Multimodal Medical Data Learning Approaches for Digital Healthcare

Oleh Basystiuk[a], Nataliia Melnykova[a]

[a] *Department of Artificial Intelligence, Institute of Computer Science and Information Technologies, Lviv Polytechnic National University, Lviv, 79000, Ukraine*

Abstract

The integration of multimodal data has emerged as a game-changing strategy in advancing smart healthcare, allowing for a holistic comprehension of patient health and tailored treatment strategies. This exploration delves into the journey from raw data to insightful wisdom, emphasizing the fusion of various data modalities, notably in CT scans or retinal photographs, to drive smart healthcare innovations.

Within this review, we comprehensively examine the fusion of diverse medical data modalities, aiming to unlock a deeper understanding of patient health. Our focus spans various fusion methodologies—from feature selection to rule-based systems, machine learning, deep learning, and natural language processing. Furthermore, we explore the challenges inherent in fusing multimodal data in healthcare settings.

The central focus revolves around determining the most efficient and accurate approach, crucial for future research endeavors in Ukrainian language audio-to-text conversion systems. The goal is to ascertain the most effective strategy that will serve as the foundation for further advancements in this domain.

Keywords 1

Speech-to-Text, speech recognition, video recognition, audio recognition, multimodal data, machine learning, deep neural network, hybrid approaches

1. Introduction

In the rapidly evolving landscape of smart healthcare, where innovation and data-centric methodologies are reshaping the industry, the convergence of multimodal data stands out as a pivotal force. This paper delves deep into the realm of multimodal medical data fusion, offering a thorough investigation into how diverse data streams converge to generate meaningful insights. It navigates through the intricate process—from initial data collection to translating it into actionable intelligence—depicted through the detailed four-level pyramid showcased in Figure 1.

The Data Information Knowledge Wisdom (DIKW) model serves as a conceptual roadmap illustrating how data evolves into profound wisdom [1]. It elucidates a transformative journey where raw data undergoes a metamorphosis into meaningful information, knowledge, and eventually, wisdom—empowering informed decision-making and adept problem-solving [2]. This model recognizes the inadequacy of raw data in driving insights and actions; it underscores the necessity to process, structure, and contextualize data to extract invaluable information. This synthesized information converges with existing knowledge, fostering an understanding that begets knowledge itself. This accrued knowledge becomes a practical tool for making informed decisions and navigating intricate challenges, ultimately culminating in the attainment of wisdom.

IDDM'2023: 6th International Conference on Informatics & Data-Driven Medicine, November 17 - 19, 2023, Bratislava, Slovakia
EMAIL: oleh.a.basystiuk@lpnu.ua (A. 1); nataliia.i.melnykova@lpnu.ua (A. 2);
ORCID: 0000-0003-0064-6584 (A. 1), 0000-0002-3257-3677 (A. 2);

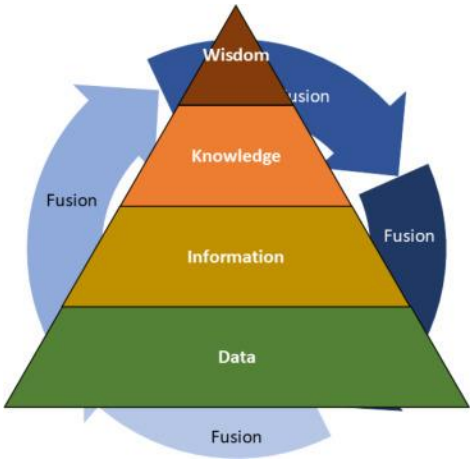

Figure 1: Multimodal data fusion conceptual model

This study delves into diverse methodologies encompassing feature selection, machine learning, deep learning, and natural language processing, aiming to fuse multimodal medical data effectively. Additionally, it confronts a spectrum of challenges, spanning data quality, privacy, security, processing, analysis, clinical integration, ethics, and result interpretation. Emphasizing the transformative capacity of multimodal medical data fusion, this paper acts as a springboard for future research and advancements in smart healthcare. It lays the foundation for enhancing patient care outcomes and propelling personalized healthcare solutions to the forefront of medical innovation [3].

