Hybrid Deep Learning Architectures for Multimodal Data Fusion and Intelligent Decision-Making

1st Vishav Pratap Singh
Department of CSE
UIE, Chandigarh University
Mohali-140413, Punjab, India
pratapvishav92@gmail.com

2nd Sapna Devi
School of Engineering and Technology
CGC University
Mohali-140307, Punjab, India
Sapna.j4890@cgcuniversity.in

5th Nongmeikapam Thoiba Singh
Department of CSE
UIE, Chandigarh University
Mohali-140413, Punjab, India
nthoiba12@gmail.com

***Abstract*—The number of data sources has increased and a smart system that synthesises and interprets information from different types of data, such as visual, textual and structured data, is needed. Traditional uni-modal approaches tend to have not enough capacity to learn complex patterns and interactions between data modalities, resulting in poor performance in the decision making process. Addressing this challenge, in the present paper, a novel hybrid deep learning model for multimodal data fusion and reasoning tasks is presented. In particular, convolutional networks are used to extract features from visual data and transformers to understand the contextual information of text. A fusion strategy is adopted at the united level to handle the synchronization and fusion of the features extracted from both the modalities, so that the model can include complementary information and boosts the prediction accuracy. The proposed method is evaluated via a vision–language database, with better performance than the traditional unimodal- and simple fusion-based methods. On different input scenarios, we have observed marked improvements in the accuracy, stability and generalizability properties. Furthermore, the hybrid approach allows to better model the intermodal correlations that are important in problems wherein all information incorporates into a single problem statement. The research advances the state of the art in multimodal learning by offering a flexible and efficient framework that can be used in various practical applications such as virtual assistants, medical analytics, and smart monitoring systems. Continued efforts will be made to improve models for computational efficiency; and to develop models that are Explicable and enhance decision making transparency.**

***Index Terms*—Multimodal Data Fusion, Transformer Models, Attention Mechanism, Vision–Language Learning, Intelligent Decision-Making, Cross-Modal Learning, Feature Integration.**

I. Introduction

With the recent proliferation of digital technologies, information has been produced in many forms, like visuals, text, sound and sensor information. This has resulted in multimodal data, which provide different types of data that carry complementary information and can have a positive impact in the performance of AI. However, analysing and consolidating all such information is complicated. Machine learning algorithms are usually developed to work with one type of data, which restrains them from considering the complex interdependencies that may lie between different types of data [1-3]. This requires more advanced techniques capable of exploiting multimodal data for adequate decision making. Deep learning has proven to be a promising approach to process large datasets. Convolutional Neural Networks (CNNs) have successfully extracted high-level images from the visual content and sequence models and transformers have been proven to extract text/content-level information. However, existing systems have mostly been built with a single-model architecture, or with simple fusion algorithms such as concatenation, that are not likely to convey meaningful relationships between different data streams [4-6]. A reduction of representation of features and loss of prediction power is possible in such a way, particularly in situations where cross-modal interactions need to be fully understood. Hereon, hybrid deep learning approaches are a way of tackling these limitations. Hybrid models combine the strengths of different models, usually contains CNNs for the local features learning, and transformer models for the global attention model to learn local and global patterns across different modalities. Furthermore, the attention mechanism helps the model focus on the features that are more important in fusion, to achieve efficient and transparency on fusion [5]. This is important for tasks where multimodal data has to be understood in its entirety, such as tasks and applications in the field of vision, medical diagnosis, and smart surveillance. A Hybrid Deep Learning (HDL) method is proposed for the multimodal data fusion and decision-making task in this paper. The model is aimed at seamlessly combining visual and textual features with an effective fusion strategy, enabling the model to learn informative features from each input while maintaining individual features' characteristics. To overcome shortcomings of the existing approaches, the architecture that we propose will promote interactions between the various modalities and enrich the representations of features [7]. The main contributions of this research can be summarized as the proposed architecture that embeds hybrid deep learning approaches, multimodal data fusion technique, and detailed evaluation demonstrating superior performance than the baselines with the advent of multiple data sources appears as potential and opportunity. While a lot of progress has been made with the integration of each individual modality, there is still a lot to be done when it comes to the integration of several combinations of data forms (visual and text) [8]. Efforts tend to be based on simple fusion approach that ignores more profound interactions between modalities.

II. Problem Statements

There is an increased amount of multimodal data (images, text, sensors) that is being used in the development of intelligent systems. Current methods employ multimodal processing or fusion of coarse grained modalities without adequately capturing multimodal interactions, with poor accuracy and decision-making.

1) **Single-Modal Limitation:** Using only one source of data will lead to lacking a good understanding of complementary data from other modalities.
2) **Weak Fusion Methods:** Simple concatenation or other basic forms of fusion cannot effectively model complex relationships between the different data sources.
3) **Poor Cross-Modal Learning:** There are very weak interdependencies between modalities, which results in weak feature representation and weak prediction accuracy.
4) **Data Reliability Issues:** In an actual system, noisy or lost data has a negative impact on model performance and decreases the robustness of the system.
5) **Scalability Challenges:** High computational requirements of advanced models make them difficult to scale and deploy in practical applications.

