RESEARCH SERIES.

UNIVERSITY OF TEXAS AT ARLINGTON

WAYNE STATE UNIVERSITY

Florida Atlantic University

Performance Analysis of Hybrid Deep-Transfer Learning Approaches with Machine Learning Methods for Face Recognition

Mohsen Mohammadagha[1]✉*, Atena Khoshkonesh[1], Shayan Sharifi[2] Vahid Ghanbarizadeh[3], Farbod Bigdeli[1]

[1] University of Texas at Arlington, Arlington, Texas, USA.
[2] Wayne State University, Detroit, Michigan, USA.
[3] Florida Atlantic University, Boca Raton, Florida, USA.

Abstract: This study addresses the pressing challenge of supervised face recognition under unconstrained conditions by systematically integrating classical machine learning, deep learning, and transfer learning approaches. While existing literature demonstrates significant progress, particularly with hybrid and transfer learning models, a gap remains in unified, detailed benchmarking across diverse techniques. The primary objective is to holistically compare Support Vector Machines (SVM) with PCA, Artificial Neural Networks (ANN), XGBoost, a custom Convolutional Neural Network (CNN), MobileNetV2-based transfer learning, and stacked hybrid meta-models using the Labeled Faces in the Wild (LFW) dataset. The methodology encompasses data preprocessing, parallel feature extraction, dimensionality reduction, ensemble learning, and interpretability analysis. Experimental results show that stacking hybrid models achieves the highest test accuracy (87.9%) and macro ROC-AUC (0.983), with MobileNetV2 transfer learning also excelling in sample efficiency and performance. Future research should expand interpretability diagnostics and benchmark these pipelines on more diverse or occluded datasets for greater real-world applicability.

Keywords: *Machine Learning, Deep Learning, Convolutional Neural Network, Support Vector Machine, XGBoost, Meta-Models.*

✉ **Corresponding author**
Mohsen Mohammadagha
University of Texas at Arlington
701 S. Nedderman Drive
Arlington, Texas
USA
Email: mxm4340@mavs.uta.edu
Orcid: 0009-0007-0394-353X

1 INTRODUCTION

The rapid advancement of artificial intelligence and machine learning has significantly transformed the field of computer vision [1], particularly in the domain of face recognition and image classification [2]. Face recognition systems are now widely deployed in diverse applications, ranging from security [3] and surveillance to personal device authentication and social media tagging [4]. At the heart of these systems lies the challenge of accurately identifying or verifying individuals from digital images [5], often captured under unconstrained, real-world conditions. This challenge has spurred the development of increasingly sophisticated algorithms that can handle variations in pose, illumination, occlusion, and expression. The integration of deep learning, along with Convolutional Neural Networks (CNNs) [6], with classical machine learning models such as Support Vector Machines (SVMs) [7], and Artificial Neural Networks (ANNs), has opened new avenues for constructing robust and scalable face recognition pipelines. These hybrid approaches [8] [9] leverage the feature extraction power of deep learning [10] and the classification strengths [11] of traditional algorithms, promising improved accuracy and generalization [12] [13].

A key resource for progress in this field is the availability of large, labeled datasets like the Labeled Faces in the Wild (LFW) dataset. The LFW dataset, which includes over 13,000 face images of 5,749 individuals [14], is specifically created to support research on unconstrained face recognition. Each image in the dataset is labeled with the person's identity, and the collection contains multiple images of many individuals, enabling the development and testing of both identification and verification algorithms. The images are collected from the web and reflect a wide range of real-world conditions, making the dataset a benchmark for evaluating face recognition models. Using deep-funneled images enhances the dataset's effectiveness for accurate face verification by aligning facial features more precisely across samples [15].

Modern face recognition systems typically follow a multi-stage pipeline [16]. The process begins with face detection, where algorithms such as Haar Cascades [17] or deep learning-based detectors identify and localize faces within an image or video frame. Following detection, feature extraction is performed to generate a compact digital representation—or "faceprint"—of each face [18]. This representation captures key attributes such as the distance between the eyes, the shape of the jawline, and other distinctive facial landmarks. In this context, deep learning models—such as custom CNNs and transfer learning architectures like MobileNetV2—excel at extracting robust, discriminative facial representations from pixel data. Classical classifiers like SVMs [19], used in conjunction with dimensionality reduction via Principal Component Analysis (PCA), remain highly effective at leveraging extracted features for accurate identity assignment [20]. Ensemble and meta-models, such as stacking combinations of SVM and XGBoost, are gaining traction for their potential to harness complementary strengths of multiple algorithms.

The emergence of transfer learning [21] has further propelled the performance of face recognition and image classification [22] systems. Transfer learning enables the adaptation of pre-trained deep learning models, originally trained on large, generic datasets—to the specific task of face recognition by fine-tuning them on domain-specific data such as LFW. This approach not only accelerates the training process but also enhances model accuracy by leveraging previously learned visual features. In hybrid models, CNNs are often used for feature extraction, while SVMs [23] serve as

the final classifier, creating a synergistic system that combines the strengths of both paradigms. The code which have been used in this study exemplifies this methodology, employing MobileNetV2 for transfer learning, a custom CNN for direct image classification, and an SVM with PCA for traditional machine learning-based recognition. Such hybrid pipelines are evaluated using a suite of metrics—including accuracy, precision, recall, F1-score [24], confusion matrices [25], and ROC curves—to provide a comprehensive assessment of their effectiveness [26].

In summary, the integration of hybrid transfer learning approaches, combining CNNs and SVMs, represents strategy for supervised face recognition and image classification [27]. By leveraging large, diverse datasets like LFW and employing advanced machine learning and deep learning techniques, it can address the inherent challenges of real-world face recognition. Notably, this research introduces a comprehensive suite of diagnostics—including per-class error tables, advanced visualizations (Grad-CAM, saliency maps), and contrastive supervised CNNs—offering deeper insight into model behavior and error analysis [28], than standard benchmarks. This introduction sets the stage for an exploration of the methodologies, experimental results, and performance analysis [29], that follow.

The basic structure of this paper is as follows. Section 2 reviews the most pertinent recent works on machine learning methodologies for face recognition, highlighting advances in hybrid and transfer learning models under unconstrained and masked conditions. Section 3 details the overall methodology, including unified data preprocessing, the hybrid meta-model framework, and the suite of evaluation metrics used to interpretable benchmarking. Section 4 presents the experimental results, providing both an integrated comparison of stacking ensembles hybrid meta-model and the performance of individual classical, deep learning, and transfer learning approaches. Section 5 discusses the empirical findings in depth, including error analysis, per-class diagnostics, and interpretability visualizations that reveal strengths and weaknesses of each pipeline. Finally, Section 6 concludes the paper and offers recommendations for practical deployment and future research directions in robust and transparent face recognition systems.

2 RELATED WORKS

Hybrid and ensemble approaches have proven effective in boosting the accuracy and robustness of face recognition systems. Opanasenko et al. (2024) demonstrated that combining algorithms such as SVMs and component-based models in an ensemble yields a notable accuracy of 98.84%, outperforming the best individual method at 95.31% [30]. Conversely, Shi et al. (2025) found that advanced multimodal large language models (MLLMs) like GPT4V and Gemini achieved only moderate accuracy in face anti-spoofing, with GPT4V reaching 33.1% and Gemini 25.6% in zero-shot settings, indicating persistent challenges in generalization [31].

