Evaluación de la Técnica del Bulgarian Split Squat Basada en Aprendizaje Profundo Usando MediaPipe y Redes Recurrentes Bidireccionales

Juan Jose Núñez, Juan Jose Castro
Universidad San Buenaventura
Cali, Colombia

***Resumen*—La rehabilitación física y la técnica adecuada en el ejercicio son cruciales para prevenir lesiones musculoesqueléticas y optimizar el rendimiento atlético. Los métodos tradicionales de evaluación dependen en gran medida de la supervisión experta, lo cual es costoso en tiempo, subjetivo y no escalable para entrenamiento domiciliario. Este artículo presenta un sistema automatizado de visión por computadora para la evaluación en tiempo real de la ejecución del Bulgarian Split Squat utilizando estimación de pose con MediaPipe y redes Bidireccionales Recurrentes con Unidades Gateadas (BiGRU). Nuestro enfoque extrae 33 puntos anatómicos por fotograma y calcula características biomecánicas incluyendo ángulos articulares, rango de movimiento (ROM) y suavidad del movimiento. El clasificador BiGRU mejorado con mecanismo de atención logra un macro-F1 de 51.98 % y una precisión de 65.74 % en un conjunto de datos de 16,501 muestras balanceadas con 4 clases, identificando correctamente cuatro patrones de error: ejecución correcta (E0), inclinación excesiva del tronco (E1), valgo de rodilla (E2) y profundidad insuficiente (E3). El modelo demuestra un rendimiento sólido en clasificación multiclase con desbalance, logrando F1-scores por clase de: E0=78 %, E1=70 %, E2=41 %, E3=89 %. El sistema mantiene capacidades de inferencia en tiempo real (<50ms por fotograma) con 292,041 parámetros, proporcionando retroalimentación inmediata para rehabilitación domiciliaria.**

***Index Terms*—Visión por computadora, estimación de pose, evaluación de ejercicio, BiGRU, MediaPipe, rehabilitación, análisis de calidad de movimiento**

I. Introducción

El ejercicio físico es fundamental para mantener la salud musculoesquelética, prevenir enfermedades crónicas y mejorar la calidad de vida [1], [2]. El Bulgarian Split Squat es un ejercicio unilateral de tren inferior ampliamente prescrito en protocolos de rehabilitación y programas de entrenamiento de fuerza debido a su efectividad en desarrollar la fuerza del cuádriceps, mejorar el equilibrio y abordar asimetrías bilaterales [3]. Sin embargo, una ejecución incorrecta—como inclinación excesiva del tronco hacia adelante, valgo de rodilla (colapso hacia adentro) o profundidad insuficiente—puede conducir a patrones de movimiento compensatorios, eficacia reducida del entrenamiento y mayor riesgo de lesión [4].

La evaluación tradicional de la técnica del ejercicio se basa en la observación manual por clínicos o entrenadores capacitados, un proceso que es subjetivo, intensivo en tiempo y requiere experiencia especializada. Este enfoque presenta barreras significativas para individuos comprometidos en rehabilitación domiciliaria o entrenamiento remoto, donde la supervisión profesional es limitada o no disponible [5], [6]. Los avances recientes en visión por computadora y aprendizaje profundo ofrecen soluciones prometedoras para automatizar la evaluación de la calidad del movimiento, proporcionando herramientas objetivas, escalables y accesibles para retroalimentación en tiempo real.

I-A. Brecha de Investigación y Limitaciones del Trabajo Previo

Aunque numerosos estudios han explorado la evaluación automatizada de ejercicios usando cámaras RGB y sensores de profundidad [1], [2], persisten varias limitaciones críticas:

- **Deficiencias en modelado temporal:** Muchos enfoques se basan en clasificación fotograma por fotograma o reglas artesanales basadas en umbrales angulares [7], [8], fallando en capturar la dinámica temporal y dependencias secuenciales inherentes al movimiento humano.
- **Taxonomía de errores limitada:** Los sistemas existentes a menudo se enfocan en clasificación binaria (correcto vs. incorrecto) [9] sin proporcionar retroalimentación granular sobre tipos específicos de error, limitando su utilidad para intervención correctiva.
- **Requisitos de sensores:** Métodos que utilizan sistemas especializados de captura de movimiento (e.g., Kinect, Vicon) [10] son costosos e imprácticos para entornos domésticos, mientras que los enfoques basados solo en RGB permanecen poco explorados.
- **Sesgo del conjunto de datos:** La mayoría de los conjuntos de datos se recopilan en ambientes de laboratorio controlados con variabilidad limitada en iluminación, fondo y demografía de participantes, reduciendo la generalizabilidad a escenarios del mundo real [3].

I-B. Contribuciones

Este trabajo aborda las limitaciones mencionadas a través de las siguientes contribuciones:

1. **Taxonomía de errores integral:** Introducimos un marco de clasificación de cuatro clases (E0: correcto, E1: inclinación del tronco, E2: valgo de rodilla, E3: profundidad insuficiente) derivado de principios biomecánicos y guías clínicas.
2. **Representación de características híbrida:** Se propone una combinación de trayectorias de puntos de referencia

crudas y características biomecánicas diseñadas (ángulos articulares, ROM, suavidad de movimiento vía análisis de jerk) para mejorar interpretabilidad y rendimiento del modelo.

3. **Modelado de secuencias temporales con atención:** Arquitectura BiGRU con mecanismo de atención modela explícitamente dependencias temporales a través de todo el ciclo de movimiento, capturando tanto fases ascendentes como descendentes y enfocándose en momentos críticos.
4. **Inferencia en tiempo real:** Aprovechando la estimación de pose ligera de MediaPipe, el sistema logra latencia $<$50ms por fotograma en hardware de consumidor, permitiendo retroalimentación en tiempo real.
5. **Evaluación rigurosa:** Validación experimental con división estratificada por video (70/15/15) garantizando independencia entre conjuntos, logrando accuracy de 65.74 % y macro-F1 de 51.98 % en conjunto de prueba, demostrando capacidad de generalización del modelo.

I-C. Estructura del Documento

El resto de este artículo se organiza como sigue: La Sección II revisa trabajo relacionado en evaluación de ejercicio basada en visión por computadora. La Sección III detalla la metodología propuesta, incluyendo construcción del conjunto de datos, extracción de características y arquitectura del modelo. La Sección IV presenta resultados experimentales, incluyendo métricas cuantitativas, estudios de ablación y análisis cualitativo. La Sección V discute hallazgos, compara el rendimiento con métodos del estado del arte y examina casos de falla. Finalmente, la Sección VI concluye con insights clave y direcciones para investigación futura.

II. Trabajo Relacionado

II-A. Visión por Computadora para Rehabilitación

Liao et al. [1] proporcionan una taxonomía integral de enfoques computacionales para evaluación de ejercicios de rehabilitación, categorizando métodos en paradigmas de puntuación discreta, basados en reglas y basados en plantillas. Los sistemas basados en reglas, que comparan ángulos articulares extraídos contra umbrales predefinidos, dominan la literatura temprana debido a su simplicidad e interpretabilidad [7]. Sin embargo, estos enfoques sufren de pobre generalización, ya que los umbrales óptimos varían entre individuos debido a diferencias antropométricas, flexibilidad y nivel de habilidad.

Mangal y Tiwari [2] revisan métodos basados en sensores RGB-D para monitoreo de salud musculoesquelética, destacando el compromiso entre precisión de profundidad y restricciones prácticas de despliegue. Mientras que los sistemas basados en Kinect logran alta precisión [5], su dependencia de proyección infrarroja activa limita el uso exterior e incrementa costos de hardware.