The key contributions of this paper are:

- Utilizing and adapting the established DIKW conceptual model to delineate the evolution from data to information to knowledge to wisdom, specifically in the domain of multimodal fusion for intelligent healthcare.
- Current techniques for representing multimodal data in applications utilizing machine learning methodologies, based on Ukrainian language-based dataset.
- Proposing a comprehensive workflow for handling multimodal medical data, such as video, audio, and textual sources, using the Sequence-to-Sequence model flow.
- Analyzing the challenges and proposing solutions for the algorithmic time complexity associated with multimodal data handling, focusing on video, audio, and textual medical data, while aligning with the proposed approach, to steer future research paths.

The following sections of this paper are structures in the following order: Section 2 overview existing techniques of multimodal data representation in applications, based on machine learning approaches. Section 3 propose multimodal handling flow with encompassing video, audio, and textual medical data, using Sequence-to-Sequence model flow. Section 4 introduces a comparison of multimodal handling algorithm time complexity, based on video, audio, and textual medical data. Section 5 and Section 6 stands for discussion and comparison and evaluation the results.

2. Related works

This research project will use a qualitative research approach. Data will be collected through a review of relevant literature, case studies, and interviews with experts in the field. The data will be analyzed using thematic analysis to identify key themes related to the use of multimodal and artificial intelligence approaches in the distance education process, namely to recognize video streams of completed course assignments [4].

Current AI applications in medicine have typically focused on specific tasks using singular data sources, like a CT scan or retinal photograph. However, clinicians rely on a multitude of data types and modalities to make diagnoses, evaluate prognosis, and determine treatment plans [5]. Moreover,

existing AI assessments capture only a snapshot in time, failing to perceive health as an ongoing continuum.

The potential for AI models extends far beyond these limitations, envisioning their ability to leverage diverse data sources, including those beyond most clinicians' scope [8]. Multimodal AI models integrating video, imaging, text and audio clinical medical data promise. They stand to bridge this gap by enabling personalized medicine, real-time pandemic surveillance, digital clinical trials, and virtual health coaching on an unprecedented scale [6, 7].

This review examines the vast potential of multimodal datasets in healthcare, emphasizing the transformative possibilities they offer. By incorporating audio and video data into this multimodal landscape, leveraging machine learning techniques, medical research could attain a new level of depth and accuracy [9]. For instance, integrating video data from patient interactions or audio data from diagnostic interviews could enrich these models, leading to more holistic and precise healthcare solutions.

For video content processing, the application and categorization of methods are proposed, including sequential comparison, clustering-based global comparison, and event/object-based methods. The most valuable compare and contrast techniques include sequence finding, classification, frame decoding, and evaluation feature detection. The most optimal use of methods is based on artificial intelligence and machine learning, where deep learning methods will be more effective than conventional methods.

However, this integration presents significant challenges. Nonetheless, with innovative strategies and advancements in machine learning, overcoming these hurdles becomes increasingly feasible. The potential benefits of multimodal data utilization in medical research are vast, heralding a paradigm shift toward more comprehensive, personalized, and effective healthcare solutions [10].

The research project is expected to provide a system for evaluating and predicting student success based on feedback. It will identify the different approaches used, their effectiveness, and the challenges and opportunities associated with their use. It is expected that the implementation of the proposed approaches will significantly improve information systems for evaluating the quality of learning outcomes, which will be used to analyze video content and textual content of answers. The project will create an approach for providing relevant feedback, based on which suggestions for improving the courses and evaluating their overall effectiveness. In addition, to influence the success of students during the distance learning process and the quality of their assimilation of subject materials.

Our roadmap includes devising an expert-level system tailored specifically for hybrid language translation within the medical sphere. By incorporating multimodal data, including audio and video, alongside machine learning techniques, we aim to pioneer a transformative approach that could redefine diagnostic accuracy, treatment protocols, and overall healthcare practices. This pursuit represents an innovative frontier poised to revolutionize medical research and application.

