Multimodal Deep Learning & Visually Explainable AI Applications: A Literature Review

Ajay Raigaga, Keren Chen, Rohaan Raheja

Abstract

The integration of Multimodal Deep Learning (MMDL) in insurance subrogation is transforming claim evaluations by incorporating textual descriptions, image-based evidence, and structured metadata. While these models significantly improve fraud detection and automated claim validation, their lack of transparency poses a challenge for adoption in regulatory-driven industries. This literature review explores recent advancements in MMDL for insurance claim processing and investigates the role of Visually Explainable AI (XAI) in enhancing model interpretability. Techniques such as SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations), and saliency mapping provide insights into feature importance and decision-making processes, ensuring fairness and accountability. We analyze state-of-the-art multimodal architectures, including AutoFraudNet and AIML, and discuss their applications in fraud detection and subrogation. Additionally, we highlight key challenges such as bias in multimodal fusion, dataset limitations, and computational inefficiencies that hinder real-world deployment. Future research should focus on hybrid explainability frameworks and dataset optimization to enhance transparency, fairness, and efficiency in AI-driven subrogation. As a next step, this study aims to develop a Python-based interface integrating MMDL with XAI, enabling interpretable decision-making for insurance professionals.

Introduction

Subrogation is a crucial process in the insurance industry, allowing insurers to recover costs by pursuing third parties responsible for damages after a claim has been paid. Traditionally, subrogation decisions have relied on manual claim assessments, requiring

extensive human effort to analyze textual claim descriptions, photographic evidence, and structured metadata such as customer history and policy details. However, manual processes are often time-consuming, prone to errors, and susceptible to inconsistencies in claim validation. The integration of artificial intelligence, particularly MMDL, has emerged as a promising solution to enhance the efficiency and accuracy of subrogation decisions. By leveraging multiple data modalities—including text, images, and structured data—MMDL models can capture complex relationships between different sources of information, leading to improved fraud detection, automated claim approvals, and better decision-making.

Despite these advancements, a significant challenge in deploying AI-driven subrogation models lies in the lack of explainability. Deep learning models, while highly accurate, often function as “black boxes,” making it difficult for insurers, regulators, and policyholders to understand how specific claims are assessed. To address this, explainable AI (XAI) techniques, such as Shapley Additive Explanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and saliency mapping, have been introduced to provide transparency into model predictions. These methods help identify key factors influencing claim outcomes, ensuring fairness, reducing bias, and increasing trust in AI-driven subrogation decisions. This literature review explores recent developments in multimodal deep learning for subrogation and investigates the role of visually explainable AI in making these models more interpretable and reliable. By analyzing existing research and methodologies, this study aims to highlight the potential of MMDL and XAI in optimizing insurance claim processing while addressing key limitations and future research directions.

Multimodal Deep Learning

The adoption of multimodal deep learning in the insurance industry has introduced significant improvements in claim approval processes by integrating textual descriptions, visual evidence, and structured metadata to enhance accuracy and efficiency. Recent research has focused on developing frameworks that systematically extract, process, and fuse features from different data sources to automate claim validation and fraud detection.

Asgarian et al. (2023) proposed AutoFraudNet, a multimodal deep learning framework designed for fraud detection in auto insurance claims. The model employs a cascaded slow fusion approach to gradually integrate features from different modalities, reducing computational complexity and preventing overfitting. BLOCK Tucker fusion blocks are used to effectively combine high-dimensional data, maintaining model efficiency while ensuring optimal performance. AutoFraudNet processes visual inputs using pre-trained CNNs, while BERT extracts features from textual claim descriptions. Structured tabular data undergoes one-hot encoding before being fused with the other modalities. The model is trained using cross-entropy loss and an Adam optimizer, with early stopping and data augmentation techniques to enhance generalizability. By leveraging these structured fusion methods, AutoFraudNet improves fraud detection while maintaining scalability for real-world claim processing.

