Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Adaptive AI Platform for Processing Breast Radiology Reports in Polish Medical Language

Maciej Bobowicz

maciej.bobowicz@gumed.edu.pl

Medical University of Gdańsk

Dariusz Szplit

Medical University of Gdansk

Beata Graff

Medical University of Gdańsk

Anna Szyndler

Medical University of Gdańsk

Julia Bogdan

Medical University of Gdańsk

Kinga Słomińska

Medical University of Gdańsk

Anna Dąbkowska

Medical University of Gdańsk

Mariusz Budzisz

eTrust Medical Ltd

Franciszek Górski

Gdańsk University of Technology

Marta Zielonka

Gdańsk University of Technology

Józef Kotus

Gdańsk University of Technology

Tomasz Stefaniak

Medical University of Gdansk

Andrzej Czyżewski

Gdańsk University of Technology

Krzysztof Narkiewicz

Medical University of Gdańsk

Article

Keywords: breast, radiology, radiology reporting, medical text annotation, description, speech recognition, Polish Medical Language, BERT topics

Posted Date: May 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-7274269/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

An accurate, well-organised report from a radiological examination, in our case a breast ultrasound, is an important part of a radiologist's work [1]. As part of the ADMEDVOICE research consortium [6], we have developed an AI platform adapted to Polish medical language, which generates structured reports based on freely dictated text during or after a breast ultrasound examination. We developed a report structure, on the basis of which we created a contextual division of information and a dictionary of synoptic data. Then, using data from real radiology reports, we trained a language model to recognise speech and correctly classify descriptions. We added algorithms for numerical and dictionary data to the grouped information and then put everything into an application that is convenient for doctors to use.

Focusing on Polish breast radiology reports, we used BERTopic for thematic modelling to categorise the text. We found that the topics extracted by the model were closely aligned with the categories defined by clinical experts. However, the accuracy of the solution was not satisfactory, reaching 73%. We then tested the creation of a classification using a large language model (LLM) as an information classifier. The LLM model was used as a learning model during the training of a smaller encoder-type language model for text classification. We used two LLM models, the multilingual Llama3.1-70B-Instruct and Bielik-11B-v2.3-Instruct, achieving a classification accuracy of 94%. In the next step, to improve the results, we added appropriate dictionaries of medical terms, abbreviations, and units of measurement used in the descriptions.

Introduction

Ultrasound examination is one of the most frequently used imaging modalities in modern clinical practice due to its accessibility, safety, and diagnostic value. In many healthcare systems, including Poland, the documentation of ultrasound findings relies heavily on free-text dictation or manual note-taking. This approach, while flexible, often results in unstructured and inconsistent reports, which can hinder clinical communication, reduce the quality of patient records, and limit the potential for data reuse in research and quality monitoring [21, 22].

The study aimed to develop an application for radiologists that generates structured report documentation in Polish during the breast examination. Support for daily work enhances the well-being of doctors. This is crucial for high-quality healthcare. Excessive workload, which doctors across all specialties complain about, causes stress, job dissatisfaction, and ultimately burnout [18, 19]. This not only impacts the health and well-being of individual doctors but also negatively affects the quality of healthcare services [20]. To address this issue, it is essential to create a supportive work environment that alleviates doctors' workload [2, 3, 4, 5, 7].

The study was conducted as a part of the larger research and development project with an acronym ADMEDVOICE - Adaptive intelligent speech processing system of medical personnel with the structuring of test results and support of therapeutic process. It was run by the ADMEDVOICE consortium consisting of Gdańsk University of Technology, Medical University of Gdańsk, University Clinical Centre

and eTRUST company. It has developed and is implementing an artificial intelligence (AI) solution that enables doctors to create dedicated electronic health records (EHR) entries using voice recognition and commands. This includes filling out selected patient medical documents and prescribing treatment as necessary [6]. Ten clinical situations have been identified in which the language model will serve as a valuable tool to support doctors' interaction with EHRs. For these scenarios, ontologies have been defined, the structure of medical information has been designed, training datasets for the language model have been collected, training methods have been tested, hundreds of hours of recordings for thousands of texts have been made, and an application has been developed to generate complete medical documentation (EHR), alongside the necessary equipment for use by doctors in their daily clinical practice.

Our goal was to develop a support tool for doctors, allowing them to quickly and conveniently prepare a structured report from a breast radiology examination for cancerous lesions. In the future, such reports will be used for further research using structured Real-World Data. Currently, the use of radiological descriptions for further research is challenging. The analysis of such unstructured reports is inefficient, hindering effective information extraction and making it difficult to analyze. Therefore, our team of medical specialists (including radiologists, oncologists and breast surgeons) defined structured formats for several clinical situations, including radiological breast studies, creating a clinically approved documentation structure. To obtain the expected effect of structured data from unstructured information, we investigated whether automated methods, such as topic modelling, can reproduce the text structure defined by medical specialists.

Focusing on Polish breast radiological reports, we used BERTopic for topic modelling to categorise the text. We found that the themes extracted by the model closely aligned with the categories defined by clinical experts. However, the accuracy of the solution was not satisfactory. Mapping these themes to our predefined categories initially achieved 73% accuracy in categorising text excerpts, demonstrating the potential of thematic modelling for automation.

We then tested classification creation using a large language model (LLM) as an information classifier. The LLM was employed as a learning model during the training of a smaller, encoder-type language model for text classification. We used two LLMs, the multilingual Llama3.1-70B-Instruct and Bielik-11B-v2.3-Instruct, achieving a classification accuracy of 94%. In a subsequent step, to improve the results, we added relevant dictionaries of medical terms, abbreviations, and units of measurement used in the descriptions.

