ISSN: 1692-7257 - Volume 1 – Number 45 - 2025

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Digital Object Identifier: 10.24054/rcta.v1i45.3751

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

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

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.*

Correspondence: *chenriquezm@unimagdalena.edu.co*

Received: *april 02, 2025.* ***Accepted:*** *may 05, 2025.* ***Published:*** *may 13, 2025.*

How to cite: *J. D. Ríos Pérez, G. Sánchez Torres, and C. Henríquez Miranda, "A Multi-Modal ViT-Based Deep Learning Architecture for Binary Classification of Traffic Accident", RCTA, vol. 1, no. 45, pp. 225–239, May 2025. Recovered from https://ojs.unipamplona.edu.co/index.php/rcta/article/view/3751*

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.

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.

1. INTRODUCTION

According to the World Health Organization (OMS) [1], each year, approximately 1.19 million people die worldwide due to traffic-related collisions. Additionally, 20 to 50 million people suffer non-fatal injuries, often leading to long-term disabilities. Traffic accidents are the leading cause of death among individuals aged 5 to 29 and the eighth leading cause of death across all age groups. This situation has worsened in countries like Colombia, where traffic accidents are the second leading cause of violent deaths, with the number of fatalities increasing each year [2]. Although 60% of vehicles are concentrated in middle- and low-income countries, 92% of traffic-related fatalities occur in these regions. These accidents also result in economic losses for individuals, families, and nations as a whole. Additionally, there is a significant weakness in the timely response following collisions—delays in detecting the need for assistance and in providing aid increase the severity of injuries. In emergency response to these accidents, reaction time plays a vital role: just a few minutes of delay can determine whether a person lives or dies, a 10-minute reduction in medical response time is statistically associated with a one-third decrease in the probability of deaths on the road [3].

On the other hand, data fusion refers to the integration of data from different sources or modalities to obtain multiple perspectives on a common phenomenon and address a specific problem. These modalities are complementary, as they provide information from different viewpoints of the phenomenon. The objective of these fusion strategies is to leverage the complementarity, redundancy, and cooperative characteristics among different modalities. Recently, these Multi-Modal Machine Learning (MMML) approaches have been increasingly studied and applied across various fields [4]-[9].

Deep Learning models have not only demonstrated significant technological advancements but have also expanded into applications of MMDL. Today, these methods are at the forefront of innovation, addressing complex challenges in fields such as audio-visual speech recognition and multimedia content retrieval for health analysis and social interaction studies.

An MMDL approach presents several challenges, including representation, translation, alignment, fusion, and co-learning when learning from two or more modalities [10], [11]. MMDL models combine heterogeneous data from multiple sources, enabling more accurate predictions. However, the accuracy and flexibility of these systems are not optimal due to the insufficient amount of labeled data. [12]. Moreover, this multimodal approach has been researched since the 1970s and has been categorized into four distinct eras: the Behavioral Era (1970s to 1980), the Computational Era (late 1980s to 2000), the Interactional Era (2000 to 2010), and the Deep Learning Era (from 2010 to the present). Likewise, MMDL has been applied to various fields such as understanding human multimodal behaviors during social interaction, multimodal emotion recognition [13], [14], Visual Question-Answering (VQA) [15], audio-visual speech recognition (AVSR) [16], image and video captioning [17], [18], multimedia content indexing and retrieval [19]-[21], and health analysis [22].

There are three types of fusion in MMDL: Early Fusion, where all modalities are combined in the initial stage, and the model learns from these combined modalities; Middle Fusion, where modalities are transformed into a common space instead of merely being concatenated—this type of fusion is applied in filtering projects such as recommendation systems; and Late Fusion, where the modalities are learned independently by the model and then combined before making a final decision—this approach is important when one modality is dominant [23]. See Fig 1.

