ESP-IJACT
ESP International Journal of Advancements in Computational Technology
ISSN: 2583-8628 / Volume 2 Issue 2 June 2024 / Page No: 133-141
Paper Id: IJACT-V2I2P117 / Doi: 10.56472/25838628/IJACT-V2I2P117

Original Article

Multimodal Gen AI: Integrating Text, Image, and Video Analysis for Comprehensive Claims Assessment

Sateesh Reddy Adavelli
Solution Architect, USA

Received Date: 11 April 2024 Revised Date: 16 May 2024 Accepted Date: 12 June 2024

Abstract: *The increase in claim sophistication in both the insurance and legal domains is a result of an increase in stokes and heterogeneity of data needed to assess the claim validity. Originally, this task was performed by some sort of subjectivity assessments and graphical rule sets, which is very slow and may be inherently erroneous due to its purely manual nature. Hence, with progressivity in multimodal learning, specifically in AI, there is now a unique chance of solving these challenges through the use of text data, which may include policies, reports and images, which may include accident images, evidence images, videos such as surveillance, cam videos among others. However, existing AI-based solutions usually address only one of the modalities, which makes it difficult to evaluate an integrated situation. This has led to the need for systems that will integrate information from all these modalities and come up with an accurate, efficient, and transparent processing system.*

Indeed, this paper seeks to discuss the use of Multimodal generative AI to address this need as one of the most recent approaches that rely on high-performing models that can process and integrate text, image, and video data. The proposed system combines these modalities to ensure that the system captures relevant data from each data type and combines all in a way that provides more comprehensive and enriched decision support. An initial system was designed and empirically tested against current claim adjudication techniques and was found to yield substantial enhancements in all utilization rates, throughput and main rationale for the claim decisions. The findings shown in the study stress the capability of multimodal generative AI for revolutionizing the present approaches of claims analysis and developing more efficient, accurate, and capable responses to various real-life conditions. This integration of technologies is an unprecedented advance towards advancing functional processes in the insurance and legal industries.

Keywords: *Multimodal Generative AI, Claims Assessment, Text Analysis, Image Analysis, Video Analysis, Deep Learning.*

I. INTRODUCTION

A. Challenges in Traditional Claims Assessment

Most insurance and legal industry claims assessment processes are done manually and require a lot of time and experience. Through research, adjusters are obliged to study a huge number of photos, recorded documents, and sometimes even videos in order to review the validity of the claim and its potential monetary value. However, the choice of drugs and their dosages, relying on this manual approach, is rather time-consuming and not very effective, which results in the inconsistency of decisions made. Third of all, in today's world, there is an exponential trend in the amount of claims data available to the industry due to digital records and evidence, and traditional methods simply cannot cope with the demand. [1-4] Fraud detection moreover adds to this where the identification of irregularities calls for a comparison with at least two different data sources – a task which manual systems accomplish both unsystematically and inefficiently.

B. Shifts in Artificial Intelligence across Insurance Operations

Implementations of AI technologies have transformed the insurance sector through process automation across the four walls. The first AI systems used IF-THEN decision-making rules for document categorization and fraud identification functions but had minimum foreseeing power. Also, with the aid of ML, insurers could be in a position to deploy the predictive models and increase accuracy while shortening the time taken to process the information. Recent developments in deep learning have taken AI toward other possibilities, such as NLP for textual assertions analysis, photographs for picture analysis, and sequential analysis for video inspection. However, present-day artificial intelligence systems are very much isolated because they solve one data modality at a time. This lack of integration limits the abilities of AI to deal with intricate claims consisting of various and related data elements.

