Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 4, No. 3, 2025

Medical Entity-Driven Analysis of Insurance Claims Using a Multimodal Transformer Model

Xiaokai Wang

Santa Clara University, Santa Clara, USA

shawnxkwang@gmail.com

Abstract: This paper proposes a multimodal Transformer-based model for medical insurance claim adjudication, aiming to enhance the accuracy of claim decisions and improve risk control in the medical insurance domain. By integrating text data with structured data, the model can comprehensively analyze various data sources, such as customer information, medical records, and claim applications, to capture potential risks, particularly those related to fraudulent claims and overclaims. Using the Transformer architecture, the model leverages the self-attention mechanism to perform a weighted fusion of different data modalities, making information extraction more efficient and accurate. Experimental results show that the proposed model significantly outperforms traditional machine learning algorithms and deep learning models, such as XGBoost, random forests, and VGG16, in metrics like AUC, accuracy, and F1-Score, validating the advantages of multimodal learning in medical insurance claim adjudication. Additionally, the study explores hyperparameter tuning, examining the impact of factors such as learning rate and data modalities on model performance. Ultimately, the integration of multimodal data improves the accuracy of claim adjudication, offering insurance companies more scientific and reliable risk management tools. This research provides important theoretical foundations and practical guidance for the application of multimodal learning in the financial sector, particularly in the medical insurance industry, and lays the groundwork for future related studies.

Keywords: Multimodal learning, Transformer model, medical insurance, claims discrimination

1. Introduction

With the development of the social economy and the intensification of population aging, the importance of health insurance in modern society has become increasingly prominent. Especially in the context of rising health risks, health insurance, as a part of the social security system, is shouldering an increasing share of risk management responsibilities. In particular, the process of determining whether health insurance claims meet the criteria for reimbursement has become a huge challenge for insurance companies. Avoiding moral hazard, adverse selection, and other uncertainties has become crucial. Against this backdrop, the use of advanced artificial intelligence technologies, especially the Transformer model in deep learning, for insurance claim adjudication has become a hot research topic both in academia and the industry [1, 2]. By integrating multimodal data, this approach can enhance the accuracy of models' predictions and improve risk control capabilities, providing the insurance industry with more reliable tools for risk management [3].

The core issue in health insurance claim adjudication is how to accurately determine whether the claim meets the reimbursement conditions [4]. Traditional adjudication methods typically rely on manual review or rule-based systems [5]. These methods are often inefficient when handling complex and unstructured information and are vulnerable to human bias. With the development of big data technology, the insurance industry has gradually adopted machine learning and deep learning techniques to assist in claim adjudication.

However, effectively extracting valuable information from vast, complex, and diverse data remains a challenging problem. Especially when faced with various uncertain factors in health insurance claims data, quickly responding while ensuring accuracy and managing risks is still a research difficulty [6].

Multimodal learning, as an emerging AI technology, can integrate information from different data sources, overcoming the information scarcity problem of single-modal systems. In the context of health insurance claim adjudication, common modalities include clients' basic information, historical claims data, medical diagnoses, treatment plans, and medical expenses. Traditional claim adjudication models typically rely on one type of data modality, such as structured client information or medical expense data, often neglecting other potentially important information. This one-sided approach can lead to missing critical details and biases in the results. Multimodal Transformer models can, through deep learning, fuse these different modalities, forming a more comprehensive basis for judgment. This enhances the accuracy and efficiency of claim adjudication and helps identify potential risks more effectively [7].

Risk is a core factor that cannot be ignored in the health insurance industry. From the perspective of insurance companies, excessive claims, fraudulent claims, and other risks often put a huge financial burden on them. This not only affects their profitability but also threatens their long-term development. For policyholders, unfair or inaccurate claim adjudication may result in an inability to cover actual medical expenses, affecting their quality of life [8]. Therefore, balancing claim adjudication efficiency and accuracy, and controlling moral hazard and information asymmetry, has become a significant challenge in the insurance industry. By introducing multimodal Transformer-based models, insurance companies can consider various risk factors during the claim adjudication process, such as a client's historical claim

behavior, the reasonableness of medical expenses, and the authenticity of medical services. This improves the robustness of the model and reduces risks [9].