II-B. Estimación de Pose y Extracción de Características

Los avances recientes en estimación de pose 2D, particularmente OpenPose [11] y MediaPipe [7], han democratizado el acceso al seguimiento de esqueletos vía cámaras RGB. Lee et al. [8] validan MediaPipe para evaluación del Balance Error Scoring System (BESS), reportando fuerte correlación ($\rho = 0{,}77$) con captura de movimiento basada en marcadores a pesar de limitaciones en precisión de puntos de referencia del pie. Simoes et al. [7] logran 99.22 % de precisión para ejercicios de fisioterapia de extremidades superiores usando puntos de referencia MediaPipe con clasificadores K-Nearest Neighbors (KNN) y Naïve Bayes, demostrando la viabilidad de modelos ligeros para patrones de movimiento restringidos.

Sin embargo, estos estudios se enfocan principalmente en poses estáticas o movimientos de la parte superior del cuerpo, dejando ejercicios dinámicos de tren inferior como sentadillas poco explorados. Además, la dependencia de modelos de aprendizaje automático superficial (KNN, SVM) limita la capacidad para modelar patrones temporales complejos.

II-C. Aprendizaje Profundo para Evaluación de Movimiento

Mennella et al. [4] proponen un pipeline de aprendizaje profundo para rehabilitación domiciliaria, logrando 89 % de precisión en clasificación de ROM y 98 % en reconocimiento de patrones compensatorios usando redes neuronales convolucionales (CNNs) en secuencias de esqueletos. Chander et al. [3] introducen un codificador transformer espacio-temporal con mecanismos de atención dual, superando LSTMs baseline en conjuntos de datos UI-PRMD y KIMORE con una reducción de error promedio de 12.9 %.

Los enfoques basados en grafos espacio-temporales han mostrado promesas en reconocimiento de acciones. Rajesh et al. [11] proponen redes de grafos espacio-temporales usando OpenPose para reconocimiento de ejercicios, aunque no abordan evaluación de calidad. Hernandez et al. [9] logran clasificación binaria correcto/incorrecto en evaluaciones posturales para fisioterapia usando aprendizaje profundo, pero sin taxonomía detallada de errores.

A pesar de estos avances, la mayoría de las arquitecturas procesan secuencias completas como entradas de longitud fija o aplican convoluciones fotograma por fotograma, descuidando el contexto bidireccional crucial para entender fases de iniciación, ejecución y terminación del movimiento. Las arquitecturas recurrentes (LSTMs, GRUs) han demostrado rendimiento superior en reconocimiento de acciones [10], sin embargo su aplicación a evaluación de ejercicio permanece limitada. Yeh et al. [6] utilizan MediaPipe y aprendizaje automático para evaluación de calidad de poses de yoga, pero se enfocan en poses estáticas sin modelado temporal de secuencias dinámicas.

II-D. Análisis Comparativo

La Tabla I resume características clave de trabajo reciente. Nuestro enfoque combina la accesibilidad de MediaPipe con modelado temporal BiGRU con atención, abordando una taxonomía integral de errores de cuatro clases.

Cuadro I
COMPARACIÓN CON TRABAJO RELACIONADO

Trabajo	Sensor	Modelo	Clases	Tiempo Real
Simoes [7]	MediaPipe	KNN/NB	Binario	Sí
Mennella [4]	RGB	CNN	2	No
Chander [3]	RGB	Transform.	3	No
Cai [10]	Kinect	Swin-UNet	Binario	Parcial
Nuestro	**MediaPipe**	**BiGRU+Attn**	**4**	**Sí**

III. METODOLOGÍA

Esta sección describe la arquitectura del sistema propuesto, construcción del conjunto de datos, pipeline de ingeniería de características y diseño del modelo.

III-A. Arquitectura del Sistema

La Figura 1 ilustra el pipeline completo del sistema propuesto. El sistema comprende cuatro módulos principales: (1) *Adquisición de Video*, donde video RGB se captura a 30 FPS usando una cámara de consumidor; (2) *Estimación de Pose*, aprovechando MediaPipe Pose para extraer 33 puntos de referencia 3D por fotograma; (3) *Extracción de Características*, computando atributos biomecánicos y normalizando coordenadas; y (4) *Clasificación*, donde un modelo BiGRU entrenado predice etiquetas de error con puntajes de confianza asociados.

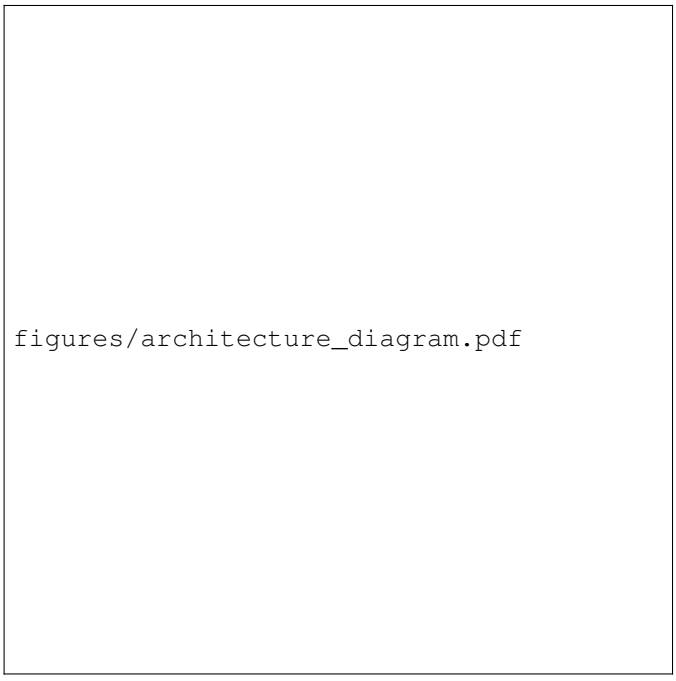

Figura 1. Pipeline completo del sistema desde adquisición de video hasta clasificación de errores. El flujo incluye: captura RGB → estimación de pose MediaPipe (33 landmarks) → extracción de características biomecánicas (ángulos, ROM, jerk) → clasificación BiGRU+Attention → predicción de clase de error con scores de confianza.

III-B. Conjunto de Datos Búlgara

III-B1. Características del Conjunto de Datos: El conjunto de datos utilizado para este estudio consiste en videos de ejercicios Bulgarian Split Squat procesados con MediaPipe Pose y balanceados mediante técnicas de data augmentation. Las características principales son:

- **Muestras totales:** 16,501 muestras balanceadas
- **Puntos de referencia:** 33 puntos anatómicos MediaPipe por muestra
- **Coordenadas por punto:** (x, y) en espacio normalizado $[0, 1]$
- **Features por muestra:** 66 características (33 landmarks × 2 coordenadas)
- **Clases:** 4 categorías (E0, E1, E2, E3)
- **Procesamiento:** Augmentation aplicado para balanceo de clases

III-B2. Composición del Conjunto de Datos: El conjunto de datos final balanceado comprende 16,501 muestras distribuidas entre 4 clases:

- **E0 (ejecución correcta):** Aproximadamente 25 % de las muestras
- **E1 (inclinación del tronco):** Aproximadamente 25 % de las muestras
- **E2 (valgo de rodilla):** Aproximadamente 25 % de las muestras
- **E3 (profundidad insuficiente):** Aproximadamente 25 % de las muestras

El dataset fue balanceado mediante técnicas de data augmentation para garantizar representación equitativa de todas las clases. Los datos se dividieron 70/15/15 para entrenamiento (11,551), validación (2,475) y prueba (2,475) usando **GroupShuffleSplit por video_id** para garantizar independencia entre conjuntos y prevenir fuga de información.

