Journal of Theoretical and Applied Information Technology
15th May 2026. Vol.104. No.9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FACT-CHECKER: A MULTIMODAL ARCHITECTURE FOR FACT CHECKING OF SOCIAL MEDIA ARTICLES IN INDIA USING A HYBRID ATTENTION-CNN AND LSTM

C. VISHNU MOHAN[1,2], R. CHENNAPPAN[3]*

[1]Research Scholar, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu

[2]Assistant Professor, Department of Computer Science, Sacred Heart College, Thevara, Kerala

[3]Assistant Professor, Department of Artificial Intelligence & Data Science, Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu

Co-author Email: vishnumohanc01@gmail.com
*Corresponding Email: chennappan.rajendran@kahedu.edu.in

ABSTRACT

The rapid rise of misinformation within India's multilingual and multimedia digital environment has intensified the demand for advanced automated fake-news detection systems. Most existing solutions primarily rely on text-only analysis, failing to capture the critical relationship between textual statements and their accompanying images. Even current multimodal models that integrate visual information struggle with domain generalization and often lack robust attention mechanisms capable of identifying subtle yet meaningful visual inconsistencies. To overcome these limitations, this research introduces Fact-Checker, a hybrid multimodal architecture designed to enhance misinformation detection by jointly analyzing visual and textual cues. The proposed framework employs an attention-empowered Convolutional Neural Network (CNN) to identify distortions, manipulations, and salient patterns in images, while a Bidirectional Long Short-Term Memory (BiLSTM) network is used to extract contextual semantics from textual content. Both modalities are processed in parallel, transformed into discriminative feature representations, and subsequently fused at the feature level to support a unified classification process. The model is evaluated using the Indian Fake News Dataset (IFND), a culturally and linguistically relevant resource, ensuring strong applicability to Indian social-media contexts. Experimental results demonstrate that Fact-Checker achieves an overall accuracy of 81%, significantly outperforming existing unimodal and multimodal baselines. These findings underscore the effectiveness of integrating attention-based visual analysis with deep contextual text modeling for detecting misinformation. Furthermore, the proposed architecture establishes a scalable foundation for future enhancements, including multilingual adaptation, incorporation of additional modalities, and deployment across varied social-media platforms to strengthen the fight against digital misinformation.

Keywords: *Fake News Detection, Convolutional Neural Network, Attention Mechanisms, Residual Learning, Bidirectional LSTM.*

1. INTRODUCTION

The rapid expansion of digital communication across India has significantly amplified the spread of misinformation, creating a complex challenge for online safety and information integrity [1]. Social-media platforms such as X (Twitter), Facebook, and WhatsApp serve as the primary channels through which misleading and manipulated content circulates at high speed [2]. Studies indicate that nearly 77% of misinformation shared in the country originates from these platforms, where users frequently engage with visually appealing and emotionally charged content [3]. Images, memes, and modified visuals often play an influential role in directing public sentiment, especially during elections, public-health crises, and other socially sensitive periods [4].

Detecting misinformation in the Indian context is inherently difficult due to the country's vast linguistic diversity [5]. With more than 22 official languages and a widespread practice of mixing English with regional languages, automated systems face challenges in understanding varied styles of

expression [6]. Furthermore, the multimodal nature of misinformation often combining text with manipulated images or short videos adds additional layers of complexity. This multimedia-driven diffusion reduces the effectiveness of traditional text-only fake-news detection approaches [7].

Another major obstacle lies in the structural characteristics of platforms such as WhatsApp, where encryption and frictionless forwarding mechanisms enable rapid and untraceable dissemination of false information [8]. Limited availability of large, culturally appropriate, and multimodal datasets further restricts the development of robust detection models. Moreover, the boundaries between satire, opinion, propaganda, and genuine misinformation are often blurred, creating ambiguity that complicates automated classification [9]. Combined with relatively low digital literacy levels in parts of the population, misinformation can spread unchecked and influence public perception and behavior [10].

Existing fake-news detection techniques typically depend on unimodal approaches, most commonly analyzing text using machine learning or deep learning. While effective in certain contexts, these methods fail to account for visual cues, manipulated imagery, or potential inconsistencies between text and image [11]. More recent multimodal models attempt to integrate image and text features but often rely on shallow fusion methods or simplistic visual pipelines that lack attention mechanisms [12]. These limitations reduce their ability to handle India's highly diverse, mixed-language, and image-rich misinformation landscape [13].

Despite the increasing urgency to combat misinformation in India's multilingual and visually rich digital landscape, existing automated fake-news detection systems remain largely inadequate. Current unimodal approaches rely heavily on textual features and fail to capture inconsistencies between text and associated images, which are often manipulated to reinforce deceptive narratives. Even available multimodal models struggle due to shallow fusion techniques, weak visual-processing pipelines, and the absence of attention mechanisms capable of identifying subtle distortions. These limitations result in poor generalization across diverse linguistic contexts and domain variations found in Indian social-media content. Furthermore, the lack of robust fusion strategies prevents models from effectively leveraging complementary signals across modalities. Therefore, there is a critical need for a comprehensive multimodal solution that can jointly analyze text and images, focus on salient visual cues, interpret deep semantic patterns, and integrate these features in a balanced manner. Addressing this gap is essential for building an effective, scalable, and culturally relevant fake-news detection system for India.

This research offers the following key contributions:

- A novel multimodal hybrid architecture that integrates attention-guided CNNs for image analysis with BiLSTM networks for textual understanding, enabling robust detection of misinformation in a multilingual, multimedia-rich environment.
- An attention-enhanced visual processing pipeline with residual learning, designed to capture subtle manipulations in images and highlight salient spatial regions relevant to misinformation.
- A balanced feature-level fusion strategy that effectively combines visual and textual representations, enabling the model to leverage cross-modal relationships often overlooked in prior works.
- A scalable foundation for future multilingual and cross-platform fake-news detection, supporting potential extensions to transformer-based text encoders, multilingual datasets, and additional modalities.

The following paper is organized as follows: Section 2 reviews existing research on fake-news dissemination and summarizes automated detection models and datasets. Section 3 introduces the proposed multimodal hybrid framework tailored to the Indian context, outlining its objectives and system architecture. Section 4 presents the experimental results and provides an in-depth discussion of the findings.

2. RELATED WORKS AND BACKGROUND

2.1 Definition, Evolution and Dissemination of Fake News

Online misinformation, commonly referred to as fake news, has emerged as a major area of concern because of its growing influence on public opinion and individual behaviour. With digital connectivity becoming an integral part of everyday life, people encounter misleading content from a very young age [14]. Addressing this issue requires a broad response that combines media literacy, critical thinking, responsible journalism, and improved technological measures from social media platforms to limit the spread of deceptive information.

The circulation of fake news poses a serious threat to informed decision-making. Studies show that false content often follows distinctive and rapid dissemination patterns compared to legitimate news, appearing almost immediately on platforms such as Twitter and Weibo. These early bursts of activity can serve as indicators for identifying misleading posts [15][16]. Researchers have also noted that how users judge online information depends on factors such as the presentation of the content, the underlying intention, and the extent to which the information can be verified. Platform interventions, such as warning labels, may not always help; in some cases, they may even encourage high-influence users to share the content more vigorously when limitations are imposed [17][18]. These behavioural characteristics also pose challenges for automated detection systems, since misleading material frequently blends emotional language with manipulated images that exploit cognitive biases. As a result, effective detection models must incorporate both textual and visual cues to capture deceptive patterns more accurately.

During the COVID-19 pandemic, misinformation spread widely and at times matched the reach of accurate information. Social media users became increasingly confined within echo chambers, amplifying confirmation biases and intensifying the impact of unreliable content. The resulting "infodemic" contributed to confusion, anxiety, and even physical harm, demonstrating the serious consequences of unchecked health-related misinformation [19][20]. These findings highlight how quickly misinformation can adapt during crises, reinforcing the importance of detection systems that can respond to shifting themes and evolving forms of online content.

Demographic patterns have also shaped the spread of fake news, as illustrated during the 2016 U.S. Presidential election. Research reported that nearly one in ten Americans shared at least one false story. Two user groups were particularly active: politically conservative users promoting content aligned with their preferred narratives, and older adults who were more likely to believe and distribute fabricated information [21]. This example shows that social and demographic tendencies play a major role in how misinformation grows and persists, offering important behavioural indicators for modelling its spread.

Taken together, these observations provide useful guidance for the development of computational approaches to fake news detection. Insights into how misinformation circulates, who participates in its spread, and what psychological triggers it relies upon can be translated into features that help machine-learning models identify false content more effectively. This underscores the need for data-driven, multimodal detection frameworks that analyse both content-level signals such as text and images and dissemination patterns that distinguish misleading narratives from authentic information.