The advantage of multimodal Transformer models lies not only in their ability to efficiently process data from different sources but also in their self-attention mechanism, which allocates weights to different modalities of data. This enables better identification of key information in the data. In practical applications of insurance claim adjudication, such models can use deep learning techniques to automatically identify potential risk factors from a client's behavior, medical records, and claim history. For instance, by analyzing the client's past claim history, the model can reveal whether the client has a tendency for malicious claims. By analyzing the reasonableness of medical expenses, the model can detect if there are inflated costs. At the same time, unstructured information in medical reports can provide insight into the accuracy and rationality of the diagnosis and treatment. The fusion of multimodal information allows insurance companies to more comprehensively assess the risks in claim applications, leading to more informed and scientific decisions.

In summary, the research on health insurance claim adjudication based on multimodal Transformer models aims to enhance the accuracy of claim assessments by integrating multimodal data and applying deep learning techniques. This reduces human error, controls claim risks, and improves the efficiency and accuracy of insurance claim adjudication. This research has theoretical significance by offering new insights into the application of multimodal learning in the insurance industry, as well as practical importance by providing insurance companies with a more accurate and intelligent tool for claim adjudication, helping the industry better manage increasingly complex risk challenges.

2. Related Work

In recent years, with the widespread application of artificial intelligence in the financial industry, more and more research has focused on optimizing insurance claim adjudication using machine learning and deep learning methods. Early studies primarily concentrated on traditional models based on structured data, such as decision trees and random forests [10]. These models mainly relied on basic client information, medical expenses, and historical claim data to determine the reasonableness of claims. However, these approaches often failed to fully consider the importance of unstructured data, such as medical reports and diagnostic texts, leading to lower accuracy and robustness when dealing with complex health insurance data. With the rise of deep learning, an increasing number of studies have attempted to improve claim adjudication accuracy by using neural networks (such as convolutional neural networks and recurrent neural networks). However, these methods are generally limited to processing single-modal data, making it difficult to comprehensively analyze multi-source information [11].

Recently, as deep learning technology has advanced, multimodal learning has become a research hotspot. Multimodal learning refers to the integration of information from different data sources into a unified model. In the insurance claim field, researchers have begun to combine structured data with unstructured data to enhance model performance [12]. For example, some studies have proposed combining natural language processing (NLP) techniques with traditional classification models to analyze keywords and semantic structures in claim texts, aiding in the assessment of claim validity. Additionally, other studies have employed convolutional neural networks to fuse medical imaging and text data, improving disease diagnosis accuracy. Although these methods have improved claim adjudication accuracy to some extent, how to effectively integrate different modalities and handle the specific characteristics of each remains a key research challenge [13].

The Transformer model, as a new deep learning architecture, has achieved significant success in processing sequential data, particularly in natural language processing tasks. Unlike traditional recurrent neural networks, Transformer can capture global dependencies in input data through its self-attention mechanism, making it highly effective for handling long texts and complex data. In the insurance domain, some studies have applied Transformer models to problems such as client credit risk assessment and policy fraud detection, with promising results [14].

3. Method

In this study, we proposed a method for medical insurance policy claim discrimination based on a multimodal Transformer model. We assume that claim discrimination can be achieved by joint modeling of multimodal data, including text data, structured data, etc. In order to take advantage of the Transformer model, we appropriately encode the multimodal data and perform weighted fusion of data of different modalities through a self-attention mechanism to improve the accuracy of discrimination and risk control capabilities. The model architecture is shown in Figure 1.