Integrating convolutional neural networks (CNNs) and support vector machines (SVMs) has also been shown to enhance face recognition performance. Serengil and Ozpinar (2024) reported that hybrid CNN-based models such as FaceNet-512d achieved up to 98.4% accuracy on the LFW dataset, surpassing human-level benchmarks [32]. Tao et al. (2016) introduced a locality-sensitive SVM that fuses local CNN features, which demonstrated superior robustness to occlusion and illumination, outperforming traditional detectors on datasets like CMU+MIT and FDDB [33].

The challenge of recognizing masked and unconstrained faces has prompted the development of more resilient models. George et al. (2024) introduced EdgeFace, a hybrid CNN-Transformer model, achieving 99.73% on LFW and 94.85% on IJB-C [34], while Wang et al. (2023) showed that ArcFace's accuracy drops from 90.66% on MFR-ALL to 72.74% on RMFRD under mask occlusion. This highlights the need for hybrid, transfer learning-based approaches that maintain high performance in diverse scenarios [35].

Large-scale and practical deployments further emphasize the value of hybrid pipelines. Srinivasan et al. (2024) achieved 95% accuracy with an IoT-based OpenCV system for smart hospitality [36], while Zhu et al. (2022) used the WebFace42M dataset and a hybrid deep learning pipeline to reach 99.83% on LFW and 97.70% TAR@FAR=1e-4 on IJB-C, significantly reducing masked face error rates. Collectively, these works demonstrate that combining CNNs, SVMs, and transfer learning leads to superior and more robust supervised face recognition and image classification, motivating the analysis undertaken in this study [37].

3 METHODOLOGY

The methodological foundation of this research centers on a comprehensive, comparative evaluation of hybrid transfer learning and deep learning approaches for supervised face recognition using the Labeled Faces in the Wild (LFW) dataset. Methodologically aligning with standards set by contemporary journals, the study is structured as a sequential multi-stage pipeline: robust data pre-processing [38], and normalization, hybrid model design and training, implementation of both traditional and state-of-the-art neural network architectures, and model evaluation using statistically validated metrics. By grounding all experiments within the LFW dataset—widely recognized for its challenging unconstrained capture conditions and demographic diversity—this work can facilitate direct benchmarking and advance the science of real-world face recognition. The design prioritizes fair comparison, cross-validated training/test splits [39], and multiple experimental runs to ensure reproducibility and statistical validity.

The core models as shown in **Figure 1** assessed include Support Vector Machine (SVM) with Principal Component Analysis (PCA) for feature reduction, Artificial Neural Network (ANN), XGBoost , a custom Convolutional Neural Network (CNN), transfer learning with MobileNetV2, and a stacked meta-model combining SVM and XGBoost in an mixture framework. Classic machine learning pipelines extract and compress features using PCA before classification, capitalizing on SVM's robustness in high-dimensional settings. The custom CNN [40] is architected to learn hierarchical representations directly from pixels, while MobileNetV2 [41] leverages pre-trained weights from large general-purpose datasets and is fine-tuned for faces from LFW. The stacking meta-model is designed to harness complementary advantages of individual algorithms, improving overall generalization. All models are trained using stratified splitting, with hyperparameters optimized via cross-validation to prevent overfitting and maximize applicability [42].

Figure 1. Workflow for Performance Analysis of Hybrid Transfer and Deep Learning Approaches.

Statistical and theoretical rigor underpin the evaluation process, drawing on five foundational performance formulas commonly employed in the face recognition literature. These criteria ensure interpretability across studies and include: (1) Accuracy [43]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

which quantifies the proportion of correct predictions [44]; (2) Precision:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

reflecting the correctness of positive predictions; (3) Recall (Sensitivity):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

measuring the ability to detect all positive instances; (4) Specificity [45]:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

indicating the ability to identify negative instances correctly; (5) F1-score:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

offering a balanced harmonic mean; Area Under the ROC Curve (AUC), quantifying classifier discrimination capability over all thresholds; and Confusion Matrix analysis, which visually and quantitatively details per-class prediction strengths and weaknesses. These metrics are complemented by advanced visualizations, including ROC curves and t-SNE feature space embeddings, ensuring granular and interpretable analysis for each method.

A crucial methodological element lies in the careful adaptation of the LFW dataset. As recommended by the literature and dataset authors, the study uses funneled versions of images, ensuring optimal facial alignment and feature extraction quality. The dataset used in this report is the Labeled Faces in the Wild (LFW) Funneled Dataset. It consists of over 13,000 face images collected from the web, each labeled with the name of the person depicted. The dataset is organized into folders by individual, with a total of 5,749 identities and at least 1,680 people having two or more images. The images are primarily JPEG files, originally sized at 250 x 250 pixels, but often resized for analysis (commonly to 62 x 47 pixels for machine learning tasks). Each image is centered on a single face and was detected using the Viola-Jones face detector [15]. Preprocessing pipelines ensure scaling, normalization, and augmentation (for deep models). Main performance evaluation in this study is based on a single split of the dataset into training and test sets. The test set remains separate and is used only for the final evaluation to ensure a fair assessment. Although 5-fold splits are used internally for some visualizations and for training the meta-model, the overall comparison is performed using this train-test division. The experimental protocol routinely compares each modeling approach under identical data splits and conditions, allowing for robust, generalizable insights into their comparative strengths.

This research distinguishes itself through the integrated and evaluation of traditional machine learning, deep learning, transfer learning, and collective (stacked meta-model) approaches within a single, detailed framework. While prior studies often isolate specific modeling families or focus on homogeneous datasets or metrics, this

work fills a gap by concurrently benchmarking hybrid pipelines using a comprehensive suite of modern and classic performance measures, under consistent and challenging test conditions. This research lies not only in the hybridization strategy—combining CNN feature extraction with SVM and XGBoost classification—but also in the transparent, statistically evaluated design, bridging the gap identified in prior literature for unified, fair, and deeply comparative face recognition assessment. The outcomes of this study thus advance both the technical and methodological state of the art, supporting the design of future systems and theoretical analyses in face recognition.

4 RESULTS

The results of this study provide a comprehensive evaluation of classical, deep, and hybrid machine learning models for supervised face recognition on the LFW dataset. Performance is assessed using a range of metrics and visualization techniques, including confusion matrices, ROC curves, t-SNE projections, and analysis of recognition rates across varying dimensionalities. By comparing Support Vector Machines (SVM), Artificial Neural Networks (ANN), XGBoost, custom CNN architectures, MobileNetV2-based transfer learning, and composite stacking, the analysis highlights both individual strengths and complementarities among these methods. The following section details these empirical findings, offering insights into the interplay between model design and real-world face recognition challenges.

