Initial Study on Cancer Diagnosis using AI/NLP and Expert System Integration

Mariusz Pelc[1][0000−−0003−−2818−−1010],
Slawomir Stemplewski[2][0000−−0001−−6874−−1719],
Waldemar Bauer[3][0000−−0002−−8543−−0995],
Tomasz Kajdanowicz[1][0000−−0002−−8417−−1012], and
Aleksandra Kawala-Sterniuk[1][0000−−0001−−7826−−1292]

[1] Department of Artificial Intelligence, Wroclaw University of Science and Technology, Wroclaw, Poland
{mariusz.pelctomasz.kajdanowicz,aleksandra.kawala-sterniuk}@pwr.edu.pl
[2] Institute for Computer Science, University of Opole, Opole, Poland
slawomir.stemplewski@uni.opole.pl
[3] Dept. of Automatic Control & Robotics, AGH University of Krakow, Krakow, Poland bauer@agh.edu.pl

Abstract. In this paper, we propose a modular system architecture leveraging Natural Language Processing to process diverse medical data. The framework integrates outputs through an expert system layer (based on fuzzy logic) to improve explainability and deliver clinically actionable recommendations. Preliminary text-based analysis of patient interviews demonstrates potential to differentiate cancer types and support diagnostic workflows. By addressing integration of multiple analytical approaches, this work contributes to AI systems aimed at accelerating cancer diagnosis and improving clinical decision-making.

Keywords: Multimodal data · Cancer Diagnosis supporting systems · Medical data processing · Artificial intelligence.

1 Introduction

Widely understood artificial intelligence (AI) has been increasingly adopted in many areas of daily life [20, 4, 17] . As AI-based data analysis methods advance, the challenges faced by AI systems become more complex [2, 22, 26]. This results from the increasing complexity of problems and the data structures describing them [21, 3].

One of the fields where such tools are especially valuable is medical science [3, 14]. In clinical practice, patient records usually consist of diverse data types, such as [18]:

- medical test results enabling comparison with thresholds,
- doctors' interviews containing textual information on visits, recommendations, and outcomes,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_29

– image data (e.g., X-ray or MRI),
– other textual data (e.g., ICD-10 codes).

Medical sciences have specific characteristics, where diagnostic errors may lead to severe consequences [7]. In this context, cancer is a disease where time is crucial. Many cases show that even malignant tumours can be successfully treated if diagnosed early [25]. Therefore, the diagnostic process must be efficient, with all necessary information obtained as quickly as possible [25]. These may include biopsy or surgery, which should be applied only when necessary due to associated risks [12]. The use of AI to analyse patient data can shorten the diagnostic process and enable earlier treatment, increasing recovery chances. Through integrating diverse approaches (AI-/NLP-based, fuzzy, LLM), as proposed in this paper, we aim at a more robust and comprehensive solution combining deep learning for image analysis, expert systems for reasoning, and NLP for health record analysis can improve diagnostic accuracy [11, 13, 8].

Our analysis is performed on a real-world dataset (in XML format) from Opole Oncology Center (OCO) in Opole, Poland, including records from 10 patients undergoing cancer diagnostics. The dataset was fully anonymized and analysed offline without any interaction with participants.

2 Background to the Study

AI-/NLP-based data analysis is a rapidly evolving field integrating information from images, text, audio, and sensor data to better understand complex phenomena [8, 13, 23]. The use of multiple modalities, or different AI methods on limited data, often improves performance compared to single-modality analysis, addressing ambiguity, incompleteness, and noise. This section reviews advances in multimodal data analysis, focusing on ML, DL, LLMs, and CNNs.

Early multimodal approaches relied on traditional machine learning. Support vector machines (SVMs) combined features from different modalities in tasks such as sentiment analysis [15]. Probabilistic models, including Hidden Markov Models (HMMs) and Bayesian Networks, modeled temporal dependencies in multimodal data, e.g., human activity recognition [10]. However, these methods depended on hand-crafted features and struggled with high dimensionality.

Deep learning enabled automatic feature extraction and learning from raw data. Convolutional Neural Networks (CNNs) achieved strong results in image recognition [9] and object detection [19], and were combined with recurrent neural networks (RNNs) for sequential data. For example, in video captioning, CNNs extract visual features while RNNs generate text [24]. Attention mechanisms further improved performance by focusing on relevant inputs, e.g., in visual question answering (VQA) [1].