Similarly, Yang et al. (2023) introduced the AIML framework, which integrates textual, visual, and structured data for fraud detection and claim assessment. NLP models extract key information from textual accident descriptions, while CNN-based networks process image inputs of reported damages. Structured metadata is encoded using one-hot representations and fused with the textual and visual features. The framework applies concatenation and Bilinear Pooling strategies to optimize feature combination, ensuring that relevant information from each modality is preserved. To enhance robustness, stratified sampling and class weighting techniques are used to handle class imbalance, reducing biases in the model. The model's cross-modal interactions allow it to effectively detect inconsistencies in claim descriptions,

making it applicable to automated claim approvals by validating the authenticity of text-based insurance claims against supporting data.

In the healthcare domain, Kline et al. (2022) reviewed the role of multimodal machine learning in decision-making, particularly in insurance-related healthcare claims. Their study highlights the advantages of early fusion techniques, where different data types—such as clinical notes (text), X-rays (images), and patient demographics (structured data)—are combined before model training, leading to more accurate predictions. The NLP-based models process text descriptions of medical diagnoses and treatment plans, while CNNs analyze medical imaging, and structured metadata provides additional context for insurance claims. The review suggests that such cross-modal fusion strategies could be transferred to insurance claim approval, particularly in complex cases such as medical or property insurance claims, where text-based claim narratives need to be validated against structured and visual records.

Visually Explainable Methods for Multimodal Deep Learning

Multimodal Deep Learning integrates multiple types of data, such as tabular data, text, images, and even audio. However, the inherent complexity of these models may make interpretability difficult, especially for non-technical users. Thus, in business contexts, it is crucial to develop explainable AI (XAI) techniques that supplement these models with more understandable & intuitive insights into the decision making processes of MMDL. These methods can leverage visualization techniques to highlight feature importance, cross-modal interactions, and attention distributions, enabling both skilled and non-skilled users to enhance their trust in AI systems. This section explores various visually explainable methods useful for MMDL.

Shapley Additive Explanations (SHAP)

SHAP is a common interpretability technique that assigns importance scores to input features based on their importance to the model's prediction. It uses calculated scores (Shapley values) that measure how much each feature contributes to a model's prediction by comparing the model's output with and without that feature across many different combinations of scenarios. In MMDL, SHAP is particularly valuable because it not only allows researchers to quantify the contributions of different features, but also different modalities in a model's decision-making process. Since MMDL models integrate multiple data types, understanding whether the model is relying more on one modality over another is critical for debugging and reducing bias. For example, in a classification model on whether Country Financial should pursue subrogation on a paid claim, if a model prioritizes customer information (tabular) over a claim description (text), it may make a biased decision on whether to pursue subrogation. SHAP enables a visual overview of which features and modalities are influencing the model the most. Additionally, SHAP enables instance-specific explanations, meaning it not only provides general insights into how a model operates but also helps analyze individual predictions.

Figures 1 & 2 show a SHAP bar plot and corresponding summary plot, respectively, for a hypothetical insurance claims model. From the bar plot, we can determine that Claim Amount (0.65) had the strongest positive impact on its decision, increasing the prediction, while Claim History Score (-0.48) significantly decreased it, possibly indicating past claim behavior matters. Days Since Last Claim (0.39) and Adjuster Notes Tone (0.25) positively influenced the outcome to a lesser degree, suggesting that longer gaps between claims and positive adjuster notes increase the likelihood of a favorable decision. Meanwhile, Claim Description Sentiment (-0.31) negatively impacted the prediction, implying that the wording in the claim description reduces approval chances. The summary plot shows how many instances of each feature are present in

the model's data, as well as each instance's individual SHAP score. The color gradient reveals how feature values affect outcomes.