2.2 Unimodal Fake News Detection Approaches

Unimodal fake news detection methods rely on a single source of information, either text or image, to classify content as real or false. The most widely used category is text-based analysis, where social media posts are collected and examined using linguistic or statistical features. Studies by Reddy et al. and Ahmed et al. represent this line of work [22][23]. Reddy et al. applied writing-style cues and word-vector features, reporting improved performance through ensemble learning, particularly boosting techniques. Ahmed et al. adopted vectorization approaches such as Term Frequency (TF) and TF-IDF combined with n-gram analysis. Although these classical text-based models are straightforward and often effective on smaller datasets, they struggle to generalize well across varied writing styles and rapidly evolving online content. Taken together, these works show that while ML-based text methods provide a strong baseline, they offer limited robustness when applied to diverse or large-scale datasets.

Advances in deep learning (DL) have expanded the scope of unimodal text-based detection. Models such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks have been widely adopted for textual analysis [24][25][26]. One study combined CNN and LSTM architectures to incorporate user behaviour and the social context surrounding the posts, achieving an accuracy of 72.50%, outperforming traditional ML methods. However, the model did not incorporate metadata, limiting its applicability. Girgis et al. evaluated various RNN- and CNN-based architectures and concluded that CNN models were the most effective within the tested setting. Another study comparing CNN and RNN-LSTM achieved high accuracy of 99.90% in English and 92.48% in Turkish, but still misclassified a considerable proportion of real news as fake, indicating overfitting issues and language-specific limitations. Overall, DL approaches consistently outperform traditional ML techniques, yet they continue to face difficulties in handling

multilingual data and maintaining consistency across different textual domains.

Sentiment-based models have also been explored as unimodal solutions. Alonso et al. highlighted the limitations of sentiment analysis for fake news detection, noting the lack of consistent evaluation metrics and the reliance on qualitative assessments in several studies [27]. These shortcomings become more pronounced when dealing with multilingual content or posts combining text with multimedia elements. In contrast, a study using a sentiment-aware classifier on the PHEME dataset achieved an accuracy of 86% on textual inputs, especially for negative sentiment cases [28]. Despite this improvement, the model struggled to generalize to publicly available datasets beyond the training environment. These contrasting findings suggest that sentiment alone is insufficient for reliable fake-news detection and may capture emotional tone without fully addressing factual accuracy.

Several studies have compared ML and DL models directly. One investigation evaluated the COVID Fake News Dataset using eight ML classifiers and four deep neural network (NN) architectures [29]. While ML models performed reasonably well, neural networks, particularly BiLSTM, achieved a higher accuracy of 97%. Yet, the NN models still exhibited limited adaptability when applied to multilingual or complex datasets. Another study proposed WELFake, a model combining linguistic features with word embeddings and trained on a newly curated dataset [30]. The authors reported an accuracy of 96.73%, surpassing the performance of BERT and CNN baselines. However, the dataset lacked external verification, raising concerns about the reliability of the generated labels. These comparisons indicate that even high-performing text-only models face notable constraints, particularly in dataset credibility, cross-lingual adaptation, and real-world generalizability.

Beyond textual cues, image-based studies have also explored fake-news detection, focusing on the manipulation or fabrication of visuals shared online. Existing approaches range from traditional digital-forensic methods and basic image transformations (such as splicing and mirroring) to machine-learning and deep-learning models aimed at identifying tampered content [31]. While forensic techniques provide high accuracy in controlled settings, they often struggle when images contain multiple layers of editing. CNN-based methods, in contrast, have shown stronger capabilities in detecting visual inconsistencies. Some studies examined image metadata or simple visual features, such as RGB channel averages in social-media images. Others developed custom ML or neural-network classifiers for Twitter images, reporting high accuracy scores [32][33]. However, despite their promise, these image-only approaches commonly fail to generalize to unseen manipulations or diverse image sources, underscoring the limitations of treating visual evidence in isolation.

2.3 Multimodal Fake News Detection Approaches

Multimodal fake news detection typically combines textual and visual information by designing separate processing branches for each modality and then integrating their representations in a unified model. This integration can occur either at the feature level, where embeddings from text and images are merged before classification, or at the decision level, where predictions from independent classifiers are combined. In both cases, the effectiveness of the model depends heavily on the quality of feature extraction. Image features often involve colour distributions, texture patterns, and spatial descriptors, while text features range from traditional representations such as Bag-of-Words (BoW) and TF–IDF to distributed embeddings and transformer-based encodings [34].

Feature-level fusion approaches form the largest body of work in multimodal FND. Singh et al. conducted one such study using eight custom ML models on the Fakeddit dataset, first evaluating the modalities separately and then merging their features [35]. Although text-only and image-only models reached 86.24% and 93.84% accuracy respectively, the fused model corrected several errors missed by unimodal systems. Another feature-level approach on the Weibo dataset combined CNN-based image features with RNN-based textual sequence encodings, capturing both spatial and temporal structure in multimodal posts. This model outperformed earlier architectures such as VGG and ResNet, despite limited fine-tuning capabilities [36]. These works demonstrate that early fusion allows richer cross-modal interactions, but the quality of fusion depends strongly on balanced learning between modalities.

A second category of research relies on decision-level fusion, where each modality is processed independently and the outputs are aggregated. Several studies combine pretrained BERT representations for text with CNN variants for images, including VGG16, VGG19, and Xception. On the Fakeddit dataset, a model combining BERT with a CNN achieved an accuracy of 78%, while another pairing fine-tuned BERT with Xception reported 91.94% accuracy in tweet classification

[37]. However, these models generally did not incorporate metadata, user information, or comment structures, which limits their contextual robustness. Another study on the same dataset used BERT for text and CNN for multi-class classification, also reporting an accuracy of 78% [38]. Decision-level systems are typically easier to implement but may miss deeper correlations between textual cues and visual elements.

Beyond individual pairings of text and image models, several studies experiment with expanded transformer-based frameworks. For example, hybrid combinations of BERT with VGG16 or VGG19 have been evaluated on Twitter and Weibo datasets, including similarity-based approaches comparing title embeddings with image tag embeddings [39][40]. Although these methods showed improvements over unimodal baselines, their gains were often marginal, reflecting the difficulty of aligning semantic and visual information when the two modalities are processed largely independently.

More advanced multimodal designs include frameworks such as TTEC, which utilizes a three-stage pipeline incorporating back-translation, BERT-based multimodal feature aggregation, and contrastive learning [41]. While this approach outperformed earlier BERT-driven models, its evaluation was confined to COVID-19 datasets, limiting generalizability. Another study trained a customized BERT–CNN architecture on the Indian Fake News Dataset (IFND), achieving an accuracy of 70% and demonstrating comparatively better performance across languages and contexts [42]. These studies highlight a growing trend toward deeper multimodal interaction, yet most systems still struggle with alignment between modalities and with consistent performance across diverse datasets.

Overall, existing multimodal methods reveal clear gains over unimodal systems, but they often face challenges in effective cross-modal fusion, handling multilingual data, and capturing subtle visual–textual inconsistencies. These limitations in multimodal alignment motivate the attention-based hybrid architecture proposed in this study.

2.4 Attention Mechanisms in Neural Networks

Attention mechanisms have been increasingly incorporated into neural network architectures to enhance their ability to focus on the most informative parts of input data. In the context of image processing, attention has shown promise in tasks such as image denoising. A CNN equipped with attention layers demonstrated strong quantitative and visual performance by simultaneously reducing noise and extracting salient features more effectively than standard convolutional models [43]. This underscores the usefulness of attention in guiding the network toward meaningful visual regions.

In another area, attention has played a key role in improving Visual Question Answering (VQA). Here, attention modules were integrated into a CNN architecture through a configurable convolution operation, allowing the model to selectively prioritize regions of an image relevant to the posed question [44]. The incorporation of attention not only improved interpretability but also highlighted its capacity to enhance multimodal alignment in vision–language tasks.

Applications outside the vision domain also demonstrate the versatility of attention mechanisms. In energy load prediction, an attention-enhanced neural network was developed to capture the complementary patterns present in different energy vectors, as well as the spatial relationships influencing consumption trends [45]. Similarly, for fire detection, adding attention modules to a CNN improved both detection accuracy and localization performance, enabling the model to better isolate visually significant cues associated with fire outbreaks [46].

Together, these studies illustrate the broad value of attention mechanisms in strengthening feature extraction, improving localization, and enabling more effective focus on relevant information across diverse tasks. This evidence supports the incorporation of attention within fake news detection systems, where identifying critical textual expressions and subtle visual distortions is essential for robust multimodal analysis.

