Facial Expression Classification

Deepak K Srivastava, Marcelino Bautista, Xinyi Zhu
{dsri, marcelinob20, xinyi_zhu}@ischool.berkeley.edu
GitHub: https://github.com/UC-Berkeley-I-School/ds-281-facial-recognition

Abstract

This project investigates facial expression recognition using the FER-2013 dataset, focusing on a mix of traditional computer vision methods and modern deep feature embeddings. The system extracts handcrafted features using Histogram of Oriented Gradients (HOG) and facial landmark geometry to capture structural patterns in facial movements. These are combined with high-quality deep embeddings generated from the *Buffalo-L* ArcFace model, which provides a stronger representation of identity-invariant facial characteristics. The extracted features are used to train and compare multiple classifiers, including logistic regression, support vector machines, and a multilayer perceptron (MLP) model tuned for expression analysis. Model performance is assessed using standard accuracy metrics and confusion matrices to evaluate consistency across all emotion classes. Results show that ArcFace embeddings significantly boost classification reliability compared to traditional descriptors alone. The project demonstrates an effective, practical pipeline for emotion recognition.

1 Introduction

This project aims to design facial expression classification models capable of categorizing facial emotions into seven distinct classes: angry, disgust, fear, happy, sad, surprise, and neutral. Such emotion inference can be extended to applications for improving safe driving (by detecting drowsiness) or in advertising targeting and content personalization on social media platforms, where emotional cues are leveraged to optimize message relevance and user engagement.

These models are trained on the FER-2013 dataset, a widely recognized benchmark for facial emotion recognition. The FER-2013 dataset is a standard benchmark for facial expression recognition, consisting of 35,887 pre-cropped grayscale facial images of size 48×48 pixels. The dataset is labeled into seven emotion classes - angry, disgust, fear, happy, sad, surprise, and neutral-and is partitioned into training and test sets, requiring no additional face cropping or color preprocessing.

2 Data

The FER-2013 dataset is partitioned into training and test sets, each reflecting the natural variability in emotional expression. In the original training set, the 'happy' emotion is the most prevalent with 7215 instances, followed by 'neutral' (4965) and 'sad' (4830). 'Fear' (4097) and 'angry' (3995) expressions also constitute substantial portions of the training data. 'Surprise' has 3171 samples, while 'disgust' is significantly underrepresented with only 436 samples, indicating a notable class imbalance. The test set mirrors these distribution trends. 'Happy' remains the most frequent emotion with 1774 instances, with 'sad' (1247), 'neutral' (1233), 'fear' (1024), and 'angry' (958) following. 'Surprise' accounts for 831 samples and 'disgust' is again the least represented class with a mere 111 samples. This consistent imbalance across both sets underscores the challenge of robustly classifying emotions, particularly 'disgust,' and highlights the need for strategies to address this disparity during model training and evaluation.

We partitioned the original training set into 70% training and 30% validation subsets using stratified sampling, preserving the original (imbalanced) class proportions. This ensured that each emotion category was represented in the same relative proportions across splits, allowing reliable performance monitoring and fair model selection without altering the underlying class distribution.

Given the varied distribution, particularly the low count for 'disgust,' our model will employ specific image features to accurately distinguish one class from the other. These image features, described below, include (HOG) for local shape and texture

Table 1: Variation in class distribution of the FER-2013 dataset across training, validation, and test sets.

Class	Train	Validation	Test
angry	2797 (13.9%)	1198 (13.9%)	958 (13.3%)
disgust	305 (1.5%)	131 (1.5%)	111 (1.5%)
fear	2868 (14.3%)	1229 (14.3%)	1024 (14.3%)
happy	5050 (25.1%)	2165 (25.1%)	1774 (24.7%)
sad	3381 (16.8%)	1449 (16.8%)	1247 (17.4%)
surprise	2220 (11.0%)	951 (11.0%)	831 (11.6%)
neutral	3475 (17.3%)	1490 (17.3%)	1233 (17.2%)

information, Dlib facial landmarks for structural facial cues, and InsightFace embeddings for robust deep facial representations. Through this comprehensive approach of combining multiple feature extraction techniques and reducing dimensionality using Principal Component Analysis (PCA) and combining features from pre-trained models, we aim to build an effective classifier despite the inherent challenges of the dataset.

3 Feature Extractions

Combining geometric features (landmarks), texture-based descriptors (HOG), and deep semantic embeddings (ArcFace) provides complementary information. This multi-feature strategy strengthens the model's ability to distinguish subtle facial expressions. In the composite overlay of emotions below, we can see some of the common features from the overlay composite of each emotion.

3.1 Image Processing

The images provided by FER-2013 are cropped to and grey scaled to only the person face would be in the image. Furthermore, looking through many images the person's face is centered even if they are not completely facing the camera. Processing for us was creating a pipeline for each feature. The landmark features were normalized due to the models not learning any face size, geometry, or camera distance. Furthermore the images were resized to 96x96 for the landmark to ensure better extraction of the features but not for the ArcFace or HOG as it would make the features inconsistent. The ArcFace classifier used was *buffalo_l* which normalizes the vectors within the function per image. These three vectors were stored in npy files for faster distribution to the team.

The total image count for the training dataset is 28,709 with the distribution of the for each emotion having a minimum of 3,000 dataset, with the exception of disgust having only 436 in the training dataset. While in the test dataset we can see a similar distribution of each class having about 1,000 but disgust having 111. To maintain the same distribution of examples across our training, validation, and test data set we split using sklearn's train_test_split function with the stratify parameter to ensure there was similar distribution between the training and validation.

Conversion from RGB to Greyscale was performed using the OpenCV method *cv2.COLOR_BGR2GRAY*. This ensured images went from 3 color space to a single channel grayscale image. Many feature extraction could be biased by various lighting or clothing color as some photos came from film. Although many of the modern day packages considered did not require grey scale images it was best not to allow any unnecessary noise.

The features we anticipated gathering have to do what humans normally use to express emotion. These contain but are not limited to mouth openness, wrinkles, eye openness, and jawlines. Having looked through the data and noticing various overlap of features that could lead to misclassification. For instance, in the example photos of each emotion above, they are angry/disgusted and neutral/sad. These two pairs have the possibility of being labeled inverted. This could affect how the images could be interpreted differently by other people.

3.2 Simple Feature: Histogram of Oriented Gradients (HOG)

HOG is a powerful feature that can be used for many tasks, including object recognition. HOG feature vectors show the magnitude and orientation of gradients at different locations in the image. We chose this feature for our image set be-