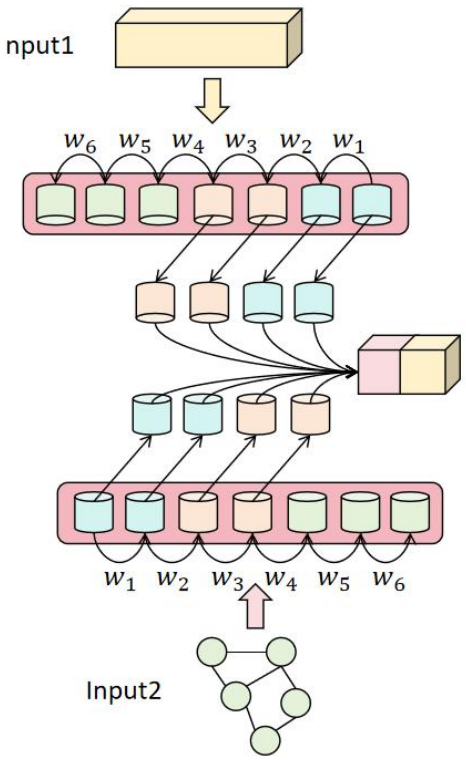

Figure 1. Multimodal Architecture

First, we preprocess the input data. Assume that our input data includes text data X_{test} and structured data X_{struct}. $X_{test} = \{x_1, x_2, ..., x_n\}$ represents text data containing n

words, and $X_{struct} = \{s_1, s_2, ..., s_m\}$ is a data set containing m structured features, which may include age, gender, historical claims records, etc.

In the process of processing text data, we use the pre-trained Transformer model to encode the text data and obtain the embedding representation of each word. Assume that the representation obtained after the text data passes through the Transformer model is $E_{text} = \{e_1, e_2, ..., e_n\}$, where each e_1 is the representation of the word vector, which contains the semantic information of the word. We weight each word vector through the self-attention mechanism to obtain the expression of global semantic information:

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

Among them, Q is the query vector, K is the key vector, V is the value vector, and d_k is the dimension of the key vector. In this way, Transformer can capture the long-distance dependencies between words in the text, thereby effectively processing complex text information related to medical insurance claims.

For the structured data part, we use a standard multi-layer perceptron (MLP) to map the structured data into a high-dimensional space to obtain the representation E_{struct} of the structured data. Suppose we use a simple MLP model whose output is:

$$E_{struct} = MLP(X_{struct})$$

Next, we fuse the representations of text data and structured data to form a unified multimodal input. In order to fuse the representations of these two modalities, we use the multi-head self-attention mechanism of the Transformer model. The calculation formula of the multi-head self-attention mechanism is:

$$MultiHead(Q, K, V) = concat(head_1, head_2, ..., head_h)W^o$$

Among them, h is the number of attention heads, $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$ is the output of the i-th attention head, W_i^Q, W_i^K, W_i^V are weight matrices, and E is the final output weight matrix.

After combining the representations E_{text} and E_{struct} of text and structured data, we get a new representation E_{multi} that represents the synthesis of the two modal information:

$$E_{multi} = MultiHead(E_{text}, E_{struct}, E_{struct})$$

This multimodal representation can be used for subsequent claims discrimination tasks. In order to perform claims discrimination, we pass the multimodal representation E_{multi} into a fully connected layer (FC) for classification:

$$y' = sigmoid(W_{out} \cdot E_{multi} + b_{out})$$

Among them, y' is the predicted output of the model, indicating whether the claim is reasonable (1 means reasonable, 0 means unreasonable). W_{out} and b_{out} are the weights and biases of the output layer.

To train this model, we use the cross entropy loss function to measure the difference between the predicted results and the true labels. The loss function is:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(y'_i) + (1 - y_i)\log(1 - y'_i)]$$

Where N is the number of samples, y_i is the true label of the i-th sample, and y'_i is the model's predicted value for the i-th sample.

Through the above method, the medical insurance policy claim discrimination model based on multimodal Transformer can make full use of the advantages of text data and structured data, and combine the self-attention mechanism and multi-head attention mechanism to improve the accuracy and robustness of claim discrimination. The model can effectively identify potential risks in medical insurance claims, such as false claims, excessive claims, etc., thereby providing insurance companies with more reliable risk control measures.

