E-ISSN: 2962-6110
P-ISSN: 2964-7843
JSRET
jsret.knpub.com
Journal of Scientific Research, Education, and Technology


# Bayesian-optimized support vector machine for Indonesian ethnicity classification based on FaceNet facial features

**Alfito Juanda, Umar Zaky**
Universitas Teknologi Yogyakarta,Yogyakarta, Indonesia

**ABSTRACT**

The classification of ethnic groups in Indonesia based on facial images faces significant challenges due to high morphological diversity and the limitations of existing computational methods in handling local ethnic variations. This research developed a hybrid classification system to address this problem. The system was built through several stages: collecting a primary dataset of 550 facial images from five ethnic groups (Acehnese, Batak, Florenese, Javanese, and Papuan), extracting facial features using the FaceNet (InceptionResnetV1) model to generate face embeddings, and classification using a Support Vector Machine (SVM). To achieve maximum precision, the SVM model's hyperparameters were automatically tuned using Bayesian Optimization. The model's capabilities were confirmed using an 80/20 training-testing split, resulting an impressive 94.55% of accuracy. Its high discriminative power was further solidified by a stellar 0.9930 AUC-ROC score. Closer inspection showed a fascinating dichotomy: the model pinpointed the Papuan ethnicity with perfect precision, though it occasionally faltered when faced with the subtle morphological overlaps found in other ethnic groups. This study demonstrates that the combination of deep learning feature extraction with an optimized SVM classifier is an effective and robust approach for complex ethnicity classification, successfully providing an accurate and objective classification solution.

***Keywords:*** *Ethnicity Classification, Facial Image, Deep Learning, FaceNet, Support Vector Machine, Bayesian Optimization*

***Corresponding author***
***Name: Alfito Juanda***
***Email: alfitojuanda14@gmail.com***

## INTRODUCTION

The archipelagic nature of Indonesia has fostered a staggering tapestry of human diversity. The nation serves as a homeland to more than 1,300 distinct ethnic groups, each bearing its own unique cultural identity and physical attributes.(Baydhowi et al., 2023). This diversity reflected in a wide range of morphological facial features, which have traditionally been used for identification. However, due to increasing inter-ethnic assimilation and lifestyle changes, manual identification based on visual observation has become impractical(Das et al., 2025). This situation highlights a clear problem: the absence of an objective, automated system capable of accurately classifying Indonesia's diverse

ethnicities from facial images. The primary challenge lies in the fact that most state-of-the-art face analysis models are trained on global datasets, which fail to capture the subtle and specific facial variations of the Indonesian population, resulting in poor performance when applied locally(Patel & Ranjan Kisku, n.d.; Wirianto & Mauritsius, 2021).

Developing such a computational system, however, presents a formidable technical challenge. The vast majority of modern face recognition and analysis models are developed and trained on large-scale global datasets, which primarily feature Caucasian, East Asian, and African individuals(Kotwal & Marcel, 2025; Yucer et al., n.d.). When these generalized models are applied to the specific and highly varied facial morphologies of the Indonesian population, they exhibit a significant drop in performance and reliability(Melzi et al., 2024). The nuanced features that distinguish one Indonesian ethnic group from another are often lost or misinterpreted. Early attempts to solve this using classical machine learning methods, such as Local Binary Pattern (LBP) or Principal Component Analysis (PCA) combined with standard classifiers, have also proven insufficient, yielding low accuracy for this complex task(Putri et al., 2020; Wirayuda et al., 2023). While modern deep learning, particularly Convolutional Neural Networks (CNNs), has demonstrated superior performance in feature extraction, standard end-to-end implementations can struggle with poor generalization, especially when faced with the limited and imbalanced local datasets typical for this domain(Hancock & Burton, 1996; Septyono et al., n.d.).

The literature on computational ethnicity classification reflects these challenges. Globally, research has advanced to sophisticated architectures; for instance, Kalkatawi et al. explored the use of Multi-Axis Vision Transformers (MaxViT), achieving 77.2% accuracy on six broads, globally defined ethnic categories(Kalkatawi & Saeed, 2024). Other studies, such as Pardede et al., have focused on comparing standard CNN architectures like ResNet-152 and DenseNet-121 on public datasets, confirming the superiority of certain architectures for general racial classification(PARDEDE & KLEB, 2024). However, these studies do not address the specific, fine-grained task of intra-national classification required for the Indonesian context. Research focused specifically on Indonesian ethnicities is limited. Putriany et al. demonstrated the feasibility of the task, achieving a high accuracy of 98.65% by using classical methods (GLCM and Color Histogram) but focused only on the periorbital (eye) region(Putriany et al., 2021). While effective, this approach does not leverage the rich, holistic feature representations captured by deep learning. Conversely, a study by Putri using a Mask R-CNN model for Indonesian racial categories reported a high training accuracy of 97.34% but a very low validation accuracy of 66.28%, perfectly illustrating the critical problem of overfitting and poor generalization when standard deep learning models are applied to limited local data(Delta Pantika putri, n.d.). This reveals a clear research gap: a lack of a system that combines the powerful feature representation of deep learning with a classification strategy that ensures robust generalization for this specific and complex domain.

To address this gap, this paper proposes and validates a hybrid classification system specifically designed for Indonesian ethnicities. Our approach bifurcates the problem into two stages: representation and classification. For representation, we leverage

a pre-trained FaceNet model, specifically the InceptionResnetV1 architecture. This model, which was pre-trained on the massive VGGFace2 dataset, is used as a fixed feature extractor to convert raw facial images into highly discriminative 512-dimensional vector representations, known as face embeddings(Cao et al., 2018). For classification, these high-dimensional embeddings are then fed into a Support Vector Machine (SVM) classifier. An SVM is explicitly chosen for its proven efficacy in handling high-dimensional feature spaces, where it excels at finding an optimal, precise decision boundary (hyperplane) that separates complex classes(Kokare & Ghisare, n.d.).

The primary innovation of this work lies in the rigorous optimization of this hybrid model. Instead of relying on default parameters or manual tuning, we employ Bayesian Optimization to automatically tune the SVM's crucial hyperparameters, specifically C (regularization) and gamma (kernel coefficient)(Elshewey et al., 2023). This probabilistic optimization approach is significantly more efficient than Grid Search or Random Search, as it intelligently explores the parameter space to find the global optimum with fewer iterations(Greif et al., 2025). This research focuses on developing a robust automated tool designed specifically to classify Indonesian ethnic groups. We investigated two core aspects: the effectiveness of a hybrid FaceNet and Support Vector Machine approach compared to standard baselines, and the actual impact of Bayesian Optimization on model accuracy versus standard tuning. By addressing these points, the study provides a validated method for distinguishing between Acehnese, Batak, Florenese, Javanese, and Papuan ethnicities.

