VOLUMEN 48
XXXI Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

Detección de parálisis facial con aprendizaje de maquina utilizando características de textura y geométricas

Jesús Eduardo González Trigueros[1], Mariana Jaqueline Peñaran Prieto[1], Emilio Muñoz Rivera[1], Jonathán de Jesús Estrella Ramirez[1], Juan Carlos Gómez Carranza[1*].

[1]Departamento de Ingeniería Electrónica, División de Ingenierías del Campus Irapuato – Salamanca, Universidad de Guanajuato.

{je.gonzaleztrigueros, mj.penaranprieto, e.munozrivera, jdj.estrellaramirez, jc.gomez} @ugto.mx

Resumen

Este trabajo propone una metodología computacional híbrida para la detección automatizada de la parálisis facial, diseñada para superar las limitaciones de los enfoques puramente geométricos ante alteraciones musculares sutiles. Se desarrolló un vector unificado de 60 dimensiones que fusiona métricas morfológicas, extraídas de una malla facial densa, con descriptores de textura superficial (LBP y Haralick) para capturar la pérdida de tono cutáneo. El sistema se evaluó mediante validación cruzada anidada en ocho escenarios experimentales utilizando cuatro bases de datos (CK+, YFP, MEEI y NeuroFace). Entre los clasificadores probados, el perceptrón multicapa (MLP) obtuvo el mejor desempeño, alcanzando una exactitud global de 0.957 (99.33% en condiciones ideales). La prueba de Wilcoxon $p = 1.36 \cdot 10^{-9}$ confirmó la superioridad sistemática de esta fusión bimodal sobre un modelo estrictamente geométrico. Se concluye que integrar geometría y textura provee una representación holística del déficit neuromotor, consolidando una herramienta diagnóstica altamente precisa y robusta.

Palabras clave: Parálisis facial, aprendizaje automático, geometría facial, análisis de textura, fusión multimodal, perceptrón multicapa.

1. Introducción

La parálisis facial, ocasionada por la disfunción del séptimo nervio craneal, deriva de múltiples etiologías como traumatismos y tumores. No obstante, la parálisis de Bell destaca como la causa idiopática más frecuente y de inicio agudo (*Peitersen, 2002; Tiemstra & Khatkhate, 2007*). Esta condición genera graves secuelas funcionales (como el riesgo de úlceras corneales) y un profundo impacto psicológico (*Butler & Grobbelaar, 2017*). Su evaluación clínica se divide en escalas de inspección visual y mediciones electrofisiológicas (*Lou et al., 2020*). Como estándar visual, la escala de House-Brackmann clasifica la disfunción en seis grados. Sin embargo, su dependencia de la observación subjetiva reduce su reproducibilidad y limita la detección de cambios sutiles (*House, 1983; Lou et al., 2020; Wang et al., 2025*). Por su parte, la escala de Sunnybrook (SFGS) ofrece una evaluación regional más detallada (*ten Harkel et al., 2023*). A pesar de ello, su aplicación prolongada la vuelve ineficiente y sigue arrastrando la misma subjetividad visual (*Lou et al., 2020; Wang et al., 2025*). Paralelamente, las alternativas electrofisiológicas (EMG y ENOG) brindan precisión, pero su naturaleza invasiva, costosa y dependiente de especialistas dificulta su uso rutinario (*Vrochidou et al., 2023; Wang et al., 2025*). Este contraste evidencia una clara brecha diagnóstica. Para resolverla, este trabajo presenta una metodología híbrida para la detección automatizada de la parálisis facial mediante la integración de geometría facial densa y textura superficial. Se propone un vector unificado de 60 dimensiones evaluado con cuatro clasificadores (SVM, RF, LGBM y MLP) sobre múltiples conjuntos de datos (CK+, YFP, MEEI y NeuroFace). Mediante un riguroso protocolo de validación cruzada anidada, se demuestra estadísticamente la superioridad de esta fusión multimodal frente a modelos puramente geométricos. El documento detalla los trabajos relacionados (Sección 2), el flujo metodológico (Sección 3), los resultados (Sección 4) y las conclusiones (Sección 5).

VOLUMEN 48
XXXI Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

2. Trabajos relacionados

En la evaluación computacional de la parálisis facial, los enfoques han evolucionado desde la extracción manual de características hasta la implementación de modelos de aprendizaje profundo. Históricamente, el paradigma predominante se ha basado en métodos geométricos que cuantifican la asimetría mediante el cálculo de distancias euclidianas y ángulos. En este contexto, la literatura reporta un alto rendimiento cuando se analizan características morfológicas; por ejemplo, la metodología propuesta por Parra-Domínguez et al., que emplea un vector de 29 métricas de simetría procesadas por un perceptrón multicapa, alcanzó un 97.22% de exactitud (*Parra-Dominguez et al., 2021*). Sin embargo, la limitación intrínseca de los enfoques geométricos puros radica en que, a pesar de su precisión cercana al 97%, son funcionalmente "ciegos" ante la atonía del músculo en reposo. Para subsanar la incapacidad geométrica de detectar alteraciones musculares sutiles, la investigación se ha dirigido hacia las arquitecturas de aprendizaje profundo. No obstante, la evidencia demuestra que la sofisticación de las redes neuronales convolucionales (CNN) no garantiza una superioridad diagnóstica y tiende a generar un sobreajuste severo (overfitting) cuando la disponibilidad de datos clínicos es limitada. Los estudios comparativos revelan un rendimiento inestable en tareas complejas de clasificación: la arquitectura 3DPalsyNet, desarrollada por Storey et al., alcanzó únicamente un 82.4% de exactitud al ser entrenada con un conjunto clínico reducido (*Storey et al., 2019*). De igual forma, modelos jerárquicos como la Red Jerárquica Profunda (HDN) de Hsu et al. vieron frenado su desempeño con una precisión máxima del 93.0% (*Hsu et al., 2019*). Estos hallazgos reafirman que, frente a bases de datos médicas escasas, los clasificadores clásicos ofrecen mayor estabilidad que los modelos profundos de "caja negra".