Integration of text and vision remains a key research area. Tasks such as image-text retrieval use joint embeddings to align semantic representations [6]. Transformer-based models significantly advanced this field, with approaches such as CLIP [16] enabling strong multimodal representations and zero-shot generalization.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_29

Large Language Models (LLMs) extend these capabilities to tasks requiring reasoning across modalities, including dialogue systems and instruction following [5]. Although primarily text-based, LLMs incorporate other modalities through projection layers or specialized tokens, enabling multimodal understanding.

Despite progress, challenges remain, including missing modalities, data heterogeneity, and robust fusion strategies. Interpretability, explainability, and ethical issues such as bias also require attention. Future work will focus on more generalizable multimodal architectures and improved fusion methods.

3 Materials and Methods

In this study, we analyze a real-world dataset from an oncological clinic, containing medical information from 10 patients during the diagnostic process. Originally in Polish, it was translated into English. The dataset includes patient visits, laboratory results, doctor consultations, diagnostic reports, and treatment plans, providing a comprehensive view of each case.

Due to the complexity of the data, including diverse attributes such as disease stages, laboratory tests, and specialist consultations, careful preprocessing was required. The variability between patient records reflects the heterogeneous nature of cancer diagnosis and treatment, with each case presenting different analytical challenges.

3.1 Patients' Records

Each patient has a record in the database, and the record is designed to store the most relevant information, including: *DateOfBirth*, *Gender*, *EpisodeCollection* (meaning all episodes related to the patient stored in the database). Each *EpisideCollection* constitutes a separate entry distinguishable by *EpisodeDate*, *ProcedureName* (e.g., visit, prescription, etc.), *ICDCodes* (ICD-10 codes describing a disease group and a specific type), *VisitExamination*, *VisitDiagnosis*, *VisitInterview/Interview*, *VisitRecommendations/Recommendations*, *LabTestValue*, *Observations* and more.

For AI-/NLP-based analysis, a key issue is that some records may be missing or repeated depending on the patient or cancer type (e.g., multiple visits with similar or different outcomes). As a result, the data cannot be easily organized into a coherent table structure required by many analytical tools. Consequently, many AI (ML/DL) algorithms are not directly applicable, requiring substantial restructuring. This may include data splitting by type (e.g., image vs. text) or transformation (e.g., flattening) to enable further processing.

Considering data structure, complexity, and completeness, we focused on the *Episodes* containing interview or consultation results, including symptoms, prior tests, and doctors' observations. From the XML data, we extracted four text chunks describing visit examination results and applied our framework to evaluate its use as a decision-support system.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_29

```
<Episode>
  <VisitExamination>General condition, good nutritional status, WHO-1</VisitExamination>
  <VisitDiagnosis>Tumor of the left lung and mediastinal lymphadenopathy</VisitDiagnosis>
  <VisitPerformed>The patient was informed that the Oncology Outpatient Clinic does not perform
  diagnostics of lung tumors</VisitPerformed>
  <VisitInterview>Patient presents for consultation for lung tumor diagnostics, referred by a
  Vascular Surgeon.</VisitInterview>
  <VisitRecommendations>Referral issued for diagnostics to the Pulmonology Department.
  An e-Referral was issued to the Department of Pulmonary Diseases. Code: 1802</
  VisitRecommendations>
  <EpisodeDate>2024-09-03T13:16:00</EpisodeDate>
  <ProcedureName>Consultation Visit</ProcedureName>
  <ICDCodes>
  (...)
  </ICDCodes>
</Episode>
```

(a) First Example of the *Episode* Record

```
<Episode>
  <EpisodeDate>2024-10-15T12:18:00</EpisodeDate>
    <ProcedureName>Visit - Anamnesis Clinic</ProcedureName>
      <ICDCodes>
        <ICDCode>
          <CODE>94.131</CODE>
          <TYPE>ICD9</TYPE>
          <DESCRIPTION>Interview before admission for treatment</DESCRIPTION>
        </ICDCode>
        <ICDCode>
          <CODE>D48.9</CODE>
          <TYPE>ICD10</TYPE>
          <DESCRIPTION>Neoplasm of uncertain or unknown behavior, unspecified</DESCRIPTION>
        </ICDCode>
      </ICDCodes>
</Episode>
```

(b) Second Example of the *Episode* Record

Fig. 1: Differences in the *Episode* Record Structure

To illustrate the differences in patient data, example data structure fragments are presented in Fig. 1a and Fig. 1b.