2.5 Datasets in Fake News Detection

The performance and reliability of fake news detection models depend greatly on the quality, diversity, and modality of the datasets used for training and evaluation. Several publicly available datasets were examined to assess their suitability for this study.

The LIAR dataset [47] consists of more than 12,800 text-based statements sourced from PolitiFact, annotated across six truthfulness levels ranging from “pants-on-fire” to “fully true.” Although the dataset is widely used, its short, politically focused statements and uneven class distribution limit its usefulness for models requiring richer contextual information. Another commonly referenced dataset is Fakeddit, a large multimodal collection of over one million Reddit posts offering six-category classification [48]. Its scale and inclusion of both text and images make it attractive

for multimodal research, yet its automated labelling process and reliance on unverified user-generated content raise concerns about annotation reliability.

Jindal et al. introduced two additional multimodal datasets - NewsBag and the expanded NewsBag++ containing approximately 589,000 samples aggregated from outlets such as The Wall Street Journal and The Onion [49]. Despite their size, these datasets suffer from issues such as limited label granularity and inconsistencies between textual content and associated images. Another dataset, "Bend the Truth," contains 900 Urdu-language articles [50]. While valuable for regional studies, its small size and single-language focus restrict its applicability for broader multilingual or multimodal tasks.

Given the limitations of existing datasets whether related to label reliability, cultural bias, insufficient multimodality, or constrained linguistic diversity, this research employs the Indian Fake News Dataset (IFND) to ensure greater relevance to the linguistic, social, and cultural characteristics of misinformation circulated in the Indian context.

Although extensive research has explored fake-news detection using unimodal text or image-based approaches, these methods fail to capture the cross-modal inconsistencies that frequently characterize misinformation, especially within India's multimedia-driven ecosystem. Existing multimodal models offer improvements but still face major limitations, including shallow or unbalanced fusion techniques, weak alignment between textual and visual features, and inadequate attention mechanisms for identifying subtle manipulations in images. Many studies rely on datasets that are either small, linguistically narrow, or globally oriented, making them unsuitable for India's highly diverse, multilingual, and culturally nuanced misinformation landscape. Furthermore, transformer-based and CNN-based multimodal systems often overlook Indian-specific content styles, code-mixed language patterns, and region-specific meme formats, resulting in limited generalizability. In addition, available models seldom incorporate mechanisms capable of focusing on salient visual cues or deep contextual semantics in text simultaneously. Taken together, these gaps reveal a critical need for a robust, attention-driven multimodal framework grounded in an Indian context and supported by a culturally representative dataset such as IFND. This motivates the development of a hybrid architecture that more effectively integrates attention-based visual processing with deep textual understanding to overcome the limitations of existing fake-news detection systems.

3. PROPOSED METHODOLOGY

The proposed methodology introduces Fact-Checker, a multimodal hybrid framework developed to address the limitations of existing fake-news detection systems, which often rely solely on textual analysis or employ shallow multimodal fusion techniques that fail to capture cross-modal inconsistencies. Many current approaches lack robust visual modeling, do not incorporate attention mechanisms, and struggle with generalization across India's diverse linguistic and multimedia content. To overcome these challenges, Fact-Checker begins with a comprehensive preprocessing phase, where textual data is cleaned, tokenized, and padded, while all associated images are downloaded, validated, and resized to maintain uniform input quality. The system then employs two parallel deep-learning pipelines: an attention-guided CNN with residual connections that enhances the extraction of salient visual features and highlights manipulated regions, and a Bidirectional LSTM network that captures long-range semantic dependencies in the text. These refined features are transformed into equal-sized vectors and integrated through an effective feature-level fusion strategy, enabling the model to learn complementary relationships between image and text that unimodal systems typically miss. The fused representation is finally processed by a dense classification layer to determine the authenticity of the news. Through deeper visual attention, richer contextual text modeling, and balanced multimodal fusion, Fact-Checker successfully overcomes the shortcomings of prior methods and provides a more reliable framework for identifying misinformation in India's complex digital ecosystem. Figure 1 presents a schematic view of this workflow.

The image branch extracts visual cues using an attention-enhanced convolutional architecture, while the text branch captures contextual patterns and word-level dependencies using a BiLSTM network. Once the feature representations have been generated separately, they are converted into vectors of equal dimension and combined to form a unified multimodal feature space. The fused representation is then passed to a fully connected layer that produces a binary output. The figure therefore highlights four major stages: data preparation, feature learning in the image and text branches, feature fusion, and binary output generation. This modular design ensures that both modalities contribute to the final decision and enables the model

Journal of Theoretical and Applied Information Technology
15th May 2026. Vol.104. No.9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

to learn from complementary characteristics present in the dataset.

The subsections that follow describe each component of this workflow and its implementation in detail.

Prior research has investigated various deep learning architectures for misinformation detection, including CNN- and RNN-based models for textual analysis, hybrid approaches combining linguistic and latent representations, and convolutional frameworks for detecting visual tampering in images [23–25] [27–29]. A limited number of studies have also explored multimodal solutions that integrate text–image features through joint or hierarchical fusion mechanisms [34–37][42]. Building upon these findings and addressing the shortcomings identified in earlier sections, the proposed Fact-Checker framework adopts a structured multimodal design.

The methodology is organized into two parallel processing pipelines—the image branch and the text branch each responsible for extracting discriminative features from its respective modality. These independent feature vectors are subsequently fused to perform final fake-news classification, as illustrated in Figure 1. To streamline the integration of preprocessing, feature extraction, and batching, a custom Python class was developed to process each image–text–label triplet throughout the pipeline. Each dataset record is encoded as a tuple consisting of:

(1) an image tensor in the form (Channels × Height × Width),

(2) a text tensor representing a fixed-length sequence of token indices, and

(3) a label encoded as 0 (Fake) or 1 (True).

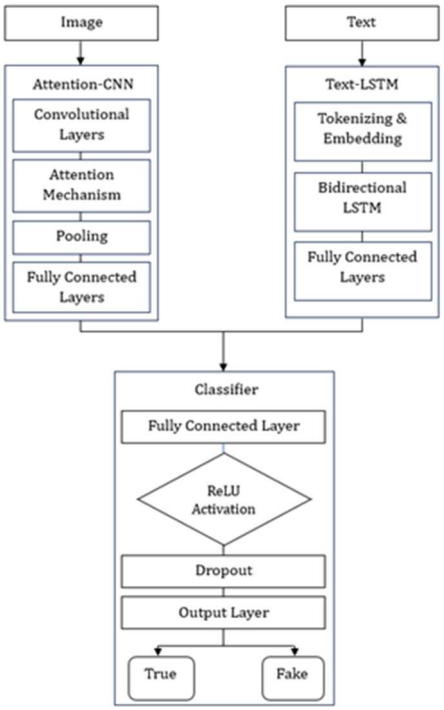

Fig. 1. Overview of the Proposed Fact-Checker Architecture

The complete preprocessing and data-flow procedure is summarized in Figure 2. This structured representation ensures seamless multimodal learning and efficient computation across the entire training pipeline.

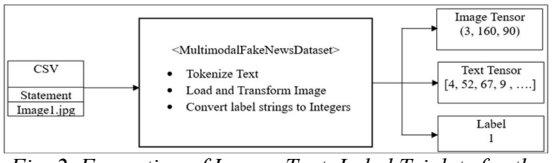

Fig. 2. Formation of Image–Text–Label Triplets for the Multimodal Pipeline

3.1 Dataset Description

In this study, we use the Indian Fake News Dataset (IFND) as the primary data resource. The IFND comprises news articles and corresponding images collected between 2013 and 2021, covering real and fake content in the Indian context. It is publicly available and described in the journal article [51]. The dataset was scraped from multiple Indian fact-checking sites and news portals, and the authors indicate it is accessible to the research community. Where available, we adhered to the dataset's licensing terms (the original publication indicates a CC BY 4.0 license) for usage in our experiments. Using IFND ensures cultural and linguistic relevance to our Indian-social-media context, particularly in enabling multimodal analysis of both text and image content.

IFND contains both textual and image information related to news consisting of 44,843 records from various domains. Basically, IFND dataset is a Comma Separated Values file (.csv) that contains textual information and hyperlinks that contain the URL of the images. The attributes of the original IFND CSV file are Id, Statement, Image, Web, Category, Date, and Label. Among the 44,843 records, 22,431 are labeled as real and 22,412 as fake. Figure. 3 shows the sample from the original dataset.

