https://doi.org/10.22201/dgtic.30618096e.2026.4.ESPECIAL.162

Núm. especial, abril 2026, págs. 48 - 63

Reconocimiento de señales dinámicas del Lenguaje Mexicano de Señas mediante redes LSTM para contextos de emergencia

Dynamic mexican sign language recognition using LSTM networks for emergency response

Información del reporte:

Para citar este reporte técnico:

Luviano Soto, I., Raya, A. y Flores Fernández, G.C. (2026). Reconocimiento de señales dinámicas del Lenguaje Mexicano de Señas mediante redes LSTM para contextos de emergencia [número especial]. *Cuadernos Técnicos Universitarios de la DGTIC, 4*, páginas (48 - 63). https://doi.org/10.22201/dgtic.30618096e.2026.4.ESPECIAL.162

Itzel Luviano Soto
Universidad Michoacana de San Nicolás de Hidalgo
itzel.luviano@umich.mx
ORCID: 0009-0003-1292-5537

Alfredo Raya
Universidad Michoacana de San Nicolás de Hidalgo
alfredo.raya@umich.mx
ORCID: 0000-0002-5394-8634

Giovanni Carlo Flores Fernández
Universidad Michoacana de San Nicolás de Hidalgo
giovanni.flores@umich.mx
ORCID: 0000-0002-8678-0196

Resumen

La integración de disciplinas como las matemáticas, los métodos numéricos, la computación y el modelado matemático ha impulsado el desarrollo de herramientas tecnológicas capaces de identificar patrones y predecir fenómenos con alta precisión. Entre estas herramientas, la inteligencia artificial ha generado soluciones innovadoras en diversos sectores industriales y de ingeniería; sin embargo, su aplicación, orientada al fortalecimiento de poblaciones vulnerables, aún es limitada.

En este contexto, se explora el uso de inteligencia artificial como herramienta de apoyo para la detección temprana de situaciones de riesgo que afectan a la comunidad sorda usuaria del Lenguaje Mexicano de Señas. El objetivo es desarrollar y evaluar un modelo basado en redes neuronales recurrentes tipo LSTM capaz de reconocer, en

https://doi.org/10.22201/dgtic.30618096e.2026.4.ESPECIAL.162 **Núm. especial**, abril 2026, págs. 49 - 63

tiempo real, señas del Lenguaje Mexicano de Señas asociadas a contextos de emergencia. La metodología propuesta se basa en una red neuronal recurrente entrenada para el reconocimiento en tiempo real de un conjunto de señas asociadas a contextos de emergencia. El sistema procesa secuencias de video cuadro por cuadro, identifica patrones temporales en los movimientos de manos, rostro y cuerpo, así como estima la probabilidad de ocurrencia de palabras vinculadas a situaciones de riesgo. El modelo alcanzó una precisión de hasta el 100% en la identificación de señas críticas en los conjuntos de entrenamiento y validación. No obstante, durante las pruebas, se identificaron errores en el reconocimiento de aquellas con alta similitud gestual. A pesar de las limitaciones actuales del sistema, como la latencia en la inferencia y el tamaño reducido del conjunto de datos, los resultados evidencian el potencial de este enfoque como una herramienta de apoyo para la identificación temprana de situaciones de riesgo. Asimismo, se identifican oportunidades de mejora futura orientadas a optimizar la velocidad de procesamiento, ampliar el vocabulario reconocido y avanzar hacia una implementación operativa en entornos reales.

Palabras clave: Aprendizaje automático, redes neuronales recurrentes, lenguaje mexicano de señas, situaciones de emergencia, inteligencia artificial aplicada.

Abstract

The integration of disciplines such as mathematics, numerical methods, computing, and mathematical modeling has driven the development of technological tools capable of identifying patterns and predicting phenomena with high accuracy. Among these tools, artificial intelligence has generated innovative solutions across several industrial and engineering sectors; however, its application aimed at supporting vulnerable populations remains limited. In this context, the use of artificial intelligence is explored as a support tool for the early detection of risk situations affecting the deaf community that uses Mexican Sign Language. The objective is to develop and evaluate a model based on Long Short-Term Memory recurrent neural networks capable of recognizing in real time Mexican Sign Language signs associated with emergency contexts. The proposed methodology is based on a recurrent neural network trained for real-time recognition of a set of signs related to emergency situations. The system processes video sequences frame by frame, identifies temporal patterns in hand, face, and body movements, and estimates the probability of occurrence of words linked to risk situations. The model achieved an accuracy of up to 100% in the identification of critical signs in the training and validation datasets. However, during testing, errors were identified in the recognition of signs with high gestural similarity. Despite the current limitations of the system, such as inference latency and the small dataset size, the results demonstrate the potential of this approach as a support tool for the early identification of risk situations. Furthermore, opportunities for future improvement are identified, including optimizing processing speed, expanding the recognized vocabulary, and advancing toward operational implementation in real-world environments.

Keywords: *Machine learning, recurrent neural networks, mexican sign language, emergency situations, applied artificial intelligence.*

https://doi.org/10.22201/dgtic.30618096e.2026.4.ESPECIAL.162 **Núm. especial**, abril 2026, págs. 50 - 63

1. INTRODUCCIÓN

La seguridad pública en México ha enfrentado desafíos significativos durante las últimas dos décadas. Según el Instituto Nacional de Estadística y Geografía (2023), la tasa de incidencia delictiva alcanzó 29,582 delitos por cada 100,000 habitantes en 2022, lo que refleja un incremento sostenido en comparación con años anteriores. Esta situación ha generado que la población mexicana considere la inseguridad como uno de los principales problemas nacionales (INEGI, 2023). Las zonas urbanas, por mayor densidad poblacional y actividad económica, concentran tanto la incidencia delictiva como la demanda de servicios de emergencia eficientes.

