BMC Medical Informatics and Decision Making

RESEARCH

Open Access

MISTIC: a novel approach for metastasis classification in Italian electronic health records using transformers

Livia Lilli[1,2†], Mario Santoro[3†], Valeria Masiello[1], Stefano Patarnello[1*], Luca Tagliaferri[1,2], Fabio Marazzi[1] and Nikola Dino Capocchiano[1]

Abstract

Background Analysis of Electronic Health Records (EHRs) is crucial in real-world evidence (RWE), especially in oncology, as it provides valuable insights into the complex nature of the disease. The implementation of advanced techniques for automated extraction of structured information from textual data potentially enables access to expert knowledge in highly specialized contexts. In this paper, we introduce MISTIC, a Natural Language Processing (NLP) approach to classify the presence or absence of metastasis in Italian EHRs, in the breast cancer domain.

Methods Our approach consists of a transformer-based framework designed for few-shot learning, requiring a small labelled dataset and minimal computational resources for training. The pipeline includes text segmentation to improve model processing and topic analysis to filter informative content, ensuring relevant input data for classification.

Results MISTIC was evaluated across multiple data sources, and compared to several benchmark methodologies, ranging from a pattern-matching system, composed of regex and semantic rules, to BERT-based models implemented in a zero-shot learning setup and Large Language Models (LLMs). The results demonstrate the generalization of our approach, achieving an F-Score above 87% on all the sources, and outperforming the other experiments, with an overall F-Score of 91.2%.

Conclusions MISTIC achieves high performance in the Italian metastasis classification task, outperforming rule-based systems, zero-shot BERT models, and LLMs. Its few-shot learning setup offers a computationally efficient alternative to large-scale models, while its segmentation and topic analysis steps enhance explainability by explicitly linking predictions to key textual elements. Furthermore, MISTIC demonstrates strong generalization across different data sources, reinforcing its potential as a scalable and transparent solution for clinical text classification. By extracting high-quality metastatic information from diverse textual data, MISTIC supports medical researchers in analyzing unstructured and highly informative content across a wide range of medical reports. In doing so, it enhances data accessibility and interpretability, addressing a critical gap in health informatics and clinical practice.

†Livia Lilli and Mario Santoro contributed equally to this work.

*Correspondence:
Stefano Patarnello
stefano.patarnello@gemelligenerator.it

Full list of author information is available at the end of the article

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Keywords Metastatic breast cancer, Natural language processing, Sentence transformer, Large language model, Few shot learning, Electronic health record

Introduction

Real World Evidence (RWE) studies are becoming relevant in healthcare, as they derive insights directly from routine medical practice. To ensure reliability, a large-scale analysis of patients is typically required, necessitating the extensive collection of healthcare data. In this context, processing of Electronic Health Records (EHRs) is crucial in RWE [1], especially in oncology [2], as it provides valuable insights into the disease's complex nature, encompassing patient history, symptoms, treatments, and laboratory parameters. However, inherent variability in EHR formats, medical terminologies, and linguistic nuances poses a significant hurdle to extracting consistent and reliable information. The introduction of advanced techniques for automated comprehension of free text reports and extraction of structured information from textual data presents both an opportunity and a challenge, potentially enabling access to expert knowledge in highly specialized contexts at large scale. An example of application is the NLP-based algorithm developed to extract the occurrence of Breast Cancer metastasis from medical reports of different types (e.g., diagnostics, radiotherapy treatment, follow-up visits) [3]. Indeed, within this complex realm of knowledge, identification of events related to Metastatic Breast Cancer represents a pivotal factor in determining disease stage and treatment strategies in clinical practice [4, 5].

Artificial Intelligence (AI) has significantly advanced human language comprehension, particularly in Natural Language Processing (NLP), which is experiencing rapid growth. Within NLP, various methodologies are employed for tasks such as text classification, named entity recognition, text summarization, and simplification. In this context, convolutional neural networks (CNNs), pattern-matching systems, and Large Language Models (LLMs) are the principal methods, proposed in the literature, to automate and enhance EHR analysis.

In this direction, several NLP studies highlight the adaptability of Convolutional Neural Networks (CNNs), known in the oncological domain for their efficiency in sequential tasks like image classification [6–9], and also extended to NLP applications, demonstrating their effectiveness in processing textual data within healthcare contexts [10]. However, CNNs may face challenges in effectively capturing contextual relationships across distant word tokens in text. Additionally, the computational complexity inherent in designing and fine-tuning ad hoc CNNs for text processing makes us exclude them as a comparison method in this paper.

State-of-the-art approaches also cover regex and pattern-matching systems for extracting information from free text [11–13]. Specifically for the Italian language, a regex-based information extraction pipeline was used [14], with the development of a domain-specific ontology to identify events as main diagnoses or prescribed drugs. Regex and pattern-matching systems are undoubtedly helpful in building explainable information extraction pipelines. However, they present several disadvantages, like the need for complex human intervention in rule development and the low generalization over different data sources and outcomes. Indeed, the same system is not easily reusable for data with new semantic structures, and rules are not valid for other kinds of outcomes.

The above methodologies, such as CNNs and pattern matching, have long been used for text analysis tasks, each with strengths and limitations. In this paper, we explore novel approaches that recently emerged in the NLP literature, thus introducing the fundamental concepts of transformers, sentence transformers, and large language models. Unlike traditional methods, transformers are deep learning model architectures renowned for capturing long-range dependencies in text. Sentence Transformers (ST) represent a specialized type of transformer model tailored for encoding entire sentences or text passages into fixed-dimensional vectors, enabling tasks like semantic similarity measurement and text classification. When trained over a large amount of data and parameters, transformer-based models are also known as large language models, and they represent the predominant NLP approach in recent research, especially in the clinical domain, for the extraction of information from EHRs [15–22]. In literature, the English language is certainly the most explored in the NLP clinical domain, but there exist several works also implemented in other languages, such as Spanish [23], Portuguese [24], or Chinese [25, 26], with no broad evidence of Italian approaches.

