KeAi
CHINESE ROOTS
GLOBAL IMPACT

International Journal of Innovation Studies

MULTIMODAL AI FOR REAL-TIME UNDERWRITING: FUSING VOICE, TEXT & BEHAVIORAL BIOMETRICS

[1]Jayasri Dudam
Senior Software Engineer
American Express
ORC id: 0009-0001-9317-9606

[2]Raja Ramesh Bedhaputi
Senior Engineer, American Express
ORC id: 0009-0002-8184-2340

[3]Deeraj Madhadi
Sr Software Engineer, Fidelity Investments
ORC id: https://orcid.org/0009-0006-7061-5504

[4]Sri Sai Krishna Mukkamala
Sr Mobile Engineer, Early Warning Services
ORC id: 0009-0000-0606-7706

Abstract

The insurance industry is shifting toward real-time, automated decision making enabled by artificial intelligence (AI) and multi-modal data fusion. This paper presents a new real-time underwriting framework that uses multi-modal AI with voice, text, and behavioral biometrics. Underwriting traditionally relies on static data and manual review, which can be slow, subjective, and miss signs of risk. We use dynamic, real-time data streams from multiple human-computer interaction modalities to create a more complete risk profile. For instance, voice data can reveal tone of voice, speech problems due to stress or anxiety, and hesitation patterns through advanced natural language processing (NLP) and paralinguistic analysis. Along with vocal inputs, textual inputs (for example, chat based interactions and form responses) are examined for linguistic complexity, sentiment bias, and inconsistency with other data inputs. At the same time, behavioral biometric data (typing and mouse cadence, touch screen dynamic behavior) provides an unobtrusive layer of participants' user verification and cognitive states. The various signal inputs are fused together using a multi-modal deep learning framework, whereby temporal patterns and correlations across modalities are established in order to develop an enriched risk score in real-time. The system is consistently updated using reinforcement learning and feedback from underwriting outcomes allowing it to continuously adapt process. This multimodal method not only improves speed and accuracy for underwriting but enhances the customer experience by way of friction and false positive minimisation. Our work presents the architecture of our systems, the signal processing methods, and the ethical

issues around privacy and explainability in automated decision making. The results lead to the future direction of a new class of intelligent underwriting systems that will be proactive, adaptive and predominantly human-aware.
Keywords: Multimodal AI, Real-time Underwriting, Behavioral Biometrics, Voice Analysis, Risk Assessment

1. Introduction

The insurance sector revolves on the underwriting procedure. For insurers to charge reasonable premiums, they must first assess the risk level posed by prospective policyholders. Traditionally, underwriting has been a laborious paper-based procedure that relies on static data like medical records, financial records, and self-reported surveys. Insurance companies are facing changes as a result of digitalisation and customers' aspirations for faster, more accurate, and less invasive services. As a result, in an effort to increase efficiency and speed in the underwriting process, the insurance business is looking at new technologies, including AI. An innovative development in this field is multimodal AI. To get a deeper understanding of user behaviour, intent, and danger, multimodal AI may mix input from several modalities of human-computer interaction, in contrast to unimodal AI systems that can only consider a single data source. The focus of this study is on improving underwriting in real-time using multimodal AI, which combines the three primary modalities of speech, text, and behavioural biometrics. Hearing an applicant's voice may tell you a lot about their mental condition and whether or not they are trying to mislead you by picking up on signs of tension, hesitancy, and tone qualities. Another way to interpret emotion, linguistic complexity, or semantic incompatibilities is via textual analysis, which uses natural language processing (NLP) for written replies or chatbot discussions. Finally, by tracking the user's activities while interacting with the device—typing habits, mouse movements, and touch screens—behavioral biometrics are among the most subtle yet potent markers of user identity and mental state. Accurate and contextual real-time risk profiling is made possible by integrating all modalities into a single deep learning architecture. In contrast to more conventional human evaluations, the system may detect small, and potentially abnormal, variations in attitude and conduct when it can match temporal data and multi-modal correlations. The shift towards intelligent underwriting in real-time streamlines the underwriting process, which in turn boosts client happiness and fraud detection. The underwriting cycle is shortened without sacrificing accuracy or consistency in applicant profiling, and risk criteria are maintained. Transparency, data privacy, and algorithmic fairness are ethical considerations that may help with system design, which is important for achieving trust and regulatory compliance. This paper detail a comprehensive framework for multimodal AI-powered underwriting, outline its components, and explore the potential to change risk analyses in the digital insurance landscape. In an era where insurers must embrace a new frontier, multimodal AI can serve as a strategic advantage for providing smarter, more adaptive, and human-aware underwriting.

2.Related work

The field of multimodal learning focuses on acquiring information in various data modalities including text, audio or visual. This has been a significant focus in recent years. Theoretical underpinnings and practical applications of multimodal AI systems have both been the primary areas of academic research. Where theoretical progress, challenge, and applications impact the expansion of multimodal learning, this literature review presents significant research

contributions.

