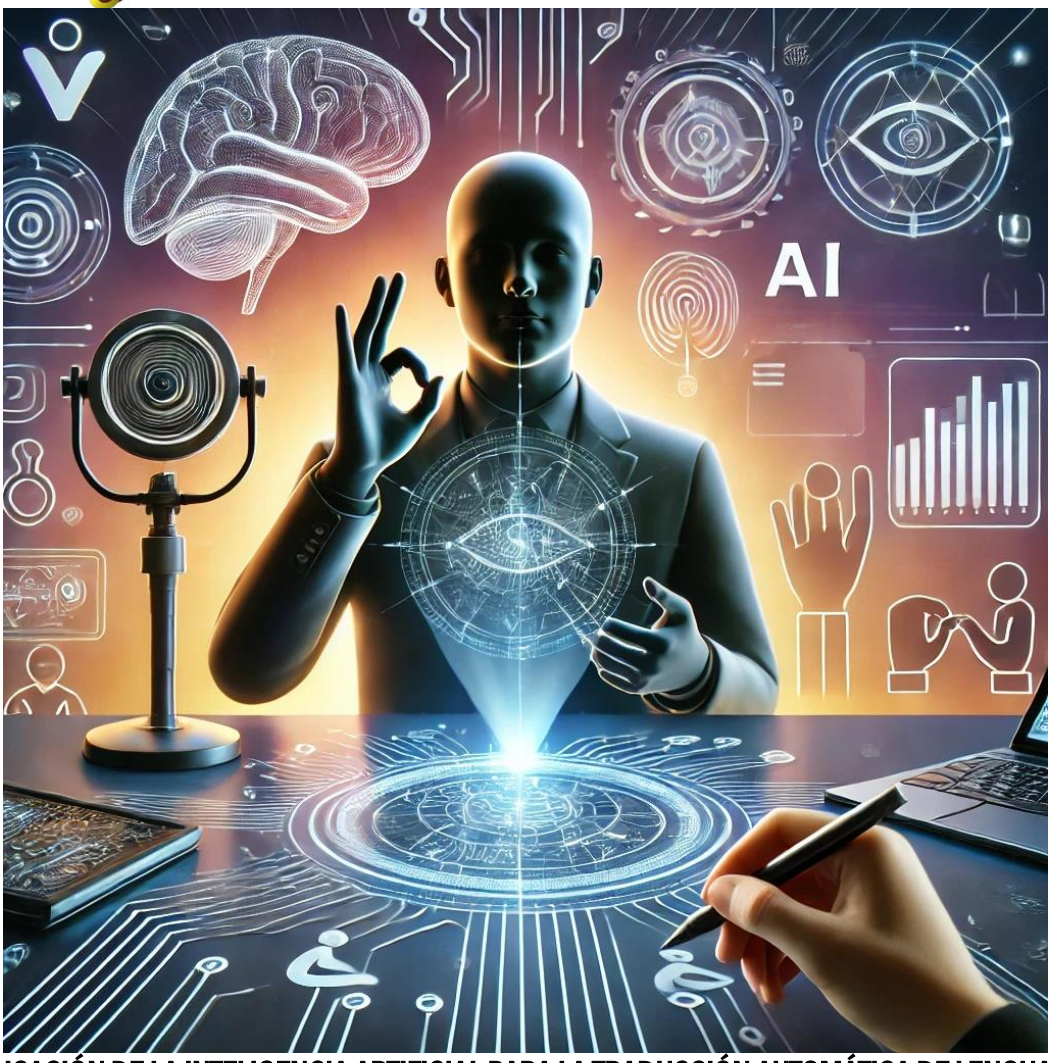
AÑO 2. No. 4
ISSN:2992-7978



APLICACIÓN DE LA INTELIGENCIA ARTIFICIAL PARA LA TRADUCCIÓN AUTOMÁTICA DE LENGUA DE SEÑAS

Josué S. Armenta[1]
ORCID: 0000-0003-0187-5338
Rodrigo Cadena Martínez[2]
ORCID: 0000-0001-9323-6132
Marcela D. Rodríguez[3]
ORCID: 0000-0002-6943-7812

[1] Doctorando Universidad Americana de Europa

[2]Profesor-Investigador Universidad Americana de Europa

[3] Profesor-Investigador Universidad Autónoma de Baja California

Resumen

La comunicación verbal es esencial para los seres humanos, pero condiciones como la sordera y la mudez impiden que un sector de la población haga uso de esta herramienta. Las lenguas de señas, como la Lengua de Señas Mexicana (LSM), permiten la comunicación no verbal, aunque una serie de barreras culturales continúa limitando la integración de los usuarios de lenguas de señas a su sociedad. Este trabajo desarrolló un sistema portátil de reconocimiento de lengua de señas (SLR) para traducir expresiones de la LSM al español en tiempo real, con motivo de reducir la brecha comunicativa entre las poblaciones sordas y oyentes. Los objetivos incluyeron desarrollar un sistema de adquisición de datos gestuales mediante sensores, construir y entrenar un modelo de clasificación, e integrar los componentes anteriores para evaluar el desempeño del traductor. Como metodología para conseguir el reconocimiento de signos se empleó un brazalete embebido de sensores y se desarrolló una aplicación portátil para capturar datos gestuales, preprocesarlos y convertirlos en imágenes que se alimentan a una red neuronal convolucional (CNN). Los resultados muestran que el modelo de clasificación alcanzó una exactitud por reconocimiento cruzado del 97% al utilizar todos los sensores del brazalete. El tiempo promedio de respuesta fue de 0.70 segundos para la traducción de oraciones compuestas de hasta cuatro signos, validando su capacidad de operar en tiempo real. Los resultados confirman la factibilidad de aplicar una novedosa metodología para llevar a cabo traducciones automáticas de la LSM con baja latencia y alta exactitud.

Palabras clave: aprendizaje profundo, reconocimiento de lengua de señas, redes neuronales artificiales

Abstract

Verbal communication is essential for human beings, but conditions such as deafness and muteness prevent a sector of the population from using this tool. Sign languages, such as Mexican Sign Language (LSM), allow non-verbal communication, although a series of cultural barriers continue to limit the integration of sign language users into their society. This work

AÑO 2. No. 4
ISSN:2992-7978

developed a portable sign language recognition (SLR) system to translate LSM expressions into Spanish in real time, in order to reduce the communication gap between deaf and hearing populations. The objectives included developing a sensor-based gesture data acquisition system, building and training a classification model, and integrating the above components to evaluate the translator's performance. As a methodology to achieve sign recognition, a bracelet embedded with sensors was used and a smart phone application was developed to capture gestural data, preprocess it and convert it into images that are fed to a convolutional neural network (CNN). The results show that the classification model achieved a cross-recognition accuracy of 97% when using all sensors on the bracelet. The average response time was 0.70 seconds for the translation of sentences composed of up to four signs, validating its ability to operate in real time. The results confirm the feasibility of applying a novel methodology to carry out automatic translations of the LSM with low latency and high accuracy.

