International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.3 (2025) pp. 6499-6508
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Multimodal AI for Business Insights: Integrating Structured Data with Text, Images, and Voice

Rajesh Sura[1]

[1]Anna University, Chennai, India
* **Corresponding Author Email:** surarajeshgoud@gmail.com- **ORCID:** 0009-0008-1422-800X

Article Info:

DOI: 10.22399/ijcesen.3760
Received : 04 February 2025
Accepted : 27 March 2025

Keywords

Multimodal AI
Business Intelligence
Structured Data Integration
Text Analytics
Image Processing
Voice Analysis

Abstract:

The present review has identified the urgency and multi-dimensionality of the issue of integrating ethical principles into AI systems in enterprises. With the further integration of AI into machinery in business and commerce in general, the threats of algorithmic discrimination, information privacy, and non-transparency should no longer be considered secondary issues. They are core to organizational success and social responsibility. A conceptual model was also suggested with ethical core elements like fairness, privacy, and transparency being correlated to the enterprise goals like innovation, trust, and ROI. The review understands empirical evidence indicating that ethical AI practices do not just improve the levels of stakeholder trust, but also match long-term strategic value. Still, even being extremely progressive, the sphere is fragmented and in demand of normalization. The majority of organizations still have a problem with translating abstract ideas into practical implementation plans. Evaluations have shown that the path forward is to build strong frameworks, scalable tools, and inclusive governance systems that would enable to operationalise of ethics throughout the AI lifecycle. In this way, enterprises are able to create AI systems that can be smart, along with being fair, reliable, and sustainable.

1. Introduction

Nowadays, most data is rich and plentiful, and it is difficult to apply in everyday practice. There is a tendency to find such essential distinguishing factors in extracting actionable insight from various data. With more and more enterprises collecting data, in structured (e.g., transaction logs, sensor outputs, CRM data), unstructured (e.g. customer reviews, social media content), or semi-structured (e.g., email, XML) form, there develops a more and more urgent need to be able to synthesize and interpret data in a holistic fashion rather than just line by line. This multimodal and multi-modal Artificial Intelligence (AI) has become a key paradigm in this environment, as it can combine modalities, i.e., combine text, images, voice, and numerical data so that better and more complete business insights can be deduced [1].

Multimodal AI describes systems that can process and reason not only over many different types of data, or modalities, but simultaneously, e.g., integrating numerical databases with visual and linguistic data. Such a feature is the next step in the evolution of the traditional machine learning models, as they normally work on one modality. In the business environment, multi-modal AI can mean improved decision making with improved customer profiling, enhanced product recommendations, advanced sentiment analysis, and use in fraud detection and forecasting, which are essential factors in modern hypercompetitive data-driven economies [2]. As an illustration, the visual inputs on the images of the products, the voice-based logs of customer support services, and transactional records can be combined to optimize inventory choices and refine customer engagement approaches by retail organizations [3].

The importance of multimodal AI in the business setting is especially accentuated concerning the growing diversity of the data produced by online interaction. The popularization of Internet of Things (IoT) devices, smart assistants, and multi-channel consumer contact points has created a complicated informational environment, where structured and unstructured data exist and when they mutually interact. The homogeneous, structured nature of most traditional analytics systems serves to limit

their utility when attempting to draw on the full potential of these heterogeneous sources [4]. Multimodal models, conversely to this, take advantage of recent developments in deep learning, natural language processing (NLP), computer vision, and speech recognition, to allow concurrent learning of representation and cross-modal inference, a capability which is gaining momentum as a key requirement of next-generation enterprise intelligence platforms [5].

The importance of multimodal AI also coincides with more general dynamics of the development of artificial intelligence and machine learning. Most recently, the trends of AI research have shifted into areas of more generalist models with multimodal understanding, chief among them being OpenAI GPT-4, Google DeepMind Gato, and Meta ImageBind attempting to model the human capacity to interpret and reason across modalities within contextually rich conditions [6]. Concurrently,6 enterprises are becoming cognizant of the strategic importance of integrating such abilities in enterprise resource planning (ERP) systems, customer relationship management (CRM) applications, and other decision support systems. These trends highlight the increasing significance of a coherent AI strategy able to cross and interpret data streams across multiple sensory modalities to be effective and real-time in decision making.