The system is currently at TRL 7 (technology readiness level, indicating a prototype developed for use in a clinical setting). It provides stability and consistency, achieving the expected speech and context recognition values during extensive testing. Polish medical language is rich in specific names and terms borrowed from other languages, mainly Latin and English, as well as abbreviations and various accents. The tests considered different pronunciations and types of content by supporting a variety of input data, such as regional accents, complex terminology, and noisy environments, supported by error correction

mechanisms and a specialized acoustic probe. The next steps include optimising the application and integrating it with the hospital's HIS system. The importance of structured descriptions is enormous. Structured data from electronic health records (EHR) supports real-world data (RWD) research, verified by medical professionals using evidence-based medicine (EBM).

In this paper, we present a part of the entire project, focusing on breast ultrasonography reporting. For this purpose, based on the hospital database of ultrasound reports, the team of medical experts made voice recordings, which were then transcribed, and divided into text fragments later placed in context using a machine learning system trained through knowledge distillation, consisting of Llama3.1-70B-Instruct as the teacher model and distilbert-base-multilingual-cased as the student model.

To teach the model to create structured information from the recognized spoken text, in the first stage, we focused on extracting information from radiology reports. It typically contains descriptions of breast tissue, specific test results, measurements, BIRADS categories, lymph node assessments, and clinical recommendations. However, this information is usually presented in an unstructured free-text format, which makes the automatic extraction and analysis difficult.

The main challenge in processing radiology reports is that various types of information are often presented together in the same narrative. To enable structural analysis, it is necessary to divide the text into coherent fragments and classify them into predefined categories. In our study, we addressed this challenge by dividing the reports into smaller text units—based on paragraph boundaries—and applying a classification approach to assign each fragment an appropriate label [8, 9, 10].

To support this task, we tested BERTopic, a thematic modeling technique that combines transformation-based embedding with clustering algorithms and LLM as a classifier. Unlike traditional models such as Latent Dirichlet Allocation (LDA), which rely on shallow statistical patterns (e.g. TF-IDF), BERTopic captures deeper semantic relationships between words and phrases, making it better suited to the subtle language of medical texts [11]. Although BERTopic is primarily designed for unsupervised topic modelling, we adapted it for use in a classification context by matching the topics it generates to a set of predefined labels developed by doctors [12, 13].

Our motivation for investigating BERTopic is its ability to discover hidden structures in unstructured text without the need for large amounts of labelled data - a common limitation in clinical NLP applications. We also investigated whether the approach could suggest meaningful structure for radiology reports, which often follow implicit conventions rather than rigid templates.

The aim of our project was to develop an optimal algorithm for automatic tagging of medical texts using voice recognition algorithms and LLM. The result of this tagging process is the structuring of the input text in the form of a radiology report for breast examinations, so that it could be saved in the application as a structured radiological report.

Results

Unsupervised BERTopic baseline

The first approach to text context classification was based on the topic modelling method. The BERTopic-based approach demonstrated promising alignment with the structure proposed by clinical experts, successfully identifying key sections of the breast ultrasound reports that correspond to the main headers used by radiologists. Despite being trained in a fully unsupervised setting—without any access to labelled data during training—the model was able to extract topics that closely reflect the intended informational structure of the reports.

Figure 1 shows the normalised confusion matrix obtained with the BERTopic model on the held-out test set.

When evaluated against a human-validated subset of the data, the model achieved an average accuracy of 73% across all predefined labels. A closer look at the confusion matrix [Fig. 1] reveals robust performance in certain categories:

- Breast composition description (breast structure) with 94.2% recall (223 samples);
- Conclusions with 100% recall (199 samples);
- Milk ducts descriptors with 70.1% recall (147 samples);
- Breast structure (glandular tissue content) with 57.4% recall (373 samples);
- Lymph nodes with 49.1% recall (169 samples).

Supervised LLM-based classifier

The second approach to text context classification was based on the LLM as an information classifier. A significant amount of tagged training data, containing several thousand examples, is necessary to complete the classifier model training process. Labelling such a large volume of data is time-consuming and resource-intensive. It requires the participation of highly specialised doctors from various disciplines as members of the tagging team. To address the challenges posed by the large size of the training and validation datasets and the limited availability of radiology specialists able to label the entire dataset, we employed an automated approach supported by AI. A large language model was used as a teacher model in the process of training a smaller encoder language model for text classification. The entire process of data tagging and model training is shown in Fig. 2.

Two large language models were used as the teacher model - the multilingual Llama3.1-70B-Instruct model and the Bielik-11B-v2.3-Instruct model, the latter explicitly developed for Polish. Both models feature very high capabilities in understanding the Polish language and generating content in this language based on set instructions. For the role of the learner model, it was decided to use the distilbert-base-multilingual-cased model, which is a basic encoder model supporting multiple languages, including Polish.

Figure 3 presents the normalised confusion matrix for the supervised LLM-based classifier on the test set, illustrating the distribution of true (rows) versus predicted (columns) labels across all classes.

From the analysis of the results for the breast radiology report, it can be observed that the accuracy of the classifier models decreases when the following tags are recognised:

L_BIRADS, R_BIRADS, LYMPH_NODES_SUSP.

The reasons why the results in classifying the annotations mentioned above are lower than the average can be named as:

1. 1. The under-representation of the training data for these tags, both quantitatively and qualitatively.
2. 2. The high similarity of the texts representing the tags in question with those of other annotations, but semantically similar, e.g. the L_BIRADS and R_BIRADS annotations are often confused by the model with the BIRADS annotation, intended to denote a collective description of the BIRADS system for both breasts of the patient.

3.1 Data Collection and Preparation

Breast radiology reports, while often loosely structured, typically follow implicit patterns that can be broken down into distinct informational components. In our study, we leveraged this characteristic by segmenting entire reports into smaller, coherent fragments, based on the original paragraph structure. This process was applied to a dataset of 6,269 breast ultrasound reports sourced from real hospital documentation, resulting in a total of 48,721 text fragments. These fragments served as the training data for our topic modeling algorithm and LLM classifier.