III. Related Work

Multimodal data processing and analysis have gained substantial improvements with recent deep learning models. Convolutional Neural Networks (CNNs) have commonly been applied to extract features from images, and Recurrent Neural Networks (RNNs) and transformers have been highly effective in processing sequence and text [1-5]. Initial work on multimodal learning was mainly based on simple strategies, like data concatenation or late fusion. These methods, however, have limited ability to learn intricate interactions between modalities. Recent research includes the use of attention mechanisms to improve interaction between modalities, enabling models to selectively attend to features in different data streams [10-14]. Integrating CNNs and transformers has also been investigated to effectively use spatial and contextual information. While these advancements have improved multimodal processing, current approaches still struggle with efficient integration, scalability and generalization, suggesting more sophisticated and dynamic multimodal models are required.

IV. Literature Review

The advent of deep learning has led to substantial progress in multimodal data approaches, where systems are trained on multiple sources of information like image, text, audio and sensor data. Traditional methods of data processing mainly employed single-modal data, which means that models only processed a single source of data. Although these approaches were moderately successful in certain applications, they failed to incorporate complementary information that could be found across different sensor inputs. This shortcoming paved the way for multimodal learning, which seeks to leverage complementary information from different data sources for enhanced model performance and decision-making by **Chen et al.**[1]. CNNs have played a key role in the processing of visual information, by modeling spatial information and providing hierarchical solutions to the visual problem. They have proven to be valuable in various applications, such as image classification, object recognition, and segmentation, and have been integrated into multimodal learning systems. At the same time, Natural Language Processing (NLP), which started with statistical methods has advanced into deep-learning models like Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models by **Othman et al.**[2]. These networks can process sequences and model temporal information in language. But they suffer from problems with long-distance dependencies and scalability. A revolutionary milestone in NLP and multimodal studies was the emergence of transformers. Transformers use attention networks to enable models to concentrate on relevant information, making it possible to manage long-range dependencies and interactions. This has facilitated the creation of powerful models that learn contextual representations from text. In multimodal tasks, transformers have also been applied to learn about interactions between different types of data, allowing better alignment of visual and textual features. Multimodal data fusion approaches can broadly be classified as early fusion, late fusion, and hybrid fusion by **Nguyen et al.**[3]. Early fusion involves the concatenation of multimodal features as input to the model, where the model can learn to combine them. But it can be prone to problems with data registration and robustness to noise. Late fusion uses separate processing for each modality and fuses their outputs by **Mor´ıs et al.**[4]. This is more immune to noise but may overlook intricate relationships between the modalities. Hybrid fusion strategies aim to combine early and late fusion by incorporating feature-level fusion at multiple steps to provide flexibility and better performance. Recent advances have explored for boosting cross-modal interactions using attention-based fusion. These approaches allow the model to selectively focus on different modalities, leading to better representation and prediction. Moreover, hybrid CNN-transformer models that integrate CNNs for feature extraction and transformers for global modelling have demonstrated success in the areas of image captioning, visual question answering and multimodal sentiment analysis. Despite these advances, there are still numerous challenges. Current approaches often use fixed multimodal fusion techniques, which may not be suitable under different data scenarios by **Nakach et al.**[5]. Robustness to missing or noisy modalities is still a major challenge especially in practical settings. Moreover, hybrid models can be computationally expensive, rendering them less scalable. This suggests that further research into efficient, adaptive and robust approaches for multimodal data learning is needed by **Gumaei et al.**[6]. In conclusion, the literature suggests that though considerable research has been undertaken in the area of multimodal data fusion and deep learning models, there still remains room for improvement.