Frente a este escenario, el análisis de textura ha emergido como el biomarcador complementario necesario para capturar el alisamiento de la superficie cutánea asociado a la atonía. Sin embargo, investigaciones recientes, como los análisis comparativos de Himawan et al., han demostrado que las metodologías sustentadas de forma exclusiva en descriptores de textura presentan un techo de exactitud que ronda el 85%, evidenciando una carencia de robustez estructural (*Himawan et al., 2025)*. En consecuencia, existe una brecha metodológica evidente en el estado del arte: la geometría pura es precisa pero incapaz de identificar la atonía sutil; el aprendizaje profundo se sobreajusta ante la escasez de datos; y la textura por sí sola carece del soporte morfológico necesario. Estas limitaciones justifican la necesidad de una metodología híbrida. Integrar mallas faciales densas con descriptores de textura permite cuantificar de manera simultánea la deformación estructural y la pérdida biomecánica de los pliegues cutáneos (el efecto de "planchado"), superando las deficiencias de los modelos unimodales.

3. Metodología

La estrategia experimental se estructura como un flujo de trabajo secuencial diseñado para garantizar la estabilidad diagnóstica y la explicabilidad clínica del modelo. Este proceso, ilustrado en la Fig. 3.0.1, opera como una tubería continúa dividida conceptualmente en tres bloques interconectados: el procesamiento y normalización de imágenes, la extracción híbrida de biomarcadores geométricos y de textura, y finalmente, el protocolo de validación cruzada anidada. Los detalles técnicos, algoritmos y parámetros específicos de cada una de estas etapas se describen a profundidad en las subsecciones siguientes.

VOLUMEN 48
XXXI Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

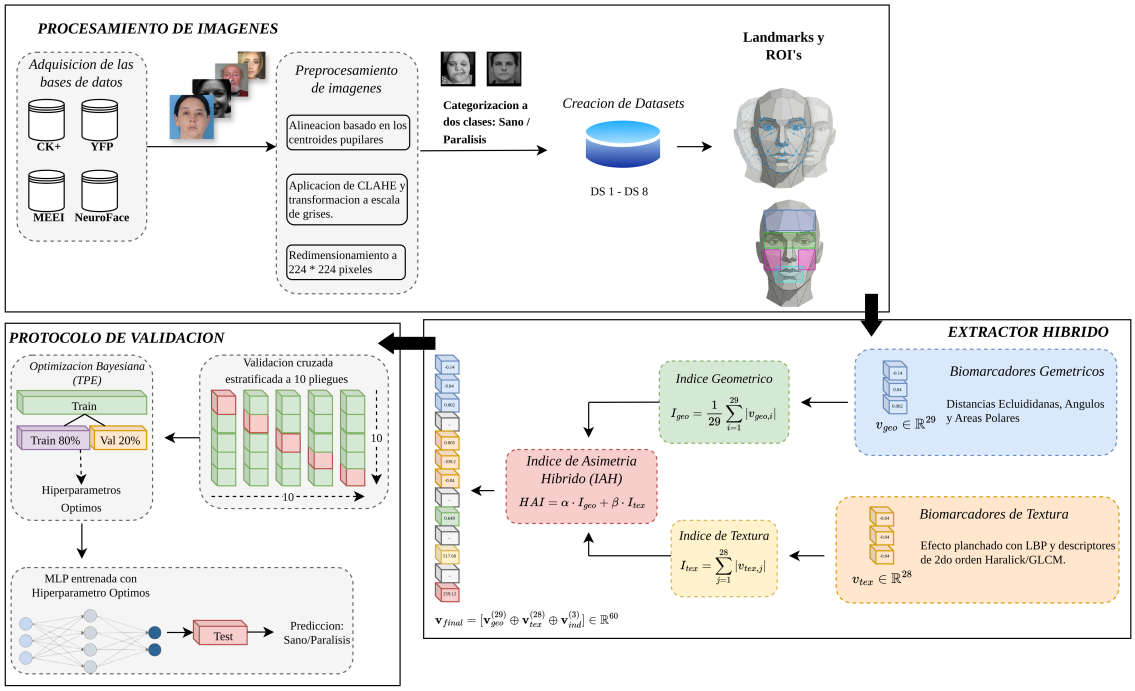


Figura 3.0.1. Metodología híbrida propuesta para la detección automatizada de parálisis facial.

3.1 Adquisición de datos

Para extraer biomarcadores confiables y garantizar la aplicabilidad del modelo en escenarios reales, el primer paso consistió en construir un repositorio consolidado a partir de cuatro bases de datos públicas, donde cada una aporta una pieza clave para asegurar la heterogeneidad fenotípica y mitigar los sesgos de selección. En primer lugar, la base de datos extendida de Cohn-Kanade (CK+) aporta el "control sano" mediante secuencias con expresiones estandarizadas bajo el Sistema de Codificación de Acción Facial (FACS), lo que permite que el sistema aprenda a diferenciar una asimetría emocional normal de una asimetría patológica (*Lucey et al., 2010*). Por su parte, el conjunto YouTube Facial Palsy (YFP) introduce la variabilidad no controlada de un entorno "en la calle" (in the wild), aportando cambios drásticos de iluminación y postura para poner a prueba la robustez del modelo frente al ruido ambiental (*Hsu et al., 2019*). En el ámbito estrictamente médico, el conjunto de la Massachusetts Eye and Ear Infirmary (MEEI) funciona como la "verdad de referencia" clínica (ground truth), ya que incluye pacientes evaluados previamente con las escalas estándar de House-Brackmann y Sunnybrook (*Greene et al., 2020*). Finalmente, Toronto NeuroFace (NF) aporta casos de etiologías neurológicas más complejas, como los accidentes cerebrovasculares (ACV) y la esclerosis lateral amiotrófica (ELA), lo que permite comprobar que los biomarcadores generalizan de manera efectiva más allá de la parálisis de Bell (*Bandini et al., 2021*).