Figure 2 visually summarizes the comparative performance and diagnostic analyses conducted for supervised face recognition using hybrid transfer learning and classical machine learning models, as detailed in the accompanying image. The top left panel shows the SVM confusion matrix heatmap, illustrating class-wise prediction errors and correctly classified samples for the seven-face class LFW dataset; this matrix validates that some identities—such as Colin Powell and George W. Bush—are far less frequently misclassified compared to others like Ariel Sharon, revealing disparities due to both class imbalance and inherent inter-class similarity (as further highlighted in the normalization and per-class error analyses). The top right t-SNE plot projects the high-dimensional SVM feature space into two dimensions, revealing the separability of classes in the learned embedding; the observed overlap and cluster spread further corroborate the confusion matrix's findings, notably for more confusable identities. At lower left and right, dual line plots compare model robustness under dimensionality reduction: macro ROC-AUC (left) and recognition rate (accuracy, right) as functions of PCA-reduced input feature set size, for SVM+PCA, ANN+PCA, and XGBoost+PCA. SVM+PCA consistently achieves the highest AUC and accuracy, demonstrating its strength in leveraging structurally meaningful, low-dimensional facial embeddings; ANN+PCA performs slightly below but climbs closer as dimensionality increases, leveraging its capacity for non-linear feature learning given more information; XGBoost+PCA trails both, showing a more variable curve and plateauing at lower recognition rates, indicating its relative weakness with linearly-compressed features. Each curve's distinct trajectory stems from the unique interplay between model bias-variance tradeoff and capacity for extracting discriminative cues under compression: SVM thrives on optimal margin separation in compact spaces, ANN benefits from greater dimensionality for feature interaction, and XGBoost, designed for tabular, high-cardinality attributes, underperforms on principal component-compressed data. The stacking meta-model result, reflected in the figure captions, attests to the Integrated ability to consolidate complementary strengths, which achieves superior macro ROC-AUC (0.983) and accuracy (87.9%) by mitigating individual model

weaknesses and enhancing generalization across challenging, unconstrained face scenarios.

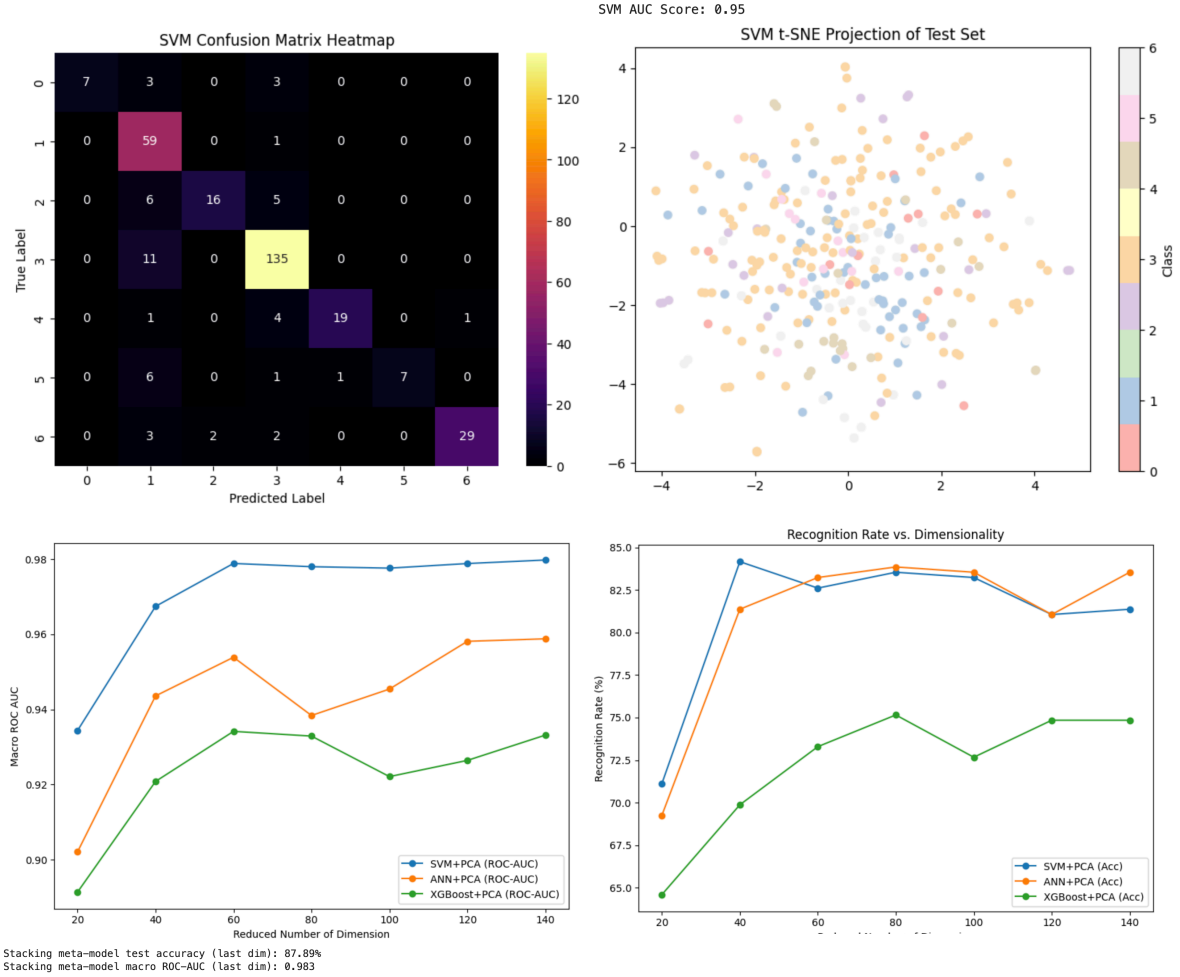

Figure 2. Performance, diagnostic comparison and confusion matrix.

For providing a comprehensive visual summary of the comparative performance and diagnostics for supervised face recognition using hybrid deep learning and classical machine learning approaches, as outlined in **Figure 3**. The upper left ROC curve plot compares the discriminative power of four pipelines: SVM+PCA, ANN+PCA, XGBoost+PCA, and their committee stacking. Each curve is distinct due to the differing learning strategies—SVM+PCA emphasizes optimal margin separation in reduced dimension spaces and achieves the highest area under the curve (AUC=0.98), indicating strong generalization after PCA compression; ANN+PCA, equipped with nonlinear activation functions, improves as more principal components are included, but is marginally less effective when dimensions are low; XGBoost+PCA performs best with higher-cardinality tabular features and underperforms after PCA due to loss of granular structure in the features, reflected by its lower AUC; the stacking integrates predictions, leveraging complementary strengths and yielding robust, superior aggregate performance. The top right confusion matrix visually demonstrates per-class trade-offs, highlighting which identities (rows) are most often misclassified as others (columns) and revealing underlying dataset imbalance and inter-class similarity; darker cells along the diagonal indicate higher correct classification rates for classes with ample data or distinct facial structure. The mid and lower panels further dissect SVM performance: the per-class ROC curves (lower left) evidence that certain identities are intrinsically easier to distinguish—curves close to the top left reflect stronger

discrimination—due to pronounced facial uniqueness or larger sample counts. The CNN ROC curves (lower right) show more modest AUCs, attributable to the architectural differences: deep CNNs, while powerful, can underperform with limited or imbalanced data. The precision-recall-F1 bar plot and sample count histogram contextualize these findings by quantifying the relationship between sample availability and classification reliability—classes with fewer samples (e.g., Ariel Sharon, Hugo Chavez) experience lower F1 scores, while abundant classes (e.g., George W. Bush) benefit from more stable estimates. The distinct shape and position of each curve or bar in these graphs directly stem from model capacity, bias-variance trade-offs, and the effectiveness of feature extraction and dimensionality reduction schemes employed in the hybrid recognition pipeline, underlining why no single approach or metric suffices for a complete assessment.