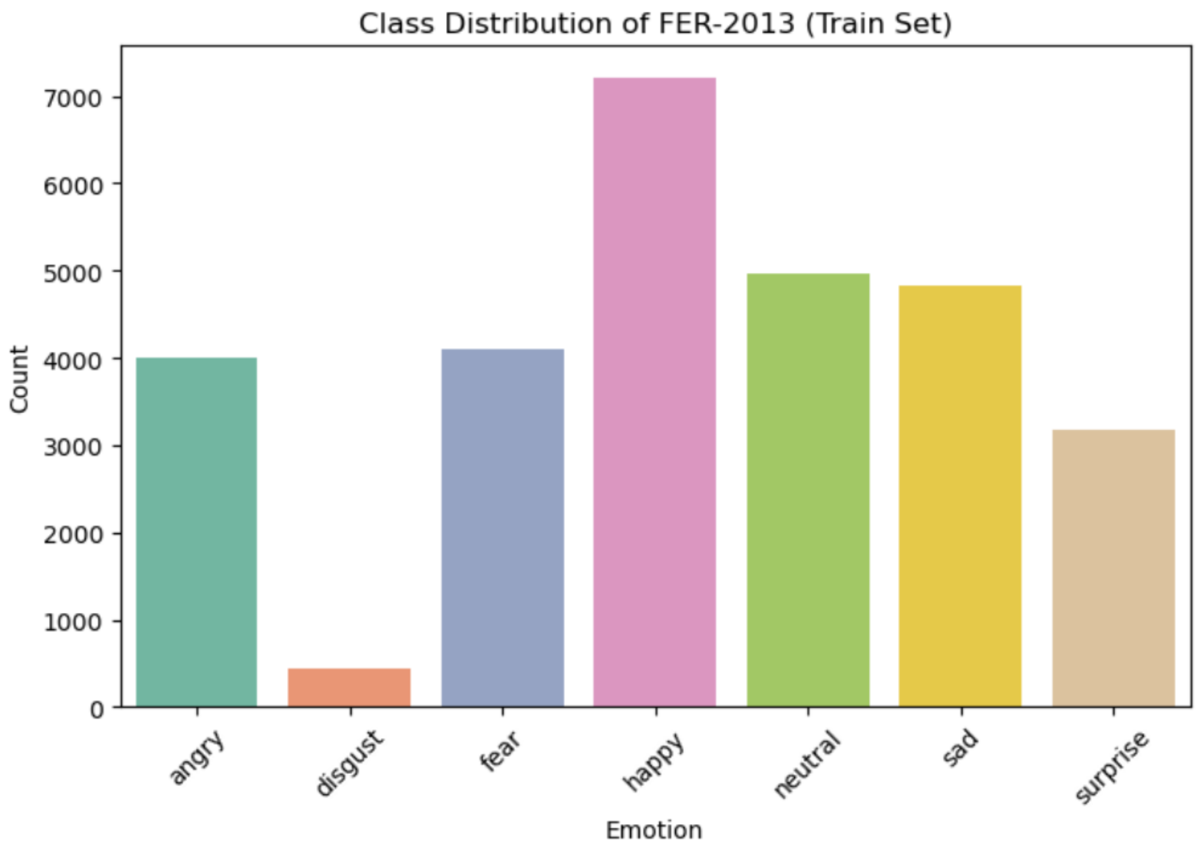

Figure 1: Variation in class distribution of the FER-2013 training set.

cause our image classes appear to have distinct gradients. For example, facial regions such as the eyebrows, eyes, and mouth produce strong horizontal and vertical gradients, while expressive muscle movements—like smiles, frowns, or raised eyebrows—introduce diagonal and curved gradient patterns that HOG effectively captures.

This project performed HOG feature extraction followed by Principal Component Analysis (PCA) for the analysis and visualization of the FER-2013 facial emotion dataset. The methodology encompasses data loading, HOG feature engineering, and subsequent dimensionality reduction using PCA for both quantitative assessment of variance retention and qualitative visualization of emotion clusters. Key findings include the number of principal components required to retain significant variance and visual representations of the reduced-dimensional data.

For each preprocessed image, HOG features were extracted. The HOG algorithm captures local object appearance and shape by counting occurrences of gradient orientation in localized portions of an image. The HOG features were extracted using the following parameters: *orientations* = 9, *pixels_per_cell* = (8, 8), *cells_per_block* = (2, 2), and *block_norm* = 'L2-Hys'.

This process resulted in a feature matrix *X_hog* of shape (28709, 900), indicating 28,709 samples each represented by a 900-dimensional HOG feature vector.

3.3 Simple Feature: Facial Landmark

Facial landmarks allowed us to get explicit information such as eye location or mouth width. This gave us a geometry of the face structure and facial muscles that make the emotions. The feature is also robust to various lighting skin tones, and image background noise. The strength of the feature comes from allowing the model to evaluate the differences in key face points from different emotions. These key points include eye openness, mouth width, jaw line and eyebrow height. These key points are the focal point used to express emotions by people. While the landmarks have a weakness of not capturing some of the subtle text detail some people have while expressing an emotion, it makes up for in geometry of the face.

The pipeline used to generate the land marks, *dlib*, open source package and a C++ toolkit for ma-

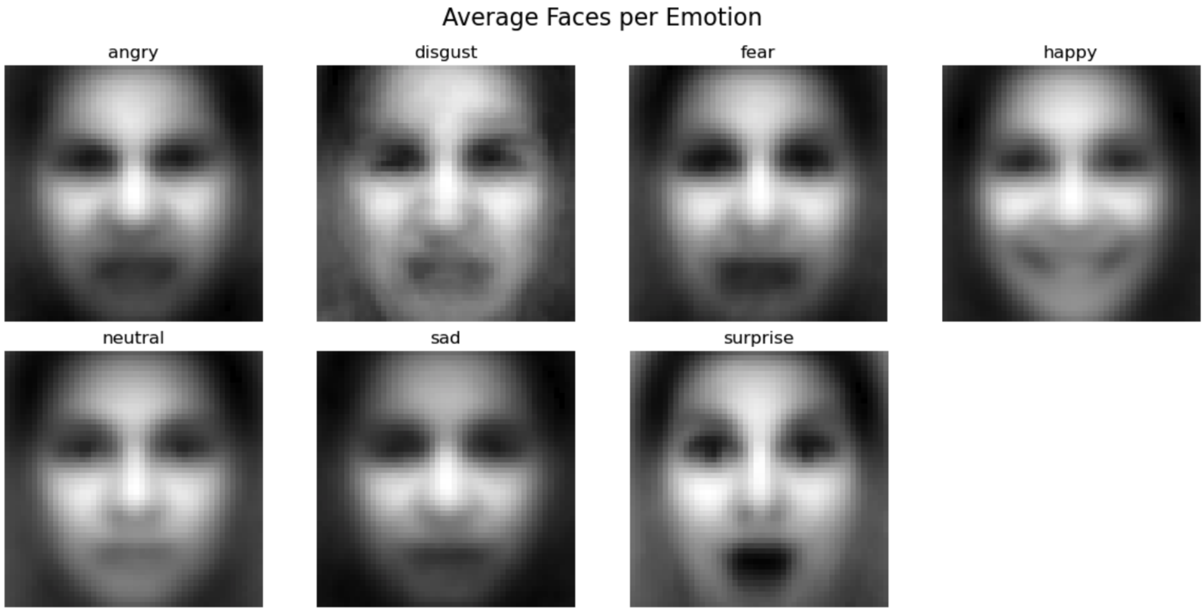

Figure 2: Composite overlay of facial expressions representing different emotions.

Figure 3: Example images representing each emotion.

chine learning. Many of the dependencies required caused lots of issues for importing the package into a local computer and for google colab due to the lack of easy to understand documentation. Once installed creating a pipeline can be followed through their documentation. The image below displays all the key points of a person's face. As we can see the point includes only the face while excluding the person's hair and other noise in the background.

3.4 Complex Feature: ArcFace Embeddings

ArcFace is a deep-learning–based embedding method developed by researchers at InsightFace, a leading computer vision group in China, with the main maintainer affiliated with Imperial College London (`https://jiankangdeng.github.io`). It generates embeddings that capture a rich combination of facial texture, geometry, and semantic attributes, producing highly discriminative representations of faces.

In this work, we use the InsightFace *buffalo_l* model, a large, high-capacity pretrained backbone trained with the ArcFace loss on large-scale face datasets. The model generates highly discriminative embeddings that capture a rich combination of facial texture, geometry, and high-level semantic attributes, and it achieves state-of-the-art performance on face verification and identification benchmarks.

Compared to traditional features such as HOG, which capture local texture and edge patterns, and facial landmarks, which encode the geometric structure of key facial points, ArcFace provides more abstract, high-level representations. When combined, HOG, landmarks, and ArcFace embeddings allow downstream classifiers such as MLPs or SVMs to leverage complementary low-level, mid-level, and high-level information, improving robustness and overall accuracy for facial expression recognition on datasets like FER-2013.

3.5 PCA Decomposition of Combined Features

Principal Component Analysis (PCA) was employed to reduce the dimensionality of the high-dimensional feature space, which combines HOG,

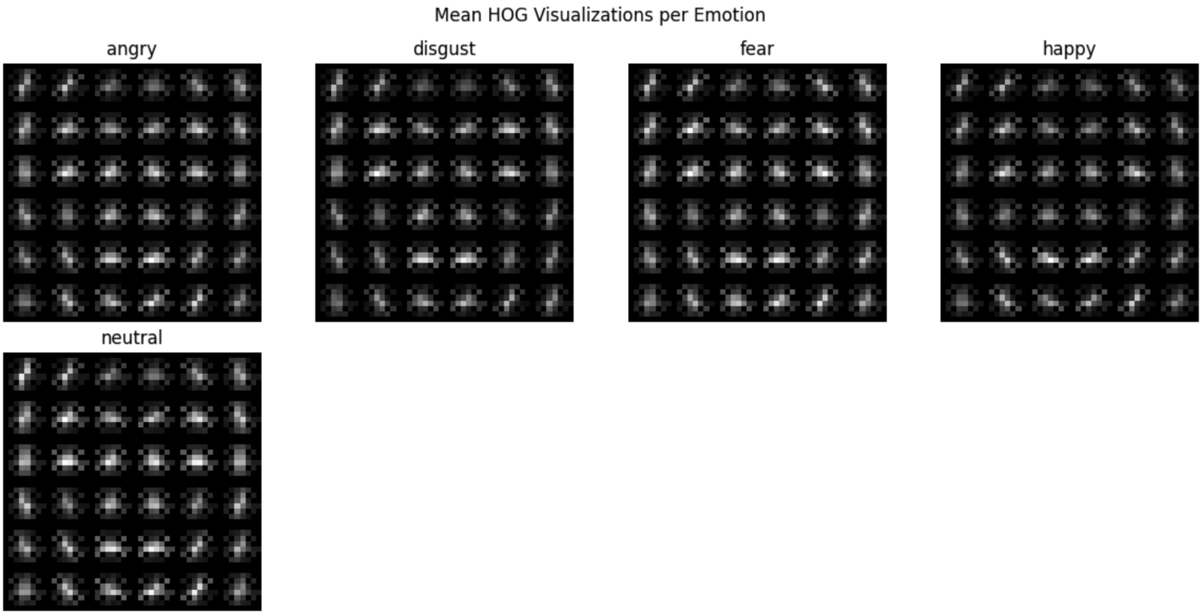

Figure 4: Mean HOG visualizations for several emotions.