Despite the great potential of these models, transformers are limited in terms of maximum text length, a critical point in EHRs, that can contain a large number of word tokens [27]. Additionally, training these models on a desired task can be computationally expensive, and a large amount of labelled data is typically required. Therefore, exploiting NLP techniques with enhanced generalization capabilities at low computational costs is crucial for generating RWE from unstructured knowledge. Moreover, in the absence of specialized large models trained on Italian clinical data for information extraction tasks, there is a pressing need for innovative

solutions to navigate the semantic complexity of Italian EHRs.

In this paper, we introduce Metastases Italian Sentence Transformers Inference Classification (MISTIC), a NLP approach which leverages a transformer-based model specifically trained on breast cancer metastasis classification within Italian EHRs, in a low computational environment (Fig. 1). The study addresses the issue of limited labelled data for training, running in a few-shot learning setup. Additionally, MISTIC wants to provide a comprehensive pipeline, where transformer-based classification is integrated with a pre-processing stage including corpus selection, data segmentation and topic detection, in order to enhance the quality of input data.

Our work aims to demonstrate the power of MISTIC in efficiently extracting breast cancer metastasis from Italian EHRs, by comparing it with several benchmarks, ranging from a rule-based system to zero-shot classification BERT models and generative LLMs.

Materials and methods

Data corpus

Data for this study were selected from an extensive collection of clinical reports of the Gemelli Hospital in Rome, specifically focusing on patients diagnosed with breast cancer [28]. During the training and evaluation phases, a team of physicians guided the selection of Electronic Health Records (EHRs), prioritizing the most informative sources for extracting details related to tumor metastasis. The selected documents included clinical diaries, medical histories, and radiodiagnostic notes, which provide direct insights into the patient's historical and current health status. These sources are particularly valuable as metastasis-related information is typically explicitly mentioned within their textual content.

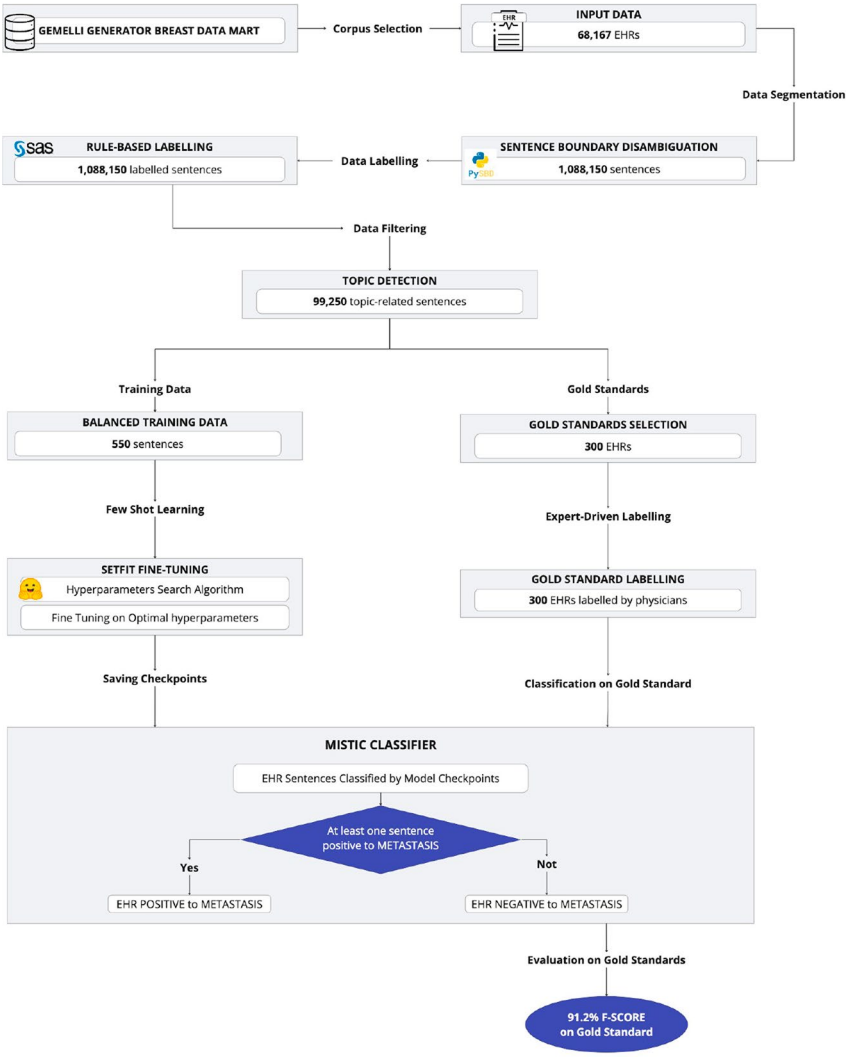

Fig. 1 MISTIC pipeline

Data segmentation

The length of EHRs vary significantly, and longer texts often present conflicting information regarding the presence of metastasis. A single clinical report may contain both positive and negative indications, depending on the anatomical location, which can mislead the overall classification of the EHR. Additionally, transformer-based models impose a maximum token limit, restricting the amount of text that can be processed at once. While techniques such as chunking or sliding windows can mitigate this constraint by splitting long documents into smaller segments, they introduce the risk of losing semantic coherence. Fragmenting the text in this manner may prevent the model from capturing contextual dependencies, potentially leading to incomplete or inaccurate extraction of the target information.

To address these challenges, we implemented a text segmentation approach on our EHR corpus, enabling the processing of clinical reports at the sentence level. For segmentation, we utilized the Python package PySBD [29], a rule-based sentence boundary disambiguation tool that supports multiple languages, including Italian. PySBD provides a dedicated model for Italian, ensuring more accurate sentence segmentation by accounting for language-specific linguistic structures and punctuation rules.