3. Methods

One standout deep learning model, known as Seq2Seq (sequence-to-sequence), has demonstrated remarkable proficiency in tasks like machine translation and text summarization. These models operate on the principle that the decoder's attention layers can access only preceding words in the input sequence, while the encoder's attention layers can access the entirety of the original phrase, fostering connections that enable an RNN (Recurrent Neural Network) to retain and replicate an entire sequence of reactions to a stimulus [11].

Initially, the sequence flows through the encoder, comprising RNNs, culminating in a final embedding at its conclusion. Subsequently, the decoder utilizes this embedding to predict subsequent sequences. This process involves using prior hidden states to forecast the succeeding instances within the sequence, as illustrated Sequence-to-Sequence model flow showcased in Figure 2.

To optimize accuracy and time efficiency tailored to our specific context, an in-depth investigation is essential, exploring the methods delineated earlier. In our prior research, the Sequence-to-Sequence approach based on Recurrent Neural Networks emerged as the most effective method, particularly when examining the libraries utilized in constructing machine learning methodologies. Notably, TensorFlow, Keras, and PyTorch stand out as the most widely employed machine learning libraries in RNN-based

language translation methods. Model play a crucial role in enhancing and effectiveness of sequence prediction and language translation tasks within our research domain [12, 13].

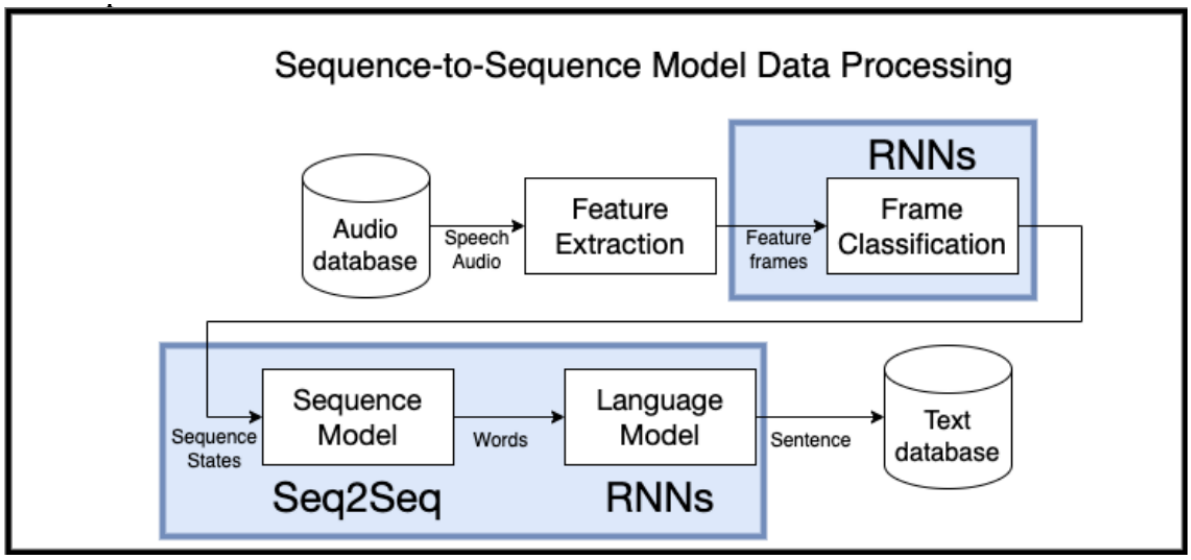

Figure 2: Sequence-to-Sequence model flow

4. Results

The research project is expected to provide a system for interconnect different data sources, from multimodal data approach, such as audio, video and text into our structure system, with ability to reinforce and aggregate these data. Advancements in speech-to-text technology are rapidly progressing, opening avenues for scaling its application beyond its current horizons. This expansion is crucial not just for accuracy and reliability but also for the future credibility and coherence of research endeavors across diverse fields. The groundwork laid by this research serves as a cornerstone for subsequent investigations, setting the stage for future platforms to tackle multifaceted challenges.