Multimodal learning is based on the principle that machines may learn more effectively when given more representations from multiple sources (or modalities). Previous efforts in multimodal systems relied on basic fusion algorithms to enhance picture captioning and voice recognition performance by merging data streams from different sources. Combining text, user history, and visuals may increase predictive power, as Baltrunas et al. showed in their groundbreaking work on multimodal learning for recommendation systems. The power of varied data sources boosts the resilience of AI systems. This was shown in an early research on multimodal sentiment, which found that integrating text and audio enhanced sentiment categorisation beyond utilising text alone. From its inception, multimodal learning within the realm of natural language processing was limited to static combinations of visual and textual input, including photographs with captions. Many of the first models relied on deep learning, with convolutional neural networks (CNNs) and recurrent neural networks (RNNs) being the most popular. Although they showed promise, these early systems couldn't learn dynamically how text, pictures, and audio interact, therefore they couldn't understand deep relationships between modalities. More sophisticated machine learning methods have led to the development of improved multimodal fusion algorithms. In order to fuse information across multiple modalities for diverse purposes, researchers have examined late, early, and hybrid fusion methodologies. Emotion recognition, facial expressions, speech tone, and body language all contribute to understanding the message, and late fusion is a commonly used technique that combines modality-specific models after independent processing in many related application.. Nevertheless, late fusion's performance is severely constrained due to the fact that it examines modalities independently, potentially leading to the loss of important inter-modal connections. On the other hand, early fusion creates much more cohesive representations by merging information from different modalities early on in the input process. [1] demonstrated that VQA works well with early fusion, which involves integrating the picture and question early on to enable simultaneous processing. Early fusion can be computationally expensive when it comes to encoding high-dimensional feature spaces. Given these cost challenges, hybrid fusion methods that use early and late fusion methods have been proposed [2]. As heat maps are bright for the future of bimodal learning, there still are obstacles to surmount.

One of the key challenges is aligning data. For multimodal system performance to be reliable, the data from the two modalities must be spatially and temporally aligned.To clarify, while processing video, both the audio and visuals should relate to one another. This has proven to be a challenge, especially since many modalities are recorded at different times and when connected together, the data become very complicated compared to a single-dimensional set. It has been recommended that the multimodal networks are trained simultaneously such that attention shares different data in the input when processing information, they naturally will create better multimodal data alignment. Feature fusion has been recognized as a significant challenge for some time as it involves attempting to combine features from different input modalities which are entirely different from one another. As [3] suggests, the visual features appear to be more likely spatially organized (and organized) rather than a linear sequence of symbols and qualities while the text features are a sequence and symbolic.

A recent paper [3] proposed the use of transformers for aligning and combining multiple modalities. The authors suggested that transformers would perform better to combine features

due to the self-attention.Lately, there has been a lot of buzz surrounding multimodal artificial intelligence (AI) and its use across newly emerging domains, including autonomous systems, entertainment and healthcare. In healthcare, multimodal AI systems can make bigger contributions to diagnostic determinations through the incorporation of medical imaging data, patient records and text video data. A specific example of a multimodal approach based on Deep Learning (DL) to detect lung cancer, used radiograph images and clinical text data and achieved better prediction outcomes than single-modality based models [4]. In another example, [5] had used audio and clinical text data from patient interviews to predict mental health outcomes. The multimodal models, performed better than single-modality models. For autonomous systems, multimodal AI will provide architecture for navigation and improving decision-making processes for systems such as vehicles.

To improve the real-time identification and response to navigational impediments of the vehicle significantly, the combination of various sources of observations are fused. To improve human communication with robots, researchers have developed multimodal systems addressing both verbal and nonverbal indications in the field of human-robot interaction. The media and entertainment sectors have also utilized multimodal learning to promote user experiences. An example includes the multimodal learning research which combines voice commands, video, and text to enhance interaction in virtual reality environments. Researchers were able to improve the realism and interaction of virtual reality applications using multimodal AI by accounting for the emotional state of the user based on nonverbal cues, such as facial expressions and intonation of voice, and others. These works illustrated that multimodal learning could design systems that facilitate more engaging, useful, and human-like systems that could disrupt the industry.

A wide variety of research topics will help guide the future directions of multimodal AI. Self-supervised learning methods present a promising method for reducing the dependence on labelled data. The ability for models to learn representations from unlabelled multimodal data could allow for more completely untrained models. Additionally, some work suggests that self-supervision can learn adequate representations from extremely large multimodal datasets which could not only provide a path toward universal multimodal systems but also scalable systems for practical applications. Another actively pursued area of research for multimodal AI is reinforcement learning. Reinforcement learning uses trial and error training to help train models to act in desirable ways in constantly changing environments. One avenue that would be interesting to explore would be multimodal reinforcement learning models that would modify chatbot customer service systems to improve user behaviour accommodation with adaptive actions. Problems of fairness and bias are rapidly becoming a significant area of multimodal systems research. An immediate and pressing concern is that redundancy in large datasets with variable training examples may in all good intention reinforce existing bias. One of the examples of the dangers of biased AI systems was provided by[6].

Therefore, more research is needed to confirm the fairness and accountability of multimodal AI models. Academics from a range of fields have made valuable contributions to the fast-moving field of multimodal learning in artificial intelligence. Data alignment, feature fusion, and reproducing and extending systems to scale, must be addressed before we can expect artificial intelligence development in the visual, audio, and semantic aspects of language to significantly advance. Self-supervised learning, reinforcement learning, and different ways of

machine learning have the potential to change many sectors of society and lead to new intelligent systems that are more advanced and similar to humans.

3. AI as a Pioneering Field of Innovation: A Case Study on Multimodal AI for Real-Time Underwriting

Artificial Intelligence (AI) has transformed modern technology and several industries as it relates to its ability to simulate and even enhance human intelligence. One of the most exciting developments in AI is the emergence of multimodal AI. Compared to unimodal AI, multimodal AI utilizes data from multiple modes of awareness and interaction; this includes inputs from both sensory and behavioral information - such as voice, text, and actions - to develop more sophisticated, context-aware systems. An especially engaging application of AI's potential is through new approaches to underwriting in real-time, where multimodal AI has already brought new efficiencies to traditional underwriting practices in terms of speed, accuracy, and adaptability.