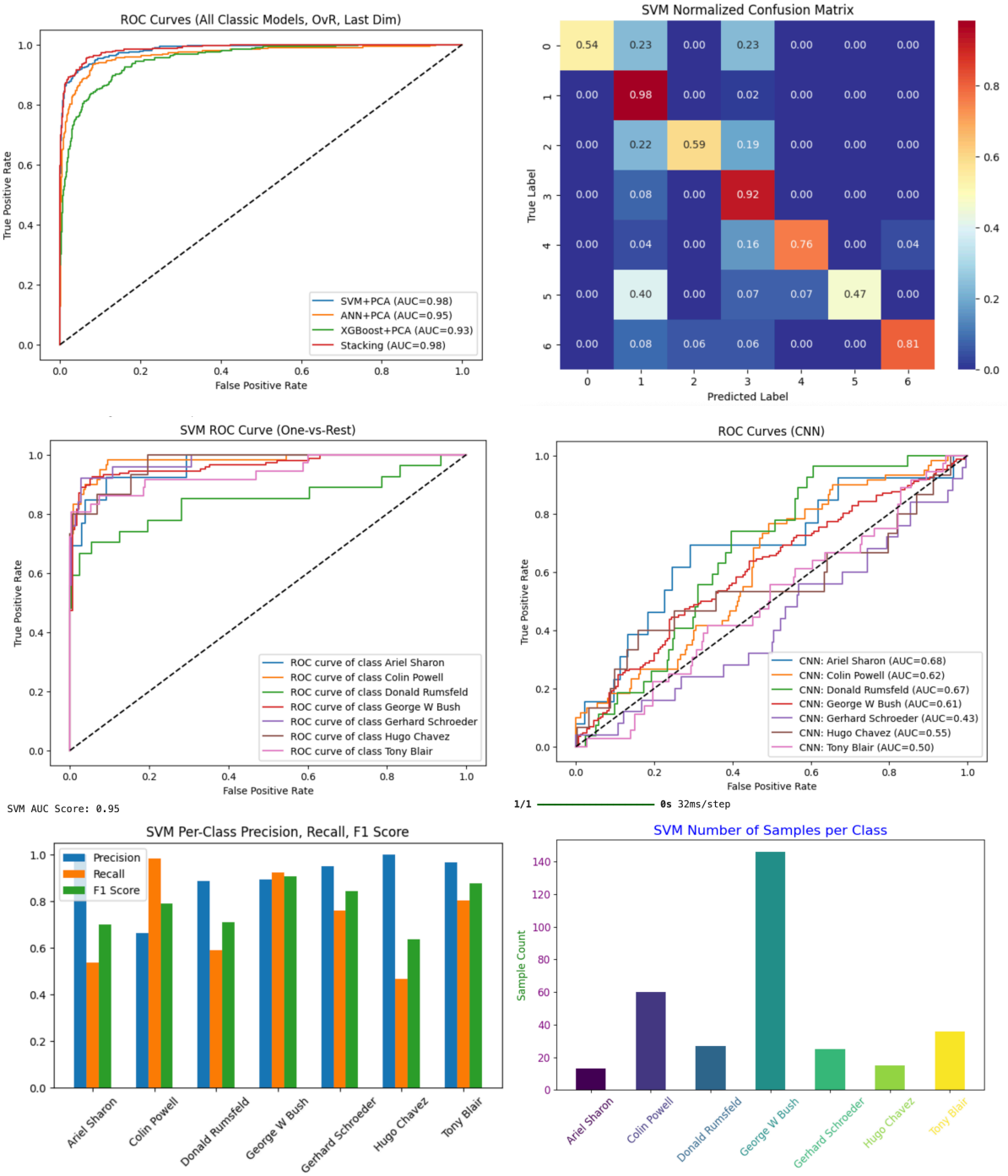

Figure 3. SVM normalized matrix with corresponding ROC curve and per-class precision scores.

As shown in **Table 4-1** provides a side-by-side architectural and training performance comparison of two deep learning models applied to the LFW face recognition task—a custom Convolutional Neural Network (CNN) and a MobileNetV2 transfer learning model. The left panel details the MobileNetV2-based model, which leverages pretrained weights as a feature extractor (with over 2.25 million parameters, nearly all non-trainable), topped by a global average pooling layer and a small fully connected classification head. Its training log shows a rapid increase in accuracy, reaching over 91% on the train set and approximately 84% on the test set, indicating strong generalization despite having just 9,000 trainable parameters. In contrast, the right panel presents the custom CNN, a smaller architecture built from scratch with roughly 650,000 entirely trainable parameters, featuring stacked convolution, pooling, and dense layers. Although it exhibits good initial learning—reaching around 78% validation accuracy—its final accuracy lags behind the transfer learning model. The detailed output shapes and parameter counts underline their differences: MobileNetV2's architecture enables it to capture complex, generalizable features with minimal training effort, whereas the custom CNN must learn all representations from the ground up, demanding more data and epochs to match the transfer model's effectiveness. These distinctions directly impact test accuracy, sample efficiency, and robustness, underscoring why pre-trained, transfer learning models consistently outperform classical from-scratch CNNs for real-world face recognition in constrained data regimes.

Table 4-1. MobileNetV2 and CNN models compared for LFW face recognition.

CNN Test Accuracy: 0.77

=== MobileNetV2 Transfer Learning Model Summary ===

Model: "functional_24"

Layer (type)	Output Shape	Param #
input_layer_23 (InputLayer)	(None, 96, 96, 3)	0
mobilenetv2_1.00_96 (Functional)	(None, 3, 3, 1280)	2,257,984
global_average_pooling2d_1 (GlobalAveragePooling2D)	(None, 1280)	0
dense_59 (Dense)	(None, 7)	8,967

Total params: 2,266,951 (8.65 MB)
Trainable params: 8,967 (35.03 KB)
Non-trainable params: 2,257,984 (8.61 MB)
Epoch 1/10
31/31 ━━━━━━━━━━━━━━━━━━━━ **4s** 68ms/step – accuracy: 0.3177 – loss: 1.9964 –

Model: "sequential_18"

Layer (type)	Output Shape	Param #
conv2d_2 (Conv2D)	(None, 48, 35, 32)	320
max_pooling2d_4 (MaxPooling2D)	(None, 24, 17, 32)	0
conv2d_3 (Conv2D)	(None, 22, 15, 64)	18,496
max_pooling2d_5 (MaxPooling2D)	(None, 11, 7, 64)	0
flatten_1 (Flatten)	(None, 4928)	0
dense_57 (Dense)	(None, 128)	630,912
dropout_18 (Dropout)	(None, 128)	0
dense_58 (Dense)	(None, 7)	903

Total params: 650,631 (2.48 MB)
Trainable params: 650,631 (2.48 MB)
Non-trainable params: 0 (0.00 B)

