A Novel Real-World Dataset of Italian Clinical Notes for NLP-based Decision Support in Low Back Pain Treatment

Agnese Bonfigli[1,2,†], Ruben Piperno[1,2,†], Luca Bacco[1,2,*], Felice Dell'Orletta[2], Dominique Brunato[2], Filippo Crispino[1], Giuseppe Francesco Papalia[3], Fabrizio Russo[3], Gianluca Vadalà[3], Rocco Papalia[3], Mario Merone[1,3] and Leandro Pecchia[1,3]

[1]Research Unit of Intelligent Health-Technologies, Department of Engineering, Università Campus Bio-Medico di Roma, Via Alvaro del Portillo 21, 00128 Rome, Italy

[2]ItaliaNLP Lab, Institute of Computational Linguistics "Antonio Zampolli", National Research Council, Via Giuseppe Moruzzi 1, 56124 Pisa, Italy

[3]Fondazione Policlinico Universitario Campus Bio-Medico, Via Alvaro del Portillo 200, 00128 Rome, Italy

Abstract

Low back pain represents a leading source of disability worldwide and poses a significant challenge for evidence-based clinical decision support. In contexts where Italian-language resources for diversified therapeutic pathways are lacking, we have assembled a novel, annotated dataset comprising up to three pre-treatment documents per patient (MRI report, X-ray report, and patient visit notes), alongside demographic information (age and sex). The cohort consists of 176 patient records, stratified into three therapeutic groups: 50 conservative, 92 regenerative, and 34 surgical.

The primary aim is to investigate whether the collected dataset can be harnessed to predict which of the three treatment modalities is most appropriate. To this end, six document-combination scenarios were defined, evaluating each single-report modality as well as all possible pairings. For each scenario, two modeling strategies were contrasted: a traditional Support Vector Machine classifier leveraging TF–IDF features based on unigrams, bigrams, and trigrams, and a fine-tuned Italian BERT model adapted to our corpus.

Experimental results indicate that classic n-gram–based approaches achieve the highest performance (macro–F_1 up to 71.3%). The BERT model, while outperforming the baseline, encounters limitations in this low-resource scenario.These findings suggest that the present dataset has the potential to catalyze the development of Italian-language clinical decision support systems that account for the distinct signatures of treatment pathways.

Keywords

Italian Medical Corpus, Decision Support Systems, Clinical Natural Language Processing, Treatment Prediction, NLP in healthcare

1. Introduction

Low back pain (LBP) represents one of the most prevalent medical conditions globally, significantly impacting both individual well-being and healthcare systems [1, 2]. It is a considerable health problem in all developed countries and is most commonly treated in primary healthcare settings. LBP is usually defined as pain, muscle tension, or stiffness localized below the costal margin and above the inferior gluteal folds, with or without leg pain. Up to 84% of the general population will experience an episode of LBP during its lifetime, and recurrence rates are high [3].

Despite extensive research and clinical experience, determining optimal treatment strategies remains challenging due to the diverse range of available therapeutic interventions. LBP management has been extensively studied considering the aforementioned impacts on the individual patient and the community. However, there is still a gap between this information and its applications in clinical practice, particularly in the area of detailing conservative (non-invasive) management. As surgeries and interventional therapies are not recommended in most patients with acute LBP, it is important for primary care physicians (PCPs) to know the details of non-invasive treatment.

The complexity of treatment selection is compounded

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

*Corresponding author. Email: l.bacco@unicampus.it

†These authors contributed equally.

agnese.bonfigli@unicampus.it (A. Bonfigli); ruben.piperno@unicampus.it (R. Piperno); l.bacco@unicampus.it (L. Bacco); felice.dellorletta@ilc.cnr.it (F. Dell'Orletta); dominique.brunato@ilc.cnr.it (D. Brunato); f.crispino@unicampus.it (F. Crispino); g.papalia@unicampus.it (G. F. Papalia); fabrizio.russo@policlinicocampus.it (F. Russo); g.vadala@policlinicocampus.it (G. Vadalà); rocco.papalia@policlinicocampus.it (R. Papalia); m.merone@unicampus.it (M. Merone); leandro.pecchia@unicampus.it (L. Pecchia)

0009-0008-7092-2875 (A. Bonfigli); 0009-0007-7399-2636 (R. Piperno); 0000-0001-5462-2727 (L. Bacco); 0000-0003-3454-9387 (F. Dell'Orletta); 0000-0003-3256-4794 (D. Brunato); 0000-0002-4140-738X (G. F. Papalia); 0000-0002-8566-8952 (F. Russo); 0000-0001-7142-1660 (G. Vadalà); 0000-0002-6069-4871 (R. Papalia); 0000-0002-9406-2397 (M. Merone); 0000-0002-7900-5415 (L. Pecchia)

by the need to consider multiple patient-specific factors, including clinical presentation, radiological findings, and demographic characteristics.