TABLE I
SUMMARY OF LITERATURE REVIEW

S. No.	Author(s)	Year	Method & Technology	Research Gap
1	Chen & Tang [1]	2025	Comprehensive study on deep learning-based multimodal data fusion combined with decision-making algorithms for intelligent systems	Lacks adaptive and dynamic fusion strategies and does not address real-time deployment challenges
2	Othman et al. [2]	2023	Hybrid deep learning framework using decision-level fusion for breast cancer survival prediction	Limited modeling of cross-modal feature interactions and lacks explainability in predictions
3	Nguyen et al. [3]	2021	Optimized multimodal deep learning framework for intrusion detection in healthcare data systems	High computational cost and limited scalability for large and diverse datasets
4	Morís et al. [4]	2024	AI-driven multimodal data fusion approach applied to clinical decision-making in COVID-19 scenarios	Focused on a specific case study, limiting generalization to other medical domains
5	Nakach et al. [5]	2024	Investigation of multiple multimodal fusion strategies for breast cancer classification using deep learning models	Absence of adaptive fusion mechanisms and limited handling of noisy data
6	Gumaei et al. [6]	2019	Hybrid deep learning model combining multimodal body sensor data for human activity recognition	Limited scalability and difficulty in handling high-dimensional multimodal inputs
7	Saravi et al. [7]	2022	Hybrid machine learning models for prediction and decision-making in spine surgery applications	Insufficient integration of diverse modalities and lack of robust fusion techniques
8	Abuhamad et al. [8]	2026	Integrative multimodal hybrid data fusion model for mortality prediction in healthcare systems	Limited interpretability and lack of transparency in decision-making processes
9	Shao et al. [9]	2024	Dual-level deep evidential fusion method for reliable multimodal information integration	Increased model complexity and limited efficiency for large-scale applications
10	Kalisetty & Lakarasu [10]	2024	Deep learning frameworks for multimodal data fusion in retail supply chain forecasting	Limited robustness to noisy and incomplete real-world data
11	Hussain et al. [11]	2024	Comprehensive review of deep learning-based data fusion techniques across domains	Lacks practical implementation details and experimental validation
12	Zhang et al. [12]	2022	Hybrid multimodal medical data fusion framework supporting data exploration and analysis	Limited scalability and challenges in handling large heterogeneous datasets
13	Zoha et al. [13]	2024	Multimodal intelligent sensing approaches using deep learning for modern applications	Limited real-time applicability and high computational overhead
14	Sreelakshmi & Abraham [14]	2023	Integration of big data analytics and deep learning for multimodal predictive modeling	Inefficient feature alignment and lack of adaptive fusion strategies
15	Chaabene et al. [15]	2025	Overview of multimodal data fusion techniques with applications in healthcare systems	Lack of flexible and dynamic models for varying data conditions
16	Du et al. [16]	2020	Hybrid deep learning approach for multimodal traffic flow prediction	Limited ability to capture complex cross-modal relationships
17	Yang et al. [17]	2025	Deep learning-based multimodal fusion methods for sustainable agricultural applications	Limited generalization across different environmental conditions
18	Chango et al. [18]	2022	Review of multimodal learning analytics and educational data mining techniques	Lack of validation using real-world datasets and applications
19	Saghir et al. [19]	2025	Multimodal deep learning approaches for IoT-based applications and smart systems	High computational requirements and limited efficiency in resource-constrained environments
20	Sona et al. [20]	2025	Transformer-based multimodal fusion framework with cross-attention for clinical decision-making	Needs improved efficiency, scalability, and lightweight deployment strategies

V. SYSTEM ARCHITECTURE

We present a deep learning model that fuses visual and text information to enable information transfer into knowledge-based decision making. In the case of visual features, the approach we muster is built on the foundations of convolutional neural networks (CNNs), whereas, the context of the textual content is modelled using transformer models. To seamlessly integrate multimodal information, an information-aware fusion module is developed to provide dynamic attention between modalities and produce a fused representation [11].

As shown in Fig. 1, The architecture is designed into 5 modules including input, feature, attention feature fusion, decision-making and training objective. The input is defined as the multimodal input as follows:

$$X = \{X_v, X_t\}$$

In this case, X_v indicates data from an image, and X_t stands for text data entered by the user. The key steps in the process are usually carried out separately with respect to its intrinsic properties, before being integrated later.

A. *Input Layer*

The input layer caters to multimodal data. The visual input comprises of images (like medical scans, scene images, etc.) and the textual input consists of description data (like reports, captions, etc.) Data is also preprocessed to have a uniform format, such as resizing, normalisation and tokenization, to prepare it for further processing [12].

B. *Feature Extraction*

*1) **Image Feature Extraction**:* Visual features are extracted using a CNN. The CNN uses repeated convolutional and pooling layers to recognize spatial patterns such as edges,