4. Experiment

4.1 Dataset Introduction

In this study, the dataset used for health insurance claim adjudication consists of real policy and claim records collected from multiple insurance companies. The dataset includes extensive customer information, medical history, claim applications, medical diagnostic reports, and detailed expense records. The customer information section contains age, gender, insurance type, insured amount, and policy duration. The medical history section includes disease type, treatment methods, medical institutions, and consultation dates. The claim application section records claim details such as claim amount, claim time, and claim status. Additionally, the dataset includes auxiliary features such as claim reason classifications and records of denied claims. These features help analyze claim risks more accurately.

The multimodal nature of this dataset makes it suitable for the proposed multimodal Transformer model. The textual data mainly comes from medical diagnostic reports and claim application forms submitted by customers. These texts contain a large amount of unstructured information, such as disease descriptions, treatment processes, and medical advice. The length of these texts varies, and they include many medical terms and industry-specific vocabulary. Therefore, effective preprocessing and feature extraction using natural language processing techniques are necessary. The structured data consists of customer demographics, historical claim records, and medical expenses. These data are highly standardized and provide clear numerical features for model processing. By integrating these two types of data, the model can better assess the validity and risk of claim applications.

The dataset's size and complexity enhance its representativeness and practical value. To ensure data quality and accuracy, we conducted rigorous cleaning and

preprocessing. We removed samples with excessive missing values or evident anomalies. Additionally, data augmentation techniques were used to increase the number of samples in minority classes. The dataset also exhibits label imbalance, as certain categories, such as fraudulent claims and excessive claims, are relatively rare. To address this issue, we applied oversampling and undersampling techniques to balance the data. This ensures that the model can effectively handle these rare but critical risk cases during training.

4.2 Experimental setup

In the experiment, we carefully tuned the hyperparameters of the multimodal Transformer model to ensure its optimal performance in health insurance claim adjudication tasks. First, for processing textual data, we used a BERT-based pre-trained model. The word vector dimension was set to 768, with a hidden layer size of 512, and a 10-layer Transformer encoder structure was employed. For handling structured data, we used a multi-layer perceptron (MLP) with two layers. Each layer had 256 hidden units to effectively extract features from the structured data.

During the fusion phase, we combined the representations of textual and structured data using a multi-head self-attention mechanism. The number of attention heads was set to 8 to capture interactions between different modalities. In the training process, we employed the Adam optimizer with an initial learning rate of 1e-4, and used a learning rate decay strategy to prevent overfitting. The batch size was set to 32. Early stopping was applied to avoid overfitting and ensure optimal performance on the validation set. Additionally, the cross-entropy loss function was used as the objective function, optimized through backpropagation to enable the model to learn potential risk features in claim adjudication effectively.

4.3 Comparative Experiment

To validate the effectiveness of the proposed multimodal Transformer model, we conducted comparison experiments with several classic machine learning and deep learning models. These models include ResNet50, XGBoost, random forests, decision trees, and VGG. ResNet50, as a deep convolutional neural network, excels at extracting image features and has strong expressive power for complex tasks. XGBoost, a highly efficient gradient boosting tree model, is widely used for classification and regression problems, with excellent generalization capabilities. Random forests and decision trees, as traditional tree-based models, process data by constructing multiple decision trees. They offer good interpretability and robustness. VGG is a classic convolutional neural network mainly used for image processing tasks. Although its architecture is relatively simple, it still performs well in certain application scenarios. By comparing with these models, we can comprehensively assess the advantages of the multimodal Transformer in health insurance claim adjudication, verifying its superior performance in terms of accuracy, robustness, and risk control. The experimental results are shown in Table 1.

