1. Introduction

**Candidate Number: 262935**

## This report outlines the results of a binary classification task aimed at classifying photos as happy or sad. The dataset includes features extracted Convolutional Neural Network (CNN). The training data has two sets: one complete and one with missing values, accompanied by confidence labels. The test data also contains missing values.

## Approach

For this classification task, we use several machine learning models: Logistic Regression, Random Forest, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). These models were selected since their ability to handle high-dimensional data is proven and effective in various classification tasks.

Logistic Regression: A linear model for binary classification, efficient for large datasets.

Random Forest: An ensemble method that uses multiple decision trees, improving robustness and accuracy.

SVM: Effective in high-dimensional spaces and uses a subset of training points in the decision function.

MLP: A type of neural network suitable for capturing complex relationships in data

## Methods

3.1 Data Preprocessing

Missing Data: used K-Nearest Neighbors (KNN) imputation, which calculates missing values based on the values of the nearest neighbors, providing a more enhanced approach compared to simple mean imputation.

Feature Selection: Recursive Feature Elimination (RFE) with Random Forest as the estimator to select the most important features, reducing the feature space to 50 components.

Feature Scaling: Standardized features to have zero mean and unit variance using StandardScaler.

3.2 Model Training and Selection

Training-Validation Split: Combined and split into 80% training and 20% validation sets.

Cross-Validation: 5-fold cross-validation to evaluate the performance of each classifier.

3.3 Classifiers Evaluation

Logistic Regression: Optimized with default hyperparameters.

Random Forest: Used default hyperparameters with 100 trees.

SVM: Used RBF kernel with default settings.

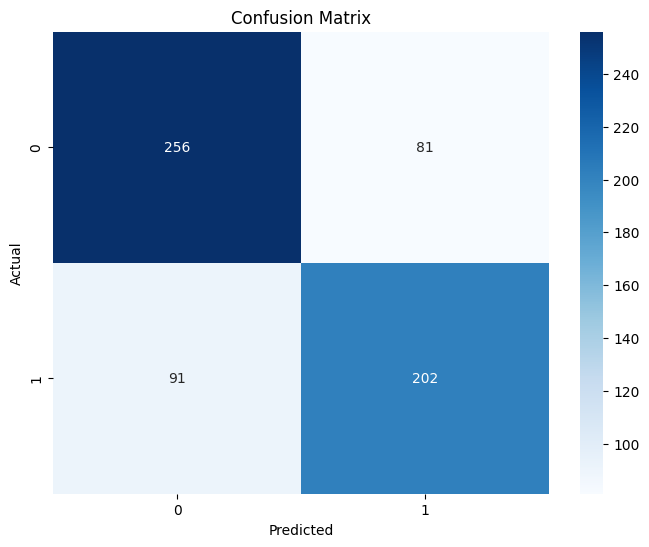
MLP: Set maximum iterations to 500 to ensure convergence.

4. Results and Discussion

4.1 Cross-Validation Scores

We performed cross-validation on four classifiers: Logistic Regression, Random Forest, SVM, and MLP. The bar plot below shows the mean cross-validation scores along with their standard deviations for each classifier. Logistic Regression and SVM showed the most consistent performance, with mean scores around 0.74 and 0.75, respectively. Random Forest had slightly more variability, while MLP had the lowest mean performance and the most variability.

4.2 Confusion Matrix



The confusion matrix for the best-performing model (Random Forest) on the validation set showed:

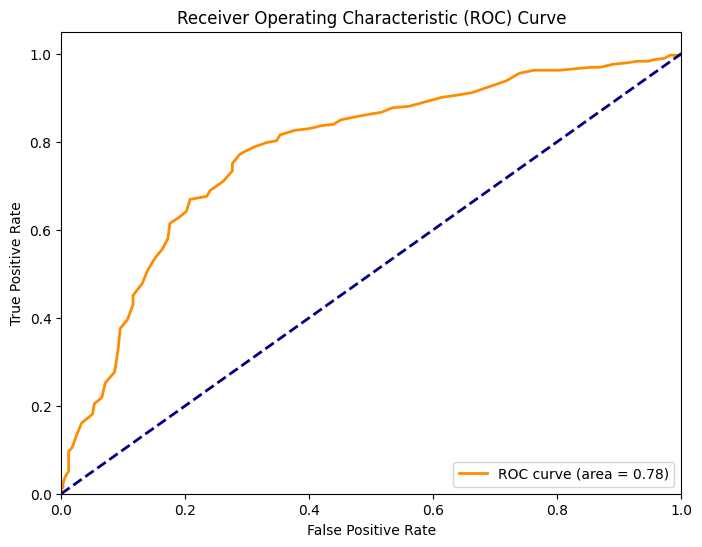
True Positives (TP): 202

True Negatives (TN): 256

False Positives (FP): 81

False Negatives (FN): 91

4.3 ROC Curve and AUC



The ROC curve and the area under the curve (AUC) for the best-performing model (Random Forest) indicated an AUC of 0.78, suggesting a good but improvable discrimination between happy and sad images.

4.4 Discussion

Performance Analysis: Logistic Regression and SVM provided the best accuracy

5. Conclusion

The classification task demonstrated the effectiveness of traditional classifiers

References

Brownlee, Jason. “Recursive Feature Elimination (RFE) for Feature Selection in Python.” Machine Learning Mastery, 24 May 2020, machinelearningmastery.com/rfe-feature-selection-in-python/.

Suh, Heajung, and Jongwoo Song. “A Comparison of Imputation Methods Using Machine Learning Models.” Communications for Statistical Applications and Methods, vol. 30, no. 3, 31 May 2023, pp. 331–341, https://doi.org/10.29220/csam.2023.30.3.331. Accessed 30 May 2024.

‌

‌