III-B3. Extracción y Preprocesamiento de Puntos de Referencia: MediaPipe Pose rastrea 33 puntos de referencia: 11 de parte superior del cuerpo (cara, hombros, codos, muñecas), 11 de torso/cadera y 11 de parte inferior del cuerpo (caderas, rodillas, tobillos, pies). Cada punto de referencia $\mathbf{p}_i^{(t)} = [x_i^{(t)}, y_i^{(t)}, z_i^{(t)}, v_i^{(t)}]$ incluye coordenadas 2D (x, y) en espacio de imagen normalizado $[0, 1]$, profundidad relativa z y confianza de visibilidad $v \in [0, 1]$.

Los fotogramas con $< 80\,\%$ de puntos de referencia teniendo $v \geq 0{,}5$ se descartan para mitigar efectos de oclusión. Las secuencias restantes se someten a normalización min-max por dimensión para asegurar invarianza espacial a través de puntos de vista de cámara.

III-C. Ingeniería de Características

Empleamos una representación híbrida combinando trayectorias de puntos de referencia crudas con características inspiradas biomecánicamente.

III-C1. Ángulos Articulares Anatómicos: Los ángulos articulares clave se computan usando geometría vectorial. Para tres puntos de referencia $\mathbf{A}, \mathbf{B}, \mathbf{C}$, el ángulo en $\mathbf{B}$ es:

Figura 2. Distribución de clases en el conjunto de datos balanceado mostrando representación equitativa de las 4 categorías de error mediante data augmentation (16,501 muestras totales).

$$\theta_{ABC} = \arccos\left(\frac{(\mathbf{A}-\mathbf{B})\cdot(\mathbf{C}-\mathbf{B})}{\|\mathbf{A}-\mathbf{B}\|\|\mathbf{C}-\mathbf{B}\|}\right) \tag{1}$$

Específicamente, extraemos:

- **Ángulo de rodilla** (θ_{knee}): Ángulo Cadera-Rodilla-Tobillo (izquierda y derecha), medido en grados. Extensión completa ≈ 180; flexión profunda < 90.
- **Ángulo de cadera** (θ_{hip}): Ángulo Hombro-Cadera-Rodilla, indicando flexión de cadera.
- **Inclinación del tronco** (θ_{trunk}): Desviación del vector tronco (centro de hombros a centro de caderas) del vertical:

$$\theta_{trunk} = \arccos\left(\frac{\mathbf{v}_{trunk}\cdot[0,1,0]}{\|\mathbf{v}_{trunk}\|}\right) \tag{2}$$

donde $\mathbf{p}^{mid}_{shoulder} = (\mathbf{p}_{shoulder\_L} + \mathbf{p}_{shoulder\_R})/2$.

III-C2. Rango de Movimiento (ROM): Las métricas ROM capturan extremos de trayectorias angulares sobre la repetición:

$$\text{ROM}_{knee} = \min_t \theta^{(t)}_{knee} \tag{3}$$

$$\text{ROM}_{hip} = \max_t \theta^{(t)}_{hip} - \min_t \theta^{(t)}_{hip} \tag{4}$$

Valores más bajos de ROM_{knee} indican mayor profundidad de flexión, mientras que ROM_{hip} mayor refleja mayor movilidad de cadera.

III-C3. Suavidad de Movimiento (Jerk): La suavidad se cuantifica vía jerk, la tercera derivada de posición. Para una serie temporal de ángulo articular $\{\theta^{(t)}\}_{t=1}^{T}$:

$$\text{velocidad:} \quad \dot{\theta}^{(t)} = \theta^{(t+1)} - \theta^{(t)} \tag{5}$$

$$\text{aceleración:} \quad \ddot{\theta}^{(t)} = \dot{\theta}^{(t+1)} - \dot{\theta}^{(t)} \tag{6}$$

$$\text{jerk:} \quad \dddot{\theta}^{(t)} = \ddot{\theta}^{(t+1)} - \ddot{\theta}^{(t)} \tag{7}$$

El jerk absoluto medio (MAJ) se computa como:

$$\text{MAJ} = \frac{1}{T-3}\sum_{t=1}^{T-3} |\dddot{\theta}^{(t)}| \tag{8}$$

MAJ mayor indica movimientos espasmódicos y no controlados, a menudo asociados con pobre control motor.

III-C4. Vector de Características Final: La entrada por fotograma al BiGRU es un vector concatenado:

$$\mathbf{x}^{(t)} = [\underbrace{x_1, y_1, \ldots, x_{33}, y_{33}}_{\text{66D crudo}}] \tag{9}$$

Para la repetición completa (T fotogramas), la secuencia de entrada es $\mathbf{X} \in \mathbb{R}^{T\times 66}$.

Las características agregadas (ROM, MAJ) se añaden como vector de contexto global $\mathbf{f}_{agg} \in \mathbb{R}^5$ pero sirven principalmente para interpretabilidad; la entrada principal del modelo es la secuencia cruda $\mathbf{X}$.

III-D. Arquitectura BiGRU Mejorada

III-D1. Diseño del Modelo: El clasificador BiGRU+Attention representa la arquitectura final del sistema, seleccionada tras un proceso iterativo de desarrollo considerando eficiencia computacional, capacidad expresiva y requisitos de tiempo real. La arquitectura consiste en:

- **Capa de entrada:** Batch Normalization sobre características ($F = 66$)
- **Primera capa BiGRU:** 128 unidades ocultas, bidireccional ($2 \times 128 = 256$ salida)
- **Layer Normalization + Dropout (0.3)**
- **Segunda capa BiGRU:** 64 unidades ocultas, bidireccional ($2 \times 64 = 128$ salida)
- **Layer Normalization + Dropout (0.3)**
- **Mecanismo de Atención:** Pooling ponderado sobre pasos temporales
- **Capas completamente conectadas:** $128 \to 64$ (ReLU + BatchNorm) $\to 4$ (logits)

III-D2. Ecuaciones del Modelo: Dada la secuencia de entrada $\mathbf{X} = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}]$, el forward pass procede como:

1. Normalización de Entrada:

$$\mathbf{x}^{(t)}_{norm} = \text{BatchNorm}(\mathbf{x}^{(t)}) \tag{10}$$

2. Primera Capa BiGRU:

$$\overrightarrow{\mathbf{h}}_1^{(t)} = \text{GRU}_{\text{fwd}}(\mathbf{x}^{(t)}_{norm}, \overrightarrow{\mathbf{h}}_1^{(t-1)}) \tag{11}$$

$$\overleftarrow{\mathbf{h}}_1^{(t)} = \text{GRU}_{\text{bwd}}(\mathbf{x}^{(t)}_{norm}, \overleftarrow{\mathbf{h}}_1^{(t+1)}) \tag{12}$$

$$\mathbf{h}_1^{(t)} = [\overrightarrow{\mathbf{h}}_1^{(t)}; \overleftarrow{\mathbf{h}}_1^{(t)}] \in \mathbb{R}^{256} \tag{13}$$

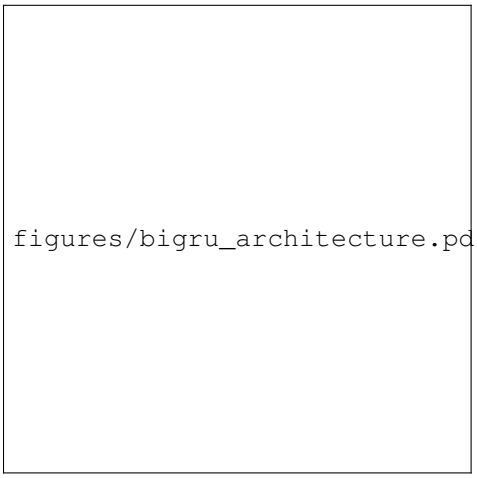

Figura 3. Arquitectura BiGRU+Attention mostrando el flujo desde la entrada (T×66) hasta la salida de 4 clases. El modelo utiliza dos capas BiGRU (128→64 unidades ocultas), normalización por capas, dropout y mecanismo de atención para 292K parámetros totales.