## METHOD

This research methodology employs a sequential pipeline, as illustrated in the framework (Figure 1), commencing with an 80/20 partitioning of the image dataset into training and testing subsets. A Multi-task Cascaded Convolutional Network (MTCNN) is first applied for precise face detection, after which the localized facial images are processed by a pre-trained FaceNet model for feature extraction, generating high-dimensional embeddings. These embeddings subsequently serve as the input for the Tuning & Training stage, where a Support Vector Machine (SVM) classifier is optimized using Bayesian Optimization (BayesSearchCV) to identify the best hyperparameters. Finally, the optimized model's performance is validated on the unseen test set through a comprehensive Model Evaluation, utilizing metrics such as classification reports and the ROC AUC to confirm the final system's accuracy.

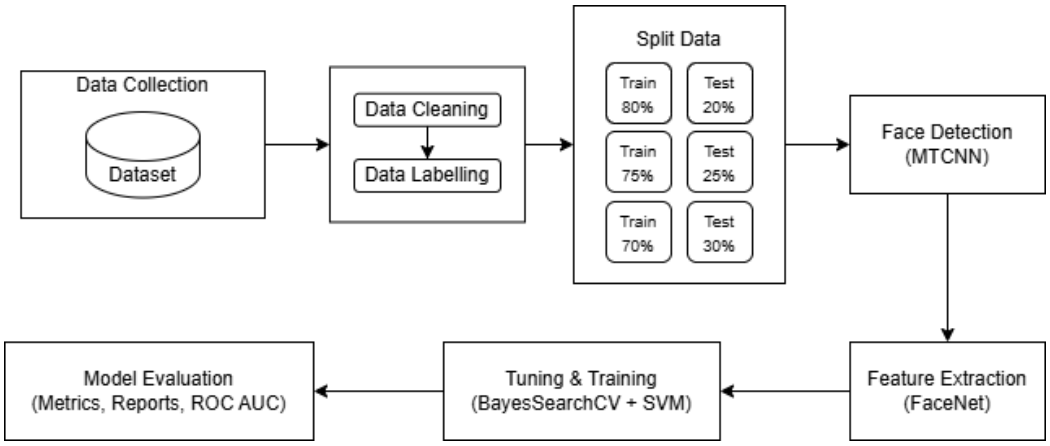


**Figure 1.** Research Framework

**Data Collection and Dataset**

This research is based on a primary dataset developed by the researcher, as no suitable public dataset for specific Indonesian ethnicities was available. The data acquisition process involved two primary methods: direct photo capture in public locations (specifically university campuses and surrounding workplaces) and contributions from the private collections of the researcher's family and colleagues. This sampling strategy resulted in a dataset primarily composed of subjects within an 18-40 year age range, with no representation of children or elderly subjects.

To ensure the integrity of the ethnic labels, a strict verification standard was applied. Each label was confirmed directly with the subject, who had to satisfy two criteria: (1) both biological parents were from the same ethnic group, and (2) the subject was born in their respective cultural region. This rigorous process replaced verification from a mere "trusted source". To ensure compliance with ethical research standards, informed consent was obtained from all subjects prior to their inclusion in the study. This included presenting each participant with a written statement guaranteeing that their facial images would be used solely for this research purpose and would not be misused.

The final dataset comprises a total of 550 images, specifically classified into five Indonesian ethnic categories: Acehnese (109 images), Batak (110 images), Florenese (107 images), Javanese (109 images), and Papuan (115 images). Figure 2 provides visual examples of facial images from each of these five defined categories. The dataset is relatively balanced, which helps prevent the model from being significantly biased toward any single class. However, it is noted that while gender was not a controlled variable, the resulting collection contains a probable bias toward male subjects over female subjects. This collection serves as the foundational data for training and validating the proposed model.

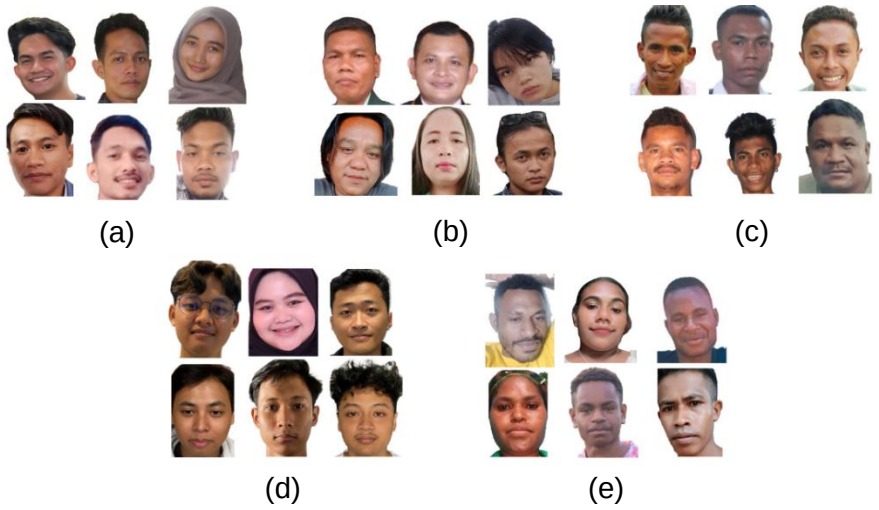


**Figure 2.** Image samples from ethnic (a) Acehnese (b) Batak (c) Florenese (d) Javanese (e) Papuan.

**Ethical Considerations and Limitations**

Ethical compliance was ensured by obtaining written informed consent from every subject prior to data collection. Participants were fully briefed on the research intent and signed agreements limiting the use of their images to academic research under strict confidentiality. However, the study faces demographic constraints inherent to the sampling context. While ethnically verified, the dataset leans towards males (60% vs. 40% female). Regarding age, the population is heavily concentrated in the 18–25 bracket due to the academic environment, though it does include a smaller subset of individuals aged 30–50 sourced from personal networks. Consequently, the dataset lacks data on children and the elderly. This implies that while the model is robust for young and middle-aged adults, future deployment across wider demographics would necessitate fine-tuning with a more age-inclusive dataset.