Paralelamente, las personas con alguna discapacidad continúan enfrentando barreras estructurales que limitan su participación plena en la sociedad. De acuerdo con el Censo de Población y Vivienda 2020 (INEGI, 2021), aproximadamente 7.1 millones de personas en México presentan algún tipo de discapacidad, lo que representa el 5.7% de la población total. Dentro de este grupo, cerca de 2.4 millones reportan dificultades relacionadas con la audición o el lenguaje. La tasa de participación laboral de personas con discapacidad auditiva es significativamente inferior al promedio nacional, situándose en torno al 38.5%, en contraste con el 60.2% de la población sin discapacidad (INEGI, 2021). Esta exclusión económica se relaciona directamente con barreras comunicativas persistentes en los espacios laborales, incluso en aquellos que implementan políticas de inclusión.

El Lenguaje Mexicano de Señas (LMS) constituye la lengua natural de la comunidad sorda en México y fue reconocido oficialmente en 2005 mediante la Ley General para la Inclusión de las Personas con Discapacidad. Sin embargo, el dominio del LMS por parte de la población oyente sigue siendo limitado. Aunque no existen estadísticas oficiales que cuantifiquen con precisión las competencias en LMS dentro de la población general, diversos estudios y reportes institucionales señalan que el conocimiento de esta lengua entre personas oyentes es muy reducido y se concentra principalmente en entornos especializados, como el ámbito educativo o familiar de personas sordas. Esta situación genera vacíos comunicativos en instituciones públicas, espacios educativos y centros de trabajo. En contextos de emergencia, donde la rapidez y precisión en la transmisión de información resultan críticas, estas barreras pueden tener consecuencias graves para la seguridad del personal sordo y la efectividad operativa de las instituciones.

El Centro de Comando, Control, Cómputo, Comunicaciones y Contacto Ciudadano (C5) de Morelia, Michoacán, representa un caso paradigmático de esta problemática. Como eje central de la respuesta a emergencias en la capital de este estado, el C5 coordina servicios de policía, bomberos, protección civil y atención médica hospitalaria. Históricamente, esta institución ha incorporado personas sordas debido a sus habilidades visuales superiores para el monitoreo continuo de sistemas de videovigilancia. No obstante, la integración de personas sordas al personal de trabajo no ha sido plenamente efectiva: pocos de los operadores oyentes dominan el LMS (C5 Morelia, comunicación personal, 2024), y el personal sordo presenta dificultades para comunicarse fluidamente mediante texto escrito debido a que el español representa su segunda lengua. Esta asimetría comunicativa ralentiza la respuesta ante incidentes, incrementa el riesgo de malentendidos críticos y limita la participación del personal sordo en actividades operativas de alto valor. En este contexto, el sistema propuesto en este trabajo se desarrolló y evaluó como una prueba piloto experimental orientada a explorar el potencial del reconocimiento automático del LMS como herramienta de apoyo para mejorar la comunicación en entornos de respuesta a emergencias.

https://doi.org/10.22201/dgtic.30618096e.2026.4.ESPECIAL.162 **Núm. especial**, abril 2026, págs. 51 - 63

Por otro lado, durante las últimas dos décadas, los avances en inteligencia artificial (IA), visión por computadora y aprendizaje profundo han transformado múltiples sectores industriales y sociales. Técnicas basadas en redes neuronales convolucionales (CNN) y recurrentes (RNN), así como algoritmos de IA han demostrado capacidad para procesar información visual compleja, reconocer patrones temporales y asistir en tareas que previamente requerían intervención humana intensiva (Goodfellow *et al*., 2016). En el ámbito del reconocimiento de lenguas de señas, diversos estudios internacionales han reportado tasas de exactitud superiores al 90%, utilizando arquitecturas profundas y conjuntos de datos robustos (Huang *et al*., 2015; Koller *et al*., 2020).

En el contexto específico del LMS, investigaciones recientes han explorado diferentes enfoques metodológicos. Rodríguez *et al.* (2025) alcanzaron una precisión del 92% en el reconocimiento de señas estáticas empleando Máquinas de Soporte Vectorial (SVM) y características extraídas mediante MediaPipe (una paquetería de uso libre de Python). Morfín-Chávez *et al.* (2023) reportaron un F1-score de 0.98 en la clasificación de letras del alfabeto utilizando algoritmos de aprendizaje automático clásicos. Sánchez-Vicinaiz *et al.* (2024) propusieron un sistema de detección de dactilología en LSM mediante fotogramas de MediaPipe y CNNs, extendiendo así las capacidades de reconocimiento hacia señas estáticas del alfabeto. Por su parte, Martínez-Seis *et al.* (2019) lograron un 92% de exactitud en la identificación de señas tanto estáticas como dinámicas mediante CNNs, demostrando la viabilidad de técnicas de aprendizaje profundo para este dominio. Solís *et al.* (2016) obtuvieron una tasa de reconocimiento del 93%, empleando momentos normalizados y redes neuronales artificiales (ANN, por sus siglas en inglés) para señas estáticas. Mejía-Pérez *et al.* (2022) desarrollaron un sistema de reconocimiento automático de LSM mediante cámara de profundidad y redes recurrentes, comparando arquitecturas LSTM y GRU con coordenadas 3D de manos, rostro y cuerpo, alcanzando una precisión del 97 % en datos limpios.

A pesar de estos avances, la mayoría de los sistemas desarrollados se han enfocado en vocabularios generales o alfabetos, con escasa atención a dominios especializados como el vocabulario de emergencias. Adicionalmente, los trabajos previos han priorizado señas estáticas sobre señas dinámicas, a pesar de que estas últimas constituyen la mayoría del vocabulario funcional del LMS. Las señas dinámicas, caracterizadas por movimientos complejos y transiciones temporales, representan desafíos técnicos adicionales que requieren arquitecturas capaces de modelar dependencias secuenciales.

En este contexto, el objetivo general del presente trabajo fue desarrollar e implementar un sistema de reconocimiento automático de señas del LMS orientado a la identificación de palabras clave relacionadas con situaciones de riesgo, utilizando RNNs con arquitectura *Long Short-Term Memory* (LSTM). El sistema fue diseñado para operar en tiempo real, procesar señales dinámicas mediante la extracción de puntos clave corporales y facilitar la comunicación entre personal sordo y oyente en el C5 de Morelia. El desarrollo del sistema inició en febrero de 2025 y se evaluó mediante una prueba piloto a finales del mes de noviembre de 2025. Los objetivos específicos fueron: (1) construir un conjunto de datos de señas dinámicas asociadas a contextos de emergencia, (2) diseñar y entrenar una arquitectura LSTM optimizada para el reconocimiento de secuencias temporales, y (3) evaluar el desempeño del sistema mediante pruebas con múltiples usuarios en condiciones reales de uso.