Keywords: artificial neural networks, deep learning, sign language recognition

INTRODUCCIÓN

La comunicación es una herramienta esencial para los seres humanos, ya que a través de ella las personas expresan sus pensamientos, comparten sus sentimientos, resuelven conflictos y construyen relaciones (Knapp et al., 2014). La comunicación verbal, aquella que involucra el uso de palabras habladas, suele ser empleada para transmitir ideas detalladas o abstractas de manera rápida y precisa (Adler et al., 2023).

Desafortunadamente, no todas las personas tienen la oportunidad de desarrollar sus habilidades verbales. Las personas sordas y las personas mudas, por ejemplo, presentan condiciones físicas que dificultan o impiden el uso de oídos y voz. La Organización Mundial de la Salud estima que en la actualidad 1.500 millones de personas experimentan algún tipo de pérdida auditiva, y pronostica que para el año 2050 esta cifra podría alcanzar los 2.500 millones (World Health Organization, 2024). Ante la limitación o impedimento del uso de lenguas habladas, los grupos anteriores emplean la comunicación no verbal. En esta forma de comunicación, en lugar de palabras habladas, se recurre al uso expresiones faciales, gestos y posturas corporales (Wadhawan & Kumar, 2019).

AÑO 2. No. 4
ISSN:2992-7978

Las lenguas de señas son lenguajes que se valen de todos los elementos de la comunicación no verbal para la transmisión efectiva de pensamientos y sentimientos. Como todos los lenguajes, poseen sus propias reglas lingüísticas para la estructuración de palabras y oraciones (Mayberry & Squires, 2006). Los principales usuarios de lenguas de señas son las personas que presentan alguna limitación auditiva o del habla, pero pueden ser aprendidas por la población en general.

Sin embargo, factores como la falta de contacto con usuarios de lenguas de señas o escasez de intérpretes disuaden al grueso de la población de aprender alguna de las variantes existentes (Kudrinko et al., 2021). Sumado a esto, la mayor parte de los servicios públicos carecen de opciones de accesibilidad para personas con estas discapacidades (Bai & Bruno, 2020). Las anteriores barreras en la comunicación pueden suponer una serie de desventajas sistemáticas para los grupos vulnerables: la falta de acceso a la educación limita las opciones laborales, lo cual a su vez repercute en la calidad de vida.

El reconocimiento de lengua de señas (SLR) es la tecnología diseñada para identificar e interpretar de manera automática los signos ejecutados por un usuario de lengua de señas, con motivo de traducir su significado a una lengua hablada o texto escrito (Nimisha & Jacob, 2020). El propósito de los sistemas SLR es derribar las barreras de comunicación entre los usuarios de lenguas de señas y la sociedad (Rastgoo et al., 2021). A menudo estos sistemas son desplegados en servicios públicos como transporte, oficinas, museos, etc., con la finalidad de hacerlos inclusivos (Garg Pragati et al., 2009).

Los primeros esfuerzos en el reconocimiento de gestos, incluidos los relacionados con las lenguas de señas, se remontan a principios de los años noventa. Desde entonces, la evolución de la inteligencia artificial ha impulsado de manera significativa el desarrollo de sistemas SLR. Un ejemplo de esta evolución son las redes neuronales artificiales (ANN). A diferencia de las técnicas clásicas de aprendizaje automático, como *k* vecinos más próximos o árboles de decisión, las ANN pueden capturar de manera más eficiente patrones no lineares en grandes conjuntos de datos. Esto las hace aptas para resolver tareas complejas, entre ellas el procesamiento de lenguaje natural, el reconocimiento de imágenes, el reconocimiento del habla, y el reconocimiento de lenguas de señas.

Otro factor significativo en la modernización de los sistemas SLR son los avances en hardware. En la última década se ha favorecido el uso de cámaras de video y sensores vestibles para la captura de gestos, mientras que otros dispositivos, como los guantes, han caído en desuso. Por su parte, los teléfonos inteligentes cuentan en la actualidad con el poder de procesamiento suficiente para llevar a cabo el reconocimiento de signos de manera portátil e instantánea (tiempo real).

AÑO 2. No. 4
ISSN:2992-7978

El estado del arte en sistemas SLR revela que la mayoría de los desarrollos tecnológicos en los últimos 5 años han sido diseñados para reconocer gestos de lenguajes como la Lengua de Señas Americana (ASL) y la Lengua de Señas China (CSL). La ASL y la CSL cuentan, además, con múltiples conjuntos de datos públicos, lo cual facilita el desarrollo de nuevos sistemas SLR. Otras variantes, como la Lengua de Señas Mexicana (LSM), carecen tanto de conjuntos de datos públicos como de implementaciones modernas de sistemas SLR.

Con lo anterior, se identifica que no se ha explorado el desarrollo de sistemas SLR portátiles y que, mediante la aplicación de técnicas de aprendizaje profundo, realicen traducciones en tiempo real para reconocer gestos de la LSM. Se desconoce entonces cual es el desempeño de tales técnicas para detectar signos de esta lengua, así como sus limitaciones.

El objetivo general de esta investigación es desarrollar un sistema SLR portátil que produzca traducciones en tiempo real desde la LSM hacia el español, cuyo uso facilité la comunicación entre personas sordas y oyentes. Para lograr el anterior cometido, se identifican los siguientes objetivos específicos:

1. Desarrollar un sistema de adquisición de datos gestuales basado en sensores vestibles que permita recolectar y manejar datos específicos de la LSM.

2. Desarrollar un modelo de reconocimiento de lengua de señas considerando requerimientos asociados a la portabilidad del sistema e inferencia en tiempo real.

3. Integrar los componentes desarrollados como resultado de los objetivos previos para evaluar su exactitud.

Se justifica que la creación de tal sistema de traducción automatizada ayudaría a subsanar el rezago social que viven los principales usuarios de lenguas de señas; tanto ellos como el restante de la población se verían beneficiados de contar con una herramienta de comunicación sincrónica y mutua. Adicionalmente, este trabajo dejaría un precedente para iniciar futuras líneas de investigación que retomen esta problemática, aportando así una base de conocimiento en el área.

ESTADO DEL ARTE Y MARCO TEÓRICO

El desarrollo de sistemas SLR es un área de investigación multidisciplinar, ya que interviene ámbitos como el reconocimiento de patrones, el procesamiento del lenguaje natural, la visión artificial y el procesamiento de señales. El objetivo de estos sistemas es construir algoritmos y métodos capaces de identificar signos producidos y percibir su significado (Wadhawan & Kumar, 2019).

AÑO 2. No. 4
ISSN:2992-7978

El proceso para llevar a cabo el reconocimiento de signos, desde su identificación hasta su traducción, puede ser generalizado en unas cuantas etapas, nominalmente adquisición de datos, pre-procesamiento, extracción de características, clasificación y traducción (Adeyanju et al., 2021; Cheok et al., 2019; Nimisha & Jacob, 2020; Safeel et al., 2020).