facial landmarks, and InsightFace embeddings. The primary goal of PCA is to retain as much of the original dataset's variance as possible in a lower-dimensional representation. Prior to applying PCA, the combined feature set was standardized using StandardScaler to ensure that all features contribute equally to the variance calculation. The cumulative explained variance plot (generated in the notebook) for the combined HOG, Landmarks, and InsightFace features shows a steep initial rise, indicating that the first few principal components capture a substantial amount of the data's variability. The curve then gradually flattens, signifying diminishing returns from adding further components.

For the subsequent modeling phases, the data was reduced to 596 components, thereby capturing 95% of the total variance, striking a balance between dimensionality reduction and information preservation. This process is crucial for mitigating the curse of dimensionality, improving computational efficiency, and potentially enhancing model generalization by reducing noise in the feature space.

To visually demonstrate the impact of dimensionality reduction, pixel-level PCA was performed on a subsample of images. Original images were compared against their reconstructions using 200 principal components in pixel space. This qualitative assessment shows that even with significant dimensionality reduction at the pixel level, the reconstructed faces retain recognizable facial features, underscoring PCA's ability to capture dominant variations.

3.6 t-SNE Visualization of Combined Features

The t-SNE plot of feature vectors allows us to see how the feature vectors are displayed in a 2D plot. Each class is displayed in a 2D embedding resulting in classes clustering revealing easily separated classification. We can see in the t-SNE plot we generated there is a lack of clustering from any class. This revealed how these classes are not easily linearly separated and there might be a higher dimensional separation. If we step back and think about how this visual represents the differentiation between features means there are a lot of overlapping features among the classes. This could also mean the t-SNE plot might not be the best way to represent the class separation for facial emotions.

4 Classification

In this segment, we detail the construction and evaluation of three distinct classifiers. The first approach employed logistic regression, which is well suited for categorization into one of six emotion classes. As an alternative, we implemented a support vector machine (SVM), chosen for its strong baseline performance in high-dimensional feature spaces. Finally, we incorporated a multilayer perceptron (MLP) to capture nonlinear relationships within the data.

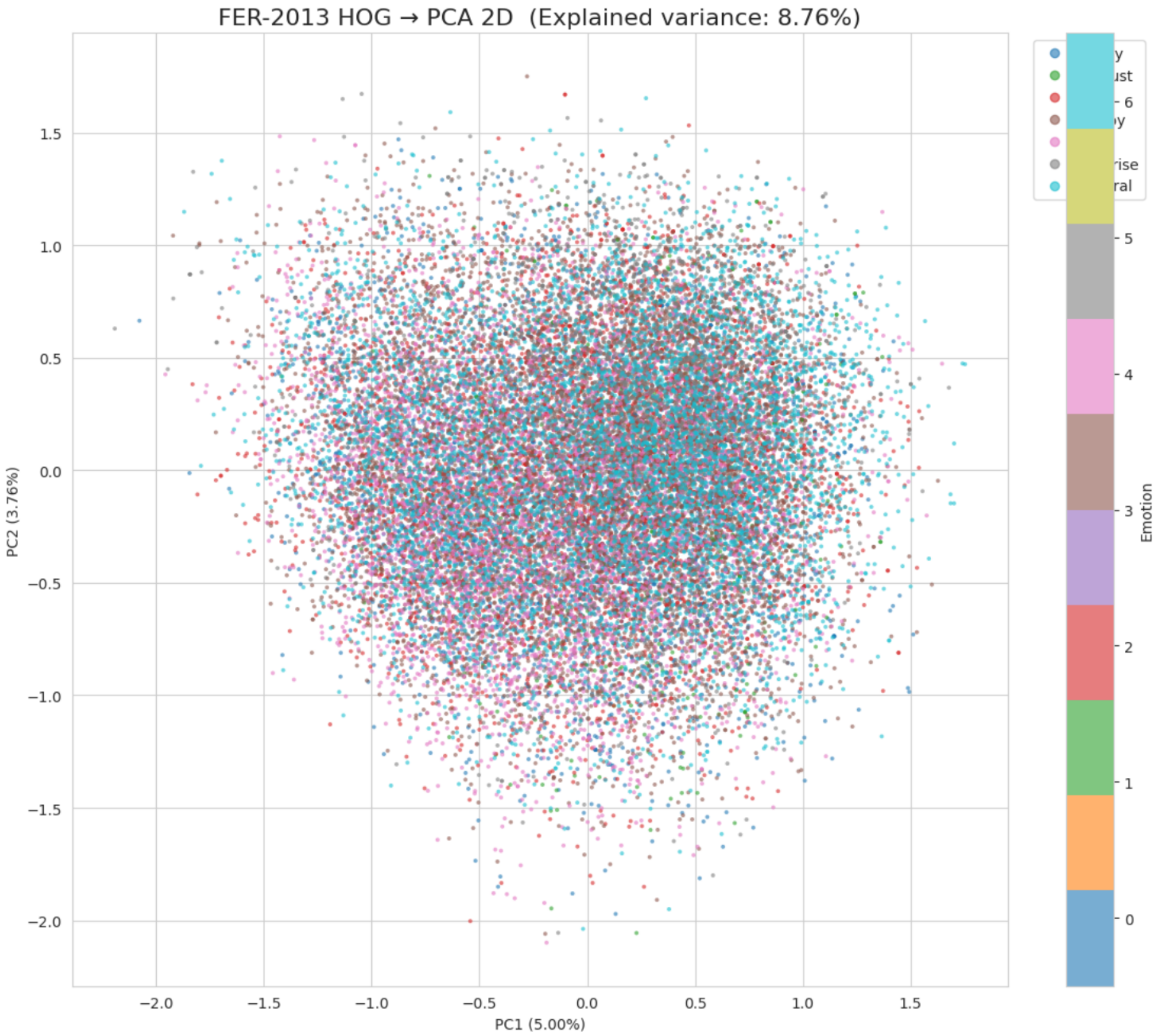

Figure 5: 2D PCA of FER-2013 HOG features, showing 8.76% of explained variance.

4.1 Logistic Regression

Using a logistic regression is most commonly used for when a feature space can be separated by its class. With this in mind we decided to test using a logistic regression model provided by sklearn. Once performing a simple logistic regression for the base model getting around 57% F1 accuracy gave us a good starting point.

Having a multi-class classification problem of identifying which class we decided to create a hyperparameter search with the intention of finding the correct regularization and solver combination, with the max iterations set to 2000. Giving a max iterations allowed the classifier to have a high ceiling for the about 600 features after conducting a PCA. We also ensured to use multimodal and class balanced parameters to ensure none of the classes were weighted more for having more examples.

4.1.1 Logistic Regression with Combined Features

As seen in the confusion matrix the label best classified was happiness. While the class with the worst was disgusting. There was also a confusion between sad and neutral. This ended with an overall F1 score of 49%.