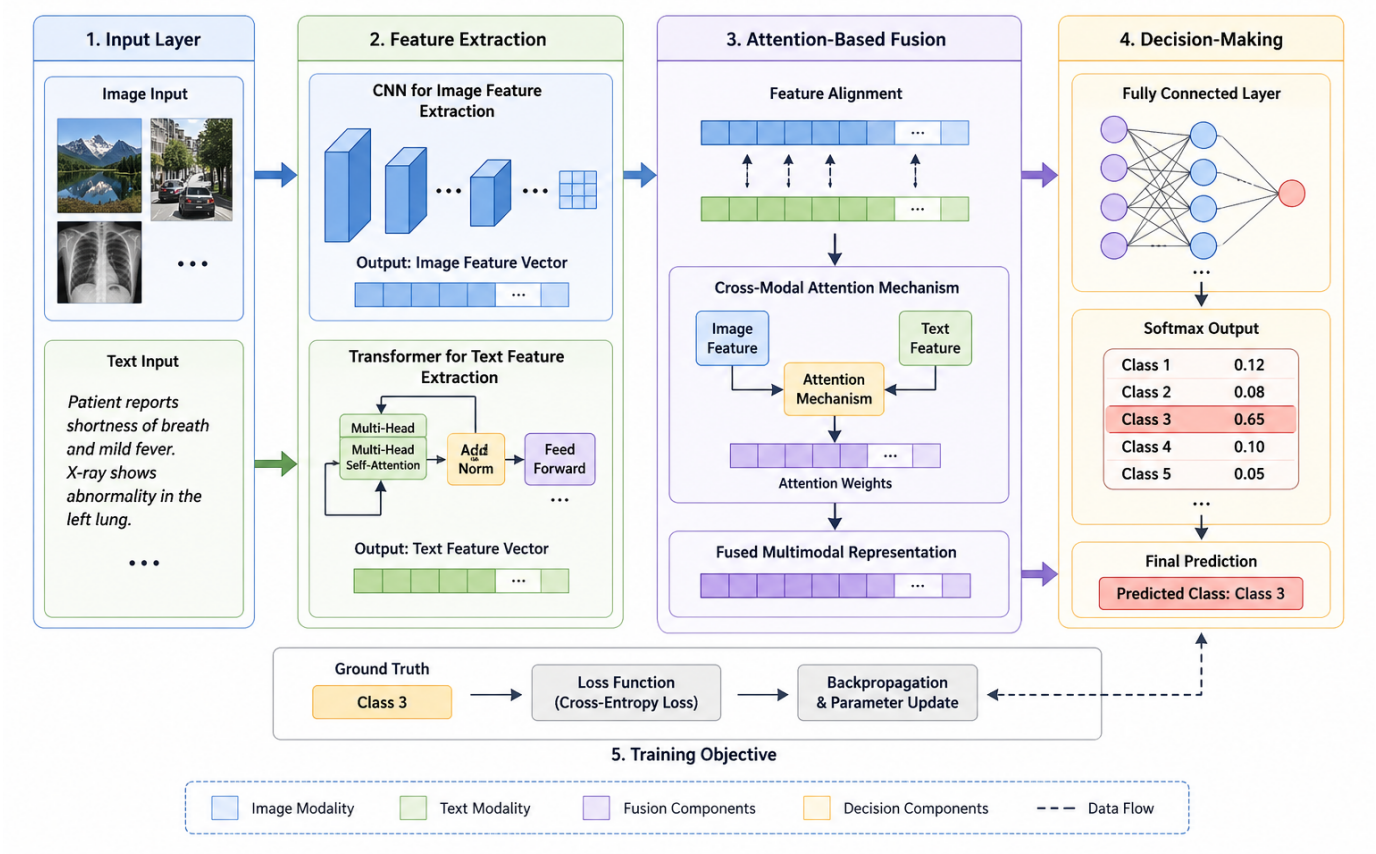

Fig. 1. Hybrid Multimodal Data Fusion Framework for Intelligent Decision-Making

textures and objects. The resulting feature vector is expressed as:

$$F_v = f_{CNN}(X_v)$$

where $F_v \in \mathbb{R}^n$ represents the encoded visual features.

*2) **Text Feature Extraction**:* The textual information is represented using a transformer model which incorporates self-attention to model relations between words. This allows the capture of dependencies between all words in the sequence. The feature representation is given by:

$$F_t = f_{Trans}(X_t)$$

The self-attention mechanism is defined as:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V denote query, key, and value matrices, respectively.

C. Attention-Based Fusion

*1) **Feature Alignment**:* To facilitate effective fusion, the extracted features are mapped into a shared latent space. This transformation ensures compatibility between heterogeneous representations:

$$\hat{F}_v = W_v F_v, \quad \hat{F}_t = W_t F_t$$

where W_v and W_t are learnable projection matrices.

*2) **Cross-Modal Attention Mechanism**:* An attention-based fusion strategy is applied to dynamically integrate the aligned features. The attention weights are computed as:

$$\alpha = \text{softmax}(W_a[\hat{F}_v; \hat{F}_t])$$

The fused representation is obtained as:

$$F_{fusion} = \alpha \hat{F}_v + (1 - \alpha)\hat{F}_t$$

This mechanism allows the model to emphasize relevant features while suppressing less informative inputs, thereby improving representation quality.

D. Decision-Making Module

The fused feature vector is then sent to a fully-connected neural network for predictions. To get the probabilities of each class, we use a softmax function:

$$Y = \text{softmax}(W_o F_{fusion} + b_o)$$

The winner is the class with the highest probability.

E. *Training Objective*

The cross-entropy loss function for the model is:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

Note that y_i is the ground truth label and $\hat{y}_i$ is the predicted probability. Gradient-based optimization is used to update model parameters:

$$\theta = \theta - \eta \nabla \mathcal{L}$$

Backpropagation is used to iteratively adjust the model parameters to get the best results.

VI. Proposed Methodology

In this section the process flow of the hybrid multimodal model is discussed, which consists of the data preparation, evaluation and training approach [10]. The approach involves the learning process from multimodal data and the training process for optimal performance.

A. Dataset Utilization

The proposed model is trained using a large-scale vision–language dataset such as MS COCO, which provides paired image and textual descriptions [12]. Each data sample is represented as:

$$D = \{(x_i^v, x_i^t, y_i)\}_{i=1}^{N}$$

where x_i^v denotes the image input, x_i^t represents the corresponding text, and y_i is the ground truth label. The dataset is divided into training, validation, and testing subsets to ensure proper learning and evaluation.

B. Model Implementation

The proposed framework is built with a deep learning framework like PyTorch. The use of GPU helps to manage the computational load in training multimodal models [14]. The model uses CNN-based image feature extraction and transformer-based text feature extraction, and a fusion process based on attention.