2. DESARROLLO TÉCNICO

El desarrollo del sistema se fundamentó en la integración de técnicas de visión por computadora, extracción automatizada de características espaciotemporales y modelado mediante aprendizaje

https://doi.org/10.22201/dgtic.30618096e.2026.4.ESPECIAL.162 **Núm. especial**, abril 2026, págs. 52 - 63

profundo. La arquitectura general del proyecto fue diseñada para capturar secuencias de video en tiempo real, procesarlas mediante una RNN optimizada y generar como salida la predicción de la palabra ejecutada en LMS con su respectiva probabilidad asociada. En este contexto, los autores participaron directamente en el desarrollo de la arquitectura de la red neuronal y la construcción del conjunto de datos utilizado para el entrenamiento con el apoyo de personal del C5 para la correcta ejecución de señas y la evaluación del modelo.

El enfoque metodológico adoptado se sustentó en investigaciones previas que han demostrado la efectividad de las redes LSTM para el reconocimiento de lenguas de señas. Huang *et al*. (2015) implementaron una arquitectura LSTM bidireccional para el reconocimiento de Lengua de Señas Americana (ASL), logrando una exactitud del 91.7% en un vocabulario de 100 palabras. De manera similar, Koller *et al*. (2020) desarrollaron un sistema híbrido que combinaba redes convolucionales con LSTM para el reconocimiento continuo de Lengua de Señas Alemana, alcanzando tasas de reconocimiento superiores al 89% en secuencias de señas encadenadas. Estos antecedentes demostraron que las arquitecturas recurrentes son particularmente adecuadas para capturar la naturaleza temporal y secuencial inherente a las lenguas de señas. Para el reconocimiento del lenguaje de señas, Sheth *et al*. (2023) compararon el desempeño de redes LSTM y GRU en la clasificación de señas dinámicas, encontrando que ambas arquitecturas recurrentes son efectivas para capturar patrones temporales en secuencias gestuales. Samaan *et al*. (2022) demostraron la viabilidad del flujo de trabajo basado en *landmarks* de MediaPipe combinados con una RNN para el reconocimiento de señas, reportando resultados satisfactorios en términos de precisión y velocidad de procesamiento. Por su parte, Ravikiran (2025) implementó un sistema de reconocimiento en tiempo real utilizando MediaPipe para la extracción de puntos de referencia o *landmarks*, y LSTM para la clasificación.

Por su parte, González-Rodríguez *et al*. (2024) desarrollaron un sistema con una exactitud del 98%, basado en una arquitectura bidireccional, la cual demostró un alto desempeño en la traducción de LSM.

Como se aprecia en los trabajos citados con anterioridad, una de las RNN más utilizadas para el reconocimiento de señas son las redes LSTM, las cuales están diseñadas específicamente para modelar dependencias temporales de largo alcance en datos secuenciales. A diferencia de las RNN convencionales, las LSTM incorporan mecanismos de compuertas —de entrada, olvido y salida— que regulan el flujo de información a lo largo del tiempo, permitiendo conservar información relevante y mitigar problemas como el desvanecimiento o explosión del gradiente durante el entrenamiento (Hochreiter & Schmidhuber, 1997). Estas características hacen que las LSTM sean particularmente adecuadas para el análisis de secuencias de video, donde la correcta interpretación de una seña depende no sólo de un instante aislado, sino de la evolución temporal de los movimientos de manos, expresiones faciales y posturas corporales. En consecuencia, el uso de esta arquitectura permite capturar patrones dinámicos complejos, mejorando la precisión y robustez del reconocimiento de señas asociadas a contextos de riesgo.

Dado lo anteriormente descrito en el sistema propuesto en este trabajo, se diseñó una LSTM utilizando *landmarks* para detectar la postura de la ejecución de la seña. El modelo fue implementado en Python 3.9, utilizando Jupyter Notebook dentro del entorno de desarrollo Visual Studio Code. Esta configuración facilitó la integración de múltiples bibliotecas especializadas: OpenCV para la captura y procesamiento de video, MediaPipe para la extracción de puntos clave corporales, NumPy para operaciones matriciales, y TensorFlow con Keras para el diseño, entrenamiento y evaluación del modelo de aprendizaje profundo. La selección de estas herramientas se basó en su amplia adopción en la comunidad científica, su

https://doi.org/10.22201/dgtic.30618096e.2026.4.ESPECIAL.162 **Núm. especial**, abril 2026, págs. 53 - 63

documentación robusta y su interoperabilidad probada en tareas de visión por computadora (Graves *et al*., 2013).

2.1 METODOLOGÍA

El procedimiento metodológico seguido para el desarrollo del modelo se organizó en tres etapas principales: (1) generación de la base de datos mediante captura de señales dinámicas, (2) diseño y entrenamiento de la red LSTM, y (3) validación del sistema mediante pruebas con usuarios múltiples. Cada una de estas etapas se describe detalladamente a continuación.

2.1.1 GENERACIÓN DE LA BASE DE DATOS

La identificación de palabras clave fue realizada mediante un proceso colaborativo con personal operativo del C5 de Morelia. Durante sesiones de trabajo realizadas en abril de 2025, se analizaron las barreras principales de comunicación para identificar las categorías de emergencia más frecuentes. Con base en esta información, se seleccionaron cinco palabras en LMS para conformar el vocabulario inicial del sistema: violencia, peligro, asesinato, ayuda y accidente. Esta selección consideró tanto la frecuencia de uso como la relevancia operativa para la detección temprana de situaciones críticas.