4.1.2 Logistic Regression with ArcFace

The confusion matrix and ROC curve results indicate that combining multiple feature representations produced performance nearly identical to using ArcFace embeddings alone. Class-wise prediction patterns and overall discriminative ability remained consistent across both approaches, suggesting that ArcFace embeddings already capture the most relevant facial information for expression classification. Additional features such as HOG descriptors and facial landmarks did not provide

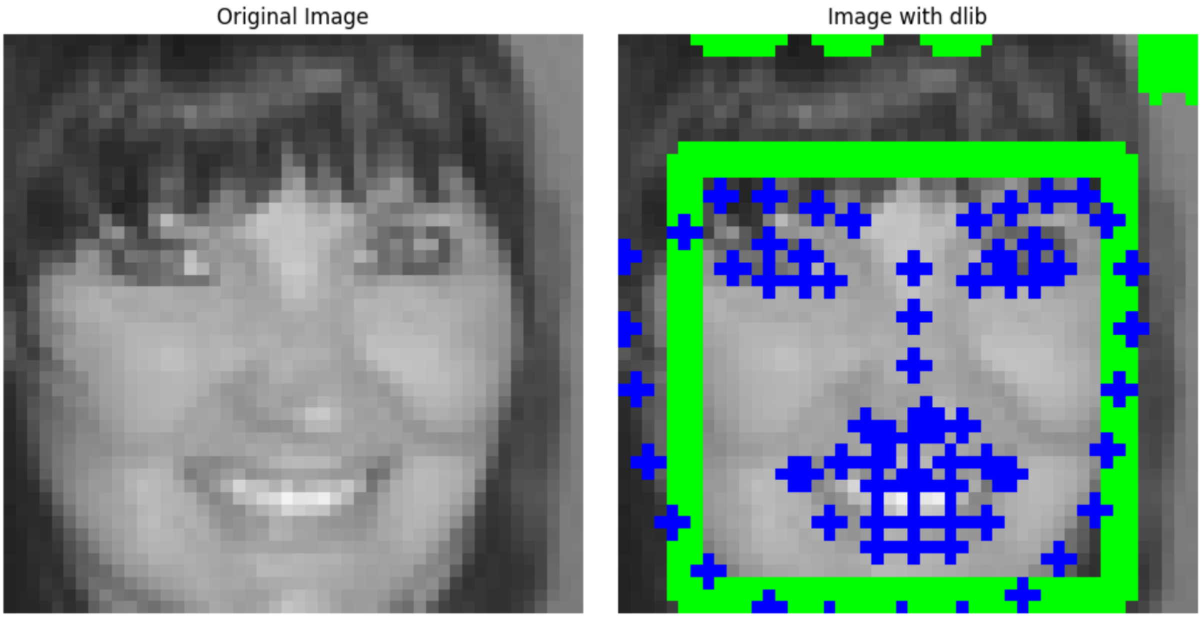

Figure 6: Example of facial landmarks overlaid on the original image.

Table 2: PCA component analysis of the combined features.

Variance Retained	Number of PCA Components
90%	486
95%	596
99%	785

meaningful performance gains, likely due to redundancy with the embedding representation or added noise. These results demonstrate there is feature overlap among the features. This ended with an overall F1 score of 50%.

4.2 SVM

SVM classifier are well suited for high dimensional data i.e. HOG, facial landmarks, and deep embeddings (InsightFace) and often outperform other algorithms by finding clear decision boundaries. Additionally, SVMs can incorporate class_weight parameters to give more importance to minority classes as in the FER-2013 dataset is imbalanced (e.g., 'disgust' has far fewer samples than 'happy'). This section summarizes the performance of Support Vector Machine (SVM) models i.e. Nonlinear PCA-SVM, applied to the FER-2013 dataset using HOG and Landmark features.

4.2.1 Nonlinear SVM on Combined Features

This model employed a fused feature representation combining HOG descriptors, normalized facial landmarks, and ArcFace embeddings, with dimensionality reduced to 276 principal components using PCA based on optimal tuning. A nonlinear SVM with an RBF kernel was used as the classifier, with hyperparameters optimized via *RandomizedSearchCV* over 15 iterations and 3-fold cross-validation. The best configuration (C $\approx$ 2.80, $\gamma \approx$ 0.00117, RBF kernel) reflected the most effective balance for this feature set. Owing to the nonlinear kernel and expanded search space, training was computationally intensive (~1,576 seconds), with prediction times of ~43 seconds on validation data and ~55 seconds on the test set. The model achieved a validation accuracy of 68.6% and weighted precision of 68.8%, while test performance dropped to 55.3% accuracy and 56.6% weighted precision, indicating some loss in generalization.

Class-wise analysis from the classification report and confusion matrix shows strong recognition of more distinctive emotions such as Surprise, Angry, and Happy. Disgust exhibited high separability (AUC) but low recall, reflecting difficulty in detecting its relatively rare instances. Fear was consistently misclassified, suggesting insufficient dis-

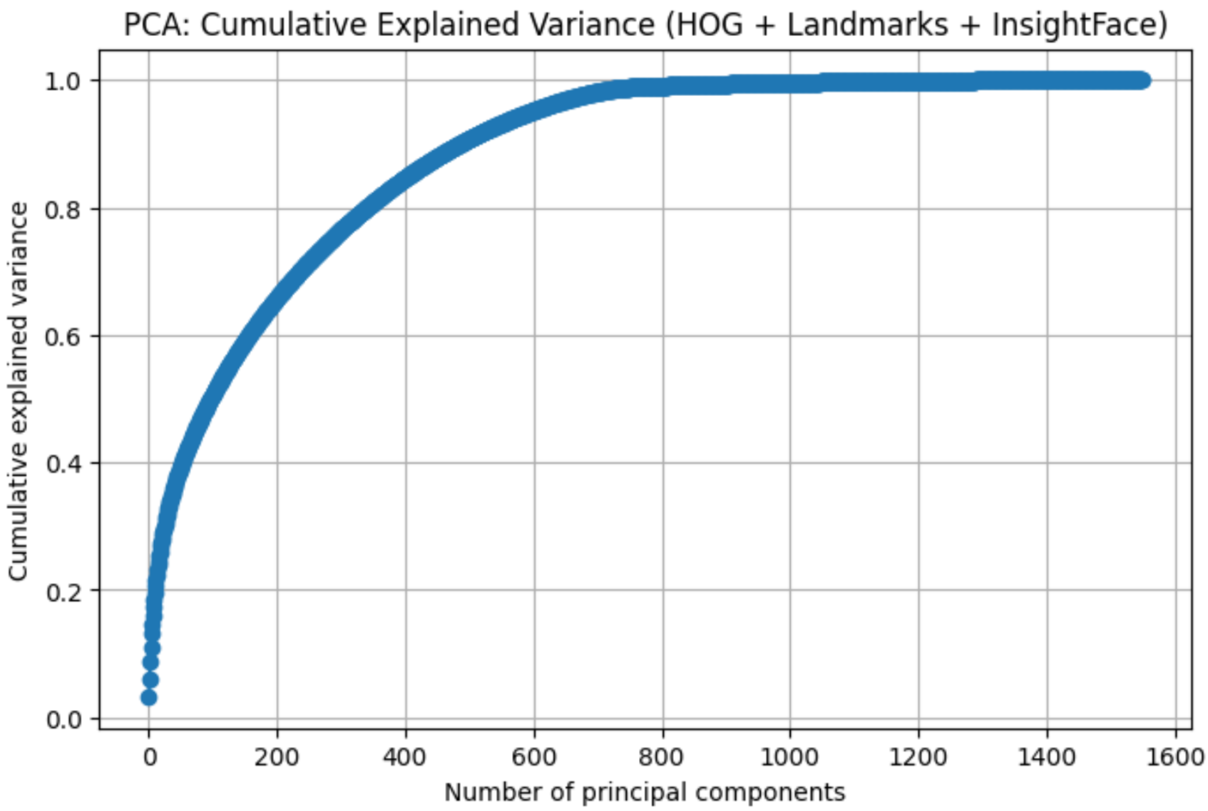

Figure 7: PCA of the combined training data.

Table 3: Hyperparameter search values for logistic regression.

Parameter	Search Values
Regularization	0.001, 0.01, 0.1, 1, 10, 100
Solver	Lbfgs, saga

criminative power for capturing its subtle characteristics. Additional confusion was observed among visually similar expressions such as Happy, Sad, and Neutral, pointing to overlapping feature representations.

4.2.2 Nonlinear SVM with ArcFace

The nonlinear SVM model using ArcFace embeddings achieved a test accuracy of approximately 63% with a weighted precision of about 63%, reflecting balanced overall performance across emotion classes. The model demonstrated strong discriminative capability for certain emotions, showing very high precision for specific classes and consistently strong precision, recall, and F1-scores for others, indicating reliable recognition where facial cues are more distinctive. However, several emotions exhibited moderate F1-scores and higher confusion, particularly for classes with subtler or overlapping expressions, suggesting room for improvement in class separability. Analysis of the confusion matrix and ROC characteristics further confirms that while emotions such as Disgust, Happy, and Surprise are well separated, the model struggles more with Angry, Fear, Sad, and Neutral. Overall, the results indicate reasonable generalization performance, with test-set inference completing in approximately 86 seconds.

