ISSN: 1692-7257 - Volumen 1 – Número 45 - 2025

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Digital Object Identifier: 10.24054/rcta.v1i45.3751

Una arquitectura de aprendizaje profundo multimodal basada en ViT para la clasificación binaria de accidentes de tráfico

A Multi-Modal ViT-Based Deep Learning Architecture for Binary Classification of Traffic Accident

Ing. Jesús David Ríos Pérez[1], PhD. German Sánchez Torres[1], PhD. Carlos Henriquez Miranda[1]

[1] ***Universidad del Magdalena***, *Grupo de Investigación y Desarrollo en Sistemas y Computación, Santa Marta, Magdalena, Colombia.*

Correspondencia: *chenriquezm@unimagdalena.edu.co*

Recibido: *02 abril 2025.* ***Aceptado:*** *05 mayo 2025.* ***Publicado:*** *13 mayo 2025.*

Cómo citar: *J. D. Ríos Pérez, G. Sánchez Torres, y C. Henríquez Miranda, «Una arquitectura de aprendizaje profundo multimodal basada en ViT para la clasificación binaria de accidentes de tráfico», RCTA, vol. 1, n.º 45, pp. 225–239, may 2025. Recuperado de https://ojs.unipamplona.edu.co/index.php/rcta/article/view/3751*

Resumen: Cada año, más de un millón de personas mueren debido a accidentes de tráfico, y un tercio de estas vidas podrían salvarse reduciendo el tiempo de respuesta médica. El aprendizaje profundo multimodal (MMDL) ha surgido en los últimos años como una poderosa herramienta que integra diferentes tipos de datos para mejorar las capacidades de toma de decisiones en los modelos. Además, los Transformadores Visuales (ViT) son un enfoque de aprendizaje profundo para procesar imágenes y videos que ha mostrado resultados prometedores en varias áreas del conocimiento. En este proyecto, proponemos una arquitectura basada en ViT para la clasificación binaria de accidentes de tráfico utilizando datos de múltiples fuentes, como datos ambientales e imágenes. La integración de un enfoque MMDL basado en ViT puede mejorar la precisión del modelo en la clasificación de accidentes y no accidentes. Este proyecto explora un enfoque MMDL integrando ViT para la monitorización de accidentes de tráfico en el contexto de las ciudades inteligentes, logrando un recall del 91%, lo que evidencia una alta robustez del modelo en la identificación de casos positivos. Sin embargo, la escasez de datos multimodales representa un gran desafío para el entrenamiento de este tipo de modelos.

Palabras clave: Multimodal, Aprendizaje profundo, Transformadores visuales, Accidentes de Tránsito.

Abstract: Each year, more than 1 million people die due to traffic accidents, and one-third of these lives could be saved by reducing medical response time. Multi-Modal Deep Learning (MMDL) has emerged in recent years as a powerful tool that integrates different types of data to enhance decision-making capabilities in models. Additionally, Vision Transformers (ViT) are a Deep Learning approach for processing images and videos that has shown promising results in various fields of knowledge. In this project, we propose a ViT-based architecture for binary classification of traffic accidents using data from multiple

sources, such as environmental data and images. The integration of an MMDL approach based on ViT can improve the model's accuracy in classifying accidents and non-accidents. This project explores a MMDL approach integrating ViT for traffic accident monitoring in the context of smart cities, achieving a recall of 91%, which evidences a high robustness of the model in identifying positive cases. However, the scarcity of multimodal data represents a major challenge for training these types of models.

Keywords: Multimodal, Deep Learning, Vision Transformers, Traffic Accident.

1. INTRODUCCIÓN

Según la Organización Mundial de la Salud (OMS) [1], cada año, aproximadamente 1,19 millones de personas mueren en todo el mundo debido a accidentes de tráfico. Además, entre 20 y 50 millones sufren lesiones no mortales, que a menudo resultan en discapacidades a largo plazo. Los accidentes de tráfico son la principal causa de muerte entre las personas de 5 a 29 años y la octava causa principal de muerte en todos los grupos de edad. Esta situación se ha agravado en países como Colombia, donde los accidentes de tráfico son la segunda causa principal de muertes violentas, con un aumento anual del número de víctimas mortales [2]. Aunque el 60% de los vehículos se concentran en países de ingresos medios y bajos, el 92% de las muertes relacionadas con el tráfico ocurren en estas regiones. Estos accidentes también resultan en pérdidas económicas para personas, familias y naciones en su conjunto. Además, existe una deficiencia significativa en la respuesta oportuna tras las colisiones: las demoras en detectar la necesidad de asistencia y en brindar ayuda aumentan la gravedad de las lesiones. En la respuesta de emergencia a estos accidentes, el tiempo de reacción es vital: tan solo unos minutos de retraso pueden determinar la vida o la muerte de una persona; una reducción de 10 minutos en el tiempo de respuesta médica se asocia estadísticamente con una disminución de un tercio en la probabilidad de muerte en la carretera [3].

Por otro lado, la fusión de datos se refiere a la integración de datos de diferentes fuentes o modalidades para obtener múltiples perspectivas sobre un fenómeno común y abordar un problema específico. Estas modalidades son complementarias, ya que proporcionan información desde diferentes perspectivas del fenómeno. El objetivo de estas estrategias de fusión es aprovechar la complementariedad, la redundancia y las características cooperativas entre las diferentes modalidades. Recientemente, estos enfoques de aprendizaje automático multimodal (MMML) se han estudiado y aplicado cada vez más en diversos campos [4]-[9].

Los modelos de aprendizaje profundo no solo han demostrado avances tecnológicos significativos, sino que también se han expandido a las aplicaciones de MMDL. Hoy en día, estos métodos están a la vanguardia de la innovación, abordando desafíos complejos en campos como el reconocimiento de voz audiovisual y la recuperación de contenido multimedia para el análisis de la salud y los estudios de interacción social.

Un enfoque MMDL presenta varios desafíos, incluida la representación, la traducción, la alineación, la fusión y el coaprendizaje cuando se aprende a partir de dos o más modalidades [10], [11]. Los modelos MMDL combinan datos heterogéneos de múltiples fuentes, lo que permite predicciones más precisas. Sin embargo, la precisión y la flexibilidad de estos sistemas no son óptimas debido a la cantidad insuficiente de datos etiquetados [12]. Además, este enfoque multimodal se ha investigado desde la década de 1970 y se ha categorizado en cuatro eras distintas: la Era del Comportamiento (de 1970 a 1980), la Era Computacional (de finales de la década de 1980 a 2000), la Era Interaccional (de 2000 a 2010) y la Era del Aprendizaje Profundo (de 2010 a la actualidad). Asimismo, el MMDL se ha aplicado a diversos campos, como la comprensión de los comportamientos multimodales humanos durante la interacción social y el reconocimiento multimodal de emociones [13], [14], Preguntas y respuestas visuales (VQA) [15], reconocimiento de voz audiovisual (AVSR) [16], subtítulos de imágenes y vídeos [17], [18], indexación y recuperación de contenido multimedia [19]-[21], y análisis de salud [22].

Hay tres tipos de fusión en MMDL: Fusión temprana, donde todas las modalidades se combinan en la etapa inicial y el modelo aprende de estas modalidades combinadas; Fusión media, donde las modalidades se transforman en un espacio común en

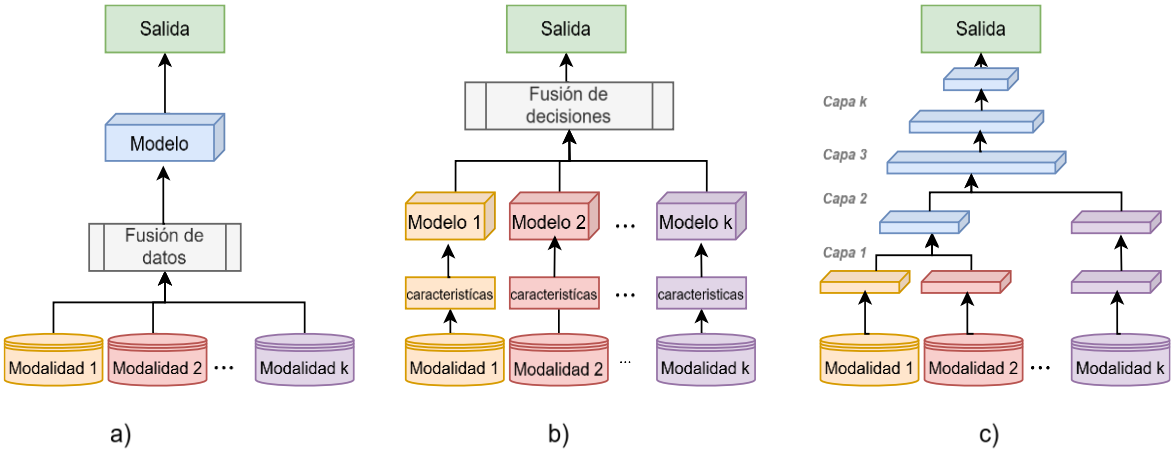

***Fig. 1**. En problemas multimodales, hay tres tipos de fusión: a) Fusión temprana, que combina las modalidades al inicio en un solo vector; b) Fusión intermedia, que aprende representaciones de estas modalidades antes de fusionarlas para que el modelo pueda aprender de ellas en conjunto; y c) Fusión tardía, que combina las predicciones hechas por cada modelo para sus modalidades correspondientes.*
***Fuente**: adaptado de [7].*

lugar de simplemente concatenarse (este tipo de fusión se aplica en proyectos de filtrado como sistemas de recomendación); y Fusión tardía, donde el modelo aprende las modalidades de forma independiente y luego las combina antes de tomar una decisión final (este enfoque es importante cuando una modalidad es dominante) [23]. Ver Fig 1.