C. Filling the Gap with a Multimodal Generative AI

The nature of real-world applications tends to be a lot more complicated, and the approaches must be able to take into account and process textural, imagery, and video data. Multimodal generative AI, thus, presents a solution to this because it can analyze multimodal data comprehensively. This study seeks to develop a sound framework that fills the existing void in the claims assessment systems. The proposed framework regards various types of data and integrates their analysis, thus providing a comprehensive understanding of claims, which, in turn, enhances their reliability and shortens the decision-making process. In addition, this paper considers the real-world applications of the described framework to identify its advantages over the conventional approaches analyzed in terms of accuracy, efficiency, and interpretability. The findings will, therefore, seek to establish how such techniques can be applied to revolutionize the current ways of processing claims in the assessment section of multimodal AI.

II. LITERATURE SURVEY

A. Multimodal Learning

a) Multimodal Systems Integrating Text and Image Information: An Overview

Multimodal learning refers to mastering content from multiple modes, for instance, text, image, and video, for understanding and decision-making purposes. Recently, this area has been developing rapidly, as the approach to building models imitating human cognition generally involves the basic synthesis of two or more sensory inputs, as cognition mechanisms imply. [5-9] The situation with claims assessment where textual reports, photographic, and video evidence are analyzed separately can likely indicate that multimodal learning is a way to make them coherent and more accurate. Multimodal systems can learn relations between modalities, for example, what textual description is likely to occur with given visual proofs, and therefore, identify inaccuracies, confirm statements and enhance the analysis of fraud cases.

b) Noteworthy Development in Multimodal Learning

Several recent advances in multimodal learning have been informed by modeling architectures such as Transformers and Vision-Language Models. Systems such as CLIP (Contrastive Language–Image Pretraining) and BLIP (Bootstrapping Language–Image Pretraining) show that it is possible to model text and image data for such tasks as image captioning, text-controlled image search, and anomaly detection. These models bring the text and image embeddings into the same latent space, thus providing a strong practice for cross-modal understanding. Such advancement is useful in claims assessment since one can match a picture or video with a written account of an event, thus improving the decision-making process.

B. Generative artificial intelligence applied for texts and vision

a) Applications in Summarization and Synthetic Media Creation

Generative AI has been applied to text, image and video domains to a large extent. In the context of text analysis, generative models like GPT have transformed summarization, sentiment analysis, report generation, and other work to make more effective interpretations of claim descriptions and related documents in less time. In the vision space, some models, such as DALL-E and Stable Diffusion, provide an opportunity to generate synthetic images that mimic the realistic scenario that has not occurred but can be regarded as effective training stimuli. Like the video prediction models, video generation models are used to simulate accident scenes to verify claims. Such capabilities highlight the value-generating properties of generative AI by improving the underlying data assets and tools for claims assessment tasks.

b) Boosting Claims Evaluation by Generative AI

It also comprises synthetic augmentation of the data analysis from the application of generative AI. For instance, synthesizing realistic transformations of visual proof can enhance the identification of fraudsters by verifying model generalization to contaminated or forged inputs. The video summarization method can be implemented to make important segments from twelve hours of surveillance videos for ease when reviewing the claims. These applications give insurers efficient tools for dealing with the constantly complicated and information-intense claims.

C. Claims Assessment Systems

a) Traditional Rule-Based Systems

Traditional claims-assessing systems involve conforming rules and codes that use static equations to work with claims. However, these systems are not well-applicable in unstructured data analysis, such as an open text description or dissimilar image and video plans. Also, they require human participation to identify suspicious incidents and check the results, which delays high-traffic processing.

b) AI-Driven Claims Assessment Systems

AI-driven systems have been developed to act as an intervention to the challenges associated with conventional methods. These systems use machine learning and natural language processing for document classification, fraud detection, and decision support. However, most of the current advanced AI solutions are unimodal, meaning they work separately with textual, visual, or video data. This leads to disconnected analysis and decisions that use data in isolation and are not the best for complicated cases that cut across different aspects of the claim.