The objective of this study is to design a multimodal architecture based on Deep Learning for binary classification of traffic accidents, which has traditionally been applied only to videos or input images. The project's hypothesis is that integrating multimodal resources such as tabular data and images could significantly improve the accuracy of

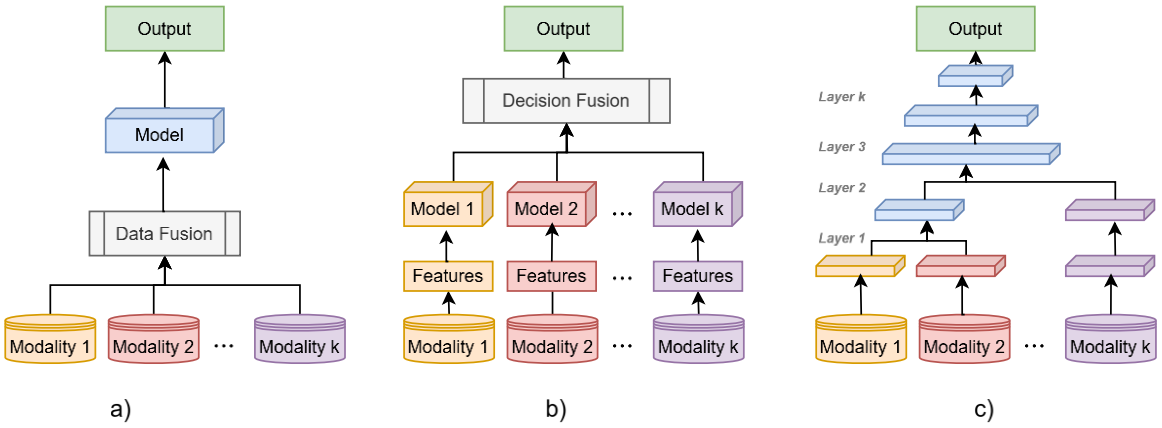

Fig. 1*. In multimodal problems, there are three types of fusion: a) Early fusion, which combines the modalities at the beginning into a single vector; b) Intermediate fusion, which learns representations of these modalities before merging them so the model can learn from them together; and c) Late fusion, which combines the predictions made by each model for their corresponding modalities.*
Source*: adapted from [7].*

traffic accident classification. Additionally, Vision Transformers (ViT) are used to leverage the potential of visual data. This project demonstrates outstanding results for binary classification of accidents and non-accidents, achieving a recall above 90%. On the other hand, the scarcity of multimodal data remains a current challenge for training these models. Techniques such as Transfer Learning and data augmentation help mitigate this issue, enabling the model to generalize more effectively.

This article is organized as follows: Section 2 reviews the related work conducted in this area. Section 3 outlines the proposed methodology used to define the architecture of the deep neural network for traffic accident detection. Section 4 presents the experimental results and the process for determining the hyperparameters of the proposed model. Section 5 provides a discussion on bias, generalizability, and key ethical considerations. Lastly, Section 6 summarizes the main conclusions and outlines directions for future research.

2. PREVIOUS WORKS

MMDL has been widely explored in diverse applications, demonstrating its adaptability and potential to enhance model performance by effectively integrating information from multiple data modalities. This versatility allows for a more comprehensive understanding of complex scenarios, as it leverages complementary features from different sources. In [12], they present a taxonomy to identify the most relevant methods and their areas of application: Multimodal Image Description [24], [25]-[31], Multimodal Video Description [32]-[45] , Multimodal Visual Question Answering [46]-[57], Multimodal Speech Synthesis [58]-[75] and other MMDL applications.

Additionally, different fields incorporate multimodal data in their problems, such as clinical applications (text, speech, images and videos) [76]-[91]; Remote Sensing Data Fusion (Panchromatic [92]-[96], Multispectral [97]-[99], Hyperspectral [100]-[101], Light Detection and Ranging [92], [102], [103], Synthetic Aperture Radar [104]-[109], infrared, night time light and satellite video data) [110]; emotion recognition (text, visual and audio) [73]. This highlights the extensive exploration and versatility of this method across different areas of knowledge, achieving outstanding results in each of these fields.