Electronic health records (EHRs) provide a rich source of clinical data that can inform LBP treatment decisions, particularly through unstructured texts such as imaging reports (e.g., Magnetic Resonance Imaging (MRI) and X-rays) and physician notes [4, 5]. Recent advancements in natural language processing (NLP) have demonstrated significant potential in extracting meaningful clinical insights from these texts, thereby supporting data-driven, informed, and personalized decision-making in healthcare [6]. This progress has been supported by large-scale English-language datasets, such as MIMIC-CXR [7] and MIMIC-IV-Note [8], which provide radiology reports related to central and lower body axial regions. However, the development of NLP-based clinical decision support systems for LBP is significantly limited by the lack of annotated datasets, especially in languages other than English. Building language-specific datasets is critical to promoting equitable access to AI-driven healthcare innovations [9, 10] adapted to different healthcare contexts, like the Italian one.

The primary objective of this work is to develop and release a novel dataset of manually annotated Italian clinical notes for low back pain management, created in close collaboration with medical experts. This resource addresses a significant gap in biomedical NLP for the Italian language, where publicly available annotated datasets are extremely limited.

To demonstrate the potential of this dataset as a valuable tool for the BioNLP community, we conduct a set of preliminary analyses focused on the task of automated treatment recommendation. Specifically, we compare the performance of traditional machine learning methods (i.e., Support Vector Machines) and Transformer models [11] like BERT [12], with the goal of exploring how this resource can support physicians decisions.

This work thus provides two main contributions:

- The release of a new annotated dataset of Italian clinical notes for LBP treatment, offering the BioNLP community a much-needed resource for conducting research in biomedical language processing in Italian.
- A preliminary comparative study designed to evaluate the dataset's capacity to support different NLP techniques and modeling strategies, thereby validating its role as a foundation for further investigation in clinical decision support and related tasks.

2. Dataset

Data Acquisition This study is based on a retrospective analysis of anonymized clinical records collected during routine care for patients with LBP enrolled at the spine clinic of the *Fondazione Policlinico Campus Bio-Medico* in Rome. The dataset represents a pilot collection curated through a rigorous manual selection process carried out in collaboration with board-certified orthopaedic specialists. All records were obtained prior to any therapeutic intervention and reflect real-world clinical decisions made during standard care.

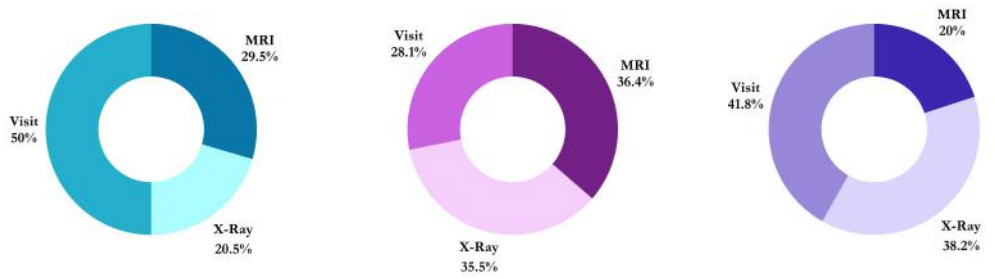
Each case was annotated by the attending physician responsible for the patient's care, linking each patient to a treatment label reflecting the therapeutic decision. Consequently, no additional annotation was necessary. For each patient, we selected the corresponding pre-treatment documents, thus creating a realistic decision-support scenario in which models are trained to predict treatment strategies based solely on clinical text available prior to intervention.

Dataset Composition The dataset reflects the real-world distribution of therapeutic strategies typically employed in orthopedic practice, clustering into three patient groups:

- **Conservative.** Patients managed non-invasively through physiotherapy, pharmacological pain control, and rehabilitative interventions designed to restore muscular strength and joint mobility;
- **Regenerative.** Patients treated with minimally invasive biologic therapies, including growth-factor injections, stem-cell preparations, or platelet-rich plasma, aimed at promoting tissue regeneration and functional recovery;
- **Surgical.** Patients who underwent operative procedures, such as spinal stabilization, to address severe pathology or persistent symptoms unresponsive to conservative care.