**Data Preprocessing**

To ensure data integrity and compatibility with the deep learning architecture, the entire dataset of 550 images underwent a rigorous, standardized preprocessing pipeline. The process commenced with a manual curation phase, where each image was reviewed to crop the primary subject and eliminate complex backgrounds, effectively reducing visual noise. Following this initial cleaning, a Multi-task Cascaded Convolutional Network (MTCNN) was deployed to automatically detect the precise bounding box of the face. The images were then cropped to this specific region to isolate facial features and discard non-essential data [20].

Subsequent transformations were applied to align the visual data with the specific input requirements of the InceptionResnetV1 architecture. Each cropped facial image was uniformly resized to dimensions of 160 \times 160 pixels. These images were then converted into PyTorch tensors, where the pixel values—originally ranging from [0, 255]—

were normalized to a range of [-1, 1]. This normalization was achieved by applying a standard transformation with both a mean and standard deviation of [0.5, 0.5, 0.5].

Finally, the fully processed and labeled dataset was partitioned into training and testing subsets using an 80:20 ratio, yielding 440 images for training and 110 images for evaluation. This partition was executed using the train_test_split function from scikit-learn. Crucially, the stratification parameter was enabled during this process to ensure that the proportional representation of each of the five ethnic classes was strictly maintained across both sets, thereby preventing sampling bias during model performance evaluation.

**Feature Extraction**

This study employs a hybrid architecture, separating the task of feature representation from classification. For feature representation, we utilize the FaceNet model, specifically the InceptionResnetV1 architecture. FaceNet is a deep convolutional network (DCN) specifically designed for facial recognition. Its primary function is to efficiently map face images directly into a low-dimensional Euclidean vector space, creating a numerical "embedding"(Schroff & Philbin, n.d.).

The model is trained using a triplet loss function, which optimizes the embedding by processing three image types simultaneously: an anchor (reference face), a positive (another face of the same person), and a negative (a face of a different person)(Schroff & Philbin, n.d.). The objective is to minimize the Euclidean distance between the anchor and positive vectors (intra-class compactness) while maximizing the distance between the anchor and negative vectors (inter-class separability). Because the resulting vector distance directly corresponds to facial identity, this embedding is highly effective for tasks like face classification, verification, and clustering.

A typical FaceNet-style configuration uses an Inception-ResNet backbone (or similar Inception-style architecture) which in many implementations takes a cropped aligned face of moderate resolution (commonly 160×160) as input and maps it to a compact embedding, most commonly 512-dimensional (though 128-dimensional versions also exist). Studies show performance degrades markedly when operating on very low-resolution inputs (for example < 30×30 pixels).

**Table 1. FaceNet (InceptionResnetV1) Configuration.**

| Function | Configuration |
|---|---|
| Input Image Size | 160 x 160 x 3 |
| Network Architecture | Inception-Style Network |
| Optimizer | SGD (Stochastic Gradient Descent) + AdaGrad + Learning Rate: 0.05 |
| Training Configuration | Loss Function: Triplet Loss |

Feature extraction in this study utilizes the FaceNet model, which produces a 512-dimensional embedding vector output (Schroff et al., 2015). Many implementations of the model adopt a moderate-resolution input (for example 160 × 160 × 3) to the underlying

Inception-style backbone network(Chun & Wang, n.d.). The backbone is built on an Inception-style architecture—in particular Inception-ResNet-V1 in some implementations—which is known for its parameter-efficiency and ability to learn multi-scale features in a single stage(Li et al., 2020).

The model's high discriminative power is derived from its Training Configuration (Table 1), which employs a Triplet Loss function. This loss function is specifically designed to optimize the embedding space by enhancing inter-class separability (increasing distance between different identities) and improving intra-class compactness (decreasing distance between the same identity) To achieve this robust feature representation, the model was trained on a massive dataset comprising 200 million images from eight million facial identities(Wang et al., n.d.). The optimization was performed using a combination of SGD (Stochastic Gradient Descent) and AdaGrad optimizers, with an initial Learning Rate of 0.05.

### Classification Model (Support Vector Machine)

This research employs a hybrid architecture that separates the task of feature representation from the final classification. While the pre-trained FaceNet (InceptionResnetV1) model serves as a powerful feature extractor (as described in section 2.3), the subsequent classification of these features is performed by a Support Vector Machine (SVM). SVM was specifically chosen for its proven efficacy and robustness in handling high-dimensional feature spaces. The 512-dimensional embedding vectors generated by FaceNet are complex and reside in a high-dimensional space, making SVM an ideal classifier for this task. Unlike standard end-to-end classifiers that might overfit on limited datasets, SVM is highly effective at finding a precise and optimal decision boundary (hyperplane) that maximizes the margin between classes, thereby enhancing generalization. This mechanism is illustrated in Figure 3.

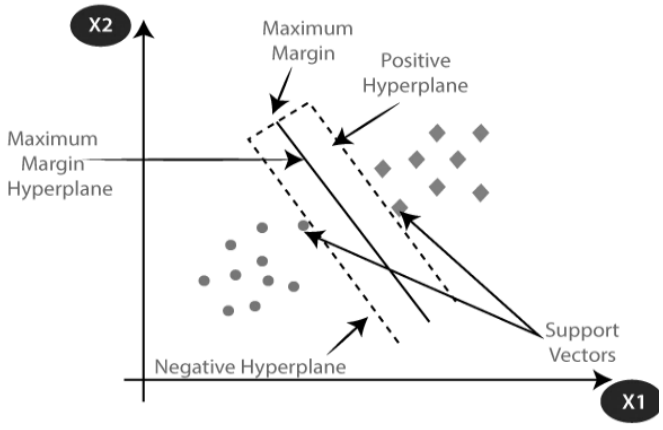


**Figure 3.** Architecture Support Vector Machine

As shown in Figure 3, the SVM works by identifying an optimal hyperplane that creates the largest possible distance, or "margin," between the data points of different classesThe model's decision boundary (hyperplane) is determined by the data points positioned closest to it. These crucial points are known as 'support vectors' and are instrumental in establishing the final position and orientation of the hyperplane(Montesinos López et al., 2022).

For complex classification tasks where data is not linearly separable in its original space, SVM employs a "kernel trick". as illustrated in Figure 4 This technique maps the input data into a higher-dimensional space where a linear separator can be found(Ningsih et al., 2024; PilarGautama et al., 2025). Given the non-linear complexities of facial morphology, this research utilizes the Radial Basis Function (RBF) kernel. The RBF kernel is a popular choice as it is highly flexible and effective at handling complex, non-linear relationships between data points.