La seña de la palabra violencia se caracteriza por un movimiento repetitivo de golpeo con dos dedos en la zona centro de la frente; la palabra peligro involucra dos dedos desplazándose lateralmente sobre la palma empuñada de la otra mano; la palabra asesinato consiste en la inserción perpendicular de dos dedos sobre otros dos en orientación vertical; la palabra ayuda se ejecuta mediante un movimiento ascendente de la mano abierta desde el pecho hacia el frente; y la palabra accidente requiere el choque lateral a los costados del torso con la palma abierta. Estas descripciones fueron validadas con un intérprete certificado de LMS para garantizar su correcta ejecución durante la fase de captura de datos.

Debido a que las palabras seleccionadas corresponden a señas dinámicas, se requirió capturar secuencias completas que representan la variación espacial y temporal de cada gesto. Las señas dinámicas no pueden caracterizarse mediante una sola imagen estática, puesto que su significado depende del movimiento y la transición entre posiciones. Por esta razón, la captura de datos se realizó *frame* por *frame*, mediante las posiciones corporales para cada instante de la ejecución.

La captura se realizó mediante la biblioteca MediaPipe Holistic, que integra detección simultánea de rostro, manos y cuerpo. Este modelo permitió identificar:

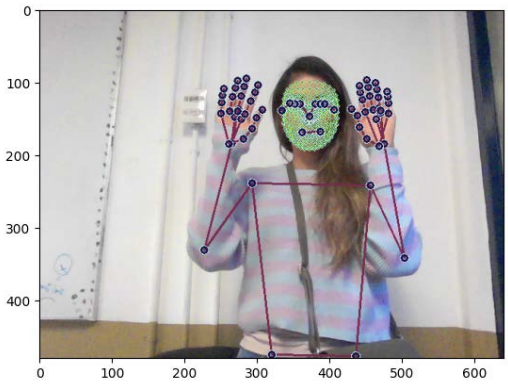
- 468 puntos faciales, incluyendo contornos, labios, cejas y puntos clave en los ojos;
- 21 puntos por mano, basados en el modelo MediaPipe Hands (Zhang *et al*., 2020), abarcando falanges, articulaciones y orientación de los dedos;
- 33 puntos corporales, principalmente en torso, hombros, codos, cadera y rodillas.

En la Figura 1, se presenta un ejemplo de los *landmarks* o puntos clave utilizados en la ejecución de cada una de las palabras seleccionadas.

https://doi.org/10.22201/dgtic.30618096e.2026.4.ESPECIAL.162 **Núm. especial**, abril 2026, págs. 54 - 63

Figura 1

Detección de landmarks o puntos claves en el proceso de recolección de datos



Estos 543 puntos clave por *frames* se extrajeron en tiempo real mientras la persona ejecutaba la seña frente a una cámara web. Para cada palabra, se registraron 30 *frames* por repetición y cada una de las cinco palabras se repitió 30 veces, lo que generó un conjunto robusto de ejemplos con variabilidad natural entre repeticiones. En total, cada ejecución produjo una matriz volumétrica con [150, 30, 1662] coordenadas individuales (543 puntos × 30 *frames*), que, posteriormente, se almacenaron como vectores numéricos en archivos independientes. El conjunto de datos utilizado para el entrenamiento fue generado por un sólo ejecutante. No obstante, el modelo se entrenó a partir de las coordenadas de los *landmarks* o puntos clave corporales extraídos mediante MediaPipe, por lo que el aprendizaje se basa en la dinámica espacial y temporal de estos puntos y no en características visuales específicas del individuo. Sin embargo, esta característica podría limitar la capacidad de generalización del modelo frente a variaciones gestuales entre diferentes usuarios o introducir cierto sesgo en el proceso de reconocimiento.

Este enfoque basado en coordenadas permitió reducir significativamente la carga computacional del entrenamiento. En lugar de trabajar con imágenes completas —que requieren mayor memoria, procesamiento y redes más profundas— se utilizaron únicamente los puntos esenciales que describen las posiciones corporales relevantes. Esta estrategia permitió optimizar el aprendizaje, acelerar los tiempos de entrenamiento y favorecer el entrenamiento del modelo.

2.1.2 DISEÑO Y GENERACIÓN DE LA LSTM

Para procesar las secuencias temporales generadas por los *frames* de cada seña, se seleccionó una RNN basada en la arquitectura LSTM. Este tipo de red es particularmente adecuado para gestionar dependencias temporales y evitar problemas clásicos de las RNN como el desvanecimiento del gradiente.

La clave de las LSTM radica en su capacidad para almacenar y actualizar información relevante a corto plazo mediante compuertas de memoria. En el contexto del reconocimiento de señas, esto permite que la red analice no sólo las posiciones individuales de los puntos, sino su evolución en el tiempo; por ejemplo, cambios en la orientación de la mano, trayectorias de movimiento o variaciones en la postura.

La arquitectura diseñada se implementó en TensorFlow/Keras y estuvo compuesta por:

- Tres capas LSTM configuradas con 64, 128 y 64 unidades respectivamente.

https://doi.org/10.22201/dgtic.30618096e.2026.4.ESPECIAL.162 **Núm. especial**, abril 2026, págs. 55 - 63

- Tres capas densas con 64, 32 y 5 neuronas. La última capa contiene cinco neuronas porque corresponde a las cinco palabras del vocabulario estudiado.

En cuanto a funciones de activación:

- Las capas internas utilizaron ReLU, que acelera el aprendizaje y estabiliza el gradiente.
- La capa final utilizó Softmax, permitiendo obtener probabilidades normalizadas para la clasificación multiclase.

La selección de la arquitectura de la red responde a criterios heurísticos basados en múltiplos de cuatro, utilizados comúnmente en problemas de secuencias por su estabilidad. En la Tabla 1, se presenta la configuración de la arquitectura empleada

Tabla 1

Arquitectura empleada en la RNN

Tipo de Capa	Unidades / Neurona	Función de Activación	Parámetros
LSTM 1	64	—	442,112
LSTM 2	128	—	98,816
LSTM 3	64	—	49,408
Densa 1	64	ReLU	4,160
Densa 2	32	ReLU	2,080
Densa Final	5	Softmax	165

La preparación del conjunto de datos incluyó su división en subconjuntos de entrenamiento y validación, permitiendo evaluar la capacidad del modelo para generalizar antes de realizar predicciones en tiempo real.