The dataset includes a total of **176** patients, distributed as follows: 50 conservative, 92 regenerative, and 34 surgical cases. This imbalanced distribution mirrors actual clinical practice, where non-invasive approaches are generally preferred over surgical interventions when clinically appropriate.

Each record consists of textual data from three primary clinical sources: radiological reports (MRI and X-ray) and consultation notes. MRI reports describe spinal anatomy and pathology; X-ray reports focus on vertebral alignment and bone structure; consultation notes provide narrative summaries written by orthopedic specialists during outpatient visits. Demographic variables, including age and sex, are also available for each patient.



(a) Conservative (b) Regenerative (c) Surgical

Figure 1: Percentage distribution of MRI, X-ray, and clinical visit reports across treatment categories.

Treatment Class	MRI		X-ray		Clinical Visit	
	Chars	Tokens	Chars	Tokens	Chars	Tokens
Surgical	1520.36	348.82	492.81	115.19	676.57	176.74
Regenerative	968.84	220.53	486.06	105.95	603.18	151.06
Conservative	1058.73	239.65	452.78	99.67	523.02	135.36

Table 1
Average length (in characters and tokens) of medical text across different treatment classes and note types.

An example of these reports is provided in Appendix A. Overall, the corpus is a multi-source, domain-specific collection that integrates radiologic descriptions with unstructured clinical narratives of varying information density.

The detailed composition of our dataset reveals varying distributions of textual data across treatment categories. Specifically, Figure 1 illustrates the percentage distribution of MRI, X-ray, and clinical visit reports across the three groups, while Table 1 presents the average report lengths for each category. Notably, X-ray reports and clinical visit notes exhibit similar average lengths across the treatment categories, while MRI reports show a marked difference, with surgical patients having significantly longer reports. This suggests that MRI documentation may be particularly relevant in distinguishing surgical from non-surgical cases in clinical practice [13, 14]. However, this hypothesis should be interpreted with caution, given the relatively small and imbalanced nature of the dataset, which may affect the generalizability of such findings.

3. Methods

The classification of clinical reports for LBP treatment poses specific challenges due to the linguistic complexity and domain-specific nature of medical documentation. These texts often feature highly specialized terminology, diverse narrative styles, and intricate links between diagnoses and recommended therapies. To address these challenges and to assess the suitability of our dataset, we adopted a modeling strategy that integrates both traditional machine learning techniques and modern deep learning approaches.

Our aim was to evaluate whether the combination of unstructured text and demographic data provides sufficient signal for a multiclass classification task focused on LBP treatment decisions. The classification task involves assigning each case to one of the three treatment classes, reflecting typical therapeutic pathways for LBP.

To explore how different modeling paradigms handle the specificities of the Italian medical language and the integration of heterogeneous inputs, we implemented and compared two approaches: a Support Vector Machine (SVM) with TF–IDF vectorization, and a BERT-based model fine-tuned on our dataset.

We chose these two models to contrast a strong classical method with a state-of-the-art contextual model. A linear-kernel SVM remains highly effective for text classification, especially on small or imbalanced clinical datasets where lexical cues often suffice [15]. In contrast, BERT [12] uses Transformer architectures [11] to capture deep contextual and semantic relationships, making it better suited for narrative clinical notes where meaning

depends heavily on context.

SVM Approach We developed a multiclass classification pipeline based on a SVM with a linear kernel, leveraging traditional NLP techniques to process clinical text and predict the appropriate treatment category. The pipeline begins with standard text pre-processing steps, including tokenization, stop-word removal, and lemmatization, aimed at normalizing the clinical narratives and reducing linguistic variability [16]. For feature representation strategy, we applied Term Frequency–Inverse Document Frequency (TF–IDF) vectorization using a combination of unigrams, bigrams, and trigrams. This n-gram approach enables the model to capture both individual medical terms and short multi-word expressions that frequently occur in clinical language. The TF–IDF transformation converts the unstructured reports into structured numerical representations by emphasizing terms that are particularly informative within the context of the corpus. To incorporate demographic information, patient age and sex were appended to the TF–IDF feature vectors, allowing the SVM to integrate both textual and structured data in the classification process.

BERT Approach We developed a multiclass classification pipeline based on the ***bert-base-italian-xxl-uncased model*** on Hugging Face made by Bavarian State Library[1], fine-tuned on our dataset to capture the semantic complexity of Italian clinical narratives. Each instance is constructed by concatenating one or more clinical free-text reports with patient age and sex, forming a single input sequence. No additional feature engineering is required, as the transformer architecture learns deep, context-aware representations of the sequence through self-attention mechanisms. The embedding of the [CLS] token is passed to a classification head that outputs the predicted treatment category via a softmax activation.