Some projects have focused on Vision Transformers (ViT) to address the complexity of image analysis and processing. Xu et al [111] drove the foundations of ViT design and its application in various high- and low-level vision tasks, such as image generation and multimodal learning. This work preserves the ability of these models to capture and represent long-term information, highlighting their performance across different types of tasks. In [112], a comprehensive review of ViT was conducted, focusing on fundamental computer vision tasks, highlighting significant improvements in benchmarks compared to traditional Convolutional Neural Networks (CNNs). ViT has also been used in a multimodal context [113], this study primarily proposes methods to reduce the computational complexity and the number of model parameters. In [114], A hierarchical variant of ViT is proposed to improve computational efficiency by clustering visual tokens, which also positively impacts the model's scalability across different dimensions, including depth, width,

and resolution. In [115], The limitations of conditional position encodings are addressed by introducing an adaptively managed scheme for variable-length input sequences, enhancing the flexibility of these models. Fang, J [116], an innovative architecture was proposed, integrating specialized tokens for local spatial information exchange between regions of an image. This helps reduce computational demands while enhancing performance in critical areas such as image classification and object detection. Based on the above, the image representation is redefined through tokens focused on ViT to delineate semantic relationships, ultimately leading to a significant improvement in the proposed classification and semantic segmentation in [117]. Likewise, the development of visualization and interpretation methods has peeled back the layers on the operational intricacies of Vision Transformers [118], providing important insights into the decision-making process [119]. Considering the above, these contributions help improve the classification of models with ViT by enhancing their performance, efficiency, versatility, and interpretability in Vision Transformers.

In [120], a Deep Learning architecture based on Conv-LSTM was used for the automatic detection of traffic accidents with a based on road videos dataset. On the other hand, a Deep Learning model using CNN, LSTM, and AE based on multimodal sensors for accident detection was proposed in [121]. Additionally, a model based on pretrained LLM and VLM was found for the automatic detection of traffic accidents [122].

Multimodal approaches have been identified in various fields, increasing model accuracy. Additionally, ViT has been used to solve different vision problems, enhancing performance, versatility, and interpretability. Specifically, in vehicular accident detection, several Deep Learning and multimodal network-based approaches have been implemented.

3. METHODOLOGY

The methodology for this binary classification project using Multimodal Deep Learning ViT is divided into five stages: Modalities (selection of input variables), Architecture Design, Preprocessing, Implementation, and Evaluation (see Fig 2).

***Fig. 2**. Methodology for binary classification project using Multimodal Deep Learning ViT.*
***Source**: own elaboration.*

3.1. Modalities

In this section, the dataset is presented along with the input modalities or variables used in the binary classification system for vehicle accident detection, which will be useful for the proposed architecture based on *section 2* results *(*Previous Works).

Roadway image provides critical real-time visual information for accident detection on the road, including traffic conditions, signage, driver behavior, potential accidents, and more. On the other hand, Tabular Traffic Data supplies information related to a traffic accident regarding road conditions, such as average speed, vehicle density, and other metrics. Additionally, Weather Data provides real-time information about weather conditions that could impact the roadway, such as rain, fog, snow, and other factors. The importance and improvement in architecture of each modality is presented in Table 1. To collect images of accidents/non-accidents, the Web Scraping technique was used to extract images from Google using Equation (1). Then, this dataset was curated by selecting images of acceptable quality.

Search Equation = "((accident OR incident OR collision OR crash) AND (traffic OR vehicle OR automobile)) OR ((traffic) AND (vehicle or automobile))" (1)

3.2. Architecture Design

The architecture design was based on an intermediate fusion architecture (see Fig 1), incorporating the modalities mentioned in Section 3.1. Each modality was processed individually before being fused. ViT was used for image processing to enhance the network's decision-making, while tabular data was preprocessed using a (Multi-Layer Perceptron) MLP.

3.3. Preprocessing

In the preprocessing stage, min-max normalization was applied to each input using Equation (2), where X represents the input variable, and *min(I)* and *max(I)* are its minimum and maximum values, respectively:

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

Due to the scarcity of data, data augmentation techniques were used to improve the network's performance and generalize from the training dataset to unseen data. Another advantage is that data augmentation helps reduce overfitting and enhances the model's robustness.