Los principales hiperparámetros definidos fueron:

- Optimizador Adam: elegido debido a su capacidad para ajustar automáticamente la tasa de aprendizaje durante la propagación del error.
- *Learning rate* de 0.0001: un valor pequeño que favorece un aprendizaje más estable y reduce oscilaciones en la actualización de pesos.
- 2000 épocas de entrenamiento: este número se definió con el objetivo de permitir que el modelo alcanzara una convergencia estable considerando el tamaño relativamente reducido del conjunto de datos y la naturaleza secuencial del problema.
- *Batch size* adaptado dinámicamente según el tamaño de los vectores.
- Parámetros: 596,741 como total de parámetros entrenables.

Durante el proceso de entrenamiento, se monitoreo la *accuracy*, utilizada como métrica principal para medir la exactitud en la predicción además de la función de pérdida, basada en *categorical cross entropy*, la cual cuantifica el error entre las predicciones y las etiquetas reales. El monitoreo simultáneo de estas

https://doi.org/10.22201/dgtic.30618096e.2026.4.ESPECIAL.162 **Núm. especial**, abril 2026, págs. 56 - 63

métricas permitió observar y evaluar el comportamiento del aprendizaje, la presencia de sobreajuste o subajuste, y la estabilidad del modelo durante las iteraciones.

2.1.3 PRUEBA Y VALIDACIÓN

La evaluación del sistema se estructuró en dos fases complementarias: análisis cuantitativo de métricas durante el entrenamiento y validación cualitativa mediante pruebas con usuarios en tiempo real.

Las métricas principales monitoreadas fueron la exactitud (*accuracy* en inglés), la precisión, el *recall, F1-score* y la función de pérdida, calculadas tanto para el conjunto de entrenamiento como para el de validación. Las métricas de validación como *accuracy*, precisión, *recall* o sensibilidad y *F1-score* se definen como se presentan en las Ecuaciones (1-4):

(1)

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

(2)

$$precisión = \frac{TP}{TP + FP}$$

(3)

$$recall = \frac{TP}{TP + FN}$$

(4)

$$F1\text{-}score = \frac{2 \cdot recall \cdot precisión}{recall + precisión}$$

donde TP (*True Positives*), TN (*True Negatives*), FP (*False Positives*) y FN (*False Negatives*) representan el número de predicciones en cada categoría. En clasificación multiclase, esta métrica se generaliza como la proporción de muestras correctamente clasificadas respecto al total.

https://doi.org/10.22201/dgtic.30618096e.2026.4.ESPECIAL.162 **Núm. especial**, abril 2026, págs. 57 - 63

La función de pérdida *categorical cross-entropy* se calculó mediante la Ecuación (5),

$$CCE(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad (5)$$

donde *C* es el número de clases; y_i es la etiqueta verdadera (1 si la clase es correcta, 0 en las demás); así como "y_i es la probabilidad que predice el modelo de pertenecer a la clase i. Valores bajos de pérdida indican que las probabilidades predichas por el modelo se concentran en las clases correctas, mientras que valores altos señalan incertidumbre o predicciones incorrectas. El conjunto de validación correspondió al 10% del total del *dataset*, mientras que el 90% restante se utilizó para el entrenamiento del modelo.

3. RESULTADOS

El análisis del proceso de entrenamiento reveló un comportamiento de aprendizaje progresivo y estable. Durante las primeras cinco épocas, el modelo mostró una exactitud inicial entre 0.13 y 0.22, con pérdida aproximada de 1.59. Este desempeño inicial es esperado dada la inicialización aleatoria de los pesos de la red. Entre las épocas 5 y 50, se observó un incremento sostenido en la exactitud, alcanzando valores de 0.50 hacia la época 10 y 0.70 hacia la 50. Simultáneamente, la función de pérdida disminuyó de forma consistente, estabilizándose alrededor de 0.45. Este patrón indica que el modelo comenzó a identificar características discriminativas relevantes de las señas.

A partir de la época 300, el modelo alcanzó una exactitud de 1.0 (100%) en el conjunto de entrenamiento, manteniéndose en este valor durante las 1,700 épocas restantes. No obstante, considerando el tamaño reducido del conjunto de entrenamiento, este resultado podría indicar cierto grado de sobreajuste, fenómeno común en problemas de aprendizaje profundo con bases de datos limitadas. Por esta razón, posterior al entrenamiento, se realizó una prueba piloto para determinar el desempeño de esta red en un ambiente real y con varios usuarios. En la Tabla 2, se presentan las métricas complementarias de precisión, *recall y F1-score* por clase, obtenidas durante la fase de validación del conjunto de datos. Éstas permiten evaluar con mayor detalle el desempeño del modelo frente a datos no utilizados directamente en el proceso de entrenamiento.

Tabla 2

Métricas de validación obtenidas en el conjunto de validación

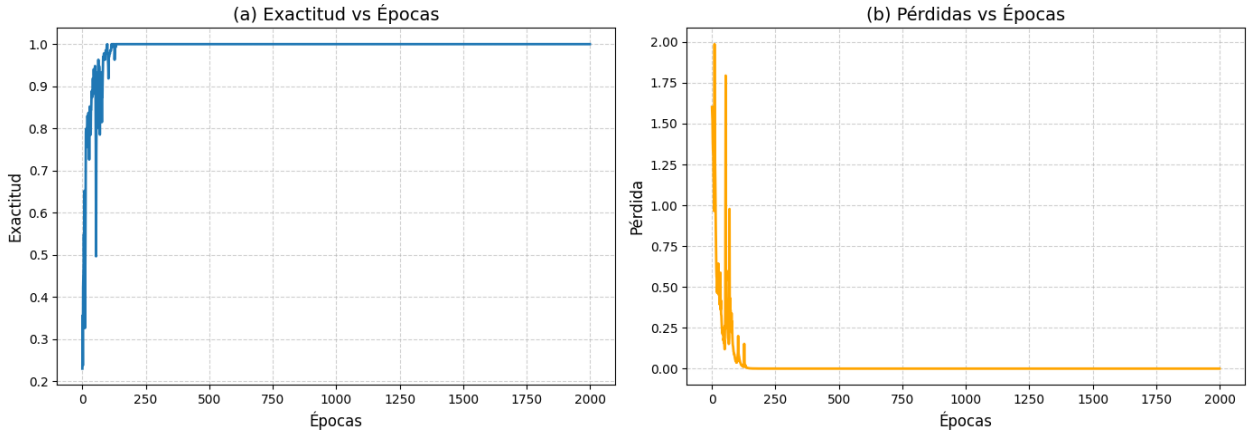
Clase	Precisión	Recall	F1-score
Violencia	0.75	1.00	0.86
Peligro	1.00	1.00	1.00
Asesinato	1.00	1.00	1.00
Ayuda	1.00	0.75	0.86
Accidente	1.00	1.00	1.00