An example of a radiological report that was analyzed is as follows:

Breasts with an fatty-glandular structure. Status after FNAB changes at 12 noon - mild change, in good condition. present without visible pathology. Suprafascial hypoechoic focal lesion, smoothly contoured, size 5.5x2.5mm - like a dense cyst - comparable to before. There were no solid or cystic lesions in the right breast. No features of the infiltration process are visible. Milk ducts not dilated on both sides. Physiological absorbents in the armpits on both sides. Conclusions and recommendations: BIRADS 3 for review.

3.2 Metodology

The LLM model has proven to be a better classifier. It was used as a learning model in the training process of a smaller encoder-type language model. We used two large language models: the multilingual Llama3.1-70B-Instruct model and the Bielik-11B-v2.3-Instruct model. To reinforce the correct structure of information for EHR recording, a dictionary containing 601 abbreviations used by physicians was prepared from real hospital records (EHR). Within this dictionary, 110 abbreviations were related to the description of the breast radiology examination. The dictionary contained abbreviations and their correct expansions, along with rules for when an abbreviation should be used in expansion and when it should

be used in letter form. This dictionary was recorded in its entirety and added to the speech recognition model and context. Text with recognized context and recognized dictionary expressions is algorithmized, i.e. assigned to the form fields. Selected excerpts are placed in the form, as text fields and selected excerpts are marked and placed as fields in the synoptic form, e.g. as checkboxes to indicate the selected clinical scale or the type of lesion found.

To assess our approach, we used a target structure for breast ultrasound reports defined by our clinical experts, consisting of eight key informational categories: referral data (sourced from HIS system), breast structure, breast description, milk ducts assessment, lymph node assessment, BIRADS classification, recommendations, and other or miscellaneous content. This expert-defined structure is shown in Fig. 4 and served as a post hoc reference for evaluating the outputs of the topic modeling process. A list of 20 clinically relevant information-topics has been prepared for both topic modelling and LLM classification: 1) Type of Imaging and Sequences, 2) Examination Features, 3) Image Characteristics, 4) Echogenicity Descriptors, 5) Internal Structure of Mammary Gland, 6) Description of Anatomical Structures, 7) Evaluation of Internal Breast Structures, 8) Presence of Pathological Features, 9) Tissue (or Lesion) Description Features, 10) Characteristics of Pathological Lesions, 11) Solid Tumor Details/Descriptors, 12) Lesion Location, 13) Lesion Shape, Margin and Contour, 14) Additional Remarks, 15) Fat Tissue Structure Abnormalities, 16) Blood Vessel Characteristics, 17) Blood Flow Patterns, 18) Lymph Node Assessment, 19) Lymph Node Location and Count, 20) Lymph Node Size.

Importantly, we deliberately excluded any human-validated labels from the training phase to maintain a fully unsupervised setting. This design choice reflects a real-world scenario in which labeled data is limited or unavailable - common in many clinical and medical text domains. By evaluating the extracted topics against a separate, manually annotated by our clinicians dataset only after training, we aimed to assess whether an unsupervised model like BERTopic can still recover a meaningful structure from unstructured clinical texts [14]. Data was annotated using Label Studio [14].

3.3 Integrating BERTopic for Classification

To model the semantic structure of the text fragments, we configured a BERTopic pipeline. For generating embeddings, we used the paraphrase-multilingual-MiniLM-L12-v2 model from the SentenceTransformers library [16], which is well-suited for multilingual data and performs efficiently even on relatively short text segments.

Dimensionality reduction was applied to the resulting embeddings using Principal Component Analysis (PCA), reducing the vector space to three dimensions. This step facilitates more effective clustering by eliminating noise and focusing on the most informative features. Subsequently, we used KMeans clustering to group the fragments into topics. The number of clusters was set to seven, reflecting the number of categories defined by our domain experts (omitting the referral section which is not part of an image test itself).

This choice enabled us to assess how well the learned topics aligned with the known components of breast ultrasound report structure. As shown in Fig. 5, the extracted topics were color-coded to reflect their alignment with the reference structure. The results revealed some degree of overlap, with certain topics - clusters 0 and 1, as well as 2 and 4—frequently merging into single sections. While clusters 0 and 1 both fall under the clinician-defined category of breast description, closer inspection shows that they are semantically distinct. Cluster 0 primarily captures descriptive findings, whereas cluster 1 tends to reflect measurements and the localization of those findings. Similarly, clusters 2 and 4 can be mapped primarily to the conclusion section of the report, though they also contain elements related to recommendations, indicating some thematic overlap between these two categories, which is reasonable as these sections are often written in the same paragraph boundary in a breast ultrasound report.

3.4 Annotation of training data and LLM training

Annotations were carried out using the two large language models described, Llama3.1-70B-Instruct and Bielik-11B-v2.3-Instruct. Each model was given a prompt, the content of which varied slightly depending on the clinical situation, containing information about the role the model was to play, the content of the command, the set of tags to be used, as well as examples of input data and the expected responses for these data within so-called in-context examples. These examples were prepared by specialists (clinicians) in the field and contain all the tags used to create the form for the description of a radiological examination. Before the model training was implemented, each collection was divided into a training collection and a validation collection. The tags generated by the Llama3.1 model and the tags from the Bielik model were used for training. It was noted that there was a disparity in the amount of data representing each tag in each set. This problem was addressed in the training of the classifier models by assigning weights to the individual tag sets, thus reducing their impact on the learning process.