Table 1. *Resume of the features and its modalities with its importance and impact in the architecture.*

Data Modality	Importance	Type of modality	Improvement in Architecture
Roadway Image	Provides critical visual information on traffic conditions, driver behaviors, and potential incidents. Essential for identifying movement patterns, obstructions, and accident dynamics.	Visual	Facilitates real-time analysis of the road environment, allowing early detection of anomalous or hazardous conditions through image processing and computer vision.
Traffic Tabular Data	Offers a quantitative view of traffic state, including average speeds, vehicle densities, and other relevant metrics. Essential for understanding overall traffic conditions and identifying deviations from normal patterns.	Tabular	Enables the architecture to correlate traffic conditions with the likelihood of incidents, enhancing prediction accuracy.
Weather Data	Significantly affects road safety. Integration of this data helps contextualize visual and tabular observations.	Tabular	Provides additional context for interpreting visual and tabular data, al-lowing for adjustments in accident detection based on weather conditions.

Source: *own elaboration.*

In this project, geometric transformations were applied to images, and the Synthetic Minority Over-Sampling Technique (SMOTE) was used for tabular data.

3.4. Implementation

The model was improved through experimentation by adjusting the following hyperparameters: Batch size, Learning rate, Optimizer, Weight decay, Number of epochs, Patch size (ViT), Number of heads (ViT), Number of transformer layers (ViT), Dropout rate, and Activation function. Note that the hyperparameters labeled with (ViT) apply only to the image processing layers using Transformers. Additionally, the project was primarily developed using the libraries NumPy, PyTorch and Torchvision for data processing and model training.

3.5. Evaluation

The model evaluation was based on three metrics: Accuracy, which measures the proportion of correct predictions over the total number of predictions (see Equation 3); Recall, which measures the proportion of correctly identified positive cases by the model (see Equation 4); and F1-Score, which calculates the harmonic mean of precision and recall, providing a balance between the two metrics. This is especially useful when a trade-off between precision and recall is needed (see Equation 5). Additionally, the Cross-Validation technique was used to address data scarcity, overfitting, and poor generalization of the model.

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

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

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

4. RESULTS

4.1. Modalities

In this section, the results of each stage of the methodology are presented: Modalities, Architecture Design, Preprocessing, Implementation, and Evaluation (see Section 3).

Based on Table 1, Web Scraping was used to extract images from the Internet using search Equation (1). After cleaning the dataset, 635 accident images and 663 non-accident images were collected, resulting in a total of 1,298 images.

For the synthetic generation of tabular data, we used the LLaVA-v1.5-7b model [123] based on the tabular variables from Table 2. The LLaVA model was evaluated using images and the prompt from Equation (6). These descriptive variables were selected because they are directly related to the probability of an accident occurring: ‘daytime’, ‘weather_conditions’, ‘road_condition’, ‘traffic_volume’, ‘traffic_signs’, ‘road_obstacles’,

and 'lighting_road_condition'. The generated data was manually verified by a human.

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

Variable	Categories	Type
'daytime'	Day = 1, Night = 0	Binary
'weather_conditions'	Clear = 0, Cloudy = 1, Rain = 2, Snow = 3, Fog = 4, Other = 5	Nominal, One-hot
'road_conditions'	Dry = 0, Wet = 1, Frozen = 2, Water accumulation = 3, Snow accumulation = 4	Nominal, One-hot
'traffic_volume'	Low = 0, Moderate = 1, High = 2	Ordinal
'traffic_signs'	Visible = 1, Not visible = 0	Binary
'road_obstacles'	None = 0, Debris = 1, Animals = 2	Nominal, One-hot
'lighting_road_condition'	Adequate = 1, Unsuitable = 0, Visible defects on the road = 2	Ordinal

Source: *own elaboration.*

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

4.2. Architecture Design

The architecture was designed with the goal of creating a model that classifies whether an accident has occurred based on an image and its metadata (tabular data: daytime, weather conditions, road condition, traffic volume, traffic signs, road obstacles, and road lighting condition). These modalities are explained in Section 3.1.

The architecture is based on an intermediate fusion approach, where each modality (road images and tabular data) is processed and learned by the model individually. Then, they are fused through an embedding, and the network learns from this fused data using an MLP-based layer to make a decision (accident or no accident). See Fig 3.