https://doi.org/10.22201/dgtic.30618096e.2026.4.ESPECIAL.162 Núm. especial, abril 2026, págs. 58 - 63

La pérdida se estabilizó en valores mínimos entre 0.01 y 0.03, con fluctuaciones atribuibles al proceso estocástico de optimización. Las curvas de aprendizaje, presentadas en la Figura 2, evidencian claramente este comportamiento: la Figura 2(a) muestra la maximización de la exactitud, mientras que la Figura 2(b) ilustra la minimización del error.

Figura 2

Curvas de aprendizaje en el proceso de entrenamiento (2a) y validación de la red (2b)



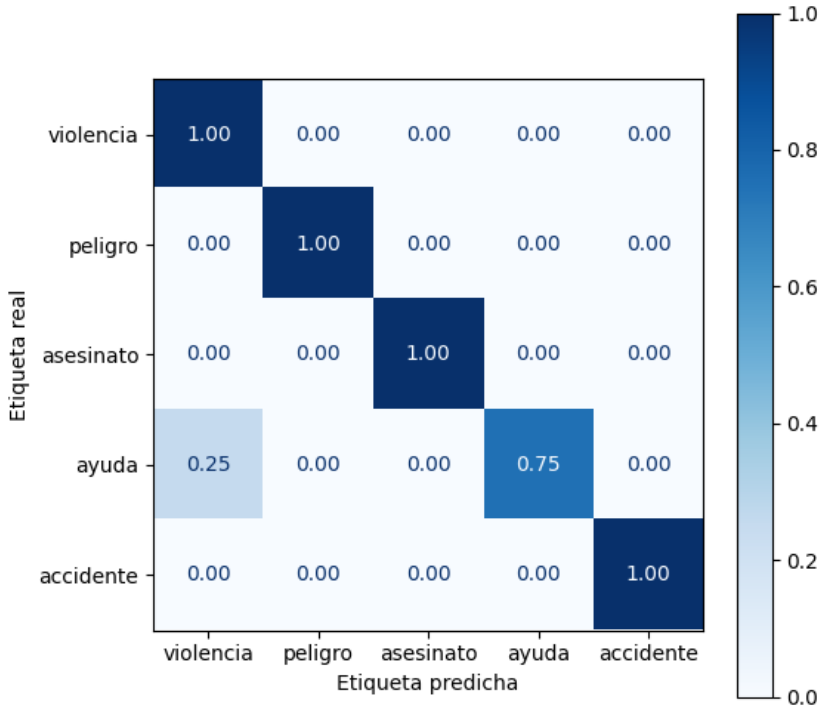
Es importante destacar que no se observaron señales de sobreajuste durante el entrenamiento, ya que las métricas de validación siguieron trayectorias similares a las de entrenamiento sin divergencias significativas. La ausencia de fluctuaciones abruptas o inestabilidades en las épocas finales sugiere que el modelo converge a un mínimo local estable.

Adicionalmente, también se presenta en la Figura 3 la matriz de confusión en los datos de validación, que muestra un desempeño general alto del modelo en la identificación de las señas evaluadas. Las clases violencia, peligro, asesinato y accidente presentan un reconocimiento correcto del 100%, lo que indica que todas las muestras pertenecientes a estas categorías fueron clasificadas adecuadamente. En contraste, la clase ayuda obtuvo una tasa de reconocimiento del 75%, observándose que el 25% de sus instancias fue clasificado erróneamente como violencia. Este comportamiento sugiere la existencia de similitudes gestuales entre ambas señas o variaciones en su ejecución, lo que puede generar ambigüedad en el proceso de clasificación. No obstante, el modelo demuestra una alta capacidad para distinguir la mayoría de las clases consideradas en el conjunto de validación.

https://doi.org/10.22201/dgtic.30618096e.2026.4.ESPECIAL.162 **Núm. especial**, abril 2026, págs. 59 - 63

Figura 3

Matriz de Confusión Normalizada para el conjunto de datos de validación



Por otro lado, las pruebas realizadas con tres participantes independientes revelaron un desempeño diferenciado según la palabra evaluada. Las palabras peligro, violencia y accidente fueron reconocidas correctamente en más del 90% de las ejecuciones respectivamente, demostrando que el modelo generaliza adecuadamente para estas clases cuando son ejecutadas por personas diferentes al usuario de entrenamiento. En la Figura 4, se presentan varios ejemplos de la ejecución de palabras realizada en la prueba piloto.

Figura 4

Prueba piloto de ejecución de palabras

https://doi.org/10.22201/dgtic.30618096e.2026.4.ESPECIAL.162 **Núm. especial**, abril 2026, págs. 60 - 63

Sin embargo, se identificaron confusiones sistemáticas para la palabra asesinato, la cual fue clasificada incorrectamente como la palabra peligro en las ejecuciones. Este error puede atribuirse a la similitud gestual entre ambas señas: mientras que la palabra peligro involucra dos dedos desplazándose lateralmente sobre la palma empuñada, la palabra asesinato requiere la inserción perpendicular de dos dedos sobre otros dos en orientación vertical. Ambas señas comparten configuraciones manuales similares (dos dedos extendidos) y movimientos en planos próximos, lo que dificulta su diferenciación cuando la orientación exacta no se captura con precisión suficiente.