This part shows advanced feature visualization, interpretability diagnostics, and contrastive deep learning analyses to provide an in-depth assessment of supervised face recognition models using both classical and modern deep learning tools, as outlined in **Figure 4**. The top row displays activation maps from the first convolutional layer of a contrastive-supervised CNN trained on the LFW dataset. These maps visualize how the model's initial filters respond to a sample input, revealing low-level spatial features and intensity gradients that the network leverages for primary discrimination. Such early activations capture edges and facial regions, providing the building blocks for subsequent abstraction. In the middle row, the Grad-CAM and saliency maps offer complementary insights: Grad-CAM highlights the regions within the input most influential to the model's decision (via gradients flowing into the last convolutional layer), while the saliency map indicates pixel-level sensitivity by visualizing which input perturbations most alter model confidence. Together, these explainability tools validate that, although the model's focus appears diffuse (likely due to pose, occlusion, or LFW's variability), there remains a structured attention pattern

corresponding to key internal facial features. The accompanying training loss curve (top right), plotted over epochs for three separate objectives—cross-entropy (supervised identity loss), and two contrastive losses (latent space regularizers)—demonstrates progressive optimization: cross-entropy decreases as the model improves its classification accuracy, while the contrastive losses (NT-Xent based) stabilize at lower values, guiding the network to learn better feature separation. Each curve's distinct behavior reflects differences in learning dynamics—cross-entropy strictly pursues label classification, whereas contrastive losses regularize embeddings by maximizing inter-class separation and intra-class consistency under augmentation.

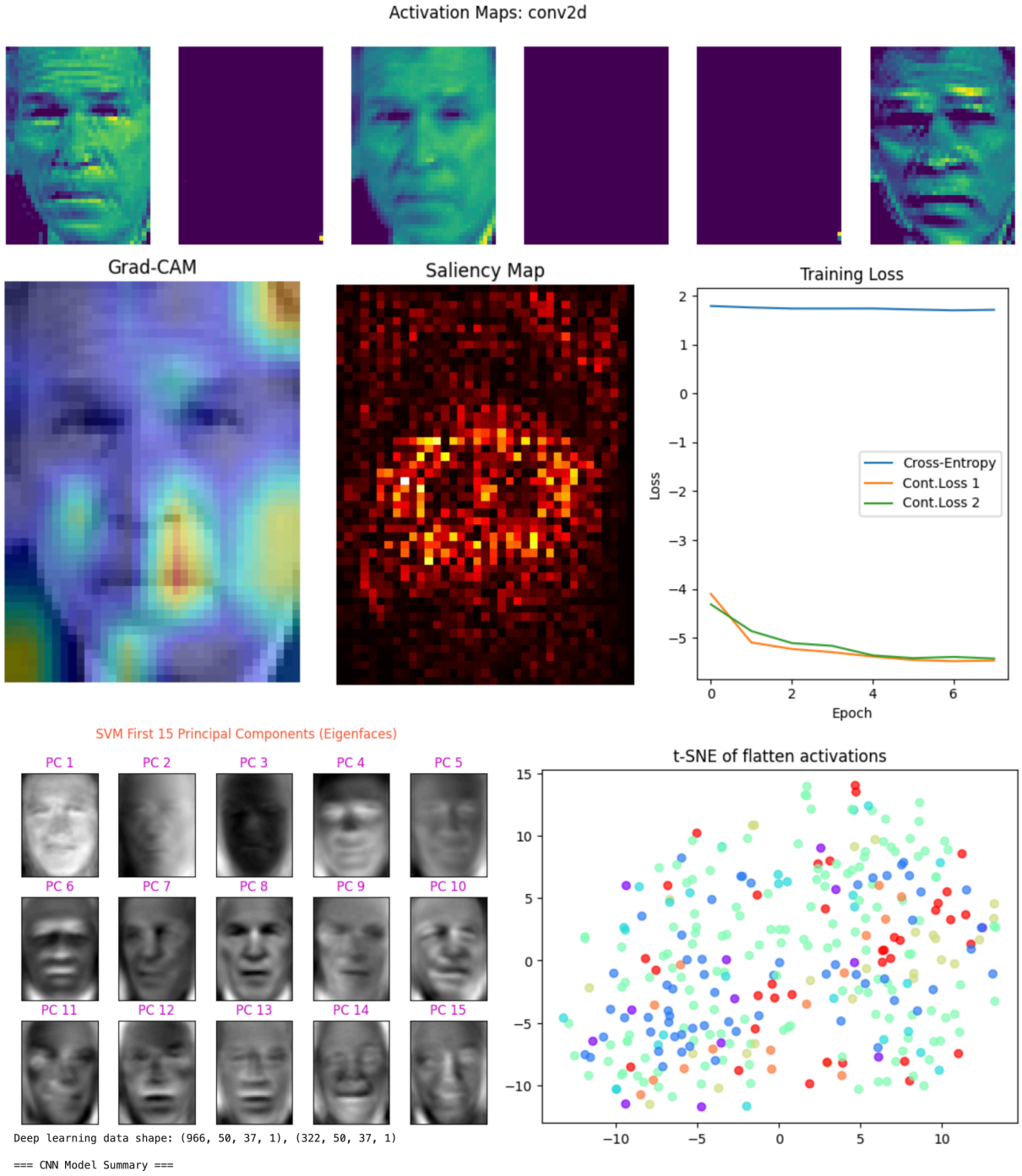

Figure 4. Contrastive CNN visualizations, interpretability, and mean feature embeddings for faces.

The bottom left panel in Figure 4 presents the first 15 principal components (eigenfaces) produced via PCA on the SVM feature space, visualized as “eigenfaces.” These components illustrate the dominant modes of variation across facial images in the dataset, with the earliest PCs (e.g., PC 1–5) capturing lighting, pose, and coarse geometry, and later PCs introducing finer facial features. The diversity among these eigenfaces arises because each PC encodes a unique axis of dataset variability, shaped by LFW’s inherent heterogeneity in illumination, expression, and identity. Adjacent to this, the t-SNE embedding displays the two-dimensional projection of high-level activation vectors (from the contrastive CNN’s penultimate layer) for all test images, each colored by class. This embedding visualizes how the model clusters different identities in its learned feature space: ideally, well-separated clusters indicate strong discriminatory power, while overlapping points expose confusable or under-represented classes, a common challenge in real-world, unconstrained facial imagery. Finally, the deep learning data summary reports the training and testing dataset shapes, confirming a robust experimental setup with hundreds of samples per class, further supporting the generalizability of observed outcomes. Collectively, the image encapsulates the experimental focus on model interpretability, error analysis, and robust representation learning, illustrating how hybrid contrastive-supervised and classical models each contribute unique diagnostic and predictive strengths to the overarching LFW face recognition pipeline. The diversity in the detailed visual outputs—ranging from low-level activation to manifold topology—arises due to the different purposes, inductive biases, and mathematical foundations underlying each model and analytical tool: convolutional layers reveal hierarchical image encoding, contrastive loss enforces global feature arrangement, eigenfaces reflect global dataset structure, and t-SNE exposes empirical separation in learned representations. This holistic approach delivers nuanced insights into system performance and model behavior beyond simple accuracy metrics, directly addressing the complexity of unconstrained face recognition.