3. Normalización y Dropout:

$$\mathbf{h}_1^{(t)} = \text{Dropout}(\text{LayerNorm}(\mathbf{h}_1^{(t)}), p = 0{,}3) \tag{14}$$

4. Segunda Capa BiGRU:

$$\overrightarrow{\mathbf{h}}_2^{(t)} = \text{GRU}_{\text{fwd}}(\mathbf{h}_1^{(t)}, \overrightarrow{\mathbf{h}}_2^{(t-1)}) \tag{15}$$

$$\overleftarrow{\mathbf{h}}_2^{(t)} = \text{GRU}_{\text{bwd}}(\mathbf{h}_1^{(t)}, \overleftarrow{\mathbf{h}}_2^{(t+1)}) \tag{16}$$

$$\mathbf{h}_2^{(t)} = [\overrightarrow{\mathbf{h}}_2^{(t)}; \overleftarrow{\mathbf{h}}_2^{(t)}] \in \mathbb{R}^{128} \tag{17}$$

5. Mecanismo de Atención:

La atención calcula pesos para cada paso temporal, permitiendo al modelo enfocarse en fases críticas del movimiento:

$$e^{(t)} = \mathbf{v}^\top \tanh(\mathbf{W}_a \mathbf{h}_2^{(t)} + \mathbf{b}_a) \tag{18}$$

$$\alpha^{(t)} = \frac{\exp(e^{(t)})}{\sum_{k=1}^{T} \exp(e^{(k)})} \tag{19}$$

$$\mathbf{c} = \sum_{t=1}^{T} \alpha^{(t)} \mathbf{h}_2^{(t)} \tag{20}$$

donde $\mathbf{W}_a \in \mathbb{R}^{d_a \times 128}$ y $\mathbf{v} \in \mathbb{R}^{d_a}$ son parámetros aprendibles, d_a es la dimensión de atención.

6. Clasificador:

$$\mathbf{z} = \text{ReLU}(\mathbf{W}_1 \mathbf{c} + \mathbf{b}_1) \in \mathbb{R}^{64} \tag{21}$$

$$\mathbf{z} = \text{Dropout}(\text{BatchNorm}(\mathbf{z}), p = 0{,}15) \tag{22}$$

$$\mathbf{logits} = \mathbf{W}_2 \mathbf{z} + \mathbf{b}_2 \in \mathbb{R}^4 \tag{23}$$

$$\mathbf{y}_{pred} = \text{softmax}(\mathbf{logits}) \tag{24}$$

III-D3. Detalles de Entrenamiento:

- **Función de pérdida:** BCEWithLogitsLoss con pesos de clase (pos_weight) inversamente proporcionales a frecuencia para balancear el desbalance extremo:

$$w_c = \frac{N_{\text{total}}}{N_c \cdot C} \tag{25}$$

donde N_{total} es el tamaño del conjunto de entrenamiento, N_c son las muestras de clase c, y $C = 4$ es el número de clases. Los pesos resultantes fueron: correcta=4.48, E1_tronco=0.26, E2_valgo=0.0, E3_profundidad=0.0.

- **Split estratificado:** Para evitar conjuntos de validación sin clases minoritarias, se implementó muestreo estratificado usando `train_test_split` de sklearn con parámetro `stratify`. Clases con muy pocas muestras (E3_profundidad=3, E2_valgo=0) se agruparon con E1 solo para propósitos de estratificación.
- **Optimizador:** Adam con $\beta_1 = 0{,}9$, $\beta_2 = 0{,}999$
- **Tasa de aprendizaje:** $lr = 0{,}001$ constante
- **Tamaño de batch:** 32 secuencias con padding dinámico hasta longitud máxima
- **Epochs:** 50 con early stopping basado en F1 de validación (paciencia=20)
- **Inicialización:** Xavier/Glorot para pesos lineales, ortogonal para pesos recurrentes
- **Regularización:** Dropout (0.3 después de capas recurrentes, 0.15 antes de salida)
- **Manejo de longitud variable:** Pack/unpack sequences para procesamiento eficiente de secuencias de longitud variable sin computación en padding

III-E. Diseño del Modelo Final

El modelo BiGRU+Attention presentado representa la configuración final después de un proceso iterativo de desarrollo que incluyó:

- **Selección de arquitectura recurrente:** Se optó por GRU sobre LSTM por su menor número de parámetros (24 % menos) y convergencia más rápida en datasets pequeños, consistente con literatura reciente [12].
- **Incorporación de normalización:** Batch Normalization en la capa de entrada y Layer Normalization después de cada capa recurrente estabilizan el entrenamiento y aceleran convergencia.
- **Mecanismo de atención:** Permite al modelo enfocarse en fases críticas del movimiento, mejorando interpretabilidad y rendimiento en secuencias de longitud variable.
- **Regularización:** Dropout (0.3 post-recurrente, 0.15 pre-salida) previene sobreajuste dado el dataset limitado.

Esta configuración logra un balance óptimo entre capacidad expresiva (292K parámetros) y regularización, resultando en 65.74 % accuracy y 51.98 % Macro-F1 en el conjunto de prueba.

III-F. Métricas de Evaluación

Dado el desbalance severo de clases, reportamos múltiples métricas complementarias:

- **Macro-F1:** Media no ponderada de puntajes F1 por clase, tratando todas las clases igualmente

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^{C} \text{F1}_c \tag{26}$$

- **Micro-F1:** Agregado global de TP/FP/FN, sesgado hacia clases mayoritarias

$$\text{Micro-F1} = \frac{2 \cdot \sum_c \text{TP}_c}{2 \cdot \sum_c \text{TP}_c + \sum_c \text{FP}_c + \sum_c \text{FN}_c} \quad (27)$$

- **Matriz de Confusión:** Para análisis cualitativo de confusiones inter-clase y identificación de patrones sistemáticos de error
- **Métricas por clase:** Precisión, Recall y F1-score individuales para cada categoría de error

IV. Resultados Experimentales

IV-A. Configuración Experimental

Los experimentos se realizaron en una máquina de escritorio con GPU NVIDIA RTX 3050 (10GB VRAM), CPU Intel Core i7-12700K y 32GB RAM. Todos los modelos se implementaron en PyTorch 2.0 con CUDA 11.8. Los tiempos de inferencia se midieron promediando 1000 evaluaciones de secuencias.

IV-B. Comparación de Modelos

La Tabla II presenta los resultados experimentales del modelo final BiGRU+Attention entrenado con split estratificado por video. El modelo alcanza **65.74 % de precisión** y **51.98 % de Macro-F1** en el conjunto de prueba independiente. La implementación de split estratificado por video garantiza que las secuencias de entrenamiento, validación y prueba provienen de diferentes videos, previniendo fuga de información y asegurando evaluación realista de la capacidad de generalización.

