Whitening, Not the Classifier: Revisiting the Eigenfaces-SVM Pipeline on Labeled Faces in the Wild

Ștefan-Daniel Wagner and Eduard-Cristian Popovici

***Abstract*—The eigenface pipeline, principal component analysis (PCA) followed by a classifier, is a standard baseline whose components are usually reported as a bundle. We isolate them on closed-set identification over the seven Labeled Faces in the Wild identities with at least 70 images, running every cell of a whitening-by-classifier grid with paired exact tests and repeated splits. Whitening the projected coordinates moves a nearest-centroid classifier from 58.9% to 84.1% on one pinned split; the gain survives every split seed we tried. The same transform leaves a grid-searched radial basis function support vector machine statistically unchanged at this sample size, 87.2% whitened against 88.4% unwhitened (paired $p = 0.58$). Contrast-limited adaptive histogram equalization, class weighting, color, and whitening for one-nearest-neighbor produce no detectable difference. Whitening is a metric change whose value depends on the classifier downstream. We document a hyperparameter-transfer trap: a bandwidth tuned at one subspace dimension and reused at another makes the classifier appear to collapse. Separately, the smallest admitted identity clears the benchmark's selection threshold by one image, so the class count depends on the dataset snapshot. This is a measurement study: one narrow, demographically unrepresentative benchmark, no new method, and no conclusion about deep face recognition.**

***Index Terms*—face recognition, principal component analysis, whitening, support vector machines, reproducibility**

I. Introduction

The eigenface pipeline projects face images onto their leading principal components and hands the result to a classifier. It is thirty years old, it is taught everywhere, and on the closed-set identification task we study here it still reaches 87.2% accuracy over seven identities. This paper contributes no new method, no new dataset, and no accuracy record. It reports what happens when the pipeline's standard components are switched off one at a time, on a single pinned split, with paired significance tests. We have not found such a separation carried out systematically (Section II).

A textbook implementation carries at least four switches: whitening of the projected coordinates, contrast-limited adaptive histogram equalization (CLAHE) of the inputs, class weighting for the imbalanced classes, and the choice of color or grayscale. Each is defensible, each is usually left at its default, and each is usually reported as part of a bundle whose members are never separated. We separate them, and find that they are not equals. Our claim is that on this benchmark principal component analysis (PCA) whitening is decisive for the nearest-centroid classifier and has no detectable effect on a bandwidth-tuned support vector machine with a radial basis function kernel (RBF-SVM), while every other switch measures at the noise level. Whitening is a metric change, Eq. (3), whose value depends on whether the downstream classifier can already fit its own length scale.

We support three refutable claims. First, whitening moves the nearest-centroid classifier from 58.9% to 84.1% accuracy, an effect so large that it survives every split seed we tried and a paired exact test at $p < 10^{-13}$ (Section V-B). Second, the same transform leaves a grid-searched RBF-SVM statistically unchanged, 87.2% whitened against 88.4% unwhitened with a paired $p = 0.58$; the selected bandwidth shifts in a way consistent with compensating for the change in scale (Sections V-B and VI). Third, CLAHE, class weighting, color, and whitening applied to one-nearest-neighbor all produce differences that fall inside the noise of a 258-image test set (Sections V-B and V-E). A fourth finding is methodological: holding the kernel bandwidth fixed while varying the subspace dimension makes the SVM appear to collapse at $k = 300$, which a re-tuned control identifies as an artifact of hyperparameter transfer rather than of the subspace (Section V-C).

We also report a reproducibility failure in the benchmark's own construction. The smallest admitted identity clears the identity-selection threshold used throughout this literature by one image, so the number of classes depends on which snapshot of the dataset a reader downloads (Section VII).