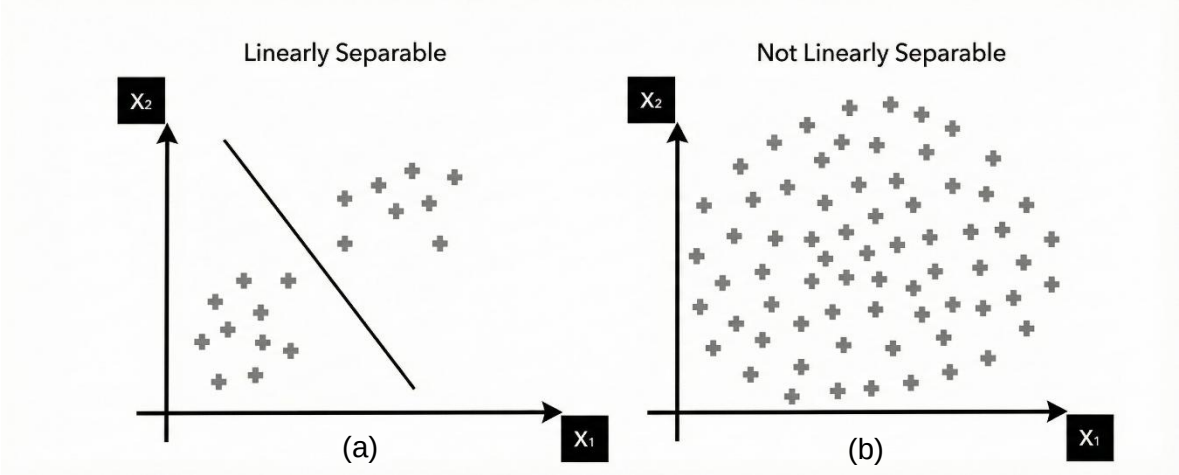


**Figure 4.** Illustration of (a) Linearly Separable data versus (b) Non-Linearly Separable data

In the context of this study's pipeline, the 512-dimensional face embeddings serve as the input vector (X) for the SVM model(Nuraeni & Faisal, 2025). The model is then trained to map these vectors to one of the five target ethnic classes (y): Acehnese, Batak, Florenese, Javanese, or Papuan. The crucial hyperparameters of the RBF kernel—specifically the regularization parameter C and the kernel coefficient gamma—are not manually set. Instead, they are systematically tuned in the subsequent phase using Bayesian Optimization (as described in section 2.5) to achieve maximum classification precision and robustness.

### Hyperparameter Optimization (Bayesian Optimization)

The performance of a Support Vector Machine (SVM) classifier, particularly with a Radial Basis Function (RBF) kernel, is critically dependent on the selection of its hyperparameters(Rizkallah, 2025). The main parameters, the regularization parameter C and the kernel coefficient gamma, collectively control the model's complexity and the

flexibility of the decision boundary(Fajri & Primajaya, 2023). An improper choice can easily lead to a model that either underfits (is too simple to capture the data's complexity) or severely overfits (is too complex and memorizes noise in the training data). A robust tuning strategy is necessary to find optimal hyperparameter combinations, as exhaustive methods like Grid Search are computationally inefficient, and Random Search lacks a systematic approach.

To overcome these limitations, this research employs Bayesian Optimization, a more intelligent and efficient global optimization strategy(Hakim et al., 2025; Saradhi, 2025). This method efficiently tunes costly functions (e.g., cross-validated SVMs) by using a probabilistic surrogate model to intelligently balance exploitation (testing known high-score areas) and exploration (testing uncertain areas)(Malu et al., n.d.). This informed approach allows the algorithm to converge upon the optimal hyperparameters in significantly fewer iterations.

In this study, the optimization was implemented using the BayesSearchCV function from the scikit-optimize (skopt) library, which integrates seamlessly with the scikit-learn pipeline. The optimization process was configured to find the best-performing C and gamma values for the RBF kernel. The specific configuration, search spaces, and validation strategy used for this optimization are detailed in Table 2.

**Table 2. Configuration of the Bayesian Optimization Process**

| Parameter | Value |
|---|---|
| Optimization Tool | BayesSearchCV (scikit-optimize) |
| Kernel | RBF |
| Search Space C | Real (1e-2 to 1e+2, log-uniform) |
| Search Space gamma | Real (1e-4 to 1e+0, log-uniform) |
| Total Iterations | 50 (n_iter=50) |
| Validation Strategy | 5-Fold Cross-Validation (cv=5) |

The final classifier used for evaluation in the "Results" section is the best_estimator_ returned by this BayesSearchCV process, representing the SVM model configured with the hyperparameter combination that yielded the highest mean cross-validation accuracy.

## Model Evaluation

The final assessment phase is designed to strictly validate the model's generalization capability using the unseen test data. To ensure an unbiased review, the stratified test set remains isolated during the training process and is used exclusively for this final check. The classification performance is quantified using four fundamental metrics derived from the standard Confusion Matrix components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

1. Accuracy
   Represents the global success rate of the classifier, computed as the proportion of accurate predictions relative to the entire dataset size (1).

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \quad (1)$$

2. Precision
   Assesses the exactness of positive class predictions. It is defined as the fraction of true positive instances among all instances labeled as positive by the model (2).

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

3. Recall
   Measures the completeness of the classification. It is calculated as the fraction of true positive instances that were correctly retrieved from the total actual positive samples (3).

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

4. F1-Score
   Delivers a composite metric that harmonizes Precision and Recall via their harmonic mean, providing a balanced view of performance (4).

$$F_1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Beyond these point-metrics, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is utilized to gauge the system's discriminative power across varying thresholds. This quantitative analysis is further substantiated by qualitative visual inspections using t-SNE for feature embedding separation and a Confusion Matrix to pinpoint specific inter-class misclassification patterns.

## FINDING AND DISCUSSION

### RESEARCH RESULT

This section presents the performance evaluation of the proposed hybrid model for Indonesian ethnicity classification. The proposed hybrid classification system, combining FaceNet feature extraction with a Bayesian-optimized Support Vector Machine (SVM), demonstrated robust performance in classifying Indonesian ethnicities. Upon evaluation on the unseen test set (20% of the total dataset), the model achieved a high Test Accuracy of 94.55% and an outstanding ROC AUC score of 0.9930. These results, validated through the multi-class Receiver Operating Characteristic (ROC) curve shown in Figure 5, confirm the efficacy of the proposed pipeline in effectively distinguishing between Acehnese, Batak, Florenese, Javanese, and Papuan facial features despite limited data availability.