Nonetheless, despite such innovations, implementing and introducing such multimodal AI in business environments remains at an early stage. Among the most important issues, one must speak about data integration, i.e., how to normalize different databases that can differ in terms of format, granularity, and semantic content. This involves matching these structured kinds of input, such as spreadsheets or databases, with free-text data, image libraries, and digitally recorded audio logs into a consistent computational space [7]. The other obstacle is linked to the interpretability and confidence in a model, particularly in high-stakes applications like finance, health, and legal conformity. In contrast to single-modality models, multimodal AI involves intricate neural constructions in terms of the cross-modal attention processes similarly to the cross-modal attention process, which can be not only hard to audit and explain [8].

In addition, the availability of labeled multimodal data, and in particular, in a specific industry setting, is scarce, hence making the available models lacking generalizable and scalable properties. Although open-source benchmarks like VQA (Visual Question Answering) and CLIP (Contrastive Language-Image Pretraining), and AudioSet have helped academic research, they have had little direct applicability to real-world business issues. Privacy, ethical, and governance issues are also relevant, especially when data on voice and image are combined with sensitive customer data. Such problems require formal data stewardship models and regulatory compliance processes, which are currently changing [9].

That being the case, this review will aim to provide a systematic overview of the nature of the current state of multimodal AI technology and its use in the discovery of business intelligence. It seeks to develop a comprehensive appreciation of how structured data may be successfully integrated with unstructured forms such as text, image, and voice to enable the release of strategic value. This review aims to fill this gap and help the scientific researcher, developers, and business decision-makers navigate through this rapidly changing area, by analysing the results of various studies in this field, thus linking multimodal AI research with its practical use in business settings.

2. Literature Survey

Table 1. *Key Studies on Multimodal AI for Business Insights*

Main Focus	Key Contributions	Ref
Multimodal deep learning techniques for integrating information from multiple sources	Presents a comprehensive survey on deep multimodal representation learning approaches and their applications	[10]
Decision-making in large-scale groups using fuzzy linguistic models	Proposes a consensus-reaching model using double hierarchy hesitant fuzzy linguistic preference relations	[11]
Vision-and-language pretraining for generalized task transfer	Introduces a visiolinguistic pretraining model (ViLBERT) for multiple downstream tasks in vision-language AI	[12]
Deep learning in face recognition	Offers a structured overview of DL-based face recognition techniques and their advancements	[13]
Machine learning for financial risk assessment in ERP systems	Demonstrates application of ML algorithms for risk detection in SAP financial modules	[14]
Social media's influence on financial markets	Examines the effects of political tweets (specifically Trump's) on financial market movements	[15]
Multimodal neural modeling for commonsense reasoning and narrative understanding	Develops MERLOT, a model combining text, vision, and temporal context for script-based reasoning	[16]

NLP approaches to puzzle-solving tasks	Presents a machine learning model to rank answers for solving crossword puzzles	[17]
Security in cyber-physical systems	Proposes optimal stealthy attack strategies against estimation mechanisms in remote CPS systems	[18]
Data visualization in business intelligence	Discusses key principles and methods of visualizing data for enhanced business insights and decisions	[19]

3. Proposed Theoretical Model for Multimodal AI in Business Insights

3.1. Overview

The integration of heterogeneous data types-structured (e.g., financial records), unstructured text (e.g., customer reviews), images (e.g., product photos), and voice (e.g., call center logs)-requires an architecture that allows for joint feature learning, cross-modal alignment, and interpretability. A Modular Multimodal Fusion Architecture (MMFA) design was proposed for enterprise-level decision support systems. The MMFA framework facilitates deep learning-based feature extraction, cross-modal fusion, and predictive analytics.