The landscape of health data is diverse, posing multifaceted challenges in gathering, linking, and annotating these multidimensional datasets. These medical datasets vary across several dimensions—sample size, depth of phenotyping, follow-up intervals, participant interactions, heterogeneity, standardization, and data linkage. While advancements in science and technology facilitate data collection and phenotyping, striking a balance among these dataset features remains a challenge. For instance, while larger sample sizes are ideal for training AI models, achieving deep phenotyping and sustained longitudinal follow-up escalates costs significantly, making it financially impractical without automated data collection methods.

In the realm of medicine, current AI applications tend to focus on specific tasks using singular data sources like CT scans or retinal photographs. This contrasts starkly with clinicians who rely on a diverse array of data sources and modalities to diagnose, forecast outcomes, and devise treatment plans. Moreover, existing AI assessments typically offer singular snapshots, capturing a moment in time rather than perceiving health as an ongoing continuum.

Biomedical data often grapples with a common issue: a notable prevalence of missing information. Although excluding patients lacking data prior to training is feasible in certain scenarios, doing so might introduce selection bias when external factors contribute to these gaps. Consequently, employing statistical methods like multiple imputation becomes a preferable approach to tackle these voids. Imputation stands as a crucial preprocessing stage in numerous biomedical disciplines, spanning from genomics to clinical data analysis.

However, in theory, AI models possess the potential to harness all available data sources, including those beyond the reach of most clinicians, like genomic medicine. The evolution of multimodal AI models that amalgamate data across various modalities—spanning biosensors, genomics, epigenomics, proteomics, microbiomes, metabolomics, imaging, textual, clinical, social determinants, and environmental data—holds the promise of narrowing this gap. These advanced models pave the way

for diverse applications, including handling multimodal medical data, such as video, audio, and textual sources, using the Sequence-to-Sequence model flow showcased in Table 1.

Table 1
Comparison of multimodal handling algorithm time complexity

Case 1	Case 2	Case 3
0.7020	0.4150	0.5800
0.8026	0.4375	0.5975
0.8163	0.4208	0.5308
0.8441	0.3920	0.6200
0.7511	0.5644	0.5644
0.8806	0.6100	0.6050
0.7507	0.5835	0.5835
0.6514	0.5665	0.5665
0.6672	0.4520	0.5520
0.7535	0.5390	0.5290

5. Discussion

The research project is expected to provide a system for interconnect different data sources, from multimodal data approach, such as audio, video and text into our structure system, with ability to reinforce and aggregate these data. Advancements in speech-to-text technology are rapidly progressing, opening avenues for scaling its application beyond its current horizons. This expansion is crucial not just for accuracy and reliability but also for the future credibility and coherence of research endeavors across diverse fields. The groundwork laid by this research serves as a cornerstone for subsequent investigations, setting the stage for future platforms to tackle multifaceted challenges [14].

In our preliminary steps, we delineate specific fields and tasks, anticipating the demands future platforms will confront. Additionally, we delve into analyzing neural network training techniques that extend beyond machine translation applications. Our findings from Section 3 highlight the potential of a hybrid approach employing recurrent networks within a sequence-to-sequence model. This approach holds promise for yielding optimal outcomes, coupling high time efficiency with commendable accuracy rates.

The utilization of recurrent neural networks (RNNs) has garnered significant traction, particularly in audio-to-text translation. Foreseeing advancements in this domain, we anticipate substantial enhancements in the near future. Harnessing multimodal data—integrating not only audio but also video data—alongside machine learning techniques stands as a promising frontier in medical research. This synergy could revolutionize diagnostic accuracy, treatment methodologies, and overall healthcare practices, paving the way for groundbreaking advancements in the field [15].

6. Conclusion

In conclusion, our exploration into multimodal Ukrainian medical data showcased the immense potential of integrating audio and video data within machine learning-based systems, particularly employing RNNs (Recurrent Neural Networks). This convergence offers a transformative pathway for reinforcing medical research endeavors, elevating diagnostic accuracy, treatment methodologies, and healthcare practices.