4.3 MLP

In deep learning, a multilayer perceptron (MLP) is a kind of modern feedforward neural network consisting of fully connected neurons with nonlinear activation functions, organized in layers, notable for being able to distinguish data that is not linearly separable. We applied the same hyperparameter search strategy across all MLP models to ensure a fair and consistent comparison. The specific hyperparameter ranges and selected values for each model are summarized in the corresponding table

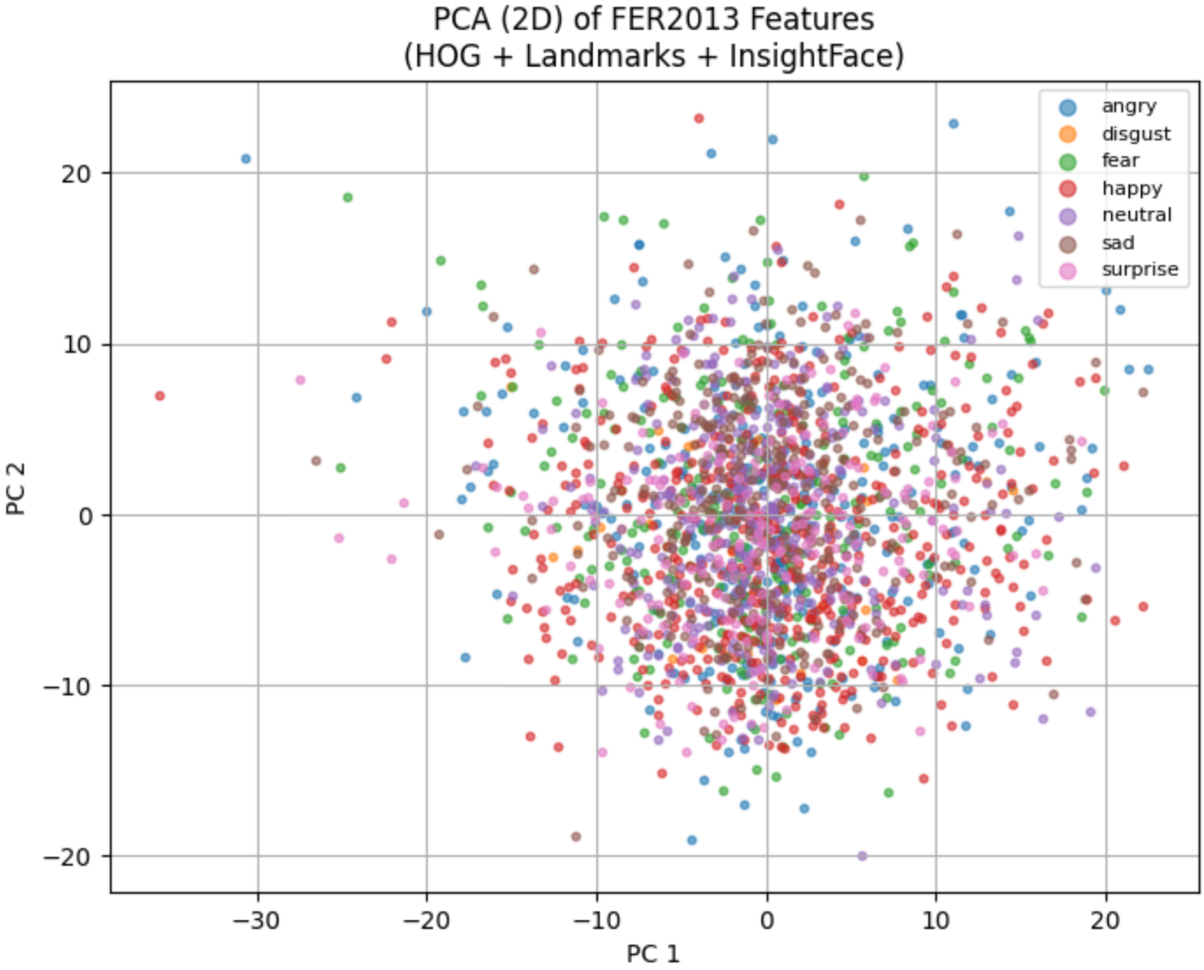

Figure 8: 2D PCA of the combined features.

below.

4.3.1 MLP on Combined Features with PCA

The multiclass ROC curve for the MLP illustrates how effectively the model distinguishes among emotion classes, with higher AUC values indicating stronger separability. Emotions such as surprise and anger show strong performance with AUC scores of approximately 0.93, while fear emerges as the most difficult class with an AUC of about 0.64. The confusion matrix further reinforces these patterns by showing how frequently each class is correctly predicted versus misclassified. The model performs best on happy, neutral, and surprise, which is reflected in the strong diagonal values. In contrast, fear and sadness display greater confusion with similar neighboring emotions, consistent with their lower AUC scores. Overall, the MLP exhibits solid classification capability, though certain emotion categories remain challenging to differentiate.

4.3.2 MLP on ArcFace Embeddings

The MLP model trained on raw ArcFace features demonstrates strong overall multiclass emotion recognition performance. The ROC analysis shows consistently high discriminative ability across classes, with AUC scores exceeding 0.90 for surprise (0.926), disgust (0.920), and happy (0.910), indicating excellent separability, while angry (0.834), sad (0.816), and neutral (0.819) achieve solid performance. Fear (0.766) remains the most challenging class. The confusion matrix further reveals that happy, neutral, sad, and surprise are classified most accurately, with high true positive counts, whereas notable confusion occurs between semantically similar emotions, particularly fear–sad, angry–sad, and neutral–happy. Overall, the results suggest that the MLP effectively leverages ArcFace embeddings for emotion classification, performing especially well on expressive emotions, while more subtle or overlapping emotional states remain harder to distinguish.

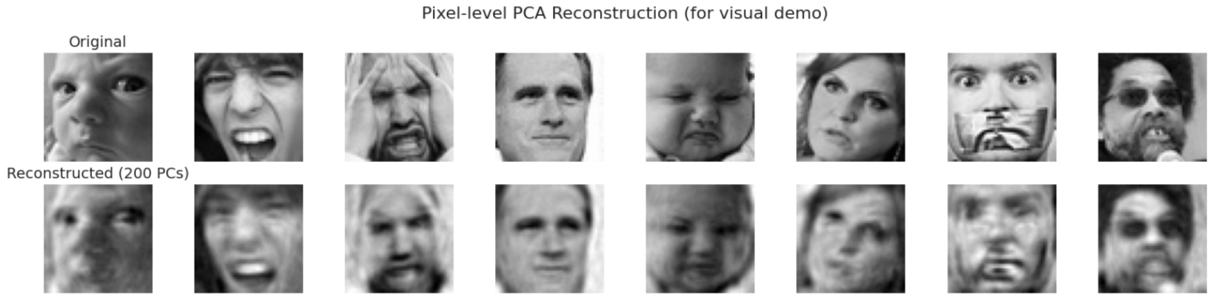

Figure 9: Pixel-level PCA demonstrating visual reconstruction of emotions.

Table 4: MLP hyperparameter search values.

Parameter	Search Values
hidden_layer_sizes	(32,), (64,), (32,16), (64,32)
activation	"relu", "tanh"
alpha	0.01, 0.05, 0.1, 0.2
learning_rate	"adaptive"
beta_1	0.7, 0.8, 0.9
beta_2	0.99, 0.999

4.4 Error Discussion Per Model

4.4.1 SVM Limitations

Support Vector Machines (SVMs) are well suited for problems with high-dimensional feature spaces and clear class boundaries, as they maximize the margin between classes and can model complex nonlinear decision surfaces through kernel functions. This makes them effective at separating visually distinctive emotions such as Disgust, Happy, and Surprise. However, SVMs are less effective when class distributions overlap significantly or when intra-class variability is high, which explains the reduced performance observed for Angry, Fear, Sad, and Neutral, where facial expressions are more subtle and less distinctly separable in the feature space.

4.4.2 Logistic Regression Limitations

The Logistic Classifier model performed adequately when using ArcFace embeddings, demonstrating clear diagonal dominance in the confusion matrix and consistent ROC performance across classes. While the model did not achieve good classification, it provided a reasonable performance for facial expression recognition with a 50% accuracy with the ArcFace embedding and 49% with the combination. This could be as the many of the features that make up emotions are not separable linearly.