La primera etapa del proceso, adquisición de datos, implica capturar los gestos manuales y/o no manuales del usuario de lengua de señas. Para esto se emplean dispositivos digitales, los cuales a través de cámaras o sensores son capaces de cuantificar el movimiento de las partes del cuerpo involucradas en la ejecución de signos. Cuando se emplean dispositivos del tipo cámara, los gestos son capturados en fotogramas individuales o secuencias de fotogramas (video). Por su parte, el uso de sensores implica capturar los gestos mediante señales o series temporales.

Entre los dispositivos del tipo cámara más populares para llevar a cabo SLR se encuentran las cámaras web, Kinect y LeapMotionController. A su vez, existe una amplia variedad de dispositivos dotados de sensores, entre ellos guantes, brazaletes y relojes inteligentes (Wadhawan & Kumar, 2019). Cada enfoque de captura (cámaras, sensores) tiene sus ventajas y limitaciones, y debido a esto no hay una alternativa mejor a la otra.

La información que han sido capturada directamente de un dispositivo digital es también llamada datos en crudo, y necesita ser tratada antes de ingresar a un modelo de clasificación. Esta etapa se conoce como pre-procesamiento y suele incluir la aplicación de algoritmos como la segmentación o la aplicación de desenfoque gaussiano y filtro de mediana (Cheok et al., 2019)

Los datos pueden contener características innecesarias o redundantes, por lo que estas son reducidas durante la extracción de características. El producto de este proceso contiene sólo la información relevante de los gestos previamente capturados, ahora representada en una versión compacta, que sirve como identidad del signo que se va a clasificar aparte de otros signos. Las redes neuronales convolucionales y el análisis de componentes principales (PCA) son las técnicas más empleadas en esta etapa (Nimisha & Jacob, 2020).

Con motivo de identificar a que signo corresponden, los datos previamente compactados deben ser ahora alimentados a un modelo de clasificación, lo cual ocurre en la etapa de clasificación. Los modelos de clasificación o inferencia son obtenidos cuando se procesa un conjunto de datos de entrenamiento mediante algoritmos de aprendizaje automático. Los conjuntos de datos pueden provenir de fuentes públicas o recolección propia. Las técnicas de aprendizaje automático más empleadas en la actualidad son derivadas del aprendizaje profundo, entre ellas las redes neuronales convolucionales (CNN) y las redes neuronales

recurrentes (RNN). Otras técnicas de aprendizaje automático clásicas, como las máquinas de vectores de soporte (SVM) y la deformación dinámica del tiempo (DTW), siguen gozando de popularidad (Adeyanju et al., 2021).

Una vez identificados, resta percibir el significado de los signos. La última etapa del proceso, traducción, introduce técnicas como la traducción automática basada en reglas (RBMT), traducción automática estadística (SMT) y traducción automática neuronal (NMT) para convertir palabras u oraciones en lengua de señas a su equivalente en una lengua hablada o texto. En la última década, una cantidad considerable de desarrollos tecnológicos SLR ha sido diseñada para reconocer signos de la Lengua de Señas Americana, la Lengua de Señas China, y la Lengua de Señas Alemana (Papastratis et al., 2021).

El desempeño de los sistemas SLR es comúnmente reportado mediante las métricas de exactitud y Word Error Rate (WER). Ambas métricas son expresadas en porcentaje. La exactitud indica la cantidad de predicciones correctas de entre todas las predicciones realizadas durante una solicitud de traducción. En cambio, el WER considera en su fórmula la cantidad de palabras incorrectas (como sustituciones y eliminaciones) dentro de una traducción. En general, es deseable que los sistemas SLR posean altas tasas de exactitud y bajas tasas de WER. El estado del arte revela que la mayor parte de los sistemas SLR construidos en los últimos cinco años reportó tasas de exactitud mayores al 90% y WER menores al 5%.

La literatura relacionada a sistemas SLR revela que existe una reciente tendencia a construir sistemas portátiles y con traducciones en tiempo real. Estos sistemas suelen ser implementados mediante la combinación de teléfonos inteligentes y sensores embebidos en dispositivos vestibles. Los sistemas SLR que obtienen datos de entrada mediante cámaras también son capaces de realizar inferencias en tiempo real, pero no suelen ser portátiles.

Tras la exploración del estado del arte, surge entonces la siguiente hipótesis: Con el empleo de sensores vestibles y técnicas de aprendizaje profundo, es posible implementar un sistema SLR portátil que identifique signos de la LSM en tiempo real, con una exactitud mayor al 90%.

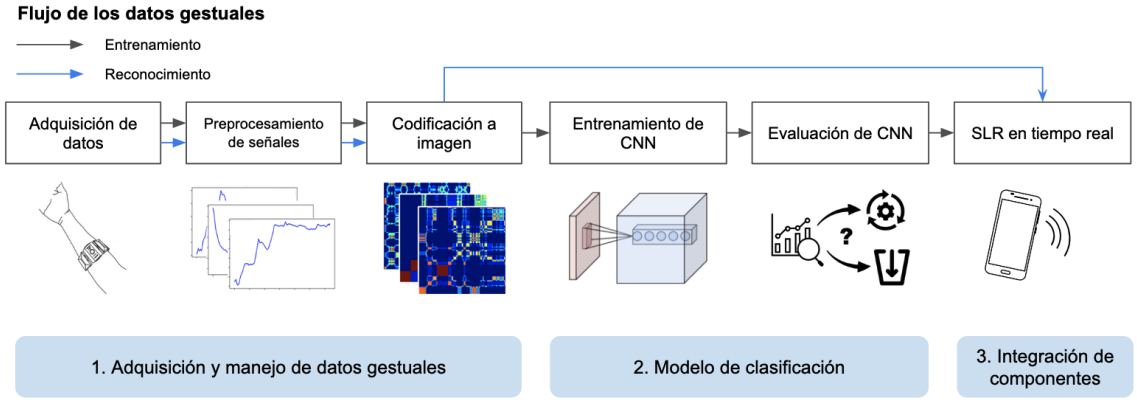
METODOLOGÍA

Se desarrolló una metodología inspirada en el proceso generalizado de SLR para implementar y evaluar el sistema propuesto. La Figura 1 describe de manera gráfica el recorrido de los datos gestuales a través de este sistema, desde su adquisición hasta su traducción en tiempo real. Como puede observarse en la figura, las etapas del proceso propuesto están estrechamente relacionadas al cumplimiento de los objetivos

específicos del proyecto, y con ello del objetivo general. La metodología para alcanzar estos objetivos será detallada a continuación.

Figura 1

Etapas del sistema SLR



Nota. El flujo de la información en el sistema difiere durante el entrenamiento del modelo de clasificación y el reconocimiento de signos.