Table 1: Comparative experiment

Model	AUC	ACC	F1-Score	Precision
Random Forest	0.850	0.784	0.803	0.820
Decision Tree	0.792	0.762	0.779	0.793
XGBOOST	0.871	0.805	0.819	0.830
VGG16	0.876	0.815	0.825	0.840
ResNet50	0.883	0.821	0.830	0.852
Ours	0.901	0.841	0.855	0.879

The experimental results show that the proposed model outperforms all other comparison models across all metrics, especially in AUC, accuracy, F1-Score, and precision. Compared to traditional tree models, random forests and decision trees fall significantly behind in all metrics. This is particularly evident in F1-Score and precision, highlighting the limitations of these models when handling complex multimodal data. While XGBoost performs well on several metrics, there remains a noticeable gap compared to the proposed model, especially in AUC and precision. This indicates that XGBoost is somewhat lacking in risk identification capability for claim adjudication.

Among deep learning models, VGG16 and ResNet50 perform slightly worse than the proposed model. Although ResNet50, as a powerful convolutional neural network, is close to the proposed model in AUC and precision, it still has limitations in processing multimodal data. The proposed model fully leverages the complementary nature of multimodal data, combining text and structured data. The weighted fusion through the self-attention mechanism further enhances risk control and predictive accuracy. In contrast, while VGG16 performs reasonably well, its relatively simple structure leads to weaker performance in handling complex data interactions.

Overall, the experimental results validate the effectiveness of the proposed model for health insurance claim adjudication. The multimodal Transformer architecture significantly improves claim adjudication accuracy, particularly in precision and F1-Score, showing strong advantages. This suggests that, in the health insurance domain, advanced deep learning methods, especially multimodal data fusion techniques, can effectively enhance the model's robustness and risk control capabilities.

4.4 Ablation experiment

Next, this paper gives the results of the ablation experiment, as shown in Table 2. The ablation settings are ablation of different modalities.

The experimental results show that models using only textual data and structured data each perform well, but still lag behind the proposed model in all metrics. The model using only textual data performs well in AUC, accuracy, and precision, particularly with a relatively high precision score (0.833). This suggests that textual data plays a crucial role in understanding the semantic information in claim adjudication. However, the textual data model has a slightly lower F1-Score, indicating that it may struggle in handling data imbalances or certain edge cases.

For the model using only structured data, although its AUC and accuracy are similar to those of the textual data model, its F1-Score and precision are lower, at 0.804 and 0.818, respectively. This result suggests that while structured data provides effective numerical features and historical records, it lacks the capacity to handle complex semantic and contextual information. As a result, it shows limitations in more complex claim adjudication tasks. In particular, when handling claims that require contextual understanding and detailed judgment, the structured data model cannot match the depth of semantic analysis provided by the textual data model.

In contrast, the proposed model combines both textual and structured data, fully leveraging the strengths of both modalities. This significantly improves performance across key metrics such as AUC (0.901), accuracy (0.841), F1-Score (0.855), and precision (0.879). This indicates that multimodal data fusion can better capture the potential risks and complex patterns in claim adjudication. The experimental results validate the effectiveness of the multimodal Transformer model in health insurance claim adjudication and further demonstrate that integrating different types of data significantly enhances the model's predictive ability and robustness.

Table 2: Ablation experiment

Model	AUC	ACC	F1-Score	Precision
Text Data	0.872	0.803	0.815	0.833
Structured Data	0.865	0.799	0.804	0.818
Ours	0.901	0.841	0.855	0.879

4.5 Hyperparameter sensitivity experiments

This paper also conducts hyperparameter sensitivity experiments, mainly to explore the impact of different learning rates on the experimental results. The experimental results are shown in Table 3.

The experimental results indicate that the learning rate has a significant impact on model performance. As the learning rate gradually decreases, the model shows noticeable improvements across all metrics. With an initial learning rate of 0.004, the model performs well in AUC, accuracy, F1-Score, and precision. However, it does not reach its optimal state. Specifically, the precision is 0.843 and the F1-Score is 0.825, which are slightly lower compared to the results achieved with smaller learning rates.

When the learning rate is reduced to 0.003 and 0.002, the model's performance continues to improve. Notably, the F1-Score and precision increase to 0.839 and 0.855, respectively. This indicates that as the learning rate decreases, the model can fine-tune its parameters more effectively, enhancing its ability to recognize complex patterns. With a learning rate of 0.002, the model's performance approaches its optimal state. Further reduction of the learning rate results in final performance optimization.