II. Related Work

Representing faces in a low-dimensional linear subspace goes back to Sirovich and Kirby, who characterized human faces with a small set of basis images [1], and to Turk and Pentland, who turned that representation into the eigenface recognizer [2]. What a recognizer built on that subspace does next is a choice of metric, and the choice has been studied systematically: Moon and Phillips evaluated variations of the PCA pipeline, including the similarity measure, under the FERET protocol [3], and Perlibakas compared distance measures for PCA-based face recognition [4]. Belhumeur, Hespanha, and Kriegman changed the projection itself, replacing PCA with a class-specific discriminant one [5]; we hold the projection fixed and vary only the metric inside it. Whitening enters that literature as one option among many. The pipeline's most-taught form today is the scikit-learn worked example that pairs eigenfaces with an RBF-SVM on exactly the Labeled Faces in the Wild (LFW) subset we use; there, whitening is a hardcoded default of the projection step [6], [7], reported

as part of a bundle and never varied. We measure that bundle one component at a time.

The classifier we place downstream is the kernel support vector machine of Boser, Guyon, and Vapnik [8] in its soft-margin form [9], whose behavior under the radial basis function kernel is governed by a bandwidth and a penalty [10]. The pairing of an eigenface projection with an RBF-SVM is a standard teaching pipeline and a standard baseline. Precisely because it is standard, the interaction between the projection's optional whitening step and the kernel's own free length scale is easy to leave unexamined. This paper examines it.

Face recognition has since moved to learned embeddings. FaceNet learns "a mapping from face images to a compact Euclidean space where distances directly correspond to a measure of face similarity" [11], ArcFace adds an angular margin to the classification loss [12], and the survey by Wang and Deng records how deep learning "reshaped the research landscape of face recognition (FR) since 2014" [13]. Those systems dominate the benchmarks and are the right tool for the problem. This paper does not enter that regime. We work in the classical pipeline because it is small enough to ablate exhaustively on a laptop, and because a component that is applied by default deserves to be measured at least once.

III. Background

Write each image as a vector $\mathbf{x} \in \mathbb{R}^d$ with $d = 1850$, and let $\boldsymbol{\mu}$ and $\mathbf{C}$ be the mean and covariance estimated on the $n = 1030$ training images. Principal component analysis [14]–[16] diagonalizes that covariance,

$$\mathbf{C} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top, \qquad \mathbf{C}\mathbf{u}_j = \lambda_j \mathbf{u}_j, \quad (1)$$

with eigenvalues ordered $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$. Retaining the leading k eigenvectors as the columns of $\mathbf{U}_k$ gives the eigenface basis [1], [2], which sidesteps the sample-density problem that afflicts direct estimation in $\mathbb{R}^d$ [17].

Two projections onto that basis differ only in scale. The plain projection is $\mathbf{U}_k^\top(\mathbf{x} - \boldsymbol{\mu})$, whose j-th coordinate has variance λ_j. The whitened projection divides each coordinate by its standard deviation,

$$\mathbf{z} = \boldsymbol{\Lambda}_k^{-1/2}\mathbf{U}_k^\top(\mathbf{x} - \boldsymbol{\mu}), \qquad \boldsymbol{\Lambda}_k = \operatorname{diag}(\lambda_1, \ldots, \lambda_k), \quad (2)$$

so that every retained direction carries unit variance. Whitening is the experimental factor of this paper, and its consequence is metric rather than representational: the two projections span the same subspace and differ only in how they measure distance inside it. Substituting (2) into a squared Euclidean distance makes that explicit,

$$\|\mathbf{z}_a - \mathbf{z}_b\|^2 = (\mathbf{x}_a - \mathbf{x}_b)^\top \mathbf{U}_k \boldsymbol{\Lambda}_k^{-1} \mathbf{U}_k^\top (\mathbf{x}_a - \mathbf{x}_b), \quad (3)$$

which is the Mahalanobis distance induced by the rank-k covariance $\mathbf{U}_k\boldsymbol{\Lambda}_k\mathbf{U}_k^\top$ through its pseudoinverse, since that covariance is singular in $\mathbb{R}^d$. Euclidean distance in the whitened subspace is therefore Mahalanobis distance in the original one. Unwhitened, the leading components dominate every distance in proportion to their eigenvalues; whitened, all k directions contribute equally. Which behavior helps depends on the classifier.