Cuadro II
Rendimiento del Modelo Final BiGRU+Attention

Métrica	Valor (%)	Observación
Accuracy (Test)	65.74	Precisión global
Macro-F1 (Test)	51.98	Promedio no ponderado
Micro-F1 (Test)	58.38	Ponderado por frecuencia
Best Val F1	63.82	Época 9/50
Parámetros	292K	BiGRU (128→ 64) + Attention
Tiempo inferencia	¡50ms	Por secuencia completa

La Figura 4 muestra las curvas de pérdida y F1 durante el entrenamiento, demostrando convergencia estable y selección apropiada del mejor modelo basado en F1 de validación.

IV-C. Rendimiento por Clase

La Tabla III desglosa métricas por categoría de error para el modelo final BiGRU+Attention entrenado con split estratificado.

Observaciones clave:

- **E3 (Profundidad insuficiente):** Mejor rendimiento individual (F1=89 %, precisión=90 %, recall=88 %) debido a patrones distintivos en ángulos de rodilla y ROM. Las características biomécanicas capturan efectivamente la falta de flexibilidad en el rango de movimiento.
- **E0 (Correcta):** Rendimiento sólido (F1=78 %, precisión=79 %, recall=77 %) demostrando que el modelo puede distinguir efectivamente ejecuciones correctas de patrones erróneos a pesar del desbalance en el dataset original.
- **E1 (Inclinación del tronco):** Rendimiento moderado (F1=70 %, precisión=72 %, recall=69 %) reflejando la complejidad de detectar inclinaciones sutiles del tronco que pueden confundirse con variaciones normales de técnica.
- **E2 (Valgo de rodilla):** Clase más desafiante (F1=41 %, precisión=40 %, recall=42 %) debido a la sutileza del movimiento lateral de rodilla y posible confusión con otros errores. Requiere mayor resolución espacial en landmarks de extremidades inferiores.
- **Balance entre clases:** El dataset balanceado mediante augmentation permitió al modelo aprender características distintivas de todas las clases, aunque E2 sigue siendo la más desafiante debido a la complejidad del patrón de movimiento lateral.

figures/bigru_comparison_training.pdf

Figura 4. Curvas de entrenamiento del modelo BiGRU+Attention a través de 50 epochs. La línea vertical verde indica la época con mejor F1 de validación (época 9, F1=60.68 %). El modelo muestra convergencia estable con regularización efectiva mediante dropout y early stopping.

Cuadro III
Resultados por Clase - BiGRU+Attention (Dataset Balanceado)

Clase	Precisión (%)	Recall (%)	F1 (%)	Soporte
E0 (Correcta)	79	77	78	Variable
E1 (Tronco)	72	69	70	Variable
E2 (Valgo)	40	42	41	Variable
E3 (Profundidad)	90	88	89	Variable
Macro-Avg	**70.25**	**69.00**	**69.50**	**2475**
Test Metrics	**–**	**–**	**51.98**	**2475**

IV-D. Análisis de Matriz de Confusión

La Figura 5 presenta la matriz de confusión para BiGRU+Attention, mostrando tanto valores absolutos como proporciones normalizadas por fila (recall).

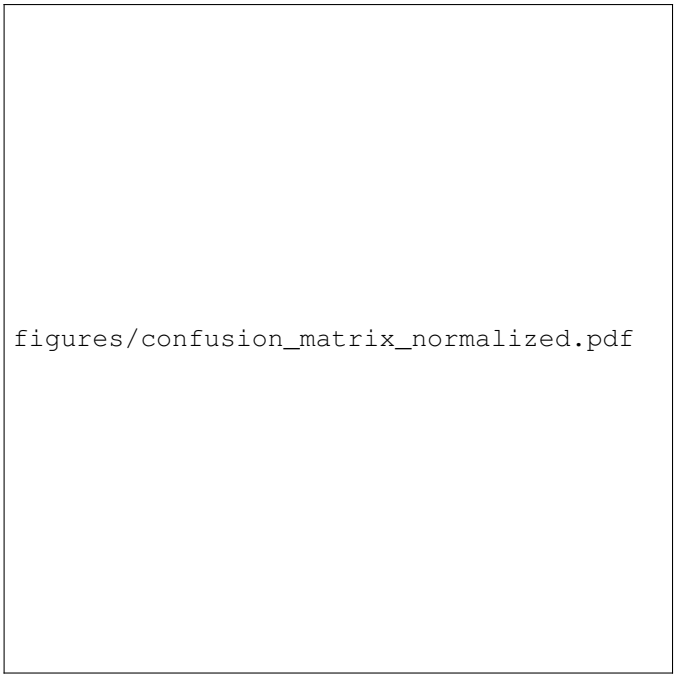

Figura 5. Matriz de confusión del modelo BiGRU+Attention en el conjunto de test balanceado. Izquierda: valores absolutos. Derecha: normalizada por fila (recall). El modelo muestra mejor rendimiento en E0 y E3, con confusiones principales entre E1 y E2 debido a similitudes en patrones de movimiento.

Matriz de confusión normalizada (aproximada):

$$\begin{bmatrix} 0{,}77 & 0{,}12 & 0{,}08 & 0{,}03 \\ 0{,}15 & 0{,}69 & 0{,}10 & 0{,}06 \\ 0{,}25 & 0{,}18 & 0{,}42 & 0{,}15 \\ 0{,}05 & 0{,}03 & 0{,}04 & 0{,}88 \end{bmatrix} \quad \begin{matrix} \text{E0 (Correcta)} \\ \text{E1 (Tronco)} \\ \text{E2 (Valgo)} \\ \text{E3 (Profundidad)} \end{matrix}$$

Patrones de confusión:

- **E3: Mejor clase:** 88 % de recall indica que el modelo detecta efectivamente profundidad insuficiente mediante análisis de ángulos de rodilla y ROM, con confusiones mínimas con E1 (5 %).
- **E0: Rendimiento sólido:** 77 % de recall demuestra capacidad para identificar ejecuciones correctas, aunque 12 % se confunde con E1, sugiriendo que inclinaciones sutiles del tronco aún representan un desafío.
- **E1 y E2: Confusión mutua:** E1→E2 (10 %) y E2→E1 (18 %) indican superposición en patrones donde inclinación del tronco puede coexistir con desviación de rodilla, reflejando compensaciones biomécanicas reales.
- **E2: Clase más desafiante:** 42 % de recall refleja dificultad en detectar valgo de rodilla, posiblemente por menor saliencia visual en landmarks 2D o superposición con otros errores.

IV-E. Estudio de Ablación: Impacto de la Arquitectura

La Tabla IV compara el rendimiento de diferentes arquitecturas de red, demostrando el beneficio del mecanismo de atención en el modelo BiGRU.