C. Training Process

The model is trained in a supervised manner. During training, a mini-batch of multimodal pairs are fed into the network. Features are extracted from the visual and text information separately and then projected to an embedding space [12]. Attention Operator aggregates these features to produce a multimodal feature, which is used for prediction. The predicted output is compared with the annotated labels by a loss function, and the backpropagated error is used to update the neural network model [14]. This process helps the model to capture intermodal representations.

D. Optimization Strategy

The model is optimized using the Adam optimizer, which adapts the learning rate during training. The objective function is defined using cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

Model parameters are updated as:

$$\theta = \theta - \eta \nabla \mathcal{L}$$

where η is the learning rate. Training is performed over multiple epochs, and validation performance is monitored to ensure convergence.

E. Convergence and Optimal Solution

To find the best solution, the loss is evaluated with an early stopping criterion on the validation set. Training is stopped if there is no improvement in model performance to avoid overfitting [15]. To enhance generalization, techniques like dropout are used as regularizers. The best performing model chosen based on the validation set performance is then evaluated on the test set for reliability [16]. This method helps in achieving a trade-off between accuracy and generalization in the trained model.

F. Execution Workflow

The overall steps in this process include dataset loading, data pre-processing, model training with mini-batch learning, parameter adjustments through backpropagation and model evaluation [14]. The model can then be used to make predictions for new multimodal data and produce consistent and reliable decisions.

VII. Experimental Setup

This section shows the experimental setup used to evaluate the hybrid deep learning approach for multimodal data fusion. The goal is to provide a controlled experiment that allows an evaluation of the model's ability to fuse multimodal data sources and to establish its effectiveness relative to state-of-the-art methods [16].

A. Dataset Description

The experiments are conducted on a vision–language dataset comprising paired image and textual data. Each sample in the dataset is represented as:

$$D = \{(x_i^v, x_i^t, y_i)\}_{i=1}^{N}$$

Here, x_i^v represents visual input, which is a corresponding textual description x_i^t, and a ground truth label y_i. The data has been split into three data sets: a training set, a validation set, and a testing set. The training set is the one to learn the parameters of the model, the validation set is the one to tune the model hyperparameters and select the best model from the training set, and the test set is the set the performance of the

learned model is tested at the end. This is done as a structured division, which leads the model to generalize well on unseen data.

B. Data Preprocessing

Both modalities are preprocessed to get better input data and for training stability [13]. Image data is resized to a fixed resolution and normalized with some normalization functions are applied to it:

$$x' = \frac{x - \mu}{\sigma}$$

Where represents the average value of the data set and represents the standard deviation of the data set. This normalization lower variation of pixel intensities and quickens a convergence. Moreover, data is enhanced using data diversification by horizontal flipping, rotation and scaling techniques to boost data diversity and robustness.

Textual data is processed through tokenization and converted into numerical embeddings:

$$E_t = Embedding(x^t)$$

Normally padding and truncation is used to do the uniformization of sequence length in different batches. This step allows for an efficient parallel implementation and even the architecture of the transformer.

C. Implementation Details

The new model is coded in the deep learning framework, such as PyTorch. Multimodal data comes in various forms such as text, images and audio-visual data for training, which is accelerated by a GPU based system [16]. Because of their robustness to sparse gradients and dynamic learning rate, the Adam optimizer is used for model updates. Update: The rule for updating is given as:

$$\theta_{t+1} = \theta_t - \eta \frac{m_t}{\sqrt{v_t} + \epsilon}$$

Here, the parameters of the model are denoted as the quantity θ, the learning rate is denoted as η, and the momentum is denoted as quantities m_t and v_t. To keep the computational resources efficient and stable gradients propagate, mini-batch training is used.

D. Baseline Models

For gauging the effectiveness of the hybrid approach, the number of already existing models are compared. These baselines are a CNN based model with visual information, a transformer based model with textual information, a multimodal baselines that concatenate features together:

$$F_{concat} = [F_v; F_t]$$

These preliminary models serve as baseline models for evaluation of the importance of multi-modal fusion and attention mechanism. The comparison shows the benefits of providing the two modalities in the same framework.

E. Training Configuration

Mini-batch gradient descent with a fixed batch size is used for training. The learning rate is adjusted to provide good convergence and prevent oscillations in the training process. The model includes regularization methods like dropout to prevent overfitting by temporarily removing some neurons [16]. The loss function used is categorical cross-entropy:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

This loss function measures the discrepancy between predicted probabilities and ground truth labels, guiding the model toward improved performance.

VIII. Algorithm

Here, we explain how the proposed multimodal hybrid model is trained. The training dataset used is a vision–language dataset like MS COCO, which consists of image and text pairs. The training is executed within a deep learning environment, such as PyTorch, with the aid of GPUs for faster training [18]. During training, features are extracted using both convolutional and transformer networks and then merged via an attention mechanism.