3.2 Configuración experimental

Una vez consolidado el corpus de imágenes, el diagnóstico se modeló computacionalmente como una tarea de clasificación binaria (parálisis vs sano) con el objetivo de evaluar el sistema como una herramienta automatizada de cribado. Para aislar variables clínicas y ambientales específicas, la experimentación se estructuró en ocho escenarios de evaluación (DS1 a DS8). Estos conjuntos fueron diseñados para abarcar un espectro completo de dificultades: desde entornos controlados con separabilidad ideal, pasando por pruebas de generalización multifuente y réplicas exactas del estado del arte para realizar comparativas directas, hasta pruebas de estrés extremo que evalúan la robustez del modelo ante un desequilibrio de clases severo (proporción 5:1). La experimentación se estructuró en ocho escenarios de evaluación (DS1-DS8) para aislar variables clínicas y operativas específicas:

VOLUMEN 48
XXXI Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

- DS1 (Separabilidad ideal): Establece el límite superior de rendimiento en condiciones controladas utilizando 1,962 muestras balanceadas (positivos de YFP y negativos de CK+).
- DS2 (Generalización multifuente): Evalúa la capacidad de procesar datos heterogéneos simultáneamente con 1,600 muestras balanceadas de YFP, MEEI y Toronto NF.
- DS3 (Validación en video): Compara secuencias temporales utilizando 800 fotogramas balanceados obtenidos íntegramente de MEEI.
- DS4 (Réplica MEEI): Replica las evaluaciones de trabajos previos (Parra-Domínguez et al.) empleando 640 muestras desbalanceadas (400 positivas y 240 negativas).
- DS5 (Unificación a escala máxima): Prueba la estabilidad general del sistema agrupando muestras únicas de DS1, DS3 y DS7, totalizando 4,762 imágenes balanceadas.
- DS6 (Réplica Toronto NF): Evalúa casos complejos de accidentes cerebrovasculares conservando el desequilibrio original de la base de datos (1,036 muestras: 817 positivas y 219 negativas).
- DS7 (Toronto NF Balanceado): Versión equilibrada de DS6 para medir el rendimiento en etiologías complejas sin sesgo estadístico, usando 2,000 muestras.
- DS8 (Estrés por desequilibrio): Evalúa el comportamiento del modelo bajo una proporción natural extrema de 5:1 (400 positivas y 80 negativas) empleando el conjunto estático de MEEI.

3.3 Preprocesamiento y estandarización

Dada la heterogeneidad de los datos recopilados, se implementó un proceso de normalización diseñado para separar la variabilidad patológica de los artefactos generados durante la captura de la imagen. En primer lugar, se aplica una transformación afín rígida para corregir la pose (inclinación) y la escala del rostro. Utilizando los centroides pupilares (P_L, P_R) como anclajes anatómicos, se calcula el ángulo de rotación θ para alinear horizontalmente la línea bipupilar. Esta alineación por centroides pupilares resuelve un problema técnico fundamental: corrige el ángulo de la cabeza para garantizar que el sistema computacional no confunda un rostro ladeado con una asimetría facial real. Simultáneamente, se deriva un factor de escala s para establecer una distancia interpupilar constante (d_{obj}) de 100 píxeles. Esta normalización por distancia interpupilar es lo que le otorga invariancia de escala al modelo; es decir, garantiza que una fotografía tomada de cerca y otra capturada desde lejos produzcan medidas biométricas matemáticamente comparables. Ambos cálculos se formalizan de la siguiente manera:

$$\theta = \arctan\left(\frac{\Delta y}{\Delta x}\right)$$

$$s = \frac{d_{obj}}{\sqrt{\Delta x^2 + \Delta y^2}}$$

Estos parámetros definen la matriz de transformación M que proyecta la imagen en un espacio anatómico canónico mediante interpolación bicúbica:

$$M = \begin{bmatrix} s\cos\theta & -s\sin\theta & (1 - s\cos\theta)C_x - s\sin\theta\, C_y \\ s\sin\theta & s\cos\theta & s\sin\theta\, C_x + (1 - s\cos\theta)C_y \end{bmatrix}$$

Posteriormente, se extrae una región de interés (ROI) adaptativa con un factor de expansión (padding) de 0.5 para preservar el contexto de la mandíbula y la frente. Para maximizar la capacidad de discriminar texturas, esta ROI se convierte a escala de grises y se procesa mediante un algoritmo de ecualización adaptativa de contraste (CLAHE), configurado con un límite de recorte de 2.0 y una cuadrícula de 8x8. La aplicación de CLAHE resulta fundamental porque resuelve el desafío de resaltar las altas frecuencias locales (detalles finos como arrugas y surcos) sin amplificar el ruido estocástico de la imagen. Este realce es un paso crítico en la metodología, ya que la preservación de estos detalles es precisamente lo que permitirá que el extractor detecte más adelante el "efecto planchado" (la pérdida patológica de pliegues cutáneos). Finalmente, la imagen se estandariza a una resolución de 224×224 píxeles utilizando interpolación de área.