Rule-based labelling

The model fine-tuning was performed using a silver standard dataset, an automatically generated annotated corpus that, while not manually labelled, provides a useful training resource despite some degree of labelling noise. This dataset was created through a rule-based system implemented with SAS® Visual Text Analytics 8.3 [30], allowing for efficient and scalable annotation of segmented EHRs.

The system was designed to annotate each input text by determining the presence or absence of metastasis, leveraging rules developed in Language Interpretation for Textual Information (LITI). These rules were crafted to identify linguistic patterns, considering word proximity, sentence structure, and regular expressions to enhance accuracy. The lemmas listed in Table 1 serve as the core patterns searched within texts to extract metastasis-related information. These lemmas are categorized into five distinct semantic groups, encompassing lesions, nodules, high metabolic activity, staging terminology, and direct mentions of metastasis, ensuring comprehensive detection across different linguistic expressions. The SAS pipeline was implemented on a secure hospital server, accessible exclusively within the institution's internal network, restricting external access. However, Expression 1 provides an example of the rule-based approach used for detecting nodules. This rule set accounts for all possible morphological variations of the lemma to ensure robust identification.

Table 1 Domain ontology with the relations among semantic categories and related Italian lemmas

Semantic category	Italian Lemma
Metastasis	*metas, secondar*
Lesion	*lesion*
Nodule	*nodul*
High metabolic activity (HMA)	*elevata attività metabolica*
TNM staging	*m +*

Listing 1 Rules for the nodule concept

```
CONCEPT: nodulo@
CONCEPT: nodulazion@
CONCEPT: formazione@ nodulare@
REGEX: m[ai] cronodul [io]
CLASSIFIER: linfoadenopatia
```

Further examples of rules are reported in Appendix A, in the Supplementary Materials (Subsect. 4), where all the concepts used to assess the presence of nodules are shown.

Topic detection

A regex-based topic detection was implemented in the pipeline to enhance the classifier. It ensured that only informative text segments containing the anchor lemmas from Table 1 were used as input to the model training and inference. By filtering out non-informative content, this method aimed to reduce the risk of false positives and prevent irrelevant data from misleading the model.

Modeling

For the modeling step, we employed the Sentence Transformer Fine-Tuning (SetFit) Framework [31], an advanced few-shot learning approach for text classification. SetFit allows fine-tuning a pre-trained sentence transformer on a specific classification task using only a small number of training samples. In our study, we trained the model on the silver sentences, selected in the previous topic detection phase (Subsect. 2.4) to distinguish between the presence and absence of metastasis.

The classification process was then performed at the sentence level using the fine-tuned model. To determine the final classification of each EHR, we applied an OR logic approach: if at least one sentence was classified as positive for the presence of metastasis, the entire EHR was labelled as positive. Conversely, if no positive sentences were detected, the EHR was classified as negative for metastasis.

Our approach is implemented independently of the patient-level view, classifying each EHR separately without considering previous or future visits for the same patient. However, from a clinical perspective, our method

can help track a patient's history by classifying all its sequence of visits. The first report classified as positive marks the time point when metastasis is first detected. Subsequent reports may reference metastasis in different ways, reflecting disease progression or treatment effects. Classifying each report individually is then essential to accurately reconstruct the patient's medical history over time.

Results

Dataset

We applied our MISTIC pipeline by selecting a set of 68,167 EHRs from the Breast Data Mart of the Italian Gemelli Hospital of Rome [28]. The dataset consists of different sources, including clinical diaries, medical history, and radio-diagnostic reports, considered the most informative texts to extract metastasis information. An overview of the EHRs distribution over the three sources is given in Fig. 2, where part (a) shows a prevalence of radio-diagnostic reports with a coverage of 50.64% overall the dataset, followed by a 34.06% of medical histories and a 15.3% of clinical diaries.

Text segmentation

In the pre-processing phase, we implemented a segmentation step using the Python PySBD package [29]. From this phase, MISTIC generated a total of 1,088,150 sentences, where EHRs present a median of 12 sentences, with a first and third quartile equal to 7 and 17 respectively (Fig. 2, part (b)). We also analyzed the token distribution to assess the advantages of processing sentences compared to full EHRs. For this purpose, we used the BERT tokenizer from the Sentence Transformer library [32]. Our findings show that EHRs have a median count of 308, with the first and third quartiles at 205 and 510 tokens, respectively. In contrast, sentences contain