Adquisición y manejo de datos gestuales

Se empleó el brazalete OYMotion gForcePro+ como medio de adquisición de datos gestuales (Figura 2). Este dispositivo vestible cuenta con sensores del tipo acelerómetro, giroscopio, magnetómetro y electromiografía de superficie. Tras ser colocado en el antebrazo de una persona, el brazalete hace uso de dichos sensores para computar y reportar la orientación y rotación del brazo, así como la actividad muscular de los dedos.

El dispositivo emite la lectura de sus sensores de manera continua mediante el estándar de comunicación electrónica Bluetooth Low Energy (BLE) 4.0; esto quiere decir que cualquier dispositivo que soporte el mismo protocolo de comunicación puede hacer uso de las señales transmitidas.

El brazalete gForcePro+ esta provisto de sensores EMG (electromiografía de superficie) e IMU (unidad de medida inercial). Las lecturas al sensor EMG son entregadas a través de ocho canales de datos, mientras que el IMU, compuesto por los sensores acelerómetro, giroscopio, y magnetómetro, emplea tres canales para cada uno, con un total de nueve. Este dispositivo vestible también cuenta con la capacidad de

muestrear en tiempo real la rotación del brazo, la cual es expresada a través de los sistemas matemáticos de ángulos de Euler (3 canales) y cuaterniones (4 canales).

Figura 2

Dispositivo de adquisición de datos

Nota. El Brazalete gForcePro+ es capaz de detectar el movimiento y actividad muscular del brazo dominante.

La aplicación de adquisición y manejo de datos gestuales fue desarrollada mediante el IDE Android Studio Hedgehog, SDK 33, y fue desplegada en un teléfono inteligente Android Motorola G32. La aplicación preprocesa los datos gestuales una vez adquiridos. Este preprocesamiento comienza con el remuestreo de las señales, pues estas provienen de sensores con diferentes tasas de muestreo (50 y 200 Hz). Esta operación se realizó aplicando la técnica de *interpolación spline cúbica*, igualando así el número de muestras para todos los canales de datos. Se estableció 1024 como la cantidad de muestras objetivo.

Tras el remuestreo, las señales fueron escaladas a través de la técnica de estandarización. Las series temporales estandarizadas son el resultado de eliminar la media y escalarlas a la varianza unitaria. Es natural que los rangos para los valores en el eje "Y" de todas las señales varíen, ya que proceden de las lecturas de distintos sensores. La estandarización es una medida para asegurar consistencia y uniformidad en dichos rangos.

Las señales emitidas por sensores suelen contener ruido, lo que puede afectar el desempeño del modelo de clasificación. Para reducir el ruido en todas las series temporales, se optó por aplicar el algoritmo de la transformada wavelet, concretamente la *ondícula spline biortogonal 3.9*, sugerida en el trabajo de (Wang et al., 2020),

Con base a los prometedores resultados de (C.-L. Yang et al., 2019), donde las señales de entrada son transformadas a imágenes y alimentadas a una CNN para aprovechar la extracción automática de

AÑO 2. No. 4
ISSN:2992-7978

características, se decidió aplicar esta técnica sobre los datos gestuales para concluir su preparación. Sugerido por el trabajo de (Wang & Oates, 2015) primero se aplicó el algoritmo de *aproximación agregada por partes* (PAA) a todas las series temporales, con una ventana de tamaño 8, con motivo de sub-muestrearlas a 128 intervalos (1024 ÷ 8 = 128).

Después, las señales fueron codificadas como imágenes a través de tres técnicas: *campo de suma angular gramiano* (GASF), *campo de diferencia angular gramiano* (GASF) y *campo de transición de Markov* (MTF). Todas las técnicas de codificación a imagen producen representaciones bidimensionales de 128 × 128 píxeles para cada señal. Para finalizar la preparación de los datos gestuales se procede a la concatenación vertical de las series temporales anteriormente codificadas. La concatenación vertical se emplea para construir una sola imagen a partir de múltiples series temporales, generando así una representación bidimensional única de los datos gestuales. Esta representación bidimensional es apta para el entrenamiento y consumo del modelo de clasificación.

Modelo de clasificación

El desarrollo de cualquier modelo de clasificación implica contar de antemano con un conjunto de datos de entrenamiento. Ya que no existen conjuntos de datos públicos que cumplan con los requerimientos de este sistema, se llevó a cabo una recolección propia.

Para encontrar voluntarios que participaran en la recolección de datos, se lanzó una convocatoria de reclutamiento en redes sociales dirigida a usuarios de la Lengua de Señas Mexicana (LSM) en la ciudad de Guaymas, Sonora, México. Como parte del proceso, un intérprete oficial de la LSM evaluó a los candidatos con base en criterios específicos. Estos incluyeron su capacidad para comunicarse fluidamente en LSM y su conocimiento de un vocabulario básico que abarcaba términos esenciales para el estudio.

Adicionalmente, se consideraron aspectos como el nivel de uso cotidiano de la LSM (por ejemplo, si eran usuarios nativos o aprendices avanzados) y su disponibilidad para participar durante la campaña de recolección. Tras el proceso de selección, trece candidatos fueron aceptados: ocho mujeres y cinco hombres, con edades comprendidas entre los 25 y 50 años, todos zurdos. Un modelo entrenado con datos diversos es más capaz de manejar las variaciones que encontrará en la práctica, como gestos realizados por personas de distintas edades o niveles de habilidad. Todos los voluntarios firmaron un consentimiento informado, donde se estipuló clara y concisamente el propósito del experimento, el uso y manejo de los datos gestuales recolectados, y los derechos de los participantes.

AÑO 2. No. 4
ISSN:2992-7978

El conjunto de datos de entrenamiento fue conformado por 35 clases o signos, ejecutados con una sola mano. Estos incluyeron las 29 letras del alfabeto LSM (A-Z, LL, RR) y 6 palabras de uso cotidiano ("Comida", "Mucho", "Rápido", "Saber", "Tu", "Yo").

Durante la recolección se colocó el brazalete en la misma posición y orientación del brazo derecho. La piel de los voluntarios y el brazalete fueron limpiados antes de cada recolección. Cada voluntario repitió la ejecución de cada uno de los signos 10 veces, con descansos de 5 segundos entre repeticiones y de 5 minutos entre signos diferentes, para evitar la fatiga muscular. Los voluntarios permanecieron cómodamente sentados durante la ejecución de los signos, y se les pidió que empezaran y terminaran cada ejecución en la misma posición. Al final de la recolección, el conjunto de datos de entrenamiento constó de 4,550 ejecuciones de signos (35 signos × 13 voluntarios × 10 repeticiones).

