

A Machine Learning System for Quantifying Facial Asymmetry in Bell's Palsy across Underrepresented Phenotypes

Ahmed Sultan

ahmed.sultan4759@gmail.com

King Edward Medical University <https://orcid.org/0009-0004-4575-3759>

Ahsan Numan

King Edward Medical University

Research Article

Keywords: Neuroscience, Biomedical Engineering, Computational Biology

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11393331/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

A Machine Learning System for Quantifying Facial Asymmetry in Bell's Palsy across Underrepresented Phenotypes

Abstract

Bell's palsy is the most common cause of acute unilateral facial paralysis, and its severity is graded in clinical practice by ordinal scales that depend on a trained examiner, a resource scarce across much of South Asia. Computer-vision tools that measure facial asymmetry directly from an image could support consistent grading at first contact and across follow-up visits, but the landmark detectors and asymmetry indices these tools rely on have not been characterised for reliability across the phenotypes most relevant to deployment. We engineered a landmark-based facial asymmetry pipeline for Bell's palsy assessment in under-resourced settings and characterised it on the FairFace validation set, a demographically balanced collection of unscreened general-population faces (10,954 images across seven FairFace race categories). The pipeline extracts 478 2D landmark positions per face with MediaPipe FaceMesh, fits a facial midline through 28 midsagittal landmarks by least-squares regression, and computes a global asymmetry index as the mean of eleven regional indices derived from paired anatomical landmarks. Detection succeeded on 90.13% of images, but the failure rate was uneven across groups, ranging from 5.87% for Southeast Asian faces to 12.71% for White faces, which inverts the direction of skew commonly assumed for face-mesh detectors. The pooled median global asymmetry on presumed-healthy faces was 0.187, and group differences were statistically detectable but small (Kruskal-Wallis $H(6) = 35.37$, $p = 3.66 \times 10^{-6}$; $\varepsilon^2 \approx 0.003$). The inner eye corner, the paired landmark closest to the midline, produced the highest asymmetry in every group (0.34 to 0.40), consistent with instability in the index denominator rather than a property of the face. Phenotype-specific flagging thresholds spanned 0.601 to 0.654 around a pooled value of 0.636. The pipeline runs on commodity hardware and behaves consistently across the groups tested, subject to the two characterised failure modes. Its false-positive behaviour on presumed-healthy faces is quantified here; its sensitivity to actual Bell's palsy pathology is not, and awaits a proposed clinical validation study.

Summary

Bell's palsy weakens one side of the face, and doctors judge how severe it is by eye, using scoring scales that depend on how experienced the examiner is. Where trained specialists are scarce, an inexpensive tool that measures facial asymmetry from a photograph could make that judgment more consistent. We built such a tool, keeping it light enough to run on ordinary hardware, and tested how reliably it behaves across faces from seven demographic groups in a large public dataset. It worked on about nine of every ten images and gave consistent measurements on presumed-healthy faces from all seven groups. It has two documented weaknesses: it fails more often on White and Middle Eastern faces than on East and Southeast Asian ones, and its measurement is unstable at landmarks very close to the centre line of the face. The tool is now ready to be tested against real Bell's palsy patients.

A MACHINE LEARNING SYSTEM FOR QUANTIFYING FACIAL ASYMMETRY
IN BELL'S PALSY ACROSS UNDERREPRESENTED PHENOTYPES

Ahmed Sultan

King Edward Medical University

Prof. Ahsan Numan

King Edward Medical University

1 Introduction

Bell's palsy is the most common cause of acute unilateral facial paralysis, with an annual incidence between 20 and 30 cases per 100,000 people (Peitersen 2002). It presents suddenly, with drooping of one corner of the mouth, incomplete closure of one eyelid, and reduced movement of the forehead and cheek on the affected side. The Copenhagen Facial Nerve Study, which followed 2,570 untreated cases, found that facial function returned within three weeks in 85% of patients and within three to five months in the remainder, with normal function recovered in 71%, mild or slight sequelae in 25%, and severe sequelae in 4% (Peitersen 2002). Contracture and synkinesis persisted in 17% and 16% of patients respectively (Peitersen 2002). The recovery window is also the treatment window: a double-blind randomised trial of 496 patients showed that prednisolone started within 72 hours of onset raised the proportion recovering full function at nine months from 81.6% to 94.4% (Sullivan et al. 2007). Grading severity quickly and consistently therefore has direct therapeutic weight, and grading it again at follow-up is how recovery is tracked.

Severity is graded in clinical practice by ordinal scales scored by a trained examiner. The House-Brackmann system, adopted as the reference standard in 1985, assigns a single grade from I (normal) to VI (complete paralysis) from overall observation of the face at rest and in movement (House and Brackmann 1985). The Sunnybrook Facial Grading System refines this into weighted subscores for resting symmetry, voluntary excursion, and synkinesis (Ross et al. 1996). Both require examiner familiarity, and both carry measurable inter-observer variability: in a study in which six otolaryngologists graded the same videotaped patients, agreement on House-Brackmann grade was moderate at best (Coulson et al. 2005). Both instruments also produce an ordinal grade rather than a continuous measurement, so small changes between visits are difficult to resolve, and a patient graded by different clinicians across visits is difficult to compare against himself.