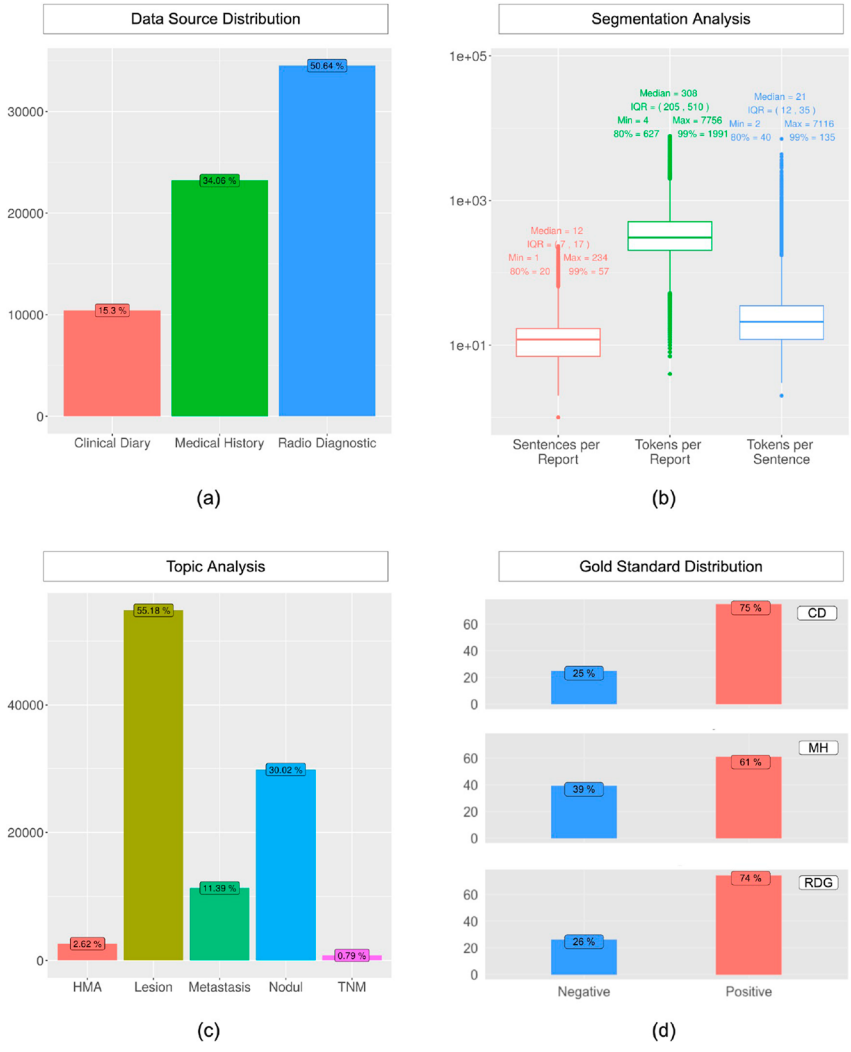

Fig. 2 (**a**) Distribution of the input EHRs over the three sources; (**b**) distribution of EHRs in terms of: sentences per report, tokens per report and tokens per sentence; (**c**) distribution of the semantic categories over the topic-filtered sentences; (**d**) gold standard outcome distribution by data source

significantly fewer tokens, with a median of 21 and an interquartile range (IQR) between 12 and 35. Notably, only 222 sentences (0.02% of the total sentences) exceed the 512-token limit, while 16942 reports (24.85% of the total EHRs) exceed the predefined 512 tokens. The above distributions are shown in Fig. 2, part (b). Building on this comparison, our approach aims to reduce internal contradictions by generating shorter and more concise texts. Additionally, it helps to mitigate the risk of text truncation caused by the token limitations of the sentence transformer model.

Topic analysis

In the topic detection phase, we used Italian lemmas from Table 1 to filter the most informative EHR sentences containing information about metastasis. This process resulted in the selection of 99,250 sentences whose lemma distribution is shown in Fig. 2, part (c). The figure shows that lemmas about lesion and nodule semantic categories are the most frequent overall the selected sentences, with percentages of 55.18% and 30.02% respectively. On the contrary, metastasis, high metabolic activity and staging represent for the minimal part the presence or absence of metastasis, with percentages of 11.39%, 2.62% and 0.79% respectively.

Training

For the training phase, we experimented two fine-tuning setups, using 550 and 1100 segments, on the paraphrase-multilingual-mpnet-base-v2[1] [32] ST model, available on Huggingface [33]. To build the training set, we started with silver-labeled sentences, which were first annotated by the rule-based classification system and then filtered through topic detection. The rule-based system assigned to each sentence a label indicating either the presence or absence of metastasis according to the five semantic categories defined in Table 1. These labels served as the basis for sampling and constructing the final training set. We then applied a balanced sampling approach, selecting either 50 or 100 sentences per category, depending on the training setup. For example, in the case of the "nodule" category in the 550-dataset setup, we selected 100 sentences (50 labeled by the rule-based system as indicating the presence of nodules and 50 as indicating their absence). In the 1100-segment setup, we doubled the sample size to 200 sentences (100 for presence and 100 for absence). This sampling strategy was applied across all five semantic categories, ensuring both positive and negative instances were well represented. Additionally, the familiarity category was considered as an extra sampling category to help the model distinguish between actual metastasis cases and references to metastases in family history. This process resulted in two training sets (with 550 and 1100 silver-labeled sentences respectively) both ensuring a balanced mix of positive and negative examples.

For SetFit fine-tuning, we used the official GitHub code repository[2] and first performed a grid search to find the best hyperparameters for the 550-segment training setup. The best-performing hyperparameters were directly applied also to the 1100-sample setup, without further tuning. The grid search was conducted on a machine with 10 CPU cores and 64GB RAM, with no GPU involvement. Further details on the grid search implementation, including parameter ranges and selected values, are provided in Appendix B, in the Supplementary Materials (Subsect. 4).

Evaluation

We evaluated our approach on a gold standard set of 300 EHRs, with 100 reports per data source (clinical diaries, medical histories, and radiodiagnostic reports). The reports were randomly sampled while ensuring that each source maintained a balanced outcome distribution, with at least 20% vs. 80% between positive and negative cases (Fig. 2, part(d)). Additionally, we ensured that sentences from the selected reports were not included in the training set. The reports were manually annotated by a team of physicians with expertise in oncology using a dedicated annotation dashboard. Further details on the annotation process and dashboard functionalities are provided in Appendix C, in the Supplementary Materials (Subsect. 4). During inference, each EHR was preprocessed through segmentation and topic detection, and then the sentence-level classifications were aggregated at the EHR level using a logical OR. The evaluation was finally performed by comparing the MISTIC classification against the manual annotations.

The MISTIC performance results, evaluated on both 550-segment and 1100-segment setups, are shown in Table 2, presenting overall scores as well as results divided by data source. Performance metrics are reported in terms of Precision, Recall, F1-score, and Accuracy.

Comparison models

In order to assess the performance of MISTIC, we compared it with several alternative approaches. These methods do not undergo preprocessing (segmentation and topic detection) before classification, allowing us to evaluate the effectiveness of our full MISTIC pipeline compared to isolated classification approaches.