4.4.3 MLP Limitations

The MLP is effective at learning nonlinear decision boundaries from feature representations; however, the chosen parameter settings prioritize stability and computational efficiency. In particular, the limited model capacity resulting from small hidden layers restricts its ability to learn complex, high-level facial patterns required for robust emotion recognition.

4.5 Error Discussion Across Models

Across all evaluated models, we observe similar patterns of misclassification, indicating shared underlying challenges rather than model-specific weaknesses. Consequently, the following discussion provides a high-level analysis of error trends that are consistent across classifiers, focusing on common sources of confusion and structural limitations that affect performance regardless of the chosen model.

4.5.1 Class-Dependent Performance Differences

Performance differences across emotion classes likely stem from the distinctiveness of facial cues and feature representation. Emotions such as happy, disgust and surprise exhibit more pronounced facial patterns, making them easier for the classifier to separate, while fear, sad, and neutral share subtler and overlapping expressions that lead to higher

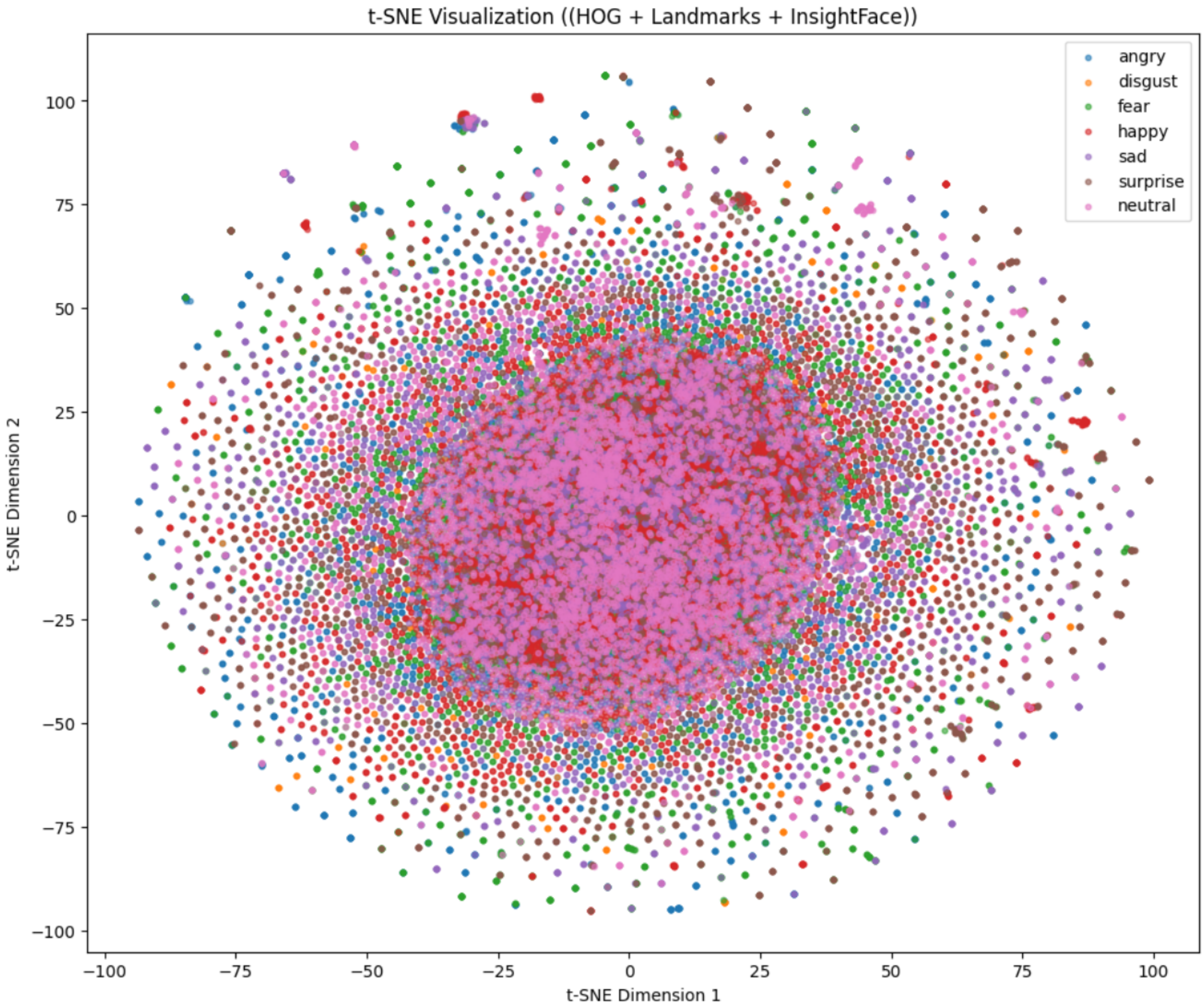

Figure 10: t-SNE visualization of training data using combined features.

confusion.

4.5.2 PCA Limitations

Applying PCA to combined feature embeddings can discard useful information because PCA retains components with the highest variance rather than those most relevant to emotion discrimination. PCA may shrink or remove small but important facial changes that signal different emotions. This makes different emotions harder to tell apart and reduces classification performance.

4.5.3 ArcFace Limitations

ArcFace is fundamentally optimized for identity recognition, as its training objective explicitly encourages features from the same person to cluster tightly while pushing features from different people far apart in embedding space. This design promotes invariance to facial expressions, treating emotion-related changes as secondary variation rather than primary signals. We observe this limitation consistently across different classifiers, suggesting that the main performance constraint arises from the identity-focused feature representation rather than the downstream model choice. Consequently, while highly expressive emotions such as happy or surprise are still captured reasonably well, more subtle and overlapping emotions remain challenging to distinguish—a difficulty that also persists in human inspection, where such emotions are often ambiguous even to trained observers.

5 Generalizability

The data were split into training, validation, and test sets before training the classifier, with hyperparameter tuning performed only on the validation set. This procedure avoids overfitting and ensures that test performance reflects true generalization to unseen data.

The models generally struggle to generalize when recognizing more subtle emotions such as

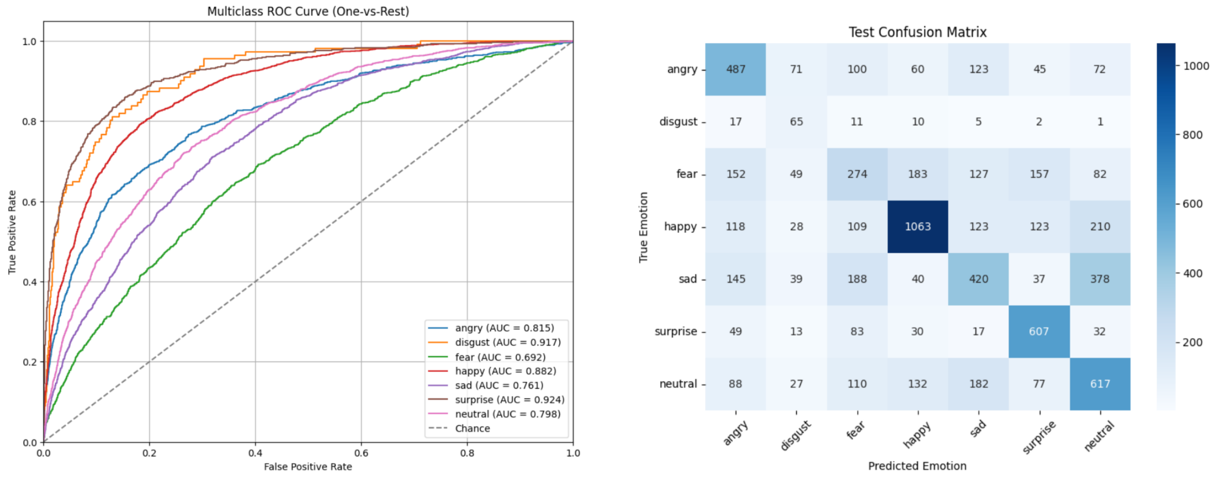

Figure 11: Logistic regression on combined features with PCA (test set): multiclass ROC curve (left) and comparative confusion matrices (right).