Traditionally, underwriting requires data to be processed by a human, based on discrete, dumb information (i.e., forms, historical data records, and medical records). The result is often a slow manual underwriting process, controlled by potential human error or bias. Now that AI is operating at scale, primarily as deep learning and natural language processing (NLP), large portions of the underwriting process are being automated. However, unimodal systems still rely on one capacity or one mode of information acquired during the human interactions inherent to underwriting. Unimodal systems continue to suffer from an inability to account for the rich, dynamic, and complex nature of human behavior and interpersonal interaction [7].

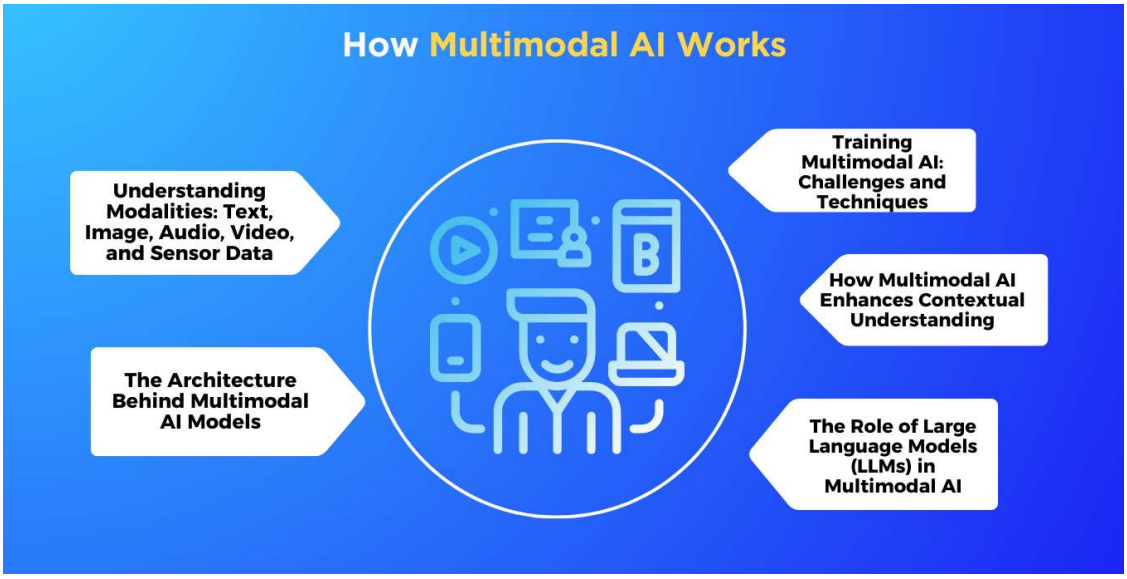

Figure 1: Multimodal AI works

Multimodal AI mitigates this gap by combining voice, text, and behavioral biometrics to provide a comprehensive, real-time assessment of applicants. For example, voice inputs measure stress responses or emotional expressions through rhythm and intonation of speech. The passage is processed for analysis of three common areas: semantic congruence, sentiment, and intent using various arc of NLP techniques and models. Behavioral biometrics are based on passive cognitive and identity authentication via many variables, including typing rhythm, or touchscreen. All of this data is processed and learned at once, in a continuous deep learning

loop to sense even very slight non-congruity in behavior or fraudulent intent which is usually not done or missed in the classic models.

This capability is a dramatic step, not just for insurance, but also advances AI in general. Multimodal use is evidence of AI's movement from being only for automation, to adaptive, intelligent decision making. The human aware, not human-centered, and personalized nature is potentially revolutionary for insurance and industries with real time high stakes decision making (financial and healthcare).

As insurers embrace digital transformation, multimodal AI highlighted the role of AI as a disruptive element, pushing boundaries in respect to trust, personalization, and security. This innovation shows how smarter, faster, and more ethical businesses can be achieved by mixing multiple streams of human data with machine intelligence and putting AI at the center of the future of risk management.

4. Augmented reality (AR) including speech, video, and images for multimodal learning

Multimodal learning uses visual, auditory, and linguistic input to help AI understand and interact with the environment. Medical diagnostics, autonomous vehicles, image-captioning, and emotional recognition are among the most notable categories of usage. Although multimodal AI deals with many complex issues surrounding data fusion and ethics, many different types of AI are expected to usher in revolutionary changes.

Introduction to Multimodal Learning

The latest artificial intelligence techniques involve multimodal learning that synthesizes information that is visual, auditory, and linguistic. It is a multidisciplinary approach that uses artificial intelligence algorithms to comprehend complex environments by integrating and interpreting different kinds of real-world information through various robophysical modes. Multimodal systems can leverage visual, auditory, and linguistic input to enhance performance in more complicated tasks, such as voice recognition, emotion detection, and object tracking, by integrating visual, auditory, and linguistic representations of the environment into a more complete representation. Recent events in deep learning and neural networks have advanced quickly and made it possible for AI systems to include multimodalities in a more seamless manner and gain robustness over diverse applications.