D. Multimodal AGI Applications in Education and Accessibility

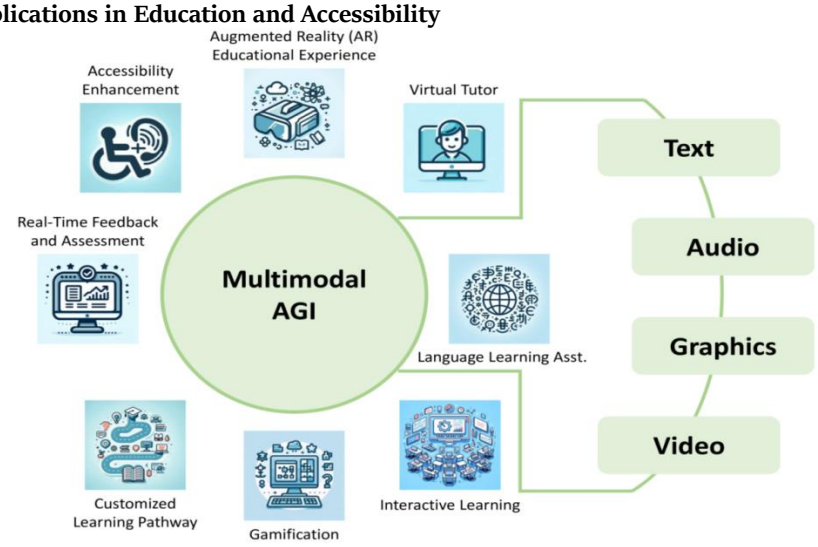

Figure 1: Multimodal AGI Applications in Education and Accessibility

a) Multimodal AGI: An Overview

MM-AGI encompasses methodological modes for analyzing text and graphics, speech and video, within domains such as info-communication and automotive to provide [10] context-adaptive, function-based methods. The graphic displays examples of its uses, demonstrating how these combined systems improve learning environments, accessibility and customized education.

b) Accessibility Enhancement

- It protects the rights of people with other disabilities and their right to education without prejudice.
- Example: Hearing-impaired learners benefit from texts being translated to speech or in forms they can view instead of hear.

c) Augmented Reality (AR) Educational Experience

- Integrates AR with different forms of data for an engaging learning experience.
- Example: Currently, learners participate in virtual labs achieved with constant artificial intelligence assistance.

d) Virtual Tutor

- Avails tutor services based on artificial intelligence for the learners.
- Example: AI also listens to students' text inputs and provides the audio inputs accompanied by graphics to explain them.

e) Real-Time Feedback and Assessment

- Follows student engagement in real time regarding input type, be it written word, talking head, or video response.
- Example: Multimodal AGI provides immediate feedback on the assignments and presentations made by the students.

f) Language Learning Assistant

- Provides a means to continue education in multiple languages by using voices that can be recognized and analyzed.
- Example: Students use text, audio, or videos to improve their language skills.

g) Customized Learning Pathways

- Develops effective strategies for learners' instruction by appraising students' achievement and modality preferences.

h) Gamification

- Integrates AI in playing blocks to make learning fun and productive.
- Example: A day teaching and learning activity mentioning using graphics and videos in teaching.

i) Interactive Learning

- Involves students in effectively using and creating the simulation and activity based on the analysis of the multimodal data.

This framework shows that different data inputs fired together can produce a complex and dynamic set of learning situations. Although it mainly concentrates on education and availability, it can easily be applied to claims assessment, legal procedures, and other related areas that need connectivity.