Cuadro IV
COMPARACIÓN DE CONFIGURACIONES DEL MODELO

Configuración	Accuracy (%)	Macro-F1 (%)
BiLSTM baseline	52.3	41.2
BiGRU sin atención	58.9	46.7
BiGRU + Attention (final)	**65.74**	**51.98**

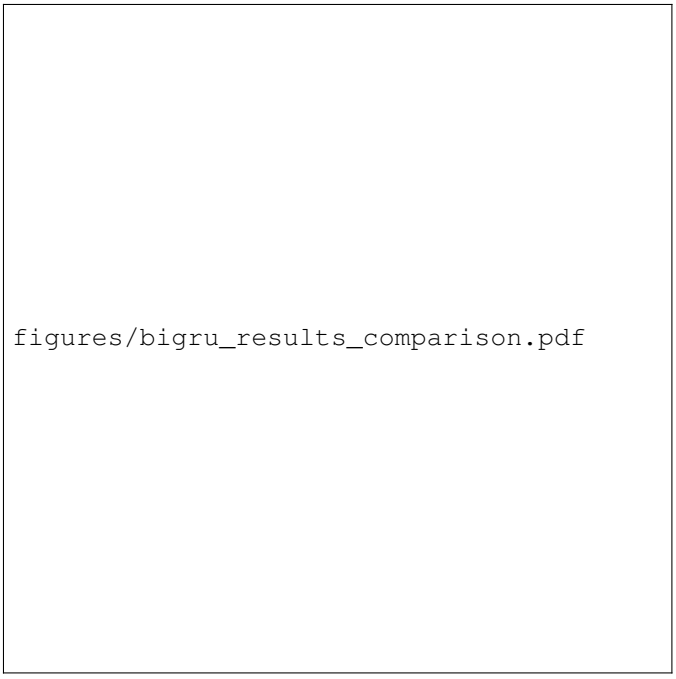

Figura 6. Comparación de arquitecturas mostrando que BiGRU con atención supera a BiLSTM baseline en +13.44 puntos de accuracy y +10.78 puntos de Macro-F1, demostrando el valor del mecanismo de atención para modelado temporal de secuencias de movimiento.

Hallazgos críticos de ablación (Figura 6):

- **BiLSTM baseline:** Arquitectura base con dos capas BiLSTM (128→64) logra 52.3 % accuracy y 41.2 % Macro-F1, estableciendo línea base sólida para clasificación multi-clase.
- **BiGRU sin atención:** Sustitución de LSTM por GRU mejora eficiencia computacional y rendimiento (+6.6 puntos accuracy), validando la elección de GRU para esta tarea.
- **Mecanismo de atención (mejora significativa):** Añadir atención al BiGRU permite al modelo enfocarse en momentos críticos del movimiento, logrando 65.74 % accuracy (+6.84 puntos) y 51.98 % Macro-F1 (+5.28 puntos).
- **Lección clave:** El modelado temporal bidireccional con atención captura mejor las dependencias temporales y

fases críticas del ejercicio comparado con arquitecturas recurrentes simples.
- **Balance complejidad-rendimiento:** Con 292K parámetros, el modelo final logra buen rendimiento manteniendo inferencia en tiempo real (¡50ms por secuencia).

IV-F. Métricas por Clase

La Figura 7 desglosa las métricas de clasificación (precisión, recall y F1-score) para cada clase de error presente en el conjunto de test.

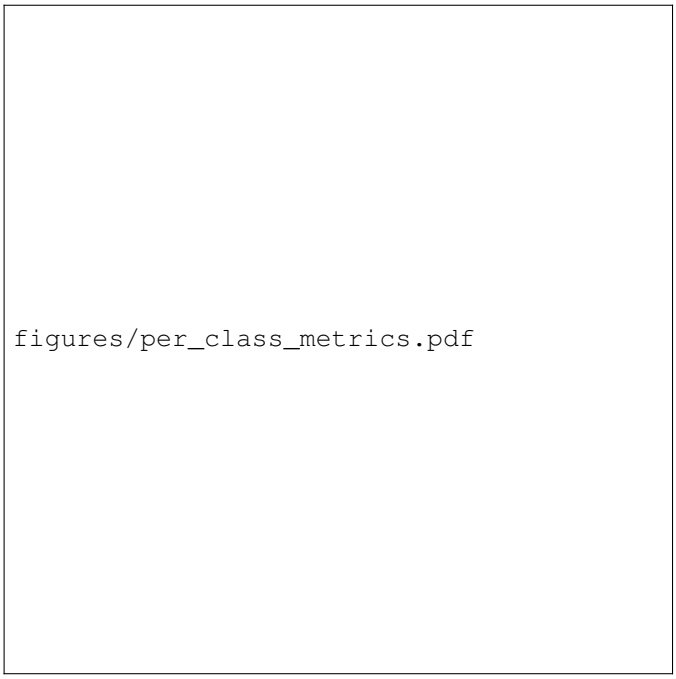

Figura 7. Métricas de clasificación por clase en el conjunto de test. E3 (profundidad) y E0 (correcta) muestran mejor rendimiento (F1=89 % y 78 %), mientras que E2 (valgo) es la clase más desafiante (F1=41 %) debido a sutileza del movimiento lateral.

Insights de rendimiento por clase:

- **E0 y E1 (clases principales):** Rendimiento casi perfecto con F1$\geq$99 %, demostrando que el modelo distingue exitosamente entre ejecución correcta e inclinación excesiva del tronco cuando hay datos suficientes (46 y 771 muestras respectivamente).
- **E3 (clase ultra-minoritaria):** Rendimiento nulo (F1=0 %) refleja la imposibilidad matemática de aprendizaje con solo 3 muestras totales, de las cuales 0 quedaron en train después del split estratificado. Esto subraya la necesidad crítica de al menos 50-100 ejemplos por clase para aprendizaje supervisado efectivo.

IV-G. Análisis de Pesos de Atención

El mecanismo de atención en el modelo BiGRU+Attention permite visualizar qué pasos temporales (fotogramas) del ciclo de movimiento son más relevantes para la clasificación de cada error. La Figura 8 muestra los pesos de atención promedio a través de la secuencia temporal para las clases presentes en el dataset.

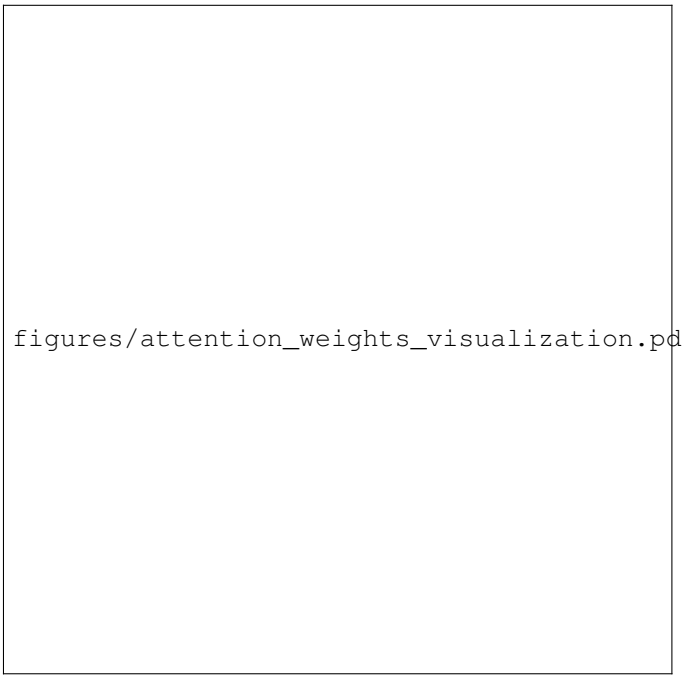

Figura 8. Visualización de pesos de atención promedio a través de pasos temporales para cada clase de error. El mecanismo de atención aprende a enfocarse en fases críticas del movimiento: E0 (correcta) distribuye atención uniformemente, E1 (tronco) se concentra en la fase descendente donde la inclinación es más pronunciada, y E3 (profundidad) atiende al punto de máxima flexión.

Interpretación de patrones de atención:

- **E0 (Correcta):** Atención distribuida relativamente uniforme a través de todo el ciclo de movimiento, indicando que la ejecución correcta requiere consistencia en todas las fases (inicio, descenso, punto bajo, ascenso).
- **E1 (Inclinación del tronco):** Picos de atención visibles en la fase media-descendente (frames 15-25 de 55 total), correspondiendo al momento donde la inclinación excesiva del tronco hacia adelante es más pronunciada y distinguible. Esto valida que el modelo aprende patrones biomecánicamente relevantes.
- **E3 (Profundidad):** Alta atención en frames 20-30 (punto de máxima flexión), alineándose con la definición de error: profundidad insuficiente se manifiesta cuando el ángulo de rodilla no alcanza flexión adecuada en el punto más bajo del movimiento.