El objetivo de este estudio es diseñar una arquitectura multimodal basada en aprendizaje profundo para la clasificación binaria de accidentes de tráfico, que tradicionalmente se ha aplicado únicamente a vídeos o imágenes de entrada. La hipótesis del proyecto es que la integración de recursos multimodales, como datos tabulares e imágenes, podría mejorar significativamente la precisión de la clasificación de accidentes de tráfico. Además, se utilizan Transformadores de Visión (ViT) para aprovechar el potencial de los datos visuales. Este proyecto demuestra resultados excepcionales en la clasificación binaria de accidentes y no accidentes, logrando una recuperación superior al 90%. Por otro lado, la escasez de datos multimodales sigue siendo un reto actual para el entrenamiento de estos modelos. Técnicas como el aprendizaje por transferencia (Transfer Learning) y la aumentación de datos ayudan a mitigar este problema, permitiendo que el modelo se generalice con mayor eficacia.

La estructura del presente artículo se detalla a continuación: la Sección 2 revisa el trabajo relacionado realizado en esta área. La Sección 3 describe la metodología propuesta para definir la arquitectura de la red neuronal profunda para la detección de accidentes de tráfico. La Sección 4 presenta los resultados experimentales y el proceso para determinar los hiper parámetros del modelo propuesto. La Sección 5 analiza el sesgo, la generalización y las consideraciones éticas clave. Finalmente, la Sección 6 resume las principales conclusiones y describe las líneas de investigación futuras.

2. ANTECEDENTES

MMDL se ha explorado ampliamente en diversas aplicaciones, demostrando su adaptabilidad y potencial para mejorar el rendimiento del modelo mediante la integración eficaz de información de múltiples modalidades de datos. Esta versatilidad permite una comprensión más completa de escenarios complejos, ya que aprovecha características complementarias de diferentes fuentes. En [12], presentan una taxonomía para identificar los métodos más relevantes y sus áreas de aplicación: Descripción de Imágenes Multimodales [24], [25]-[31], Descripción de Vídeo Multimodal [32]-[45], Preguntas y Respuestas Visuales Multimodales [46]-[57], Síntesis de Voz Multimodal [58]-[75] y otras aplicaciones MMDL.

Además, diferentes campos incorporan datos multimodales en sus problemas, como aplicaciones clínicas (texto, voz, imágenes y vídeos) [76]-[91]; Fusión de Datos de Teledetección (pancromática

[92]-[96], Multiespectral [97]-[99], Hiper espectral [100]-[101], Detección y Medición de Distancias por Luz [92], [102], [103], Radar de Apertura Sintética [104]-[109], datos de infrarrojos, luz nocturna y vídeo satelital) [110]; reconocimiento de emociones (texto, imágenes y audio) [73]. Esto coloca de relieve la amplia exploración y versatilidad de este método en diferentes áreas del conocimiento, logrando resultados sobresalientes en cada uno de estos campos.

Algunos proyectos se han centrado en los Transformadores de Visión (ViT) para abordar la complejidad del análisis y procesamiento de imágenes. Xu et al [111] sentó las bases del diseño de ViT y su aplicación en diversas tareas de visión de alto y bajo nivel, como la generación de imágenes y el aprendizaje multimodal. Este trabajo preserva la capacidad de estos modelos para capturar y representar información a largo plazo, destacando su rendimiento en diferentes tipos de tareas. En [112], se realizó una revisión exhaustiva de ViT, centrándose en tareas fundamentales de visión por computadora, destacando mejoras significativas en los puntos de referencia en comparación con las redes neuronales convolucionales (CNN) tradicionales. ViT también se ha utilizado en un contexto multimodal [113], este estudio propone principalmente métodos para reducir la complejidad computacional y el número de parámetros del modelo. En [114], se propone una variante jerárquica de ViT para mejorar la eficiencia computacional mediante la agrupación de tokens visuales, lo que también impacta positivamente en la escalabilidad del modelo en diferentes dimensiones, incluida la profundidad, el ancho y la resolución. En [115], las limitaciones de las codificaciones de posición condicional se abordan mediante la introducción de un esquema gestionado de forma adaptativa para secuencias de entrada de longitud variable, lo que mejora la flexibilidad de estos modelos. Fang, J [116], se propuso una arquitectura innovadora que integra tokens especializados para el intercambio local de información espacial entre regiones de una imagen. Esto ayuda a reducir la demanda computacional y a mejorar el rendimiento en áreas críticas como la clasificación de imágenes y la detección de objetos. Con base en lo anterior, la representación de la imagen se redefine mediante tokens centrados en ViT para delinear relaciones semánticas, lo que finalmente conduce a una mejora significativa en la clasificación y segmentación semántica propuestas [117]. De la misma manera, el desarrollo de métodos de visualización e interpretación ha desvelado las complejidades operativas de los Transformadores de Visión [118], proporcionando información importante sobre el proceso de toma de decisiones [119]. Considerando lo anterior, estas contribuciones ayudan a mejorar la clasificación de modelos con ViT al mejorar su desempeño, eficiencia, versatilidad e interpretabilidad en Vision Transformers.

En [120], se utilizó una arquitectura de aprendizaje profundo basada en Conv-LSTM para la detección automática de accidentes de tráfico con un conjunto de datos basado en vídeos de carretera. Por otro lado, se propuso un modelo de aprendizaje profundo que utiliza CNN, LSTM y AE basado en sensores multimodales para la detección de accidentes [121]. Además, se encontró un modelo basado en LLM y VLM preentrenados para la detección automática de accidentes de tráfico [122].

Se han identificado enfoques multimodales en diversos campos, lo que aumenta la precisión de los modelos. Además, ViT se ha utilizado para resolver diversos problemas de visión, mejorando el rendimiento, la versatilidad y la interpretabilidad. En concreto, en la detección de accidentes de tráfico, se han implementado diversos enfoques basados en aprendizaje profundo y redes multimodales.

3. METODOLOGÍA

La metodología para este proyecto de clasificación binaria utilizando Deep Learning Multimodal ViT se divide en cinco etapas: Modalidades (selección de variables de entrada), Diseño de la Arquitectura, Preprocesamiento, Implementación y Evaluación (ver Fig 2).

Fig. 2*. Metodología para proyecto de clasificación binaria utilizando Aprendizaje Profundo Multimodal ViT.*
Fuente*: elaboración propia.*

3.1. Modalidades

En esta sección se presenta el conjunto de datos junto con las modalidades o variables de entrada utilizadas en el sistema de clasificación binaria para la detección de accidentes de vehículos, que serán de utilidad para la arquitectura propuesta con base en los resultados de la sección 2 (Antecedentes).

Las imágenes de la carretera proporcionan información visual crucial en tiempo real para la detección de accidentes, incluyendo las condiciones del tráfico, la señalización, el comportamiento del conductor, los posibles accidentes y más. Por otro

lado, los datos tabulares de tráfico proporcionan información relacionada con un accidente de tráfico en relación con las condiciones de la carretera, como la velocidad media, la densidad de vehículos y otras métricas. Además, los datos meteorológicos proporcionan información en tiempo real sobre las condiciones meteorológicas que podrían afectar la carretera, como la lluvia, la niebla, la nieve y otros factores. La importancia y la mejora en la arquitectura de cada modalidad se presentan en la Tabla 1. Para recopilar imágenes de accidentes/no accidentes, se utilizó la técnica de web scraping para extraer imágenes de Google mediante la ecuación (1). Posteriormente, este conjunto de datos se optimizó seleccionando imágenes de calidad aceptable.