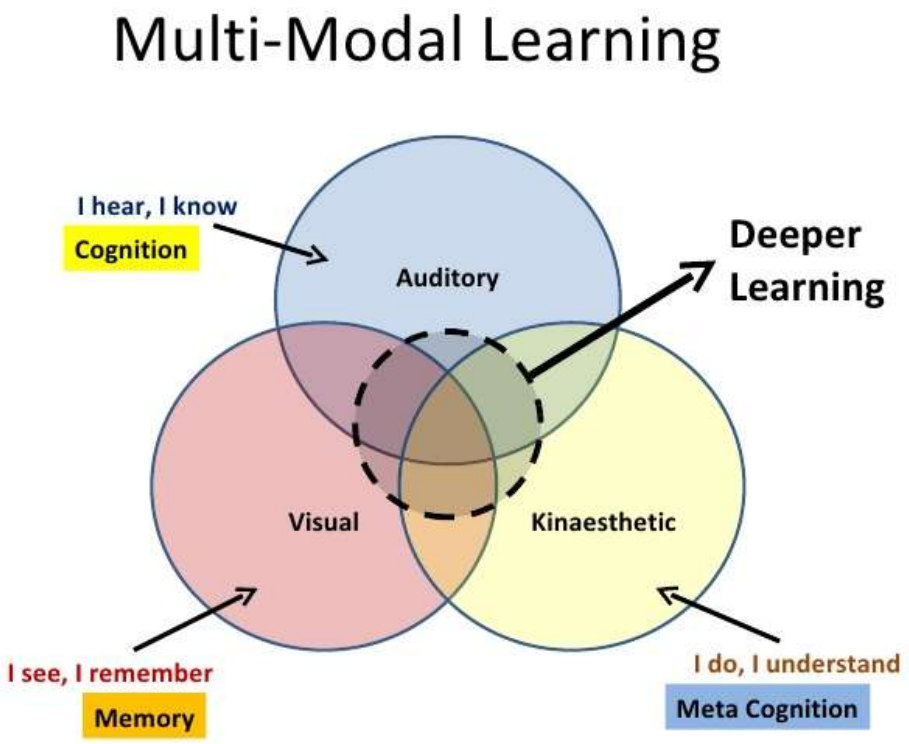

Figure 2: Multi-Modal Learning

Artificial Intelligence for Emotion and Multimodal Speech

Emotion and voice recognition systems are one of the few applications of multimodal AI. In this situation we combine visual, auditory, and linguistic input to better understand human emotion and build better communication products. Companies like Affectiva employ multimodal AI for the purposes of decoding people's emotions by studying their facial expressions, and the tone of their speech. Affectiva's intended applications were in automated customer service in which the AI was meant to decode client facial expressions, and vocal intonation for in turn provide an appropriate response. By combining audio cues with text content, the speech analytics, sentiment analysis, and real time activities of a virtual assistant or chatbot is enhanced. Even companies like IBM Watson have begun to build multimodal functions into their platforms. All of this has greatly increased the scope of AI encounters that are more compassionate, and perceptive, valuable in many spaces, including customer service and health care.

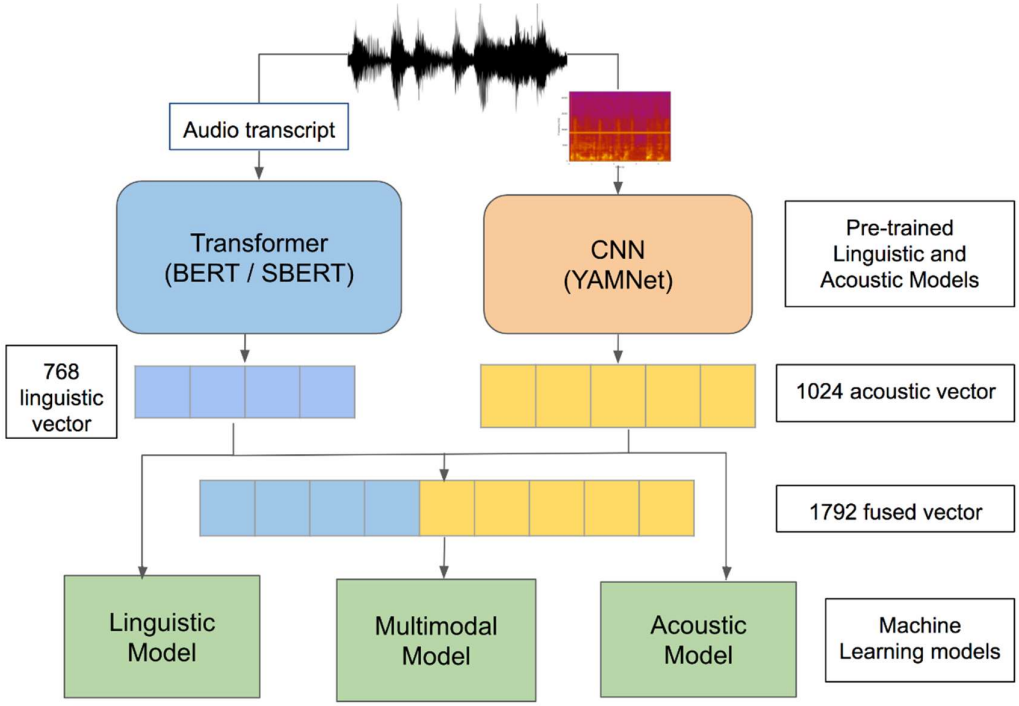

Figure 3: Speech Emotion Recognition

Adding Text to Images for Image Captioning

Image captioning is another well-researched field that is near where AI can fuse vision and language. This involves convolutional neural networks (CNNs) for image processing and recurrent neural networks (RNNs) for text generation. To illustrate, CaptionBot, a part of Microsoft's picture recognition software, uses AI to label images with a caption, or provably based on the image's contents and surrounding context. Accessibility may be an area to be improved with such a multimodal system, because it can generate a written subtitles form to encapsulate the photograph for a visually impaired person. Content moderation has also leverage image captioning, in which algorithms made by AI have been applied to classify and label photographs on social media and to identify unwanted content.