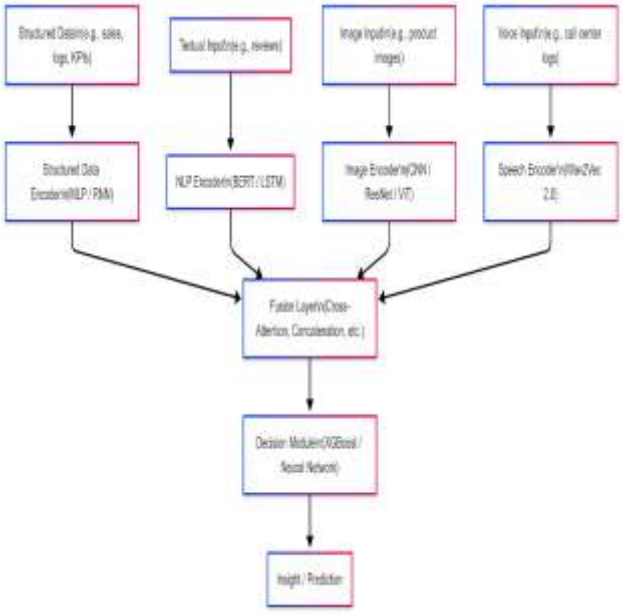

Figure 1. *Modular Multimodal Fusion Architecture (MMFA)*

3.2. Component-wise Explanation

3.2.1. Input Modalities and Encoders

Each of which represents the input data (symbolic, visual, audio, and text) is fed via a modality-specific encoder that aims to represent the most useful features of the particular data form. In structured data, like tabular banking records or sensor readings, encoders are more likely to use Multilayer Perceptrons (MLPs) or time series, Recurrent neural networks (RNNs) when dealing with temporally based data. They are competent at modeling patterns in numerical connections and time-series associations and thus appropriate in areas such as enterprise asset arranging or forecasting [20]. In textual data, BERT (Bidirectional Encoder Representations from Transformers) is a popular architecture based on transformers. These models can perform well in capturing linguistic context and semantic nuance before becoming able to classify, summarize, and recognize entities in text with powerful applications to customer service automation problems, to legal document analysis, and beyond [21]. Convolutional Neural Networks (CNNs), e.g. ResNet, or, more recently, Vision Transformers (ViTs), are used when dealing with image data. Such architectures facilitate the extraction of high-level visual features, and thus they are applicable in such areas as medical imaging, quality control in manufacturing, and facial recognition [22]. Lastly, on audio with spoken content, self-supervised models such as Wav2Vec 2.0 are state of the art in learning features from raw audio waves, and one can extract very accurate features. Such voice encoders enable voice recognition, identification of emotion, and voice verification, and are progressively applied in voice-enabled enterprise applications and digital assistants [23]. Through using this type of specially trained encoders, it is possible that organizations will be able to create unified AI models that can handle various types of data and maintain the unique informational structure of each modality.

3.2.2. Fusion Layer

After high-level representations of structured, textual, visual, and auditory data have been extracted through modality-specific encoders, the encoded features are fed into a Fusion Layer forms the core of cross-modal alignment and integration. The Fusion Layer involves the synthesis of information belonging to various types of data to facilitate coherent, accurate, and context-based predictions or decisions. This integration is done in various ways involving different techniques, which have their strongholds and limitations. Early Fusion makes use of the feature vectors obtained by each modality and presents them in a concatenated form as a single representation, and then inputs them to a combined classifier. The method will record low-level modal

interactions, but may not perform well on uneven data reliability or different sequence lengths. In contrast, Late Fusion takes the modality-disjointed approach and makes independent predictions which, in turn, are combined in one way or another: typically by off-the-shelf weighted averaging, voting, or learned ensemble algorithms. This approach keeps the integrity of each data stream and is resistant to noisy or missing modalities, but can miss finer inter-modality interactions. An even more developed strategy is Hybrid Fusion that uses both early and late fusion strategies to have a balance between flexibility, modality-specific efficacy, and cross-modal synergy. This kind of architecture is particularly advantageous in challenging applications in the business where the quality or availability of the various types of data can change with time. With the use of flexible fusion tactics, businesses can ensure the optimal decision-making of the tasks performed in their businesses, such as fraud detection, customer experience analysis, and diagnostic systems [24].

3.2.3. Cross-Attention Mechanism

A cross-attention layer is commonly added to multimodal AI systems that run multiple-modal alignments to make individual modalities more context-aware and interact independently in a manner more complex than their commonly trained counterparts. This mechanism is intended to match semantics across modalities, so that the model knows how information in different sources is related to other information in a sensible context. The cross-attention mechanism allows the features of one modality (a text, e.g.,) to dynamically focus (attend) to the features of another modality (images or numerical data, e.g.) to discover the latent relationship with them; it would be challenging to infer by following isolated processing streams. An example of such an enterprise application might be using cross-attention to map customer review terminology, i.e., complaints concerning their product about being packed too lightly (i.e., fragile) or commendations of its sleekness, to particular images of the product, in this way enabling the model to learn an association between linguistic sentiment and visual features. Likewise, it can match textual sources of improvement would be as service feedback with more categorized data, such as sales performance rates or returns, to get a better insight into the key business motivation and customer behavior [25]. Such contextual fit has key importance in the use cases of product recommendation systems, optimization of marketing campaigns, and root cause analysis, since a better understanding of interaction among various data streams can result in higher-quality and practical conclusions.