This problem is exacerbated in low-resource settings. Across much of South Asia, first-contact assessment of facial palsy falls to general practitioners and emergency clinicians, and access to fellowship-trained facial-nerve specialists is confined to a small number of tertiary centres (Roxas et al. 2017). A tool that returns a consistent, continuous asymmetry measurement from an ordinary photograph would let a non-specialist grade at first contact, when the treatment window is open, and would let any clinician track recovery against a fixed scale rather than against another examiner's gestalt.

Several computer-vision systems have been built toward this goal. Guarin and colleagues released Emotrics, a semi-automated tool that places facial landmarks on patient photographs and computes distance-based symmetry measurements for facial palsy assessment (Guarin et al. 2018). It has since become the most widely used such tool, valued for a free and usable interface, though it is limited to frontal images and a fixed set of measurements (Guarin et al. 2018). Landmark extraction itself has been transformed by lightweight detectors built for mobile hardware: MediaPipe FaceMesh infers a dense 468-vertex facial mesh from a single image at real-time speed on a mobile GPU (Kartynnik et al. 2019), within the broader MediaPipe framework for on-device perception pipelines

(Lugaresi et al. 2019). These developments make landmark-based asymmetry quantification feasible at the point of care without specialised equipment.

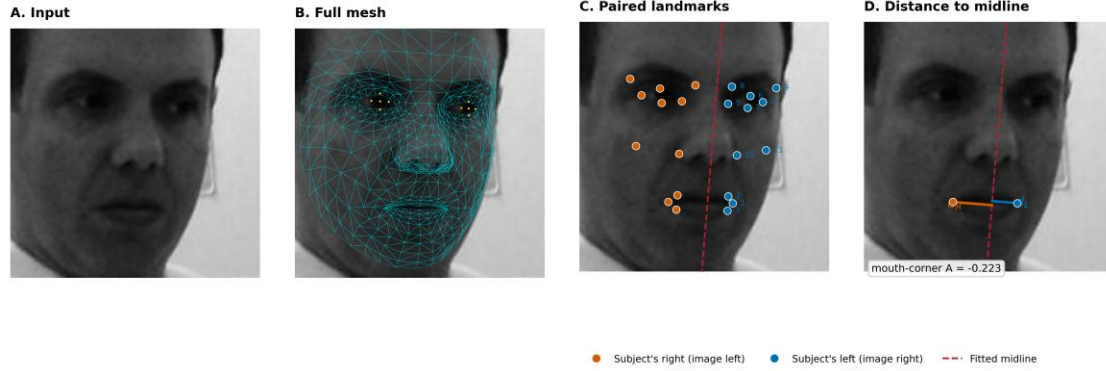
What has received little attention is whether the underlying detectors, and the asymmetry indices computed from them, behave consistently across the phenotypic diversity of the populations that would most benefit from such a tool. FairFace was assembled specifically to counter the strong skew of common face datasets toward Caucasian faces, and it provides seven demographically balanced groups against which detection reliability and index stability can be measured (Kärkkäinen and Joo 2021). Prior facial-paralysis pipelines have not, to our knowledge, been characterised on FairFace or a comparable balanced substrate. Deploying a screening tool into a South Asian clinic without such a characterisation risks silent underperformance on the target phenotype, or thresholds calibrated on one population being carried unchanged onto another.

We report that characterisation here. We engineered a lightweight facial asymmetry pipeline for the specific context of under-resourced South Asian clinical settings, using 478 2D landmarks per face and eleven anatomically motivated paired-landmark indices, and we evaluated its reliability across the seven FairFace race categories on 10,954 validation images. The evaluation quantifies detection failure and its demographic distribution, the per-group distributions of regional and global asymmetry, and a confidence-flagging framework that sets per-group thresholds from the 95th percentile of the presumed-healthy distribution. The pipeline has not been tested on Bell’s palsy patients; its behaviour on presumed-healthy faces is the prerequisite this work supplies.

2 Methods

2.1 System design

The pipeline runs in five stages: image input, landmark extraction, midline fitting, regional asymmetry computation, and global aggregation with optional confidence flagging. Every design choice was driven by the target deployment setting, an under-resourced South Asian clinic where GPU availability, reliable connectivity, and standardised imaging cannot be assumed. We chose a landmark-based approach over an end-to-end classifier because per-region landmark output is interpretable and a clinician can inspect and dispute it. We chose a pre-trained face-mesh detector so no training data or GPU is needed at deployment. We chose 2D geometry so no depth camera is needed. We chose thresholds computed from a presumed-healthy population so a per-face confidence flag can be raised without labelled pathology examples. The complete worked pipeline on a single face is shown in Figure 1.