Ecuación de búsqueda = "((accident OR incident OR collision OR crash) AND (traffic OR vehicle OR automobile)) OR ((traffic) AND (vehicle or automobile))" (1)

3.2. Diseño de la Arquitectura

El diseño de la arquitectura se basó en una arquitectura de fusión intermedia (véase la Fig. 1), que incorpora las modalidades mencionadas en la Sección 3.1. Cada modalidad se procesó individualmente antes de la fusión. Se utilizó ViT para el procesamiento de imágenes con el fin de optimizar la toma de decisiones de la red, mientras que los datos tabulares se preprocesaron mediante un Perceptrón Multicapa (MLP).

3.3. Preprocesamiento

En la etapa de preprocesamiento, se aplicó la normalización min-max a cada entrada utilizando la ecuación (2), donde X representa la variable de entrada y *min(I)* y *max(I)* son sus valores mínimo y máximo, respectivamente:

$$X' = \frac{X - min\,(X)}{max(X) - min\,(X)} \qquad (2)$$

Debido a la escasez de datos, se emplearon técnicas de aumento de datos para mejorar el rendimiento de la red y generalizar el conjunto de datos de entrenamiento a datos no vistos. Otra ventaja es que el aumento de datos ayuda a reducir el sobreajuste y mejora la robustez del modelo.

Tabla 1. *Resumen de las características y sus modalidades con su importancia e impacto en la arquitectura.*

Modalidad de datos	Importancia	Tipo de modalidad	Mejora en la Arquitectura
Imagen de la Carretera	Proporciona información visual crucial sobre las condiciones del tráfico, el comportamiento de los conductores y posibles incidentes. Esencial para identificar patrones de movimiento, obstrucciones y la dinámica de los accidentes.	Visual	Facilita el análisis en tiempo real del entorno de la carretera, permitiendo la detección temprana de condiciones anómalas o peligrosas a través del procesamiento de imágenes y visión artificial.
Datos Tabulares del Tráfico	Ofrece una visión cuantitativa del estado del tráfico, incluyendo velocidades promedio, densidad de vehículos y otras métricas relevantes. Esencial para comprender las condiciones generales del tráfico e identificar desviaciones de los patrones normales.	Tabular	Permite que la arquitectura correlacione las condiciones del tráfico con la probabilidad de incidentes, mejorando la precisión de la predicción.
Datos Meteorológicos	Afecta significativamente la seguridad vial. La integración de estos datos ayuda a contextualizar las observaciones visuales y tabulares.	Tabular	Proporciona contexto adicional para interpretar datos visuales y tabulares, lo que permite realizar ajustes en la detección de accidentes según las condiciones climáticas.

Fuente: *elaboración propia.*

3.4. Implementación

El modelo se mejoró mediante experimentación, ajustando los siguientes hiper parámetros: Batch size, Learning rate, optimizador, Weight decay, número de épocas, Patch size (ViT), número de cabezales (ViT), número de capas de transformador (ViT), Dropout rate y función de activación. Cabe destacar que los hiper parámetros etiquetados con (ViT) solo se aplican a las capas de procesamiento de imágenes que utilizan transformadores. Además, el proyecto se desarrolló principalmente utilizando las bibliotecas NumPy, PyTorch y Torchvision para el procesamiento de datos y el entrenamiento del modelo.

3.5. Evaluación

La evaluación del modelo se basó en tres métricas: Exactitud, que mide la proporción de predicciones correctas sobre el número total de predicciones (véase la ecuación 3); Recall, que mide la proporción de casos positivos correctamente identificados por el modelo (véase la ecuación 4); y F1-Score, que calcula la media armónica de la precisión y el recall, proporcionando un equilibrio entre ambas métricas. Esto resulta especialmente útil cuando se requiere un equilibrio entre la precisión y el recall (véase la ecuación 5). Además, se empleó la técnica de validación cruzada para abordar la escasez de datos, el sobreajuste y la mala generalización del modelo.

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F1\_Score = 2 \times \frac{precision \times Recall}{Precision + Recall} \quad (5)$$

4. RESULTADOS

4.1. Modalidades

En esta sección se presentan los resultados de cada etapa de la metodología: Modalidades, Diseño de Arquitectura, Preprocesamiento, Implementación y Evaluación (ver Sección 3).

Con base en la Tabla 1, se utilizó el Web Scraping para extraer imágenes de internet mediante la ecuación de búsqueda (1). Tras depurar el conjunto de datos, se recopilaron 635 imágenes de accidentes y 663 imágenes de otros tipos, lo que resultó en un total de 1298 imágenes.

Para la generación sintética de datos tabulares, utilizamos el modelo LLaVA-v1.5-7b [123] basado en las variables tabulares de la Tabla 2, el modelo LLaVA se evaluó utilizando imágenes y la indicación de la Ecuación (6). Estas variables descriptivas se seleccionaron por su relación directa con la probabilidad de accidente: ‘tiempo_del_día’, ‘condiciones_meteorológicas’, ‘estado_de_la_vía’, ‘volumen_del_tráfico’, ‘señales_de_tráfico’, ‘obstáculos_viales’ y ‘estado_de_la_iluminación_de_la_vía’. Los datos generados fueron verificados manualmente por un profesional.

Table 2. *Used Tabular Variables for the Multimodal Accident Binary Classification Model.*

Variable	Categorías	Tipo
tiempo_del_día	Día = 1, Noche = 0	Binario
condiciones_meteorológica	Claro = 0, Nublado = 1, Lluvioso = 2, Nevado = 3, Niebla = 4, Otro = 5	Nominal, One-hot
condiciones_de_la_vía	Seco = 0, Mojado = 1, Congelado = 2, Acumulación de agua = 3, Acumulación de nieve = 4	Nominal, One-hot
traffic_volume	Bajo = 0, Moderado = 1, Alto = 2	Ordinal
traffic_signs	Visible = 1, No visible = 0	Binario
obstaculos_de_la_vía	Ninguna = 0, Escombros = 1, Animales = 2	Nominal, One-hot
condición_de_la_iluminación_en_la_vía	Adecuado = 1, Inadecuada = 0, Defectos visibles en la carretera = 2	Ordinal

Fuente: *elaboración propia.*

$$Prompt = \text{"}Evaluates\ the\ image\ and\ assigns\ the\ integer\ value\ corresponding\ to\ each\ characteristic\ according\ to\ the\ present\ conditions.\text{"} \quad (6)$$

4.2. Diseño de la Arquitectura

La arquitectura se diseñó con el objetivo de crear un modelo que clasifique si se ha producido un accidente a partir de una imagen y sus metadatos (datos tabulares: día, condiciones meteorológicas, estado de la vía, volumen de tráfico, señales de tráfico, obstáculos en la vía y estado del alumbrado público). Estas modalidades se explican en la Sección 3.1.

La arquitectura se basa en un enfoque de fusión intermedia, donde el modelo procesa y aprende cada modalidad (imágenes de carretera y datos tabulares) individualmente. Posteriormente, se fusionan mediante una incrustación, y la red aprende de estos datos fusionados utilizando una capa basada en MLP para tomar una decisión (accidente o no accidente). Véase la Fig. 3.

Las imágenes de la carretera se procesan mediante una capa de extracción de parches y se transforman en un espacio semántico mediante la incrustación de parches. Posteriormente, se procesan mediante capas de transformación, cada una compuesta por dos capas de normalización, un mecanismo de atención multicabezal y un MLP. Estas características se fusionan con los datos tabulares para aprovechar al máximo las características de la imagen, ya que es la modalidad semánticamente más importante para la detección de accidentes. Por otro lado, los datos tabulares se procesan mediante un MLP bidimensional antes de fusionarse con las

imágenes de la carretera. Estos datos se concatenan y se pasan por una incrustación fusionada, lo que permite a la red aprender de los datos combinados mediante un MLP.

4.3. Preprocesamiento

En esta etapa, se aplicó primero la normalización a todas las modalidades de entrada, como se define en la Sección 3.3. Además, todas las imágenes se redimensionaron a un ancho y una altura de 224. Debido a la escasez de datos, se realizó un aumento de datos para aumentar la variabilidad del conjunto de datos y mejorar la generalización del modelo.