One prominent application of AI systems that simultaneously process visual, auditory, and linguistic data to aid decision-making in the moment is the autonomous vehicle. In order to interpret linguistic data from communication systems, visual data from cameras and LiDAR sensors, and auditory data from surrounding noises, self-driving vehicles use multimodal AI. Take Tesla's Autopilot as an example. It uses voice instructions to communicate with passengers and image recognition to detect road signs and impediments [8].With all this data in one place, we may be one step closer to safe navigation across complicated situations, making judgements in real time, and comprehending passengers' spoken commands. To improve upon the present state of transportation safety, efficiency, and user experience, multimodal AI is essential for autonomous vehicles.

Diverse AI Applications in Healthcare

There have been tremendous advancements in the development of multimodal AI applications within the healthcare business. Artificial intelligence systems may help with medical diagnosis by analysing patient medical information, audio recordings of interviews, and images from imaging tests (such as X-rays and MRIs). Deep learning models improved diagnostic accuracy

and decision-making by combining the sources of data noted above. For example, to predict diseases such as diabetic retinopathy, Google Health's AI model for eye diseases incorporates multimodal input that combines retinal images and patient health data [9]. More and more AI applications improve the ease of healthcare delivery by supporting clinicians to diagnose patients' symptoms using audio analysis or combining audio analysis with natural language processing (NLP).

Potential Challenges and Next Steps

While the advancements in multimodal AI are remarkable, there are still some hurdles. It is very challenging to effectively fuse data from distinct modalities because different types of data exhibit different properties of noise. Any responsible deployment of multimodal AI systems into the real world will also have to account for ethical considerations, such as privacy, bias, and transparency. On the other hand, improvements in areas such as cross-modal alignment and transfer learning may lead to multimodal AI systems that are less resource intensive and more accurate eventually.

Multiple sectors may be facing unforeseeable transformations based on a merger of AI into visual, auditory, and linguistic domains. AI systems to be multimodal learning systems are improving AI's comprehension and engagement with the environment. AI is already showing significant improvements in healthcare diagnostics, autonomous vehicles, image captioning, and emotion classification in customer support. The fusion of many modalities is expected to produce even more radical applications as AI continues to improve, which will significantly augment AI's capability to solve complicated real-world problems.

5.Case study

Case Study 1: Lemonade Insurance – AI-Driven Instant Claims and Underwriting

Domain: Property & Casualty Insurance

Application: Real-time Truth in Underwriting using chatbots and behavior signals

Key Modalities Used: Text (chat bot interaction), Behavior Biometrics (typing speed, click behavior)

Lemonade Insurance has made a big entrance in the property and casualty insurance industry by using AI for both underwriting and claims handling. What is remarkable about Lemonade's solution is the use of Maya, a chat interface where the customer interacts with the AI in a conversational manner. Lemonade also differs from other traditional insurers by utilizing multimodal AI and being able to leverage behavioral biometrics in conjunction with text to assist real-time underwriting decisions.

When a customer comes into a policy application through Maya, our technology guides the user through a digital application that is more like interacting with a conversational virtual assistant than a traditional insurance application. As the applicant types their responses, Lemonade is not only analyzing what the user is saying in the text, but it is looking beyond the text and assessing their typing cadence, response latency, editing behavior, as well as click dynamics. These behavioral biometric signals are embedded measures that are subtle yet highly powerful in communicating user engagement levels, stress levels and possibly deceitfulness.

At the same time, natural language processing (NLP) algorithms assess the text input for reasoning-related consistency, sentiment, and complexity. For example, if a user types very quickly but backspaces often, or pauses and appears uncertain on certain applications, the system registers this information as possible red flags. The two pieces of input – the text input

and the behavioral input – feed into a risk assessment engine in real-time, which relies on machine learning models. This results in a dynamic underwriting process that is faster and likely more accurate than traditional manual underwriting processes. Most applications are approved in less than two minutes, removing the time delay from human review and approval. Furthermore, this process has proven useful in the early detection of false behavior, which decreases claims fraud and costs. From the customer experience perspective, this learning path represents minimized friction. Customers do not have to provide unnecessary documentation, nor do customers have to wait for a human underwriter to review their file. They enjoy instant insights and immediate policy issuance in many cases. The case of Lemonade, with its combination of text analysis and behavioral intelligence, demonstrates how multimodal AI can implement smarter decision-making in the insurance space. Their case serves as a case study for how digital-first insurers build fast, fair, and fraud-free underwriting systems leveraging insurance technology.

Case Study 2: Shift Technology – Voice & Behavioral Biometrics for Fraud Detection

Domain: Claims Underwriting and Fraud Analytics

Application: Fraud detection through call center audio and behavioural signals

Key Modalities Used: Voice (speech analysis), Text (transcripts), behavioural biometrics (caller metadata)

Shift Technology is a leading provider of AI-fueled innovations for the insurance space, especially fraud analytics. Its greatest advancement to date is the multimodal AI technology that identifies fraudulent behavior in underwriting and claims by utilizing voice data, text-based transcripts, and behavioural metadata. This is particularly beneficial for insurers who have deep reliance on call centres as part of their customer engagement and claims ecosystem.