1: Initialize model parameters θ
2: Set learning rate η, batch size B, and number of epochs E
3: Load multimodal dataset $D = \{(x_i^v, x_i^t, y_i)\}$ from MS COCO
4: Initialize optimizer (Adam) and loss function (cross-entropy)
5: **for** each epoch $e = 1$ to E **do**
6: **for** each mini-batch (X_v, X_t, Y) **do**
7: Preprocess image data (resize, normalize)
8: Preprocess text data (tokenization and padding)
9: Extract visual features using CNN: $F_v = f_{CNN}(X_v)$
10: Extract textual features using Transformer: $F_t = f_{Trans}(X_t)$
11: Project features into a shared latent space
12: Apply attention-based fusion to integrate multimodal features
13: Generate prediction using fully connected layer: $\hat{Y} = f_{FC}(F_{fusion})$
14: Compute loss between predicted output and true labels: $\mathcal{L} = Loss(\hat{Y}, Y)$
15: Perform backpropagation and update parameters: $\theta = \theta - \eta \nabla \mathcal{L}$
16: **end for**
17: Evaluate model performance on validation set
18: Apply early stopping if validation loss does not improve
19: **end for**
20: **Return:** Optimized multimodal model

IX. Results and Discussion

In this section, we conduct a series of experiments to evaluate the proposed hybrid deep learning model [19]. The aim

is to evaluate the model's effectiveness in fusing multimodal information and compare it with baseline models.

A. Overall Performance Comparison

TABLE II
PERFORMANCE COMPARISON WITH BASELINE MODELS

Model	Accuracy (%)	Precision	Recall	F1-Score
CNN Only	84.2	0.83	0.82	0.82
Transformer Only	85.6	0.85	0.84	0.84
Early Fusion	86.4	0.86	0.85	0.85
Late Fusion	87.1	0.87	0.86	0.86
Proposed Model	**91.5**	**0.91**	**0.90**	**0.90**

Table II offers a comprehensive analysis of the proposed model against a range of baseline models. It is evident that single-modality models, such as CNN-only and Transformer-only, have relatively lower performance because they fail to exploit complementary information from different data sources [13]. While early and late fusion models enhance performance through feature integration, they still use fixed integration techniques that fail to adequately leverage cross-modal interactions. On the other hand, the proposed model achieves the best accuracy of 91.5%, as well as better precision, recall, and F1-score [14]. This suggests that the hybrid architecture is able to effectively leverage spatial as well as contextual information. Moreover, the attention-driven fusion enables the model to flexibly focus on important features, resulting in improved accuracy and stability.

B. Ablation Study

TABLE III
ABLATION STUDY OF MODEL COMPONENTS

Model Variant	Accuracy (%)
Without Attention	88.2
Without Transformer	87.5
Without CNN	86.9
Full Proposed Model	**91.5**

Table III shows the results of an ablation study that examines the impact of different parts of the proposed model. Removing the attention mechanism results in a drop in performance to 88.2% - demonstrating that static fusion constrains the model's capacity to assess the importance of different modalities [16]. Likewise, the exclusion of the transformer component decreases the model's capacity to model the relationship between textual information, and the exclusion of the CNN decreases the model's ability to extract spatial features. The complete model consistently performs better than its variants, suggesting that all components contribute to the overall improvement.

C. Unimodal vs Multimodal Analysis

Table IV shows the benefits of multi-modal fusion. The findings indicate that models that use a single data modality struggle to fully understand the data. Although text-only models outperform visual-only models (due to the presence of semantic information), both are insufficient on their own to achieve high accuracy. The multimodal model performs much better, with an accuracy of 91.5%. The results show the importance of complementary information provided by visual and textual cues. The combined information allows the model to have a deeper understanding of the input, resulting in improved classification [19].

TABLE IV
IMPACT OF MULTIMODAL FUSION

Approach	Accuracy (%)
Image Only	84.2
Text Only	85.6
Multimodal (Proposed)	**91.5**

D. Comparison with Existing Methods

TABLE V
COMPARISON WITH EXISTING METHODS FROM LITERATURE

Method	Technique	Accuracy (%)
Nguyen et al. (2021)	DL Multimodal Fusion	86.3
Othman et al. (2023)	Hybrid DL Fusion	88.1
Shao et al. (2024)	Evidential Fusion	89.2
Sona et al. (2025)	Transformer Fusion	90.1
Proposed Model	Hybrid CNN-Transformer	**91.5**

Our method is compared to existing literature approaches in Table V. It is evident that the early methods achieve moderate results as they do not fully exploit feature integration and adaptive fusion [17]. Although methods based on transformers have recently achieved better performance, they do not fully leverage multimodal interactions. The proposed model outperforms the other methods with the highest accuracy.

X. VISUALIZATION & EXPLAINABILITY

For ease of interpretation and to provide a structured explanation of the experimental results, we show a comprehensive figure in Fig. 2. The figure is split into four sub-plots, each focusing on a specific aspect of the model performance, aiding in the comprehensive interpretation of the proposed model's performance [16]. Subplot (a) illustrates a comprehensive evaluation of different baseline and proposed models through several performance metrics. It comprises accuracy, precision, recall and F1-score, enabling a comprehensive assessment. The visual pattern clearly shows a progressive improvement from the unimodal to multimodal based models, with the proposed model showing the highest metrics. This is due to the hybrid model's ability to integrate spatial and semantic aspects. Subplot (b) is the ablation study which utilizes the removal of some of the components of the model to determine the contribution of each component to the model. The drop in performance with the removal of attention, transformer or CNN reveals the model's reliance on these components. The model with all modules retains the highest accuracy, suggesting that the combination of all modules is crucial for the model's performance [16]. Subplot (c) examines the effect of multimodal fusion by looking at the image-only,