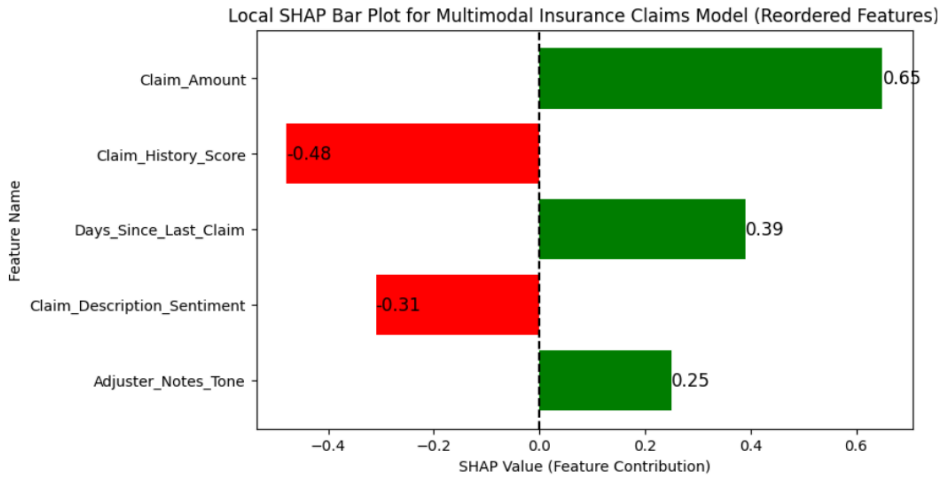

Figure 1: Bar plot of average SHAP values

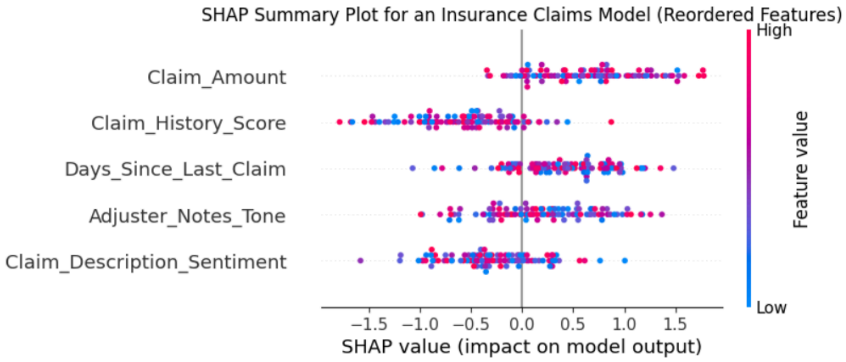

Figure 2: Corresponding summary plot of individual SHAP values

Local Interpretable Model-agnostic Explanations (LIME)

LIME is an explainable AI technique that helps interpret machine learning models by approximating their predictions with locally interpretable, simpler models. LIME works by generating multiple variations of a single observation in the data, slightly altering feature values, and observing how these changes affect the model's output. By fitting a lightweight, interpretable model (such as linear regression) to these perturbed samples, LIME highlights which features contributed most to a specific prediction. For instance, imagine a deep learning

model with potentially millions of parameters & observations that predicts whether a tweet is fake news or real news. LIME takes one tweet, generates slightly modified versions of it, sees how the model's prediction changes, and then builds a small, easy-to-read model (like a weighted list of words) to explain which words had the most influence in classifying that specific tweet. As opposed to SHAP, which shows a broad view of the most important features of the whole dataset, LIME offers a more nuanced model interpretation, breaking down importance by observation. As a visual explainable AI tool, LIME depends on the data type. For tabular data, it typically presents feature importance scores using bar charts or heatmaps, where the most influential features are highlighted. For text data, LIME visualizes influential words through color-coded text overlays or word importance lists.

Figure 3 shows a LIME table of a hypothetical model that predicts whether a news article is fake or real based on its headline. Based on the predicted probabilities and weights of each word, LIME shows us that sentences with political names like "Hillary" and "Clinton" are more likely to be classified as fake, while those with neutral or geopolitical terms are seen as real. Some sentences show mixed results, meaning the model is unsure. Words related to social and religious topics, like "LGBT," slightly increase the fake probability.