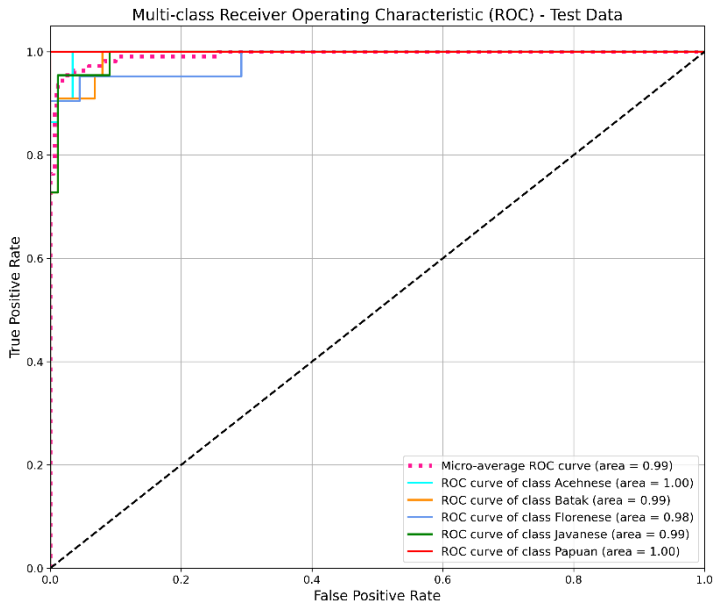


**Figure 5.** Multi-class Receiver Operating Characteristic (ROC) Curve on Test Data

For a more detailed analysis of the model's performance on each ethnic class, a classification report was generated, as shown in Table 3. This report includes the metrics of precision, recall, and F1-score for each class. Precision measures the accuracy of positive predictions, while recall measures the model's ability to identify all relevant instances of a class [4, 5]. The F1-score provides a harmonic mean of precision and recall, offering a single metric to evaluate their balance.

**Table 3. Classification report of the model on the test data.**

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Acehnese | 0.8800 | 1.0000 | 0.9362 | 22 |
| Batak | 0.9091 | 0.9091 | 0.9091 | 22 |
| Florenese | 1.0000 | 0.8571 | 0.9231 | 21 |
| Javanese | 0.9545 | 0.9545 | 0.9545 | 22 |
| Papuan | 1.0000 | 1.0000 | 1.0000 | 23 |
| Accuracy | | | 0.9455 | 110 |
| Macro avg | 0.9487 | 0.9442 | 0.9446 | 110 |
| Weighted avg | 0.9487 | 0.9455 | 0.9453 | 110 |

The results in Table 3 confirm the model's strong and consistent performance. The 'Papuan' class achieved a perfect F1-score of 1.0000, indicating flawless classification for that group. The 'Javanese' class also demonstrated very high performance with an F1-score of 0.9545. The 'Acehnese' class achieved a perfect recall of 1.0000, meaning no true 'Acehnese' samples were missed, although its precision was 0.8800, suggesting it misclassified some other ethnicities as 'Acehnese'. The weighted average F1-score of 0.9453

further confirms that the model is robust and well-balanced, performing accurately across the entire dataset without significant bias. The analysis indicates that the primary classification challenges occurred with the 'Florenese' class, which had the lowest recall (0.8571), and the 'Batak' class, which had the lowest F1-score (0.9091). These specific errors will be examined in the following error analysis section.

### Detailed Performance and Error Analysis

To gain deeper insights into the model's behavior and identify specific areas of confusion, a confusion matrix was generated Figure 6 presents a heatmap of this matrix, which visualizes the relationship between the true labels and the predicted labels for each class in the test set. The diagonal elements represent the number of correct predictions, while the off-diagonal elements indicate misclassifications.

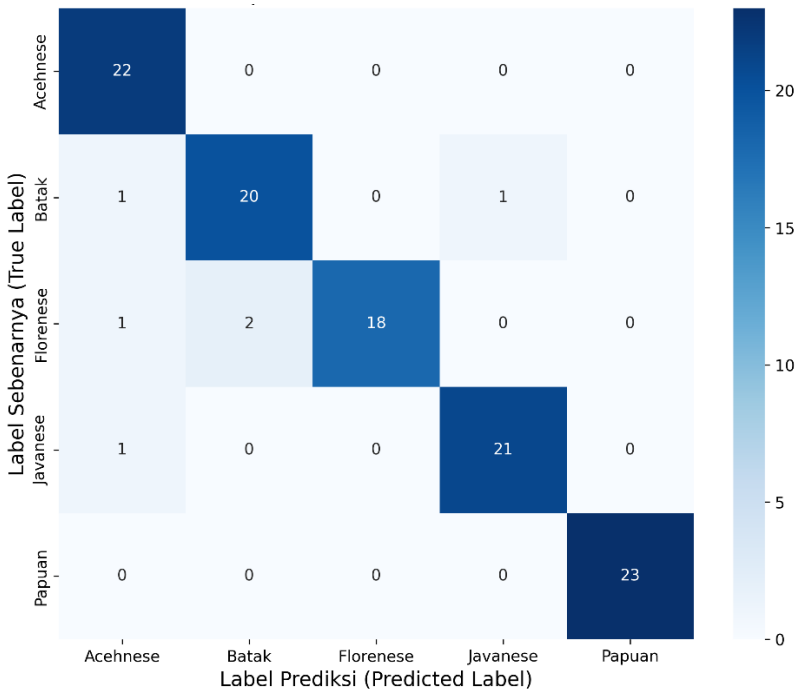


**Figure 6.** Heatmap of the Confusion Matrix on Test Data

The confusion matrix presented in Figure 6 substantiates the robust performance observed in the classification metrics, evidenced by the dense concentration of values along the main diagonal. This visual distribution confirms that the vast majority of predictions were accurate. Notably, the model demonstrated flawless execution on the 'Papuan' class, correctly identifying all 23 samples without a single misclassification. Furthermore, the 'Florenese' category achieved perfect precision; while the model missed some instances, every single sample it predicted as 'Florenese' was indeed correct, indicating an absence of false positives for this group.