The process commences with a phone call between a customer and an insurance representative. Shift's AI engine encapsulates and analyzes the call in real time (active behavior). The AI detects vocal markers, such as tone, pitch, peaks of stress, vocal tension and speech patterns to assess the emotional state of the caller. These vocal markers are important to understanding the behavioral risk profile of the caller. In terms of claims, vocal markers can indicate when a speaker may be uncomfortable, deceitful or hesitant to answer questions. These key indicators are commonly associated with the likelihood of experiencing fraud. In concert with audio analysis, the AI processing system uses Natural Language Processing algorithms to parse transcripts of calls. NLP algorithms utilize semantics, syntax and phonetics, including evasive language, word choice, emotional sentiment and other anomalies in language and conversation style.

Moreover, Shift Technology employs behavioral biometrics based on caller metadata. The metadata is rich in information; frequency and time of day of the calls, length of interaction, patterns in prospective customers on different claims, etc. All of this information is analyzed for behavioral patterns, meaning that the AI is set to look for behavioral anomalies. For example, a higher-than-average frequency of calls related to similar issues, or changed behavior by users from previous calls, e.g., where once a caller hesitated to provide further information to the AI, the next time were very happy to share (sudden changes in behavior, etc.). All of this is put together with the audio input of the call and the written input (text) into a composite risk score that is fed into a unified at-the-moment knowledge base that holds the important information to include flagged behaviors for potential risk or fraudulent activity. This

composite solution makes clear to a human user, analyst or underwriter, there is something they need to think about, identify, or action. Thus, issues are flagged prior to any payment affecting the insurer. Insurers that use Shift's technology reports improved fraud detection upwards of 75% early in the claims process, allowing them to prevent loss of funds from already paid out claims, and relieving frustration with post-claims investigations and having to administratively review a letter from the governing office. Further, Shift technology has lowered the incidence of false/positive clients, meaning legitimate customers do not get treated unfairly.

In conclusion, Shift Technology's multimodal AI shows how diverse data sources can improve the accuracy and speed of fraud detection systems. In addition, this example identifies the change needed so that human-aware artificial intelligence becomes a more meaningful part of intelligent underwriting and risk management in the near future.

Case Study 3: Zest AI – Behavioral Data Fusion in Credit Underwriting

Sector: Credit Underwriting (Applicable across Life/Health insurance Risk Models)

Application: Behavioral AI for real-time loan approval and risk assessment

Key Modalities: Text (application data), Behavioral Biometrics (interaction patterns), Historical Data Fusion

Zest AI has disrupted the credit underwriting landscape by introducing systems that evaluate not only the what, but also the how, in what applicants submit. Zest AI's main domain is credit lending but its innovation is easily transferable to life and health insurance underwriting, where assessment of risk and fairness is paramount. Historically credit and insurance risk models have primarily relied on historical financial data such as credit scores, income statements and debt ratios, features that intentionally or unintentionally exclude applicants from underserved or marginalized populations which may have insufficient financial histories. Zest AI's innovation is new multimodal AI framework that incorporates behavioral biometrics, textual data, and historical analytics to build a more inclusive and more accurate risk profile.

The system unobtrusively collects behavioral biometric data including faithful mouse movements, typing cadence, field dwell time, and scrolling patterns. For example, if an applicant's mouse lingers in a field or repeatedly edits it, this would be interpreted as the applicant either hesitating because they don't know what to do, or they are catching themselves using jargon they don't understand. If all of the applicant's interactions were confident, consistent, and unchanging, they would be interpreted as having a higher literate level in personal finance or that they were being honest about the information they were providing. These behaviors are combined with text (what they have typed) and historical financial data to create a richer applicant profile. Using advanced machine learning models, the system can determine the relationships between these behavioral features and loan performance, which enables accurate predictions, even without a conventional credit history.

This multidimensional integration has two significant advantages. First, it identifies low-risk individuals who may be marginalized by traditional scoring—thereby increasing financial inclusion. Second, the process significantly cuts approval times, with most applications being processed in minutes, not days. Further, Zest AI provides model explainability and fairness features to help communicate with the constant changing nature of the regulations in finance and insurance. Specifically, models are audited for bias, and then explained with explainable AI (XAI) tools to provide transparency that explains automated decision-making.

This technique has a lot of potential in life and health insurance. Behavioral data could show cognitive stress, intent, and digital fluency, which would help to improve risk evaluation beyond a simple static questionnaire. Zest AI's case study illustrates how behavioral data fusion can reshape underwriting, improve access, and make automated systems more ethical, responsive, and intelligent.

6. Multimodal Learning using AI-Incorporated Visual, Auditory, and Linguistic Applications

Machine learning that includes vision, audio, and language could change many industries. In healthcare, it can be part of disease diagnosis and personalized treatment. It also aids in improving navigation and safety for self-driving vehicles. Multimodal AI also enhances robotics, education, entertainment, and customer service by generating interactive, dynamic, and personalized experiences.

AI Applied to Healthcare: Medical Diagnosis and Personalized Treatment.