By harnessing the power of RNNs alongside multimodal data, we've unveiled new horizons for enhanced understanding and analysis within medical research. The fusion of audio and video data within this framework promises a more comprehensive view of patient health, potentially revolutionizing diagnostic precision and personalized treatment plans. Increasing accuracy of handling

medical data and cutting medical time to handle the paperwork after medical research, such as retinal photograph, or scans based on computed tomography (CT) results.

Moving forward, leveraging these machine learning approaches in multimodal Ukrainian medical data sets the stage for groundbreaking advancements, paving the way for innovative applications and more effective healthcare solutions. The fusion of audio, video, and textual data within RNN frameworks not only strengthens the foundations of medical research but also offers a promising avenue for the future of healthcare practices.

Funding Statement: This research is funded by the EURIZON Fellowship Program: "Remote Research Grants for Ukrainian Researchers", grand № 138.

7. References

[1] Smit, A. et al. CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing 1500-1519 (2020).

[2] Willemink, M. J. et al. Preparing medical imaging data for machine learning. Radiology 295, 4–15 (2020).

[3] M. Havryliuk, et. al., "Check for updates Interactive Information System for Automated Identification of Operator Personnel by Schulte Tables Based on Individual Time Series", Advances in Artificial Systems for Logistics Engineering III 180, 372.

[4] Cirillo, D. et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digit. Med. 3, 81 (2020).

[5] Damask, A. et al. Patients with high genome-wide polygenic risk scores for coronary artery disease may receive greater clinical benefit from alirocumab treatment in the ODYSSEY OUTCOMES trial. Circulation **141**, pp. 624–636 (2020).

[6] Vyas, D. A., Eisenstein, L. G. & Jones, D. S. Hidden in plain sight: reconsidering the use of race correction in clinical algorithms. N. Engl. J. Med. 383, 874–882 (2020).

[7] O. Basystiuk, N. Melnykova, Multimodal Approaches for Natural Language Processing in Medical Data, IDDM 2022 Informatics & Data-Driven Medicine, pp. 246-252.

[8] Z. Rybchak, et. al., Analysis of computer vision and image analysis technics, ECONTECHMOD: an international quarterly journal on economics of technology and modelling processes, Lublin, Poland, 2017, pp. 79-84.

[9] M. Havryliuk, I. Dumyn, O. Vovk. Extraction of Structural Elements of the Text Using Pragmatic Features for the Nomenclature of Cases Verification. In: Hu, Z., Wang, Y., He, M. (eds) Advances in Intelligent Systems, Computer Science and Digital Economics IV. CSDEIS 2022. Lecture Notes on Data Engineering and Communications Technologies, 2023, vol 158. Springer, Cham. https://doi.org/10.1007/978-3-031-24475-9_57.

[10] Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. Nat. Med. 28, 31–38 (2022).

[11] A. Esteva, A. et al. Deep learning-enabled medical computer vision. NPJ Digit. Med. 4, 5 (2021).

[12] J.N., Falcone, G.J., Rajpurkar, P. et al. Multimodal biomedical AI. Nat Med 28, 1773–1784 (2022). https://doi.org/10.1038/s41591-022-01981-2

[13] Nataliya Shakhovska, et. al. "Big Data analysis in development of personalized medical system", The 10th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN), 160, 229-234. (2019)

[14] Yaroslav Tolstyak, Myroslav Havryliuk "An Assessment of the Transplant's Survival Level for Recipients after Kidney Transplantations using Cox Proportional-Hazards Model", Proceedings of the 5th International Conference on Informatics & Data-Driven Medicine, Lyon, France, November 18 - 20, CEUR-WS.org, 2022. pp. 260-265.

[15] Kang, M., Ko, E. & Mersha, T. B. A roadmap for multi-omics data integration using deep learning. *Brief. Bioinform.* **23** (2022).