The classifier we tune is a support vector machine [8], [9] with the radial basis function kernel

$$\kappa(\mathbf{z}, \mathbf{z}') = \exp\!\left(-\gamma\|\mathbf{z} - \mathbf{z}'\|^2\right), \qquad \gamma > 0, \quad (4)$$

trained with a soft margin whose penalty C trades training violations against margin width, and extended to seven classes by the one-against-one construction [10]. Both free parameters act on the same squared distance that (3) rewrites: γ sets the length scale at which two images stop resembling each other, and C sets how hard the fit works to separate them. This is why whitening and the kernel interact at all.

Two hyperparameter-free rules serve as baselines in the same projected spaces. Nearest centroid assigns the class whose training mean is closest; one-nearest-neighbor assigns the class of the single closest training image. Neither has a bandwidth with which to compensate for the geometry it is handed, which is what makes them informative about whitening.

IV. Data and Experimental Setup

We use the LFW dataset [18] as packaged by scikit-learn [7], keeping the identities with at least 70 images. This filter yields 7 identities and 1288 grayscale images, downscaled by a factor of 0.4 to 50×37 pixels (1850 features). The class distribution is heavily skewed: George W. Bush accounts for 530 images and Hugo Chavez for 71, a 7.5:1 ratio between largest and smallest class. LFW's canonical benchmark task is pair verification [18]; we instead study closed-set identification over these seven identities, the configuration of the pipeline under study.

We apply CLAHE [19], clip limit 2.0 on a 4×4 tile grid, to the 8-bit intensities, then scale pixel values to $[0, 1]$. One ablation arm repeats the full protocol with CLAHE disabled. We split the data once, stratified by identity, into 80% training and 20% test sets (1030 and 258 images) with seed 42; Section V additionally repeats the main comparisons across five split seeds.

The pipeline projects images onto $k = 150$ principal components fitted on the training split only, with whitening on or off as the experimental factor; test images are projected with the same fitted basis. Fig. 1 shows the mean face and the leading eigenfaces of this basis. An RBF-SVM with balanced class weights is trained in the projected space. Each SVM cell in the ablation grid is tuned independently by 5-fold cross-validation (CV) on the training split, over a shared grid of $C \in \{10^2, 10^3, 5 \times 10^3\}$ and $\gamma \in \{\text{scale}, 10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}, 10^{-2}\}$, 18 candidates per cell. The shared grid matters: whitening rescales every coordinate to unit variance, so a γ range calibrated for whitened data would handicap the unwhitened arm. PCA is never refitted during hyperparameter selection, and test data enter no fitting step.

Two hyperparameter-free distance baselines run in the same projected spaces: nearest centroid (NC), which assigns the