	A	B	C	D	E	F	G
1	id	Statement	Image	Web	Category	Date	Label
2	1	WHO praises India's	https://cdn.dnaindia.com/sites/default/files/styles,	DNAINDIA	COVID-19	Oct-20	TRUE
3	2	In Delhi, Deputy US	https://cdn.dnaindia.com/sites/default/files/styles,	DNAINDIA	VIOLENCE	Oct-20	TRUE
4	3	LAC tensions: China	https://cdn.dnaindia.com/sites/default/files/styles,	DNAINDIA	TERROR	Oct-20	TRUE
5	4	India has signed 25(	https://cdn.dnaindia.com/sites/default/files/styles,	DNAINDIA	COVID-19	Oct-20	TRUE
6	5	Tamil Nadu chief m	https://cdn.dnaindia.com/sites/default/files/styles,	DNAINDIA	ELECTION	Oct-20	TRUE
7	6	Bihar Assembly Elec	https://cdn.dnaindia.com/sites/default/files/styles,	DNAINDIA	ELECTION	Oct-20	TRUE
8	7	Hathras case: CBI r	https://cdn.dnaindia.com/sites/default/files/styles,	DNAINDIA	VIOLENCE	Oct-20	TRUE
9	8	Rajasthan Crime Ne	https://cdn.dnaindia.com/sites/default/files/styles,	DNAINDIA	VIOLENCE	Oct-20	TRUE

Fig. 3. Sample records from the Indian Fake News Dataset (IFND)

3.2 Data Preparation and Preprocessing

The preprocessing phase ensured that both textual and visual inputs were transformed into clean, consistent, and model-ready formats. The first task to make the data ready for processing was to download the image corresponding to each record. The images were successfully downloaded to the drive for processing. In case the images were missing or failed to download, their corresponding records in the CSV file were excluded from further processing. After this filtering, 44,622 records remained. The next step was to identify and retain only those attributes that contributed meaningfully to the classification process. We finalized the features Id, Statement, Image (ImgName), and Label. A sample of the resulting dataset after removing irrelevant attributes is shown in Figure 4.

For text preprocessing, all statements were converted to lowercase and stripped of punctuation and special characters. The cleaned text was then tokenized, and a vocabulary was constructed containing only those tokens occurring at least twice, along with the special tokens <PAD> and <UNK>. To ensure consistent sequence length across samples, each statement was transformed into a fixed-length tensor using padding for shorter sequences and truncation for longer ones, based on a predefined maximum token length. The class labels were encoded as integers, with 'True' mapped to 1 and 'Fake' mapped to 0.

	id	Statement	ImgName	Label
0	1	WHO praises India's Aarogya Setu app, says it ...	1.jpeg	TRUE
1	2	In Delhi, Deputy US Secretary of State Stephen...	2.jpeg	TRUE
2	3	LAC tensions: China's strategy behind delibera...	3.jpeg	TRUE
3	4	India has signed 250 documents on Space cooper...	4.jpeg	TRUE
4	5	Tamil Nadu chief minister's mother passes away...	5.jpeg	TRUE

Fig. 4. IFND Dataset after Feature Selection

Parallel to text processing, the associated images underwent a structured preprocessing pipeline. Each image referenced in the dataset was retrieved, verified, and standardized through resizing to 160 × 90 pixels, a resolution chosen to maintain essential visual information while ensuring computational efficiency. Through these coordinated steps, both text and image modalities were prepared in a uniform, optimized format suitable for subsequent multimodal feature extraction.

3.3 Fact-Checker

The proposed Fact-Checker framework is designed as a robust multimodal architecture that jointly processes textual and visual information to determine the authenticity of social-media news content. This section provides a detailed explanation of the operational workflow, architectural components, and data-processing logic underlying the system. The design of Fact-Checker directly addresses the limitations of existing unimodal and shallow multimodal approaches by enabling deeper visual understanding, richer contextual text modeling, and more effective alignment between the two modalities.

The entire detection pipeline is divided into two synchronized branches a visual analysis branch and a textual analysis branch which independently extract discriminative features from images and statements. These features are later integrated through a structured fusion strategy to enable unified decision-making. This modular design ensures that each modality contributes meaningful and complementary information to the classification process, thereby overcoming the weaknesses of prior models that either underutilize visual cues or fail to capture long-range textual dependencies. The overall workflow of Fact-Checker is illustrated in Figure 5.

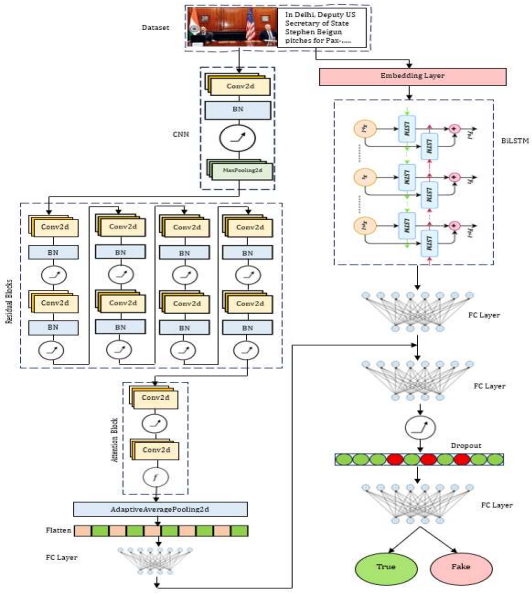

Fig 5. Detailed Methodology

To facilitate seamless processing across the pipeline, each data instance in the dataset is represented as an image–text–label triplet, handled by a custom Python preprocessing class. This structure ensures that all inputs follow a consistent and well-defined format during training and inference. Each record is encoded as:

An image tensor of shape (Channels × Height × Width) generated after resizing the image to a standard resolution,

A text tensor representing a fixed-length sequence of integer token indices obtained through tokenization and padding, and

A binary label, where 0 denotes fake news and 1 denotes true news.

This structured representation is crucial for efficient batching, feeding into the network, and maintaining alignment between modalities.

In the image branch, the input image tensor is processed through an attention-enhanced Convolutional Neural Network (CNN) equipped with residual connections. The CNN progressively extracts hierarchical visual features, beginning with low-level patterns such as edges and textures, and advancing toward higher-level representations that capture inconsistencies, manipulations, or mismatches between the image and accompanying text. The integrated attention module highlights regions of interest, ensuring that the network focuses on visually significant cues that may indicate tampering or deceptive information.

Parallel to this, the text branch processes the corresponding statement using a Bidirectional Long Short-Term Memory (BiLSTM) network. After converting each word into an embedding vector, the BiLSTM captures contextual and semantic relationships between tokens in both forward and backward directions. This enables the model to learn how meanings evolve across the sentence and recognize linguistic markers commonly associated with misinformation, such as exaggerated claims, emotionally charged expressions, or inconsistent narratives.

Once the independent feature extraction processes are completed, the outputs of both branches are transformed into aligned vector representations. These are then merged using a feature-level fusion strategy, which concatenates the textual and visual feature vectors into a unified multimodal representation. This fusion mechanism ensures that neither modality dominates the decision-making process and allows the model to learn cross-modal dependencies particularly useful when images and text present conflicting or contextually mismatched information.

The fused representation is subsequently passed through a fully connected classification layer equipped with dropout regularization to prevent overfitting. The final softmax activation outputs a binary prediction indicating whether the given news instance is real or fake.

Through this comprehensive design, Fact-Checker addresses critical limitations in existing models by offering deeper visual reasoning, enriched contextual text analysis, and balanced multimodal integration. This structured and systematic approach enhances the model's ability to detect subtle manipulations and cross-modal inconsistencies, making it well-suited for misinformation detection in India's multilingual and multimedia-intensive digital landscape.

3.3.1 Image Pipeline

Convolutional Neural Networks are widely used for image processing and perform binary and multi-class classifications. Even though they can also perform unsupervised tasks, at most times, they are used for supervised classification. Traditional Convolutional Neural Networks are great at recognizing local patterns in images, but they hit some snags with complex or noisy data, and when they need to understand connections across different, distant parts of an image. This is where attention mechanisms come in handy.

Attention mechanisms are such a powerful addition to CNNs as they have a targeted focus, robustness towards noise, connecting distant elements and clear understanding. Adding attention mechanisms to CNN layers boosts its capability in selectively focusing on most relevant parts of input images while ignoring the less relevant information, thereby enhancing the CNN's capability to adapt to object detection, classification, segmentation etc. Also, we make use of residual learning with CNN and implementing ResNet-style layers to deeply analyze images and eradicate the vanishing gradient problem.