In the training set for the radiological examination, there is a large variation between the

quantities of data for individual tags. The largest number of examples was collected for tags

directly describing the condition of the breast - the LEFT, RIGHT, BOTH tags. The fewest examples were collected for tags relating to breast structure types: the GLAND_FAT_COMP, GLAND_COMP and FAT_GLAND_COMP tags.

The prepared training and validation sets were used to train the classifier model. As a result of the experiments, it was decided to use a distilbert-base-multilingual-cased model. Experimentally, the following hyperparameters of the model training process were established: learning rate – 1e-6; training data batch size – 32; gradient accumulation cycle – 1 training step.

Appropriate weights for classes (tags) were used to balance their influence on the learning process. The learning curve functions were determined by calculating the error values - training and validation during

the post-training process of the models. On this basis, it was concluded that the training was correct - both functions systematically decreased throughout the training period.

3.5 Current status of work

The current state of work involves testing the application interface for doctors performing breast ultrasound examinations in real-life conditions. Acoustic probes have been installed in doctors' offices. These are specialised microphones developed by Gdańsk University of Technology to recognise the voice of the doctor performing the examination accurately. The accuracy of context recognition has been improved thanks to the methods described above. A method for creating clinical accuracy benchmarks for recognised and transcribed texts has been established. Tests conducted to date in real-world conditions have demonstrated the stability of the system, with consistent speech recognition accuracy and context assignment of 94% during prolonged use in hospital settings. Recognition accuracy has been confirmed in various scenarios, from controlled laboratory conditions to a typical hospital doctor's office, where the recorded sound was adapted to noise and dynamic interactions thanks to the probe. Sustainability is ensured through compatibility with hospital information systems (HIS), enabling seamless data exchange and updates without disrupting clinical activities.

Discussion

Our results indicate that BERTopic can extract clinically relevant groupings that conform to expert-defined categories, even if it does not capture finer distinctions (such as left and right breast) that require a broader awareness of context on the part of the model, i.e. analysing larger wholes than paragraph boundaries. The results obtained confirm the possibility of using BERTopic as a lightweight, unsupervised classification tool for structuring clinical texts, including texts in Polish. It is an efficient and promising method in situations where labelled training data is unavailable or difficult to obtain, making it a potentially valuable tool for exploratory structuring of other types of medical narratives. However, a more accurate tool that we use for context detection is a language model trained for a given type of study. The topic modelling application solution was replaced by a classifier model, whose task was to classify text sequences in the form of sentence fragments, complete sentences or sequences longer than one sentence. The accuracy levels achieved in tag recognition are very high, currently standing at 94%. This performance compares favourably with the cross-institutional benchmark reported in [8] — one of the few studies tackling section classification with limited in-domain labels where micro-F1 scores averaged 0.66 and peaked at 0.82 for the most favourable source–target pair. Structured reporting has repeatedly been identified as a key prerequisite for reliable AI pipelines in breast imaging [17]; by reaching 94% accuracy our classifier approaches the performance level considered adequate for clinical integration. Adding an abbreviation dictionary and selected proper names to the model allows this accuracy to be increased – testing of this solution is ongoing.

4.1 Limitations

Although the topic-based modelling approach has shown promising results in recovering domain-relevant categories, it has also revealed significant limitations. First, the BERTopic model does not take into account the global contextual information contained in reports. As a result, it is unable to distinguish clinically relevant nuances, such as laterality (e.g., left vs. right breast), which often requires document-level context or resolution of co-references. Second, some semantically similar categories (e.g., conclusions vs. recommendations) were partially merged by the model, reflecting the fact that these elements are often recorded in shared paragraphs and do not have clear boundaries. Therefore, working with a model trained for a specific type of form and clinical situation to recognize context is definitely more promising and is the basis for us to implement this solution in the real-world conditions of a university hospital.

Study results are limited by a single-centre design and the lack of external validation of the developed methods. This will be addressed at later stages of the project's development.

4.2 Future work

In the next steps, we are conducting a supervised approach study to directly assign labels to text fragments from actual records in radiology departments and compare their performance with the unsupervised baseline established in this study. The application is currently at TRL 7, where POC is being tested for ultrasound examination forms and other clinical situations. By the end of the year, the application will reach TRL 9 and will be launched at a university hospital. Finally, although this study focused exclusively on text reports, the overall research project includes audio recordings of doctor-patient interactions, which will ultimately require well-tested speech-to-text processing and context classification processes that take into account the authors of the spoken content.

Conclusions

The encoder classifier models for text sequence classification, whether in the form of sentence fragments, a complete sentence, or a sequence longer than one sentence, are more efficient than topic modelling applications for Polish medical vocabulary. Moreover, the proposed methodology offers an automatic and appealing structure for ultrasound reporting comparable to that developed by human experts. The results of the context recognition were adequate for real-world ultrasonography laboratory testing.

Methods

All tests were conducted on anonymized data, in accordance with the guidelines and regulations on personal data protection in the European Union. The data processing protocols were approved by the authorized Bioethics Committee for Scientific Research at the Medical University of Gdańsk (Resolution no. KB/508/2023). All participants in the recordings conducted by the project team gave their informed, written consent to the recording and storage of voice recordings.

Declarations

Acknowledgements

The Polish National Centre for Research and Development (NCBR) supported this research in the project: “ADMEDVOICE- Adaptive intelligent speech processing system of medical personnel with the structuring of test results and support of therapeutic process,” no. INFOSTRATEG4/0003/2022.

Funding statement declaration:

This project is funded by the Polish National Centre for Research and Development (NCBR) under the Infostrateg IV Programme.

Grant amount: PLN 7,983,375.00; Total project value: PLN 8,299,875.00

Author contributions

Conceptualization, M.B, D.S, A.C, B.G, K.S, A.D, F.G, M.Z

Methodology, M.B, D.S, A.C, A.D, K.S, F.G, M.Z

Supervision, A.C, K.N, T.S,

Data Preprocessing, D.S, B.G, A.S, J.B, A.D, F.G, M.B, M.Z, S.Z, J.K.