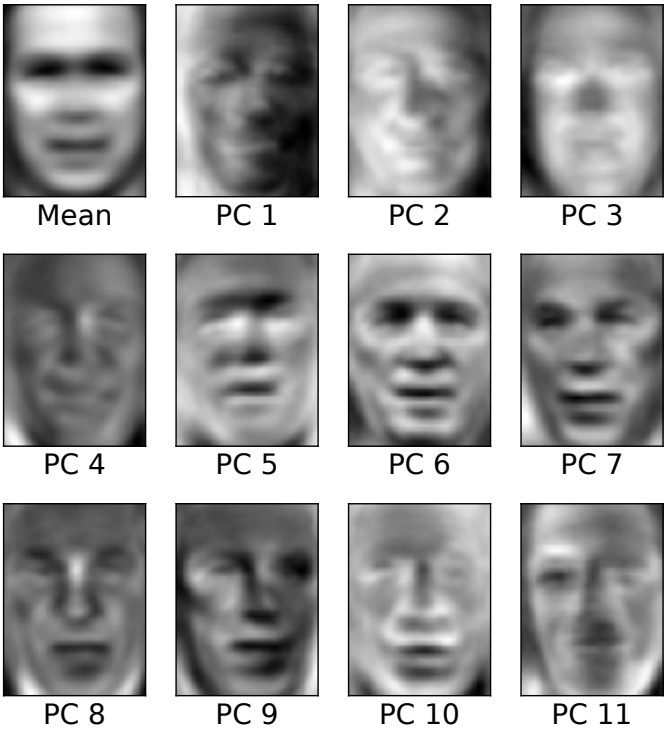

Fig. 1. Mean face and leading eigenfaces of the PCA basis fitted on the CLAHE-processed training split ($k = 150$). The leading directions carry large variances; whitening rescales the projection onto each direction to unit variance, and Section V quantifies what that rescaling does to each classifier.

class of the nearest Euclidean class mean, and one-nearest-neighbor (1-NN) under the same metric. We report test accuracy and macro-averaged F1, the latter because of the class imbalance. Near the observed operating points the binomial standard error of a single accuracy estimate is about 2.1 percentage points ($n = 258$); we use it only as a coarse screen, and rest every comparison on paired predictions and an exact test (Section V).

V. Results and Ablations

A. Headline Result

The tuned whitened pipeline, PCA ($k = 150$, whitened) followed by the RBF-SVM, reaches 87.2% test accuracy and 0.845 macro-F1 at $C = 100$, $\gamma = 0.005$, with 83.3% CV accuracy on the training split. Fig. 2 shows where the remaining errors live: 23 of the 33 misclassifications are predictions into the majority class, George W. Bush, and the smallest class, Hugo Chavez, has the lowest recall (0.57), despite the balanced class weights.

B. Whitening Versus Classifier

Table I evaluates the same two projected spaces, whitened and unwhitened, with three classifiers.

Throughout, we compare paired predictions with McNemar's exact test, two-sided [20], the appropriate choice for classifiers trained once on a single test set [21].

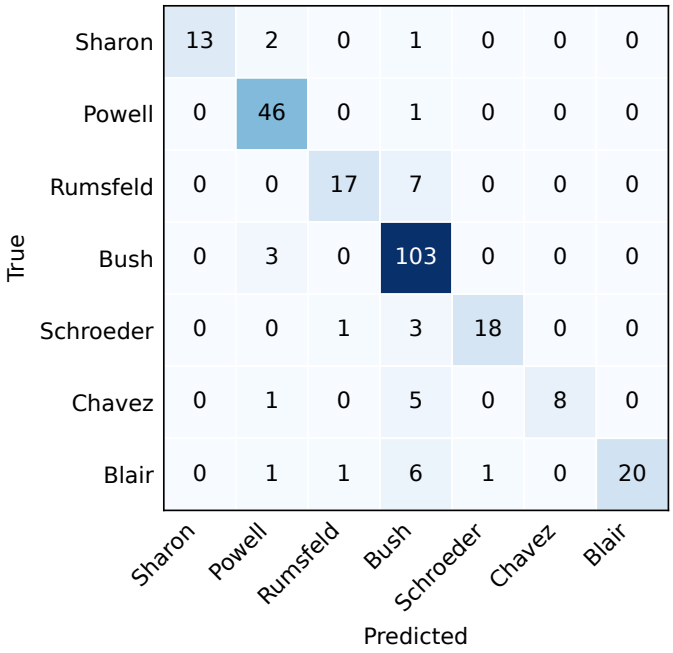

Fig. 2. Confusion matrix of the tuned whitened SVM on the seed-42 test split ($n = 258$). Errors concentrate in the majority class: 23 of 33 misclassifications are predicted as George W. Bush, and the smallest class (Hugo Chavez, 14 test images) has the lowest recall, 0.57.

TABLE I
Test accuracy (%) and macro-F1 on the seed-42 split ($n = 258$), with and without whitening of the $k = 150$ PCA space. Each SVM cell is grid-searched independently; the CLAHE-off rows rerun the whitened cells under an otherwise identical protocol. p: exact two-sided McNemar test, whitened against unwhitened for the classifier rows, CLAHE off against the matching whitened cell for the last two.