As shown in Fig. 1a and Fig. 1b, one can see where the difficulty in processing the patients' data may come from: the two *Episode* sections are in the XML file, but have completely different structures.

3.2 The Proposed Architecture for the Recommendation System

Patient data can be used in many ways; however, in cancer cases, the primary goal is to shorten the diagnostic process, potentially omitting unnecessary procedures.

In the 1*scaleati*65*t* dataset, one case shows a first visit on 03/09/2024 and a surgery recommendation on 26/02/2025, indicating a lengthy diagnostic process involving multiple consultations and tests. This higscalets 65e need for solutions supporting earlier decision-making, especially when cancer is not yet Example conDifferences in the Structure of the *Episode* Record. initial consultations have occurred. By applying ML, DL, LLMs, or CNNs to patient data and leveraging knowledge from similar cases, the system may recommend either standard diagnostics or more direct procedures, such as biopsy or surgery, thereby shortening the diagnostic timeline and improving outcomes.

Based on these considerations, we propose an AI-/NLP-based cancer diagnosis system architecture integrating a fuzzy inference system, as shown in Fig. 2.

The target system proposed in Fig. 2 should allow a detailed and multi-level analysis of the available patients' data. For this paper purposes we focus only on the text NLP-based analysis supported by a fuzzy inference system, while the overall architecture is discussed to place our findings in the proper context.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_29

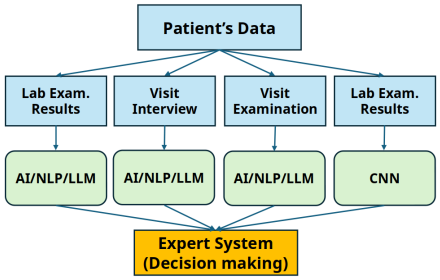

Fig. 2: Proposed Diagnostic System Architecture.

3.3 The Test Bed

In this section, we provide some implementation details as well as preliminary results indicating the way(s) in which the whole data analysis was performed. All the relevant code was developed in Python.

The first step of the analysis was based on comparing the extracted interview results against keywords that were chosen so that they specifically referred to various (*CANCER_INDICATORS*) appearing in patients suffering from a specific cancer type. Example keywords for just two cancer types are shown in Fig. 3a.

```
# Cancer type knowledge base (same as before)
CANCER_INDICATORS = {
    'breast_cancer': {
        'anatomical': ['breast', 'mammary', 'axillary', 'nipple', 'areola'],
        'symptoms': ['lump', 'mass', 'thickening', 'dimpling', 'discharge'],
    (…)
    },
    'lung_cancer': {
        'anatomical': ['lung', 'pulmonary', 'bronch', 'chest', 'thoracic', 'mediastin'],
        'symptoms': ['cough', 'hemoptysis', 'dyspnea', 'breath', 'hoarseness', 'wheez'],
    (…)
    },
```

(a) Keywords List for Cancer Indicators.

```
"""Define fuzzy IF-THEN rules for cancer identification"""
return {
    'breast_cancer': [
        {'IF': {'anatomical': 'high', 'diagnostic': 'high'}, 'THEN': 'very_high'},
        {'IF': {'anatomical': 'high', 'diagnostic': 'medium'}, 'THEN': 'high'},
        {'IF': {'anatomical': 'medium', 'symptoms': 'moderate'}, 'THEN': '
        {'IF': {'anatomical': 'low', 'symptoms': 'few'}, 'THEN': '
    ],
(…)
}
```

(b) Fuzzy Rules Set

Fig. 3: Details of the Rule-based Fuzzy Inference System.

The fuzzy reasoning is based on a fuzzy rule set shown in Fig. 3b.

The list of keywords is of utmost importance, as the the system checks the number of occurrences of cancer-type-specific keywords in the analysed text. The comparison is performed in three main categories: *Anatomical Terms*, *Symptoms* and *Diagnostic Terms*.