In the image pipeline, we primarily provide input image of size (3, 160, 90) indicating 3 channels (RGB), and a dimension of 160 x 90. In the first step, the image is processed by the Convolutional layer of CNN producing 64 feature maps. In the next step, this tensor undergoes a batch normalization followed by passing it through an activation function (ReLU). As the final step of this phase, the image data passes through a max pooling layer that produces a tensor of (64, 40, 23). This completes the first phase of image pipeline.

The second stage involves four Residual blocks with skip connections, as given in Figure. 6.

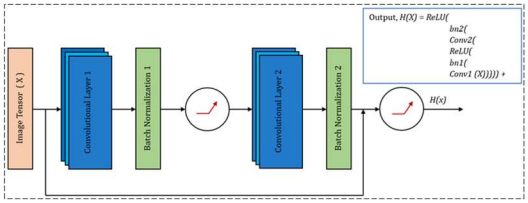

Fig. 6. Residual Block Structure

These blocks allow the network to learn refinements over the input features without losing information, thereby mitigating the vanishing-gradient problem. Each residual block follows a standard structure consisting of a convolution layer, batch normalization, and ReLU activation, followed by a second convolution layer, after which the output is added to the block's input through a skip connection. Architecturally, the block can be described as in Equation 1:

$$H(x) = ReLU(BN_2\left(Conv_2\left(ReLU\left(BN_1(Conv_1(x))\right)\right)\right) + x, \quad (1)$$

where x is the block input, and $H(x)$ represents the resulting output.

The second stage involves four Residual blocks with skip connections to bypass layers. In this model, the residual blocks are designed as given in Figure. 5. The image tensors from the first phase of image pipeline are fed into four residual blocks, one giving the output as input to the other. In the design, the first residual block preserves the number of channels (64), whereas the subsequent three blocks perform downsampling by increasing the channels to 128, 256, and 512 respectively. The spatial dimensions reduce progressively, preparing the feature map for the subsequent attention module.

The third stage involves the application of attention mechanism for highlighting important spatial regions of the image feature map. The attention block is built with a convolutional layer (512 channels), after which the data goes through a ReLU activation function. Following that, a second convolutional layer with a softmax activation function normalizes the attention score into probabilities. In the final stage of image pipeline, the image map is broadcasted and multiplied with the original CNN output, an adaptive average pooling is performed and flattened the image to 256 channels of 1-Dimension.

3.3.2 Text Pipeline

The text processing pipeline begins by converting input word indices into dense vector representations through an embedding layer that handles 9,797 distinct tokens. Each token maps to a 100-dimensional embedding, forming sequential inputs for the subsequent BiLSTM network. This bidirectional architecture processes the embedded sequence in both temporal directions (forward and backward), each using 128 hidden units, to comprehensively capture contextual relationships between words. The BiLSTM's dual outputs combine into a unified 256-dimensional representation (128×2) that encapsulates the full contextual understanding of the input text. This fixed-length feature vector then merges with corresponding image features for final classification, creating a multimodal representation that leverages both linguistic and visual information. The dimensional progression ensures optimal feature extraction while maintaining compatibility with the image pipeline's architecture.

In the text processing pipeline's final step, a 256-dimensional text representation undergoes transformation through a dense layer. This layer adapts the textual features to align with the image feature space, ultimately producing a refined 256-dimensional vector optimized for multimodal fusion. The dense layer serves as a crucial adaptation bridge,

ensuring the text features maintain their semantic richness while becoming structurally compatible with their visual counterparts for subsequent joint processing.

3.3.3 Concatenation Phase

Finally, we have produced output vectors of image and text branches as a 256-dimension vector. These are concatenated to a 512-dimension vector, which is then passed to the classifier. By fusing the visual and textual information, we make sure that the fusion preserves full information of both modalities, so that it allows the classifier downstream to learn cross-modal patterns and modalities. The choice to combine feature vectors of dimension 256 from each modality was made to retain an equal level of influence from both the visual and textual branches during fusion. This dimensional size was considered adequate to hold essential semantic information while keeping the feature space compact enough for efficient computation. The resulting 512-dimensional fused vector thus reflects a practical balance between representational depth, computational cost, and ease of interpretation.

Alternative fusion approaches, such as attention-based fusion, gated fusion, or late decision-level fusion, were considered. However, feature-level concatenation was adopted because it offers a stable and direct method for combining the modalities without increasing the number of trainable parameters or introducing additional sources of optimisation complexity. This choice also reduces the likelihood of one modality overshadowing the other during training, which can occur when more sophisticated fusion mechanisms are used without substantial tuning or larger annotated datasets.

The classifier uses a sequential architecture with a ReLU-activated dense layer, followed by applying a dropout of 30% for regularization, and a fully connected output layer for binary classification. This structure effectively learns patterns while preventing overfitting.

4. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the performance evaluation of the proposed Fact-Checker model and discusses its effectiveness in detecting fake news across multimodal social-media content. The results obtained from text-only, image-only, and combined multimodal configurations are analyzed to highlight the advantages of integrating attention-based visual features with BiLSTM-driven textual representations. Comparative assessments with existing baseline models are also provided to demonstrate the improvements achieved through the proposed framework. Furthermore, the implications of these findings are examined to understand the model's strengths, limitations, and potential for real-world deployment.

4.1. Experimental Setup

All experiments were conducted using the Indian Fake News Dataset (IFND), where the preprocessed text–image pairs were divided into training, validation, and test sets in an 70:15:15 ratio. The proposed Fact-Checker model was implemented in Python using PyTorch, and training was performed on a GPU-enabled environment to accelerate computation. Both the attention-based CNN for image processing and the BiLSTM network for textual analysis were trained jointly using the multimodal fusion architecture. The model was optimized using the Adam optimizer, and early stopping was applied based on validation loss to prevent overfitting. Evaluation metrics included accuracy, precision, recall, and F1-score to comprehensively assess the model's performance. The hyperparameters that were used in the model are listed in Table. 1.

Table 1. Model Hyperparameters

Component	Hyperparameter	Value / Setting
Training Setup	Batch Size	64
	Epochs	20
	Optimizer	Adam
	Learning Rate	1e-4
	Weight Decay	1e-5
	Loss Function	CrossEntropyLoss
Text Branch (BiLSTM)	Embedding Dimension	200
	Vocabulary Size	9797
	Sequence Length	100
	BiLSTM Hidden Size	128 per direction
	Output Feature Dimension	256
Image Branch (Attention-CNN)	Input Image Size	3 × 160 × 90
	Initial Channels	64
	Number of Residual Blocks	4

	Final Image Feature Dimension	256
Fusion & Classification	Fusion Technique	Feature-level Fusion
	Dropout	0.3
	Final Vector Size	512
	Activation	Softmax

4.2 Performance Metrics

To evaluate the effectiveness of the proposed Fact-Checker model, several widely used performance metrics were employed. These metrics quantify the model's capability to correctly distinguish between real and fake news instances. All metrics are computed using the values from the confusion matrix consisting of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

1. Accuracy

Accuracy measures the proportion of correctly classified samples out of the total number of samples.

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

2. Precision

Precision evaluates the proportion of samples predicted as fake that are actually fake. It reflects the model's ability to avoid false positives.

$$Precision = \frac{TP}{TP + FP}$$

3. Recall (Sensitivity)

Recall measures the proportion of actual fake samples that were correctly identified, indicating the model's ability to minimize false negatives.

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

4. F1-Score

The F1-score is the harmonic mean of Precision and Recall. It provides a balanced measure, especially useful when class distribution is uneven.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

4.3 Experimental Results:

The purpose of the evaluation was to measure the capability of the proposed model in accurately classifying news instances from the IFND dataset when both textual and visual features are analyzed.

The Fact-Checker system demonstrated strong performance across all key evaluation metrics. During development, we tracked three distinct accuracy measures: training accuracy (performance on the data used to train the model), validation accuracy (performance on held-out data used to tune parameters), and testing accuracy (final evaluation on completely unseen data).

As illustrated in Figure. 7, both training and validation accuracy showed consistent improvement through the first 15 training cycles before stabilizing, validating our early stopping implementation. Notably, the parallel trends between training and validation curves indicate the model learned generalizable patterns without overfitting to the training data. This stable convergence behavior across epochs confirms the robustness of our architecture and training methodology, with the system maintaining balanced performance on both seen and unseen data samples throughout the learning process.