Specialized Data Labeling, M.B, B.G, A.S, J.B, A.D, M.Z, J.K

Data Analysis, J.B, A.D, K.S, M.Z, F.G

Writing—original draft, M.B and D.S, B.G, A.D, K.S, F.G

Writing—review and editing, M.B, D.S, B.G, K.S, M.Z and A.C.

All authors have read and agreed to the published version of the manuscript.

Data availability statement

The anonymized datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

Anonimized medical audio files. You can choose set of files (divided by recording source types) or as a single to dawnload: http://www.multimed.org/admedvoice/free-access

Competing interests

The authors declare no competing interests.

Ethical approval

The anonymized datasets generated during and analyzed during the current study are available from the corresponding author upon reasonable request. Since the study involves only physicians and no patients, it is not a medical experiment. This work is a part of the broader ADMEDVOICE project, which the Bioethical Commission for Scientific Research has favorably evaluated at the Gdańsk Medical University (Resolution no. KB/508/2023 dated 2023, Sept. 15th).

Additional information

Correspondence and requests for materials should be addressed to D.S.

References

1. European Society of Radiology (ESR). ESR paper on structured reporting in radiology—update 2023. *Insights Imaging*. **14**, 199. https://doi.org/10.1186/s13244-023-01560-0 (2023).
2. Pratama, R. & Yufika, A. Physicians' Workload and Quality Healthcare in Indonesia. *Trends Infect. Global Health*. **3** (1), 43–55. https://doi.org/10.24815/tigh.v3i1.32363 (2023).
3. Jackson, N. et al. Understanding Burnout Rates for Clinicians and Physical Symptoms of Allostatic Load. *Archives Case Rep.* **9** (1), 011–015. https://doi.org/10.29328/journal.acr.1001122 (2025).
4. Smith, M. E. et al. Artificial Intelligence in Emergency Medicine: A Primer for the Nonexpert. *JACEP Open.* **6** (2), 100051 (2025).
5. Elhilali, A., Brügger, V., Tschannen, I., Hautz, W. & Krummrey, G. AI-Enhanced Speech Recognition in Triage. In Healthcare of the Future 2025 (31–34). IOS. (2025).
6. Zielonka, M. et al. *Machine Learning Tools Match Physician Accuracy in Multilingual Text Annotation*. (2024). https://doi.org/10.21203/rs.3.rs-4157610/v1
7. Tierney, A. A. et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst Innovations Care Delivery*. **5** (3), CAT–23 (2024).
8. Zhou, W. et al. Improving the Transferability of Clinical Note Section Classification Models with BERT and Large Language Model Ensembles. ClinicalNLP 2023. (2023). 10.18653/v1/2023.clinicalnlp-1.16
9. Wang, H. et al. Few-shot section classification in clinical notes with domain-aware contrastive learning. In Proceedings of EMNLP 2022, pages 6934–6947. (2022). 10.18653/v1/2022.emnlp-main.472
10. Chen, Z., Kim, J., Liu, Y. & Narayanan, S. Clinical note section classification on doctor–patient conversations in low-resourced settings. In Proceedings of the 3rd Workshop on NLP for Medical Conversations, pages 1–12. (2023). 10.18653/v1/2023.nlpmc-1.1
11. Blei, D. M., Ng, A. Y. & Jordan, M. I. *Latent Dirichlet Allocation JMLR* **3**: 993–1022. (2003).

12. Chiu, C-C., Wu, C-M., Chien, T-N., Kao, L-J. & Li, C. Predicting ICU Readmission from Electronic Health Records via BERTopic with a Long Short-Term Memory Network Approach. *J. Clin. Med.* **13** (18), 5503. 10.3390/jcm13185503 (2024).
13. Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen,C., … Chaudhari, A. S. 2024. Adapted Large Language Models Can Outperform Medical Experts in Clinical Text Summarization. Nature Medicine, 30(4), 1134–1142. doi:10.1038/s41591-024-02855.
14. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. (2022). arXiv preprint arXiv:2203.05794.
15. Tkachenko, M., Malyuk, M., Holmanyuk, A. & Liubimov, N. Label Studio: Data labeling software. Available at 301 (2024). https://github.com/HumanSignal/label-studio
16. Wolf, T. et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of EMNLP: System Demonstrations, 38–45. (2020). 10.18653/v1/2020.emnlp-demos.6
17. Busch, F. et al. Large language models for structured reporting in radiology: past, present, and future. *Eur. Radiol.* **35**, 2589–2602. https://doi.org/10.1007/s00330-024-11107-6 (2025).
18. Harolds, J. A., Parikh, J. R., Bluth, E. I., Dutton, S. C. & Recht, M. P. Burnout of Radiologists: Frequency, Risk Factors, and Remedies: A Report of the ACR Commission on Human Resources. *J. Am. Coll. Radiol.* **13** (4), 411–416 (2016). Epub 2016 Jan 4. PMID: 26768546.
19. Fawzy, N. A. et al. Incidence and factors associated with burnout in radiologists: A systematic review. *Eur. J. Radiol. Open.* **11**, 100530 (2023). PMID: 37920681; PMCID: PMC10618688.
20. Cowen, J. E., Vigneswaran, G., Bekker, J., Brennan, P. A. & Oeppen, R. S. Human factors in diagnostic radiology: practical challenges and cognitive biases. *Eur. J. Radiol.* **190**, 112248. 10.1016/j.ejrad.2025.112248 (2025). Epub 2025 Jun 17. PMID: 40543287.
21. Neri, E., Coppola, F., Miano, R. & Grassi, R. Artificial Intelligence: Who Is Responsible for the Diagnosis? *Radiol. Oncol.* **52** (2), 109–114 (2018).
22. Radiological Society of North America (RSNA) Ultrasound Task Force. Structured Reporting in Ultrasound: Improving the Quality and Communication of Diagnostic Imaging Examinations. *Radiology* **290** (2), 370–374 (2019).