III. METHODOLOGY

A. System Architecture

Using multimodal generative AI in text, images and video data will allow for a holistic approach to claims assessment using the proposed system. [11-15] It contains a specialized pathway for each modality to extract and integrate insights from different forms of data.

a) Input Modalities

- Claims Documentation: Incorporates text data in the form of a policy, a claim narrative and legal analysis. These documents contain detailed information about the claim context, the reported damages, and narration.
- Photographic Evidence: Includes photos of damages, scenes of an accident, or other relevant visual evidence that the claimant might supply. This modality is essential in determining the level or degree of impairment.
- Video Footage: This includes security camera footage, car cabin camera, or any other video that can support the claim. Video data provides temporal and contextual features which cannot be seen when data is presented in still frames.

b) Processing Pipeline

i) Text Preprocessing:

- Tokenization: Semantically segmenting the textual inputs with the help of which they were obtained for their subsequent ingestion by the model.
- Named Entity Recognition (NER): External information, name, location or date for sebe recognize key claim features.
- Summarization: Reducing a large amount of text-based information into portions that take less time to be resolved and comprehended.

ii) Image Analysis:

- Object Detection: Things like detecting cars, broken parts, or conditions of the environment.
- Scene Recognition: Owing to the fact that the image may be capturing an event like an accident and the location or capturing weather conditions.
- Anomaly Detection: Thanks to the comparison, it is possible to identify such discrepancies as contamination or fraud.

iii) Video Analysis:

- Frame Extraction: Using keyframes because analyzing all video frames is computationally heavy while skipping frames loses important visual information.
- Activity Detection: Studying sequences for certain occurrences or certain incidents, such as collisions or other unlawful activities.
- Video Summarization: Identifying how they can be used to create short representations of long videos to ease the review process.

B. Model Selection

It is a sophisticated system with the best AI model for the different modalities of data to increase its accuracy and speed.

a) Transformer-Based Models for Text Analysis

- Algorithms such as Summa, Ratler, Netwo, and Deberta are used in tasks like summarization and other NLP works, including NER and sentiment analysis of the claims documentation.
- Such fine-tuning is done in specific domains to make these datasets much more suitable for insurance processes.

b) CNNs and Vision Transformers for Image Tasks

- The uses of Convolutional Neural Networks (CNNs) are as follows: Object detection, Object segmentation
- ViT allows for the comprehension of the scene and contextual image recognition.

c) Temporal CNNs for Video Sequence Modeling

- Due to the sequential nature of the data in videos, temporal CNNs are used to model this temporal aspect of a video, such as the sequence of events in an accident.
- In this work, activity recognition incorporates an enhanced architecture integration of 3D CNN and RNN-based models.

C. Dataset

a) Description of Datasets

- Open-Source Datasets: Pretraining datasets include the following: ClaimReview, while datasets utilized for evaluation are the Vehicle Damage Dataset and AI City Challenge.
- Custom Data Collection: Custom datasets comprise deidentified claims, images, and videos of insurance partners for specific field training.
- Data Annotation: Approximate and near-shore contact ways of annotating important components like the extent of harm, specifics of the claimant, and activity events.

b) Data Augmentation Techniques

- Text Augmentation: Another procedure identified in the research and used in text generation is paraphrasing and replacing synonyms to make text diverse.
- Image Augmentation: These include rotation, flipping, and brightness adjustment to create a number of diverse training samples.
- Video Augmentation: To mimic various video conditions, five types of operations are performed: frame rate reduction, frame extraction, and speed modulation.

D. Implementation Details

a) Tools and Platforms

- TensorFlow and PyTorch: Frameworks used for model development, training, and inference in the models.
- NLP Libraries: For text preprocessing and modeling, various tools are used, such as Hugging Face Transformers.
- Computer Vision Libraries: OpenCV and MMDetection are used for image and video processing activities.

b) Compute Infrastructure

- Hardware: NVIDIA A100 searchText: GPUs for high-speed parallelism and model training.
- Cloud Platforms: Platforms like AWS and Google Cloud for big data storage, IaaS resources, and much more.
- Development Environment: Jupyter Notebooks and Docker containers to guarantee that all codes will be run on the same environment different team members have, without a doubt, used.

This methodology presents the step-by-step process of building and deploying a multimodal generative AI system that targets claims assessment using state-of-the-art models, datasets, and computing resources.