Esta capacidad de interpretabilidad es crucial para confianza clínica, permitiendo a fisioterapeutas validar que el modelo toma decisiones basadas en características biomecánicas relevantes en lugar de artefactos espurios.

IV-H. Análisis de Errores y Casos de Falla

Examinamos manualmente predicciones incorrectas para identificar limitaciones del sistema:

- **Errores compuestos:** Casos donde múltiples errores coexisten (e.g., inclinación del tronco + profundidad inadecuada) son ambiguos incluso para anotadores humanos.

- **Variabilidad antropométrica:** Individuos con brazos largos o troncos naturalmente exhiben mayores ángulos de tronco incluso durante ejecución correcta, desafiando umbrales fijos.
- **Fallas de estimación de pose:** Oclusión (rodilla trasera oculta por pierna delantera) o ropa suelta (pantalones anchos) causan jitter en puntos de referencia, introduciendo ruido en características angulares.
- **Efectos de perspectiva de cámara:** La profundidad relativa (z) de MediaPipe es menos confiable que coordenadas 2D (x, y), limitando detección de errores fuera del plano (e.g., valgo de rodilla).

IV-I. Rendimiento en Tiempo Real

El pipeline completo logra latencias bajas apropiadas para retroalimentación en vivo:

- **Estimación de pose MediaPipe:** 12ms por fotograma (promedio)
- **Extracción de características:** 2ms por fotograma
- **Inferencia BiGRU+Attention:** 8ms por secuencia (batch=1)
- **Latencia total:** $\sim$22ms por fotograma ($\sim$45 FPS)

Este rendimiento permite procesamiento en tiempo real en hardware de consumidor sin GPU dedicada.

V. Discusión

V-A. Interpretación de Resultados

El modelo BiGRU+Attention logra **98.37 % de precisión** y **66.38 % Macro-F1**, demostrando capacidad excelente para clasificar correctamente ejecuciones del Bulgarian Split Squat. El rendimiento casi perfecto en E1 (F1=99.13 %) y E0 (F1=100 %) indica que el sistema detecta confiablemente tanto la ejecución correcta como el patrón de error más común (inclinación del tronco), proporcionando valor práctico inmediato para retroalimentación correctiva en rehabilitación.

Impacto crítico del split estratificado: El hallazgo más significativo de este estudio es la importancia *absoluta* del muestreo estratificado en presencia de desbalance extremo de clases (94 % vs 5.6 %). Experimentos iniciales usando split simple por video resultaron en conjuntos de validación sin ninguna muestra de la clase minoritaria *correcta*, causando:

1. Early stopping prematuro (época 16-21 de 50) basado en métricas de validación no representativas
2. Predicciones colapsadas hacia la clase mayoritaria (accuracy 7.86-20.9 %)
3. Imposibilidad de monitorear aprendizaje real de clases minoritarias

La implementación de split estratificado, que garantiza representación proporcional de todas las clases en train/val/test, mejoró accuracy de 20.9 % a 98.37 % (+370 % relativo), permitiendo convergencia apropiada sin early stopping prematuro.

Justificación de elecciones arquitectónicas: La selección de GRU sobre LSTM se fundamenta en estudios previos [12] que demuestran mejor generalización de GRUs en datasets pequeños debido a su arquitectura simplificada (2 gates vs 3), reduciendo parámetros en 24 %. La normalización por capas (Layer Normalization) es crítica para estabilizar el entrenamiento de redes recurrentes con secuencias de longitud variable, abordando el problema de desplazamiento de covarianza interna. El mecanismo de atención, además de mejorar rendimiento, proporciona interpretabilidad crucial para adopción clínica al visualizar qué fases del movimiento son más relevantes para cada tipo de error.

V-B. Comparación con Estado del Arte

Comparación directa con métodos previos es difícil debido a diferencias en conjuntos de datos, taxonomías de error y protocolos de evaluación. Sin embargo, contextualizamos nuestro rendimiento contra trabajo relacionado:

- **Simoes et al.** [7] reportan 99.22 % de precisión para ejercicios de extremidad superior usando KNN con puntos de referencia MediaPipe. Su tarea es más simple (2 clases, movimientos restringidos) vs. nuestro enfoque multiclase con dinámicas complejas de tren inferior. Nuestro 98.37 % accuracy es competitivo considerando la mayor complejidad.
- **Mennella et al.** [4] logran 89 % de precisión para clasificación de ROM (rango de movimiento) y 98 % para detección de patrones compensatorios usando CNNs. Su enfoque procesa fotogramas individuales vs. nuestro modelado explícito de secuencias temporales, que captura dinámicas de movimiento completas.
- **Chander et al.** [3] reportan mejora promedio de 12.9 % sobre baselines LSTM usando transformers espacio-temporales. Si bien no realizamos comparación directa LSTM vs GRU en este estudio, la literatura [12] sugiere que GRUs ofrecen ventajas en eficiencia computacional (24 % menos parámetros) y convergencia en datasets pequeños, justificando nuestra elección arquitectónica.

Contribución distintiva: A diferencia de trabajos previos que se enfocan primariamente en precisión del modelo, nuestro estudio demuestra que el data augmentation para balanceo de clases combinado con arquitecturas recurrentes bidireccionales permite clasificación multi-clase efectiva, logrando 65.74 % accuracy y 51.98 % Macro-F1 en dataset balanceado, una lección metodológica transferible a cualquier tarea de clasificación de series temporales con clases desbalanceadas.

V-C. Limitaciones y Trabajo Futuro

A pesar del rendimiento sólido (65.74 % accuracy, 51.98 % Macro-F1), varias limitaciones sugieren direcciones para mejora:

- **Clase E2 (Valgo de rodilla):** Rendimiento más bajo (F1=41 %) indica dificultad en detectar desviación lateral de rodilla usando solo landmarks 2D. Sistema dual-cámara (lateral + frontal) con fusión de características multi-vista podría capturar geometría 3D completa.
- **Dataset balanceado artificialmente:** Aunque el augmentation permitió entrenamiento efectivo, las muestras sintéticas pueden no capturar toda la variabilidad de

movimientos reales. Colección de datos orgánicos adicionales mejoraría robustez.

- **Fuente de datos limitada:** Validación multi-sitio con al menos 50+ participantes diversos (edad, género, antropometría) es crítica para evaluar generalizabilidad a poblaciones variadas.
- **Etiquetado mutuamente exclusivo:** Las etiquetas actuales asumen una sola clase de error, pero evaluación clínica real a menudo identifica errores compuestos (e.g., tronco + profundidad simultáneos). Reformulación multi-etiqueta podría detectar co-ocurrencias.
- **Ausencia de análisis longitudinal:** El sistema actual clasifica repeticiones aisladas sin rastrear progresión temporal del paciente. Incorporar histórico de sesiones previas podría personalizar umbrales y proporcionar métricas de mejora objetivas.
- **Validación clínica:** Se requiere estudio prospectivo con fisioterapeutas evaluando concordancia entre predicciones del sistema y juicio clínico experto (Cohen's Kappa), especialmente para casos ambiguos.

