurmukhi character recognition. In *2014 IEEE International Conference on Computational Intelligence and Computing Research*, pages 1–5, 2014.

Simon J. Thorpe and Michèle Fabre-Thorpe. Seeking categories in the brain. *Science*, 291:260–263, 2001.