USING LLMS FOR GENERATING TEXTUAL DESCRIPTIONS OF MEDICAL DIAGNOSTICS BASED ON MULTIMODAL DATA

Minukhin Serhii, Rudoi Valerii

Department of Systems Engineering
Kharkiv National University of Radioelectronics,
Ukraine

E-mail: serhii.minukhin@nure.ua,
valerii.rudoi@nure.ua

Abstract

This research investigates an approach to automated generation of textual medical reports based on multimodal data using Large Language Models (LLMs). A two-stage architecture is proposed, consisting of a feature extraction module for medical images via Convolutional Neural Networks (CNNs) and a language generation module that integrates latent visual features with numerical clinical parameters of the patient. A combined loss function, encompassing both pathology classification and text generation, was applied for model training. Experimental evaluation on a multimodal dataset demonstrated the LLM's capability to produce grammatically correct and clinically relevant conclusions. A correlation matrix confirmed expected relationships between clinical parameters and pathology probability, ensuring model stability. The results indicate the effectiveness of integrating LLMs into multimodal analysis for automated medical reporting in a technical and computational framework.

Introduction

Modern medical information technologies increasingly adopt artificial intelligence methods to enhance diagnostic accuracy and efficiency. One promising direction is the automation of textual medical report generation based on the combination of heterogeneous (multimodal) data, including imaging, physiological signals, laboratory tests, and clinical records. However, generating grammatically correct and clinically adequate text requires a high level of semantic abstraction, which traditional deep learning architectures can only partially provide [1-3].

Large Language Models (LLMs) enable the creation of a cognitive layer in diagnostic systems capable of transforming neural network analysis results into natural language descriptions. Their integration with hybrid architectures, combining Convolutional Neural Networks (CNNs) or Recurrent/Long Short-Term Memory networks (RNNs/LSTMs) for medical image and numerical data analysis, allows the generation of coherent and interpretable diagnostic reports [4].

The objective of this research is to develop and experimentally validate a technical approach for generating textual medical descriptions of diagnostic results from multimodal data using LLMs.

Methodology

A two-stage architecture was developed for the experiment, comprising a feature extraction module and a language generation module. The feature extraction module is implemented as a CNN that converts medical images into a latent feature vector [5]:

$$f = CNN(I) \in R^n, \tag{1}$$

where I – input medical image, $CNN(I)$ – convolutional neural network function extracting image features and f – latent feature vector encoding key visual characteristics for computational analysis.

The language generation module is based on a LLM that receives both the latent feature vector and numerical clinical parameters such as glucose level, blood pressure, age, and other relevant indicators. The LLM then generates a textual description of the patient's condition [6]:

$$T = LLM([f, x]) \in R^n, \tag{2}$$

where T – generated textual report and $[f, x]$– concatenated vector of image features and numerical clinical parameters.

The clinical vector x is composed of individual numerical measurements:

$$x = [x_1, x_2, ..., x_m], \tag{3}$$

where x_m – individual numerical parameters.

Integrating vector x with latent image features provides the LLM with a full multimodal context, enabling the generation of accurate and coherent diagnostic descriptions.

To align the feature spaces of image and numerical data, a multimodal transformation is applied [7]:

$$z = W_f f + W_x x + b, \tag{4}$$

where W_f, W_x – weight matrices for the different modalities and b – bias vector.

The resulting vector z is used as the context for LLM text generation. The loss function for training the combined system is defined as [8]:

$$L = \alpha L_{\text{csl}} + (1 - \alpha) L_{txt}, \tag{5}$$

where L_{csl} – pathology classification loss (binary cross-entropy), L_{txt} – text generation loss (cross-entropy between predicted and reference tokens) and α – balancing coefficient.

For demonstration, the COVID-19 Chest X-ray Dataset was used, combining visual features and numerical patient characteristics [9]. Table 1 provides sample multimodal data entries with LLM-generated descriptions.

Table 1. Sample Multimodal Dataset Entries

ID	Glucose Level (mmol/L)	Blood Pressure (mmHg)	Image Class	Pathology Probability	Age Index	LLM Description
P001	6.4	125/80	Mild infiltration	0.18	0.42	Lungs show no pathology; minor shading is non-critical
P002	8.7	145/90	Patchy structure	0.73	0.58	Possible pneumonia; follow-up recommended in 48 hours
P003	4.9	120/75	Homogeneous structure	0.05	0.33	Normal lung tissue
P004	11.2	160/100	Dense shadow in lower lobe	0.89	0.66	Significant density in lower lung regions – further examination needed

The data preprocessing pipeline was implemented using Pandas for tabular data integration and OpenCV for medical image normalization and resizing. Each image was standardized to a 224×224 pixel input for the CNN backbone. The model training process used the AdamW optimizer with a learning rate scheduler and a batch size of 32. Training was conducted in the Jupyter Notebook environment with GPU acceleration on Google Colab. For reproducibility, the random seeds were fixed, and model checkpoints were saved using PyTorch Lightning. The language model was fine-tuned via Hugging Face Transformers with tokenization provided by the Byte Pair Encoding (BPE) algorithm [10]. These tools collectively enabled the synchronization of visual and textual data streams, ensuring robust multimodal integration [11].