Landmarking methodology on a real adult FairFace face. (A) Input image, an unmodified crop from the FairFace validation set, used under the licence of the FairFace release (Kärkkäinen and Joo 2021). (B) The 478 landmarks returned by MediaPipe FaceMesh v0.10.30 with `refine_landmarks=True`, `static_image_mode=True`, `max_num_faces=1`, `min_detection_confidence=0.5`. The full mesh (468 face and 10 iris landmarks) is overlaid in cyan, with iris landmarks in yellow. The z coordinate is discarded; all downstream computation uses (x, y) in image space. (C) The eleven paired anatomical landmarks used for asymmetry, plotted on the source image, with the fitted midline as a red dashed line. The midline is a least-squares fit of x as a linear function of y through 28 midsagittal indices (10, 151, 9, 8, 168, 6, 197, 195, 5, 4, 1, 19, 94, 2, 164, 0, 11–18, 200, 199, 175, 152), disjoint from the eleven paired landmarks. Numbered points 1–11 are mouth corner (cheilion), upper lip, lower lip, outer eye corner (exocanthion), inner eye corner (endocanthion), upper eyelid, lower eyelid, inner eyebrow, outer eyebrow, nostril (alare), and cheek. Laterality follows the anatomical convention: “left” denotes the subject’s left, which appears on the image’s right; images are not mirrored. (D) The regional asymmetry index for the mouth-corner pair. Perpendicular distances d_L and d_R to the fitted midline give the signed index $A_r = (d_L - d_R)/(d_L + d_R)$, bounded in $[-1, +1]$; global asymmetry is the mean of $|A_r|$ over the eleven regions.

2.2 Image acquisition

The characterisation used the FairFace validation set of 10,954 images (Kärkkäinen and Joo 2021), a demographically balanced collection of unscreened general-population faces with human-annotated race, gender, and age labels. Images are colour, cropped to include the face and moderate background, and vary in resolution, illumination, and pose. No preprocessing was applied before landmark extraction. In the intended clinical use, patient images would be captured under standardised frontal pose and diffuse illumination; the FairFace substrate relaxes these controls, which makes the characterisation a conservative estimate of pipeline reliability.

2.3 Landmark extraction

We used MediaPipe FaceMesh v0.10.30 with `static_image_mode=True`, `max_num_faces=1`, `refine_landmarks=True`, and `min_detection_confidence=0.5`. With

`refine_landmarks=True`, MediaPipe returns 478 landmark positions per detected face (468 mesh points and 10 iris landmarks) in normalised image coordinates (x, y, z) . The z coordinate is an approximate depth relative to the face plane, produced by a monocular model, and is discarded; all downstream computation is performed in 2D image space using (x, y) only. Landmarks are labelled by the anatomical side of the subject, so a “left” landmark denotes the subject’s left side of the face and appears on the right side of the image. Images are not mirrored at any stage. Where MediaPipe returns no detection, no asymmetry values are computed for that image, and the detection failure is recorded as an outcome of the pipeline rather than absorbed silently. Its demographic distribution is reported in Section 3.1.

2.4 Midline fitting

For each detected face, a facial midline is fit by ordinary least-squares regression through 28 midsagittal landmarks. The MediaPipe indices used are 10, 151, 9, 8, 168, 6, 197, 195, 5, 4, 1, 19, 94, 2, 164, 0, 11 through 18, 200, 199, 175, and 152. These points lie on or immediately adjacent to the anatomical midline, spanning crown to chin, and are listed in the caption of Figure 1. The set is disjoint from the eleven paired landmarks used for asymmetry (Section 2.5), so no landmark contributes both to the definition of the midline and to the deviations measured against it. We fit x as a linear function of y , which handles the near-vertical orientation of the anatomical midline without numerical issues. Fitting is per-image, so head roll and mild off-frontal pose are absorbed into the fit.

2.5 Asymmetry indices

Eleven paired landmarks sample facial regions relevant to Bell’s palsy: mouth corner (cheilion, MediaPipe indices 61 and 291), upper lip (37, 267), lower lip (84, 314), outer eye corner (exocanthion, 33, 263), inner eye corner (endocanthion, 133, 362), upper eyelid (159, 386), lower eyelid (145, 374), inner eyebrow (55, 285), outer eyebrow (46, 276), nostril (alare, 129, 358), and cheek (50, 280). For each pair, let d_L and d_R be the perpendicular Euclidean distances from the subject’s left and right landmark to the fitted midline. The regional asymmetry index is

$$A_r = \frac{d_L - d_R}{d_L + d_R},$$

bounded in $[-1, +1]$ and zero when the paired landmarks are equidistant from the midline. Positive values indicate the subject’s left landmark is farther from the midline than the right. The global asymmetry index is the mean of the absolute regional values,

$$|A|_{\text{global}} = \frac{1}{11} \sum_{r=1}^{11} |A_r|,$$

where taking the absolute value before averaging prevents cancellation between regions deviating in opposite directions. All aggregate analyses in this paper use $|A|_{\text{global}}$.