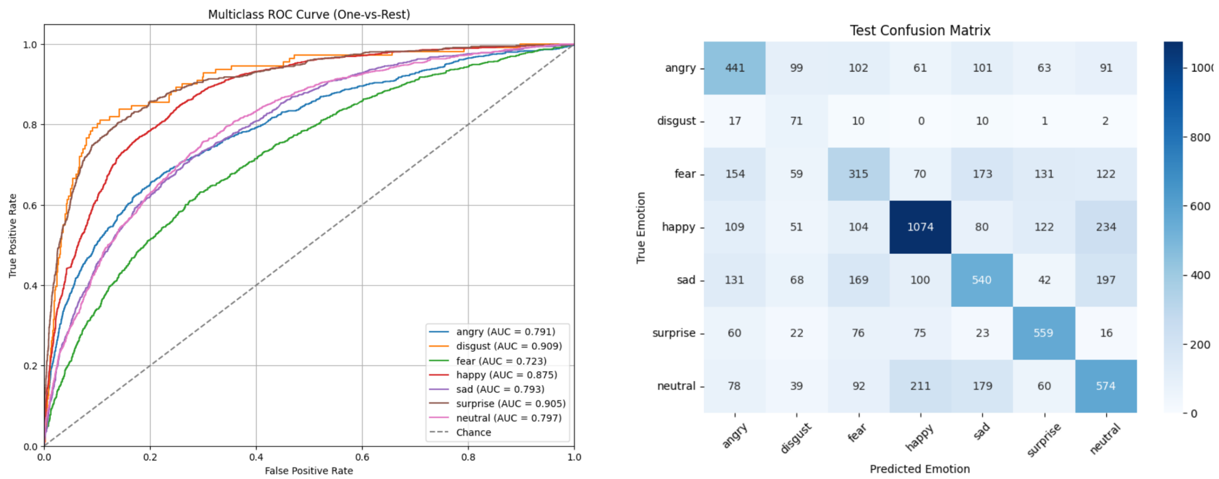

Figure 12: Logistic regression on ArcFace feature (test set): multiclass ROC curve (left) and comparative confusion matrices (right).

sadness, neutral, and fear, while performing better on more expressive emotions like happiness, disgust and surprise. This suggests that certain emotions are inherently more ambiguous, making them more difficult to distinguish. Visual inspection of the images helps explain this behavior, as sad and neutral expressions often share very similar facial features with minimal visual differences.

When using combined features (HOG + landmarks + ArcFace) with PCA, the models exhibit mixed generalization performance. The MLP maintains balanced training and validation accuracy (0.85) but drops to 0.48 on the test set, indicating difficulty generalizing to unseen data, particularly for subtle emotions. The SVM with a nonlinear kernel, while achieving extremely high training accuracy (0.99), attains the highest test accuracy (0.55) among the combined-feature models, though it still exhibits limited generalization, indicating residual overfitting. Logistic Regression generalizes most consistently across training, validation, and test splits, indicating low overfitting, though its overall accuracy is modest (test accuracy of 0.49).

Using ArcFace embeddings alone leads to improved generalization and higher test accuracy across all models. Both MLP and SVM benefit significantly from the richer, high-level representations, with the nonlinear SVM achieving the highest test accuracy of 0.63. Logistic Regression continues to demonstrate stable performance with minimal overfitting, although its accuracy remains moderate (0.50). These results indicate that ArcFace embeddings better capture facial texture, geometry, and semantic information, making the models more robust to subtle differences between emotions.

Despite this, there is room for improvement

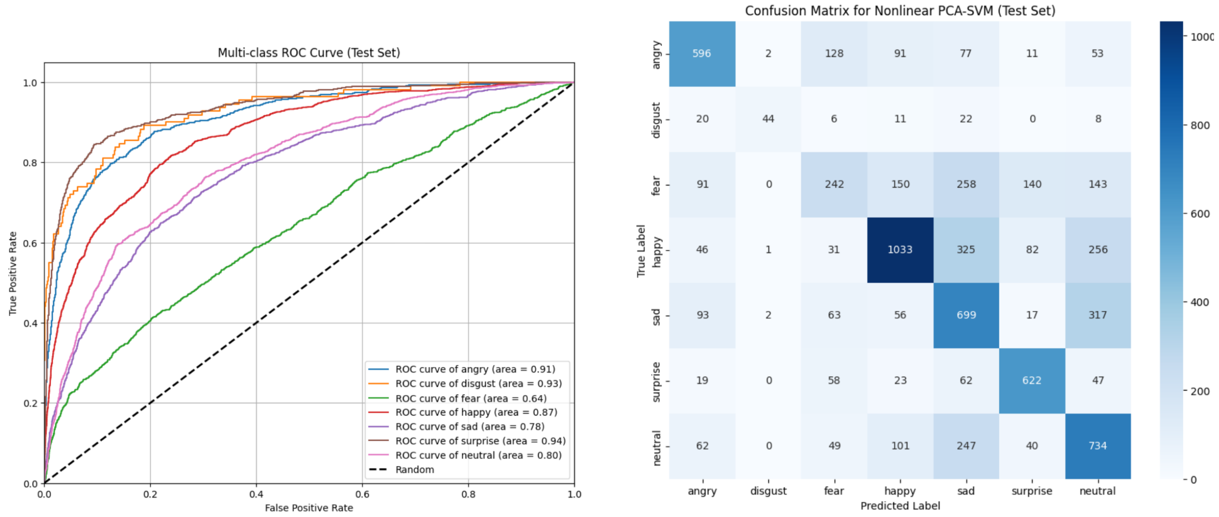

Figure 13: Nonlinear SVM on combined features with PCA (test set): multiclass ROC curve (left) and comparative confusion matrices (right).

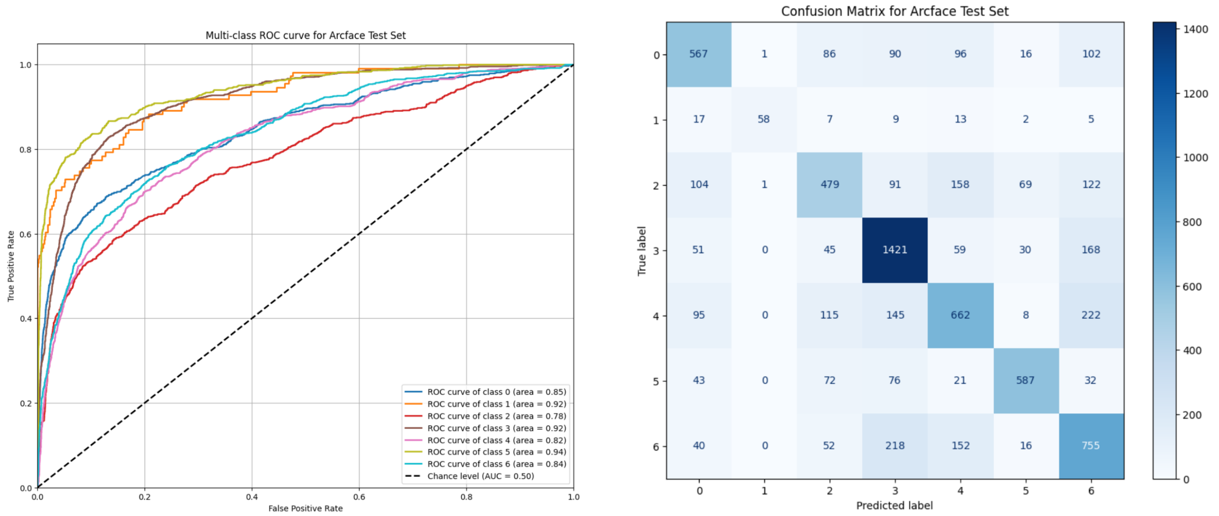

Figure 14: Nonlinear SVM on ArcFace feature (test set): multiclass ROC curve (left) and comparative confusion matrices (right).

in the training process. Generalization could be enhanced by addressing class imbalance through reweighting or data augmentation, particularly for underrepresented emotions. Additionally, trying emotion-focused embeddings for our classifier where emotion-discriminative features are emphasized, may improve sensitivity to emotion-specific cues and reduce systematic confusion on the test set.

6 Efficiency vs Accuracy

In terms of efficiency, the MLP with ArcFace stands out as a strong overall choice, offering moderate training time (78.01 s) and the fastest hyperparameter search (1 min) among the higher-performing models, while still achieving competitive test accuracy (0.56). In contrast, the SVM with ArcFace achieves the highest test accuracy (0.63), making it the preferred option when predictive performance is the primary objective. However, this improvement comes with increased computational cost, including longer training time (5.03 minutes) and a more expensive hyperparameter search (21.9 minutes). Overall, the results highlight a clear trade-off: the SVM with ArcFace provides best accuracy at a reasonable computational cost, while the MLP with ArcFace offers best efficiency with moderately lower accuracy.