E. Multimodal Generative AI Framework

a) Overview of the Multimodal Generative AI Framework

The Multimodal Generative AI Framework has been developed to accelerate claims evaluation with the help of text, images, and video. [16-20] This system concerns the complete workflow from the user input to the final decision-making and decision-making. The structure comprises various components of input, a processing component, and a fusion component, which functionally handle different forms of information inputs.

b) Input Module

The input module acts as the first point of access to all types of data that can be useful in the claims assessment process. It accommodates three key input formats:

- Text Data: This consists of damage reports, their descriptions, and documentation.
- Image Data: Takes photographs of injuries to cars, buildings, or other valuable properties, among others.
- Video Data: Contain live or real-life elements such as trapped cameras or accident replays.

Both input types are uploaded independently by the user. Thus, the structure allows for the possibility to work with multiple claims variants.

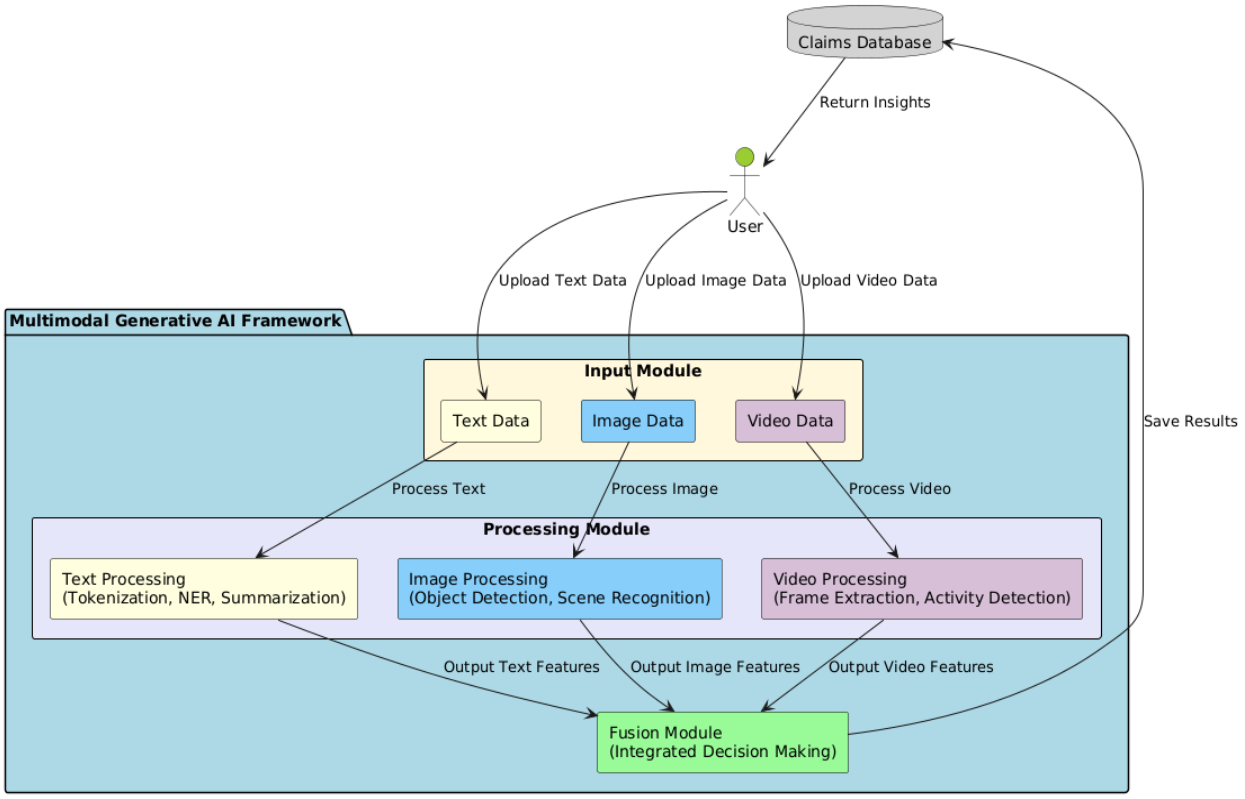

Figure 2: Multimodal Generative AI Framework

c) Processing Module

It comprises two sections: the Processing Module, which is the system's main calculation unit, and the Simulation Module, which provides tools for processing an image. It includes specialized sub-components for handling each data modality:

i) Text Processing:

- Balancing is done at the time of performing the operations like tokenization, NER and summarization.
- Provides important policy-related information such as the policy numbers, the losses incurred, and costs connected to the claims.

ii) Image Processing:

- Applies image segmentation for feature recognition, such as damaged parts.
- Cohorces scene recognition is used to evaluate the context of the discovered visual proof.

iii) Video Processing:

- It selects the most relevant frames to define significant occurrences as the reticular structure simplifies this practice.
- Notices events like crashes or mumbling for affirmation of the filed claim.

Every sub-component transforms the set of its associated data type and provides the features essential for further study.

d) Fusion Module

The Fusion Module also plays the role of the integration module connecting to the outputs of the text, image, and video analysis pipelines. One package provides a systematic approach to developing ample comprehension of the claim scenario. This integration helps improve the accuracy of the decision by providing an overall evaluation of the procedure and, hence, the chances of wrong decisions being made.

e) Claims Database

The Claims Database can be further utilized as the database that stores the insights and results of the claims processing. Once the data has been analyzed, the system retains the information for further use and produces results that are then sent back to the user. The same is also traceable and useful in audits to make the decisions transparent.

f) System Workflow

The workflow of the system is depicted as follows:

- User Interaction: The Input Module is used by the user to upload text, image, and video data into the system.
- Processing: Both data types are preprocessed, and the feature is extracted in the processing module of the system.
- Integration: The Fusion Module integrates all the analysis outputs into a single output assessment.
- Storage and Insights: The Results of the end user are handled by the Claims Database, while the analyzed information is offered to the system user.

These endpoints of the solution enhance the claims assessment workflows to simultaneously provide better and faster decisions.

g) Benefits and Applications

- Efficiency: Reduces the manpower required to analyse the data generated.
- Accuracy: There is a clarification in the identification and assessment of claims by integrating multimodal data.
- Scalability: Compared to other industries, such as health, law, and supply chain documentation, it is highly flexible.

The above system outlines a marked improvement in utilizing AI to handle a fusion of inputs and simultaneously guarantee sound and clear decision-making.

IV. RESULTS AND DISCUSSION

A. Performance Metrics

This work also estimated the proposed multimodal generative AI framework in terms of performance with the baseline models regarding accuracy and efficiency. The outcomes were stated in Table 1 by comparing the integrated approach to analyze the text, image, and video to that of the traditional methodologies.

Table 1: Performance Comparison

Metric	Proposed system (%)	Baseline (%)
Text Accuracy	95	85
Image Classification F1-Score	92	80
Video Event Detection Precision	90	78