A correlation matrix for numerical clinical parameters and pathology probability was constructed to evaluate interrelationships and assess the influence of key factors on diagnostic outcomes (Fig. 1).

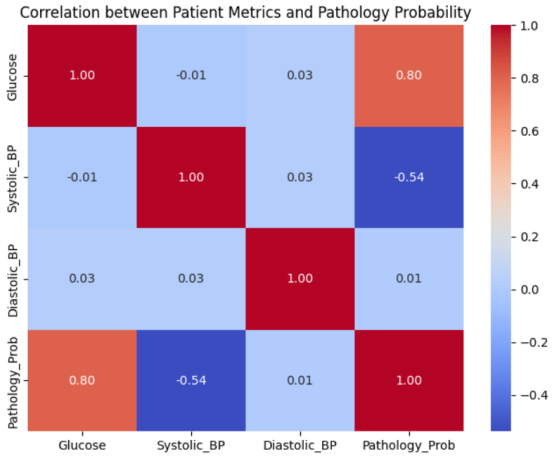

Fig. 1. Correlation matrix of clinical parameters and pathology probability

Analysis revealed a strong positive correlation between glucose level and pathology probability (0.80), consistent with clinical observations and confirming that elevated glucose reliably indicates pathological processes in respiratory or systemic disorders. This supports the view that hyperglycemia drives inflammation, tissue damage, and immune impairment, making it a key predictive factor in disease progression. The high correlation demonstrates the model's capacity to capture physiologically meaningful links between biochemical and image-derived features, validating the robustness of multimodal integration within the proposed LLM-based diagnostic system.

Systolic blood pressure showed a moderate negative correlation with pathology probability (−0.54), reflecting complex interactions where elevated pressure may accompany compensatory cardiovascular responses that slow disease progression. This highlights the importance of multimodal models capable of detecting non-linear relations that statistical methods may overlook. The negative correlation also indicates that including hemodynamic variables improves interpretability and clinical relevance of CNN–LLM predictions.

Diastolic blood pressure showed a negligible correlation with pathology probability (0.01), indicating it is not a decisive marker. Though its predictive value is limited, including it preserves data completeness and latent interactions during model training.

Mutual correlations among glucose, systolic, and diastolic pressures were minimal (−0.01; 0.03), confirming the independence of these features and supporting the stability of the multimodal LLM-based framework. Consistent metrics across dataset partitions reinforce the reproducibility of the preprocessing pipeline.

Figure 2 illustrates the multimodal training dynamics of the CNN + LLM system. The left plot shows classification loss decreasing across epochs, indicating improved abnormal pattern discrimination, while the right shows reduced text generation cross-entropy, confirming coherent diagnostic text learning. The parallel convergence of both losses demonstrates balanced optimization through the joint loss function, aligning visual recognition with linguistic synthesis.

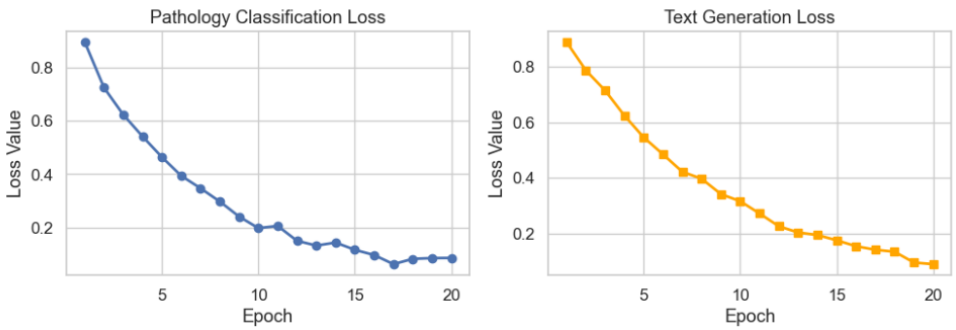

Fig. 2. Loss convergence curves for multimodal CNN + LLM architecture

Conclusions

This research presents an experimental approach for automated generation of textual medical reports from multimodal data using LLMs. The architecture, combining CNN-based image feature extraction with a language module integrating numerical clinical parameters, successfully produced grammatically correct and clinically relevant conclusions.