3.4 Extracción de características y biomarcadores descriptivos

Este estudio aprovecha una topología densa de 468 vértices obtenida mediante la malla facial de MediaPipe para analizar la microgeometría del rostro. Para evitar el ruido de alta dimensionalidad proveniente de zonas rígidas sin relevancia diagnóstica, se implementa un enmascaramiento anatómico selectivo que restringe el análisis computacional a cinco regiones de interés (ROIs) alineadas con los grupos musculares miméticos críticos. Específicamente, se aíslan los contornos correspondientes a la región ocular (m. orbicularis oculi) para evaluar anomalías como la ptosis y el lagoftalmos mediante la relación de aspecto ocular (EAR); las cejas (m. frontalis) para cuantificar su caída o aplanamiento; la región oral (m. orbicularis oris) para medir la desviación y el tono muscular; y la región nasolabial (m. zygomaticus) para calcular la profundidad del surco como indicador de hipotonía malar. Finalmente, se establece un eje de simetría central inamovible empleando los vértices de la glabela, la punta nasal, el mentón y las fosas nasales, trazando la referencia estructural exacta sobre la cual se calculan todas las asimetrías laterales. A partir de estas coordenadas anatómicas, se procede al cálculo matemático de las características morfológicas y a la conformación del vector geométrico y de textura.

3.4.1 Biomarcadores geométricos

A partir de las regiones de interés (ROIs) definidas en la topología anatómica, se calcula un vector de características geométricas de 29 dimensiones designado como v_{geo}. Para superar las limitaciones de los modelos morfológicos tradicionales, el diseño de este vector implementa tres mejoras matemáticas fundamentales:

1. El uso de diferencias con signo ($\Delta = V_R - V_L$)para preservar la lateralidad de la parálisis.
2. La inclusión de descriptores integrales, tales como áreas poligonales y curvaturas, que capturan la deformación regional más allá de simples distancias lineales.
3. Una estricta normalización estructural basada en la distancia de referencia interpupilar D_{ref}.

Esta distancia de referencia se calcula como la norma euclidiana entre los centroides oculares, garantizando que todas las métricas sean invariantes a la escala y a la distancia de la cámara:

$$D_{ref} = \left\| \overline{P}_{Reye} - \overline{P}_{Leye} \right\|_2$$

Para cuantificar la severidad de la patología, se calcula el diferencial de asimetría $\Delta_i = |Value_R - Value_L|$ para cada característica. Bajo esta lógica clínica, los valores donde $\Delta_i \approx 0$ denotan una simetría estructural conservada (fenotipo sano), mientras que las magnitudes $\Delta_i \gg 0$ indican una deformación patológica evidente. Conceptualmente, estas métricas operan como vectores macroscópicos que modelan dinámicas de deformación específicas: por ejemplo, la discrepancia en el vector cigomático cuantifica la caída gravitacional de la mejilla debido a la pérdida de tono muscular (hipotonía), mientras que el vector de apertura palpebral sirve como un indicador cuantitativo preciso del lagoftalmos. El vector geométrico de 29 dimensiones se conforma en las cuatro regiones anatómicas críticas:

- Región Frontal (*M. Frontalis*): Cuantifica la elevación, forma y curvatura de las cejas mediante el cálculo de diferencias de altura, longitudes de arco y desplazamientos, lo cual permite detectar la caída o el aplanamiento de esta zona.
- Región Ocular (*M. Orbicularis Oculi*): Evalúa la apertura, orientación y geometría palpebral utilizando parámetros como la relación de aspecto ocular (EAR), áreas poligonales, ángulos de eje y distancias cantales. Estas métricas operan como indicadores cuantitativos directos para anomalías como el lagoftalmos.
- Dinámica Perioral (*M. Zygomaticus / Risorius*): Analiza la simetría de los labios, comisuras y pliegues nasolabiales a través del cálculo de áreas, grosor labial, índices de curvatura y desplazamientos orales. Este conjunto captura la caída gravitacional de la mejilla provocada por la hipotonía muscular.
- Estructura Global: Mide las asimetrías generales en el resto del rostro, integrando discrepancias en el área de las narinas, la desviación de la punta nasal respecto al eje central, el desplazamiento malar y las diferencias en la curvatura de la mandíbula."

3.4.2 Biomarcadores de textura

Para complementar el análisis morfológico, la metodología incorpora un módulo de evaluación superficial fundamentado en la biomecánica de la piel. La atonía muscular inducida por la parálisis facial provoca el borramiento patológico de los pliegues cutáneos, un fenómeno caracterizado clínicamente como el "efecto planchado". Debido a esta pérdida de tono, el lado paralizado exhibe una suavidad anómala que contrasta drásticamente con la textura rica en detalles (arrugas y surcos) del lado sano. Para cuantificar este fenómeno, el sistema procesa de forma bilateral cuatro regiones de interés (ROIs) críticas: la frente, la zona periorbital, la mejilla y el pliegue nasolabial. A partir de estas áreas, se extrae un vector de textura de 28 dimensiones (v_{tex}) utilizando dos descriptores complementarios que operan sobre la imagen ecualizada:

1. **Patrones Binarios Locales (LBP):** Se emplea el operador $LBP^{u2}_{16,2}$ para codificar la varianza del contraste local (Ojala *et al., 2002*). En lugar de medir distancias físicas, este descriptor evalúa la discrepancia textural comparando los histogramas del lado sano (H_L) y del lado afectado (H_R). Esta diferencia se cuantifica rigurosamente mediante la distancia chi-cuadrada (χ^2):

$$f_{\mathrm{LBP}} = \chi^2(H_L, H_R) = \frac{1}{2}\sum_{i=1}^{B}\frac{(H_L[i] - H_R[i])^2}{H_L[i] + H_R[i] + \varepsilon}$$