El aumento de datos se aplicó por separado a cada modalidad (imágenes y datos tabulares). Se aplicaron tres tipos de transformaciones geométricas a las imágenes: rotación horizontal aleatorio, rotación aleatoria (con un ángulo máximo de rotación de 15 grados) y filtros de color (brillo = 0,2, contraste = 0,2, saturación = 0,2), cada uno con una probabilidad independiente del 50 %. Consulte la Tabla 3 para obtener un resumen de las transformaciones geométricas. Para los datos tabulares, se utilizó la técnica SMOTE para generar la misma cantidad de datos aumentados que en las imágenes

Tabla 3. *Resumen de transformaciones geométricas para la ampliación de datos en imágenes de carreteras.*

Transformación	Valor	Probabilidad
Rotación horizontal Random		50%
Rotación Random	15°	50%
Brillo	0.2	50%
Contraste	0.2	50%
Saturación	0.2	50%

Fuente: *elaboración propia.*

4.4. Implementación

Tras la recopilación de datos, el diseño de la arquitectura y el preprocesamiento de los datos (véanse las secciones 4.1, 4.2 y 4.3), se implementa la arquitectura multimodal basada en ViT. Los hiper parámetros utilizados en el modelo se determinaron tras varias iteraciones. La Tabla 4 presenta los hiper parámetros finales. Si bien existe una amplia variedad de parámetros posibles, aquí solo se listan los más relevantes para este problema: Batch size, Learning rate, optimizador, Weight decay, Patch size, dimensionalidad del codificador del transformador, número de cabezales, número de capas del transformador, Dropout rate y función de activación. Es importante destacar que estos hiper parámetros dependen de la cantidad de datos, la potencia computacional y las modalidades utilizadas para entrenar el modelo.

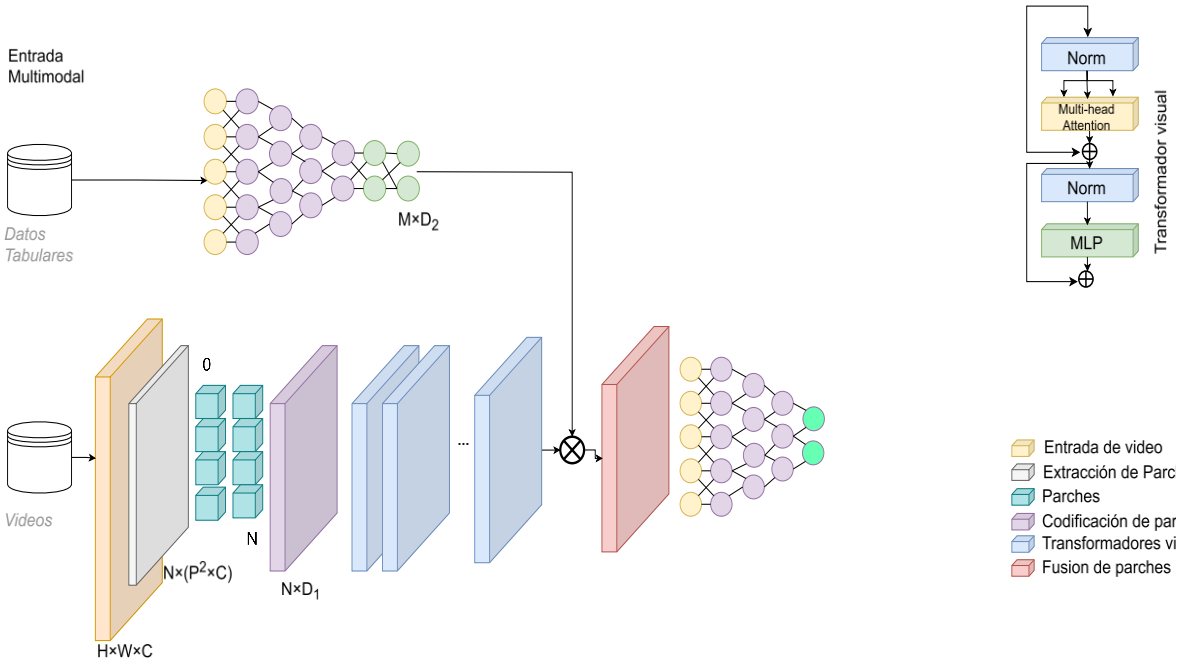

Fig. 3*. Architecture Design for a binary classification model based on Multimodal Deep Learning ViT.*
Source*: own elaboration.*

Tabla 4. *Configuraciones de hiper parámetros para el modelo de clasificación binaria de accidentes ViT de aprendizaje profundo multimodal.*

Hiper parámetro	Valor	Nota
Batch size	32	Depende de los recursos computacionales y de la memoria de la GPU.
Learning rate	1e-4	Se puede mejorar utilizando un programador de tasa de aprendizaje que disminuya la tasa con el tiempo.
Optimizador	AdamW	AdamW se utiliza comúnmente con transformadores debido a su manejo de la caída de peso.
Weight decay	0.01	Regularización para evitar el sobreajuste.
Patch size (Video)	32x32	Dependiendo de los recursos computacionales y la memoria de la GPU.
Dimensionalidad del codificador del transformador (*d_model*)	768	Dependiendo de la cantidad de datos.
Número de cabezales	8	Número de cabezas de atención en el transformador (debe ser divisible por *d_model*).
Número de capas de transformadores	8	Se puede ajustar; más capas capturan patrones más complejos, pero aumentan la demanda computacional
Dropout rate	0.1	Ayuda con la regularización (ajuste según el nivel de sobreajuste observado).
Función de activación	RELU o GELU	GELU se utiliza a menudo en modelos de transformadores, pero RELU es una buena alternativa.

Fuente: *elaboración propia.*

4.5. Evaluación

El modelo se entrenó y evaluó utilizando un 80% de datos de entrenamiento y un 20% de datos de validación, con 20 épocas de entrenamiento. Además, se aplicó la técnica K-Folds, con K establecido en 5. El modelo se evaluó utilizando las tres métricas presentadas en la Sección 3.5 (Exactitud, F1-Score y Recall). La Tabla 5 presenta los resultados promedio de estas métricas.

Los valores de pérdida comienzan en 0,3868 para el entrenamiento y 0,0766 para la validación, y terminan en 0,054077 ± 0,001109 y 0,052353 ± 0,002453, respectivamente. Por otro lado, la métrica de exactitud promedio, que mide el número de casos correctamente clasificados, muestra valores muy similares para cada división de datos, aproximadamente el 75% para el entrenamiento y el 78% para la validación. Asimismo, el F1-Score promedio oscila entre el 78% para el entrenamiento y el 80% para la validación. Finalmente, es importante destacar que la métrica de recall promedio alcanza el 93% para el entrenamiento y el 91% para la validación. Esta última métrica es crucial, ya que enfatiza la identificación de casos positivos, considerando que la detección temprana podría salvar una vida o prevenir el empeoramiento de una lesión.

La pequeña diferencia entre los valores de las métricas indica la ausencia de sobreajuste a los datos de entrenamiento en comparación con los datos de validación. El uso de la técnica K-Folds ayudó a mitigar los problemas relacionados con la división de datos y el sobreajuste al conjunto de entrenamiento.

Tabla 5. *Valores de las métricas promedio de la última época durante K Folds divididos por datos de entrenamiento y validación.*

Métrica	Entrenamiento	Validación
Avg. Pérdida	0.054077 ± 0.001109	0.052353 ± 0.002453
Avg Exactitud	0.759538 ± 0.009293	0.785385 ± 0.019822
Avg. F1-Score	0.788745 ± 0.006164	0.804784 ± 0.013133
Avg. Recall	0.931600 ± 0.003878	0.918400 ± 0.019200

Fuente: *elaboración propia.*

5. DISCUSIÓN

Los enfoques multimodales mejoran el rendimiento del modelo al aprovechar las diferentes características de la escena. Además, la incorporación de capas ViT mejora la precisión. Sin embargo, la falta de datos multimodales en la clasificación de accidentes y otros escenarios sigue siendo un desafío. Si bien la síntesis de datos ayuda a mitigar este problema, depende de la modalidad principal, lo que podría introducir sesgos en el modelo.

Por otro lado, procesar más datos para cada modalidad involucrada requiere mayor potencia computacional, lo que convierte la capacidad de procesamiento en una de las principales limitaciones. Sin embargo, el uso de redes preentrenadas podría ayudar a mitigar los problemas relacionados con la falta de un volumen sustancial de datos y reducir el tiempo de entrenamiento.

El uso de proyectos multimodales está en auge. Sin embargo, persiste un desafío importante debido a la falta de datos multimodales disponibles. Se espera que este problema se mitigue en los próximos años a medida que se disponga de más datos.