The road images are processed through a Patch Extraction layer and transformed into a semantic space using Patch Embedding. They are then processed by Transformer layers, where each layer consists of two normalization layers, a Multihead Attention mechanism, and an MLP. These features are then fused with the tabular data to fully exploit the characteristics of the image, as it is the most semantically important modality for accident detection. On the other hand, the tabular data is processed by a two-dimensional MLP before being fused with the road images. These data are concatenated and passed through a fused embedding, allowing the network to learn from the combined data using an MLP.

4.3. Preprocessing

At this stage, normalization was first applied to all input modalities, as defined in Section 3.3. Additionally, all images were resized to a width of 224 and a height of 224. Due to data scarcity, data augmentation was performed to increase dataset variability and improve the model's generalization.

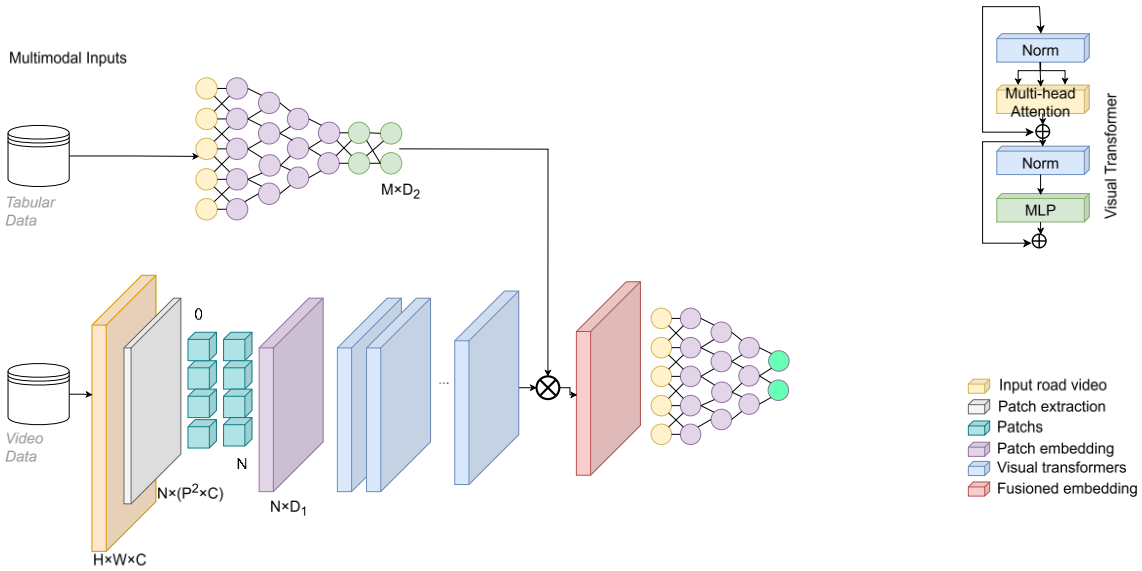

Fig. 3*. Architecture Design for a binary classification model based on Multimodal Deep Learning ViT.*
Source*: own elaboration.*

Data augmentation was applied separately to each modality (images and tabular data). Three types of geometric transformations were applied to the images: Random Horizontal Flip, Random Rotation (with a maximum rotation angle of 15 degrees), and color filters (brightness = 0.2, contrast = 0.2, saturation = 0.2), each with an independent probability of 50%. See Table 3 for a summary of the geometric transformations. For tabular data, the SMOTE technique was used to generate the same amount of augmented data as in the images.

Table 3. *Summary of Geometric Transformations for Data Augmentation in Road Images.*

Transformation	Value	Probability
Random Horizontal Flip		50%
Random Rotation	15°	50%
Brightness	0.2	50%
Contrast	0.2	50%
Saturation	0.2	50%

Source: *own elaboration.*

4.4. Implementation

After data collection, architecture design, and data preprocessing (see Sections 4.1, 4.2, 4.3), the multi-modal architecture based on ViT is implemented. The hyperparameters used in the model were determined after several iterations. Table 4 presents the final hyperparameters; although there is a wide variety of possible parameters, only those considered most relevant to this problem are listed here: Batch size, Learning rate, Optimizer, Weight Decay, Patch Size, Dimensionality of the Transformer Encoder, Number of heads, Number of transformer layers, Dropout rate, and Activation Function. It is important to highlight that these hyperparameters depend on the amount of data, computational power, and modalities used to train the model.