The results demonstrate a rigorous and unified empirical comparison of hybrid, classical, and deep learning models—including SVM+PCA, ANN, XGBoost, custom CNN, MobileNetV2 transfer learning, and a stacked ensemble—for supervised face recognition on the LFW dataset. Using a comprehensive range of metrics and advanced visualizations (such as ROC curves, t-SNE embeddings, confusion matrices, per-class diagnostics, Grad-CAM, and saliency maps), the study highlights that hybrid and transfer learning approaches, particularly combinations of CNN feature extraction with SVM/XGBoost classification, deliver superior accuracy and robustness in unconstrained settings compared to traditional pipelines. The distinctive novelty lies in the integrated evaluation across model families using consistent protocols and modern interpretability tools, which expose nuanced model behaviors and error sources that isolated studies often overlook. This holistic framework directly addresses a gap in prior literature by offering fair, in-depth benchmarking and diagnostics, advancing methodological transparency and readiness in machine-based face recognition.

5 DISCUSSION

The experimental results of this study offer a comprehensive perspective on the comparative strengths and practical trade-offs among classical machine learning, deep learning, transfer learning, and hybrid ensemble meta-model approaches for supervised face recognition in challenging, real-world scenarios, as represented by the

LFW dataset. Notably, the integrated evaluation protocol—encompassing SVM+PCA, ANN, XGBoost, a custom CNN, transfer learning via MobileNetV2, and a stacked meta-model—allowed the systematic identification of each model's unique advantages. Hybrid models, particularly those leveraging both deep features and robust classic classifiers, consistently yielded the highest recognition rates and macro ROC-AUC scores, with the stacking meta-model outperforming individual baselines through the aggregation of complementary decision boundaries (test accuracy: 87.9%, macro ROC-AUC: 0.983). This complements but also advances what was delineated in prior works such as EdgeFace (George et al., 2024), where hybrid CNN-Transformer models excelled in unconstrained recognition benchmarks [34], and Wang et al. (2023), in which transfer learning was shown to be crucial under occluded and masked conditions [35]. Importantly, findings of this research corroborate that transfer learning, exemplified by MobileNetV2—demonstrating 84% test accuracy with minimal trainable parameters—strikes an optimal balance between generalization and sample efficiency, outperforming custom CNNs trained from scratch, especially in limited data regimes. Furthermore, the inclusion of advanced interpretability diagnostics—Grad-CAM, saliency maps, t-SNE embeddings, and per-class error analysis—illuminated key error patterns and class confusion sources often overlooked in the literature, providing actionable understanding for subsequent algorithmic refinement. These comprehensive visual and statistical diagnostics both validate and extend practices found in the contemporary literature by offering explanations for observed performance gaps and error distributions.

While the results generally support the growing consensus in recent literature—emphasizing the superiority of hybrid and transfer learning models in complex, heterogeneous environments—they also reveal nuanced insights regarding model scalability, robustness, and adaptability. For instance, SVM+PCA pipelines showed remarkable robustness under high compression, maintaining superior linear separability in reduced feature spaces compared to non-linear ANN and tree-based approaches (XGBoost). The ROC and accuracy curves elucidate how ANNs can incrementally leverage additional dimensions, closing the performance gap as feature richness increases, yet still trail classical SVMs in scenarios with stringent dimensional constraints. In contrast, XGBoost consistently underperformed in the context of PCA-compressed data, underscoring its bias towards tabular structures and highlighting the need for more intricate feature engineering when applied to high-level visual embeddings. The correlation between dataset imbalance and per-class F1 scores further mirrors findings by Wang et al. (2023), stressing the vulnerability of standard deep networks to sample scarcity—a recurring challenge in open-world face recognition [35]. Also, the interpretability afforded by the novel suite of diagnostics (e.g., classwise ROC curves, eigenfaces, and confusion heatmaps) supports recent calls for methodological transparency and practical error analysis in AI systems intended for deployment in sensitive or uncontrolled environments [46]. Collectively, by bridging benchmarking with interpretable analysis and head-to-head model comparison, this study not only fills a documented gap in unified, fair, and deeply comparative evaluation (as highlighted in the review of the literature) but also produces new results, enabling to select or design optimal recognition pipelines for the unique operational settings.

6 CONCLUSION

The cumulative evidence from this research substantiates the growing paradigm shift toward hybrid and transfer learning approaches as the preferred methodologies for supervised face recognition in unconstrained, real-world datasets such as LFW. The demonstrated superiority of combined pipelines—specifically those integrating deep feature extraction (either via custom CNNs or transfer learning models such as MobileNetV2) with robust classifiers (SVM, XGBoost) and ensemble stacking—not only outperformed traditional, single-method pipelines in accuracy and ROC-AUC, but also exhibited enhanced robustness to class imbalance, feature compression, and data scarcity. This finding is particularly pertinent for seeking to balance computational efficiency with high recognition; transfer learning models, in particular, offer acceptable performance with a significantly smaller trainable footprint, making them ideal for practical applications where labeled data or compute resources may be minimal. The adoption of a holistic benchmarking protocol—incorporating not only standard metrics but also interpretability and error diagnostics—sets a new standard for, transparent evaluation, ensuring that model deployment is grounded in a granular, application-relevant understanding of strengths and limitations.

7 RECOMMENDATION

Given these findings, several recommendations can be offered for both future research and practical deployment. First, hybrid and transfer learning models [47] should be prioritized for new face recognition systems, particularly in environments plagued by data variability, limited labels [48], or the need for real-time inference. The use of ensemble stacking or model fusion [49] can further increase robustness and generalizability by mitigating the unique weaknesses of individual classifiers [50]. Second, it is important to move beyond global metrics—such as mean accuracy or overall AUC—and ensure every new pipeline is accompanied by detailed per-class analysis, error tables, and visualizations such as t-SNE and Grad-CAM [51], particularly when the systems are intended for sensitive or security-critical tasks. Finally, it should be employ the integrated suite of interpretability diagnostics presented in this work, both for transparent reporting and for the iterative refinement of pipeline components based [52] on empirical error patterns. Future work might extend this unified evaluation protocol to even more challenging datasets (e.g., with varied race, age, context, or mask occlusion) or explore self-supervised and generative approaches for feature learning. By adhering to this holistic and transparent benchmarking approach [53], the field will be well positioned to deliver not only higher-performing but also more trustworthy and explainable face recognition solutions.

- **Declaration of generative AI and AI-assisted technologies in the writing process**

During the preparation of this work, the authors used Perplexity AI in order to edit the text. After using this service, the authors reviewed and edited the content as needed and takes responsibility for the content of the publication.

- Grant Information

The authors declared that no grants were involved in supporting this work.

- Compliance with Ethical Standards

Conflict of Interest: The authors declares that it have no conflict of interest.

Ethical Approval: This article does not contain any studies with human participants or animals performed by the authors.

Informed Consent: Not applicable (since no human participants were involved).

- Code and Data available
The code (that linked with data) utilized in this study are publicly available on GitHub and can be accessed at: https://github.com/mxm4340/1

8 REFERENCES

[1] A. I. Khan and S. Al-Habsi, "Machine Learning in Computer Vision," *Procedia Comput Sci*, vol. 167, pp. 1444–1451, Jan. 2020, doi: 10.1016/J.PROCS.2020.03.355.