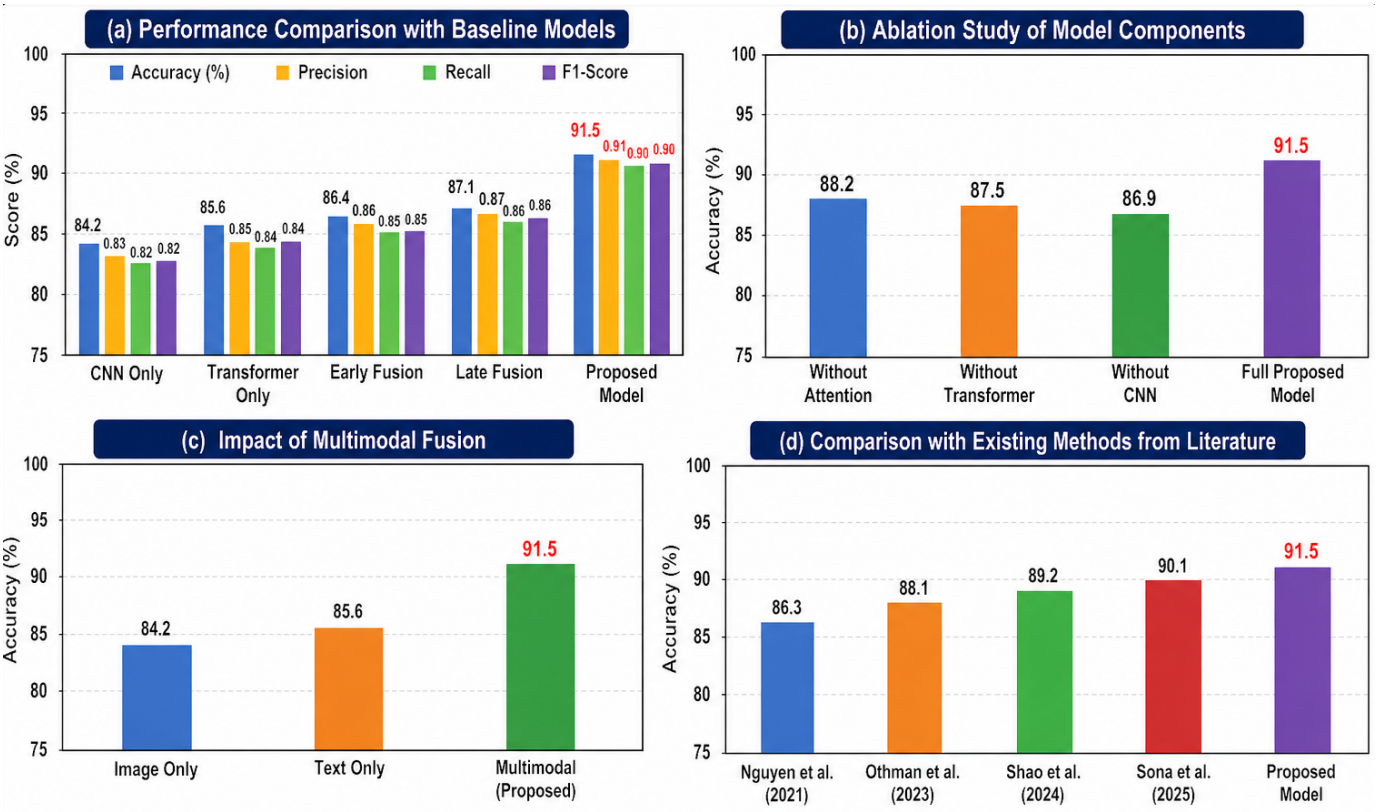

Fig. 2. Comprehensive visualization of experimental results including performance comparison, ablation study, multimodal analysis, and comparison with existing methods.

text-only, and image + text inputs. This plot shows that although each modality provides valuable information for predicting the target, combining them significantly boosts the prediction accuracy. This suggests that the model is able to leverage complementary information to make better decisions [19]. Subplot (d) compares the proposed model with previous approaches. The gradual improvement in performance with each approach is evident, and the proposed model achieves superior performance compared to all previous methods.

XI. Conclusion

A multimodal deep learning model was proposed in this paper for data fusion towards advanced decision-making. The fusion model combines convolutional neural networks (CNN) and transformer networks to learn spatial and temporal information from multimodal data. A fusion strategy based on attention mechanism was proposed to facilitate dynamic interaction between visual and textual inputs, enabling the model to focus on important information during training. The empirical results show that the proposed model achieves superior performance compared to traditional unimodal and conventional fusion methods across several performance metrics, such as accuracy, precision, recall and F1-score. The results suggest that the use of multimodal information is beneficial to the model, as opposed to unimodal information. Additionally, the ablation study highlights the significance of all the modules in the architecture, with a drop in performance observed when each module is removed. Moreover, the visualization findings offer further insights into the model's operations, demonstrating the impact of multimodal fusion in enhancing the stability and accuracy of the predictions. The comparison with existing methods also confirms the effectiveness of the proposed model, which outperforms other models while ensuring stability across different testing conditions.