   Al calcular esta distancia de manera independiente para cada una de las cuatro ROIs estratégicas, se obtienen 4 características escalares que caracterizan la disparidad global de la rugosidad facial.
2. **Matrices de Co-ocurrencia de Niveles de Gris (GLCM)**: Para capturar la complejidad de la superficie, se calculan estadísticas de segundo orden o características de Haralick (H*aralick et al., 1973*). Siguiendo el mismo principio diferencial establecido en el análisis geométrico, las métricas se calculan bilateralmente y la asimetría funcional se define como $\Delta_{Har} = Val_R - Val_L$ para aislar la discrepancia neuromotora. Se aplican seis estadísticas específicas a las cuatro ROIs, generando un total de 24 descriptores de textura con una interpretación clínica directa. Específicamente, el contraste y la disimilitud cuantifican las variaciones locales de intensidad y la reducción de la complejidad textural, las cuales disminuyen drásticamente en las zonas afectadas debido al efecto de planchado. En contraste, la homogeneidad y la energía presentan incrementos atípicos en el lado paralizado, ya que valores altos en ambas métricas indican una textura de piel extremadamente suave y uniforme por la ausencia de arrugas estructurales. Finalmente, la correlación evalúa la regularidad estructural y la dependencia lineal de los niveles de gris, mientras que el incremento del Segundo Momento Angular (ASM) funciona como un indicador directamente proporcional al grado de atonía muscular de la región analizada**.**

En conjunto, los 4 componentes derivados de LBP y los 24 descriptores generados mediante GLCM consolidan el vector de textura de 28 dimensiones (v_{tex}). Esta extracción garantiza que el modelo analice no solo la caída física del rostro, sino también la pérdida de la integridad fisiológica y biomecánica de la piel.

3.4.3 Fusión multimodal e índices globales

En la etapa final de la extracción de características, la evidencia morfológica y superficial recopilada no se procesa de manera aislada. Para maximizar la robustez de la clasificación, los descriptores individuales se sintetizan en tres biomarcadores globales. Esta abstracción de alto nivel resuelve un problema analítico fundamental: le permite al modelo computacional capturar el grado general de asimetría del paciente, yendo más allá de las variaciones locales individuales y evaluando el rostro como un sistema neuromotor unificado. Estos tres biomarcadores globales se definen de la siguiente manera:

1. **Índice Geométrico** (I_{geo})**:** Este índice se calcula como el promedio de las magnitudes absolutas de las 29 características geométricas previamente extraídas. Su propósito clínico es cuantificar la deformación estructural total a lo largo de toda la malla facial. Matemáticamente se expresa como:

$$I_{geo} = \frac{1}{29}\sum_{i=1}^{29}|v_{geo,i}|$$

2. **Índice de Textura** (I_{tex})**:** A diferencia del promedio geométrico, este biomarcador representa la sumatoria de las diferencias superficiales acumuladas a través de las cuatro regiones de interés (ROIs) procesadas. Su función es capturar y resumir la pérdida global del tono muscular y el borramiento de los pliegues cutáneos en todo el rostro:

$$I_{geo} = \frac{1}{29}\sum_{j=1}^{28}|v_{tex,j}|$$

3. **Índice de Asimetría Híbrido (HAI):** En la práctica clínica, existen patologías que pueden manifestarse de forma muy sutil en la estructura morfológica, pero que resultan sumamente evidentes en la textura de la piel (o viceversa). Para detectar estos casos, se implementa una combinación lineal ponderada de los dos índices anteriores. Para garantizar que tanto el dominio morfométrico como el textural contribuyan de manera equitativa al diagnóstico, los coeficientes de ponderación se establecen como α=0.5 y β=0.5:

$$HAI = \alpha \cdot I_{geo} + \beta \cdot I_{tex}$$

La integración definitiva del sistema se consolida mediante una estrategia de fusión temprana (*early fusion*). Los descriptores específicos de cada dominio (29 morfológicos y 28 texturales) y los índices globales (3 biomarcadores resumen) se concatenan mediante el operador de suma directa en un vector denso unificado de 60 dimensiones $(v_{final} \in \mathbb{R}^{60})$. Este proceso se define formalmente como:

$$v_{final} = \left[v_{geo}^{(29)} \oplus v_{tex}^{(28)} \oplus v_{ind}^{(3)}\right]$$

Este vector de 60 dimensiones constituye la entrada estandarizada para la posterior etapa de clasificación mediante aprendizaje automático. A través de este empaquetamiento matemático, el sistema logra una representación holística del estado fisio-neuromotor del paciente, integrando simultáneamente la simetría estructural (forma) y la integridad fisiológica de los tejidos (superficie) para emitir una predicción.

3.5 Protocolo de validación

Para garantizar una estimación imparcial del error y prevenir la filtración de datos, el diseño experimental empleó una validación cruzada anidada estratificada ($k = 10, l = 5$ \$) que desacopla la selección de hiperparámetros de la evaluación final. El ciclo externo reservó iterativamente un 10% de los datos para prueba, mientras que el ciclo interno subdividió el 90% restante en 5 pliegues para ejecutar 50 iteraciones de optimización bayesiana mediante el estimador de Parzen estructurado en árbol (TPE). Con la configuración óptima, los modelos se reentrenaron con el conjunto de ajuste completo y se evaluaron en la partición oculta. La capacidad discriminativa del vector híbrido de 60 dimensiones se evaluó mediante cuatro clasificadores, optimizando los siguientes espacios de búsqueda de hiperparámetros: para máquinas de vectores de soporte (SVM) se ajustaron los parámetros $C \in [10^{-4}, 10^{4}]$y $\gamma \in [10^{-4}, 10^{1}]$; para bosques aleatorios (RF), el número de estimadores osciló de 10 a 200 con una profundidad máxima de 2 a 32; para LGBM, la tasa de aprendizaje se buscó entre 10^{-3}y 0.1con un límite de 2 a 256 hojas; y para el Perceptrón Multicapa (MLP), se configuró una arquitectura variable de 1 a 3 capas ocultas, de 6 a 128 neuronas y una tasa de dropout de 0.0 a 0.5. Para SVM y MLP, se aplicó una estandarización z-score calculada estrictamente sobre cada pliegue de entrenamiento individual, fijando la semilla aleatoria en 42 dentro de la librería Optuna para asegurar la reproducibilidad total.