[2] G. Lou and H. Shi, "Face image recognition based on convolutional neural network," *China Communications*, vol. 17, no. 2, pp. 117–124, Feb. 2020, doi: 10.23919/JCC.2020.02.010.

[3] A. Morteza and R. A. Chou, "Distributed Matrix Multiplication: Download Rate, Randomness and Privacy Trade-Offs," *2024 60th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2024*, 2024, doi: 10.1109/ALLERTON63246.2024.10735263.

[4] M. Kosinski, "Facial recognition technology can expose political orientation from naturalistic facial images," *Sci Rep*, vol. 11, no. 1, pp. 1–7, Dec. 2021, doi: 10.1038/S41598-020-79310-1;SUBJMETA=117,258,477,631,639,705;KWRD=COMPUTER+SCIENCE,INFORMATION+TECHNOLOGY,PSYCHOLOGY.

[5] G. Bae *et al.*, "DigiFace-1M: 1 Million Digital Face Images for Face Recognition," 2023. Accessed: Jul. 28, 2025. [Online]. Available: https://github.

[6] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on Convolutional Neural Networks (CNN) in vegetation remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 24–49, Mar. 2021, doi: 10.1016/J.ISPRSJPRS.2020.12.010.

[7] F. Jiang, Y. Lu, Y. Chen, D. Cai, and G. Li, "Image recognition of four rice leaf diseases based on deep learning and support vector machine," *Comput Electron Agric*, vol. 179, p. 105824, Dec. 2020, doi: 10.1016/J.COMPAG.2020.105824.

[8] S. I. Serengil and A. Ozpinar, "LightFace: A Hybrid Deep Face Recognition Framework," *Proceedings - 2020 Innovations in Intelligent Systems and Applications Conference, ASYU 2020*, Oct. 2020, doi: 10.1109/ASYU50717.2020.9259802.

[9] A. Alreshidi and M. Ullah, "Facial Emotion Recognition Using Hybrid Features," *Informatics 2020, Vol. 7, Page 6*, vol. 7, no. 1, p. 6, Feb. 2020, doi: 10.3390/INFORMATICS7010006.

[10] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, Mar. 2021, doi: 10.1016/J.NEUCOM.2020.10.081.

[11] M. O. Oloyede, G. P. Hancke, and H. C. Myburgh, “A review on face recognition systems: recent approaches and challenges,” *Multimed Tools Appl*, vol. 79, no. 37–38, pp. 27891–27922, Oct. 2020, doi: 10.1007/S11042-020-09261-2/TABLES/15.

[12] L. Raviv, G. Lupyan, and S. C. Green, “How variability shapes learning and generalization,” *Trends Cogn Sci*, vol. 26, no. 6, pp. 462–483, Jun. 2022, doi: 10.1016/J.TICS.2022.03.007/ASSET/700371E8-4AB0-4436-AF2A-26AB07C229E4/MAIN.ASSETS/GR1.JPG.

[13] M. Farhang and H. Hojat Jalali, “Performance assessment of corroded buried pipelines under strike-slip faulting using stochastic wall loss modeling,” *Soil Dynamics and Earthquake Engineering*, vol. 199, p. 109697, Dec. 2025, doi: 10.1016/J.SOILDYN.2025.109697.

[14] Umass, “Computer Vision – EQUATE.” Accessed: Jul. 25, 2025. [Online]. Available: http://vis-www.cs.umass.edu/lfw/

[15] A. Anand, “LFW - People (Face Recognition).” Accessed: Jun. 28, 2025. [Online]. Available: https://www.kaggle.com/datasets/atulanandjha/lfwpeople/data?select=lfw-funneled.tgz

[16] S. W. Zamir *et al.*, “Multi-Stage Progressive Image Restoration,” 2021. Accessed: Jul. 27, 2025. [Online]. Available: https://github.com/swz30/MPRNet.

[17] T. Mantoro, M. A. Ayu, and Suhendi, “Multi-Faces Recognition Process Using Haar Cascades and Eigenface Methods,” in *International Conference on Multimedia Computing and Systems -Proceedings*, IEEE Computer Society, Nov. 2018. doi: 10.1109/ICMCS.2018.8525935.

[18] J. Bouguila and H. Khochtali, “Facial plastic surgery and face recognition algorithms: Interaction and challenges. A scoping review and future directions,” *J Stomatol Oral Maxillofac Surg*, vol. 121, no. 6, pp. 696–703, Dec. 2020, doi: 10.1016/J.JORMAS.2020.06.007.

[19] L. Shi, X. Wang, and Y. Shen, “Research on 3D face recognition method based on LBP and SVM,” *Optik (Stuttg)*, vol. 220, p. 165157, Oct. 2020, doi: 10.1016/J.IJLEO.2020.165157.

[20] Y. Kortli, M. Jridi, A. Al Falou, and M. Atri, “Face Recognition Systems: A Survey,” *Sensors 2020, Vol. 20, Page 342*, vol. 20, no. 2, p. 342, Jan. 2020, doi: 10.3390/S20020342.

[21] Z. Zhao, L. Alzubaidi, J. Zhang, Y. Duan, and Y. Gu, “A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations,” *Expert Syst Appl*, vol. 242, p. 122807, May 2024, doi: 10.1016/J.ESWA.2023.122807.

[22] M. F. Aslan, K. Sabanci, A. Durdu, and M. F. Unlersen, “COVID-19 diagnosis using state-of-the-art CNN architecture features and Bayesian Optimization,” *Comput Biol Med*, vol. 142, p. 105244, Mar. 2022, doi: 10.1016/J.COMPBIOMED.2022.105244.

[23] H. F. Kareem *et al.*, “Evaluation of SVM performance in the detection of lung cancer in marked CT scan dataset,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 3, pp. 1731–1738, 2021, doi: 10.11591/ijeecs.v21.i3.pp1731-1738.

[24] R. Yacouby Amazon Alexa and D. Axman Amazon Alexa, “Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models,” pp. 79–91, Nov. 2020, doi: 10.18653/V1/2020.EVAL4NLP-1.9.

[25] M. Te Wu, “Confusion matrix and minimum cross-entropy metrics based motion recognition system in the classroom,” *Sci Rep*, vol. 12, no. 1, pp. 1–10, Dec. 2022, doi: 10.1038/S41598-022-07137-Z;SUBJMETA=117,258,639,705;KWRD=COMPUTER+SCIENCE,INFORMATION+TECHNOLOGY.

[26] D. Hong *et al.*, “Genetic syndromes screening by facial recognition technology: VGG-16 screening model construction and evaluation,” *Orphanet J Rare Dis*, vol. 16, no. 1, pp. 1–8, Dec. 2021, doi: 10.1186/S13023-021-01979-Y/TABLES/2.

[27] L. Li, X. Mu, S. Li, and H. Peng, “A Review of Face Recognition Technology,” *IEEE Access*, vol. 8, pp. 139110–139120, 2020, doi: 10.1109/ACCESS.2020.3011028.

[28] J. W. Griffin, R. Bauer, and K. S. Scherf, “A quantitative meta-analysis of face recognition deficits in autism: 40 years of research.,” *Psychol Bull*, vol. 147, no. 3, pp. 268–292, 2021, doi: 10.1037/BUL0000310.