Tras adquirir el conjunto de datos de entrenamiento, se procedió a construir el modelo de clasificación. La arquitectura del modelo es una versión modificada del clasificador CNN LeNet-5 (LeCun et al., 1998). La capa superior admite imágenes con 128 pixeles de ancho y 128 $\times$ m pixeles de altura, donde m es la cantidad de canales de datos (señales de sensores) a procesar. Como será explicado más adelante, esta decisión en el diseño permite experimentar con la cantidad de características a considerar durante el proceso de inferencia.

Las capas de convolución llevan a cabo la extracción de características, transformando cada imagen en múltiples mapas de características. El submuestreo de las imágenes es realizado mediante capas de agrupación máxima intercaladas con las capas de convolución. Los mapas de características son convertidos a un vector compacto mediante una capa de aplanamiento, el cual es alimentado a una red neuronal artificial para llevar a cabo el reconocimiento del signo. Para reducir las posibilidades de sobre-entrenamiento, se configuró la red neuronal con un abandono del 20%.

Los procesos de entrenamiento y evaluación del modelo de clasificación comiezan con la división del conjunto de datos recolectados en dos partes: entrenamiento (80%) y prueba (20%). Para evitar sesgos en el aprendizaje y obtener una evaluación confiable, ambas particiones contienen el mismo número de clases.

El entrenamiento del modelo de clasificación se lleva a cabo utilizando validación cruzada 10-fold, lo cual implica dividir el conjunto de datos de entrenamiento en 10 particiones (folds) de tamaño similar. En cada iteración, nueve folds se utilizan para entrenar el modelo, mientras que el fold restante se reserva para validación. Este proceso se repite 10 veces, rotando el fold de validación en cada ciclo, de modo que cada partición actúa como conjunto de validación una vez. Al final, los resultados de las 10 iteraciones se

AÑO 2. No. 4
ISSN:2992-7978

promedian para obtener una evaluación robusta del modelo. Este enfoque reduce la dependencia de un único conjunto de validación y mejora la generalización del modelo.

Una vez que se ha afinado el modelo, se procede a evaluar el desempeño de este sobre la partición de datos de prueba a través de un análisis de sustitución. Esta herramienta permite visualizar la capacidad del modelo para detectar de manera correcta cada una de las clases o signos. La métrica más empleada para reportar el desempeño de un sistema SLR es la exactitud (Cheok et al., 2019), la cual se deriva directamente de la matriz de confusión y es aquí utilizada para reportar el poder predictivo o desempeño del modelo. Su calculo considera el número de predicciones correctas entre todas las predicciones realizadas por el modelo de clasificación.

A su vez, el desempeño del modelo puede expresarse en términos de exactitud dependiente del usuario y exactitud por reconocimiento cruzado. Para medir la exactitud exactitud dependiente del usuario, el modelo es entrenado y evaluado usando datos de entrenamiento exclusivamente de un mismo individuo. Por su parte, el reconocimiento cruzado implica tomar datos de entrenamiento de múltiples sujetos. Mientras que un modelo personzalido puede alcanzar altas tasas de reconocimiento al ser utilizado por un usuario especifico, el modelo de reconocimiento cruzado suele tener un desempeño moderado pero más adaptable a múltiples estilos gestuales. Con motivo de comparar el desempeño de este sistema con el de estudios similares, ambas variantes son tomadas en cuenta para el reporte de resultados.

Se utilizó un ordenador Apple Mac M1 con 16 GB de RAM como plataforma de desarrollo y la librería *TensorFlow 2.14.0* de Python para construir, entrenar y evaluar el modelo de clasificación.

Integración de componentes

Para concluir el proceso, el modelo es integrado a la aplicación de Android con la librería *TensorFlowLite 2.14.0*. La aplicación permite la traducción de letras, palabras y oraciones, desplegadas en una ventana tipo chat. Para hacer uso de la herramienta, el usuario de la LSM debe portar el brazalete gForcePro+ y capturar la ejecución de signos individuales, siendo cuatro el número máximo permitido. Las capturas son enviadas al modelo de clasificación, el cual identifica a que signo pertenece cada captura, y aplicando una serie de reglas gramaticales predefinidas, devuelve en tiempo real la traducción en forma de texto, la cual puede ser editada. Una vez conforme con su contenido, el usuario puede añadir el texto a un listado de traducciones.

Ya integrados todos los componentes es posible calcular la latencia de traducción, la cual se mide directamente mediante funciones de programación (tiempo Unix) que registran el momento exacto en que

AÑO 2. No. 4
ISSN:2992-7978

el sistema recibe los gestos a reconocer (tiempo de entrada) y el momento en que genera la salida en formato texto (tiempo de salida). La diferencia entre ambos define la latencia de traducción, medida en milisegundos. De esta manera, si el sistema registrara una petición para traducir los gestos correspondientes a un signo en la marca de tiempo Unix 1732212191e3 y devolviera la traducción en la marca 1732212192e3, la latencia de traducción habrá sido de 1000 milisegundos. El sistema contiene un registro de las traducciones y marcas de tiempo con motivo de facilitar el cálculo de estadísticas relativas al tiempo de latencia, como el promedio por signo.

RESULTADOS

Métricas y evaluación

Los siguientes análisis de datos tienen como objetivo comprobar las afirmaciones contenidas en la hipótesis de esta investigación. Para llevar a cabo esto, el sistema SLR es evaluado en términos de exactitud y latencia de traducción. La primera métrica mide la eficacia en el reconocimiento de señas individuales, mientras que la segunda evalúa la viabilidad de realizar traducciones en tiempo real.

A su vez, el modelo de clasificación es entrenado bajo distintas condiciones o tratamientos con motivo de estudiar como repercute sobre su poder predictivo la elección de algunas variables independientes (técnicas de codificación y clasificación). Para reportar diferencias significativas del desempeño final del modelo entre tratamientos se emplea la prueba t de Student sobre medias independientes (dos colas, $p = .05$), siendo la exactitud por reconocimiento cruzado la variable comparada.

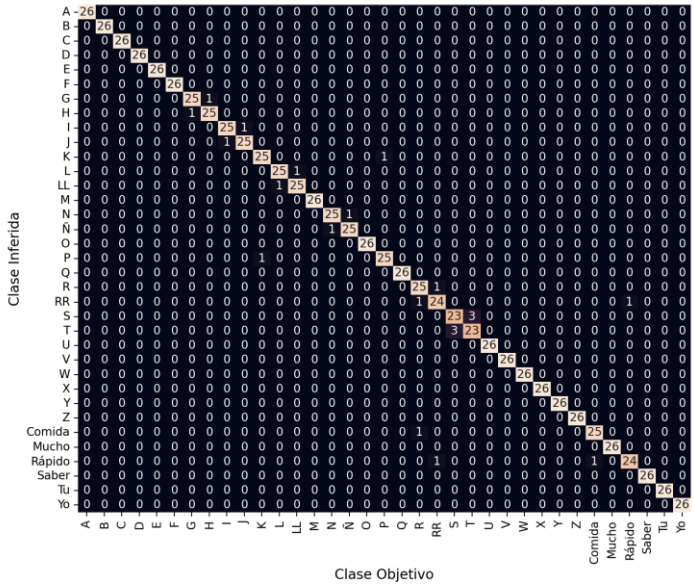
Análisis de sustitución

Como se mencionó brevemente en la metodología, el desempeño del modelo de clasificación es obtenido a través del análisis de sustitución o matriz de confusión, la cual se produce al evaluar el modelo de clasificación con la partición de datos de prueba. Este análisis no solo permite visualizar la exactitud del modelo para clasificar signos de manera independiente, sino que también detalla cuales clases suelen confundirse.