Figures

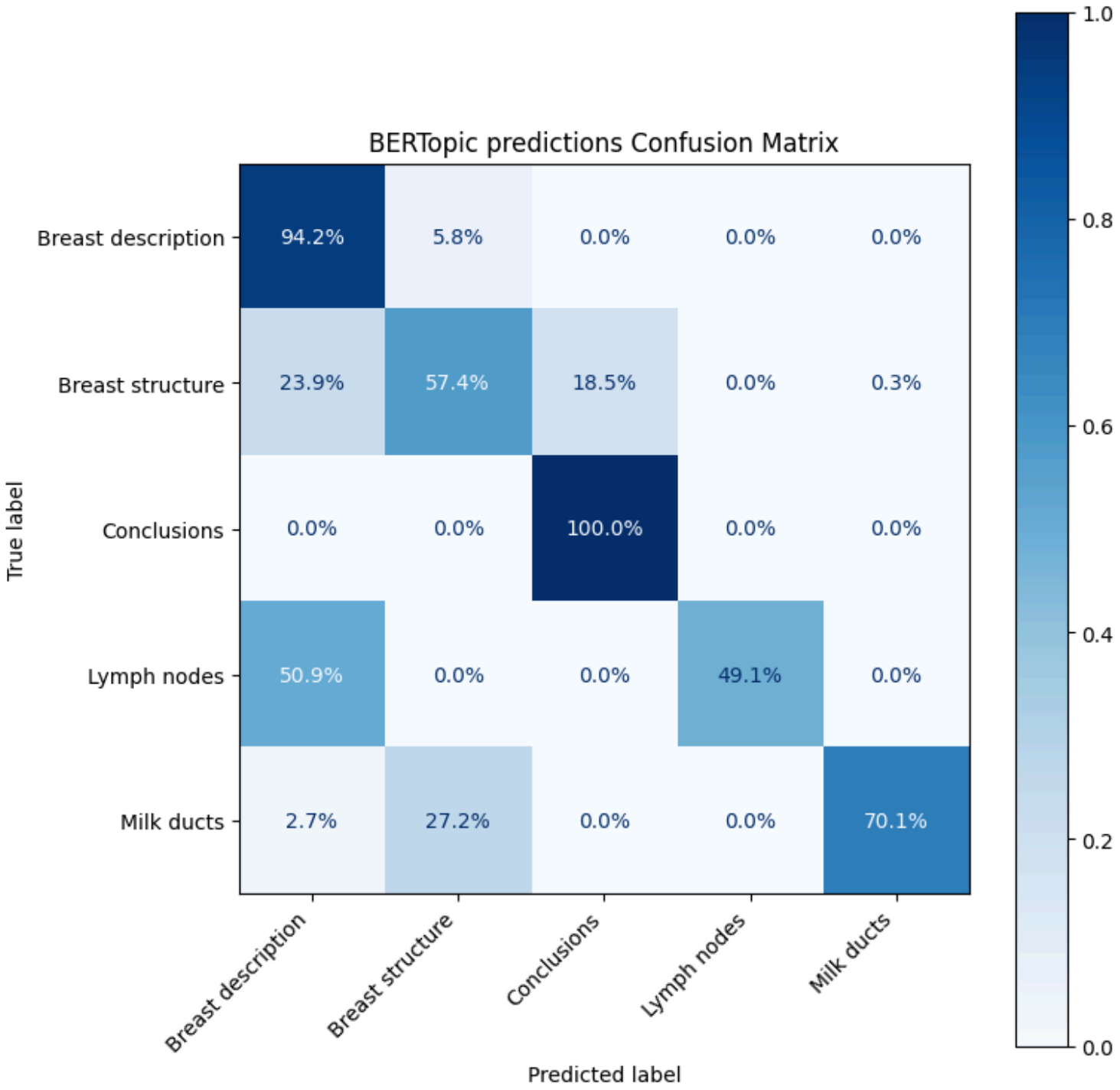


Figure 1

Normalized Confusion Matrix for BERTopic model predictions

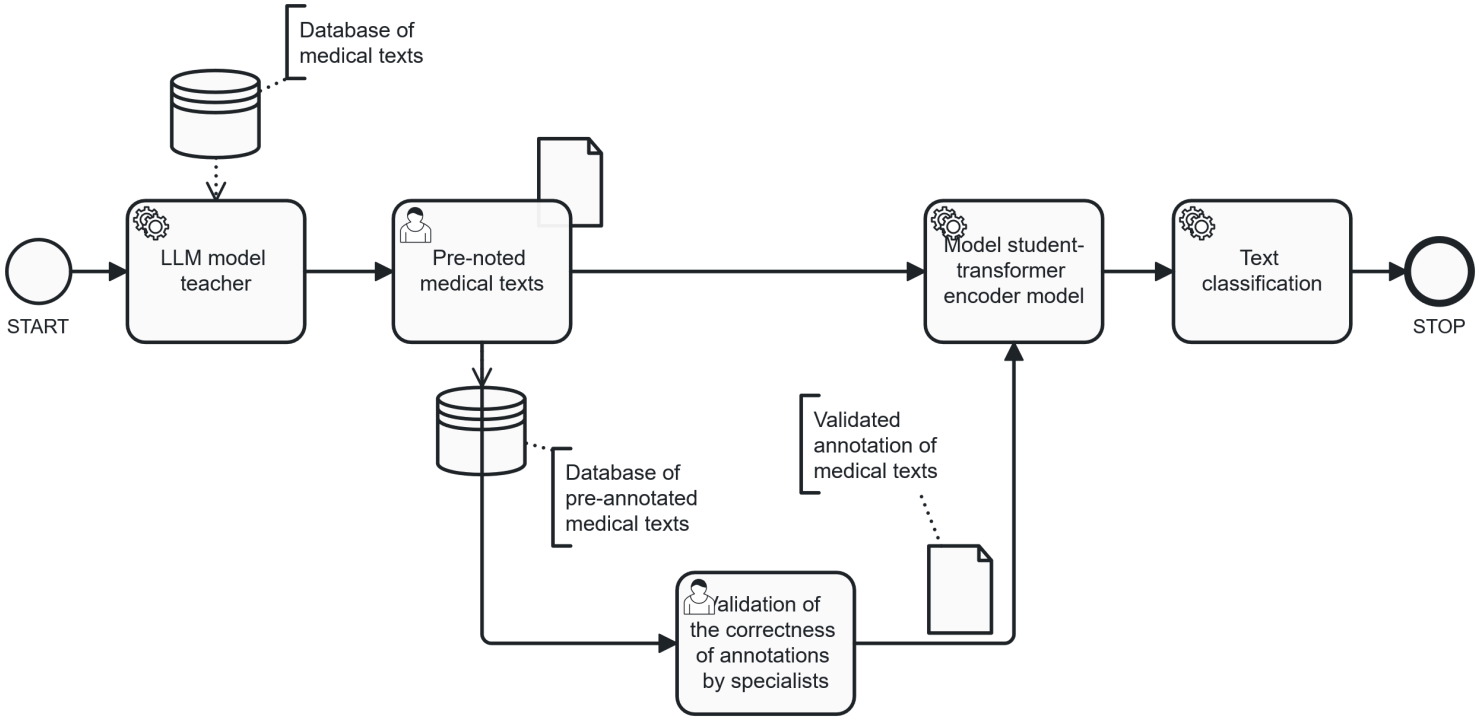


Figure 2