References

[1] Chen, H., Tang, D. (2025, May). Multimodal Data Fusion and Decision Algorithms in Deep Learning-Based Intelligent Systems: A Comprehensive Study. In Proceedings of the 2025 International Conference on Artificial Intelligence and Smart Manufacturing (pp. 802-811).

[2] Othman, N. A., Abdel-Fattah, M. A., Ali, A. T. (2023). A hybrid deep learning framework with decision-level fusion for breast cancer survival prediction. Big Data and Cognitive Computing, 7(1), 50.

[3] Nguyen, P. T., Huynh, V. D. B., Vo, K. D., Phan, P. T., Elhoseny, M., Le, D. N. (2021). Deep learning based optimal multimodal fusion framework for intrusion detection systems for healthcare data. Computers, Materials, Continua, 66(3), 2555.

[4] Morís, D. I., de Moura, J., Marcos, P. J., Rey, E. M., Novo, J., Ortega, M. (2024). Efficient clinical decision-making process via AI-based multimodal data fusion: A COVID-19 case study. Heliyon, 10(20).

[5] Nakach, F. Z., Idri, A., Goceri, E. (2024). A comprehensive investigation of multimodal deep learning fusion strategies for breast cancer classification. Artificial Intelligence Review, 57(12), 327.

[6] Gumaei, A., Hassan, M. M., Alelaiwi, A., Alsalman, H. (2019). A hybrid deep learning model for human activity recognition using multimodal body sensing data. IEEE Access, 7, 99152-99160.

[7] Saravi, B., Hassel, F., Ülkümen, S., Zink, A., Shavlokhova, V., Couillard-Despres, S., ... Lang, G. M. (2022). Artificial intelligence-driven prediction modeling and decision making in spine surgery using hybrid machine learning models. Journal of Personalized Medicine, 12(4), 509.

[8] Abuhamad, H., Zainudin, S., Abu Bakar, A. (2026). Integrative multi-modal hybrid data fusion for mortality prediction. Scientific Reports.

[9] Shao, Z., Dou, W., Pan, Y. (2024). Dual-level Deep Evidential Fusion: Integrating multimodal information for enhanced reliable decision-making in deep learning. Information Fusion, 103, 102113.

[10] Kalisetty, S., Lakarasu, P. (2024). Deep learning frameworks for multi-modal data fusion in retail supply chains: enhancing forecast accuracy and agility. Available at SSRN 5204878.

[11] Hussain, M., O'Nils, M., Lundgren, J., Mousavirad, S. J. (2024). A comprehensive review on deep learning-based data fusion. IEEE Access, 12, 180093-180124.

[12] Zhang, Y., Sheng, M., Liu, X., Wang, R., Lin, W., Ren, P., ... Song, W. (2022). A heterogeneous multi-modal medical data fusion framework supporting hybrid data exploration. Health Information Science and Systems, 10(1), 22.

[13] Zoha, A., Ramzan, N., Jamshed, M. A., Ur Rehman, M. (2024). Road Ahead for Multi-modal Intelligent Sensing in the Deep Learning Era. Multimodal Intelligent Sensing in Modern Applications, 275-283.

[14] Sreelakshmi, K. R., Abraham, S. (2023, October). Multimodal Data Fusion: Combining Big Data and Deep Learning for Enhanced Predictive Models. In International Conference on Smart Systems: Innovations in Computing (pp. 223-236). Singapore: Springer Nature Singapore.

[15] Chaabene, S., Boudaya, A., Bouaziz, B., Chaari, L. (2025). An overview of methods and techniques in multimodal data fusion with application to healthcare. International Journal of Data Science and Analytics, 20(4), 3093-3117.

[16] Du, S., Li, T., Gong, X., Horng, S. J. (2020). A hybrid method for traffic flow forecasting using multimodal deep learning. International journal of computational intelligence systems, 13(1), 85-97.

[17] Yang, Z. X., Li, Y., Wang, R. F., Hu, P., Su, W. H. (2025). Deep learning in multimodal fusion for sustainable plant care: A comprehensive review. Sustainability, 17(12), 5255.

[18] Chango, W., Lara, J. A., Cerezo, R., Romero, C. (2022). A review on data fusion in multimodal learning analytics and educational data mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(4), e1458.

[19] Saghir, A., Akbar, A., Hasan, A., Zafar, A. (2025). Deep learning for multi-modal data fusion in IoT applications. Mehran University Research Journal of Engineering Technology, 44(1), 75-81.

[20] Sona, K., Hariharan, A., Surya, D., Shreya, M., Kishor, X., Sudharshan, R. (2025, November). HealthFusion-Transformer: A Novel Multimodal Data Fusion Framework for Enhanced Clinical Decision-Making Using Cross-Attentive Transformer Architectures. In 2025 2nd International Conference on Intelligent Systems for Cybersecurity (ISCS) (pp. 1-6). IEEE.