With a learning rate of 0.001, the proposed model achieves the best performance, especially in precision (0.879) and F1-Score (0.855). This suggests that a lower learning rate helps the model converge more stably, leading to better predictive performance. In health insurance claim adjudication, the appropriate learning rate ensures that the model can effectively identify potential risks when processing complex multimodal data, resulting in higher prediction accuracy and robustness.

Table 3: Hyperparameter sensitivity experiments

Learning Rate	AUC	ACC	F1-Score	Precision
0.004	0.875	0.812	0.825	0.843
0.003	0.888	0.826	0.839	0.855
0.002	0.892	0.834	0.848	0.861
0.001	0.901	0.841	0.855	0.879

4.6 Visualizing Experimental Results

Finally, this paper gives an image of the loss function decrease, as shown in Figure 2.

As shown in the figure, the training loss decreases significantly as the number of epochs increases. During the early stages of training, the loss value is relatively high, around 0.7. However, as the training progresses, the loss curve drops rapidly, indicating that the model is effectively optimizing during the learning process. This sharp decline typically suggests that the model is able to capture the main features of the data in the initial phase and adapt quickly to the structure of the training data.

After the loss reaches a lower value, the curve flattens, indicating the model's convergence in the later stages of training. At this point, the loss continues to decrease but at a much slower rate, suggesting that the model is approaching its optimal state. Further training at this stage may not lead to significant performance improvements. This is a common phenomenon, as the loss tends to stabilize when the model is nearing the optimal solution.

Overall, the loss curve during the training process shows that the proposed model effectively learns and optimizes relevant features. The steady decrease in loss indicates that the model gradually converges after reaching a high level of accuracy. This ensures the model’s efficiency and stability in the claim adjudication task.

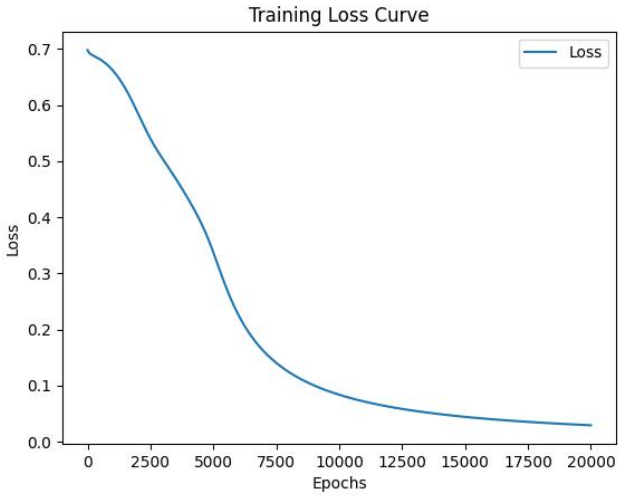

Figure 2. Loss function drop graph

Overall, the loss curve during the training process shows that the proposed model effectively learns and optimizes relevant features. The steady decrease in loss indicates that the model gradually converges after reaching a high level of accuracy. This ensures the model's efficiency and stability in the claim adjudication task.

5. Conclusion

This paper presents a multimodal Transformer-based model for medical insurance claim adjudication and validates its effectiveness across different data modalities through a series of experiments. By integrating text data with structured data, the proposed model performs exceptionally well in medical insurance claim adjudication tasks, particularly excelling in metrics such as AUC, accuracy, and F1-Score. Compared to traditional machine learning and deep learning models, the experimental results show that the model can effectively identify potential risks in medical insurance claims, such as fraudulent claims and overclaims, providing insurance companies with more accurate and efficient risk control tools. Furthermore, this study highlights the critical impact of learning rate and data modality selection on model performance through comparison experiments, further demonstrating the advantages of multimodal learning in practical applications.