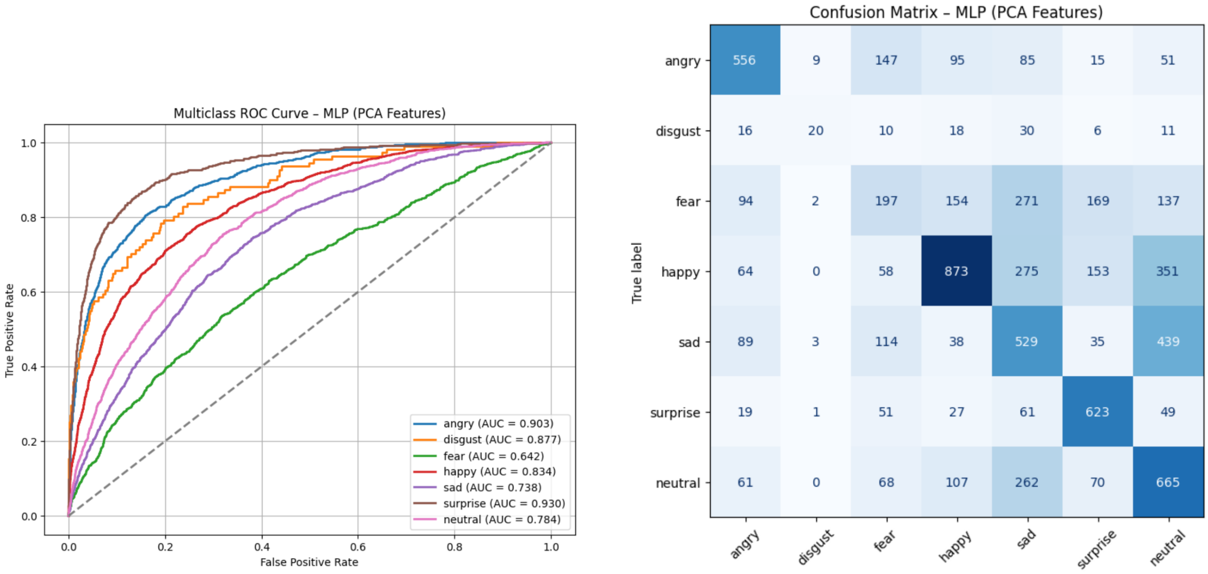

Figure 15: MLP on combined features with PCA (test set): multiclass ROC curve (left) and comparative confusion matrices (right).

Table 5: Model performance summary metrics.

Model (Features)	**Training Accuracy**	**Validation Accuracy**	**Test Accuracy**	**Training Time**	**Hyperparameter Search Time**
MLP (Combined features with PCA)	0.85	0.85	0.48	77.58 s	3.7 min
Non-linear SVM (Combined features with PCA)	0.99	0.68	0.55	144.49 s	17.6 min
Logistic Regression (Combined features with PCA)	0.58	0.58	0.49	21 s	6.7 h
MLP (ArcFace)	0.74	0.74	0.56	78.01 s	1 min
Non-linear SVM (ArcFace)	0.97	0.63	0.63	5.03 min	21.9 min
Logistic Regression (ArcFace)	0.55	0.50	0.50	0.01 s	51.6 min

7 Conclusion

Initially, we assumed that recognizing emotions from images would be straightforward, with each emotion being clearly distinguishable. However, our results showed that many emotions, such as disgust vs. anger or neutral vs. sad, appear very similar and are difficult to differentiate. All three models struggled with these subtle distinctions, highlighting the inherent challenge of facial expression recognition.

Contrary to our expectations, combining all feature sets did not produce the best results, and even with PCA reduction, the nonlinear SVM model achieved up to 0.55 test accuracy. In contrast, using only complex features such as ArcFace embeddings with SVM and MLP models led to substantially better performance, with test accuracy reaching up to 0.63 for the nonlinear SVM. This improvement is likely due to ArcFace embeddings already capturing a rich combination of texture, geometry, and semantic facial attributes, meaning that adding simpler feature sets introduced noise and reduced accuracy and precision. Additionally, concatenating simple features with high-dimensional embeddings may have contributed to the curse of dimensionality, making it harder for models to learn meaningful decision boundaries, further hindering model learning.

From the model comparisons, the MLP with ArcFace provides the best tradeoff between efficiency and performance, benefiting from moderate training time and fast hyperparameter tuning, while the SVM with ArcFace achieves the highest test accuracy, making it the preferred choice when predictive performance is the priority. Logistic Regression remains the most stable in terms of generalization but achieves modest accuracy. For future work, exploring more complex emotion-focused embeddings and experimenting with more diverse model selection may further enhance performance and help models better capture nuanced expressions.

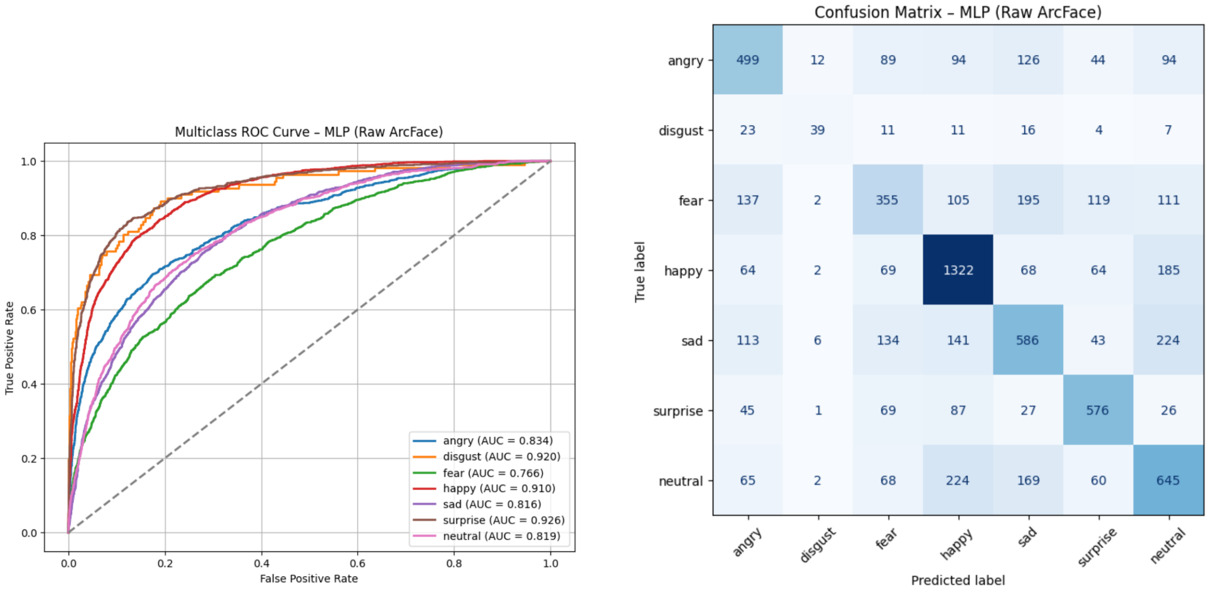

Figure 16: MLP on ArcFace feature (test set): multiclass ROC curve (left) and comparative confusion matrices (right).

8 Contributions

In this project, our team divided the workload. Deepak focused on exploratory data analysis (EDA), providing key insights into the dataset's structure and patterns, and also developed HOG features and trained the SVM classifier. Marcelino handled dimensionality reduction, applying PCA and t-SNE to visualize and preprocess the data, extracted facial landmarks, and trained the Logistic Regression model. Xinyi worked on ArcFace embeddings, capturing high-level semantic and geometric features of faces, and implemented the MLP classifier. Throughout the project, we had a great time collaborating, working together on slides and the final report.

Table 6: Model performance summary notes.

Model (Features)	**Notes**
MLP (Combined features with PCA)	Balanced training and validation performance; high efficiency; moderate generalization with noticeable test-set drop.
Non-linear SVM (Combined features with PCA)	High training accuracy; overfitting on training data; poor generalization; moderate efficiency due to long training and tuning time.
Logistic Regression (Combined features with PCA)	Stable performance across splits; low overall accuracy; fast training but extremely long hyperparameter tuning time.
MLP (ArcFace)	**Best efficiency classifier** due to fastest hyperparameter search; good tradeoff between accuracy, efficiency, and generalization.
Non-linear SVM (ArcFace)	**Best accuracy classifier** with highest test accuracy (0.63); strong generalization using ArcFace embeddings; lower efficiency than MLP due to longer training and tuning time.
Logistic Regression (ArcFace)	Extremely fast training; simple model with moderate accuracy and generalization; tuning time is relatively long.