Despite the overall success, the matrix illuminates specific zones of misclassification, particularly regarding the 'Florenese' and 'Acehnese' categories. The 'Florenese' class recorded the lowest recall, as three of its samples were mislabeled: one was incorrectly identified as 'Acehnese' and two were mistaken for 'Batak'. Conversely, the 'Acehnese' class exhibited the lowest precision. Although the model successfully retrieved all actual 'Acehnese' images (perfect recall), it was overly aggressive in its prediction, incorrectly assigning the 'Acehnese' label to single samples from the 'Batak', 'Florenese', and 'Javanese' groups.

These specific error patterns provide critical insights into the dataset's complexity. The recurring confusion between 'Florenese', 'Batak', and 'Acehnese' subjects strongly suggests a significant degree of morphological overlap in facial features among these three ethnicities. This visual similarity presents a greater challenge for the classifier compared to the 'Papuan' class, whose distinct physical characteristics allowed the model to distinguish it with absolute certainty.

### Feature Space Visualization (t-SNE)\

To qualitatively validate the quantitative findings from the confusion matrix and to understand why certain classes were confused, the 512-dimensional face embeddings were visualized using t-Distributed Stochastic Neighbor Embedding (t-SNE). This technique reduces the high-dimensional feature space into a 2D plot, allowing for a visual inspection of class separability. Visualizations were generated for both the training set (to observe how the model learned) and the test set (to observe how it generalized).

Figure 7 shows the t-SNE visualization for the Training Set. The plot clearly illustrates the model's learned feature separation. The cluster for the 'Papuan' class (magenta) is highly distinct and well-separated from all other classes on the left side of the plot. This wide separation in the feature space visually explains why the model was able to achieve a perfect F1-score for this class, as its features are highly unique. Conversely, the plot reveals a significant overlap between the data points for 'Florenese' (yellow), 'Batak' (lime green), 'Acehnese' (blue), and 'Javanese' (cyan). This confirms that the feature representations for these groups are closely situated, which directly correlates with the fundamental challenge of classifying them.

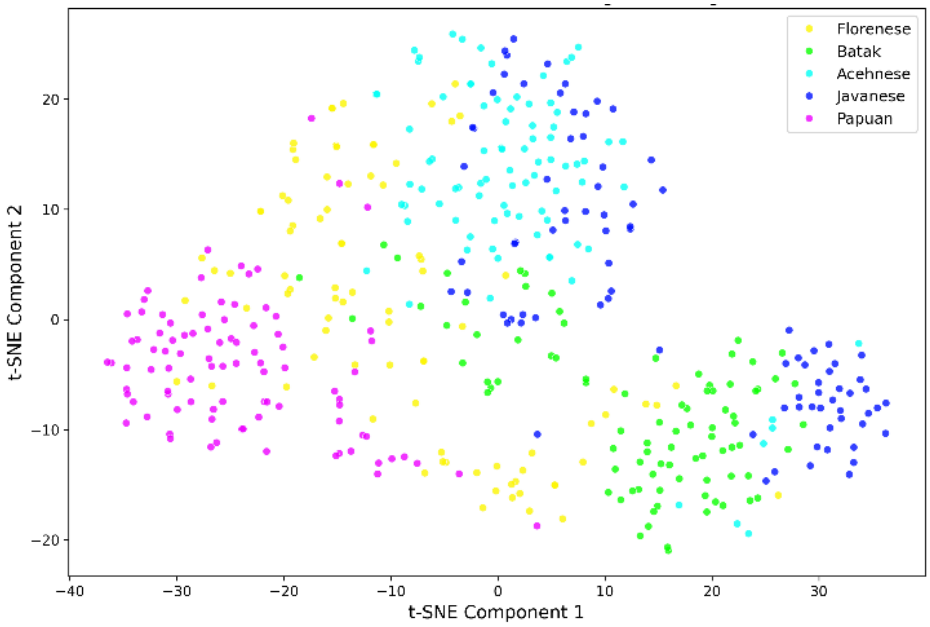


**Figure 7.** t-SNE Visualization of Face Embeddings from the Training Set

Figure 8 shows the t-SNE visualization for the unseen Test Set. This plot demonstrates how the model's learned representations generalize to new data. The findings from the training set are mirrored here: the 'Papuan' class (yellow) remains in a distinct cluster on the right side of the plot, confirming its separability and explaining its perfect classification on the test data. However, the visualization also confirms the generalization of the classification challenge, as the data points for 'Florenese' (cyan), 'Batak' (blue), and 'Acehnese' (magenta) are heavily intermixed, particularly in the large cluster on the left. In summary, the t-SNE analysis provides a compelling visual explanation for the performance patterns observed in the confusion matrix (Figure 6). The model's misclassifications are not random but are a direct result of the inherent morphological similarities between certain ethnic groups (Florenese, Batak, and Acehnese), which are captured as overlapping features by the FaceNet model.

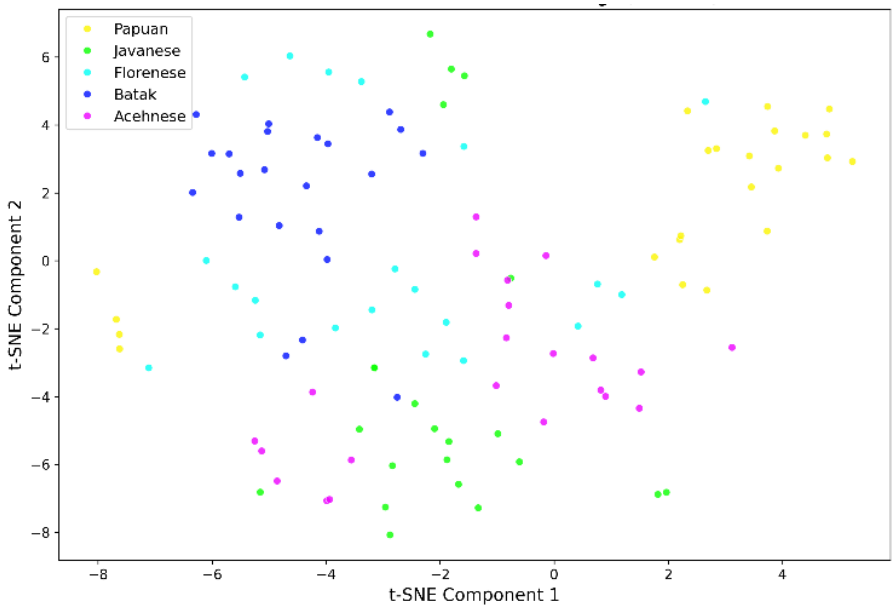


**Figure 8**. t-SNE Visualization of Face Embeddings from the Test Set

## DISCUSSION

The proposed hybrid methodology was compared against other modeling architectures and tuning strategies, as summarized in Table 4. A standard end-to-end fine-tuning approach using a ResNet-34 architecture (Exp. 1) performed poorly, achieving only 56.14% accuracy, likely due to overfitting on the limited dataset . A separate experiment to fine-tune only the classifier head of the FaceNet model, optimized with Optuna (Exp. 3), yielded a modest 81.64% cross-validation accuracy.