4. Resultados y discusión

La evaluación del sistema se llevó a cabo a través de la ejecución del protocolo de validación cruzada anidada sobre los ocho escenarios experimentales (DS1-DS8), lo que permite no solo cuantificar el rendimiento puro, sino también interpretar el comportamiento del modelo ante distintas variables clínicas y operativas. Para analizar el desempeño del vector híbrido de 60 dimensiones, se evaluó su capacidad discriminativa (ver Tabla 1) empleando cuatro clasificadores (SVM, RF, LGBM y MLP).

VOLUMEN 48
XXXI Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

***Tabla 1**. Evaluación comparativa completa del rendimiento en todos los escenarios experimentales. Los resultados se presentan como promedio y desviación estándar ($Mean_{SD}$) de los 10 pliegues. El texto en negrita indica el modelo con mejor rendimiento en cada escenario.*

ID	*Métrica*	*MLP*	*SVM*	*RF*	*LGBM*	*Promedio*
DS1	Accuracy	**$0.993_{0.004}$**	$0.989_{0.010}$	$0.973_{0.009}$	$0.988_{0.007}$	0.986
	F1-macro	**$0.993_{0.004}$**	$0.989_{0.010}$	$0.973_{0.009}$	$0.988_{0.007}$	0.986
DS2	Accuracy	$0.963_{0.015}$	**$0.964_{0.015}$**	$0.934_{0.017}$	$0.960_{0.013}$	0.955
	F1-macro	$0.962_{0.015}$	**$0.964_{0.015}$**	$0.933_{0.017}$	$0.960_{0.013}$	0.955
DS3	Accuracy	**$0.971_{0.013}$**	$0.953_{0.028}$	$0.948_{0.029}$	$0.968_{0.029}$	0.960
	F1-macro	**$0.971_{0.013}$**	$0.952_{0.028}$	$0.947_{0.030}$	$0.967_{0.029}$	0.960
DS4	Accuracy	**$0.911_{0.025}$**	$0.883_{0.058}$	$0.847_{0.040}$	$0.878_{0.040}$	0.880
	F1-macro	**$0.906_{0.028}$**	$0.867_{0.076}$	$0.835_{0.046}$	$0.869_{0.045}$	0.869
DS5	Accuracy	**$0.974_{0.006}$**	$0.958_{0.016}$	$0.922_{0.015}$	$0.965_{0.011}$	0.955
	F1-macro	**$0.974_{0.006}$**	$0.958_{0.016}$	$0.921_{0.015}$	$0.965_{0.011}$	0.955
DS6	Accuracy	**$0.969_{0.018}$**	$0.953_{0.032}$	$0.901_{0.023}$	$0.949_{0.021}$	0.943
	F1-macro	**$0.966_{0.020}$**	$0.942_{0.043}$	$0.854_{0.039}$	$0.863_{.103}$	0.906
DS7	Accuracy	**$0.987_{0.010}$**	$0.973_{0.018}$	$0.951_{0.017}$	$0.976_{0.012}$	0.972
	F1-macro	**$0.987_{0.010}$**	$0.973_{0.018}$	$0.951_{0.017}$	$0.976_{0.012}$	0.972
DS8	Accuracy	$0.885_{0.045}$	**$0.888_{0.042}$**	$0.842_{0.011}$	$0.867_{0.034}$	0.870
	F1-macro	**$0.772_{0.088}$**	$0.729_{0.113}$	$0.501_{0.025}$	$0.686_{0.114}$	0.672
Promedio Accuracy Global		**0.957**	0.945	0.915	0.944	--
Promedio F1-macro Global		**0.941**	0.921	0.864	0.911	--

Desde una perspectiva general, el Perceptrón Multicapa (MLP) se posicionó como la arquitectura óptima, alcanzando el mayor rendimiento global con una exactitud media de 0.957 y una F1-macro de 0.941. Este clasificador lideró la mayoría de los entornos, obteniendo métricas máximas en escenarios clave como DS1 (99.33%), DS5 (97.35%) y DS7 (98.73%). Por su parte, los modelos SVM y LGBM mostraron resultados competitivos y cercanos, mientras que el Bosque Aleatorio (RF) presentó el desempeño más bajo y una mayor susceptibilidad a la variabilidad de los datos. Este comportamiento sugiere que la integración de métricas de diversa naturaleza (morfológicas y texturales) genera un espacio de características altamente no lineal que el MLP logra modelar con mayor eficacia que los métodos basados en árboles de decisión, los cuales presentan una degradación más pronunciada ante la heterogeneidad. En cuanto a la complejidad de los entornos, el escenario DS1 demostró ser el más sencillo de clasificar, estableciendo el límite superior del modelo (exactitud media general de 0.986) gracias a sus condiciones de separabilidad ideal y balance perfecto, lo que permite al vector híbrido capturar inequívocamente las diferencias entre sujetos sanos y con parálisis. En escenarios de generalización como DS2 y DS5, la estabilidad de los resultados confirma que la

VOLUMEN 48
XXXI Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

inclusión de texturas superficiales logra compensar las variaciones de iluminación y protocolos de captura al unificar múltiples bases de datos. Por otro lado, el contraste entre el procesamiento de secuencias de video (DS3) y capturas estáticas (DS4) reveló una brecha analítica importante: las secuencias temporales resultan más fáciles de procesar porque ofrecen una mayor variabilidad de microgestos que enriquecen el vector, mientras que las imágenes estáticas limitan la información disponible para el extractor. Asimismo, al balancear artificialmente las muestras de una base de datos neurológica compleja (DS7 frente a DS6), la exactitud ascendió significativamente por encima del 98.73%, confirmando la alta capacidad diagnóstica del modelo en ausencia de sesgos estadísticos. Para comprender a profundidad cómo estos retos (especialmente el desequilibrio de clases) impactan la predicción, el análisis visual de los errores resulta revelador. Se examinaron las matrices de confusión del MLP en tres escenarios críticos (ver Fig. 4.0.1): separabilidad ideal (DS1), generalización máxima (DS5) y estrés por desequilibrio extremo (DS8). Los números en las matrices indican el promedio y la desviación estándar de los 10 pliegues.