Correlation analysis confirmed the clinical significance of key parameters and the stability of the model. The results indicate that integrating LLMs into multimodal diagnostic analysis enhances the accuracy and coherence of automated medical reports, offering potential benefits for clinical decision support systems within a technical and computational framework.

References

1. Рудой В.В. Оптимізація мультимодальних нейронних мереж із використанням механізмів уваги для класифікації стадій цукрового діабету // *ΛΌГОΣ: матеріали IX Міжнародної науково-практичної конференції «DÉBATS SCIENTIFIQUES ET ORIENTATIONS PROSPECTIVES DU DÉVELOPPEMENT SCIENTIFIQUE»* (31 жовтня 2025 р., Париж, Французька Республіка). – ГО «Європейська наукова платформа»; SCI SORBONNE. – Париж, 2025. – С. 123–128. – DOI: https://doi.org/10.36074/logos-31.10.2025.022

2. Мінухін С. В., Рудой В. В. Розроблення гібридної моделі прогнозування стадії захворювання на цукровий діабет на основі згорткових нейронних мереж та мереж глибинного навчання // Global Trends in Science and Education: матеріали 4-ї Міжнародної науково-практичної конференції. – SPC «Sci-conf.com.UA». – 2025. – С. 316–319.

3. Рудой В. В., Мінухін С. В. Застосування методів машинного навчання для прогнозування стадій хвороби цукрового діабету // Матеріали XXVIII Міжнародного молодіжного форуму «Радіоелектроніка та молодь у XXI столітті». Конференція «Інформаційні інтелектуальні системи». – 2025. – С. 416–419.

4. Thawkar O. Can LLMs facilitate interpretation of pre-trained biomedical models? / O. Thawkar, A. Shaker, A. Sasan, H. R. Tizhoosh, A. A. Bidgoli // Scientific Reports. – 2023. – Vol. 13, Iss. 1. – Art. 19162. – DOI: 10.1038/s41598-023-46191-z.

3. Wu C. Towards Generalist Foundation Model for Radiology by Leveraging Web-scale 2D & 3D Medical Data / C. Wu, X. Zhang, Y. Zhang, Y. Wang, W. Xie // arXiv. – 2024. – arXiv:2401.14108. – DOI: 10.48550/arXiv.2401.14108.

4. Zhang Z. A Survey on Large Language Models for Biomedical Text Data: From Pre-training to Fine-tuning / Z. Zhang, Y. Zhao, T. Zhao, J. Liu, L. Li // Journal of Biomedical Informatics. – 2024. – Vol. 153. – Art. 104240. – DOI: 10.1016/j.jbi.2024.104240.

5. Jiang L. Y. Health system-scale language models are all-purpose prediction engines / L. Y. Jiang, X. C. Liu, N. P. Nejatian [et al.] // Nature. – 2023. – Vol. 619, Iss. 7969. – P. 357–362. – DOI: 10.1038/s41586-023-06160-y.

6. Мінухін С. В., Семенець О. М. Підвищення точності моделей машинного навчання при лікуванні цукрового діабету на основі збагачення тестових даних // Матеріали XXVIII Міжнародного молодіжного форуму «Радіоелектроніка та молодь у XXI столітті». Конференція «Інформаційні інтелектуальні системи». – 2025. – С. 461–463.

7. Singhal K. Large language models encode clinical knowledge / K. Singhal, S. Azizi, T. Tu [et al.] // Nature. – 2023. – Vol. 620, Iss. 7972. – P. 172–180. – DOI: 10.1038/s41586-023-06291-2.

8. Moor M. Foundation models for generalist medical artificial intelligence / M. Moor, O. Banerjee, Z. S. H. Abad [et al.] // Nature. – 2023. – Vol. 616, Iss. 7956. – P. 259–265. – DOI: 10.1038/s41586-023-05881-4.

9. Toma A. Clinical Camel: An Open Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding / A. Toma, P. R. Lawler, J. Ba [et al.] // arXiv. – 2023. – arXiv:2305.12031. – DOI: 10.48550/arXiv.2305.12031.

10. Wang H. HuatuoGPT, Towards Taming Language Model to Be a Doctor / H. Wang, C. Liu, N. Xi [et al.] // arXiv. – 2023. – arXiv:2305.15075. – DOI: 10.48550/arXiv.2305.15075.

11. Yang Z. A Survey of Large Language Models in Medicine: Progress, Application, and Challenge / Z. Yang, L. Li, J. Wang [et al.] // arXiv. – 2024. – arXiv:2403.05268. – DOI: 10.48550/arXiv.2311.05112.