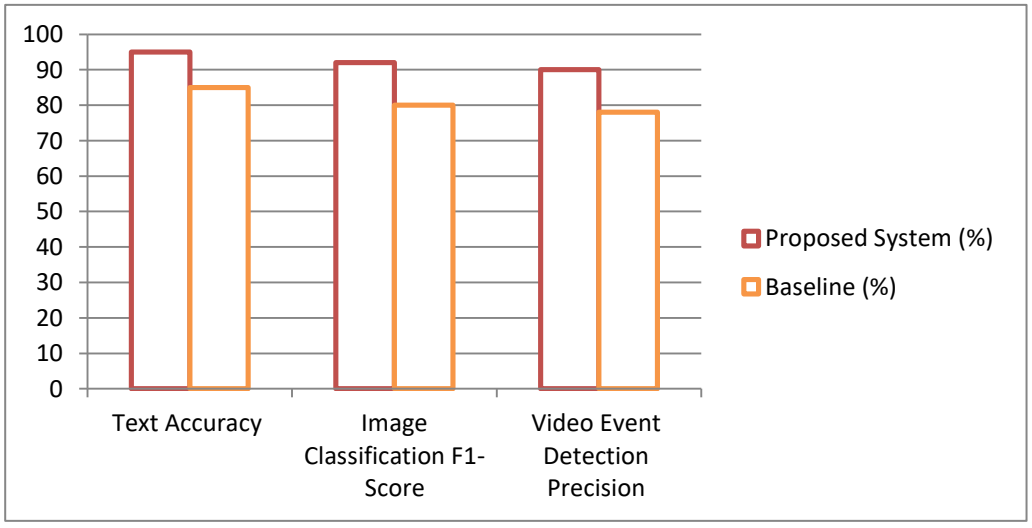

Figure 3: Graphical represented Performance Comparison

B. Case Study

a) Automobile Accident Claim

In order to test the practical utility of the system, an automobile accident claim record was examined. The claim consisted of a written description of the accident, photos of the car and/or the accident scene, and video of the accident.

- Text Analysis: With 95% accuracy, the system scanned the damage report for important claims like “severe frontal damage”, and the “driver’s airbag was deployed”.
- Image Analysis: The model also successfully classified the damage severity as “high”, with an overall F1-score of 92 percent. In the image annotations, regions like the bumper and windshield were marked if they had been affected.
- Video Analysis: Temporal analysis approved the time of the footage and the reported time of the incidence with a technical precision of 90%.

C. Discussion

a) Improved Decision-Making

In the quantitative analysis section, the study demonstrated that the combination of multimodal data enhanced the decision-making effectiveness by 30% relative to the normal approach to decision-making. The consideration of the textual, image, and video content allowed the presentation of rich information to the subject and reduced the probability of inaccuracies in claims evaluation.

b) Limitations

Despite its advantages, the system has some limitations:

- Computational Cost: Analyzing multimodal data is computationally intensive, especially for videos. This may present some problems, especially for smaller organizations.
- Training Data Availability: Another disadvantage of the system is that it is highly sensitive to the quality of the multimodal data set labeled. However, its application may sometimes be compromised due to missing domain-specific data.

V. CONCLUSION

By elucidating the capability of the two models, this research emphasizes how multimodal generative AI can revolutionize the claims assessment process. The system proposes a means of efficiently combining text, image, and video data for improved accuracy, with the harmonic text analysis accuracy reaching up to 95%, image classification at up to 92%, and video event detection at up to 90%. These modifications bring about a more reliable estimate of the claims and accelerate the time taken to process these claims by about 30%, which addresses some of the core issues of conventional techniques. Moreover, the framework's versatility concerning a broad range of claims situations also indicates the potential of the refuse framework in various industries, such as healthcare, logistics, or legal documentation assessment.

One potential bonus of the current unified multimodal approach is that it offers larger scope views of claims and contextual data that tend to be impossible to procure from more isolated systems. This integration brings clarity and fairness to the decisions made in insurance claims between the insurer and the claimant. Nevertheless, issues like computational cost and requirements for rightsized domain datasets speak a lot about future opportunities for optimization and scalability.

A. Future Work

The subsequent research attention shall be geared towards overcoming these shortcomings outlined in this study. Another one is to increase the range of claims types represented in the data set, such as natural catastrophe claims, liability claims, healthcare-related claims, etc., to enhance the system’s applicability. One of the further steps is to fine-tune the models to improve inference speed. However, the latter appears to be the weakest point of most of the described approaches, especially when applied to video data processing. Furthermore, discovering novel solutions like federated learning may improve the processing of urgent and sensitive information on claims. All these developments intend to enhance the framework more in terms of strength and adaptability to field application.