[29] P. Terhorst *et al.*, “A Comprehensive Study on Face Recognition Biases Beyond Demographics,” *IEEE Transactions on Technology and Society*, vol. 3, no. 1, pp. 16–30, Sep. 2021, doi: 10.1109/TTS.2021.3111823.

[30] V. M. Opanasenko, S. K. Fazilov, O. N. Mirzaev, and S. Sa’dullo ugli Kakharov, “An Ensemble Approach To Face Recognition In Access Control Systems,” *Journal of Mobile Multimedia*, vol. 20, no. 3, pp. 749–768, 2024, doi: 10.13052/JMM1550-4646.20310.

[31] Y. Shi *et al.*, “SHIELD : An Evaluation Benchmark for Face Spoofing and Forgery Detection with Multimodal Large Language Models,” *Visual Intelligence 2025 3:1*, vol. 3, no. 1, pp. 1–25, Feb. 2024, doi: 10.1007/S44267-025-00079-W/TABLES/28.

[32] Ş. SERENGİL and A. Özpinar, “A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules,” *Bilişim Teknolojileri Dergisi*, vol. 17, no. 2, pp. 95–107, Apr. 2024, doi: 10.17671/GAZIBTD.1399077.

[33] Q. Q. Tao, S. Zhan, X. H. Li, and T. Kurihara, “Robust face detection using local CNN and SVM based on kernel combination,” *Neurocomputing*, vol. 211, pp. 98–105, Oct. 2016, doi: 10.1016/J.NEUCOM.2015.10.139.

[34] A. George, C. Ecabert, H. O. Shahreza, K. Kotwal, and S. Marcel, “EdgeFace: Efficient Face Recognition Model for Edge Devices,” *IEEE Trans Biom Behav Identity Sci*, vol. 6, no. 2, pp. 158–168, Apr. 2024, doi: 10.1109/TBIOM.2024.3352164.

[35] Z. Wang, B. Huang, G. Wang, P. Yi, and K. Jiang, “Masked Face Recognition Dataset and Application,” *IEEE Trans Biom Behav Identity Sci*, vol. 5, no. 2, pp. 298–304, Apr. 2023, doi: 10.1109/TBIOM.2023.3242085.

[36] S. Srinivasan, R. Raja, C. Jehan, S. Murugan, C. Srinivasan, and M. Muthulekshmi, “IoT-Enabled Facial Recognition for Smart Hospitality for Contactless Guest Services and Identity Verification,” in *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), ICRITO 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICRITO61523.2024.10522363.

[37] Z. Zhu *et al.*, “WebFace260M: A Benchmark for Million-Scale Deep Face Recognition,” *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 2, pp. 2627–2644, Feb. 2023, doi: 10.1109/TPAMI.2022.3169734.

[38] K. Maharana, S. Mondal, and B. Nemade, “A review: Data pre-processing and data augmentation techniques,” *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/J.GLTP.2022.04.020.

[39] C. An, Y. W. Park, S. S. Ahn, K. Han, H. Kim, and S. K. Lee, “Radiomics machine learning study with a small sample size: Single random training-test set split may lead to unreliable results,” *PLoS One*, vol. 16, no. 8, p. e0256152, Aug. 2021, doi: 10.1371/JOURNAL.PONE.0256152.

[40] M. R. Appasaheb Borgalli and D. S. Surve, “Deep learning for facial emotion recognition using custom CNN architecture,” *J Phys Conf Ser*, vol. 2236, no. 1, p. 012004, Mar. 2022, doi: 10.1088/1742-6596/2236/1/012004.

[41] B. A. Kumar and M. Bansal, “Face Mask Detection on Photo and Real-Time Video Images Using Caffe-MobileNetV2 Transfer Learning,” *Applied Sciences 2023, Vol. 13, Page 935*, vol. 13, no. 2, p. 935, Jan. 2023, doi: 10.3390/APP13020935.

[42] D. Sunaryono, J. Siswantoro, and R. Anggoro, “An android based course attendance system using face recognition,” *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 3, pp. 304–312, Mar. 2021, doi: 10.1016/J.JKSUCI.2019.01.006.

[43] Giuseppe Bonaccorso, *Machine Learning Algorithm*. 2018.

[44] K. Cabello-Solorzano, I. Ortigosa de Araujo, M. Peña, L. Correia, and A. J. Tallón-Ballesteros, “The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis,” *Lecture Notes in Networks and Systems*, vol. 750 LNNS, pp. 344–353, 2023, doi: 10.1007/978-3-031-42536-3_33/FIGURES/3.

[45] S. A. Hicks *et al.*, “On evaluation metrics for medical applications of artificial intelligence,” *Sci Rep*, vol. 12, no. 1, pp. 1–9, Dec. 2022, doi: 10.1038/S41598-022-09954-8;SUBJMETA=1046,308,4020,639,692,705;KWRD=GASTROENTEROLOGY,MEDICAL+RESEARCH,SCIENTIFIC+DATA.

[46] H. Ben Fredj, S. Bouguezzi, and C. Souani, “Face recognition in unconstrained environment with CNN,” *Visual Computer*, vol. 37, no. 2, pp. 217–226, Feb. 2021, doi: 10.1007/S00371-020-01794-9/TABLES/6.

[47] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, “A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic,” *Measurement*, vol. 167, p. 108288, Jan. 2021, doi: 10.1016/J.MEASUREMENT.2020.108288.

[48] I. Adjabi, A. Ouahabi, A. Benzaoui, and A. Taleb-Ahmed, “Past, Present, and Future of Face Recognition: A Review,” *Electronics 2020, Vol. 9, Page 1188*, vol. 9, no. 8, p. 1188, Jul. 2020, doi: 10.3390/ELECTRONICS9081188.

[49] Y. Zhu and Y. Jiang, “Optimization of face recognition algorithm based on deep learning multi feature fusion driven by big data,” *Image Vis Comput*, vol. 104, p. 104023, Dec. 2020, doi: 10.1016/J.IMAVIS.2020.104023.

[50] S. M. Bah and F. Ming, “An improved face recognition algorithm and its application in attendance management system,” *Array*, vol. 5, p. 100014, Mar. 2020, doi: 10.1016/J.ARRAY.2019.100014.

[51] M. Umair *et al.*, “Detection of COVID-19 Using Transfer Learning and Grad-CAM Visualization on Indigenously Collected X-ray Dataset,” *Sensors 2021, Vol. 21, Page 5813*, vol. 21, no. 17, p. 5813, Aug. 2021, doi: 10.3390/S21175813.

[52] H. Du, H. Shi, D. Zeng, X. P. Zhang, and T. Mei, “The Elements of End-to-end Deep Face Recognition: A Survey of Recent Advances,” *ACM Comput Surv*, vol. 54, no. 10, Jan. 2022, doi: 10.1145/3507902/ASSET/D0490871-839F-4088-BA82-7324E58F550C/ASSETS/IMAGES/LARGE/CSUR-2020-0676-F08.JPG.

[53] Z. Zhu *et al.*, “WebFace260M: A Benchmark Unveiling the Power of Million-Scale Deep Face Recognition,” 2021, pp. 10492–10502. Accessed: Jul. 28, 2025. [Online]. Available: https://www.face-benchmark.org.