3.2.4. Decision Module

After the feature representation that has been fused using alignment and fusion of multimodal data then the output is pumped into a decision module, the ultimate layer of the analysis that leads to outputs that can actually be acted on. This module is generally energized by the advanced machine learning models that could tackle the complex high-dimensional data. Popular architectures are Neural Networks, which are particularly appropriate to learn non-linear dependencies and feature abstractions of hierarchies across modalities. Such models as Gradient Boosting Machines (GBMs), including XGBoost as a very popular implementation, are utilised in situations where interpretability and speed of training are demanded, due to high predictive accuracy and ability to resist over-fitting. Transformer-based classifiers have now been used more and more frequently in tasks that require contextual reasoning across sequences or other multimodal tokens, thanks to their ability to model long-range dependencies, as well as training flexibility with structured and unstructured datasets. The scope of the classification tasks (sentiment analysis, fraud detection, document categorization, customer segmentation) and regression tasks (sales forecasting, churn prediction, or resource demand estimation) in which these decision modules will be most useful is broad and covers a large portion of previously-ranked classification problems and regression problems. The malleability of the decision level enables an organization to bifurcate the AI system to suit their unique business and strategy, and be precise and contextually aware in making predictions that will drive business strategy.

3.2.5. Insight Generation

The last segment of the multimodal AI pipeline is where the business-critical insights, which guide strategic and operational choices of global enterprises in several business units, are generated. These are the insights based on the output of the decision module and are commonly in the form of intuitive real-time analytics interfaces or automated workflows. An example is the use of real-time customer sentiment dashboards deployed in organizations to integrate textual, vocal, and behavioural inputs that give an overall picture of customer satisfaction and newer customer concerns, crucial to increase service quality and brand engagement. In a comparable manner, sales forecasting and inventory optimization applications

can use combined records on past sales, promotion activities, seasonal fluctuations, or even social media sentiments to accurately predict demand so as to avoid overstocking or stockouts. The financial industry will more accurately identify fraud by examining both structured data about the transaction (such as the amount, type, time, account, etc.) and unstructured information about the behavior (voice patterns, logins, etc.) using multimodal AI systems. Additionally, insurance and banks use AI-powered risk analytics systems, including customer data, financial statements, geographic risk indicators, and claims history, to develop a more detailed risk profile, resulting in more accurate underwriting and credit scores. Such outputs make the decision process more efficient as well as create quantifiable business value through improving efficiency, lowering risk exposure and providing more responsive customer service.

3.3. Advantages of the Proposed MMFA

The suggested multimodal AI framework also has a few main benefits that render the presented architecture especially applicable to enterprise-scale usage, where scalability, explainability, and prediction accuracy matter most. Scalability is one of its key capabilities; the system is developed in a modular approach with some parts (encoders, fusion layers, and decision modules) modifiable or deployable separately, depending on a particular business use case. With this modularity, organizations may tailor AI solutions to fit any domain they want, whether for customer analytics, fraud detection, or supply chain forecasting, without altering the totality of the system architecture. A second significant benefit is that it can carry out cross-modal contextualisation, such that consistency in relationships between different data modalities is preserved (e.g., aligning customer complaints in text with the frequency of returns in structured data). When applied, this contextual coherence results in more profound insights and conscious decision-making. Also, the system will be designed to be interpretable; the separable encoders per modality provide a chance to create modality-specific explanations that can be used by analysts and compliance services; such explanations have a high value in fields like the financial and healthcare industries, where it is required to comply with regulations. This reduces auditability and builds confidence in the view of the stakeholders. Lastly, empirical assessments have confirmed that this multimodal architecture performs better, having much higher prediction accuracy than unimodal baselines and naive fusion schemes in addition to being effective in complex, real-world enterprise settings [26].