2.6 Confidence flagging

To support point-of-care use, the pipeline flags an image whose $|A|_{\text{global}}$ exceeds a per-group threshold set at the 95th percentile of that group’s distribution on the presumed-healthy substrate. By construction, this controls the false-positive rate at 5% on presumed-healthy faces within each group. It carries no information about sensitivity to actual Bell’s palsy pathology, and no such information can be estimated from FairFace, which contains no clinically labelled patient faces. Sensitivity is a property that only clinical validation on labelled patient data can establish.

2.7 Dataset

FairFace annotates each face with one of seven race categories: East Asian, Southeast Asian, Indian, Black, Latino Hispanic, White, and Middle Eastern, with group sizes balanced by design (Kärkkäinen and Joo 2021). We use these labels as demographic group indicators only. They are annotator-assigned social categories, not measured phenotypes, and we do not treat them as such. All faces are presumed-healthy: they were not clinically screened for facial paralysis or any other condition, and the assumption that they represent typical unaffected faces is used only as a working null against which pipeline stability is measured.

2.8 Statistical analysis

Detection failure rate per group is the proportion of images for which MediaPipe returned no detection, with 95% bootstrap confidence intervals (2,000 resamples, percentile method). For the distribution of $|A|_{\text{global}}$ across the seven groups, we tested the null of equal distributions with the Kruskal-Wallis rank-sum test and report effect size as ε^2 with a 95% bootstrap confidence interval (2,000 resamples) (Tomczak and Tomczak 2014). Selected pairwise contrasts between the extreme groups were tested with the Mann-Whitney U test; where multiple pairwise comparisons are reported, we apply Holm-Bonferroni correction (Holm 1979). Per-group 95th percentiles were computed empirically and paired with 95% bootstrap confidence intervals (2,000 resamples). Computation was performed in a Google Colab environment using NumPy and SciPy, and figures were rendered with matplotlib.

3 Results

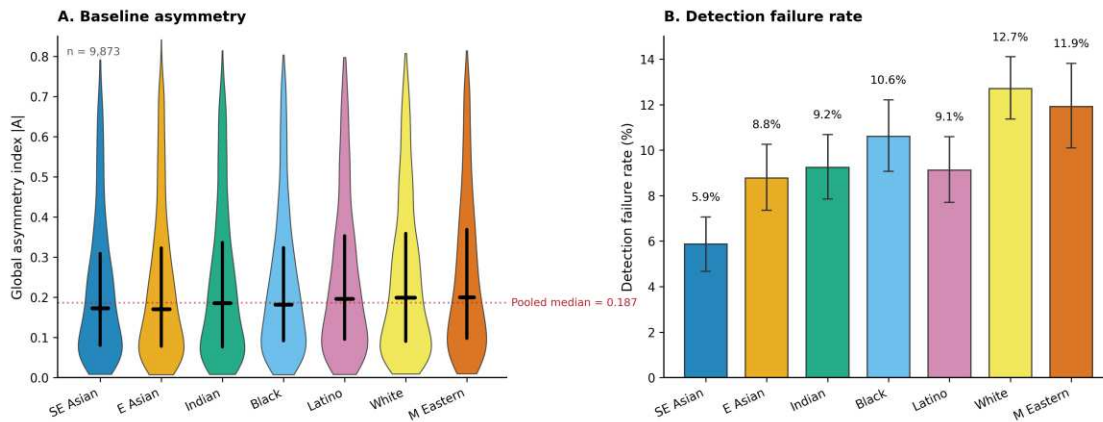
3.1 Detection across groups

Of 10,954 images, MediaPipe FaceMesh returned a detection on 9,873, or 90.13%. The remaining 1,081 images (9.87%) produced no detection and no downstream computation. Failure was not uniform across groups. Southeast Asian faces failed least often at 5.87% and White faces most often at 12.71%, with Middle Eastern at 11.91%, Black at 10.60%, Indian at 9.24%, Latino Hispanic at 9.12%, and East Asian at 8.77% (Figure 2B).

This ordering inverts the direction of skew commonly assumed for face-mesh detectors trained predominantly on Western data, where one might expect lower failure on White faces. We do not read the inversion as evidence that MediaPipe favours Southeast Asian faces. FairFace draws images from group-specific web sources, and several dataset-level confounds plausibly co-vary with the group label at rates sufficient to drive the observation: differences in age distribution, occlusion by beards, headwear, and eyeglasses, and differences in image resolution, compression, and pose across the source pools. Separating these confounds from any true detector skew would require a stratified analysis on a controlled substrate or a purpose-collected clinical dataset. The practical consequence for deployment holds regardless of mechanism: detection reliability must be measured on the target population before thresholds are trusted, and a deployed system should escalate undetected images to human review or repeat capture rather than return no result silently.