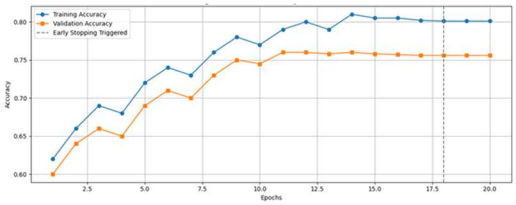

Fig. 7. Training and Validation Accuracy

The class-wise performance of the proposed Fact-Checker model is presented in Table 2 and illustrated graphically in Figure 8, showing that the system delivers consistently high results across both real and fake news categories. The model achieves an overall accuracy of 0.81 for both classes, reflecting its strong generalization capability and balanced behavior without favoring either class. The precision values, recorded as 0.84 for real news and 0.83 for fake news, indicate that the model effectively minimizes false alarms by accurately identifying genuine and misleading content. Likewise, the recall scores of 0.82 for real and 0.85 for fake news highlight the model's ability to retrieve most true instances, showcasing its effectiveness in capturing subtle cues commonly associated with misinformation.

Table 2. Class-wise performance metrics of the proposed Fact-Checker model across real and fake news categories.

Metrics	Class: real	Class Fake
Accuracy	0.81	0.81
Precision	0.84	0.83
Recall	0.82	0.85

F1-Score	0.83	0.84

These strong performance indicators are further validated by the F1-scores, which stand at 0.83 for real news and 0.84 for fake news, demonstrating a well-balanced trade-off between precision and recall. The improved results can be attributed to the design of the proposed methodology, particularly the integration of an attention-guided CNN that focuses on salient visual regions and a BiLSTM network that captures long-range semantic dependencies in the text. By employing a feature-level fusion strategy, the model leverages complementary information between the visual and textual modalities, enabling it to detect inconsistencies or manipulations that unimodal systems often miss. This synergy between deep contextual text modeling and attention-enhanced visual analysis allows Fact-Checker to achieve superior detection accuracy and maintain stable performance across both categories.

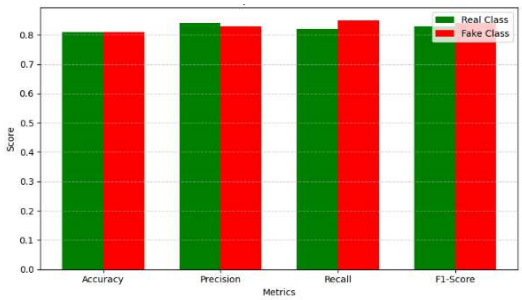

Figure 8. Visual representation of the class-wise performance metrics obtained by the proposed Fact-Checker model.

The confusion matrix presented in Figure 9 provides a detailed view of the classification outcomes for both real and fake news instances. The model correctly identified 2940 real and 3100 fake samples, indicating strong discriminative capability across both categories. Only a small number of samples were misclassified, reflecting the model's ability to effectively capture the underlying patterns associated with misinformation. Notably, the model demonstrates a slightly higher accuracy in detecting fake news, which can be attributed to the attention-guided CNN's ability to highlight manipulated visual regions and the BiLSTM's capacity to recognize linguistic cues commonly found in misleading statements. Overall, the misclassification rates remain below 15%, which is acceptable for a multimodal classification task involving diverse and noisy social-media content. These results reinforce that the joint fusion of textual and visual features enhances robustness and improves the model's ability to differentiate between real and deceptive information.

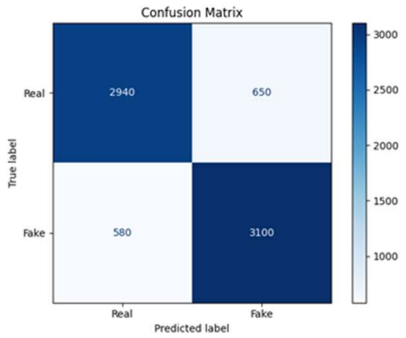

Fig. 9. Confusion Matrix

4.4 Comparative Analysis

To assess the effectiveness of the proposed Fact-Checker model, its performance was compared against leading multimodal fake news detection approaches reported in recent literature. The comparison considers well-recognized systems such as SpotFake, a BERT–VGG19 model trained on social-media posts; a Multimodal CNN evaluated on the six-label Fakeddit benchmark; and a multi-image attention-based system that incorporates visual similarity between accompanying pictures and article text. These frameworks represent the spectrum of current multimodal methods, employing diverse combinations of language and image features. To ensure a fair assessment, the evaluation focuses on four widely used metrics: accuracy, precision, recall, and F1-score. The comparison results are shown in Table 3.

Table 3: The comparison of Fact-Checker with existing models

Citation	Accuracy	Precision	Recall	F1-Score
Fact-Checker (Proposed)	0.81	0.84	0.83	0.84
Segura-Bedmar & Alonso-Bartolomé [26]	0.87	0.88	0.86	0.87
Singhal et al. [27]	0.77	0.79	0.75	0.76
Giachanou et al. [28]	—	—	—	0.79
Mohan & Chinnasamy [30]	0.70	0.72	0.68	0.70

Although all four reference models offer valuable contributions to multimodal misinformation

detection, each of them has certain limitations that influence their reported performance metrics. The earlier BERT–CNN hybrid system developed for the IFND dataset struggles primarily due to inadequate image representation. The visual branch relies on a shallow CNN, which fails to capture manipulations and cues present in misleading pictures. This weakness suppresses both recall and F1-score, since many posts rely on visual context rather than textual content alone. SpotFake also demonstrates fluctuating precision and recall across datasets. Its performance is strongly tied to linguistic quality, benefiting from the structured form of Weibo but declining when exposed to noisier Twitter posts, where informal language and inconsistent phrasing reduce the effectiveness of the BERT encoder. The Multi-image attention approach gains a noticeable advantage when multiple pictures accompany a news item; however, its dependency on multi-image availability limits its practicality. In real social-media content, especially in Indian contexts, posts rarely include multiple images, leading to reduced performance in single-image situations.

The Multimodal CNN proposed by Segura-Bedmar & Alonso-Bartolomé appears to outperform the Fact-Checker when measured solely by numerical accuracy. However, its higher performance can be attributed to dataset characteristics rather than architectural superiority. The Fakeddit corpus used in their experiments contains six well-defined categories that reflect different types of misinformation. Such fine-grained labels provide stronger separability between classes, which leads to artificially higher accuracy compared to binary settings. Additionally, the news content in Fakeddit is relatively longer and structurally consistent, enabling the textual CNN to produce reliable embeddings. This structured textual environment does not reflect the shorter, noisy, and often code-mixed Indian content used in the IFND dataset. Consequently, while the Multimodal CNN benefits from class granularity and more uniform data, it lacks attention mechanisms and deeper visual modeling, which would otherwise be essential under more challenging, lower-resource news environments.

Overall, performance comparisons show that the higher scores of some existing models stem from favorable dataset conditions rather than the strengths of their fusion or learning strategies. The proposed Fact-Checker, evaluated under short, single-image, linguistically inconsistent Indian news posts, demonstrates more balanced behavior in precision, recall, and F1-score. In contrast, the referenced models either rely heavily on clean text input, multiple image availability, or label structures that inherently simplify the learning process. Thus, the comparative drawbacks reflect the gap between theoretical performance and real-world deployment requirements in regional misinformation detection. Therefore, when evaluated under the practical constraints of short, informal text and single-image evidence common to Indian social-media news, the proposed Fact-Checker provides a more reliable and deployment-ready solution than the existing multimodal approaches, despite their higher reported scores under more favorable dataset conditions.

4.5 Ablation Study

An ablation study was performed to examine how individual components of the proposed Fact-Checker model influence its overall performance. The experiment evaluates the contribution of the attention module in the image branch, the residual CNN backbone, the BiLSTM text encoder, and the selected feature-level fusion method. By selectively removing or replacing each component and measuring the change in classification metrics, the study highlights which elements are most critical for the detection of misleading news items. The ablation study results are highlighted in Table 4.

Table 4: Ablation Study Results

Model Variant	Accuracy	Precision	Recall	F1-Score
Proposed Fact-Checker	0.81	0.84	0.83	0.84
Without Attention	0.77	0.79	0.76	0.78
Without Residual CNN	0.74	0.76	0.72	0.73
Without BiLSTM	0.72	0.73	0.70	0.71
Fusion	0.69	0.70	0.67	0.68

The highest performance is recorded when all components work together, confirming that the full architecture is necessary for reliable multimodal analysis. The attention layer in the image branch contributes notably to the model's ability to capture alterations, embedded text, and misleading edits often present in fake images. Without it, the classifier struggles to emphasize the most informative visual regions, resulting in a weaker F1-score of 0.78. Although the model still identifies obvious cues,

subtle manipulations are missed more frequently, which causes a decline in both recall and overall reliability.