La palabra ayuda también presentó confusiones ocasionales con la palabra violencia, aunque en menor medida que el caso anterior. Esta confusión resulta menos intuitiva, sugiriendo que podría deberse a variabilidad en la ejecución individual o a limitaciones en la cantidad de ejemplos de entrenamiento para capturar todas las posibles variaciones.

El análisis de pruebas en las ejecuciones en tiempo real, generadas durante las pruebas, reveló que la mayoría de los errores se concentraron en pares específicos de palabras, mientras que las confusiones entre otras combinaciones fueron prácticamente inexistentes. Por ejemplo, no se observó confusión entre la palabra accidente (choque de puños) y la palabra ayuda (movimiento ascendente con mano abierta), lo cual valida que el modelo distingue claramente diferencias gestuales marcadas.

Adicionalmente, se analizó la distribución de probabilidades generadas por la capa Softmax para las predicciones correctas. En promedio, el modelo asignó una probabilidad de 0.8 a la clase correcta cuando la predicción fue acertada. Esta confianza relativamente alta indica que, cuando el modelo clasifica correctamente, lo hace con certeza razonable. Por el contrario, las predicciones incorrectas mostraron distribuciones más uniformes, con probabilidades máximas promedio de 0.52, sugiriendo incertidumbre inherente en estos casos.

Finalmente, es importante reconocer la discrepancia entre el desempeño teórico durante el entrenamiento (100% de exactitud) y el rendimiento práctico con usuarios nuevos (aproximadamente 78% de exactitud promedio). Para la evaluación con usuarios externos, se realizaron pruebas con tres participantes independientes que no formaron parte del proceso de entrenamiento. Esta diferencia, aunque significativa, era anticipada dado el tamaño reducido del conjunto de datos (150 secuencias totales) y la limitada diversidad de ejecutantes (un único participante en el entrenamiento). Los resultados subrayan la necesidad crítica de ampliar el *dataset* con múltiples usuarios, variaciones de iluminación y contextos de ejecución diversos para mejorar la capacidad de generalización del sistema.

Es importante señalar que, en aplicaciones relacionadas con contextos de emergencia, la latencia durante la etapa de inferencia representa un factor crítico para la viabilidad operativa del sistema. Aunque el modelo desarrollado mostró un desempeño adecuado en las pruebas realizadas, el tiempo de respuesta puede verse afectado por factores como la capacidad computacional del dispositivo, el procesamiento de las secuencias de entrada y la complejidad del modelo. Por lo tanto, en trabajos futuros, se plantea

https://doi.org/10.22201/dgtic.30618096e.2026.4.ESPECIAL.162 Núm. especial, abril 2026, págs. 61 - 63

optimizar la arquitectura y explorar estrategias de aceleración y despliegue eficiente que permitan reducir la latencia y mejorar el desempeño del sistema en escenarios de operación en tiempo real.

4. CONCLUSIONES

En este trabajo, se empleó una RNN para la identificación en tiempo real de palabras de riesgo, con el objetivo de desarrollar una prueba piloto orientada a la implementación de un sistema inteligente de traducción en tiempo real. Como aportación técnica, se desarrolló la arquitectura de la red neuronal recurrente, la programación del sistema de reconocimiento basado en la extracción de puntos clave corporales mediante MediaPipe y la construcción del conjunto de datos utilizado para el entrenamiento y validación del modelo. Dicho sistema está destinado a su posible aplicación en centros de prevención y control de la violencia, con la finalidad de promover la inclusión de personas no oyentes en el ámbito laboral.

Si bien el desempeño global del modelo puede considerarse modesto, los resultados obtenidos durante la fase de entrenamiento y validación mostraron métricas satisfactorias, reflejadas en una adecuada disminución del error y una alta precisión (*accuracy*). No obstante, durante la ejecución en tiempo real, se identificaron inconsistencias en el reconocimiento de palabras cuando el sistema fue evaluado con distintas personas. Estas discrepancias pueden atribuirse, principalmente, a las variaciones en la forma de ejecución de las palabras, tanto en el conjunto de datos utilizado para el entrenamiento como en las condiciones reales de validación.

A pesar de estas limitaciones, el estudio demuestra que la identificación automática de palabras de riesgo es técnicamente viable. Sin embargo, esta tarea implica una complejidad considerable, dado que las palabras en lenguaje de señas no son estáticas, sino dinámicas y, en muchos casos, presentan similitudes en su ejecución, lo que dificulta su discriminación por parte del modelo.

Finalmente, se concluye que, mediante la construcción de un conjunto de datos más robusto y diverso, el incremento del vocabulario, así como el aumento en la profundidad y capacidad de la red neuronal, es posible mejorar sustancialmente el desempeño del sistema. En un trabajo futuro, este enfoque podría aplicarse de manera efectiva en distintos entornos laborales con la participación de personas no oyentes. Adicionalmente, este desarrollo representa una herramienta con alto potencial para reducir las barreras de comunicación, especialmente en contextos de riesgo donde las personas no oyentes son particularmente vulnerables, contribuyendo así a la inclusión social y a un mejor monitoreo y gestión de la seguridad pública. Asimismo, este trabajo contribuye al desarrollo y aplicación de Tecnologías de la Información y la Comunicación en el ámbito universitario, al integrar herramientas de IA, visión por computadora y aprendizaje profundo en un proyecto de investigación orientado a la inclusión social y a la innovación tecnológica.

AGRADECIMIENTOS

Se agradece al C5 y al Departamento de Idiomas de la Universidad Michoacana de San Nicolás de Hidalgo (UMSNH) por el apoyo brindado en la instrucción del Lenguaje Mexicano de Señas.

https://doi.org/10.22201/dgtic.30618096e.2026.4.ESPECIAL.162 **Núm. especial**, abril 2026, págs. 62 - 63

Declaración de contribución de autoría

Itzel Luviano Soto: Especialista en desarrollo tecnológico ingenieril. Realizó el desarrollo de un modelado integral para un sistema de reconocimiento basado en redes neuronales recurrentes. Llevó a cabo la generación y estructuración del conjunto de datos, incluyendo los procesos de adquisición, preprocesamiento y etiquetado de las muestras. Realizó el diseño, entrenamiento y ajuste de la red neuronal, así como la validación de su desempeño mediante métricas cuantitativas y pruebas en tiempo real. Efectuó el análisis de resultados, permitiendo identificar limitaciones del modelo y establecer líneas de trabajo futuro orientadas a la mejora del sistema.