As summarized in Table 4, the proposed hybrid model using BayesSearchCV (Exp. 5) significantly outperformed other architectures, achieving the highest accuracy of 94.55%. It surpassed both standard fine-tuning methods (Exp. 1, 3) and less rigorous SVM optimizations (Exp. 2, 4). This confirms that combining FaceNet feature extraction with Bayesian Optimization is the most effective strategy for this classification task.

**Table 4. Performance Comparison of Different Architectures and Tuning Methods**

| Exp. | Approach | Tuning Method | Test Accuracy |
|---|---|---|---|
| 1 | End-to-End Fine-tuning (ResNet-34) | Manual | 56.14% |
| 2 | Feature Extraction + SVM | Manual (Hardcoded) | 75.53% |
| 3 | Fine-tuning Classifier Head (FaceNet) | Optuna | 81.64% (CV Avg) |
| 4 | Feature Extraction + SVM | RandomizedSearch | 77.48% |
| 5 | Feature Extraction + SVM | BayesSearchCV | 94.55% |

To validate the effectiveness of the proposed hybrid architecture (FaceNet + Bayesian-optimized SVM) and the chosen 80:20 data split, a series of comparative

experiments were conducted. The model was trained and evaluated using different train_test_split ratios to determine the optimal data partition. As detailed in Table 5, the 80:20 split achieved the highest overall test accuracy (94.55%) and weighted F1-Score (0.9453). This significantly outperformed the 75:25 split (92.03%) and the 70:30 split (88.48%). Furthermore, an experiment with 2x data augmentation on the 80:20 split was also conducted. This resulted in a decrease in accuracy to 91.82%, suggesting that simple augmentation may introduce noise or reduce model generalization for this specific task. Based on these results, the 80:20 split without augmentation was confirmed as the optimal configuration.

**Table 5. Performance Comparison of Different Data Splits and Augmentation**

| Experiment | Test Accuracy | Test ROC AUC | Weighted F1-Score |
|---|---|---|---|
| 70:30 Split | 88.48% | 0.9884 | 0.8850 |
| 75:25 Split | 92.03% | 0.9893 | 0.9200 |
| 80:20 Split + Augmentation | 91.82% | 0.9922 | 0.9182 |
| 80:20 Split (Proposed) | 94.55% | 0.9930 | 0.9453 |

Apart from quantitative success, the broader implications of deployment require attention. Given the demographic imbalance in the dataset, real-world application demands strict fairness protocols to avoid disproportionate errors against women or the elderly. From a technical standpoint, the shift from standardized, cropped images to unconstrained environments poses significant challenges. Real-world systems, whether for surveillance or attendance, face issues like variable lighting and pose changes that were minimized in this study. Consequently, future development must focus on strengthening pre-processing capabilities to maintain robustness under these dynamic conditions.

## CONCLUSION

This research successfully developed and validated a robust hybrid system for classifying diverse Indonesian ethnicities from facial images, a task complicated by subtle morphological similarities and the poor performance of globally trained models. The proposed methodology, which integrates a pre-trained FaceNet feature extractor with an SVM classifier. This classifier was then rigorously optimized using Bayesian Optimization.

A comparative analysis demonstrated the superiority of this specific architecture. The hybrid FaceNet+SVM model, when trained on an 80:20 data split, achieved a high-test accuracy of 94.55% and a ROC AUC score of 0.9930. This result significantly outperformed other configurations, including different data splits (70:30 and 75:25), data augmentation (which reduced accuracy to 91.82%), and alternative architectures like end-to-end fine-tuning (56.14% accuracy). The results confirm that for this specific domain, a hybrid approach that separates feature extraction from classification, combined with an intelligent hyperparameter search, is vastly more effective than standard end-to-end models.

While the model achieved perfect classification for the 'Papuan' class, the analysis of the confusion matrix and t-SNE visualizations identified clear challenges in distinguishing between visually similar groups, specifically 'Florenese', 'Batak', and 'Acehnese' This highlights an important area for future work. Future research should focus on enhancing the dataset's diversity and exploring fine-tuning techniques for the feature extractor, which may help the model learn the more nuanced features required to separate these overlapping classes. This study serves as a strong foundation for developing more accurate and contextually-aware biometric systems tailored to Indonesia's unique population, From a practical standpoint, this technology offers significant utility for demographic analytics in smart city initiatives and tourism planning. However, its potential implementation at a policy level must be strictly governed by ethical guidelines and privacy regulations to ensure that automated ethnicity classification serves to enhance cultural understanding rather than facilitate discriminatory profiling.

**REFERENCES**


Baydhowi, B., Purwono, U., Prathama Siswadi, A. G., Ali, M. M., Syahputra, W., & Iskandar, T. Z. (2023). Perception of threat and national identity: Investigation of the mediating role of collective self esteem. *Heliyon*, *9*(6). https://doi.org/10.1016/j.heliyon.2023.e17207

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). *VGGFace2: A dataset for recognising faces across pose and age*. http://arxiv.org/abs/1710.08092

Chun, C., & Wang, W. Y. (n.d.). *Face Recognition with Sub-Sampled Images*.

Das, S. R., Champatyray, S., & Panda, D. K. (2025). Anthropometric analysis of facial dimensions using 3D imaging for forensic identification and ethnicity-specific reference models. *Forensic Science International: Reports*, *12*. https://doi.org/10.1016/j.fsir.2025.100428

Delta Pantika putri. (n.d.). *DETEKSI SUKU/RAS DI INDONESIA BERDASARKAN WAJAH MENGGUNAKAN METODE INSTANCESEGEMENTATION MASK-RCNN*.