Rule-Based Sytem We first considered a pattern-matching approach, by leveraging the same rule-based

[1] sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[2] https://github.com/huggingface/setfit

Table 2 MISTIC performance metrics fine-tuned on 550 and 1100 training samples

	FT (550 samples)				FT (1100 samples)			
Data source	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**
Clinical diary	0.923	0.960	**0.941**	0.910	0.923	0.960	**0.941**	0.910
Medical history	0.789	0.984	0.876	0.830	0.782	1.000	**0.878**	0.830
Radiodiagnostic	0.841	1.000	**0.914**	0.860	0.830	0.986	0.901	0.840
Overall	0.851	0.981	**0.912**	0.867	0.844	0.981	0.907	0.860

FT fine-tuned model

P precision, *R* recall, *F1* F1-score, *Acc* accuracy

system used to annotate the silver standard training dataset.

Zero-Shot Learning We then evaluate BERT-based models for natural language inference (NLI) in a zero-shot learning setting. To classify the texts, we provide the models with the candidate labels `presenza di metastasi` (presence of metastasis) and `assenza di metastasi` (absence of metastasis), allowing them to determine whether metastasis is present or absent. The models evaluated in this scenario include mDeBERTa-v3-base-tasksource, mDeBERTa-v3-base-MNLI-XNLI, XLM-RoBERTa-large-IT-MNLI, and Comprehend-IT. The mDeBERTa-v3-base-tasksource[3] and mDeBERTa-v3-base-MNLI-XNLI[4] models are based on the DeBERTa architecture [34, 35] and fine-tuned on multilingual zero-shot NLI tasks [36, 37]. The XLM-RoBERTa-large-IT-MNLI[5] is a multilingual RoBERTa variant [38], specifically fine-tuned on zero-shot NLI tasks using the MNLI corpus. Finally, the Comprehend-IT[6] model is a multilingual variant of T5, designed for multilingual text comprehension, including Italian.

Structured Generation We also explored large language models (LLMs) within a structured generation framework using Outlines[7] [39], a Python library designed to enforce predefined schemas and output constraints during text generation. To ensure consistency in classification, we implemented a multiple-choice setup, where the model was prompted to select between two predefined labels: “positive" (indicating the presence of metastasis) and “negative" (indicating its absence). The following prompt was used:

```
You are a physician specialized in breast cancer. Given the following clinical report written in Italian language, state if the patient is positive or negative to metastasis. Clinical report: {text}
```

[3] sileod/mdeberta-v3-base-tasksource-nli

[4] MoritzLaurer/mDeBERTa-v3-base-mnli-xnli

[5] Jiva/xlm-roberta-large-it-mnli

[6] knowledgator/comprehend_it-multilingual-t5-base

[7] https://github.com/dottxt-ai/outlines

Table 3 MISTIC pipeline results compared to the other state-of-the-art approaches

Model	***P***	**R**	**F1**	**Acc**
Rule-based system				
Pattern-matching	**0.961**	**0.710**	**0.816**	0.777
Zero-shot learning				
mDeBERTa-v3-base-MNLI-XNLI	0.746	0.810	0.776	0.673
mDeBERTa-v3-base-tasksource	0.712	0.848	0.774	0.653
Comprehend-IT	0.702	1.000	**0.825**	0.703
XLM-RoBERTa-large-IT-MNLI	0.737	0.814	0.774	0.667
Structured generation				
Llama 3.2 3B	0.706	0.595	0.645	0.543
Minerva 3B	0.723	0.648	**0.683**	0.580
Mixtral 7B	0.720	0.562	0.631	0.540
Few-shot learning				
sentence-bert-base-italian-uncased	0.873	0.952	0.911	0.870
MISTIC	**0.851**	**0.981**	**0.912**	0.867

P precision, *R* recall, *F1* F1-score, *Acc* accuracy

The generative LLMs evaluated were Llama 3.2 3B, Minerva 3B, and Mixtral 7B. The Llama 3.2 3B[8] model is a lightweight version of Meta’s advanced language models, known for their efficiency and strong NLP performance across various tasks [40]. The Minerva 3B[9] is an Italian-specific LLM, pre-trained to enhance comprehension and generation in the Italian language. Finally, the Mixtral 7B[10] is a mixture of experts (MoE) model, designed to optimize the trade-off between performance and computational efficiency [41].

Few-Shot Learning In addition to these approaches, we also evaluated the overall MISTIC pipeline with an alternative sentence encoder, sentence-bert-base-italian-uncased[11], a BERT-based model optimized for sentence embeddings in Italian, trained on diverse textual sources to enhance semantic similarity tasks. To assess its impact, we repeated the fine-tuning process in the 550-segment setup, replacing the original sentence transformer in the MISTIC pipeline with this model. These and all the above results are evaluated in terms of Precision, Recall, F1-score, and Accuracy, and are reported in Table 3.

[8] meta-llama/Llama-3.2–3B

[9] sapienzanlp/Minerva-3B-base-v1.0

[10] mistralai/Mixtral-8×7B-v0.1

[11] nickprock/sentence-bert-base-italian-uncased

Results and discussion

The results in Table 2 show that, across both MISTIC fine-tuning setups, clinical diary achieves the highest F1-score (0.941), while Medical History presents the most challenging classification scenario, with an F1-score of 0.876 in the 550-sentence setup, improving slightly to 0.878 in the 1100-sentence setup. This aligns with real-world clinical practice, where Medical History reports are typically long, semantically complex, and contain diverse types of information spanning multiple time periods. For medical history, the 1100-sentence setup performs better, suggesting that additional training data helps in complex cases. However, for clinical diary and radio diagnostic, the 550-sentence setup yields better results, also leading to a higher overall F1-score (0.912 vs. 0.907). This suggests that increasing the training size does not always improve performance, likely because additional data introduces noise rather than informative content, possibly leading to overfitting. However, MISTIC demonstrated good generalization capabilities, maintaining F1-Score above 87% over different data sources.