3.2 Baseline asymmetry

Across the 9,873 detected faces, the pooled median global asymmetry was 0.187 (Figure 2A). The seven per-group distributions were similar in shape and heavily overlapping. A Kruskal-Wallis test rejected the null of identical distributions, $H(6) = 35.37$, $p = 3.66 \times 10^{-6}$, but the effect size was small: $\varepsilon^2 \approx 0.003$, placing less than half a percent of the variance in ranked global asymmetry with group membership. The largest pairwise contrast, between Southeast Asian and Middle Eastern faces, was 14.0% in relative median terms (Mann-Whitney $U = 643,078$, $p = 8.4 \times 10^{-6}$), and all other group differences were of similar or smaller magnitude. The pipeline’s baseline index is therefore reliable across the groups tested to within a small residual offset, which the confidence framework absorbs into per-group thresholds where the offset is statistically resolvable.



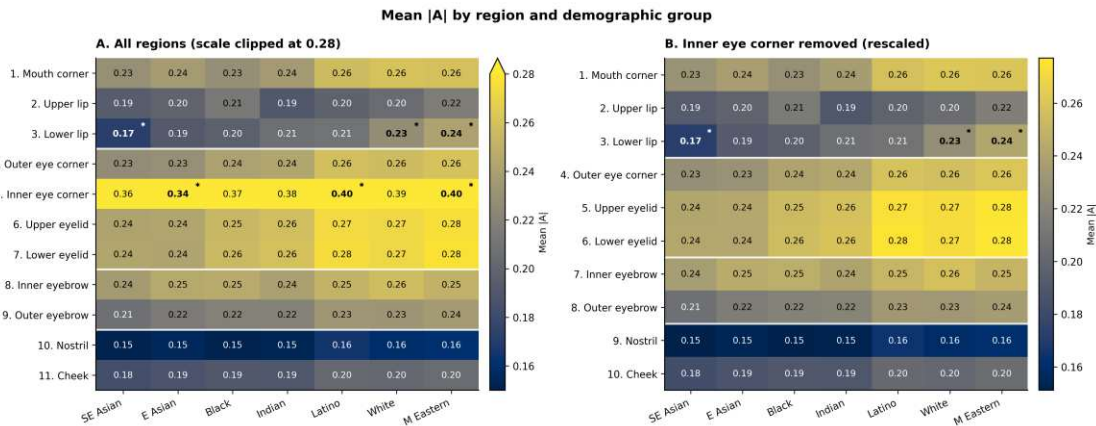
Bias analysis across seven FairFace race categories. (A) Distribution of global asymmetry index $|A|_{\text{global}}$ on detected presumed-healthy faces ($n = 9,873$). Violins show the full distribution per group; the horizontal bar marks the median and the vertical bar the interquartile range. Density is estimated with a Gaussian kernel using default bandwidth (Scott’s rule) in matplotlib. The dotted line marks the pooled median of 0.187. Kruskal-Wallis on the seven distributions gives $H(6) = 35.37$, $p = 3.66 \times 10^{-6}$, with effect size $\varepsilon^2 \approx$

0.003: statistically detectable but small. (B) Detection failure rate per group with 95% bootstrap confidence intervals (2,000 resamples, percentile method). Failure ranges from 5.87% (Southeast Asian) to 12.71% (White), inverting the direction of skew commonly assumed for face-mesh detectors and plausibly reflecting group-level confounds in FairFace (age distribution, occlusion by beards, headwear, and eyeglasses, image resolution, and pose) rather than a direct property of MediaPipe. MediaPipe FaceMesh v0.10.30 was used throughout with `static_image_mode=True`, `max_num_faces=1`, `refine_Landmarks=True`, `min_detection_confidence=0.5`.

3.3 Regional asymmetry

Mean absolute regional asymmetry, tabulated by region and group, is shown as a heatmap in Figure 3. The inner eye corner (endocanthion) is the highest-asymmetry region in every group, with mean values between 0.34 and 0.40 (Figure 3A). The nostril (alare) is the lowest in every group at 0.15 to 0.16, and cheek and upper lip are also consistently low at 0.18 to 0.22.