En la Figura 3 se ilustra la matriz de confusión para el modelo de clasificación propuesto, cuya exactitud de reconocimiento cruzado ronda al 97%. Como puede observarse en la figura, los errores de clasificación ocurren entre signos cuyos gestos son parecidos, como sucede entre las letras “S” y “T, y también entre “R” y “RR”.

Figura 3

Matriz de confusión para el modelo de clasificación de reconocimiento cruzado



Nota. La diagonal principal indica la cantidad de signos detectados correctamente para cada clase o signo durante el proceso de evaluación, siendo 26 el número máximo.

Selección de señales de entrada

En la hipótesis se enuncia la factibilidad de emplear sensores vestibles como medio de adquisición de datos gestuales para el sistema SLR propuesto. Esta afirmación asume que todos los canales de datos proporcionados por el brazalete son de utilidad para el modelo de clasificación, aunque existe la posibilidad de que algún sensor o grupo de sensores aporte datos innecesarios o redundantes. Con motivo de probar lo anterior, el modelo es entrenado en múltiples instancias, variando en cada una el tipo de sensor a tomar en cuenta. El modelo también es entrenado considerando todas las señales disponibles, incluyendo la rotación, descrita en ángulos de Euler y cuaterniones. La Tabla 1 muestra una comparativa del desempeño del modelo de clasificación en función de las señales procesadas por el sistema.

Tabla 1

Resultados de la prueba t de Student para comparar la exactitud del modelo entre señales de entrada

AÑO 2. No. 4
ISSN:2992-7978

Señal de entrada	*M*	*SD*	*t*	*p*
Todas	0.968	0.006		
IMU	0.933	0.008	11.181	< .05
EMG	0.804	0.037	13.685	< .05
Acelerómetro	0.523	0.028	49.194	< .05
Giroscopio	0.480	0.025	60.713	< .05

Nota. M = Media, *SD* = Desviación estándar, *t* = valor t, *p* = valor p.

Selección de técnicas de codificación y clasificación

Dentro de la hipótesis se sugiere también que el sistema SLR propuesto podría emplear técnicas de aprendizaje profundo para realizar inferencias con una alta tasa de exactitud. Para probar este punto, el modelo de clasificación fue entrenado de manera independiente mediante los algoritmos CNN y DTW, con motivo de comparar el efecto de emplear redes neuronales y aprendizaje automático tradicional. El modelo DTW consume las series temporales sin procesamiento previo. En cambio, la red neuronal fue entrenada para aceptar como entrada imágenes codificadas mediante las técnicas GAF y MTF, dando lugar a los modelos CNN-GADF, CNN-GASF, y CNN-MTF. La Tabla 2 presenta una comparativa del desempeño del sistema en relación a las técnicas de clasificación y codificación implementadas por el modelo.

Tabla 2

Resultados de la prueba t de Student para comparar la exactitud del modelo entre algoritmos de clasificación y codificación

Modelo	*M*	*SD*	*t*	*p*
CNN-GADF	0.969	0.002		
CNN-GASF	0.967	0.005	1.044	.309
CNN-MTF	0.965	0.005	2.099	.051
DTW	0.911	0.008	19.159	< .05

Nota. M = Media, *SD* = Desviación estándar, *t* = valor t, *p* = valor p.

Latencia de traducción

Los resultados de las pruebas anteriores son útiles para establecer la capacidad predictiva del modelo de clasificación, aunque esto representa solo una parte de los requerimientos del sistema. Otro factor importante a considerar es el tiempo que le toma a un sistema SLR devolver el significado de los signos una vez adquiridos (latencia de traducción), y es que a través de esta métrica es como se puede decretar si el sistema funciona o no en tiempo real. La reducción de la latencia de traducción mejora la fluidez de las conversaciones y la experiencia del usuario al proporcionar retroalimentación casi instantánea durante la comunicación.

Tabla 3

Tiempos de respuesta para el sistema en función de la cantidad de signos a reconocer

Cantidad de signos	Latencia de traducción (segundos)	
	M	*SD*
1	0.15	0.03
2	0.32	0.05
3	0.49	0.05
4	0.70	0.09

Nota. M = Media, *SD* = Desviación estándar, *t* = valor t, *p* = valor p.

DISCUSIÓN

Contextualización de los resultados

Los resultados de la prueba de selección de señales de entrada, desplegados en la Tabla 1, indican que el modelo de clasificación con mejor desempeño es aquel que aprovecha la información contenida en todos los canales de datos del brazalete, consiguiendo una exactitud cercana al 97%, significativamente superior a la que consiguen individualmente los otros sensores.

La elección de algoritmos o técnicas de aprendizaje automático juega un papel tan importante como la adquisición y selección de datos gestuales. Los resultados contenidos en la Tabla 2 revelan que la exactitud del modelo CNN-GADF ($M = 0.969$, $SD = 0.002$) es significativamente más alta, $t(18) = 19.159$, $p < .05$, que la obtenida por el modelo DTW ($M = 0.911$, $SD = 0.008$). La elección de técnica de codificación a imagen no repercute significativamente en el desempeño del modelo. Esto concuerda con los hallazgos de

AÑO 2. No. 4
ISSN:2992-7978

(C. L. Yang et al., 2020), donde los autores implementaron con igual grado de éxito los modelos CNN-GADF, CNN-GASF, y CNN-MTF para clasificar señales capturadas mediante sensores.

El desempeño del algoritmo DTW es también destacable, aunque, así como sucede en otros estudios que emplean esta técnica de clasificación para el reconocimiento de señas, su largo tiempo de inferencia suele suponer una limitante para su implementación en sistemas en tiempo real (Assaleh et al., 2012; Chu et al., 2021; Hou et al., 2019). En contraste, las redes neuronales son capaces de llevar a cabo esta tarea en menos de un segundo (Zhang et al., 2019).