Alfredo Raya: Especialista en física computacional. Contribuyó mediante el establecimiento de vínculos institucionales con el C5 y el Departamento de Idiomas, así como con la comunidad no oyente, lo que facilitó la participación de voluntarios para la generación del conjunto de datos. Asimismo, realizó el análisis e identificación de las palabras de riesgo consideradas en el estudio. Propuso la idea general del proyecto y participó en la definición de su enfoque conceptual, orientado a la aplicación del sistema en contextos de prevención, seguridad e inclusión social.

Giovanni Carlo Flores Fernández: Especialista en ingeniería civil. Brindó apoyo en la simulación y entrenamiento del sistema propuesto, así como en la optimización del modelo para mejorar su desempeño. Participó en el análisis estadístico de los resultados y en la evaluación de las métricas de validación empleadas en el estudio. Contribuyó en la redacción, revisión y corrección del documento, asegurando la claridad, coherencia y consistencia científica del manuscrito final.

REFERENCIAS

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

González-Rodríguez, J.-R., Córdova-Esparza, D.-M., Terven, J., & Romero-González, J.-A. (2024). Towards a bidirectional Mexican Sign Language–Spanish translation system: A deep learning approach. *Technologies, 12*(1). https://doi.org/10.3390/technologies12010007

Graves, A., Mohamed, A.-R., & Hinton, G. (2013, 26-31 de mayo). *Speech recognition with deep recurrent neural networks* [conferencia]. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canadá. https://doi.org/10.1109/ICASSP.2013.6638947

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang, J., Zhou, W., Li, H., & Li, W. (2015, 29 de junio-03 de julio). *Sign language recognition using 3D convolutional neural networks* [conferencia]. 2015 IEEE International Conference on Multimedia and Expo (ICME), Turín, Italia. https://doi.org/10.1109/ICME.2015.7177428

Instituto Nacional de Estadística y Geografía [INEGI]. (2021). *Censo de Población y Vivienda 2020: Resultados sobre discapacidad.* https://www.inegi.org.mx/programas/ccpv/2020/

Instituto Nacional de Estadística y Geografía [INEGI]. (2023). *Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública (ENVIPE) 2023.* https://www.inegi.org.mx/programas/envipe/2023/

https://doi.org/10.22201/dgtic.30618096e.2026.4.ESPECIAL.162 **Núm. especial**, abril 2026, págs. 63 - 63

Koller, O., Camgoz, N. C., Ney, H., & Bowden, R. (2020). Weakly supervised learning with multi-stream CNN–LSTM–HMMs to discover sequential parallelism in sign language videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(9), 2306–2320. https://doi.org/10.1109/TPAMI.2019.2911077

Martínez-Seis, B., Pichardo-Lagunas, O., Rodríguez-Aguilar, E. J., & Saucedo-Díaz, E.-R. (2019). Identification of static and dynamic signs of the Mexican Sign Language alphabet for smartphones using deep learning and image processing. *Research in Computing Science, 148*(11), 199–211.

Mejía-Pérez, K., Córdova-Esparza, D.-M., Terven, J., Herrera-Navarro, A.-M., García-Ramírez, T., & Ramírez-Pedraza, A. (2022). Automatic recognition of Mexican Sign Language using a depth camera and recurrent neural networks. *Applied Sciences, 12(*11). https://doi.org/10.3390/app12115523

Morfín-Chávez, R. F., Gortarez-Pelayo, J. J., & Lopez-Nava, I. H. (2023). Fingerspelling recognition in Mexican Sign Language (LSM) using machine learning [artículo de conferencia]. En H. Calvo, L. Martínez-Villaseñor, & H. Ponce (Eds.), *Advances in Computational Intelligence: 22nd Mexican International Conference on Artificial Intelligence, MICAI 2023* (pp. 110–120). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-47765-2_9

Ravikiran, V. (2025). Real-time sign language recognition and translation using MediaPipe and LSTM-based deep learning. *International Journal of Computer Applications, 187(*25), 10–14. https://doi.org/10.5120/ijca2025925415

Rodriguez, M., Oubram, O., Bassam, A., Lakouari, N., & Tariq, R. (2025). Mexican Sign Language Recognition: Dataset Creation and Performance. Evaluation Using MediaPipe and Machine Learning Techniques. *Electronics 14*(7). https://doi.org/10.3390/ELECTRONICS14071423

Sánchez-Vicinaiz, T. J., Camacho-Pérez, E., Castillo-Atoche, A. A., Cruz-Fernandez, M., García-Martínez, J. R., & Rodríguez-Reséndiz, J. (2024). MediaPipe frame and convolutional neural networks-based fingerspelling detection in Mexican Sign Language. *Technologies, 12*(8). https://doi.org/10.3390/technologies12080124

Samaan, G. H., Wadie, A. R., Attia, A. K., Asaad, A. M., Kamel, A. E., Slim, S. O., Abdallah, M. S., & Cho, Y.-I. (2022). MediaPipe's landmarks with RNN for dynamic sign language recognition. *Electronics, 11*(19). https://doi.org/10.3390/electronics11193228

Sheth, P., Rajora, S., & Makwana, Y. (2023). Sign language recognition application using LSTM and GRU (RNN). *ResearchGate.* https://doi.org/10.13140/RG.2.2.18635.87846

Solís, F., Martínez, D., & Espinoza, O. (2016). Automatic Mexican Sign Language recognition using normalized moments and artificial neural networks. *Engineering, 8(*10), 733-740. https://doi.org/10.4236/ENG.2016.810066

Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.-L., & Grundmann, M. (2020). MediaPipe Hands: On-device real-time hand tracking. *arXiv.* https://doi.org/10.48550/arXiv.2006.10214