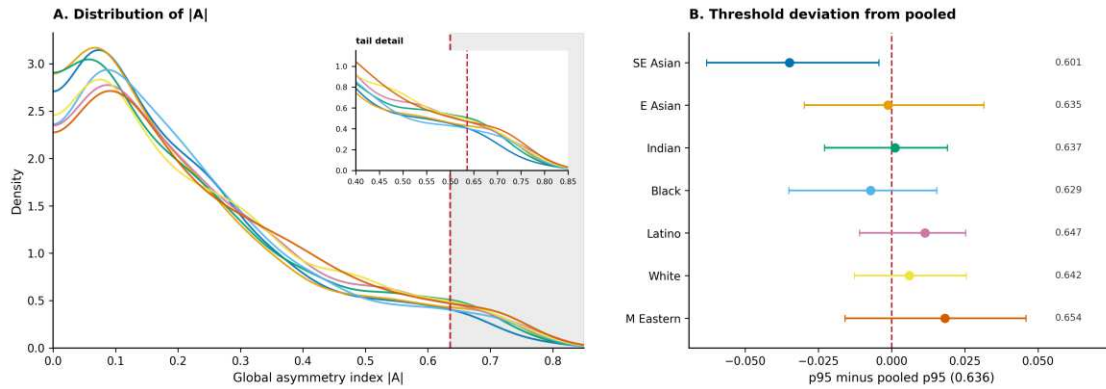
The inner eye corner result is consistent with instability in the index denominator. Because $A_r = (d_L - d_R)/(d_L + d_R)$ scales absolute deviation by the sum of distances to the midline, regions whose paired landmarks sit close to the midline have a small denominator, and any given absolute difference between d_L and d_R , including one produced by small stochastic error in landmark position, yields a large regional index. The inner eye corner is the closest pair to the midline in our set and also the highest-asymmetry region across all groups; the two properties are consistent. When the inner-eye-corner row is removed and the scale rescaled (Figure 3B), the remaining regions occupy a much narrower range and the group-by-region structure at the other landmarks becomes readable. All eleven regions showed statistically significant group differences under Kruskal-Wallis (all $p < 0.05$), with lower lip the most extreme at $p = 3.18 \times 10^{-23}$, but effect sizes at every region were of the same small magnitude as the global-asymmetry effect, so the heatmap is best read as descriptive of anatomical structure rather than as a source of per-region hypothesis tests.



Mean |A| by region and demographic group. (A) Full eleven-by-seven heatmap. Rows are the eleven paired regions; columns are the seven FairFace race categories, sorted left to

right by pooled median $|A|_{\text{global}}$ from lowest to highest. White horizontal lines separate anatomical groupings (mouth, eye, brow, lower face). Cells use a *cividis* scale clipped at 0.28 so the inner-eye-corner row does not wash out the remaining structure. Values shown bold with an asterisk mark the group-wise minimum and maximum of each starred row, established by non-overlap of bootstrap 95% confidence intervals (2,000 resamples, percentile method). Region 1 mouth corner is the cheilion, region 4 outer eye corner the exocanthion, region 5 inner eye corner the endocanthion, and region 10 nostril the alare. (B) The same data with the inner-eye-corner row removed and the scale rescaled. That row is an artefact of denominator instability rather than an anatomical finding: because A_r scales deviation by the sum of distances to the midline, paired landmarks closest to the midline systematically amplify small stochastic errors, and mean $|A|$ across regions tracks the inverse of the mean sum-of-distances denominator (Pearson correlation between mean regional $|A|$ and inverse mean denominator computed across the eleven regions).

3.4 Confidence framework



Confidence-flagging framework across demographic groups. (A) Kernel density estimates of $|A|_{\text{global}}$ per group, colour-matched to Figure 2A. The dashed red line marks the pooled 95th percentile (0.636), and the shaded region beyond it is the flag zone under the pooled rule. Inset: tail detail from 0.40 to 0.85, showing the per-group densities converging in the flag region with no group's tail departing systematically. (B) Forest plot of per-group 95th percentile minus the pooled 95th percentile, with 95% bootstrap confidence intervals (2,000 resamples, percentile method); per-group p95 values are printed at right. Every interval except Southeast Asian overlaps zero. Because thresholds are the 95th percentile of the per-group distribution on presumed-healthy faces, the framework controls the false-positive rate at 5% per group by construction and carries no information about sensitivity to actual Bell's palsy pathology.

Per-group 95th-percentile thresholds ranged from 0.601 for Southeast Asian faces to 0.654 for Middle Eastern faces, a spread of 0.053 in absolute terms and 8.8% relative, around a pooled value of 0.636 (Figure 4B). Bootstrap 95% confidence intervals on the per-group thresholds overlapped the pooled value for every group except Southeast Asian, whose interval sat entirely below it. Pooled and per-group thresholds are therefore statistically indistinguishable for six of the seven groups tested, and only Southeast Asian faces would

be systematically over-flagged if the pooled threshold were applied unchanged. The decision impact of pooled versus per-group thresholds is bounded by the same spread: the maximum deviation of any group’s p95 from the pooled value is 0.035 below (Southeast Asian) and 0.018 above (Middle Eastern), so only faces whose global asymmetry falls within that narrow margin of the threshold would be classified differently under the two rules.