	Whitened		Unwhitened		
Classifier	Acc.	F1	Acc.	F1	p
SVM (RBF, tuned)	87.2	0.845	88.4	0.848	0.58
Nearest centroid	84.1	0.808	58.9	0.543	$< 10^{-13}$
1-NN	67.8	0.558	71.7	0.630	0.20
SVM (tuned), CLAHE off	87.2	0.847	-	-	1.0
NC, CLAHE off	85.3	0.819	-	-	0.65

For the grid-searched SVM, whitening produces no detectable difference: 87.2% whitened against 88.4% unwhitened, a gap the paired test does not resolve (8 against 5 discordant errors, $p = 0.58$), and across five split seeds the two arms remain statistically indistinguishable (Section V-D). With only 13 discordant pairs, the exact test rejects nothing short of an 11-to-2 split, a shift of about 3.5 accuracy points, so smaller true differences would go undetected. What changes is the selected bandwidth, which doubles from $\gamma = 0.005$ to $\gamma = 0.01$ on the unwhitened coordinates, consistent with the kernel absorbing the coordinate scaling that whitening would otherwise perform. For the nearest-centroid classifier the picture inverts: whitening lifts accuracy from 58.9% to 84.1%, a gain of 25.2 points on 79 discordant predictions that

split 72 to 7 in its favor ($p < 10^{-13}$). This is the one effect in the study that is large and unambiguous. The whitened centroid then lands 3.1 points behind the tuned SVM, a gap the paired test does not separate (9 against 17 discordant errors, $p = 0.17$). The 1-NN baseline moves 3.9 points the other way, 67.8% whitened against 71.7% unwhitened, but that difference is also within noise (30 against 20 discordant errors, $p = 0.20$), so we read it as no measurable effect rather than as a reversal.

Whitening therefore buys a great deal for the centroid classifier, nothing measurable for its single-neighbor cousin, and nothing detectable for the tuned kernel. Section VI gives our reading of why, stated as interpretation rather than measurement, and takes up why a scalar γ cannot literally stand in for whitening.

The remaining knobs measure as noise. Disabling CLAHE leaves the tuned SVM's accuracy at exactly the same 87.2% with the same selected hyperparameters (6 against 6 discordant predictions, $p = 1.0$), and moves the whitened centroid by 1.2 points (85.3% against 84.1%; 11 against 8 discordant, $p = 0.65$). Switching the SVM's class weights from balanced to uniform at the production operating point reproduces accuracy and both F1 aggregates exactly; per-image predictions were not retained for this cell, so no paired count accompanies it.

C. Effect of Subspace Dimension

Fig. 3 sweeps k from 10 to 300 with nested subspaces: a single PCA with 300 components (full SVD solver) is fitted once on the training split and its leading columns are sliced, so every smaller space is exactly a prefix of the larger one. The centroid classifier needs no hyperparameters, and its curve is clean: accuracy climbs to 84.9% at $k = 100$ and stays within a point of that level through $k = 300$ (at $k = 150$ it returns exactly the 84.1% of Table I, the same basis by construction), tracking the flattening cumulative variance (90.1% explained at $k = 150$). The SVM curve carries a caveat: C and γ were tuned once at $k = 150$ and held fixed, so the sweep isolates the subspace while freezing the bandwidth. It peaks at 87.2% at $k = 75$ and $k = 150$, then collapses to 64.3% at $k = 300$. Re-tuning (C, γ) at $k = 300$ by the same protocol recovers 82.2%, the selected bandwidth falling from $\gamma = 0.005$ to $\gamma = 0.001$: the collapse is the frozen bandwidth, not a property of the subspace. Each added whitened component contributes unit variance, so expected squared distances grow linearly with k, and a γ selected at 150 components is far too large once distances have doubled. The flat NC curve over the same spaces shows the added components themselves cost the centroid rule nothing.