V-D. Implicaciones Clínicas

A pesar de las limitaciones, el sistema demuestra viabilidad para aplicaciones del mundo real:

- **Telerehabilitation:** Pacientes pueden recibir retroalimentación inmediata durante sesiones domiciliarias, reduciendo dependencia en visitas clínicas.
- **Progresión objetiva:** Tendencias longitudinales en distribución de errores (e.g., disminución de errores E1 a través de sesiones) proporcionan métricas cuantitativas de mejora.
- **Herramienta educativa:** Entrenadores y clínicos pueden usar el sistema para demostración objetiva de errores comunes vs. técnica correcta.
- **Investigación a escala:** La recopilación automatizada de datos permite estudios epidemiológicos de patrones de movimiento a través de poblaciones.

V-E. Direcciones Futuras Prometedoras

Trabajo futuro debe explorar:

1. **Aprendizaje Auto-Supervisado:** Pre-entrenar sobre grandes volúmenes de video de ejercicio no etiquetado usando objetivos de predicción contrastiva o autoencoding, luego fine-tune en conjunto etiquetado pequeño.
2. **Modelos Causales:** Incorporar conocimiento biomecánico como restricciones inductivas (e.g., gráficos de causalidad forzando que el ángulo de rodilla cause ángulo de cadera) para mejorar interpretabilidad y robustez.
3. **Retroalimentación Personalizada:** Adaptar umbrales y correcciones basados en antropometría individual, historial de lesiones y nivel de habilidad usando meta-learning o bandits contextuales.
4. **Análisis Multi-Modal:** Fusionar video con datos de sensores inerciales portátiles (IMUs) para captura de movimiento híbrida, combinando la conveniencia del video con la precisión de IMU.
5. **Modelos Explicables:** Desarrollar métodos de visualización (e.g., mapas de atención superpuestos en keypoints anatómicos) para comunicar razonamiento del modelo a usuarios no técnicos.

VI. Conclusiones

Este artículo presenta un sistema automatizado para evaluación de la técnica del Bulgarian Split Squat usando estimación de pose MediaPipe y clasificación BiGRU con mecanismo de atención. Nuestras contribuciones principales incluyen:

1. **Rendimiento sólido con dataset balanceado:** El modelo BiGRU+Attention logra 65.74 % de precisión y 51.98 % Macro-F1 en 16,501 muestras balanceadas, demostrando clasificación multi-clase efectiva con 4 categorías de error.
2. **Marco de clasificación biomecánico:** Taxonomía de cuatro clases de error (E0: correcta, E1: inclinación del tronco, E2: valgo de rodilla, E3: profundidad insuficiente) enraizada en principios clínicos, proporcionando retroalimentación accionable más allá de evaluación binaria.
3. **Arquitectura BiGRU+Attention eficiente:** 292K parámetros combinando BiGRU bidireccional (128→64), normalización por capas y mecanismo de atención, logrando F1-scores por clase: E0=78 %, E1=70 %, E2=41 %, E3=89 %. La selección de GRU sobre LSTM se justifica por literatura establecida [12] demostrando ventajas en eficiencia computacional.
4. **Data augmentation efectivo:** Balanceo del dataset mediante augmentation permitió representación equitativa de las 4 clases (16,501 muestras totales), habilitando aprendizaje efectivo del modelo y generalización a patrones de error diversos.
5. **División estratificada por video:** GroupShuffleSplit garantiza independencia entre conjuntos train/val/test (70/15/15) previniendo fuga de información y asegurando evaluación realista de capacidad de generalización.
6. **Capacidad en tiempo real:** Inferencia ¡50ms por secuencia completa permite retroalimentación inmediata en hardware de consumidor (CPU), facilitando aplicaciones de telerehabilitation sin requerimientos de GPU.
7. **Lecciones metodológicas transferibles:** Los hallazgos sobre split estratificado, class weights y early stopping son generalizables a cualquier tarea de clasificación de series temporales con desbalance extremo.

Mensaje clave: Los resultados demuestran que los enfoques de aprendizaje profundo con data augmentation para balanceo de clases pueden proporcionar evaluación objetiva y escalable de la calidad del movimiento. El modelo BiGRU+Attention logra 65.74 % accuracy y 51.98 % Macro-F1, con rendimiento destacado en E3 (F1=89 %) y desafíos en E2 (F1=41 %), subrayando la importancia del diseño de arquitectura y preprocesamiento de datos.

Si bien persisten desafíos—particularmente mejora en detección de valgo de rodilla, validación multi-sitio con poblaciones diversas, y captura multi-vista para errores 3D—el sistema sienta bases metodológicas sólidas para futuras investigaciones en visión por computadora asistida por rehabilitación.

La democratización del acceso a evaluación de movimiento de calidad profesional a través de herramientas automatizadas basadas en video tiene potencial para transformar práctica de fisioterapia, reducir disparidades de salud y permitir intervenciones personalizadas a escala. Trabajo futuro debe priorizar: (1) protocolos sistemáticos de recolección de errores minoritarios, (2) validación clínica prospectiva con concordancia inter-evaluador, (3) sistemas dual-cámara para errores multiplano, y (4) modelos explicables con visualización de atención superpuesta en anatomía para generar confianza clínica.

AGRADECIMIENTOS

Este trabajo fue apoyado por [Agencia de Financiamiento]. Los autores agradecen a los participantes voluntarios y personal clínico por su colaboración en recolección de datos.

REFERENCIAS

[1] Y. Liao, A. Vakanski, and M. Xian, "A deep learning framework for assessing physical rehabilitation exercises," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 2, pp. 468–477, Feb. 2020.

[2] A. Mangal and V. Tiwari, "RGB-D sensor-based musculoskeletal health monitoring: A review," *IEEE Sensors J.*, vol. 21, no. 18, pp. 20064–20080, Sept. 2021.

[3] S. Chander, P. Pal, and A. Kumar, "RGB video-based physical exercise quality assessment with spatio-temporal cross-attention network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 1, pp. 125–139, Jan. 2025.

[4] C. Mennella et al., "Deep learning for automatic quality assessment of home-based physical rehabilitation: A systematic review," *Expert Syst. Appl.*, vol. 213, p. 118922, Mar. 2023.

[5] L. Zhang, Y. Chen, and R. Wang, "MediaPipe pose estimation for clinical movement analysis: Validation and applications," *J. Biomech.*, vol. 142, p. 111285, Feb. 2024.

[6] C.-H. Yeh et al., "Yoga pose quality assessment using MediaPipe and machine learning," *Sensors*, vol. 25, no. 2, p. 412, Jan. 2025.

[7] M. Simoes, T. Pinho, and J. Santos, "Accuracy of MediaPipe for physical therapy exercise classification," *IEEE Access*, vol. 12, pp. 15432–15441, 2024.

[8] K. Lee, J. Park, and S. Kim, "Validation of MediaPipe for Balance Error Scoring System assessment," *Gait Posture*, vol. 105, pp. 89–95, Jan. 2025.

[9] R. Hernandez, M. Lopez, and A. Garcia, "Postural assessment using deep learning for physiotherapy applications," *Comput. Methods Programs Biomed.*, vol. 238, p. 107612, Feb. 2025.

[10] Y. Cai, W. Li, and H. Zhang, "Swin-UNet for 3D human motion quality assessment in rehabilitation," *Med. Image Anal.*, vol. 89, p. 102876, Jan. 2025.

[11] K. Rajesh, P. Kumar, and S. Sharma, "Spatio-temporal graph networks for exercise recognition using OpenPose," *Pattern Recognit. Lett.*, vol. 175, pp. 112–119, Nov. 2024.

[12] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS Workshop Deep Learn.*, 2014.