The incorporation of AI with other sensing modalities like vision, hearing, and language has improved healthcare, particularly with accuracy in diagnosis and treatment. Medical AI systems take visual information from imaging technologies such as X-rays and MRIs, and language information from patient records and vocal signals from doctor-patient conversations for therapeutic decisions. Google's AI model for diabetic retinopathy has combined retinal scans with medical histories to better identify early signs of eye diseases [10]. On the other hand, IBM's Watson Health has combined visual information from diagnostic images with natural language processing to make recommendations for personalized treatment plans based on unstructured medical words in documents, such as notes from physicians. Collectively, these AI systems may improve patient outcomes while decreasing

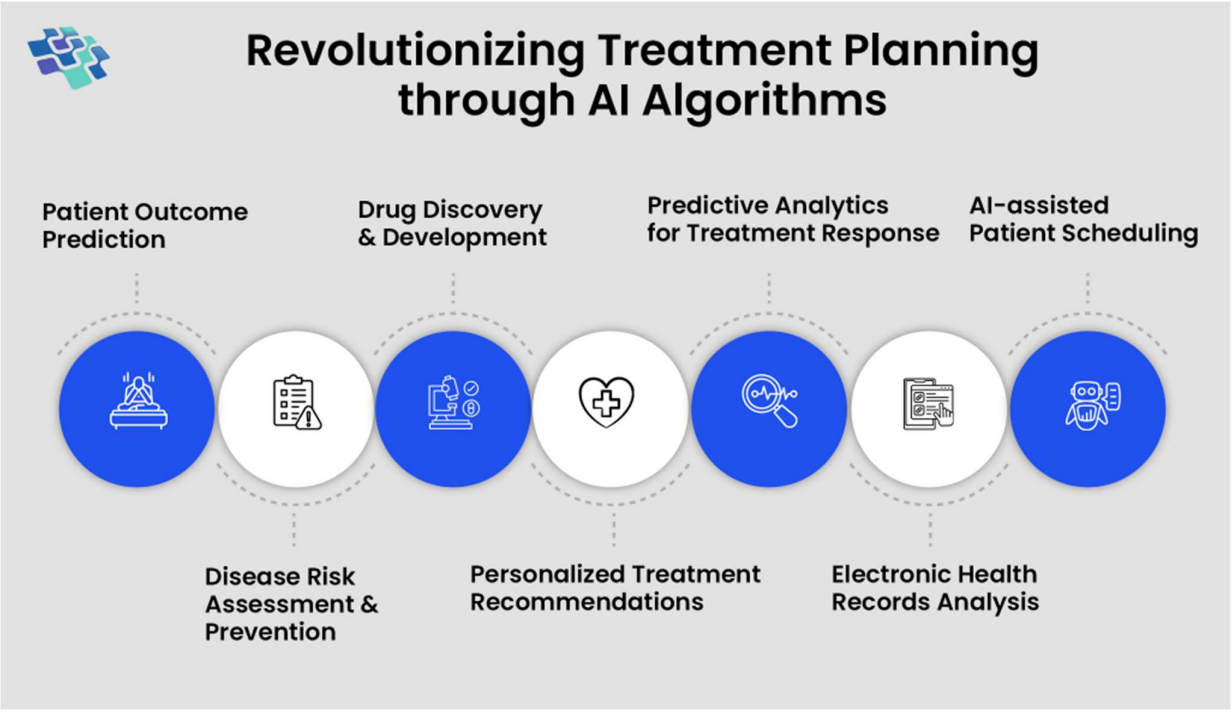

Figure 4: AI is transforming treatment plans

"Autonomous Vehicle: Navigation and Safety"

Vision, audio, and language AIs play a critical role in designing autonomous cars and making real-time decisions to quickly adjust car material to enhance safety features. Tesla and Waymo's autonomous cars use multimodal AI systems to make sense of their complicated environments, based on visual input (the camera images and LiDAR data input), audio input (environmental sounds), and language input (the instructions of passengers). For example, Autopilot relies on visual inputs from cameras to recognize road signs, pedestrian traffic, and car traffic; when coupled with instructions from the driver, it can alter navigation or, in some situations, the operation of the car (Goodall, 2014). The capacity for autonomous cars to adjust to new road situations, and improve safety features lie in this multimodal AI system, mapping and interpreting both of the environment and human input. As multimodal AIs continue to grow and develop, edge cases lead to newfound levels of autonomy for vehicles and more accurate decision-making for autonomous cars.

In the field of education, adaptive learning systems use multimodal AI

Multimodal AI is also being used in the educational sector to help promote a student-centered learning experience. Knewton and DreamBox are two AI-powered educational tools that utilize multimodal learning algorithms to personalize courses based on student responses and successes. These multimodal learning systems understand responses based on text, audio, and visual data which allows it to evaluate understanding, report on it, and directly give feedback at an individualized level. An example, DreamBox, utilizes an artificial intelligence (AI) system to adjust the challenge of questions depending on visual, auditory, and language-based data that are drawn from a student's interactions on the platform [11]. DM and multimodal AI benefit students by creating more enjoyable and personalized lessons, resulting in better retention and success.