D. Stability Across Splits

We repeat the four main cells over five stratified splits (seeds 42 to 46) with hyperparameters held fixed at the seed-42 selections. Mean accuracy is 86.8% (±1.2) for the whitened SVM, 87.1% (±2.0) for the unwhitened SVM, 83.8% (±2.0) for the whitened centroid, and 60.9% (±3.2) for the unwhitened centroid (sample standard deviations, in points). The two SVM arms differ by 0.2 points on average (86.82 against 87.05 before rounding), well inside their spread, while the whitening gain on the centroid classifier survives every seed (smallest gap 16.7 points). With five seeds the spreads are themselves rough estimates; they are comparable to the binomial standard error of a single estimate (2.1 points at 87%, 3.0 at 61%), consistent with split-to-split noise, and we draw no stronger conclusion from them.

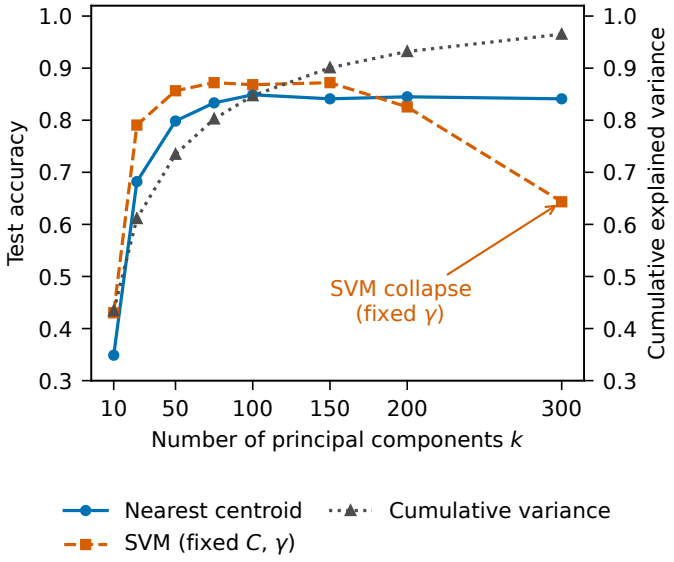

Fig. 3. Accuracy against the number of retained components k (left axis) and cumulative explained variance (right axis), nested subspaces on the seed-42 split, sampled at $k \in \{10, 25, 50, 75, 100, 150, 200, 300\}$. NC is hyperparameter-free and stays within a point of its $k = 100$ level thereafter; the SVM runs at (C, γ) tuned once at $k = 150$, so its collapse at $k = 300$ traces to the frozen bandwidth rather than the subspace (re-tuning there recovers 82.2%), whose added components the flat NC curve absolves.

E. Grayscale Versus RGB

Color runs as a separate arm at resize 0.5 (62×47 pixels) so that RGB and grayscale share an identical protocol; its absolute numbers are therefore not comparable with Table I. Grayscale reaches 82.9% and RGB 84.9%, with the grid selecting the same hyperparameters in both arms ($C = 100$, $\gamma = 0.005$). The two arms' correctness differs on 17 test images, 6 against 11 in RGB's favor, a split the exact test does not resolve ($p = 0.33$). On this benchmark, color is another knob whose effect the test cannot distinguish from zero.

VI. Discussion and Limitations

A. Why Whitening Splits the Classifiers

Equation (3) explains the centroid result. Unwhitened, a distance to a class mean is dominated by the leading components, which on face data encode illumination and pose at least as much as identity; the discriminative directions further down the spectrum contribute in proportion to their small eigenvalues and are effectively outvoted. Whitening equalizes the directions, and the centroid rule recovers 25.2 points. The 1-NN null fits the same reading: a class mean averages away the per-image noise that whitening amplifies along the trailing

directions, while a single neighbor inherits it. The data cannot distinguish that cancellation story from no whitening effect on 1-NN at all ($p = 0.20$). Distance measures in eigenface spaces have been compared systematically before [3], [4], and a whitened distance helping a mean-based rule is the expected direction; what we add is the controlled ablation that isolates whitening and the measured contrast that a bandwidth-tuned RBF-SVM shows no matching gain.