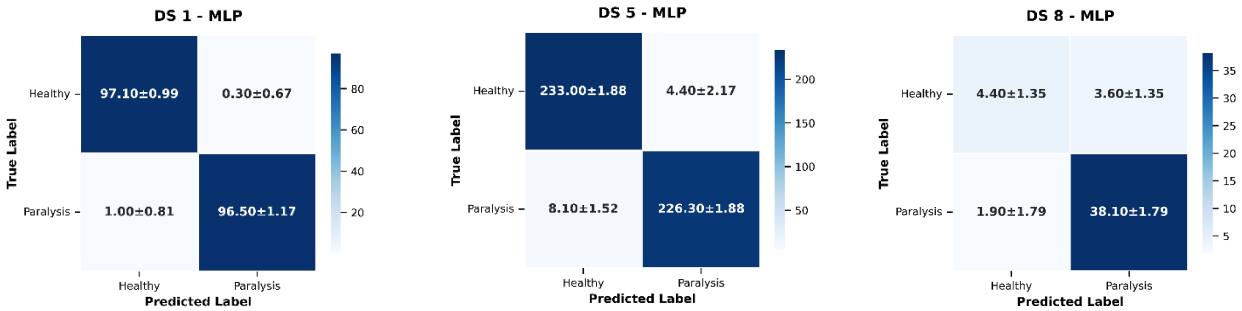


Figura 4.0.1. Matrices de confusión para tres escenarios experimentales críticos utilizando el clasificador MLP. (Healthy = Sano, Paralysis = Parálisis)

En el escenario DS1, la matriz refleja una precisión casi absoluta; la inmensa mayoría de las instancias se concentran en la diagonal principal con la presencia marginal de errores. En el escenario DS5, que representa el reto de unificar todas las bases de datos, la matriz demuestra una notable robustez; a pesar del incremento en el volumen y la heterogeneidad clínica, el sistema logró mantener altas tasas de verdaderos positivos y verdaderos negativos de manera proporcional. Sin embargo, el escenario DS8 (con una proporción 5:1, donde la parálisis sobrerrepresenta a la clase sana) se consolidó como el desafío técnico más severo. La matriz de confusión de DS8 hace visible el sesgo algorítmico inducido por esta escasez de datos: el sistema tiende a predecir "parálisis" incluso en individuos sanos, sacrificando el rendimiento sobre la clase minoritaria. Este comportamiento confirma que, incluso con biomarcadores altamente discriminativos, la carencia de muestras sesga las fronteras de decisión hacia la categoría dominante. A pesar de estas dificultades intrínsecas del desequilibrio, resulta imperativo cuantificar y validar el valor añadido de la metodología propuesta. Por ello, la arquitectura híbrida de 60 dimensiones se comparó directamente contra un estándar de referencia puramente geométrico de 29 dimensiones (*Parra-Dominguez et al., 2021*) bajo las mismas condiciones experimentales. Para determinar rigurosamente si la superioridad del vector híbrido sobre este modelo base poseía significancia generalizada, se aplicó una prueba estadística de rangos con signo de Wilcoxon de una cola sobre una muestra total de 80 métricas de evaluación pareadas.

***Tabla 2**. Análisis comparativo de la precisión media (validación cruzada de 10 pliegues): vector geométrico de referencia (v_{base}) frente al vector híbrido propuesto (v_{final}). La significancia estadística se evaluó mediante la prueba de rangos con signo de Wilcoxon.*

ID	Acc. Promedio Baseline	Acc. Promedio Híbrido (este trabajo)	Ganancia neta (Δ)	$(p - value)$
DS1	97.81%	99.33%	+1.52%	0.001
DS2	95.69%	96.25%	+0.56%	0.216
DS3	94.99%	97.12%	+2.13%	0.024
DS4	89.22%	91.09%	+1.88%	0.084
DS5	94.32%	97.35%	+3.03%	0.001
DS6	94.86%	96.93%	+2.07%	0.032
DS7	96.49%	98.73%	+2.24%	0.004
DS8	84.38%	88.53%	+4.16%	0.016
Validación estadística global (N = *80*)				
Estadístico de Wilcoxon *W* = 2797.0 Global *p-value* $= 1.36 \cdot 10^{-9}$				