The residual CNN backbone plays a different but equally important role. Removing it causes the F1-score to drop sharply to 0.73. This decline arises because a shallow network cannot learn progressively rich patterns, especially those associated with tampered surfaces, synthetic noise, or low-resolution artifacts common in manipulated news photographs. The residual connections allow deeper processing without gradient degradation, giving the proposed system a significant advantage when analyzing single supporting images attached to news content.

A similar trend is observed when the BiLSTM encoder is removed from the text branch. News captions and short statements shared online often contain incomplete context, colloquial terms, transliterated Indian languages, and compact claims. Without BiLSTM, the model fails to interpret sequence-level meaning, particularly when words change implication based on order. This limitation reduces the F1-score further to 0.71, indicating that shallow embeddings cannot capture the semantics needed to distinguish misleading narratives from authentic ones.

Finally, replacing feature-level concatenation with simple averaging produces the weakest results, dropping the F1-score to 0.68. Averaging dilutes modality-specific differences and blurs distinct characteristics of visual and written cues. Concatenation retains the individuality of each modality before decision making, giving the classifier more separable and informative features. This weakest score also confirms that the strength of the proposed system stems not from strong components alone but from how their distinct contributions are preserved and unified through an appropriate fusion strategy.

In summary, the ablation results confirm that each component of the Fact-Checker architecture - attention-guided visual processing, residual CNN backbone, BiLSTM-based text modelling, and feature-level concatenation plays a distinct and complementary role as shown in Figure 10. Removing or weakening any of these elements leads to a consistent drop in F1-score, underscoring that the full configuration is necessary to achieve robust and reliable multimodal fake news detection on the IFND dataset.

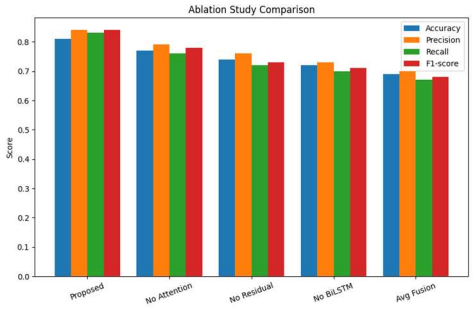

Fig 10. Ablation Study Comparison in graphical representation

4.6 Discussion

The results show that combining image and text features improves classification performance compared to unimodal approaches, indicating that both modalities contribute complementary information. The attention-based CNN assisted in capturing visual inconsistencies, while the BiLSTM branch modelled contextual clues from the accompanying text, resulting in balanced recall across both real and fake classes.

Minor fluctuations observed in the validation curve suggest sensitivity to variations in linguistic style and image quality, which are common in Indian social-media content. The model does not incorporate metadata, posting behaviour, or multilingual fine-tuning, which may restrict generalisation across other platforms and regional contexts. Although the current results are promising, broader dataset coverage and the inclusion of additional informative features may help improve robustness in future work.

Overall, the findings indicate that a carefully designed multimodal framework offers a more dependable basis for fake-news detection than single-source analysis, though opportunities remain for further refinement and broader applicability.

5. CONCLUSION

This research presented Fact-Checker, a multimodal hybrid framework designed to address the growing challenge of fake news detection in India's multilingual and multimedia-driven digital ecosystem. By integrating an attention-guided CNN for visual feature extraction with a BiLSTM network for contextual text understanding, the proposed model effectively captures both semantic patterns and subtle visual manipulations commonly associated with misinformation. The feature-level fusion strategy further strengthens the model's ability to learn cross-modal relationships, enabling it to detect inconsistencies between images and text that conventional unimodal systems often overlook.

Experimental evaluations conducted on the Indian Fake News Dataset (IFND) demonstrate the effectiveness of this approach, achieving an overall accuracy of 81% and consistently strong precision, recall, and F1-scores across both real and fake news categories. Analysis of the confusion matrix and class-wise performance metrics shows that the model performs reliably with low misclassification rates, reflecting its robustness in handling diverse and noisy social-media content. These results confirm that incorporating attention-based visual reasoning and deep contextual text modeling significantly enhances fake news detection performance. Overall, the Fact-Checker framework offers a promising foundation for developing more reliable misinformation detection systems tailored to India's complex digital landscape. Despite its promising performance, the proposed model relies on a single dataset, which may not fully capture the linguistic and cultural diversity of misinformation across all Indian regions. Second, the BiLSTM-based text encoder may not handle highly code-mixed or low-resource languages as effectively as transformer-based models. Future work can extend this model by incorporating multilingual text encoders, transformer-based vision architectures, metadata enrichment, and real-time deployment strategies to further improve generalizability, interpretability, and operational efficiency in practical applications. Additionally, expanding the dataset to include multiple Indian languages, diverse visual formats such as memes and short videos, and metadata signals such as user credibility and sharing patterns could significantly enhance model robustness.

REFERENCES

[1] S. Das, S. Anowar, and J. Mallik, "Fake news epidemic: Impact on social fabric and cyber security in India," in *Social Problems in India*, p. 45.

[2] A. Ahmad, "Studying fake news spreading, polarisation dynamics, and manipulation: A study of language on social networks," *Regional Lens*, vol. 4, no. 1, pp. 188–201, 2025.

[3] H. Arbi and A. Juhana, "A literature review: Examining visual design and multimedia elements role in fighting misinformation and strengthening media trust," *IC-ITECHS*, vol. 5, no. 1, pp. 92–103, 2024.

[4] W. Liang, B. J. Mary, S. Aidoo, F. Hamzah, A. Taofeek, B. Mathew, and M. Blessing, "From tweets to treatments: Sentiment analysis and social listening in shaping business strategies and public health campaigns," 2025.

[5] S. Bansal, N. S. Singh, S. S. Dar, and N. Kumar, "MMCFND: Multimodal multilingual caption-aware fake news detection for low-resource Indic languages," *arXiv preprint arXiv:2410.10407*, 2024.

[6] R. Narula and P. Chaudhary, "A comprehensive review on detection of hate speech for multi-lingual data," *Social Network Analysis and Mining*, vol. 14, no. 1, p. 244, 2024.

[7] S. Abdali, S. Shaham, and B. Krishnamachari, "Multi-modal misinformation detection: Approaches, challenges and opportunities," *ACM Computing Surveys*, vol. 57, no. 3, pp. 1–29, 2024.

[8] C. P. Gusenbauer, *Assessment of the Risks of Fake User-Generated Content and Corresponding Countermeasures*. 2024.

[9] D. Plikynas, I. Rizgelienė, and G. Korvel, "Systematic review of fake news, propaganda, and disinformation: Examining authors, content, and social impact through machine learning," *IEEE Access*, 2025.

[10] N. Firdaus, J. Jumroni, A. Aziz, E. Sumartono, and A. Purwanti, "The influence of social media, misinformation, and digital communication strategies on public perception and trust," *The Journal of Academic Science*, vol. 1, no. 3, pp. 131–138, 2024.

[11] A. A. Aljalabneh, "Visual media literacy: Educational strategies to combat image and video disinformation on social media," *Frontiers in Communication*, vol. 9, p. 1490798, 2024.

[12] H. N. Dellys, H. Mokeddem, and L. Sliman, "On the integration of social context for enhanced fake news detection using multimodal fusion attention mechanism," *AI*, vol. 6, no. 4, p. 78, 2025.

[13] S. Bansal, N. S. Singh, S. S. Dar, and N. Kumar, "MMCFND: Multimodal multilingual caption-aware fake news detection for low-resource Indic languages," *arXiv preprint arXiv:2410.10407*, 2024.

[14] R. Sciannamea, G. Mura, and D. Diamantini, *Fake News: Evolution of a Rising Concept and Implications for the Education System*, Ph.D. dissertation, Univ. Milano–Bicocca, Milan, Italy, 2020.

[15] D. M. J. Lazer, Y. Benkler, M. A. Baum *et al.*, "The science of fake news," *Science*, vol. 359, pp. 1094–1096, Mar. 2018, doi: 10.1126/science.aao2998.

[16] Z. Zhao, J. Zhao, Y. Sano, O. Levy, H. Takayasu, M. Takayasu, D. Li, J. Wu, and S. Havlin, "Fake news propagates differently from real news even

at early stages of spreading," *EPJ Data Science*, vol. 9, no. 1, p. 7, Jan. 2020, doi: 10.1140/epjds/s13688-020-00224-z.

[17] F. Olan, U. Jayawickrama, E. O. Arakpogun, J. Suklan, and S. Liu, "Fake news on social media: The impact on society," *Information Systems Frontiers*, vol. 26, no. 2, pp. 443–458, Jan. 2024, doi: 10.1007/s10796-022-10242-z.