The kernel result deserves more care than the claim that γ simply absorbs whitening. It cannot, exactly. Whitening applies a separate factor $\lambda_j^{-1/2}$ to each axis, whereas γ in (4) rescales all axes by a single number. A bandwidth can match the overall length scale of the two spaces, and the selected values move as one would expect, from $\gamma = 0.005$ whitened to $\gamma = 0.01$ unwhitened, but no scalar reproduces a per-axis reweighting. What the measurements show is that any effect of the reweighting sits below what this test set can detect: tuned on either geometry, the SVM lands within noise of the same accuracy (paired $p = 0.58$). For the tuned RBF-SVM on this problem, the anisotropy that rescues the centroid rule is a degree of freedom it did not need. In practice, whitening buys insensitivity to the choice of classifier, lifting a hyperparameter-free rule to within 3.1 points of the tuned SVM at no detected cost to the SVM.

The k-sweep of Fig. 3 sharpens the point from the other side. Held at a bandwidth tuned for $k = 150$, the SVM decays to 64.3% by $k = 300$, because each added whitened component contributes unit variance and inflates every squared distance until the frozen γ no longer matches the space. The centroid rule, which has no bandwidth to mismatch, stays flat across the same subspaces. The trap is hyperparameter transfer, and it generalizes: any pipeline tuned at one dimensionality and reused at another is exposed to it.

B. Limitations

The evaluation is small and closed. Seven identities with at least 70 images each is the canonical LFW subset for this exercise, but 258 test images spread over seven classes leave some of them thin: Hugo Chavez contributes 14, so one flipped prediction moves his recall by seven points. All seven identities are men, and all are public political figures photographed in the early 2000s, so nothing here speaks to demographic robustness. Accuracy near 87% carries a binomial standard error of about 2.1 points, which is why we test paired predictions instead of comparing point estimates, and why we call the remaining knobs noise-level rather than useless.

We report one dataset, one closed-set protocol, and no deep-learning baseline; that literature operates in a regime this study does not enter. Hyperparameters are selected per cell over 18 candidates by five-fold cross-validation, a coarse grid rather than an exhaustive search, and several cells are scored against the same fixed test split, so the gaps we report should be read with the usual caution about repeated comparisons. Roughly a dozen paired comparisons appear across the grid; the only effect we claim, whitening under the nearest centroid, survives any plausible correction by orders of magnitude, and no other comparison is used to assert an effect. Two further caveats: the unwhitened arm selects $\gamma = 0.01$ and every SVM cell selects $C = 100$, the upper and lower edges of the shared grid, and we did not widen either range to confirm the optima are interior; and the significance tests treat each grid-selected model as fixed, ignoring variability in the selection step itself. The color arm runs at a different resize factor, so it is internally consistent but not comparable with Table I. Our account of why whitening rescues the centroid rule remains an interpretation of Eq. (3) and of the observed pattern; a direct test would decompose the error by spectral band, which we have not run.

VII. Ethics, Data, and Reproducibility

A. Ethics and Data

This study is a methodological ablation, run for research and teaching, on a public benchmark of public figures. That the subjects are heads of state photographed at press events mitigates, but does not erase, the consent problem inherent in a dataset scraped from news photographs. We report identification accuracy on seven people and claim nothing beyond it.

The creators of LFW are explicit about the limits of their benchmark, and we quote them rather than paraphrase. LFW is "a public benchmark for face verification, also known as pair matching," and "[n]o matter what the performance of an algorithm on LFW, it should not be used to conclude that an algorithm is suitable for any commercial purpose" [22]. They further warn that "it is very difficult to extrapolate from performance on verification to performance on 1:N recognition," and that "[m]any groups are not well represented in LFW," with very few children, very few people over 80, and a relatively small proportion of women [22]. Our protocol is closed-set identification over a seven-identity, all-male subset, which is narrower still on every axis they name. Nothing in this paper supports deploying such a pipeline, and the accuracy figures here should not be read as evidence that face identification is safe or accurate in a surveillance setting, where errors fall hardest on the groups the benchmark under-represents.