Theoretically, this study explores the application of the multimodal Transformer model in medical insurance claim adjudication, offering new insights for research in this field. By combining natural language processing techniques with traditional structured data analysis methods, the model can make decisions from multiple perspectives, enhancing its ability to recognize complex data patterns. In particular, the self-attention mechanism in the Transformer architecture allows the model to precisely capture important information from large datasets, improving both the accuracy and robustness of claim adjudication. This contribution provides strong support for the further application of multimodal learning in financial risk management and offers valuable guidance for multimodal learning research in other domains.

However, despite the promising results of the proposed model in experiments, several areas still require improvement. Firstly, although the fusion of multimodal data significantly enhances model performance, how to perform more refined weighted fusion between different modalities to maximize the advantages of each data source remains an open question. Secondly, although the model performs well across multiple metrics, there is still a need to address data imbalance issues by incorporating more advanced techniques, such as Generative Adversarial Networks (GANs), to improve the model's ability to recognize minority classes. Lastly, in more complex claim scenarios, the model's interpretability and transparency need to be further improved. For insurance companies, the model should not only have high prediction accuracy but also provide clear decision-making justifications. Thus, enhancing the interpretability of the model's adjudication process remains a critical direction for future research.

Looking ahead, with the continuous growth of data in the medical insurance and other financial sectors, the application of multimodal learning and deep learning will become more widespread. While the proposed model addresses some of the limitations of existing claim adjudication systems, there is still room for improvement in handling more complex and dynamically changing financial risks. Future research could explore how to combine the multimodal Transformer with other deep learning architectures, such as graph neural networks or reinforcement learning, to better handle complex data with temporal and spatial correlations. Additionally, as technology advances, the computational efficiency and real-time responsiveness of the model will become important research areas. Enhancing computational efficiency while maintaining accuracy and adapting to large-scale data processing will be key to the successful application of multimodal learning in financial risk management.

References

[1] Kline, Adrienne, et al. "Multimodal machine learning in precision health: A scoping review." npj Digital Medicine 5.1 (2022): 171.

[2] Verma, Gaurav, et al. "Overcoming language disparity in online content classification with multimodal learning." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 16. 2022.

[3] Booth, Brandon M., et al. "Bias and fairness in multimodal machine learning: A case study of automated video interviews." Proceedings of the 2021 international conference on multimodal interaction. 2021.

[4] Ter Schure, Sophie, Caroline Junge, and Paul Boersma. "Discriminating non-native vowels on the basis of multimodal, auditory or visual information: Effects on infants' looking patterns and discrimination." Frontiers in Psychology 7 (2016): 525.

[5] Huang, Guan, et al. "Multimodal learning of clinically accessible tests to aid diagnosis of neurodegenerative disorders: a scoping review." Health Information Science and Systems 11.1 (2023): 32.

[6] Scepanovic, S., Martin-Lopez, E., Quercia, D., & Baykaner, K. (2020, April). Extracting medical entities from social media. In Proceedings of the ACM conference on health, inference, and learning (pp. 170-181).

[7] Chow, Wei, et al. "Unified Generative and Discriminative Training for Multi-modal Large Language Models." arXiv preprint arXiv:2411.00304 (2024).

[8] Xu, Peng, Xiatian Zhu, and David A. Clifton. "Multimodal learning with transformers: A survey." IEEE Transactions on Pattern Analysis and Machine Intelligence 45.10 (2023): 12113-12132.

[9] Zhang, Yiyuan, et al. "Meta-transformer: A unified framework for multimodal learning." arXiv preprint arXiv:2307.10802 (2023).

[10] Miyazawa, Kazuki, Yuta Kyuragi, and Takayuki Nagai. "Simple and effective multimodal learning based on pre-trained transformer models." IEEE Access 10 (2022): 29821-29833.

[11] Zhang, Yao, et al. "mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2022.

[12] Hu, Ronghang, and Amanpreet Singh. "Unit: Multimodal multitask learning with a unified transformer." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[13] Akbari, Hassan, et al. "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text." Advances in Neural Information Processing Systems 34 (2021): 24206-24221.

[14] Zhou, Hong-Yu, et al. "A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics." Nature biomedical engineering 7.6 (2023): 743-755.