As the reference sample dataset consists of only 10 patients' data, (insufficient to train a model from scratch) hence we decided to apply NLP-based analysis and support it with a fuzzy inference system.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_29

In order to do this, we performed a simple stylometric analysis based on comparisons of word distributions between 4 different interviews. These cases were related to: *Pelvic Cancer*, *Breast Cancer*, *Liver Cancer*, *Lung Cancer*.

Taking into account the relevant categories of analysis, we designed a fuzzy system allowing the final decision and cancer type identification to be obtained based on the NLP-based analysis results. The fuzzy system design is presented in Fig. 4.

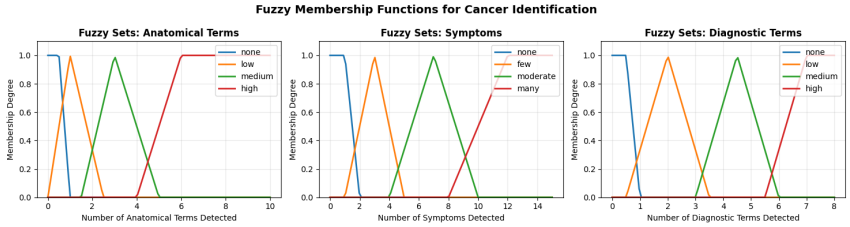

Fig. 4: Fuzzy System Design Details.

4 Results

To represent differences between interviews, we used violin plots from the *seaborn* module in Python.

Word count for the four interviews is presented in Fig. 5.

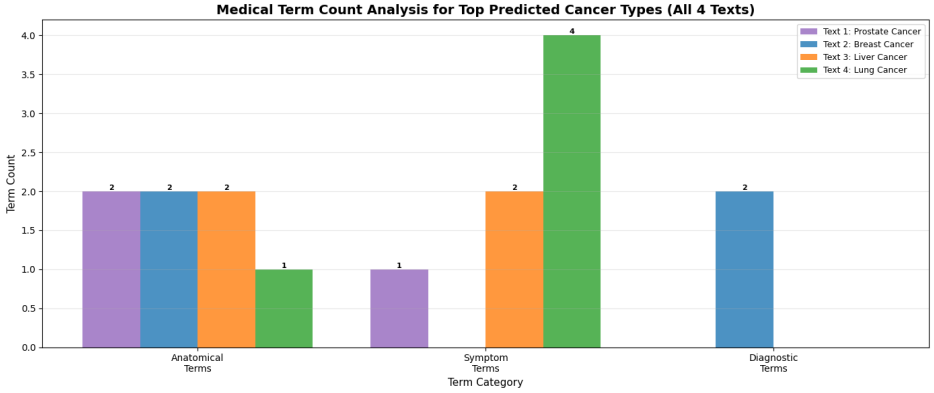

Fig. 5: Word Count Statistics for Each Interview.

Analysis of matching words in the fuzzy system allowed identification of three out of four cancer types. The highest score (nearly 55%) was observed for lung cancer while second highest score was for Liver cancer (32%).

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_29

5 Conclusion

Preliminary analysis of patient interviews shows differences in word frequency distributions between cancer types, indicating the potential of NLP tools in diagnosis. The system aims to support tailored diagnostic pathways and shorten time to interventions such as biopsy or surgery, improving clinical decision-making and patient outcomes.

The architecture addresses data heterogeneity by assigning techniques to specific data types: ML for numerical data, NLP or LLMs for text, and CNNs for imaging. Integration through an expert system supports coherent decisions and reduces risks such as hallucinations or overfitting, aligning with clinical requirements for interpretability and reliability.