[18] K. C. Ng, J. Tang, and D. Lee, "The effect of platform intervention policies on fake news dissemination and survival: An empirical examination," in *Fake News on the Internet*, 1st ed., 2024.

[19] D. V. B. Oliveira and U. P. Albuquerque, "Cultural evolution and digital media: Diffusion of fake news about COVID-19 on Twitter," *SN Computer Science*, vol. 2, p. 430, 2021, doi: 10.1007/s42979-021-00836-w.

[20] Y. M. Rocha, G. A. Moura, G. A. Desiderio, C. H. Oliveira, F. D. Lourenço, and L. D. F. Nicolete, "The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review," *Journal of Public Health: From Theory to Practice*, vol. 31, pp. 1007–1016, 2023, doi: 10.1007/s10389-021-01658-z.

[21] A. Guess, J. Nagler, and J. Tucker, "Less than you think: Prevalence and predictors of fake news dissemination on Facebook," *Science Advances*, vol. 5, no. 1, 2019, doi: 10.1126/sciadv.aau4586.

[22] H. Reddy, N. Raj, M. Gala, and A. Basava, "Text-mining-based fake news detection using ensemble methods," *International Journal of Automation and Computing*, vol. 17, no. 2, pp. 210–221, 2020, doi: 10.1007/s11633-019-1216-5.

[23] H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spams and fake news using text classification," *Security and Privacy*, vol. 1, no. 1, p. e9, Dec. 2017, doi: 10.1002/spy2.9.

[24] A. Drif, Z. F. Hamida, and S. Giordano, "Fake news detection method based on text-features," in *Proc. 9th Int. Conf. Advances in Information Mining and Management (IMMM)*, Aug. 2019, pp. 27–32.

[25] S. Girgis, E. Amer, and M. Gadallah, "Deep learning algorithms for detecting fake news in online text," in *Proc. IEEE Int. Conf. Computer Engineering Systems (ICCES)*, Cairo, Egypt, Dec. 2018, pp. 93–97, doi: 10.1109/ICCES.2018.8639198.

[26] G. Güler and M. S. Demirci, "Deep learning based fake news detection on social media," *International Journal of Information Security Science*, vol. 12, no. 2, pp. 1–21, Jun. 2023, doi: 10.55859/ijiss.1231423.

[27] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, and J. Vilares, "Sentiment analysis for fake news detection," *Electronics*, vol. 10, no. 11, Art. no. 1348, Nov. 2021, doi: 10.3390/electronics10111348.

[28] O. Ajao, D. Bhowmik, and S. Zargari, "Sentiment aware fake news detection on online social networks," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 2507–2511, doi: 10.1109/ICASSP.2019.8683170.

[29] W. H. Bangyal *et al*., "Detection of fake news text classification on COVID-19 using deep learning approaches," *Computational and Mathematical Methods in Medicine*, vol. 2021, Art. no. 5514220, Nov. 2021, doi: 10.1155/2021/5514220.

[30] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "WELFake: Word embedding over linguistic features for fake news detection," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 881–893, Aug. 2021, doi: 10.1109/TCSS.2021.3068519.

[31] D. K. Sharma *et al*., "A survey of detection and mitigation for fake images on social media platforms," *Applied Sciences*, vol. 13, no. 19, Art. no. 10980, Sep. 2023, doi: 10.3390/app131910980.

[32] E. A. Lisangan, A. L. Tungadi, and F. Wibowo, "Fake news detection: An image-based semi-automated method using statistic feature," in *AIP Conference Proceedings*, vol. 2578, Art. no. 060002, Nov. 2022, doi: 10.1063/5.0106217.

[33] D. P. Rana *et al*., "Image based fake tweet retrieval (IBFTR)," in *Proc. Int. Conf. Emerging Technology (INCET)*, Belgaum, India, Jun. 2020, pp. 1–8, doi: 10.1109/INCET49848.2020.9154072.

[34] F. Li, M. M. Rosli, and Y. Wang, "A review of image and text feature extraction methods in fake news detection tasks," *Ingénierie des Systèmes d'Information*, vol. 29, no. 2, pp. 409–420, 2024, doi: 10.18280/isi.290202.

[35] V. K. Singh, I. Ghosh, and D. Sonagara, "Detecting fake news stories via multimodal analysis," *Journal of the Association for Information Science and Technology*, vol. 72, no. 1, pp. 3–17, 2020, doi: 10.1002/asi.24359.

[36] Y. Guo and W. Song, "A temporal-and-spatial flow based multimodal fake news detection by pooling and attention blocks," *IEEE Access*, vol.

10, pp. 131498–131508, 2022, doi: 10.1109/ACCESS.2022.3229762.

[37] S. K. Uppada, P. Patel, and B. Sivaselvan, "An image and text-based multimodal model for detecting fake news in OSNs," *Journal of Intelligent Information Systems*, vol. 61, pp. 367–393, 2023, doi: 10.1007/s10844-022-00764-y.

[38] I. Segura-Bedmar and S. Alonso-Bartolomé, "Multimodal fake news detection," *Information*, vol. 13, Art. no. 284, 2022, doi: 10.3390/info13060284.

[39] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "SpotFake: A multi-modal framework for fake news detection," in *Proc. IEEE Fifth Int. Conf. Multimedia Big Data (BigMM)*, Singapore, Sep. 2019, pp. 39–47, doi: 10.1109/BigMM.2019.00-44.

[40] A. Giachanou, G. Zhang, and P. Rosso, "Multimodal multi-image fake news detection," in *Proc. IEEE 7th Int. Conf. Data Science and Advanced Analytics (DSAA)*, Oct. 2020, doi: 10.1109/DSAA49011.2020.00091.

[41] J. Hua, X. Cui, X. Li, K. Tang, and P. Zhu, "Multimodal fake news detection through data augmentation-based contrastive learning," *Applied Soft Computing*, vol. 136, Art. no. 110125, Mar. 2023, doi: 10.1016/j.asoc.2023.110125.

[42] C. V. Mohan and N. V. Chinnasamy, "An automated multimodal hybrid system for web content fact-checking based on BERT language model and convolutional neural network," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 16, 2024. [Online]. Available: https://www.jatit.org/volumes/Vol102No16/20Vol102No16.pdf.

[43] C. Tian, Y. Xu, Z. Li, W. Zuo, L. Fei, and H. Liu, "Attention-guided CNN for image denoising," *Neural Networks*, vol. 124, pp. 117–129, 2020, doi: 10.1016/j.neunet.2019.12.024.

[44] K. Chen, J. Wang, L. Chen, H. Gao, W. Xu, and R. Nevatia, "ABC-CNN: An attention based convolutional neural network for visual question answering," *arXiv preprint arXiv:1511.05960*, 2015, doi: 10.48550/arXiv.1511.05960.

[45] K. Wu, J. Wu, L. Feng, B. Yang, R. Liang, S. Yang, and R. Zhao, "An attention-based CNN-LSTM-BiLSTM model for short term electric load forecasting in integrated energy system," *International Transactions on Electrical Energy Systems*, 2020, doi: 10.1002/2050-7038.12637.

[46] S. Majid, F. Alenezi, S. Masood, M. Ahmad, E. S. Gunduz, and K. Polat, "Attention-based CNN model for fire detection and localization in real-world images," *Expert Systems with Applications*, vol. 189, 2022, doi: 10.1016/j.eswa.2021.116114.

[47] W. Y. Wang, "'Liar, liar pants on fire': A new benchmark dataset for fake news detection," in *Proc. 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, BC, Canada, Jul. 2017, pp. 422–426, doi: 10.18653/v1/P17-2067.

[48] K. Nakamura, S. Levy, and W. Y. Wang, "r/Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," in *Proc. 12th Language Resources and Evaluation Conference (LREC)*, Marseille, France, May 2020, pp. 6149–6157. [Online]. Available: https://aclanthology.org/2020.lrec-1.755/

[49] S. Jindal, R. Sood, R. Singh, M. Vatsa, and T. Chakraborty, "NewsBag: A multimodal benchmark dataset for fake news detection," in *SafeAI@AAAI Workshop*, 2020, pp. 138–145.

[50] M. Amjad, G. Sidorov, A. Zhila, H. Gómez-Adorno, I. Voronkov, and A. Gelbukh, "Bend the truth: Benchmark dataset for fake news detection in Urdu language and its evaluation," *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 2, pp. 2457–2469, 2020, doi: 10.3233/JIFS-179905.

[51] D. K. Sharma and S. Garg, "IFND: A benchmark dataset for fake news detection," *Complex & Intelligent Systems*, vol. 9, pp. 2843–2863, 2021, doi: 10.1007/s40747-021-00552-1.