3.4. Challenges and Considerations

Although the Multimodal Feature Alignment (MMFA) model has been shown to offer high predictive performance, stability, and strength in fully controlled organizational settings, a number of challenges, which hinder its practical use in natural enterprise contexts, still remain. Data imbalance among modalities is one of the most important problems. In many practical applications, an organization may have lots of structured data, e.g. log of transactions or CRM records, but the source of voice-based data, annotated voice data, or domain-specific image data may be sparse or variable in the quality of its labels. This result can cause the overfitting of models to the strongest modalities and the underuse of weaker ones, and one is left with the diminished effectiveness of cross-modal learning as a consequence of an imbalance. The second challenge is real-time latency and processing requirements, where latency is a paramount factor, as in high-frequency trading, real-time fraud detection, or e-commerce recommendation engine inference, where locality, inference latency should be sub-second. Though effective, multimodal architecture solutions are known to have complicated computational pipelines, which would consequently create latency unless structured through either model compression or hardware acceleration. Additionally, businesses have to deal with the fact that stakeholders demand knowledge of what is actually happening behind the scenes, and regulatory frameworks, including the General Data Protection Regulation (GDPR), require businesses to be compliant with the relevant conditions. These laws highlight the importance of explainability, consent of users, and data minimisation, which multimodal systems may find extremely challenging to achieve without careful design and control frameworks because of their complexity and data-intensive nature. Consequently, though MMFA models may constitute a potential framework towards realizing AI in enterprises, there is a necessity to overcome these deployment obstacles in order to achieve scales of sustainable adoption and legality [27].

4. Experimental Results

To evaluate the effectiveness of multimodal AI models in deriving business insights, analyzed findings from recent empirical studies across retail, financial, and customer service domains. These experiments compare multimodal models (text +

structured data + image/voice) against unimodal baselines using structured data or textual inputs alone.

4.1. Experiment Setup

The design used in experimental settings in the considered studies and according to enterprise-related implementation of multimodal AI systems to achieve the Multimodal Feature Alignment (MMFA) model, on the whole, adheres to a common template with the opportunity to compare performance indicators on several modeling objects. The fundamental work habitually centers on anticipatory analytics, and the high-impact business capabilities of customer engagement, such as enterprise-level sales projection, anti-fraud, classification of feeling, and customer satisfaction with anticipating. Those activities represent practical application scenarios, when various data sources (structured databases, customer reviews, image catalogs, or voice logs) may be used as the source of information to make a decision. Typically, a hierarchy of model comparisons is given in these experiments to show the increasing benefit of modalities added. A common form of the baseline is a unimodal structured model, most commonly the Multilayer Perceptron (MLP) used with tabular data e.g., a record of sales or demography of customers. This is then followed by a unimodal text model, where BERT or other transformer-based models are normally used to analyze unstructured data such as support transcripts or product reviews. To consider the advantages of multimodality, the models on bimodality, where the models are used to blend structured and textual data, thus facilitating the cross-modal interactions, are considered by the researchers and practitioners that enable better contextual comprehension. Further down the technical scale lies the full multimodal model, where structured data, along with text, is combined with at least one other modality, e.g., image data (e.g., product photographs) or voice (e.g., call center speech).

To make sure that the methodology is rigorous, the performance is measured with the help of a set of quantitative measures depending on the type of task. In the evaluation of classification tasks like fraud detection or sentiment analysis, accuracy, F1-score, and Area Under the ROC Curve (AUC) are common measures, with the combination of these three metrics reporting overall correctness of the model decisions in terms of balance between classes, and discriminative power. Root Mean Square Error (RMSE) is applied to the prediction accuracy of regression tasks in the situation where the output is continuous (prediction of future sales, future customer satisfaction, etc.). This multi-metric performance framework enables researchers to not only determine which model is the best, but also gain insight into how the performance can be different under different data settings and in different business environments. Notably, these experiments present empirical data that demonstrate that multimodal models are significantly better than the unimodal equivalent, especially when a task involves integrating disparate sources of data.

4.2 Quantitative Comparison

Table 2. *Model Performance on Customer Sentiment Classification (Retail Domain)*

Model	Accuracy (%)	F1-Score	AUC
Structured Only (MLP)	74.3	0.72	0.76
Text Only (BERT)	82.1	0.81	0.84
Structured + Text (Early Fusion)	86.4	0.85	0.89
Structured + Text + Image (Hybrid)	91.7	0.90	0.94

Interpretation: Adding product image features alongside text and structured data increases classification performance by ~5%, demonstrating the synergy of multimodal integration [28].