6. CONCLUSIONES

La exploración de enfoques multimodales en aprendizaje profundo destaca su notable evolución, versatilidad e impacto potencial en una amplia gama de ámbitos, desde la atención médica hasta los entornos de monitoreo social. Esta progresión ha abarcado desde los primeros marcos conceptuales hasta el uso actual de técnicas de aprendizaje profundo. Cabe destacar que los Transformadores de Visión (ViT) han demostrado un rendimiento excepcional en diversos tipos de implementaciones.

Nuestro diseño arquitectónico para la clasificación binaria de accidentes de tráfico emplea fusión intermedia, lo que permite procesar individualmente las características de cada modalidad antes de combinarlas para la toma de decisiones. La fusión de datos desempeña un papel crucial en el rendimiento de la red, por lo que su selección es esencial y depende en gran medida del problema específico. Estudios previos destacan la amplia gama de arquitecturas y métodos utilizados en diseños multimodales, lo que sugiere que no existe un marco único que se adapte a todas las aplicaciones.

El modelo demostró un rendimiento sólido a pesar de la cantidad limitada de datos y la ausencia de modalidades adicionales. Aumentar el volumen de datos durante la adquisición e incorporar más modalidades podría mejorar el rendimiento de la red en trabajos futuros. Además, estrategias como el aprendizaje por transferencia podrían contribuir a robustecer el modelo

REFERENCIAS

[1] «Traumatismos causados por el tránsito». Accedido: 18 de marzo de 2025. [En línea]. Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/road-traffic-injuries

[2] M. T. Pulgarín *et al.*, «Autores: Agencia Nacional de Seguridad Vial».

[3] R. Sánchez-Mangas, A. García-Ferrrer, A. de Juan, y A. M. Arroyo, «The probability of death in road traffic accidents. How important is a quick medical response?», *Accid. Anal. Prev.*, vol. 42, n.º 4, pp. 1048-1056, jul. 2010, doi: 10.1016/j.aap.2009.12.012.

[4] Y. Li, F.-X. Wu, y A. Ngom, «A review on machine learning principles for multi-view biological data integration», *Brief. Bioinform.*, vol. 19, n.º 2, pp. 325-340, mar. 2018, doi: 10.1093/bib/bbw113.

[5] C. Manzoni *et al.*, «Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences», *Brief. Bioinform.*, vol. 19, n.º 2, pp. 286-302, mar. 2018, doi: 10.1093/bib/bbw114.

[6] «Milestones in Genomic Sequencing». Accedido: 23 de noviembre de 2023. [En línea]. Disponible en: https://www-nature-com.biblioteca.unimagdalena.edu.co/immersive/d42859-020-00099-0/index.html

[7] S. R. Stahlschmidt, B. Ulfenborg, y J. Synnergren, «Multimodal deep learning for biomedical data fusion: a review», *Brief. Bioinform.*, vol. 23, n.º 2, p. bbab569, ene. 2022, doi: 10.1093/bib/bbab569.

[8] «Single-cell multiomics: technologies and data analysis methods | Experimental & Molecular Medicine». Accedido: 23 de noviembre de 2023. [En línea]. Disponible en: https://www-nature-com.biblioteca.unimagdalena.edu.co/articles/s12276-020-0420-2

[9] T. Baltrušaitis, C. Ahuja, y L.-P. Morency, «Multimodal Machine Learning: A Survey and Taxonomy», *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, n.º 2, pp. 423-443, feb. 2019, doi: 10.1109/TPAMI.2018.2798607.

[10] «The promise and challenges of multimodal learning analytics», doi: 10.1111/bjet.13015.

[11] D. Hong *et al.*, «More Diverse Means Better: Multimodal Deep Learning Meets Remote Sensing Imagery Classification», *IEEE Trans. Geosci. Remote Sens.*, vol. 59, n.º 5, pp. 4340-4354, may 2021, doi: 10.1109/TGRS.2020.3016820.

[12] S. Jabeen, X. Li, M. S. Amin, O. Bourahla, S. Li, y A. Jabbar, «A Review on Methods and Applications in Multimodal Deep Learning», *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 19, n.º 2s, p. 76:1-76:41, feb. 2023, doi: 10.1145/3545572.

[13] *Proceedings of the 2020 International Conference on Multimodal Interaction*. Association for Computing Machinery, 2020.

[14] J. Chen *et al.*, «HEU Emotion: a large-scale database for multimodal emotion recognition in the wild», *Neural Comput. Appl.*, vol. 33, n.º 14, pp. 8669-8685, jul. 2021, doi: 10.1007/s00521-020-05616-w.

[15] «Improving reasoning with contrastive visual information for visual question answering - Long - 2021 - Electronics Letters - Wiley Online Library». Accedido: 19 de noviembre de 2023. [En línea]. Disponible en: https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/ell2.12255

[16] B. P. Yuhas, M. H. Goldstein, y T. J. Sejnowski, «Integration of acoustic and visual speech signals using neural networks», *IEEE Commun. Mag.*, vol. 27, n.º 11, pp. 65-71, nov. 1989, doi: 10.1109/35.41402.

[17] S. Bai y S. An, «A survey on automatic image caption generation», *Neurocomputing*, vol. 311, pp. 291-304, oct. 2018, doi: 10.1016/j.neucom.2018.05.080.

[18] «Future Internet | Free Full-Text | Video Captioning Based on Channel Soft Attention and Semantic Reconstructor». Accedido: 19 de noviembre de 2023. [En línea]. Disponible en: https://www.mdpi.com/1999-5903/13/2/55

[19] R. Souza, A. Fernandes, T. S. F. X. Teixeira, G. Teodoro, y R. Ferreira, «Online multimedia retrieval on CPU–GPU platforms with adaptive work partition», *J. Parallel Distrib. Comput.*, vol. 148, pp. 31-45, feb. 2021, doi: 10.1016/j.jpdc.2020.10.001.

[20] P. K. Atrey, M. A. Hossain, A. El Saddik, y M. S. Kankanhalli, «Multimodal fusion for multimedia analysis: a survey», *Multimed. Syst.*, vol. 16, n.º 6, pp. 345-379, nov. 2010, doi: 10.1007/s00530-010-0182-0.

[21] C. G. M. Snoek y M. Worring, «Multimodal Video Indexing: A Review of the State-of-the-art», *Multimed. Tools Appl.*, vol. 25, n.º 1, pp. 5-35, ene. 2005, doi: 10.1023/B:MTAP.0000046380.27575.a5.

[22] A. H. Yazdavar *et al.*, «Multimodal mental health analysis in social media», *PLoS ONE*, vol. 15, n.º 4, p. e0226248, abr. 2020, doi: 10.1371/journal.pone.0226248.

[23] «Sensors | Free Full-Text | Effective Techniques for Multimodal Data Fusion: A Comparative Analysis». Accedido: 8 de diciembre de 2023. [En línea]. Disponible en: https://www.mdpi.com/1424-8220/23/5/2381

[24] «Cascade recurrent neural network for image caption generation - Wu - 2017 - Electronics Letters - Wiley Online Library». Accedido: 19 de noviembre de 2023. [En línea]. Disponible en: https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/el.2017.3159

[25] M. Chen, G. Ding, S. Zhao, H. Chen, Q. Liu, y J. Han, «Reference Based LSTM for Image Captioning», *Proc. AAAI Conf. Artif. Intell.*, vol. 31, n.º 1, Art. n.º 1, feb. 2017, doi: 10.1609/aaai.v31i1.11198.

[26] W. Jiang, L. Ma, Y.-G. Jiang, W. Liu, y T. Zhang, «Recurrent Fusion Network for Image Captioning», 30 de julio de 2018, *arXiv*: arXiv:1807.09986. doi: 10.48550/arXiv.1807.09986.

[27] J. Ji *et al.*, «Improving Image Captioning by Leveraging Intra- and Inter-layer Global Representation in Transformer Network», *Proc. AAAI Conf. Artif. Intell.*, vol. 35, n.º 2, Art. n.º 2, may 2021, doi: 10.1609/aaai.v35i2.16258.

[28] Z. Zhang, Q. Wu, Y. Wang, y F. Chen, «High-Quality Image Captioning With Fine-Grained and Semantic-Guided Visual Attention», *IEEE Trans. Multimed.*, vol. 21, n.º 7, pp. 1681-1693, jul. 2019, doi: 10.1109/TMM.2018.2888822.

[29] P. Cao, Z. Yang, L. Sun, Y. Liang, M. Q. Yang, y R. Guan, «Image Captioning with Bidirectional Semantic Attention-Based Guiding of Long Short-Term Memory», *Neural Process. Lett.*, vol. 50, n.º 1, pp. 103-119, ago. 2019, doi: 10.1007/s11063-018-09973-5.