The results in Table 3 confirm that MISTIC is the best-performing approach, achieving the highest F1-score (0.912). When replacing the sentence encoder with sentence-bert-base-italian-uncased, performance remains high (F1 = 0.911), showing that the few-shot learning pipeline remains consistently over 90%. This highlights the robustness of the MISTIC framework, even with different sentence embeddings. The structured generation approach using LLMs performs the worst, with F1-scores ranging from 0.631 (Mixtral 7B) to 0.683 (Minerva 3B). This suggests that generative models are not well-suited for text classification tasks, likely due to their tendency to hallucinate and lack of direct optimization for classification. Adaptation strategies such as in-context learning or massive fine-tuning could potentially improve their performance. The zero-shot BERT-based models outperform generative LLMs, achieving F1-scores between 0.774 and 0.825, but still lower than MISTIC. Their advantage, however, lies in the absence of training, making them faster and easier to implement, though at the cost of lower classification accuracy. Finally, the pattern-matching rule-based system achieves a competitive F1-score of 0.816, demonstrating its effectiveness as a strong baseline. However, its recall (0.710) is significantly lower than that of MISTIC, indicating that it fails to capture a portion of metastasis-positive cases, making it less reliable for comprehensive classification.

We have seen MISTIC outperforming other comparison models in terms of performance metrics, but its advantages go beyond accuracy. Table 4 highlights key factors such as training data needs, explainability, generalization, computational demand, automation and manual effort. Rule-based systems require manual rule creation, making them highly explainable but lacking generalization, as they are tailored to specific sources. Zero-shot BERT models and LLMs, implemented without training, are fully automated but not explainable, as their reasoning cannot be traced. While their context mechanisms enable generalization, they are computationally demanding, especially LLMs.

Table 4 Comparison of different approaches in terms of key factors

Criterion	Zero-shot BERT	Generative LLM	Rule-based	MISTIC
Training data	No	No	Predefined rules	Few
Explainability	No	No	Yes	Yes
Generalization	Yes	Yes	No	Yes
Computational demand	Moderate	High	Low	Moderate
Automation	Full	Full	Partial	Full
Manual effort	Low	Low	High	Low

MISTIC combines strong generalization across data sources, with moderate training needs (few-shot environment) and full automation, making it a strong alternative to rule-based, zero-shot BERT, and generative LLM models. Additionally, its segmentation and topic detection steps enhance explainability by breaking down clinical text into meaningful segments and identifying key metastasis-related lemmas. This structured approach ensures that classification decisions are made on specific, relevant portions of texts. By linking predictions to distinct textual elements, the model provides transparency into its decision-making process, allowing researchers to trace how and why a particular classification was made. The topic analysis in Fig. 2 part(c) also shows the key-lemmas distribution over text segments, demonstrating how the modelling process is controlled by focusing on informative segments in both the training and inference phases. This approach reduces noise, ensuring that the model is trained and applied only to clinically relevant text, improving both accuracy and interpretability. From a computational perspective, MISTIC maintains low resource requirements, running without a GPU. The grid search and fine-tuning took 27 h and 1 h, respectively, on a 10-core machine with 64 GB RAM, while inference required only few seconds. While not included in our experiments, there exist other few-shot learning approaches discussed in the literature, such as meta-learning [42] and prototypical networks [43], which typically prioritize performance but at the cost of less interpretability, making their decision-making process not transparent. Furthermore, although MISTIC operates in a fully automated manner, its explainability supports post-hoc validation and expert review, aligning with principles of Evaluative AI [44, 45], which emphasize the importance of transparency and clinical oversight.

These results reinforce MISTIC as a practical solution for clinical text classification, achieving a balance among generalization, computational effort, explainability and automation.

In terms of comparison with the literature, in another study [21], we applied a LLaMA [46] instruction-tuning for metastasis classification, ensembling it with a BERT-based classifier. While achieving an F-score of 88.8%, its performance was lower than MISTIC's 91.2%, highlighting the advantages of using ST in a few-shot learning pipeline, with integrated preprocessing steps. Similarly, other studies have explored language-specific models for breast cancer feature extraction, such as CancerBERT [47], an approach for the English clinical reports which achieves performance up to 90.4%, and a BERT-based approach for Spanish, with a F-Score above 93% [23]. While MISTIC does not always achieve the highest metrics compared to the other studies, it shows great potential, especially considering that Italian is a minor language, less represented in medical NLP research.

Building on these results, we aim to expand our pipeline to new clinical outcomes and explore its application in other pathological domains, while continuing to focus on the Italian language, which remains less explored compared to other major languages in medical NLP. Furthermore, we plan to integrate these tools within our hospital RWE, improving the automation and scalability of clinical text analysis to support and improve both research and patient care.

Conclusions

This paper presents MISTIC, a transformer-based classifier fine-tuned for breast metastasis classification in a few-shot learning framework. To evaluate its effectiveness, we compared it against a pattern-matching system leveraging regex-based rules, word distances, and text structure, as well as zero-shot BERT-based models, and LLMs implemented in a structured generation framework. Additionally, our pipeline was tested across different training setups of varying training size, and evaluated on multiple data sources, to verify its robustness and adaptability in diverse clinical texts.

Results show that MISTIC achieves an F1-Score of 0.912, outperforming all other approaches while requiring minimal training data. In contrast, rule-based systems demand extensive human effort for rule development and fail to generalize to new data, despite offering some level of explainability. Likewise, zero-shot BERT models and LLMs, although fully automated, function as black boxes, deliver inferior performance, and need high computational resources. Overall, our proposed pipeline combines superior performance, robust generalization, and clear explainability without extensive manual intervention, making it a potential solution for clinical text classification.

To further enhance its usability, we plan to extend MISTIC to other types of tumors and new clinical outcomes, such as disease progression and treatment changes, while continuing to focus on the Italian language, which remains underexplored in healthcare NLP. This expansion would enable the deployment of a highly adaptable tool for real-world clinical practice in the hospitals.