Test	Sentence	Probability fake	Probability real	Highlighted words	Weights
1	FBI NEW YORK FIELD OFFICE Just Gave A Wake Up Call To Hillary Clinton	0.97	0.03	1. Gave 2. Just 3. A 4. Hillary 5. Clinton	+ 0.28 + 0.25 + 0.23 + 0.21 + 0.17
2	Turkey-backed rebels in Syria put IS jihadists through rehab	0.00	1.00	1. Turkey 2. Syria 3. in 4. backed 5. jihadist	– 0.02 – 0.02 – 0.01 – 0.01 – 0.01
3	Trump looms behind both Obama and Haley speeches	0.58	0.42	1. and 2. Obama 3. Haley 4. looms 5. behind	+ 0.17 + 0.16 + 0.13 – 0.05 + 0.05
4	Pope Francis Demands Christians Apologize For Marginalizing LGBT People	0.29	0.71	1. For 2. Pope 3. People 4. Marginalizing 5. LGBT	– 0.10 – 0.08 + 0.08 + 0.08 + 0.04

Figure 3: Plot for LIME evaluation of fake news

Saliency Mapping

Saliency mapping is an explainable AI technique that highlights the most influential parts of an input that drive a model's prediction. As a visual explainable AI tool, saliency mapping is often represented as a heatmap for images, where brighter areas indicate regions that strongly contributed to the model's classification. For text-based models, saliency is visualized by highlighting important words, often using color intensity to indicate their influence. In tabular data, it assigns numerical importance scores to each feature, which can be presented as bar plots or tables. As saliency mapping is only a visualization technique, there are several models one can employ to determine importance, including LIME, Grad L2, and Integrated Gradients. The best way to map saliency would be a combination of several methods, but this can be computationally expensive. Saliency mapping is particularly useful for explaining NLP models due to its ease of interpretation with text data.

Figure 4 shows an example of how saliency mapping can be used to determine the sentiment of an IMDb movie review. As there are many more & darker red-highlighted words than blue, one can easily (and correctly) understand why a sentiment analysis would predict this excerpt to be negative.

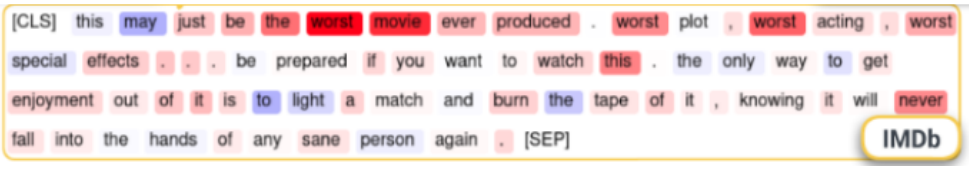

Figure 4: Saliency map highlighting key words in a negative IMDb review,

Partial Dependence Plots (PDPs)

PDPs help explain machine learning models by showing how a feature influences predictions while holding other variables constant. Unlike the aforementioned methods, PDPs are unique because they show a graphical representation of how different values of the same

feature affect model output, introducing an element of direction and linearity not present in other methods. One advantage of PDPs is their interpretability for non-experts, as they rely on simple line plots where the x-axis represents the feature value and the y-axis represents the predicted probability (in classification tasks) or predicted output (in regression tasks). In MMDL, where models can integrate tabular and text-based features, PDPs clarify how structured data and language-derived variables impact predictions.

For example, in a hypothetical customer support escalation model, PDPs can illustrate how Urgency Score (tabular) and Sentiment Score (text-based) affect escalation probability. From figure 5, we can conclude that a rising PDP curve for urgency suggests the model correctly prioritizes urgent cases, while a downward trend for sentiment may indicate negative complaints increase escalation likelihood.