[30] «Stack-VS: Stacked Visual-Semantic Attention for Image Caption Generation | IEEE Journals & Magazine | IEEE Xplore». Accedido: 19 de noviembre de 2023. [En línea]. Disponible en: https://ieeexplore.ieee.org/document/9174742

[31] L. Chen, Z. Jiang, J. Xiao, y W. Liu, «Human-like Controllable Image Captioning with Verb-specific Semantic Roles», 22 de marzo de 2021, *arXiv*: arXiv:2103.12204. doi: 10.48550/arXiv.2103.12204.

[32] B. Wang, L. Ma, W. Zhang, y W. Liu, «Reconstruction Network for Video Captioning», en *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, jun. 2018, pp. 7622-7631. doi: 10.1109/CVPR.2018.00795.

[33] W. Pei, J. Zhang, X. Wang, L. Ke, X. Shen, y Y.-W. Tai, «Memory-Attended Recurrent Network for Video Captioning», en *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA: IEEE, jun. 2019, pp. 8339-8348. doi: 10.1109/CVPR.2019.00854.

[34] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani, y A. Mian, «Spatio-Temporal Dynamics and Semantic Attribute Enriched Visual Encoding for Video Captioning», 29 de abril de 2019, *arXiv*: arXiv:1902.10322. Accedido: 19 de noviembre de 2023. [En línea]. Disponible en: http://arxiv.org/abs/1902.10322

[35] S. Liu, Z. Ren, y J. Yuan, «SibNet: Sibling Convolutional Encoder for Video Captioning», *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, n.º 9, pp. 3259-3272, sep. 2021, doi: 10.1109/TPAMI.2019.2940007.

[36] J. Perez-Martin, B. Bustos, y J. Perez, «Improving Video Captioning with Temporal Composition of a Visual-Syntactic Embedding», en *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA:

IEEE, ene. 2021, pp. 3038-3048. doi: 10.1109/WACV48630.2021.00308.

[37] M. M. Rahman, T. Abedin, K. S. S. Prottoy, A. Moshruba, y F. H. Siddiqui, «Semantically Sensible Video Captioning (SSVC)», *ArXiv*, sep. 2020, Accedido: 19 de noviembre de 2023. [En línea]. Disponible en: https://www.semanticscholar.org/paper/Semantically-Sensible-Video-Captioning-(SSVC)-Rahman-Abedin/cf2193f4e9e203fe05addffabed27e0c37a89efa

[38] Z. Fang, T. Gokhale, P. Banerjee, C. Baral, y Y. Yang, «Video2Commonsense: Generating Commonsense Descriptions to Enrich Video Captioning», en *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, y Y. Liu, Eds., Online: Association for Computational Linguistics, nov. 2020, pp. 840-860. doi: 10.18653/v1/2020.emnlp-main.61.

[39] Z. Zhang, D. Xu, W. Ouyang, y L. Zhou, «Dense Video Captioning Using Graph-Based Sentence Summarization», *IEEE Trans. Multimed.*, vol. 23, pp. 1799-1810, 2021, doi: 10.1109/TMM.2020.3003592.

[40] X. Wang, W. Chen, J. Wu, Y.-F. Wang, y W. Y. Wang, «Video Captioning via Hierarchical Reinforcement Learning», 29 de marzo de 2018, *arXiv*: arXiv:1711.11135. doi: 10.48550/arXiv.1711.11135.

[41] Y. Chen, S. Wang, W. Zhang, y Q. Huang, «Less Is More: Picking Informative Frames for Video Captioning», 4 de marzo de 2018, *arXiv*: arXiv:1803.01457. doi: 10.48550/arXiv.1803.01457.

[42] L. Li y B. Gong, «End-to-End Video Captioning With Multitask Reinforcement Learning», en *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, ene. 2019, pp. 339-348. doi: 10.1109/WACV.2019.00042.

[43] J. Mun, L. Yang, Z. Ren, N. Xu, y B. Han, «Streamlined Dense Video Captioning», 8 de abril de 2019, *arXiv*: arXiv:1904.03870. doi: 10.48550/arXiv.1904.03870.

[44] W. Zhang, B. Wang, L. Ma, y W. Liu, «Reconstruct and Represent Video Contents for Captioning via Reinforcement Learning», 3 de junio de 2019, *arXiv*: arXiv:1906.01452. doi: 10.48550/arXiv.1906.01452.

[45] W. Xu, J. Yu, Z. Miao, L. Wan, Y. Tian, y Q. Ji, «Deep Reinforcement Polishing Network for Video Captioning», *IEEE Trans. Multimed.*, vol. 23, pp. 1772-1784, 2021, doi: 10.1109/TMM.2020.3002669.

[46] H. Ben-younes, R. Cadene, M. Cord, y N. Thome, «MUTAN: Multimodal Tucker Fusion for Visual Question Answering», 18 de mayo de 2017, *arXiv*: arXiv:1705.06676. doi: 10.48550/arXiv.1705.06676.

[47] R. Cadene, H. Ben-younes, M. Cord, y N. Thome, «MUREL: Multimodal Relational Reasoning for Visual Question Answering», 25 de febrero de 2019, *arXiv*: arXiv:1902.09487. doi: 10.48550/arXiv.1902.09487.

[48] B. N. Patro, S. Pate, y V. P. Namboodiri, «Robust Explanations for Visual Question Answering», 23 de enero de 2020, *arXiv*: arXiv:2001.08730. Accedido: 19 de noviembre de 2023. [En línea]. Disponible en: http://arxiv.org/abs/2001.08730

[49] S. Lobry, D. Marcos, J. Murray, y D. Tuia, «RSVQA: Visual Question Answering for Remote Sensing Data», *IEEE Trans. Geosci. Remote Sens.*, vol. 58, n.º 12, pp. 8555-8566, dic. 2020, doi: 10.1109/TGRS.2020.2988782.

[50] Z. Yu, J. Yu, J. Fan, y D. Tao, «Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering», 4 de agosto de 2017, *arXiv*: arXiv:1708.01471. doi: 10.48550/arXiv.1708.01471.

[51] P. Anderson *et al.*, «Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering», 14 de marzo de 2018, *arXiv*: arXiv:1707.07998. doi: 10.48550/arXiv.1707.07998.

[52] Z. Yu, J. Yu, Y. Cui, D. Tao, y Q. Tian, «Deep Modular Co-Attention Networks for Visual Question Answering», 25 de junio de 2019, *arXiv*: arXiv:1906.10770. doi: 10.48550/arXiv.1906.10770.

[53] L. Li, Z. Gan, Y. Cheng, y J. Liu, «Relation-Aware Graph Attention Network for Visual Question Answering», 9 de octubre de 2019, *arXiv*: arXiv:1903.12314. doi: 10.48550/arXiv.1903.12314.

[54] P. Wang, Q. Wu, C. Shen, A. Dick, y A. van den Hengel, «FVQA: Fact-Based Visual Question Answering», *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, n.º 10, pp. 2413-2427, oct. 2018, doi: 10.1109/TPAMI.2017.2754246.

[55] K. Marino, M. Rastegari, A. Farhadi, y R. Mottaghi, «OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge», 4 de septiembre de 2019, *arXiv*: arXiv:1906.00067. doi: 10.48550/arXiv.1906.00067.

[56] J. Yu, Z. Zhu, Y. Wang, W. Zhang, Y. Hu, y J. Tan, «Cross-modal Knowledge Reasoning for Knowledge-based Visual Question Answering»,

Pattern Recognit., vol. 108, p. 107563, dic. 2020, doi: 10.1016/j.patcog.2020.107563.

[57] K. Basu, F. Shakerin, y G. Gupta, «AQuA: ASP-Based Visual Question Answering», en *Practical Aspects of Declarative Languages: 22nd International Symposium, PADL 2020, New Orleans, LA, USA, January 20–21, 2020, Proceedings*, Berlin, Heidelberg: Springer-Verlag, ene. 2020, pp. 57-72. doi: 10.1007/978-3-030-39197-3_4.

[58] Y. Wang *et al.*, «Tacotron: Towards End-to-End Speech Synthesis», 6 de abril de 2017, *arXiv*: arXiv:1703.10135. doi: 10.48550/arXiv.1703.10135.

[59] S. O. Arik *et al.*, «Deep Voice: Real-time Neural Text-to-Speech», 7 de marzo de 2017, *arXiv*: arXiv:1702.07825. doi: 10.48550/arXiv.1702.07825.

[60] S. Arik *et al.*, «Deep Voice 2: Multi-Speaker Neural Text-to-Speech», 20 de septiembre de 2017, *arXiv*: arXiv:1705.08947. doi: 10.48550/arXiv.1705.08947.