4.5. Evaluation

The model was trained and evaluated using an 80% train and 20% validation data, using 20 epochs in the training. Additionally, the K-Folds technique was applied, with K set to 5. The model was assessed using the three metrics presented in Section 3.5 (Accuracy, F1-Score, and Recall). In the Table 5, it presents the average results of these metrics.

Table 4. *Hyperparameters configurations for the Multi-Modal Deep Learning ViT Accident Binary Classification Model.*

Hyperparameter	Value	Note
Batch size	32	Depending on the computational resources and GPU memory.
Learning rate	1e-4	It can be improved using a learning rate scheduler de-crease the rate over time.
Optimizer	AdamW	AdamW is commonly used with transformers due to its handling of weight decay.
Weight decay	0.01	Regularization to prevent overfitting.
Patch size (Video)	32x32	Depending on the computational resources and GPU memory.
Dimensionality of the Transformer Encoder (*d_model*)	768	Depending of the data amout.
Number of heads	8	Number of attention heads in the transformer (it should be divisible by *d_model*).
Number of transformer layers	8	It can be adjusted; more layers capture more complex patterns but increase computational demand.
Dropout rate	0.1	It helps with regularization (adjust based on the level of overfitting observed).
Activation function	RELU or GELU	GELU is often used in transformer models, but RELU is a good alternative.

Source: *own elaboration.*

The loss values start at 0.3868 for training and 0.0766 for validation, ending at 0.054077 ± 0.001109 and 0.052353 ± 0.002453, respectively. On the other hand, the average Accuracy metric, which measures the number of correctly classified cases, shows very similar values for each data split, approximately 75% for training and 78% for validation. Likewise, the average F1-Score ranges between 78% for training and 80% for validation. Finally, it is important to highlight that the average Recall metric reaches 93% for training and 91% for validation. This last metric is crucial, as it emphasizes the identification of positive cases, considering that early detection could save a life or prevent an injury from worsening.

The small gap between the metric values indicates the absence of overfitting to the training data compared to the validation data. The use of the K-Folds technique helped mitigate issues related to data splitting and overfitting to the training set.

Table 5. *Average metrics values of the last epoch during K Folds divided by train and validation data.*

Metric	Train	Val
Avg. Loss	0.054077 ± 0.001109	0.052353 ± 0.002453
Avg Accuracy	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

Source: *own elaboration.*

5. DISCUSSION

Multimodal approaches enhance model performance by leveraging different scene features. Additionally, incorporating ViT layers improves accuracy. However, the lack of multimodal data in accident classification and other scenarios remains a challenge. While data synthesis helps mitigate this issue, it relies on the primary modality, which could introduce bias into the model.

On the other hand, processing more data for each involved modality requires greater computational power, making processing capacity one of the main limitations. However, using pre-trained networks could help mitigate issues related to the lack of a substantial data volume and reduce training time.

The use of multimodal projects is increasingly growing. However, a significant challenge remains due to the lack of available multimodal data. It is expected that this issue will be mitigated in the coming years as more data becomes available.

6. CONCLUSIONS

The exploration of multimodal approaches in Deep Learning highlights their re-markable evolution, versatility, and potential impact across a wide range of domains from healthcare to social monitoring environments. This progression has spanned from early conceptual frameworks to the current use of Deep Learning techniques. Notably, Vision Transformers (ViT) have demonstrated outstanding performance in various types of implementations.

Our architectural design for binary classification of traffic accidents employs intermediate fusion, which allows the features from each modality to be individually processed before being combined for decision-making. Data fusion plays a crucial role in the network's performance, making its selection essential and highly dependent on the specific problem. Previous studies highlight the wide range of architectures and methods used in multimodal designs, suggesting that there is no one-size-fits-all framework suitable for every application.

The model demonstrated solid performance despite the limited amount of data and the absence of additional modalities. Increasing the volume of data during acquisition and incorporating more modalities could enhance the network's performance in future work. Additionally, strategies such as Transfer Learning could contribute to making the model more robust.

ACKNOWLEDGMENTS

This is optional and where credit can be given to institutions and individuals for their contributions.

REFERENCES

[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