With its promising results, MISTIC provides a scalable and efficient solution for processing clinical information, addressing the traditionally manual and resource-intensive task of building retrospective RWE oncological datasets. By automatically extracting metastasis-related data from clinical reports, it reduces the need for extensive human effort while ensuring structured data essential for healthcare applications and observational studies. Its strong performance makes it a valuable tool for advancing research and RWE generation in oncology.

Abbreviations

MISTIC	Metastases Italian sentence transformers inference classification
NLP	Natural language processing
LLM	Large language models
ST	Sentence transformer
CNN	Convolutional neural network
LITI	Language interpretation for textual information
HER	Electronic health record
RWE	Real-world evidence

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-02994-w.

Supplementary Material 1

Acknowledgements

Not applicable.

Author contributions

LL, VM, NDC prepared the data for the study. LL, MS designed and implemented the experiments. LL, MS, VM, SP validated the results. LL, MS, SP wrote the first draft of the manuscript. All the authors wrote and reviewed the final version of the manuscript.

Funding

This study received partial funding from Italian Ministry for University and Research (MUR) under the Program PON "Research and Innovation" supporting the development of artificial intelligence platform Gemelli Generator at Policlinico Universitario A. Gemelli IRCCS".

Data availability

The code for the MISTIC implementation is available at the following GitHub repository: https://github.com/LivLilli/MISTIC.

Declarations

Ethics approval and consent to participate

The use of data for this study has been implemented in full compliance with ethics and GDPR requirements. Specifically, data usage has been approved by the Ethics Committee Policlinico Gemelli to conduct the presented research (Protocol Number: 2889256), and the de-identification of sensitive data has been performed.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests

Author details
[1]Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy
[2]Catholic University of the Sacred Heart, Rome, Italy
[3]Istituto per le Applicazioni del Calcolo "Mauro Picone", Italian National Research Council, Rome, Italy