4. Experiments

To explore the capabilities of our dataset, we conducted a series of experiments examining how varying combinations of clinical documents and different feature-extraction techniques affect system performance. Through this systematic analysis, we identified the optimal configuration for deploying our LBP treatment-planning decision support system in the Italian healthcare setting, as illustrated in Figure 2.

[1]Model available at *https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased.*

4.1. Classification Approach

- **SVM (TF–IDF N-grams):** We implemented an SVM Classifier and evaluated three n-gram configurations with TF-IDF vectorization to extract features from Italian-language LBP clinical reports: unigrams (1-gram), bigrams (2-gram), and trigrams (3-gram). This multilevel approach enabled us to capture both individual medical terms and significant multi-word expressions commonly found in diagnostic-related documentation. The n-gram analysis proved especially effective at uncovering language-specific LBP diagnostic patterns and treatment indicators in Italian medical terminology.
- **BERT:** Rather than relying on manual feature engineering, we fine-tuned a pre-trained Italian BERT model to obtain contextualized token representations. Thanks to its multi-head self-attention mechanism, BERT inherently models the sequential dependencies among tokens, such that the order of concatenated documents (e.g., X-ray $\rightarrow$ MRI vs. MRI $\rightarrow$ X-ray) can influence prediction performance. For this BERT approach, we therefore applied the full document-combination analysis described in Section 4.2 to evaluate how different report sequences affect model accuracy [17, 12].

4.2. Document Combination Analysis

To assess the impact of our Italian LBP dataset on model performance, we systematically explored the following eight input configurations, and, for each paired setup, evaluated all possible document orders:

- **Single Document Decision Support:**
  - MRI reports
  - X-ray reports
  - Clinical visit notes
- **Paired Document Decision Support:**
  - MRI reports with clinical visit notes
  - X-ray reports with MRI reports
  - X-ray reports with clinical visit notes
- **Comprehensive Decision Support:**
  - Integration of all three document types

Patient demographic (age and sex) are appended as additional input information at the end of the selected (concatenation of) documents.

Patient Cohort: As this study reflects the real-world clinical scenario, not every patient in the registry possesses the full set of imaging and clinical documents. For

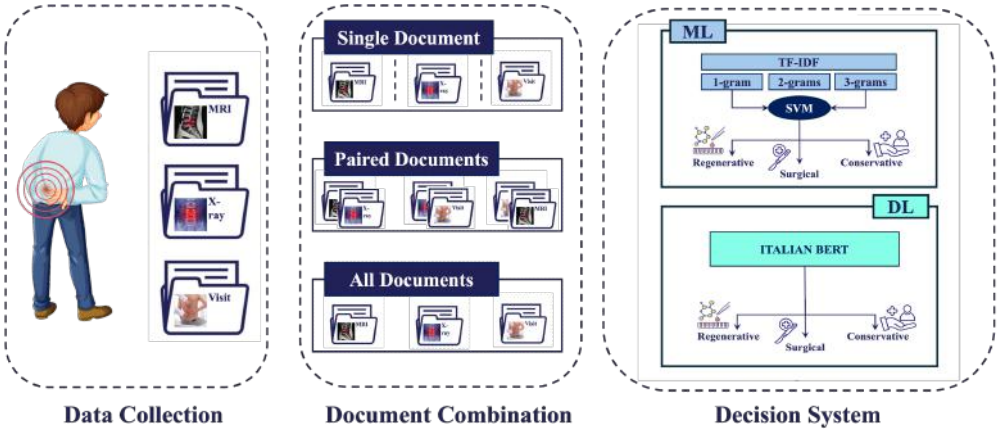


Figure 2: Overview of the LBP decision support system pipeline, from initial data collection through document combination strategies to the final treatment recommendation based on NLP and machine learning techniques.

Table 2
Performance of SVM models with different n-gram settings. F_1-Score is reported as mean $\pm$ standard deviation. The results are compared with the baseline.