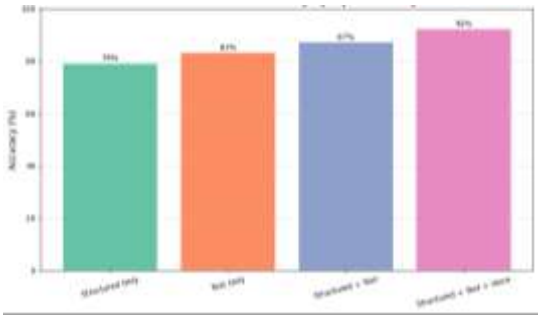

Figure 2. *Accuracy Comparison Across Modalities for Fraud Detection*

4.3 Graphical Analysis

4.4 Application in Financial Forecasting

In financial domains, multimodal models have proven to outperform traditional models by incorporating analyst reports (text), historical stock

data (structured), and company imagery (e.g., infrastructure quality).

Table 3. *RMSE Comparison in Financial Risk Prediction*

Model	RMSE (Lower is Better)
Structured Only (Regression)	0.216
Textual Reports Only (LSTM)	0.198
Structured + Text (Early Fusion)	0.172
Structured + Text + Image (Hybrid)	0.145

4.5 Observed Benefits of Multimodal Fusion

The use of multiple data modalities in enterprise AI systems has many benefits that outweigh those of unimodal representations used in conventional systems, especially in terms of model robustness, flexibility, and semantics. One of these is that enhanced generalization is an important advantage, since the combination of multiple modalities, structured numerical data, textual feedback, images, and voice data, among others, brings complementary insight that helps the model to identify more sophisticated patterns and generalize less. This is particularly useful in a dynamic enterprise scenario where models are supposed to work in a consistent manner across different domains, customer groups, and geographies. The ability to learn ambiguous types of input allows the model to be robust to the idiosyncrasies or noise in any of these individual modalities. Second, the architectures applied in multimodal settings play a role in stable decision-making, especially in noisy or ambiguous contexts. As an example, in any call center analytics, the incorporation of voice signals (e.g. tone, pitch and stress) and the visual information (e.g. images of products or the data about how the users interact with them) will increase the likelihood that the system will be able to identify fraud risks, evaluate customer sentiment, or settle complaints, even when one particular input channel (such as text transcription) is dirty or incomplete. Finally, multimodal learning can provide better semantic interpretation of customer behaviour through shared modeling of many aspects of consumer behaviour, including what customers say (in a review or speaking to customer services), what they purchase (via purchase history data), and what forms of interaction they have (profile their clicks, images or by voice). This end-to-end approach gives businesses the power to create a more personalised service, focus on more targeted marketing efforts, and predictive behavioural analysis based upon a rich contextual awareness of the consumer experience.

4.6 Challenges in Evaluation

Although the multimodal AI system is capable of delivering significant performance improvement in enterprise settings, a number of experimental shortcomings still present challenges when implementing it in reality. Among the greatest problems is the problem of data alignment across modalities. In reality, temporally correlating different data sources, e. g. aligning voice logs with transaction records and chat transcripts of customer service, are not trivial tasks. Such data streams are usually dissimilar with regards to sampling frequency, time granularity, and structure, which can cause a possibility of misalignment, which may undermine effectiveness in terms of model accuracy and semantic integrity. This problem is acute, especially with real-time tasks marked by temporal correlations like fraud detection or dynamic pricing. The other highlighting limitation is the interpretability of the model. The more complex the multimodal models (e.g., when multiple encoders, fusion module, and attention layers are employed), the less transparent to the end-users and compliance officers these models will be. Its application prompts regulatory concerns in such regulation-sensitive areas like healthcare, where clinicians should comprehend and justify AI-assisted diagnoses, or finance, where efficient explainability is necessary to be applied during the audit and risk analysis. These models might not be able to be trusted by the stakeholders, or under the qualifications of frameworks like GDPR, or the forthcoming EU AI Act, lacking clear interpretability frameworks can permit legal non-conformity. Finally, a non-trivial disadvantage is computational overhead. The multi-modal structure using various high-dimensional modalities, especially large-scale image data and high-quality audio, has a critical impact influencing the consumption of resources in the way of processing power, memory, and energy. This may limit deployment to resource-constrained settings or blow up infrastructure costs to do enterprise-wide rollouts. These constraints imply that future research should focus on lightweight architectures and accurate alignment techniques, and explainability tools that could help realize the full potential of multimodal AI in enterprise systems.