Visualization of the data tagging and model training process

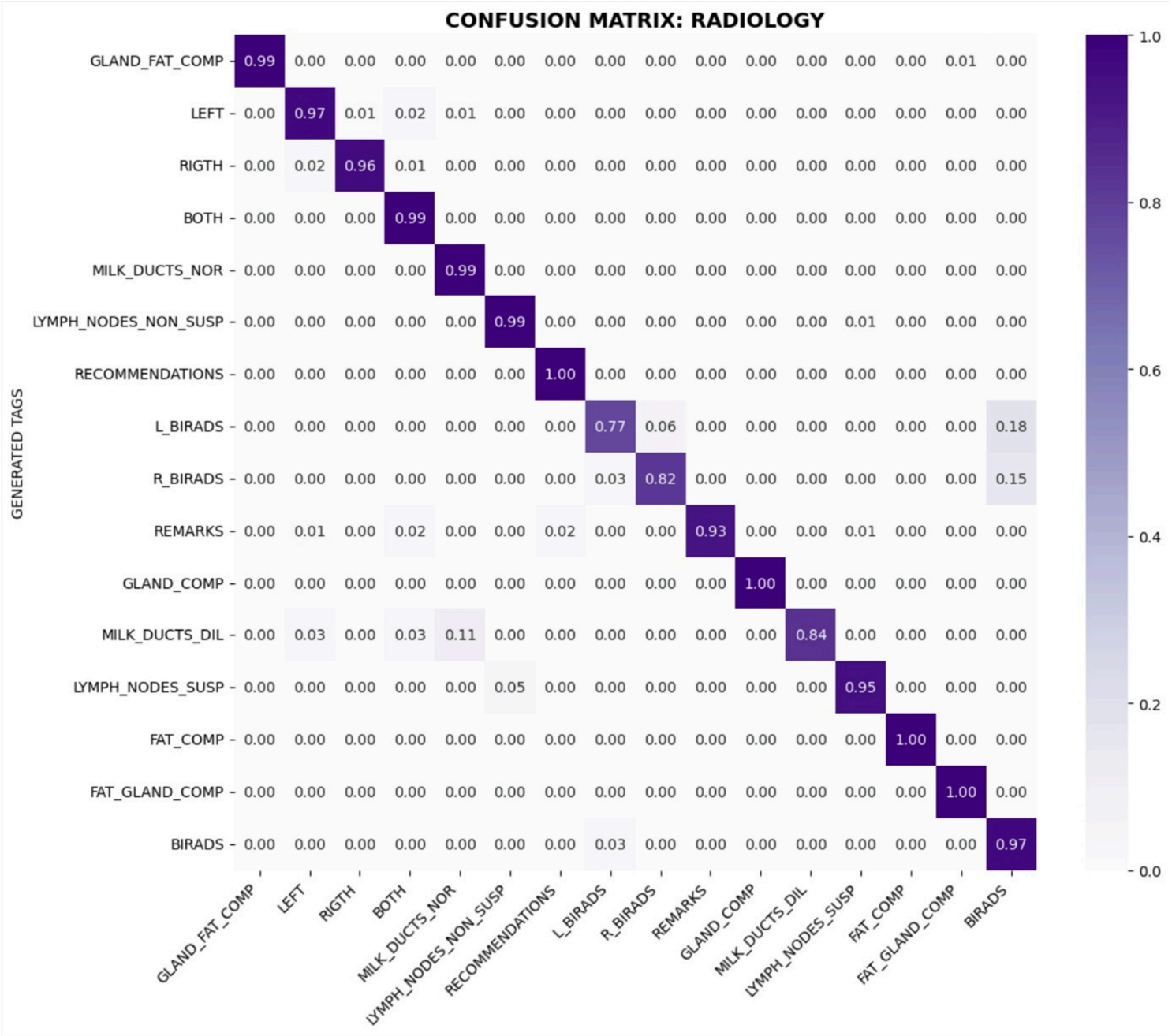


Figure 3

Confusion matrix of a trained classifier model on a test set for Radiology.