Estudios similares

Al comparar este estudio con trabajos relacionados, se destacan contribuciones y áreas de mejora en varios aspectos clave. El hardware utilizado en el estudio actual es el brazalete gForcePro+, equipado con sensores IMU y sEMG. Esta combinación captura tanto el movimiento como la actividad muscular, mejorando la precisión del reconocimiento de gestos. Configuraciones de sensores similares se utilizan en algunos trabajos relacionados, usualmente en formato de brazalete o guante (Gupta & Kumar, 2021; Khomami & Shamekhi, 2021; Ovur et al., 2021; Wang et al., 2020; Zhang et al., 2019)

Sin embargo, el enfoque en la LSM distingue este estudio de la mayoría de las investigaciones existentes, que predominantemente se centran en la americana (Hou et al., 2019; Zhang et al., 2019) o la china (Ji et al., 2023; Ovur et al., 2021; Wang et al., 2020). Al abordar la LSM, el estudio actual llena un vacío crucial en la investigación de SLR, proporcionando herramientas y metodologías específicamente adaptadas para la comunidad sorda mexicana.

En cuanto al tamaño del vocabulario, el sistema actual reconoce 29 letras del alfabeto y 6 palabras de uso común. Aunque esto es menor que algunos estudios que presentan vocabularios más amplios (como (Gupta & Kumar, 2021) con más de 100 palabras) la inclusión del alfabeto completo permite a los usuarios deletrear palabras no programadas inicialmente en el sistema. Esta característica proporciona flexibilidad y amplía la aplicabilidad del sistema más allá del vocabulario predefinido.

La precisión y dependencia del usuario son métricas críticas en los sistemas SLR. El artículo actual reporta una exactitud dependiente del usuario del 99.5% y una exactitud de reconocimiento cruzado del 97% para palabras, lo que indica alta fiabilidad y robustez entre diferentes usuarios. Estas cifras son comparables o superiores a las de los trabajos relacionados, como (Hou et al., 2019), que reporta una exactitud dependiente del usuario del 99.2% y una exactitud de reconocimiento cruzado del 89.8%. La alta precisión en escenarios de reconocimiento cruzado demuestra su potencial para una adopción generalizada.

AÑO 2. No. 4
ISSN:2992-7978

Una innovación clave en el estudio actual es la metodología novedosa que involucra la conversión de datos gestuales preprocesados en imágenes, que luego se introducen en una red neuronal convolucional para su clasificación. Este enfoque difiere de los utilizados en trabajos relacionados, que a menudo emplean métodos como redes bidireccionales de memoria a corto y largo plazo (Ji et al., 2023; Wang et al., 2020) o variaciones de CNN (Ovur et al., 2021; Suri & Gupta, 2020; Zhang et al., 2019).

En términos de portabilidad y rendimiento en tiempo real, el sistema desarrollado sobresale al lograr un retraso de traducción de solo 0.70 segundos para oraciones compuestas por hasta cuatro señas (Tabla 3), lo cual es aceptable en contextos de comunicación en tiempo real. Sin embargo, este rendimiento es inferior al reportado en los trabajos relacionados (Hou et al., 2019), que reporta un tiempo de traducción de 1.1 segundos para una oración de once palabras, y (Zhang et al., 2019), con 1.43 segundos para ocho palabras.

Limitaciones

Las limitaciones para el uso masivo del sistema SLR propuesto pueden dividirse en aspectos técnicos y sociales, cada uno con desafíos específicos que dificultan su adopción y efectividad.

Desde una perspectiva técnica, uno de los principales problemas es el tamaño reducido de la muestra utilizada para entrenar el modelo de clasificación, lo que limita su capacidad de generalización. Además, el vocabulario detectado por el sistema es insuficiente; trabajar con 29 letras y 6 palabras de uso cotidiano no satisface las necesidades del lenguaje diario, que requiere al menos mil signos para una comunicación fluida. Otro desafío importante es la segmentación manual de las señas para formar oraciones, lo cual es impráctico en un entorno de conversación en tiempo real. A esto se suma la dificultad de reconocer variaciones en las señas, que pueden depender de las expresiones regionales o las diferencias personales en estilo y ejecución.

En cuanto a las limitaciones sociales, las comunidades sordas que más podrían beneficiarse de esta tecnología, como aquellas en áreas rurales o con menos recursos, podrían enfrentar barreras significativas para acceder al dispositivo necesario. Por otro lado, existe una resistencia por parte de algunos miembros de la comunidad sorda, quienes prefieren la interacción con intérpretes humanos. Estos últimos son valorados por su capacidad de comprender mejor el contexto y las sutilezas culturales, aspectos que los sistemas automáticos todavía no logran replicar adecuadamente.

CONCLUSIONES

En esta investigación se desarrolla un sistema de reconocimiento de lengua de señas (SLR) portátil que traduce en tiempo real los signos de la Lengua de Señas Mexicana (LSM) al español. El sistema está basado en la adquisición de datos gestuales mediante un brazalete con sensores vestibles y emplea técnicas de aprendizaje profundo para realizar inferencias. Los resultados obtenidos confirman la hipótesis planteada, ya que el sistema portátil es capaz de identificar signos de la LSM en tiempo real, con una exactitud superior al 90%, cumpliendo así con el objetivo general de este trabajo.

El análisis de selección de señales revela que el modelo de clasificación que utiliza la combinación de todos los sensores disponibles en el brazalete (IMU y EMG) es el que presenta el mejor desempeño, con una exactitud por reconocimiento cruzado cercana al 97%, validando así la pertinencia del uso de sensores vestibles en este tipo de sistemas. Asimismo, el empleo de técnicas de aprendizaje profundo, como las redes neuronales convolucionales (CNN), ofrece una ventaja significativa sobre métodos clásicos como DTW, demostrando que las CNN son capaces de reconocer datos gestuales con mayor eficiencia.

Otro aspecto destacado es que la latencia de traducción del sistema, que promedia 0.70 segundos para la traducción de hasta cuatro signos, confirma la viabilidad de su uso en tiempo real. Esto lo convierte en una herramienta adecuada para facilitar la comunicación entre personas sordas y oyentes, cubriendo una necesidad crítica en la accesibilidad comunicativa.

Por último, se establece que la implementación del sistema propuesto no solo beneficia a los usuarios de la LSM, sino que también sienta un precedente para futuras investigaciones en la creación de sistemas SLR portátiles aplicados a otras lenguas de señas, ampliando el campo de estudio y desarrollo tecnológico en esta área. Se plantea como trabajo futuro incrementar el vocabulario reconocido por el traductor, identificar signos producidos por ambas manos, incorporar mecanismos de atención para el manejo de oraciones sin segmentación manual, y sustituir el uso de reglas gramaticales predefinidas por un modelo lingüístico grande (LLM).