Los resultados de esta comparativa evidenciaron una tendencia clara: el modelo puramente geométrico presenta un rendimiento sistemáticamente inferior y una mayor fluctuación frente a cambios en la composición de los datos. La evaluación estadística arrojó un valor $p = 1.36 \times 10^{-9}$, confirmando con un nivel de confianza superior al 99.9% que la ganancia discriminativa obtenida por la fusión multimodal no es producto del azar, sino el resultado de una representación diagnóstica mucho más completa. El desglose demostró mejoras estadísticamente significativas en entornos controlados y balanceados, como DS1 (donde la línea base obtuvo 97.81% frente al 99.33% del híbrido, p=0.001), DS3 (p=0.024) y DS7 (p=0.004). Asimismo, en el desafío de escala máxima (DS5), la mejora del +3.03% (p=0.001) demostró que la contribución de la textura se mantiene firme incluso al integrar orígenes heterogéneos. La debilidad del modelo estrictamente morfológico y el valor clínico de la propuesta se hicieron aún más evidentes en la prueba de estrés (DS8). En este escenario de desequilibrio severo, la exactitud de la geometría pura cayó drásticamente al 84.38%, evidenciando que, al carecer de soporte superficial, es funcionalmente "ciega" ante alteraciones musculares sutiles; en contraste, el vector híbrido logró una ganancia neta del +4.16% (p=0.016), corroborando que los descriptores texturales logran capturar el "efecto planchado" o el alisamiento anómalo de la piel donde la asimetría posicional por sí sola resulta insuficiente. Finalmente, las métricas para DS4 (p=0.084) y DS2 (p=0.216) no alcanzaron el umbral de significancia convencional; este fenómeno obedece a que la dependencia de capturas estáticas en DS4 limita la variabilidad de la señal textural, mientras que en DS2, la extrema heterogeneidad de los protocolos entre bases atenúa la contribución diferencial de la textura frente al ruido experimental dominante.

5. Conclusiones

En esta investigación se validó una metodología híbrida para la detección automatizada de la parálisis facial que integra geometría densa y textura superficial en un vector unificado de 60 dimensiones. Este enfoque supera las limitaciones de los modelos puramente morfológicos al cuantificar simultáneamente la deformación estructural y el ``efecto planchado" (alisamiento anómalo de los pliegues cutáneos por atonía muscular), logrando una representación holística del déficit neuromotor. La evaluación experimental sobre un corpus

multifuente (CK+, YFP, MEEI y Toronto NeuroFace) confirmó la robustez de la propuesta. El Perceptrón Multicapa (MLP) alcanzó el mejor desempeño con una exactitud global de 0.957 y un F1-macro de 0.941. Asimismo, la prueba de Wilcoxon ($p = 1.36 \times 10^{-9}$) demostró con un nivel de confianza superior al 99.9% que el modelo multimodal supera sistemáticamente al enfoque geométrico tradicional, siendo un factor crítico en escenarios de desequilibrio severo donde los biomarcadores unimodales pierden eficacia]. Como trabajo futuro, se plantea transicionar a un sistema multiclase capaz de estimar grados de severidad según las escalas de House-Brackmann o Sunnybrook, además de expandir la diversidad fenotípica del conjunto de datos.

Bibliografía/Referencias

Bandini, A., Orlandi, S., Green, J. R., Gunawardana, S. B., De Marchi, F., Mazzini, L., & Yunusova, J. (2021). A New Dataset for Facial Motion Analysis in Individuals With Neurological Disorders. IEEE Journal of Biomedical and Health Informatics, 25(4), 1111-1119.

Butler, D. P., & Grobbelaar, A. O. (2017). Facial palsy: what can the multidisciplinary team do?. Journal of Multidisciplinary Healthcare, 10, 377-381.

Greene, J. J., Guarin, D. L., Tavares, J., Fortier, E., Robinson, M., Dusseldorp, J., Quatela, O., Jowett, N., & Hadlock, T. (2020). The spectrum of facial palsy: The MEEI facial palsy photo and video standard set. Laryngoscope, 130, 32-37.

Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics, SMC-3(6), 610-621.

Himawan, S. N., Suheryadi, A., Cahyanto, K. A., Sitanggang, F., & Pamungkas, K. A. (2025). Comparative analysis of feature extraction techniques for facial paralysis classification. Journal of Electronics, Electromedical Engineering, and Medical Informatics, 7(2), 341-351.

House, J. W. (1983). Facial Nerve Grading Systems. Laryngoscope, 93(8), 1056-1069.

Hsu, G.-S. J., Kang, J.-H., & Huang, W.-F. (2019). Deep Hierarchical Network With Line Segment Learning for Quantitative Analysis of Facial Palsy. IEEE Access, 7, 4833-4842.

Lou, J., Yu, H., & Wang, F.-Y. (2020). A review on automated facial nerve function assessment from visual face capture. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 28(2), 488-497.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. En 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops (CVPRW) (pp. 94-101).

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 971-987.

Parra-Dominguez, G. S., Sanchez-Yanez, R. E., & Garcia-Capulin, C. H. (2021). Facial Paralysis Detection on Images Using Key Point Analysis. Applied Sciences, 11(5), 2435.

Peitersen, E. (2002). Bell's Palsy: The Spontaneous Course of 2,500 Peripheral Facial Nerve Palsies of Different Etiologies. Acta Oto-Laryngologica, 122(7), 4-30.

Storey, G., Jiang, R., Keogh, S., Bouridane, A., & Li, C.-T. (2019). 3dpalsynet: A facial palsy grading and motion recognition framework using fully 3d convolutional neural networks. IEEE Access, 7, 121655-121664.

ten Harkel, T., de Jong, G., Marres, H., Ingels, K., Speksnijder, C., & Maal, T. (2023). Automatic grading of patients with a unilateral facial paralysis based on the sunnybrook facial grading system - a deep learning study based on a convolutional neural network. American Journal of Otolaryngology, 44(3), 1-8.

Tiemstra, J. D., & Khatkhate, N. (2007). Bell's palsy: diagnosis and management. American Family Physician, 76(7), 997-1002.

Vrochidou, E., Papi, V., Kalampokas, T., & Papakostas, G. A. (2023). Automatic facial palsy detection from mathematical modeling to deep learning. Axioms, 12(12).

Wang, J., Tang, X., Zhang, J., Shao, X., Liu, Y., Zhai, Y., Xu, J., Lan, S., Yen, Y., & Wang, C. (2025). A review of the application of intelligent sensing technology in the recognition and evaluation of facial paralysis. Frontiers in Neuroscience, 19, 1646485.