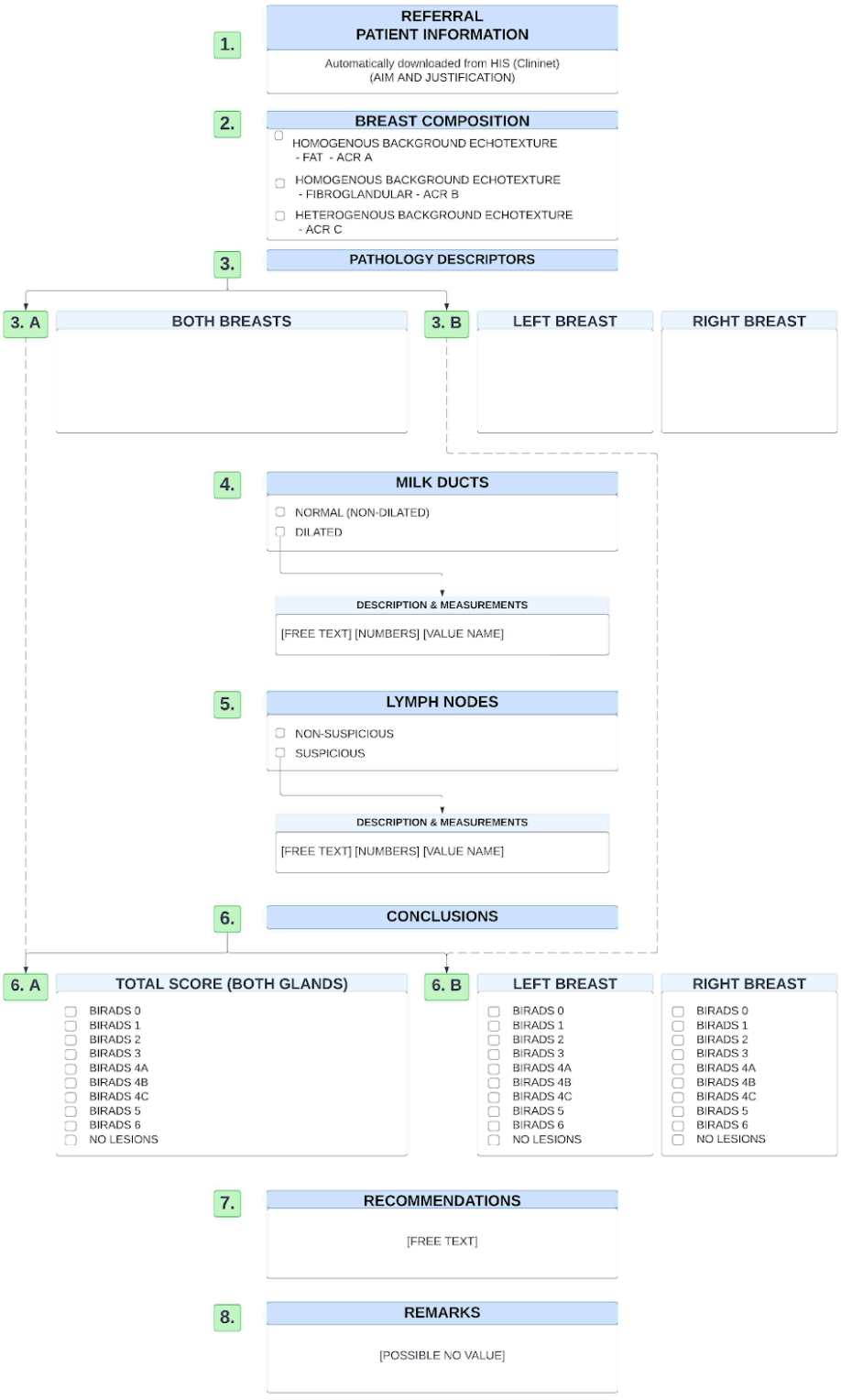


Figure 4

Structure of the form for radiological examination

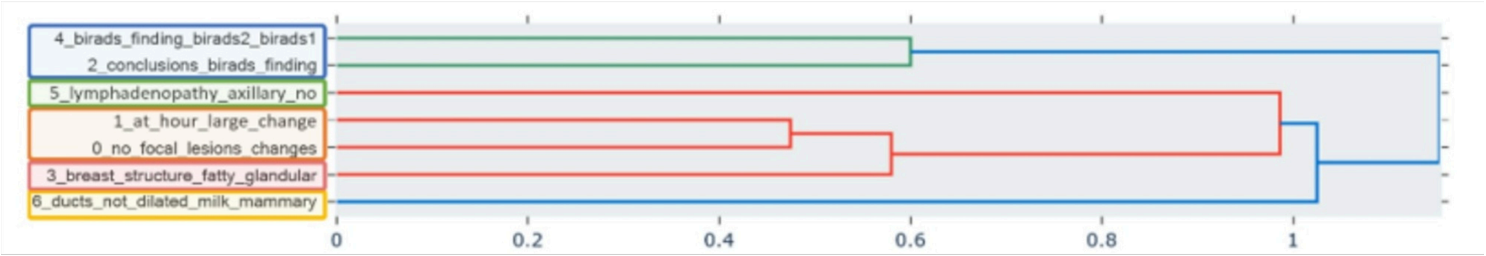


Figure 5

Defined by clinicians' structure of breast radiology reports hierarchical clustering.