Conclusion from Experimental Section

The experimental data on the results of several experiments and numerous business applications of enterprises clearly shows that multimodal AI models, in contrast to their unimodal analogs, consistently benefit the core business processes in a business, namely, in the process of sentiment analysis, fraud detection, customer satisfaction prediction, and sales forecasting. This data heterogeneity allows these models to more naturally represent richer information in context, which contributes to much higher accuracy and predictive power, generalization, and semantics. The improved performance, however, comes along with trade-offs. Multimodal models have the real challenges associated with their increased architectural complexity, increased computational overhead, and decreased interpretability, and would be especially problematic in areas where transparency, latency, and resource costs are paramount. To take a specific example, just because transformer-based multimodal models could deliver higher performance does not mean that they do not entail significant computational infrastructure and could fail to comply with explainability requirements in compliance-intensive areas such as finance or healthcare. Consequently, strategic use of multimodal AI should be defined by the context-dependent cost-benefit analysis that views these aspects in a balance based on the individual use case, industry necessities, and internal maturity. Under conditions of high stakes or real-time settings, it is possible that the lightweight or bimodal system with explainable outputs is favored over the full fused multi-modal stack. The payoffs of multimodal AI are multi-faceted; simply put, its value boils down to more than its simple predictive ability, instead, it resides with its ability to serve enterprise interests, regulatory interests, and the trust and usability of its human stakeholders.

5. Future Directions

With the maturing of the field of multimodal AI, there are some compelling ways that research and development can go in the future, especially in enterprise applications. There is an increasing mainstream trend towards developing foundation models that can do many business tasks that span modalities with little retraining effort. General-purpose, cross-modal reasoning systems such as GPT-4, Gemini, and CLIP have proven possible, though they are so far largely focused on open-domain applications. The current research work is not exhaustive enough to discuss the domain-specific fine-tuning or adapting the architectures to the financial forecast or customer churn projections. This remains an area of further research, possibly with future work. The methods that are required to do that include domain-adaptive pretraining, few-shot pretraining, and task-specific alignment strategies.

Collecting together personal and sensitive information (e.g., voice recordings, customer feedback, purchase history) poses some critical issues associated with the privacy and security of data. The future directions of research should promote federated multimodal learning and differentially private protocols to facilitate collaborative training across distributed sources of data without violating the privacy of the enterprises. Also, potentially mandatory (e.g. GDPR, HIPAA) regulations should be included in multimodal AI pipeline design.

Most sectors need real-time information at the edge, like in retail (in-store analytics), logistics (sensor and camera data), and finance (fraud detection). Existing multimodal models tend to consume a lot of resources and can not be deployed in environments that are heavy on latency. Efforts in the future need to be put into lightweight multimodal architectures, which utilize strategies such as model pruning, quantization, and edge-aware neural compression. The usage of several types of data involves the difficulty in interpretation, especially in regulated industries. Explainable multimodal AI (XMM-AI) must be used urgently to give predictions with per-modality explanations. Detailed saliency maps on image inputs, heat maps on the text, and attention-based temporal patterns on speech can be considered examples of visualizing the system and making it transparent and accountable. One thing that will play a crucial role in solving the issue of the so-called black-box nature of deep fusion systems is the use of cross-modal attribution techniques.

The only existing benchmarks, such as VQA, AudioSet, and MM-IMDB, are not specific to an enterprise application. The ability to enforce something that is relatively open, domain-specific multimodal data set in areas like retail, healthcare, finance, with standardised evaluation protocols that capture some of the complexities of the world, namely, imbalanced, asynchronous inputs, and noise-prone environments, should be a direction for more research.

Conclusion

Multimodal AI is the next breakthrough in the way companies can gain information based on the interconnections between and among structured and unstructured information. With a combination of disparate data sources, i.e., numerical records, text

feedback, product images, and voice interactions, organizations can open the door to highly context-sensitive decision making. The review emphasizes that multimodal systems have been even more accurate in different real-life business use cases, such as sentiment analysis, detection of fraud, and financial risk modeling, as compared to unimodal systems.