Received: 8 May 2024 / Accepted: 2 April 2025
Published online: 10 April 2025

References

1. Bastarache L, Brown JS, Cimino JJ, Dorr DA, Embi PJ, Payne PR, Wilcox AB, Weiner MG.Developing real-world evidence from real-world data: transforming raw data into analytical datasets. Learn Health Syst. 2022;6(1):10293.
2. Benedum CM, Sondhi A, Fidyk E, Cohen AB, Nemeth S, Adamson B, Estévez M, Bozkurt S.Replication of real-world evidence in oncology using electronic health record data extracted by machine learning. Cancers. 2023;15(6):1853.
3. Buckley JM, Coopey SB, Sharko J, Polubriaginof F, Drohan B, Belli AK, Kim EM, Garber JE, Smith BL, Gadd MA, et al. The feasibility of using natural language processing to extract clinical information from breast pathology reports. J Pathol Inform 2012;3:23.
4. Scully OJ, Bay B-H, Yip G, Yu Y. Breast cancer metastasis. Cancer genomics & proteomics. 2012;9(5):311–20.
5. Lu J, Steeg PS, Price JE, Krishnamurthy S, Mani SA, Reuben J, Cristofanilli M, Dontu G, Bidaut L, Valero V, et al. Breast cancer metastasis: challenges and opportunities. Cancer Res 2009;69:4951–53.
6. Ting FF, Tan YJ, Sim KS. Convolutional neural network improvement for breast cancer classification. Expert Syst Appl. 2019;120:103–15.
7. Gupta K, Chawla N. Analysis of histopathological images for prediction of breast cancer using traditional classifiers with pre-trained cnn. Procedia Comput Sci. 2020;167:878–89.
8. Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, Polónia A, Campilho A.Classification of breast cancer histology images using convolutional neural networks. PLoS ONE. 2017;12(6):0177544.
9. Lilli L. A calibrated multiexit neural network for detecting urothelial cancer cells. Comput Math Methods Med. 2021;2021:5569458.
10. Wang L, Luo L, Wang Y, Wampfler J, Yang P, Liu H. Natural language processing for populating lung cancer clinical research data. BMC medical informatics and decision making. 2019;19:1–10.
11. Deshmukh PR, Phalnikar R.Anatomic stage extraction from medical reports of breast cancer patients using natural language processing. Health Technol. 2020;10(6):1555–70.
12. AAlAbdulsalam AK, Garvin JH, Redd A, Carter ME, Sweeny C, Meystre SM: Automated extraction and classification of cancer stage mentions fromunstructured text fields in a central cancer registry. AMIA summits on translational science proceedings, 2018, 16 (2018)
13. Lilli L, Bosello SL, Antenucci L, Patarnello S, Ortolan A, Lenkowicz J, Gorini M, Castellino G, Cesario A, D'Agostino MA, et al. A comprehensive natural language processing pipeline for the chronic lupus disease. In: Digital Health and Informatics Innovations for Sustainable Health Care Systems. Amsterdam: IOS Press: 2024. p. 909–13.
14. Viani N, Larizza C, Tibollo V, Napolitano C, Priori SG, Bellazzi R, Sacchi L. Information extraction from italian medical reports: an ontology-driven approach. Int J Med Inform. 2018;111:140–48.
15. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H.Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare (HEALTH). 2021;3(1):1–23.
16. Santos T, Tariq A, Das S, Vayalpati K, Smith GH, Trivedi H, Banerjee I: PathologyBERT-Pre-trained Vs. a new transformer language model for pathology domain. AMIA annual symposium proceedings, Vol. 2022. (2022)
17. Yang Z, Mitra A, Liu W, Berlowitz D, Yu H.Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. Nat Commun. 2023;14(1):7857.
18. Ji S, Hölttä M, Marttinen P. Does the magic of bert apply to medical code assignment? a quantitative study. Comput Biol Med. 2021;139:104998.
19. Adamson B, Waskom M, Blarre A, Kelly J, Krismer K, Nemeth S, Gippetti J, Ritten J, Harrison K, Ho G, et al. Approach to machine learning for extraction of real-world data variables from electronic health records. Front Pharmacol. 2023;14.
20. Solarte-Pabón O, Torrente M, Garcia-Barragán A, Provencio M, Menasalvas E, Robles V: Deep learning to extract Breast Cancer diagnosis concepts. In: 2022 IEEE 35th international symposium on computer-based medical systems (CBMS). (2022)
21. Lilli L, Patarnello S, Masciocchi C, Masiello V, Marazzi F, Luca T, Capocchiano N: Llamamts: optimizing metastasis detection with llama instruction tuning and bert-based ensemble in italian clinical reports. In: Proceedings of the 6th Clinical Natural Language Processing Workshop, pp. 162–71 (2024)
22. Lilli L, Antenucci L, Ortolan A, Bosello SL, D'Agostino MA, Patarnello S, Masciocchi C, Lenkowicz J: Lupus alberto: a transformer-based approach for sle information extraction from italian clinical reports (2024)
23. Solarte-Pabón O, Montenegro O, Garca-Barragán A, Torrente M, Provencio M, Menasalvas E, Robles V. Transformers for extracting breast cancer information from spanish clinical narratives. Artif Intell Med. 2023;143:102625.
24. Schneider ETR, Souza JVA, Knafou J, Oliveira LES, Copara J, Gumiel YB, Oliveira LFA, Paraiso EC, Teodoro D, Barra CMCM: BioBERTpt-a Portuguese neural language model for clinical named entity recognition. In Proceedings of the 3rd clinical natural language processing workshop. (2020)
25. Liu H, Zhang Z, Xu Y, Wang N, Huang Y, Yang Z, Jiang R, Chen H.Use of bert (bidirectional encoder representations from transformers)-based deep learning method for extracting evidences in chinese radiology reports: development of a computer-aided liver cancer diagnosis framework. J Med Internet Res. 2021;23(1):19689.
26. Cai F, Ye H: Chinese medical text classification with RoBERTa. In: International symposium on biomedical and computational biology, pp. 223–36. (2022)
27. Gao S, Alawad M, Young MT, Gounley J, Schaefferkoetter N, Yoon HJ, Wu X-C, Durbin EB, Doherty J, Stroup A, et al. Limitations of transformers on clinical text classification. IEEE J Biomed Health Inform 2021;25:3596–607.
28. Marazzi F, Tagliaferri L, Masiello V, Moschella F, Colloca GF, Corvari B, Sanchez AM, Capocchiano ND, Pastorino R, Iacomini C, et al. Generator breast datamart-the novel breast cancer data discovery system for research and monitoring: preliminary results and future perspectives. J Personal Med 2021;11:65.
29. Sadvilkar N, Neumann M. Pysbd: pragmatic sentence boundary disambiguation. 2020. arXiv preprint arXiv:2010.09657.
30. Anandarajan M, Hill C, Nolan T, Anandarajan M, Hill C, Nolan T. Sas visual text analytics. In: Practical Text Analytics: maximizing the Value of Text Data. Belin: Springer; 2019. p. 263–82.
31. Tunstall L, Reimers N, Jo UES, Bates L, Korat D, Wasserblat M, Pereg O. Efficient few-shot learning without prompts. 2022. arXiv preprint arXiv:2209.11055.
32. Reimers N, Gurevych I. Sentence-bert: sentence embeddings using siamese bert-networks. 2019. arXiv preprint arXiv:1908.10084.
33. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac, P. P, Rault T, Louf R, Funtowicz M, et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pp. 38–45 (2020)
34. He P, Liu X, Gao J, Chen W: Deberta: decoding-enhanced bert with disentangled attention. In: International conference on learning representations (2021). https://openreview.net/forum?id=XPZIaotutsD
35. He P, Gao J, Chen W. DeBERTaV3: improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. 2021. https://arxiv.org/abs/2111.09543.
36. Sileo D. tasksource: structured dataset preprocessing annotations for frictionless extreme multi-task learning and evaluation. 2023. arXiv preprint arXiv:2301.05948.
37. Laurer M, Atteveldt WV, Casas AS, Welbers K. Less Annotating, More Classifying: addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT—NLI, Preprint. Open Science Framework: 2022. Accessed 28 July 2022.
38. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave É, Ott M, Zettlemoyer L, Stoyanov V: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp. 8440–51 (2020)
39. Willard BT, Louf R. Efficient guided generation for llms. 2023. arXiv preprint arXiv:2307.09702.
40. Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, Mathur A, Schelten A, Yang A, Fan A, et al. The llama 3 herd of models. 2024;arXiv preprint arXiv:2407.21783.

41. Jiang AQ, Sablayrolles A, Roux A, Mensch A, Savary B, Bamford C, Chaplot DS, Casas DDL, Hanna EB, Bressand F, et al. Mixtral of experts. 2024;arXiv preprint arXiv:2401.04088.
42. Zhang B, Luo C, Yu D, Li X, Lin H, Ye Y, Zhang B: Metadiff: meta-learning with conditional diffusion for few-shot learning. In: Proceedings of the AAAI conference on artificial intelligence, vol. 38, pp. 16687–95 (2024)
43. Lyu Q, Wang W: Compositional prototypical networks for few-shot classification. In: Proceedings of the AAAI conference on artificial intelligence, vol. 37, pp. 9011–19 (2023)
44. Miller T: Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. In: Proceedings of the 2023 ACM conference on fairness, accountability, and transparency. FAccT'23, pp. 333–42. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3593013.3594001. https://doi.org/10.1145/3593013.3594001
45. Barredo Arrieta A, Dã-az-rodrã-guez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F. Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai. Inf Fusion. 2020;58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012.
46. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, et al. Llama 2: open foundation and fine-tuned chat models. 2023;arXiv preprint arXiv:2307.09288.
47. Zhou S, Wang N, Wang L, Liu H, Zhang R.Cancerbert: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. J Am Med Inform Assoc. 2022;29(7):1208–16.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.