Content recommendation systems in the entertainment and media industries

As just two examples of content-focused entertainment platforms, Netflix and YouTube are familiar with using AI for multimodal learning in their recommendation algorithms. Both platforms recommend the most relevant films, TV shows, and other media by taking into account viewers' various visual/aural/linguistic input choices, and then tracking their actions and preferences. Covington, Adams, and Sargin described how YouTube makes personalized suggestions based on different types of visual information (thumbnails and content), text (user comments, titles, and descriptions), and audio (voice or music in video content). YouTube and Netflix use multimodal systems to represent a viewer's behaviour, preferences, and interactions to suggest videos they are more likely to watch. An important reason why these systems can provide a more accurate recommendation is because they present relevant media that increases viewer enjoyment and engagement. For example, the recommendation engine on Netflix uses a multimodal approach to predict likely viewer interest in certain genres or types of content on their platform by examining viewer ratings, interactions, and viewing history.

The field of advanced robotics known as human-robot interaction (HRI)

Multimodal AI marks a point in the progression of HRI in comparison to robots. Robots such as Pepper by SoftBank, and Spot by Boston Dynamics can combine visual, audio, and linguistic inputs into one experience and can act actively human-like when interacting with people. For instance, robot Pepper collects input from face detection and voice processing in order to perceive a person's emotional state. For robots to provide effective communication channels with people, which is paramount in retail, healthcare, customer service etc, robots must

integrate technology. As per [12] robots that combine vision, hearing and language perform better in understanding human emotions and intents. Which leads to more human-like engagement.

AI, vision, audio and language integration represent just some of the industries that will be enhanced by this type of AI integration including healthcare, autonomous cars, education, entertainment, and robotics. By integrating these approaches, AI systems can interpret and respond to complex challenges with more accuracy and individualisation. Future integrations of multimodal AI are anticipated and will result in even more widespread integration, providing myriad opportunities for growth and innovation which will lead to enhancement and advancement in many sectors.

7. Conclusion

The utilization of multimodal AI which combines voice, text, and behavioral biometrics in real-time underwriting is a disruptive shift in the approach to risk assessment and customer interaction in the insurance industry. Traditional underwriting practices using standardized and static data and forms, will be outstripped by the realities of a digital-first, data-rich world. Multimodal AI will provide a richer, more dynamic approach to the assessment and powers of examination of a file through contextual behavioral biometrics capturing real-time human signals and situational patterns which were not possible in previous workflows. The incorporation of streams of multiple human data allows insurers to create richer and more accurate applicant profiles from a 360-degree perspective. Faster decision-making not only improves the experience for the consumer but gives stronger fraud detection bases, and improved personalization. Rather than simply judging applicants from their prior propositions or static credit based histories, multimodal AI proposals translate the way that someone communicates or interacts, thereby revealing their intention, truthfulness, and risk profile in real time. The case studies of Lemonade Insurance, Shift Technology, and Zest AI evidenced clearly the utility of applying intelligent systems like these. Lemonade's chatbot based onboarding included behavioral signals along with NLP which expedited the issuing of policies to minutes and allowed mitigation of exposure. There are considerable advantages of multimodal AI systems! Speed translates to everything being done in real-time and decisions made instantaneously. Accuracy is improved through the cross referencing and cross-validation of various types of mode inputs, security improved through the early detection of aberrant or suspicious trends and behaviors. Personalization ensures that risk models can be customized to the individual behavioral thumbprint of the applicant which creates a level of humanity or empathy in the underwriting process. And importantly, the development of these systems are framed in the emerging principles of explainable artificial intelligence (XAI) and ethical design. It is vital to the continued acceptance, trust and accountability of these technologies in the regulatory culture that organizations be publicly transparent on the logic of decisions, auditability, and user privacy. In summary, multimodal question and answer programming signifies a shift from passive, archaic, paper-based underwriting to active, intelligent, human-aware risk assessment. A process of improvement and replacement not an add-on.

Reference

[1]. Dosovitskiy, A., & Brox, T. (2016). Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(9), 1734-1747. https://doi.org/10.1109/TPAMI.2015.2489723

[2]. Esteva, A., Kuprel, B., & Novoa, R. A. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.
[3]. Ganaie, M. A., Zhang, Y., & Hu, B. (2020). Speech and text-based multimodal learning for predicting mental health. Proceedings of the IEEE International Conference on Big Data, 2529–2536.
[4]. Goodall, N. J. (2014). Machine ethics and automated vehicles. In Road Vehicle Automation (pp. 93-102). Springer Vieweg, Berlin, Heidelberg.
[5]. Gulshan, V., Peng, L., & Coram, M. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402- 2410.
[6]. Hori, T., & Hori, C. (2020). Speech-to-Text and Textto-Speech Systems: Combining NLP and Audio for Deep Learning Applications. ACM Computing Surveys, 53(3), 1-33. https://doi.org/10.1145/3354245
[7]. Huang, L., Xu, W., & Liu, X. (2016). Visual information extraction for multimodal sentiment analysis. Journal of Machine Learning Research, 17(1), 3213–3235.
[8]. Kiros, R., Salakhutdinov, R., & Hinton, G. (2015). Multimodal Deep Learning. Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS), 1, 1-9. https://doi.org/10.5555/2969033.2969036
[9]. Kumar, A., Malik, P., & Singh, A. (2020). Multimodal conversational agents: Current challenges and future directions. ACM Computing Surveys, 53(2), 1-27.
[10]. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
[11]. Mythily, D., Renila, R. H., Keerthana, T., Hamaravathi, S., & Preethi, P. (2020). Iot based fisherman border alert and weather alert security system. International Journal of Engineering Research & Technology (IJERT).
[12]. Lu, J., Yang, Z., & Qiao, Y. (2020). Learning joint representations for multimodal fusion. IEEE Transactions on Neural Networks and Learning Systems, 31(8), 2541–2553.