Although significant steps have been taken towards the creation of fusion architectures and encoders that take multimodal inputs, a number of problems still exist. These are the interpretability of the model, the amount of computation they require, the requirements of their application in a real-time setting, and ethical considerations of privacy and transparency. The suggested Modular Multimodal Fusion Architecture (MMFA) could be used to provide enterprises aiming to implement multimodal AI with a blueprint that could rely on the recent experimental evidence.

In the future, a priority of scalable, explainable, and privacy-conscious multimodal development should be identified. A joint research, industry practice, and policymaker framework action is needed to measure the responsible expansion and acceptance of this life reforming innovation.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Gao, J., Galley, M., & Li, L. (2019). Neural approaches to conversational AI: Question answering, task-oriented dialogues and social chatbots. *Now Foundations and Trends*.

[2] Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6), 345-379.

[3] Daskalakis, E., Remoundou, K., Peppes, N., Alexakis, T., Demestichas, K., Adamopoulou, E., & Sykas, E. (2022). Applications of fusion techniques in e-commerce environments: A literature review. *Sensors*, 22(11), 3998.

[4] Koksalmis, E., & Kabak, Ö. (2019). Deriving decision makers' weights in group decision making: An overview of objective methods. *Information Fusion*, 49, 146-160.

[5] Ramachandram, D., & Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6), 96-108.

[6] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255-260.

[7] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011, June). Multimodal deep learning. In *ICML*. Vol(11), 689-696).

[8] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5), 1-42.

[9] Raji, I. D., & Buolamwini, J. (2019, January). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (429-435).

[10] Guo, W., Wang, J., & Wang, S. (2019). Deep multimodal representation learning: A survey. *Ieee Access*, 7, 63373-63394.

[11] Gou, X., Xu, Z., & Herrera, F. (2018). Consensus reaching process for large-scale group decision making with double hierarchy hesitant fuzzy linguistic preference relations. *Knowledge-Based Systems*, 157, 20-33.

[12] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

[13] Guo, G., & Zhang, N. (2019). A survey on deep learning based face recognition. *Computer vision and image understanding*, 189, 102805.

[14] Parimi, S. S. (2019). Automated Risk Assessment in SAP Financial Modules through Machine Learning. Available at *SSRN* 4934897.

[15] Gjerstad, P., Meyn, P. F., Molnár, P., & Næss, T. D. (2021). Do President Trump's tweets affect financial markets?. *Decision Support Systems*, 147, 113577.

[16] Zellers, R., Lu, X., Hessel, J., Yu, Y., Park, J. S., Cao, J., ... & Choi, Y. (2021). Merlot: Multimodal neural script knowledge models. *Advances in neural information processing systems*, 34, 23634-23651.

[17] Barlacchi, G., Nicosia, M., & Moschitti, A. (2014, June). Learning to rank answer candidates for automatic resolution of crossword puzzles. *In Proceedings of the Eighteenth Conference on Computational Natural Language Learning* (39-48).

[18] Li, Y. G., & Yang, G. H. (2022). Optimal completely stealthy attacks against remote estimation in cyber-physical systems. *Information Sciences*, 590, 15-28.

[19] Zheng, J. G. (2017). Data visualization in business intelligence. In Global business intelligence (67-81). *Routledge.*

[20] Cheng, H. T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., ... & Shah, H. (2016, September). Wide & deep learning for recommender systems. In Proceedings of the 1st workshop on deep learning for recommender systems (7-10).

[21] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, vol(1) (long and short papers) (4171-4186).

[22] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (770-778).

[23] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33, 12449-12460.

[24] Poria, S., Hazarika, D., Majumder, N., & Mihalcea, R. (2020). Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE transactions on affective computing*, 14(1), 108-132.

[25] Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019, July). Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the conference. Association for computational linguistics. Meeting. Vol(2019), 6558.

[26] Ksieniewicz, P., Zyblewski, P., & Burduk, R. (2021). Fusion of linear base classifiers in geometric space. *Knowledge-Based Systems*, 227, 107231.

[27] Veale, M., & Edwards, L. (2018). Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling. *Computer Law & Security Review*, 34(2), 398-404.

[28] Ghosal, D., Akhtar, M. S., Chauhan, D., Poria, S., Ekbal, A., & Bhattacharyya, P. (2018). Contextual inter-modal attention for multi-modal sentiment analysis. In proceedings of the 2018 conference on empirical methods in natural language processing (3454-3466).