REFERENCIAS

Adeyanju, I. A., Bello, O. O., & Adegboye, M. A. (2021). Machine learning methods for sign language recognition: A critical review and analysis. *Intelligent Systems with Applications*, *12*, 200056. https://doi.org/10.1016/j.iswa.2021.200056

Adler, R. B., Rodman, G., Du Pré, A., & Overton, B. C. (2023). *Understanding human communication* (15th ed.). Oxford University Press. https://global.oup.com/ushe/product/understanding-human-communication-9780197615638?cc=us&lang=en&

Assaleh, K., Shanableh, T., & Zourob, M. (2012). Low Complexity Classification System for Glove-Based Arabic Sign Language Recognition. *Neural Information Processing*, *7665 LNCS*, 262–268. https://doi.org/10.1007/978-3-642-34487-9_32

Bai, Y., & Bruno, D. (2020). Addressing Communication Barriers Among Deaf Populations Who Use American Sign Language in Hearing-Centric Social Work Settings. *Columbia Social Work Review*, *18*(1), 37–50. https://doi.org/10.7916/CSWR.V18I1.5928

Cheok, M. J., Omar, Z., & Jaward, M. H. (2019). A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics*, *10*(1), 131–153. https://doi.org/10.1007/s13042-017-0705-5

Chu, X., Liu, J., & Shimamoto, S. (2021). A sensor-based hand gesture recognition system for Japanese sign language. *2021 IEEE 3rd Global Conference on Life Sciences and Technologies*, 311–312. https://doi.org/10.1109/LifeTech52111.2021.9391981

Garg Pragati, Aggarwal, N., & Sofat Sanjeev. (2009). Vision Based Hand Gesture Recognition. *World Academy of Science, Engineering*. https://doi.org/doi.org/10.5281/zenodo.1074855

Gupta, R., & Kumar, A. (2021). Indian sign language recognition using wearable sensors and multi-label classification. *Computers & Electrical Engineering*, *9*. https://doi.org/https://doi.org/10.1016/j.compeleceng.2020.106898

Hou, J., Wang, Y., Li, X. Y., Qian, J., Zhu, P., Wang, Z., & Yang, P. (2019). SignSpeaker: A real-time, high-precision smartwatch-based sign language translator. *Proceedings of the Annual International Conference on Mobile Computing and Networking, MOBICOM*, *19*. https://doi.org/https://doi.org/10.1145/3300061.3300117

Ji, A., Wang, Y., Miao, X., Fan, T., Ru, B., Liu, L., Nie, R., & Qiu, S. (2023). Dataglove for Sign Language Recognition of People with Hearing and Speech Impairment via Wearable Inertial Sensors. *Sensors 2023, Vol. 23, Page 6693*, *23*(15), 6693. https://doi.org/10.3390/S23156693

Khomami, S. A., & Shamekhi, S. (2021). Persian sign language recognition using IMU and surface EMG sensors. *Measurement*, *168*, 108471. https://doi.org/10.1016/J.MEASUREMENT.2020.108471

Knapp, M. L., Vangelisti, A. L., & Caughlin, J. P. (2014). *Interpersonal Communication and Human Relationships* (7th ed.). Pearson. https://experts.illinois.edu/en/publications/interpersonal-communication-and-human-relationships

Kudrinko, K., Flavin, E., Zhu, X., & Li, Q. (2021). Wearable Sensor-Based Sign Language Recognition: A Comprehensive Review. *IEEE Reviews in Biomedical Engineering*, *14*, 82–97. https://doi.org/10.1109/RBME.2020.3019769

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2323. https://doi.org/10.1109/5.726791

Mayberry, R. I., & Squires, B. (2006). Sign Language: Acquisition. *Encyclopedia of Language & Linguistics*, 291–296. https://doi.org/10.1016/B0-08-044854-2/00854-3

Nimisha, K. P., & Jacob, A. (2020). A Brief Review of the Recent Trends in Sign Language Recognition. *Proceedings of the 2020 IEEE International Conference on Communication and Signal Processing, ICCSP 2020*, 186–190. https://doi.org/10.1109/ICCSP48568.2020.9182351

Ovur, S. E., Zhou, X., Qi, W., Zhang, L., Hu, Y., Su, H., Ferrigno, G., & De Momi, E. (2021). A novel autonomous learning framework to enhance sEMG-based hand gesture recognition using depth information. *Biomedical Signal Processing and Control*, *66*, 102444. https://doi.org/10.1016/J.BSPC.2021.102444

Papastratis, I., Chatzikonstantinou, C., Konstantinidis, D., Dimitropoulos, K., & Daras, P. (2021). Artificial Intelligence Technologies for Sign Language. *Sensors 2021, Vol. 21, Page 5843*, *21*(17), 5843. https://doi.org/10.3390/S21175843

Rastgoo, R., Kiani, K., & Escalera, S. (2021). Sign Language Recognition: A Deep Survey. *Expert Systems with Applications*, *164*, 113794. https://doi.org/10.1016/J.ESWA.2020.113794

Safeel, M., Sukumar, T., Shashank, K. S., Arman, M. D., Shashidhar, R., & Puneeth, S. B. (2020). Sign Language Recognition Techniques- A Review. *2020 IEEE International Conference for Innovation in Technology, INOCON 2020*. https://doi.org/10.1109/INOCON50539.2020.9298376

Suri, K., & Gupta, R. (2020). Convolutional Neural Network Array for Sign Language Recognition using Wearable IMUs. *2019 6th International Conference on Signal Processing and Integrated Networks, SPIN 2019*, 483–488. https://doi.org/10.1109/SPIN.2019.8711745

Wadhawan, A., & Kumar, P. (2019). Sign Language Recognition Systems: A Decade Systematic Literature Review. *Archives of Computational Methods in Engineering*, *28*(3), 785–813. https://doi.org/10.1007/s11831-019-09384-2

Wang, Z., & Oates, T. (2015). *Encoding Time Series as Images for Visual Inspection and Classification Using Tiled Convolutional Neural Networks*. http://coral-lab.umbc.edu/wp-content/uploads/2015/05/10179-43348-1-SM1.pdf

Wang, Z., Zhao, T., Ma, J., Chen, H., Liu, K., Shao, H., Wang, Q., & Ren, J. (2020). Hear Sign Language: A Real-Time End-to-End Sign Language Recognition System. *IEEE Transactions on Mobile Computing*, *21*(7), 2398–2410. https://doi.org/10.1109/TMC.2020.3038303

World Health Organization. (2024). *Deafness and hearing loss*. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

Yang, C. L., Chen, Z. X., & Yang, C. Y. (2020). Sensor Classification Using Convolutional Neural Network by Encoding Multivariate Time Series as Two-Dimensional Colored Images. *Sensors 2020, Vol. 20, Page 168*, *20*(1), 168. https://doi.org/10.3390/S20010168

Yang, C.-L., Yang, C.-Y., Chen, Z.-X., & Lo, N.-W. (2019). Multivariate Time Series Data Transformation for Convolutional Neural Network. *2019 IEEE/SICE International Symposium on System Integration (SII)*, 188–192. https://doi.org/10.1109/SII.2019.8700425

Zhang, Q., Zhao, R., Wang, D., & Yu, Y. (2019). MyoSign: Enabling end-to-end sign language recognition with wearables. *International Conference on Intelligent User Interfaces, Proceedings IUI*, *Part F147615*, 650–660. https://doi.org/10.1145/3301275.3302296