Although limited by the small dataset, results confirm semantic differences between cancer types, supporting further research with larger samples to improve accuracy and generalisability. In our view the work provides a foundation for AI-enhanced oncology diagnostics, with potential to shorten diagnostic time and improve patient care.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2425–2433 (2015)
2. Assres, G., Bhandari, G., Shalaginov, A., Gronli, T.M., Ghinea, G.: State-of-the-art and challenges of engineering ml- enabled software systems in the deep learning era. ACM Comput. Surv. **57**(10) (May 2025). https://doi.org/10.1145/3731597, https://doi.org/10.1145/3731597
3. Barbierato, E., Gatti, A.: The challenges of machine learning: A critical review. Electronics **13**(2), 416 (2024)
4. Dargan, S., Kumar, M., Ayyagari, M.R., Kumar, G.: A survey of deep learning and its applications: a new paradigm to machine learning. Archives of computational methods in engineering **27**(4), 1071–1092 (2020)
5. Dubois, E., Fan, Y., Xu, W., Yao, H., Yang, X., Wen, L., Zhang, H., Hu, Y., Xie, H., Li, Z., Li, S., Zhang, J., Wu, H., Zhu, X., Zou, W.: Alpaca-farm: A simulation framework for benchmarking instruction-following models. arXiv preprint arXiv:2305.15047 (2023)
6. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. arXiv preprint arXiv:1312.6115 (2013)
7. Kempt, H.: Ethical investigations of AI in medical diagnostics. Ph.D. thesis, Dissertation, Rheinisch-Westfälische Technische Hochschule Aachen, 2023 (2023)
8. Khalate, P., Gite, S., Pradhan, B., Lee, C.W.: Advancements and gaps in natural language processing and machine learning applications in healthcare: a comprehensive review of electronic medical records and medical imaging. Frontiers in Physics **12**, 1445204 (2024)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105 (2012)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_29

10. Lahat, D., Adali, T., Jutten, C.: Multimodal data fusion: An overview of methods, challenges, and prospects. Proceedings of the IEEE **103**(9), 1449–1477 (2015). https://doi.org/10.1109/JPROC.2015.2443015
11. Li, C., Zhang, Y., Weng, Y., Wang, B., Li, Z.: Natural language processing applications for computer-aided diagnosis in oncology. Diagnostics **13**(2), 286 (2023)
12. Linet, M.S., Slovis, T.L., Miller, D.L., Kleinerman, R., Lee, C., Rajaraman, P., Berrington de Gonzalez, A.: Cancer risks associated with external radiation from diagnostic imaging procedures. CA: a cancer journal for clinicians **62**(2), 75–100 (2012)
13. Nazir, A., Hussain, A., Singh, M., Assad, A.: Deep learning in medicine: advancing healthcare with intelligent solutions and the future of holography imaging in early diagnosis. Multimedia Tools and Applications **84**(17), 17677–17740 (2025)
14. Panahi, O.: Deep learning in diagnostics. Journal of Medical Discoveries **2**(1), 1–6 (2025)
15. Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L.P.: Multimodal sentiment analysis: A unified framework. IEEE Intelligent Systems **32**(6), 56–65 (2017). https://doi.org/10.1109/MIS.2017.4531016
16. Radford, A., Wu, J.W., Child, R., Luan, D., Amodei, D., Sutskever, I.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
17. Razzaq, K., Shah, M.: Machine learning and deep learning paradigms: From techniques to practical applications and research frontiers. Computers **14**(3), 93 (2025)
18. Recharla, M., Chakilam, C., Kannan, S., Nuka, S.T., Suura, S.R.: Revolutionizing healthcare with generative ai: Enhancing patient care, disease research, and early intervention strategies. American Journal of Psychiatric Rehabilitation **28**(1), 98–111 (2025)
19. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788 (2016)
20. Sarker, I.H.: Machine learning: Algorithms, real-world applications and research directions. SN computer science **2**(3), 160 (2021)
21. Simon, H.A.: Identifying basic abilities underlying intelligent performance of complex tasks. In: The nature of intelligence, pp. 65–98. Routledge (2024)
22. Soni, N., Nigam, N.: Recent advances in artificial intelligence and machine learning: Trends, challenges, and future directions. International Journal of Engineering Trends and Applications (IJETA) **12**(1), 9–12 (2025)
23. Swanson, K., Wu, E., Zhang, A., Alizadeh, A.A., Zou, J.: From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. Cell **186**(8), 1772–1791 (2023)
24. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2625–2633 (2015)
25. Wisniewska, K., Marku, E., Ugurbas, M.V., Hartmane, I., Shukurova, M.: Innovations in cancer diagnosis and treatment: prospects and challenges. Healthcare in Low-resource Settings (2024)
26. Zalewa, K., Olszak, J., Kapłan, W., Orłowska, D., Bartoszek, L., Kaus, M., Klepacz, N.: Application of artificial intelligence in radiological image analysis for pulmonary disease diagnosis: A review of current methods and challenges. Journal of Education, Health and Sport **77**, 56893–56893 (2025)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_29