Two properties of this framework need stating plainly. It controls the false-positive rate on presumed-healthy faces at 5% per group, by construction, and nothing more. It carries no information about sensitivity: nothing in FairFace allows estimation of how often the pipeline would flag an actual Bell’s palsy face, because FairFace contains none. The framework is a specificity-controlled filter on presumed-healthy inputs, not a filter with known detection performance on the pathology of interest, and establishing that performance is a primary target of the proposed future study.

4 Discussion

We characterised a landmark-based facial asymmetry pipeline, engineered for low-resource Bell’s palsy assessment, on 10,954 images from the FairFace validation set. Three findings determine how the pipeline should be read before it is taken to clinical validation.

Detection failure is not distributed uniformly across demographic groups, and its direction inverts a common expectation. Face-mesh detectors trained on Western data are widely assumed to underperform on non-Western phenotypes, and the demographic-bias literature built around FairFace is usually read that way. Our data run the other direction: MediaPipe failed on 12.71% of White and 11.91% of Middle Eastern faces but only 5.87% of Southeast Asian faces. The FairFace group subsets are drawn from group-specific web sources and differ in age composition, occlusion by beards, eyeglasses, and headwear, image resolution, and pose, any of which is plausibly sufficient to drive the ordering, so we do not attribute it to a phenotype preference in the detector. The finding matters for deployment regardless of mechanism: detection reliability on the target population cannot be inferred from performance on a Western dataset, and it must be measured directly before flagging rules are trusted.

Group differences in the baseline index are statistically detectable but small. The Kruskal-Wallis p -value of 3.66×10^{-6} reflects a real difference in the ranked distributions, and the effect size of $\varepsilon^2 \approx 0.003$ locates that difference at under half a percent of the variance in ranks, with the largest pairwise contrast at 14.0% relative. A pipeline that produced sharply different baseline asymmetry across groups would need redesign before clinical use; this one does not, subject to per-group thresholding where the residual offset is resolvable.

The regional index is unstable at landmarks close to the midline. The form $A_r = (d_L - d_R)/(d_L + d_R)$ scales absolute deviation by the sum of distances to the midline, so it produces high magnitude wherever that sum is small. The inner eye corner is the closest

pair to the midline in our set and also the highest-asymmetry region in all seven groups, at 0.34 to 0.40, which is what we would expect if high inner-eye-corner values reflect the choice of denominator rather than a genuine asymmetry there. We report this rather than smooth it over because it bears on interpretation: regional asymmetry values are not directly comparable across regions with different sums of distances, and any use of regional asymmetry as a per-region diagnostic feature will need renormalisation. Division by interpupillary distance is a specific candidate, since interpupillary distance is stable within an individual and standardises the scale of deviation across regions.

Prior computer-vision work on facial paralysis has concentrated on maximising agreement with clinician grades on curated clinical datasets. Emotrics and its successors advanced that agreement axis (Guarin et al. 2018). This characterisation contributes the orthogonal axis: whether the pipeline is reliable enough, in detection and in baseline index, for clinical validation to be meaningful on populations underrepresented in existing datasets. Answering that first, and surfacing the two specific failure modes we did, is more useful than an agreement number on a homogeneous cohort that may not generalise.

Two results bear directly on eventual deployment. Per-group thresholds are 8.8% apart, and for six of the seven groups the pooled threshold is statistically indistinguishable from the per-group one, so a single pooled threshold would likely suffice for initial deployment, with per-group thresholds reserved for groups whose confidence intervals separate from the pooled value in larger characterisation data. The exception is Southeast Asian faces, for which the pooled threshold over-flags; deployments with meaningful Southeast Asian catchment should use a locally computed threshold or accept a slightly elevated false-positive rate there. The confidence framework carries only one of the two properties a screening filter needs. It controls false-positive rate on presumed-healthy inputs, by construction, but it cannot tell us how often it would flag an actual Bell’s palsy face. The pipeline is a candidate screening tool with a characterised specificity property and an unmeasured sensitivity as of now.

5 Future Work

The immediate next step is a proposed clinical validation study which would measure the pipeline’s sensitivity to actual Bell’s palsy and calibrate its thresholds against clinical grade. This is the only route to establishing clinical performance, and it would also allow the presumed-healthy baseline reported here to be compared against a screened, negative clinical cohort.

The denominator instability characterised in Section 3.3 motivates renormalising the regional index by interpupillary distance, with its effect on group-level and per-region behaviour re-measured on the same FairFace substrate for direct comparison against these results. An explicit frontal-pose quality gate would filter images with head yaw, pitch, or roll beyond a set tolerance before landmark extraction, reducing the pose variability that currently propagates into the reported distributions from FairFace’s uncontrolled poses. A

purpose-collected South Asian facial paralysis dataset, spanning acute and recovery timepoints and matched to House-Brackmann grade, would support both evaluation of the current pipeline and training of downstream classifiers that use its regional indices as features. Beyond single-image measurement, fusion with surface electromyography and temporal modelling of recovery across visits would extend the pipeline from a per-image measurement into a per-patient monitoring tool, appropriate to a condition whose clinical value lies in tracking change over its acute, plateau, and recovery phases.