B. Reproducibility

Every number in this paper comes from a single scripted run on the pinned environment recorded with the artifacts (Python 3.12, scikit-learn 1.4.2, NumPy 2.2.2, OpenCV 4.13), on CPU, seconds per fit. The split is stratified with seed 42, PCA is fitted on the training partition only, and the paired predictions behind every significance test are stored alongside the results. The complete implementation and experiment scripts will be released publicly upon publication.

The pipeline selects identities by a minimum-images-per-person threshold of 70. Hugo Chavez has exactly 71 images, one above the cutoff, so the seven-identity subset we study (1288 images) sits a single image from being a six-identity one, and which of the two a given download yields depends on the exact dataset snapshot. A threshold that the smallest class

clears by a single image makes the class list, and therefore the task, a property of the snapshot. We recommend that work on this benchmark report the resulting class count and image total, not the threshold alone, as we do in Section IV.

The authors used a large language model for language editing and for restructuring drafts. The research questions, the experiments, the analysis, and the conclusions are the authors' own.

VIII. Conclusion

On closed-set identification over seven LFW identities, whitening the PCA coordinates is worth 25.2 accuracy points to a nearest-centroid classifier and has no detectable effect on a grid-searched RBF-SVM, which reaches 87.2% whitened and 88.4% unwhitened (paired $p = 0.58$ on 13 discordant predictions). CLAHE, class weighting, color, and whitening applied to one-nearest-neighbor all measure at the noise level of this test set. Whitening is a change of metric, and its value is set by whether the classifier downstream can already fit its own length scale: give it a rule with no free bandwidth and whitening is the difference between a broken classifier and a competitive one; give it a tuned kernel machine and no change is detectable at this sample size.

None of it transfers to the deep embeddings that own this problem today. It is one pipeline, one benchmark, seven men photographed at press conferences, and 258 test images. What it offers is a measurement of a default the literature usually leaves untested: a preprocessing step bundled into a projection is still a modeling decision.

References

[1] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *Journal of the Optical Society of America A*, vol. 4, no. 3, pp. 519–524, 1987.

[2] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[3] H. Moon and P. J. Phillips, "Computational and performance aspects of PCA-based face-recognition algorithms," *Perception*, vol. 30, no. 3, pp. 303–321, 2001.

[4] V. Perlibakas, "Distance measures for PCA-based face recognition," *Pattern Recognition Letters*, vol. 25, no. 6, pp. 711–724, 2004.

[5] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[6] scikit-learn developers, "Faces recognition example using eigenfaces and SVMs," https://scikit-learn.org/stable/auto_examples/applications/plot_face_recognition.html, Scikit-learn documentation, accessed 2026-07-10.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[8] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. Fifth Annu. Workshop Comput. Learn. Theory (COLT)*, 1992, pp. 144–152.

[9] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[10] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press, 2002.

[11] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 815–823.

[12] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4685–4694.

[13] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, 2021.

[14] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901.

[15] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6–7, pp. 417–441, 498–520, 1933.

[16] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., ser. Springer Series in Statistics. New York: Springer, 2002.

[17] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton University Press, 1961.

[18] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, 2007, University of Massachusetts, Amherst, Technical Report 07-49.

[19] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*. Academic Press, 1994, pp. 474–485.

[20] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[21] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[22] University of Massachusetts, Amherst, "Labeled faces in the wild: Official project page, disclaimer," http://vis-www.cs.umass.edu/lfw/, 2024, Archived: https://web.archive.org/web/20240102064857/http://vis-www.cs.umass.edu/lfw/, accessed 2026-07-10.