Elshewey, A. M., Shams, M. Y., El-Rashidy, N., Elhady, A. M., Shohieb, S. M., & Tarek, Z. (2023). Bayesian Optimization with Support Vector Machine Model for Parkinson Disease Classification. *Sensors*, *23*(4). https://doi.org/10.3390/s23042085

Fajri, M., & Primajaya, A. (2023). Komparasi Teknik Hyperparameter Optimization pada SVM untuk Permasalahan Klasifikasi dengan Menggunakan Grid Search dan Random Search. In *Journal of Applied Informatics and Computing (JAIC)* (Vol. 7, Issue 1). http://jurnal.polibatam.ac.id/index.php/JAIC

Greif, L., Hübschle, N., Kimmig, A., Kreuzwieser, S., Martenne, A., & Ovtcharova, J. (2025). Structured sampling strategies in Bayesian optimization: evaluation in mathematical and real-world scenarios. *Journal of Intelligent Manufacturing*. https://doi.org/10.1007/s10845-025-02597-2

Hakim, S. U. El, Arifianto, R., Sugiyanto, S., Pratiwi, I. A. P., Bahari, G., & Krisnaputra, R. (2025). Bamboo Diameter Detection System Based on Image Processing as a Pre-Assessment for an Automated Bamboo Splitting Technology. *Kinetik: Game*

*Technology, Information System, Computer Network, Computing, Electronics, and Control*. https://doi.org/10.22219/kinetik.v10i2.2170

Hancock, P. B., & Burton, A. M. (1996). Face processing: Human perception and principal components analysis. In *Memory & Cognition* (Vol. 24, Issue I).

Kalkatawi, A.-A., & Saeed, U. (2024). Ethnicity Classification Based on Facial Images using Deep Learning Approach. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 15, Issue 2). www.ijacsa.thesai.org

Kokare, S., & Ghisare, V. (n.d.). SVM-Based Approach for Human Face Detection and Recognition. In *International Journal on Science and Technology*.

Kotwal, K., & Marcel, S. (2025). *Review of Demographic Fairness in Face Recognition*. https://doi.org/10.1109/TBIOM.2025.3601217

Li, H. C., Deng, Z. Y., & Chiang, H. H. (2020). Lightweight and resource-constrained learning network for face recognition with performance optimization. *Sensors (Switzerland)*, *20*(21), 1–20. https://doi.org/10.3390/s20216114

Malu, M., Dasarathy, G., & Spanias, A. (n.d.). *Bayesian Optimization in High-Dimensional Spaces: A Brief Survey*. https://distill.pub/2020/bayesian-optimization/.

Melzi, P., Rathgeb, C., Tolosana, R., Vera-Rodriguez, R., Morales, A., Lawatsch, D., Domin, F., & Schaubert, M. (2024). *Synthetic Data for the Mitigation of Demographic Biases in Face Recognition*. http://arxiv.org/abs/2402.01472

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Support Vector Machines and Support Vector Regression. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction* (pp. 337–378). Springer International Publishing. https://doi.org/10.1007/978-3-030-89010-0_9

Ningsih, M. R., Unjung, J., Pertiwi, D. A. A., Prasetiyo, B., & Muslim, M. A. (2024). Optimized Support Vector Machine with Particle Swarm Optimization to Improve the Accuracy Amazon Sentiment Analysis Classification. *KINETIK*, *9*(1), 101–108. https://kinetik.umm.ac.id/index.php/kinetik/article/view/1888https://kinetik.umm.ac.id/index.php/kinetik/article/view/1888

Nuraeni, N., & Faisal, M. (2025). Classification of Sleep Disorders Using Support Vector Machine. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*. https://doi.org/10.22219/kinetik.v10i1.2054

PARDEDE, J., & KLEB, S. S. (2024). Face Race Classification using ResNet-152 and DenseNet-121. *ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika*, *12*(3), 798. https://doi.org/10.26760/elkomika.v12i3.798

Patel, S., & Ranjan Kisku, D. (n.d.). *Improving Bias in Facial Attribute Classification: A Combined Impact of KL Divergence induced Loss Function and Dual Attention*.

PilarGautama, H., Prasetiyowati, S. S., & Sibaroni, Y. (2025). Land Price Distribution Prediction in Jakarta Using Support Vector Machine with Feature Expansion and Kriging Interpolation. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*. https://doi.org/10.22219/kinetik.v10i3.2216

Putri, T. T., Rachmawati, E., & Sthevanie, F. (2020, November 10). Indonesian Ethnicity Recognition Based on Face Image Using Uniform Local Binary Pattern (ULBP) and Color Histogram. *ICICoS 2020 - Proceeding: 4th International Conference on Informatics and Computational Sciences*. https://doi.org/10.1109/ICICoS51170.2020.9299103

Putriany, D. M., Rachmawati, E., & Sthevanie, F. (2021). Indonesian Ethnicity Recognition Based on Face Image Using Gray Level Co-occurrence Matrix and Color Histogram. *IOP Conference Series: Materials Science and Engineering*, *1077*(1), 012040. https://doi.org/10.1088/1757-899x/1077/1/012040

Rizkallah, L. W. (2025). Optimizing SVM hyperparameters for satellite imagery classification using metaheuristic and statistical techniques. *International Journal of Data Science and Analytics*, *20*(5), 4945–4962. https://doi.org/10.1007/s41060-025-00762-7

Saradhi, T. V. (2025). A Study on Hyperparameter Tuning in Support Vector Machines and its Impact on Model Accuracy. In *Global Journal of Engineering Innovations & Interdisciplinary Research GJEIIR* (Vol. 5, Issue 1).

Schroff, F., & Philbin, J. (n.d.). *FaceNet: A Unified Embedding for Face Recognition and Clustering*.

Septyono, M. B., Anggraeny, F. T., & Mumpuni, R. (n.d.). *JIP (Jurnal Informatika Polinema) Pengenalan Ekspresi Wajah dengan LBP dan Multi-Level CNN*.

Wang, K., Wang, S., Zhang, P., Zhou, Z., Zhu, Z., Wang, X., Peng, X., Sun, B., Li, H., & You, Y. (n.d.). *An Efficient Training Approach for Very Large Scale Face Recognition*.

Wirayuda, T. A. B., Munir, R., & Kistijantoro, A. I. (2023). Compact-Fusion Feature Framework for Ethnicity Classification. *Informatics*, *10*(2). https://doi.org/10.3390/informatics10020051

Wirianto, & Mauritsius, T. (2021). The development of face recognition model in indonesia pandemic context based on dcnn and arcface loss function. *International Journal of Innovative Computing, Information and Control*, *17*(5), 1513–1530. https://doi.org/10.24507/ijicic.17.05.1513

Yucer, S., Tektas, F., Al Moubayed, N., & Breckon, T. P. (n.d.). *Measuring Hidden Bias within Face Recognition via Racial Phenotypes*.