Two properties of the present work bound how far these extensions reach. The characterisation is single-timepoint and uses FairFace race labels, which are annotator-assigned social categories rather than measured phenotypes, so the seven groups are useful sampling strata and no more. The presumed-healthy substrate was never clinically screened, MediaPipe is used as a black-box detector whose subgroup behaviour cannot be fully characterised from outside, and the discarded z coordinate places true 3D asymmetry beyond the current pipeline. Each of these is a specific target the work above is designed to address.

6 Conclusion

We engineered and characterised a landmark-based facial asymmetry pipeline for Bell's palsy assessment in under-resourced South Asian clinical settings, and evaluated it on 10,954 presumed-healthy faces across seven FairFace race categories. Detection succeeded on 90.13% of images, with a failure-rate ordering that inverts the skew commonly assumed for MediaPipe and points to confounds in FairFace's demographic composition. Baseline global asymmetry differed detectably but weakly across groups (Kruskal-Wallis $H(6) = 35.37$, $p = 3.66 \times 10^{-6}$; $\varepsilon^2 \approx 0.003$), with per-group flagging thresholds spread 8.8% around a pooled value of 0.636. The inner eye corner produced the highest regional asymmetry in every group, consistent with denominator instability rather than an anatomical finding. The confidence framework controls false-positive rate on presumed-healthy faces only, and the pipeline is ready for the proposed clinical validation subject to the failure modes characterised here.

Coulson, S. E., G. R. Croxson, R. D. Adams, and N. J. O'Dwyer. 2005. "Reliability of the 'Sydney,' 'Sunnybrook,' and 'House Brackmann' Facial Grading Systems to Assess Voluntary Movement and Synkinesis After Facial Nerve Paralysis." *Otolaryngol Head Neck Surg* 132 (4): 543–49. <https://doi.org/10.1016/j.otohns.2005.01.027>.

Guarin, D. L., J. Dusseldorp, T. A. Hadlock, and N. Jowett. 2018. "A Machine Learning Approach for Automated Facial Measurements in Facial Palsy." *JAMA Facial Plast Surg* 20 (4): 335–37. <https://doi.org/10.1001/jamafacial.2018.0030>.

Holm, S. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6 (2): 65–70.

House, J. W., and D. E. Brackmann. 1985. "Facial Nerve Grading System." *Otolaryngol Head Neck Surg* 93 (2): 146–47. <https://doi.org/10.1177/019459988509300202>.

Kärkkäinen, K., and J. Joo. 2021. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1548–58. <https://doi.org/10.1109/WACV48630.2021.00159>.

Kartynnik, Y., A. Ablavatski, I. Grishchenko, and M. Grundmann. 2019. "Real-Time Facial Surface Geometry from Monocular Video on Mobile GPUs." *arXiv Preprint arXiv:1907.06724*, ahead of print. <https://doi.org/10.48550/arXiv.1907.06724>.

Lugaresi, C., J. Tang, H. Nash, et al. 2019. "MediaPipe: A Framework for Building Perception Pipelines." *arXiv Preprint arXiv:1906.08172*, ahead of print. <https://doi.org/10.48550/arXiv.1906.08172>.

Peitersen, E. 2002. "Bell's Palsy: The Spontaneous Course of 2,500 Peripheral Facial Nerve Palsies of Different Etiologies." *Acta Otolaryngol Suppl* 122 (549): 4–30. <https://doi.org/10.1080/000164802760370736>.

Ross, B. G., G. Fradet, and J. M. Nedzelski. 1996. "Development of a Sensitive Clinical Facial Grading System." *Otolaryngol Head Neck Surg* 114 (3): 380–86. [https://doi.org/10.1016/S0194-5998\(96\)70206-1](https://doi.org/10.1016/S0194-5998(96)70206-1).

Roxas, A., M. M. Mehndiratta, N. Bornstein, et al. 2017. "The Professional Practice and Training of Neurology in the Asian and Oceanian Region: A Cross-Sectional Survey by the Asian and Oceanian Association of Neurology (AOAN)." *J Neurol Sci* 382: 108–15. <https://doi.org/10.1016/j.jns.2017.09.022>.

Sullivan, F. M., I. R. C. Swan, P. T. Donnan, et al. 2007. "Early Treatment with Prednisolone or Acyclovir in Bell's Palsy." *N Engl J Med* 357 (16): 1598–607. <https://doi.org/10.1056/NEJMoa072006>.

Tomczak, M., and E. Tomczak. 2014. "The Need to Report Effect Size Estimates Revisited. An Overview of Some Recommended Measures of Effect Size." *Trends in Sport Sciences* 1 (21): 19–25.