[61] W. Ping *et al.*, «Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning», 22 de febrero de 2018, *arXiv*: arXiv:1710.07654. doi: 10.48550/arXiv.1710.07654.

[62] A. van den Oord *et al.*, «Parallel WaveNet: Fast High-Fidelity Speech Synthesis», 28 de noviembre de 2017, *arXiv*: arXiv:1711.10433. doi: 10.48550/arXiv.1711.10433.

[63] Y. Taigman, L. Wolf, A. Polyak, y E. Nachmani, «VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop», 1 de febrero de 2018, *arXiv*: arXiv:1707.06588. doi: 10.48550/arXiv.1707.06588.

[64] J. Shen *et al.*, «Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions», 15 de febrero de 2018, *arXiv*: arXiv:1712.05884. doi: 10.48550/arXiv.1712.05884.

[65] F. Tao y C. Busso, «End-to-End Audiovisual Speech Recognition System With Multitask Learning», *IEEE Trans. Multimed.*, vol. 23, pp. 1-11, 2021, doi: 10.1109/TMM.2020.2975922.

[66] I. Elias *et al.*, «Parallel Tacotron: Non-Autoregressive and Controllable TTS», 22 de octubre de 2020, *arXiv*: arXiv:2010.11439. doi: 10.48550/arXiv.2010.11439.

[67] D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, y C. Fookes, «Deep Spatio-Temporal Features for Multimodal Emotion Recognition», en *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, mar. 2017, pp. 1215-1223. doi: 10.1109/WACV.2017.140.

[68] D. Nguyen, K. Nguyen, S. Sridharan, D. Dean, y C. Fookes, «Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition», *Comput. Vis. Image Underst.*, vol. 174, pp. 33-42, sep. 2018, doi: 10.1016/j.cviu.2018.06.005.

[69] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, y R. Zimmermann, «ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection», en *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, y J. Tsujii, Eds., Brussels, Belgium: Association for Computational Linguistics, oct. 2018, pp. 2594-2604. doi: 10.18653/v1/D18-1280.

[70] L. Chong, M. Jin, y Y. He, *EmoChat: Bringing Multimodal Emotion Detection to Mobile Conversation*. 2019, p. 221. doi: 10.1109/BIGCOM.2019.00037.

[71] «Multistep Deep System for Multimodal Emotion Detection With Invalid Data in the Internet of Things | IEEE Journals & Magazine | IEEE Xplore». Accedido: 19 de noviembre de 2023. [En línea]. Disponible en: https://ieeexplore.ieee.org/document/9216023

[72] H. Lai, H. Chen, y S. Wu, «Different Contextual Window Sizes Based RNNs for Multimodal Emotion Detection in Interactive Conversations», *IEEE Access*, vol. 8, pp. 119516-119526, 2020, doi: 10.1109/ACCESS.2020.3005664.

[73] R.-H. Huan, J. Shu, S.-L. Bao, R.-H. Liang, P. Chen, y K.-K. Chi, «Video multimodal emotion recognition based on Bi-GRU and attention fusion», *Multimed. Tools Appl.*, vol. 80, n.º 6, pp. 8213-8240, mar. 2021, doi: 10.1007/s11042-020-10030-4.

[74] Y. Gao, H. Zhang, X. Zhao, y S. Yan, «Event Classification in Microblogs via Social Tracking», *ACM Trans. Intell. Syst. Technol.*, vol. 8, n.º 3, p. 35:1-35:14, feb. 2017, doi: 10.1145/2967502.

[75] Z. Yang, Q. Li, W. Liu, y J. Lv, «Shared Multi-View Data Representation for Multi-Domain Event Detection», *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, n.º 5, pp. 1243-1256, may 2020, doi: 10.1109/TPAMI.2019.2893953.

[76] «Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset - ScienceDirect». Accedido: 24 de noviembre de 2023. [En línea]. Disponible en: https://www-sciencedirect-com.biblioteca.unimagdalena.edu.co/science/article/pii/S0957417419305834

[77] M. J. Rafiee, K. Eyre, M. Leo, M. Benovoy, M. G. Friedrich, y M. Chetrit, «Comprehensive review of artifacts in cardiac MRI and their mitigation», *Int. J. Cardiovasc. Imaging*, vol. 40, n.º 10, pp. 2021-2039, oct. 2024, doi: 10.1007/s10554-024-03234-4.

[78] H. Suresh, N. Hunt, A. Johnson, L. A. Celi, P. Szolovits, y M. Ghassemi, «Clinical Intervention Prediction and Understanding using Deep Networks», 23 de mayo de 2017, *arXiv*: arXiv:1705.08498. doi: 10.48550/arXiv.1705.08498.

[79] Y. Chang *et al.*, «Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature», *Sci. Rep.*, vol. 8, n.º 1, p. 8857, jun. 2018, doi: 10.1038/s41598-018-27214-6.

[80] C. Peng, Y. Zheng, y D.-S. Huang, «Capsule Network Based Modeling of Multi-omics Data for Discovery of Breast Cancer-Related Genes», *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 17, n.º 5, pp. 1605-1612, 2020, doi: 10.1109/TCBB.2019.2909905.

[81] Y. Fu *et al.*, «A gene prioritization method based on a swine multi-omics knowledgebase and a deep learning model», *Commun. Biol.*, vol. 3, n.º 1, Art. n.º 1, sep. 2020, doi: 10.1038/s42003-020-01233-4.

[82] I. Bichindaritz, G. Liu, y C. Bartlett, «Integrative survival analysis of breast cancer with gene expression and DNA methylation data», *Bioinforma. Oxf. Engl.*, vol. 37, n.º 17, pp. 2601-2608, sep. 2021, doi: 10.1093/bioinformatics/btab140.

[83] «Frontiers | SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer». Accedido: 24 de noviembre de 2023. [En línea]. Disponible en: https://www.frontiersin.org/articles/10.3389/fgene.2019.00166/full

[84] «Predicting Alzheimer's disease progression using multi-modal deep learning approach | Scientific Reports». Accedido: 24 de noviembre de 2023. [En línea]. Disponible en: https://www-nature-com.biblioteca.unimagdalena.edu.co/articles/s41598-018-37769-z

[85] O. B. Poirion, K. Chaudhary, y L. X. Garmire, «Deep Learning data integration for better risk stratification models of bladder cancer», *AMIA Jt. Summits Transl. Sci. Proc. AMIA Jt. Summits Transl. Sci.*, vol. 2017, pp. 197-206, 2018.

[86] S. Takahashi *et al.*, «Predicting Deep Learning Based Multi-Omics Parallel Integration Survival Subtypes in Lung Cancer Using Reverse Phase Protein Array Data», *Biomolecules*, vol. 10, n.º 10, p. 1460, oct. 2020, doi: 10.3390/biom10101460.

[87] O. B. Poirion, Z. Jing, K. Chaudhary, S. Huang, y L. X. Garmire, «DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data», *Genome Med.*, vol. 13, n.º 1, p. 112, jul. 2021, doi: 10.1186/s13073-021-00930-x.

[88] L. Tong, J. Mitchel, K. Chatlin, y M. D. Wang, «Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis», *BMC Med. Inform. Decis. Mak.*, vol. 20, n.º 1, p. 225, sep. 2020, doi: 10.1186/s12911-020-01225-8.

[89] T. Ma y A. Zhang, «Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE)», *BMC Genomics*, vol. 20, n.º 11, p. 944, dic. 2019, doi: 10.1186/s12864-019-6285-x.

[90] M. T. Hira, M. A. Razzaque, C. Angione, J. Scrivens, S. Sawan, y M. Sarker, «Integrated multi-omics analysis of ovarian cancer using variational autoencoders», *Sci. Rep.*, vol. 11, n.º 1, Art. n.º 1, mar. 2021, doi: 10.1038/s41598-021-85285-4.

[91] S. Albaradei, F. Napolitano, M. A. Thafar, T. Gojobori, M. Essack, y X. Gao, «MetaCancer: A deep learning-based pan-cancer metastasis prediction model developed using multi-omics data», *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 4404-4411, 2021, doi: 10.1016/j.csbj.2021.08.006.

[92] J. Huang, X. Zhang, Q. Xin, Y. Sun, y P. Zhang, «Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network», *ISPRS J. Photogramm. Remote Sens.*, vol. 151, pp. 91-105, may 2019, doi: 10.1016/j.isprsjprs.2019.02.019.

[93] G. Masi, D. Cozzolino, L. Verdoliva, y G. Scarpa, «Pansharpening by Convolutional Neural Networks», *Remote Sens.*, vol. 8, n.º 7, Art. n.º 7, jul. 2016, doi: 10.3390/rs8070594.