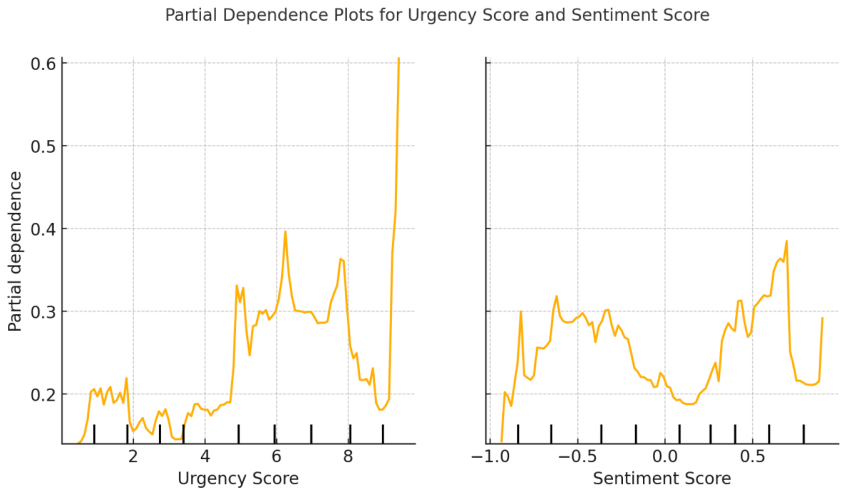

Figure 5: Partial Dependence Plots

Conclusion

The integration of multimodal deep learning into subrogation processes is revolutionizing the insurance industry by enabling more efficient and accurate claim evaluations. Through the

fusion of textual claim descriptions, image-based evidence, and structured metadata, MMDL models provide a comprehensive approach to fraud detection and automated claim validation. Research in this area has demonstrated the potential of advanced deep learning architectures, such as AutoFraudNet and the AIML framework, in improving the reliability of claim assessments. However, while these models enhance prediction accuracy, their lack of transparency remains a critical barrier to widespread adoption in regulatory-driven industries like insurance. Without explainability, AI-driven subrogation decisions may face skepticism from insurers, auditors, and legal professionals, leading to concerns over fairness and accountability.

To bridge this gap, explainable AI methods, including SHAP, LIME, and saliency mapping, have been incorporated to provide interpretability in MMDL-based subrogation models. These techniques help visualize the importance of different input features, enabling both technical and non-technical stakeholders to understand model decisions. However, challenges such as bias in multimodal data fusion, limited availability of real-world datasets, and computational inefficiencies still hinder the full-scale deployment of these models in practical insurance applications. Moving forward, future research should focus on developing hybrid explainability frameworks that combine global model interpretability with instance-level explanations to enhance transparency. Additionally, improving dataset diversity and optimizing computational performance will be crucial in ensuring the robustness of AI-driven subrogation models. As a next step, this project will aim to build a Python-based interface that integrates multimodal deep learning with visually explainable AI methods, making subrogation decision-making both accurate and interpretable for insurance professionals.

References

- **Asgarian, Azin, et al.** "AutoFraudNet: A Multimodal Network to Detect Fraud in the Auto Insurance Industry." *Association for the Advancement of Artificial Intelligence (AAAI)*, 2023.
- **Kline, Adrienne, et al.** "Multimodal Machine Learning in Precision Health: A Scoping Review." *npj Digital Medicine*, vol. 5, no. 171, 2022, https://doi.org/10.1038/s41746-022-00712-8.
- **Yang, Jiaxi, et al.** "Auto Insurance Fraud Detection with Multimodal Learning." *Data Intelligence*, vol. 5, no. 2, 2023, pp. 388-412. MIT Press, https://doi.org/10.1162/dint_a_00191.
- Trevisan, V. (2022, January 17). “Using shap values to explain how your Machine Learning Model Works.” *Towards Data Science*. https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137/
- Mavrepis, P., Makridis, G., Fatouros, G., Koukos, V., Separdani, M. M., & Kyriazis, D. (2024). XAI for All: Can Large Language Models Simplify Explainable AI? arXiv preprint arXiv:2401.13110.
- Keita, Z. (2023, May 10). Explainable AI, Lime & Shap for model interpretability: Unlocking AI’s decision-making. *DataCamp*. https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models
- Molnar, C. (2022). Interpretable machine learning: A guide for making Black Box models explainable. Christoph Molnar.