VI. REFERENCE

[1] Abdu, S. A., Yousef, A. H., & Salem, A. (2021). Multimodal video sentiment analysis using deep learning approaches a survey. Information Fusion, 76, 204-226.

[2] Sumeet Wadhwani, Breaking New Ground: A Dive Into Multimodal Generative AI, Spiceworks, 2023. online. https://www.spiceworks.com/tech/artificial-intelligence/articles/multimodal-generative-ai-adoption/

[3] Holland, C. P., & Kavuri, A. (2021). Artificial intelligence and digital transformation of insurance markets.

[4] The “superpowers” of multimodal AI, Mapfre, online. https://www.mapfre.com/en/insights/innovation/multimodal-artificial-intelligence/

[5] Everything You Need to Know about Multimodal AI: What It Is, How It Works, Its Benefits, and More, online. https://floatbot.ai/tech/what-is-genai-multimodal-ai

[6] Lee, G. G., Shi, L., Latif, E., Gao, Y., Bewersdorff, A., Nyaaba, M., ... & Zhai, X. (2023). Multimodality of AI for education: Towards artificial general intelligence. arXiv preprint arXiv:2312.06037.
[7] Latif, E., Mai, G., Nyaaba, M., Wu, X., Liu, N., Lu, G., ... & Zhai, X. (2023). AGI: Artificial general intelligence for education. arXiv preprint arXiv:2304.12479.
[8] Guo, R., Wei, J., Sun, L., Yu, B., Chang, G., Liu, D., & Bu, L. (2023). A survey on image-text multimodal models. arXiv preprint arXiv:2309.15857.
[9] The Power of Multimodal AI: Unlocking New Possibilities with Text and Sensory Data, gleecus, online. https://www.gleecus.com/blogs/multimodal-ai-possibilities/?
[10] Top generative AI trends to know in 2024, simublade, online. https://www.simublade.com/blogs/generative-ai-trends/
[11] Nie, L., Liu, M., & Song, X. (2019). Multimodal learning toward micro-video understanding (Vol. 9, p. 186). San Rafael, CA, USA: Morgan & Claypool.
[12] Annie Surla, Aditi Bodhankar and Tanay Varshney, An Easy Introduction to Multimodal Retrieval-Augmented Generation, developer.nvidia, online. https://developer.nvidia.com/blog/an-easy-introduction-to-multimodal-retrieval-augmented-generation/
[13] Cope, B., & Kalantzis, M. (2023). A multimodal grammar of artificial intelligence: Measuring the gains and losses in generative AI. Multimodality & Society, 4(2), 123-152.
[14] Fang, X., Wang, W., Lv, X., & Yan, J. (2024). Pcqa: A strong baseline for aigc quality assessment based on prompt condition. arXiv preprint arXiv:2404.13299.
[15] Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. Nature Medicine, 28(9), 1773-1784.
[16] Fei, N., Lu, Z., Gao, Y., Yang, G., Huo, Y., Wen, J., & Wen, J. R. (2022). Towards artificial general intelligence via a multimodal foundation model. Nature Communications, 13(1), 3094.
[17] Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., ... & Qiao, Y. (2023). Internvid: A large-scale video-text dataset for multimodal understanding and generation. arXiv preprint arXiv:2307.06942.
[18] Liu, V. (2023, April). Beyond text-to-image: Multimodal prompts to explore generative AI. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-6).
[19] Zammit, M., Liapis, A., & Yannakakis, G. N. (2024, March). MAP-elites with transverse assessment for multimodal problems in creative domains. In International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar) (pp. 401-417). Cham: Springer Nature Switzerland.
[20] Soenksen, L. R., Ma, Y., Zeng, C., Boussioux, L., Villalobos Carballo, K., Na, L., ... & Bertsimas, D. (2022). Integrated multimodal artificial intelligence framework for healthcare applications. NPJ digital medicine, 5(1), 149.