[94] Q. Liu, H. Zhou, Q. Xu, X. Liu, y Y. Wang, «PSGAN: A Generative Adversarial Network for Remote Sensing Image Pan-Sharpening», *IEEE Trans. Geosci. Remote Sens.*, vol. 59, n.º 12, pp. 10227-10242, dic. 2021, doi: 10.1109/TGRS.2020.3042974.

[95] T.-J. Zhang, L.-J. Deng, T.-Z. Huang, J. Chanussot, y G. Vivone, «A Triple-Double Convolutional Neural Network for Panchromatic Sharpening», *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, n.º 11, pp. 9088-9101, nov. 2023, doi: 10.1109/TNNLS.2022.3155655.

[96] H. Zhou, Q. Liu, y Y. Wang, «PanFormer: a Transformer Based Model for Pan-sharpening»,

22 de marzo de 2022, *arXiv*: arXiv:2203.02916. doi: 10.48550/arXiv.2203.02916.

[97] F. Palsson, J. R. Sveinsson, y M. O. Ulfarsson, «Multispectral and Hyperspectral Image Fusion Using a 3-D-Convolutional Neural Network», *IEEE Geosci. Remote Sens. Lett.*, vol. 14, n.º 5, pp. 639-643, may 2017, doi: 10.1109/LGRS.2017.2668299.

[98] «Physics-Based GAN With Iterative Refinement Unit for Hyperspectral and Multispectral Image Fusion | IEEE Journals & Magazine | IEEE Xplore». Accedido: 19 de noviembre de 2023. [En línea]. Disponible en: https://ieeexplore.ieee.org/document/9435191

[99] J.-F. Hu, T.-Z. Huang, y L.-J. Deng, «Fusformer: A Transformer-based Fusion Approach for Hyperspectral Image Super-resolution», *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1-5, 2022, doi: 10.1109/LGRS.2022.3194257.

[100] W. G. C. Bandara, J. M. J. Valanarasu, y V. M. Patel, «Hyperspectral Pansharpening Based on Improved Deep Image Prior and Residual Reconstruction», *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-16, 2022, doi: 10.1109/TGRS.2021.3139292.

[101] «HPGAN: Hyperspectral Pansharpening Using 3-D Generative Adversarial Networks | IEEE Journals & Magazine | IEEE Xplore». Accedido: 19 de noviembre de 2023. [En línea]. Disponible en: https://ieeexplore.ieee.org/document/9097446

[102] «Hyperspectral and LiDAR Data Fusion Using Extinction Profiles and Deep Convolutional Neural Network | IEEE Journals & Magazine | IEEE Xplore». Accedido: 19 de noviembre de 2023. [En línea]. Disponible en: https://ieeexplore.ieee.org/document/7786851

[103] «Multimodal Hyperspectral Unmixing: Insights From Attention Networks | IEEE Journals & Magazine | IEEE Xplore». Accedido: 19 de noviembre de 2023. [En línea]. Disponible en: https://ieeexplore.ieee.org/document/9724217

[104] S. Wang, D. Quan, X. Liang, M. Ning, Y. Guo, y L. Jiao, «A deep learning framework for remote sensing image registration», *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 148-164, nov. 2018, doi: 10.1016/j.isprsjprs.2017.12.012.

[105] «Remote Sensing | Free Full-Text | A Fusion Method of Optical Image and SAR Image Based on Dense-UGAN and Gram–Schmidt Transformation». Accedido: 19 de noviembre de 2023. [En línea]. Disponible en: https://www.mdpi.com/2072-4292/13/21/4274

[106] A. Meraner, P. Ebel, X. X. Zhu, y M. Schmitt, «Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion», *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 333-346, ago. 2020, doi: 10.1016/j.isprsjprs.2020.05.013.

[107] J. Hu, L. Mou, A. Schmitt, y X. X. Zhu, «FusioNet: A two-stream convolutional neural network for urban scene classification using PolSAR and hyperspectral data», en *2017 Joint Urban Remote Sensing Event (JURSE)*, mar. 2017, pp. 1-4. doi: 10.1109/JURSE.2017.7924565.

[108] J. Li, Z. Liu, X. Lei, y L. Wang, «Distributed Fusion of Heterogeneous Remote Sensing and Social Media Data: A Review and New Developments», *Proc. IEEE*, vol. 109, n.º 8, pp. 1350-1363, ago. 2021, doi: 10.1109/JPROC.2021.3079176.

[109] Z. Shao, L. Zhang, y L. Wang, «Stacked Sparse Autoencoder Modeling Using the Synergy of Airborne LiDAR and Satellite Optical and SAR Data to Map Forest Above-Ground Biomass», *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 10, n.º 12, pp. 5569-5582, dic. 2017, doi: 10.1109/JSTARS.2017.2748341.

[110] J. Li *et al.*, «Deep learning in multimodal remote sensing data fusion: A comprehensive review», *Int. J. Appl. Earth Obs. Geoinformation*, vol. 112, p. 102926, ago. 2022, doi: 10.1016/j.jag.2022.102926.

[111] Y. Xu *et al.*, «Transformers in computational visual media: A survey», *Comput. Vis. Media*, vol. 8, n.º 1, pp. 33-62, mar. 2022, doi: 10.1007/s41095-021-0247-3.

[112] «A Survey of Visual Transformers». Accedido: 22 de marzo de 2025. [En línea]. Disponible en: https://ieeexplore.ieee.org/document/10088164

[113] S. Lee, Y. Yu, G. Kim, T. Breuel, J. Kautz, y Y. Song, «Parameter Efficient Multimodal Transformers for Video Representation Learning», 22 de septiembre de 2021, *arXiv*: arXiv:2012.04124. doi: 10.48550/arXiv.2012.04124.

[114] Z. Pan, B. Zhuang, J. Liu, H. He, y J. Cai, «Scalable Vision Transformers with Hierarchical Pooling», 18 de agosto de 2021, *arXiv*: arXiv:2103.10619. doi: 10.48550/arXiv.2103.10619.

[115] X. Chu, Z. Tian, B. Zhang, X. Wang, y C. Shen, «Conditional Positional Encodings for Vision Transformers», 13 de febrero de 2023, *arXiv*: arXiv:2102.10882. doi: 10.48550/arXiv.2102.10882.

[116] J. Fang, L. Xie, X. Wang, X. Zhang, W. Liu, y Q. Tian, «MSG-Transformer: Exchanging Local Spatial Information by Manipulating Messenger Tokens», 25 de marzo de 2022, *arXiv*: arXiv:2105.15168. doi: 10.48550/arXiv.2105.15168.

[117] B. Wu *et al.*, «Visual Transformers: Token-based Image Representation and Processing for Computer Vision», 20 de noviembre de 2020, *arXiv*: arXiv:2006.03677. doi: 10.48550/arXiv.2006.03677.

[118] R. Mallick, J. Benois-Pineau, y A. Zemmari, «I Saw: A Self-Attention Weighted Method for Explanation of Visual Transformers», en *2022 IEEE International Conference on Image Processing (ICIP)*, oct. 2022, pp. 3271-3275. doi: 10.1109/ICIP46576.2022.9897347.

[119] Q. Zhang, Y. Xu, J. Zhang, y D. Tao, «ViTAEv2: Vision Transformer Advanced by Exploring Inductive Bias for Image Recognition and Beyond», *Int. J. Comput. Vis.*, vol. 131, n.º 5, pp. 1141-1162, may 2023, doi: 10.1007/s11263-022-01739-w.

[120] S. Robles-Serrano, G. Sanchez-Torres, y J. Branch-Bedoya, «Automatic Detection of Traffic Accidents from Video Using Deep Learning Techniques», *Computers*, vol. 10, n.º 11, Art. n.º 11, nov. 2021, doi: 10.3390/computers10110148.

[121] H. Hozhabr Pour *et al.*, «A Machine Learning Framework for Automated Accident Detection Based on Multimodal Sensors in Cars», *Sensors*, vol. 22, n.º 10, Art. n.º 10, ene. 2022, doi: 10.3390/s22103634.

[122] I. de Zarzà, J. de Curtò, G. Roig, y C. T. Calafate, «LLM Multimodal Traffic Accident Forecasting», *Sensors*, vol. 23, n.º 22, Art. n.º 22, ene. 2023, doi: 10.3390/s23229225.

[123] «liuhaotian/llava-v1.5-7b · Hugging Face». Accedido: 31 de marzo de 2025. [En línea]. Disponible en: https://huggingface.co/liuhaotian/llava-v1.5-7b