Document(s)	# of samples	# Train	# Test	SVM 1-gram (%)	SVM 2-grams (%)	SVM 3-grams (%)	Baseline (%)
MRI	125	100	25	27.54 ± 0.45	29.71 ± 4.54	29.71 ± 4.54	30.22 ± 0.82
X-ray	125	100	25	60.24 ± 7.36	54.20 ± 12.58	53.97 ± 10.65	28.52 ± 1.17
Visit	135	108	27	62.31 ± 8.94	67.99 ± 5.18	69.75 ± 6.18	25.84 ± 2.49
MRI+Visit	168	134	34	65.04 ± 6.30	68.74 ± 7.40	70.27 ± 8.24	22.94 ± 0.45
X-ray+MRI	142	113	29	50.15 ± 3.47	47.81 ± 6.04	45.34 ± 4.14	26.30 ± 1.02
X-ray+Visit	170	136	34	68.42 ± 8.04	68.18 ± 4.41	69.55 ± 5.95	22.56 ± 1.08
X-ray+MRI+Visit	176	140	36	68.31 ± 5.57	$\mathbf{71.34 \pm 6.05}$	68.83 ± 8.14	22.88 ± 0.32

each input configuration we therefore retain all patients who have at least one of the documents in that specific combination (e.g., any patient with an X-ray *or* an MRI is included in the X-ray+MRI setting). This choice maximizes cohort size while mirroring typical clinical availability, where documentation completeness varies across healthcare facilities.

This structured evaluation aimed to identify the most informative combination of clinical documents for LBP treatment prediction. We focused particularly on configurations that balance predictive performance with clinical availability, acknowledging that healthcare facilities may have varying access to different types of diagnostic documentation. The analysis of document combinations proved especially relevant in LBP cases, where the diagnostic value of imaging studies may vary based on specific pathology presentations and resource availability.

4.3. Evaluation Protocol

We performed 5-fold cross-validation for each configuration, maintaining consistent patient splits across all models to ensure a fair and comparable evaluation. Class distributions were preserved within each folad to retain the original class balance across splits. Model performance was evaluated using the macro-averaged F_1-score, which is particularly appropriate for imbalanced classes. All models were compared against a baseline classifier that always predicts the majority class within each fold. Results are reported as the mean $\pm$ standard deviation across the five folds.

4.4. Training Configuration Details

To ensure reproducibility and provide clarity on our modeling setup, we report below all the key hyperparameters and implementation choices for both the BERT-based and the SVM-based experiments. All hyperparameters reported were left at their default values in the respective libraries, with no manual tuning.

Table 3
F_1-Scores, reported as mean ± standard deviation, of BERT models and majority-class baseline on different document combinations.

Document(s)	# of samples	# Train	# Test	BERT (%)	Baseline (%)
MRI	125	100	25	31.93 ± 10.55 %	30.22 ± 0.82 %
X-ray	125	100	25	36.66 ± 10.32 %	28.52 ± 1.17 %
Visit	135	108	27	52.84 ± 13.95 %	25.84 ± 2.49 %
MRI+X-ray	142	113	29	52.21 ± 7.54 %	26.30 ± 1.02 %
X-ray+MRI	142	113	29	46.39 ± 5.84 %	26.30 ± 1.02 %
MRI+Visit	168	134	34	51.89 ± 15.59 %	22.94 ± 0.45 %
Visit+MRI	168	134	34	48.60 ± 2.98 %	22.94 ± 0.45 %
X-ray+Visit	170	136	34	53.51 ± 7.95 %	22.56 ± 1.08 %
Visit+X-ray	170	136	34	**55.24 ± 9.37 %**	22.56 ± 1.08 %
MRI+X-ray+Visit	176	140	36	49.65 ± 8.56 %	22.88 ± 0.32 %
MRI+Visit+X-ray	176	140	36	51.54 ± 10.94 %	22.88 ± 0.32 %
X-ray+MRI+Visit	176	140	36	44.12 ± 8.40 %	22.88 ± 0.32 %
X-ray+Visit+MRI	176	140	36	47.76 ± 7.67 %	22.88 ± 0.32 %
Visit+MRI+X-ray	176	140	36	47.67 ± 12.25 %	22.88 ± 0.32 %
Visit+X-ray+MRI	176	140	36	49.78 ± 13.25 %	22.88 ± 0.32 %

- **BERT** We use BERT's fast tokenizer to preprocess the input, applying truncation and padding to a fixed length.
  - Max sequence length: 512 tokens
  - Batch size: 16
  - Number of epochs: 6
  - Learning rate: 5×10^{-5}
  - Optimizer: AdamW
- **SVM**
  - Vectorization: TF–IDF with n-gram range $[1, N]$, $N \in \{1, 2, 3\}$
  - Classifier: `LinearSVC` with $C = 1.0$, class weights = inverse sample frequency

5. Results

Tables 2 and 3 present the results of our preliminary experiments using SVMs with *n*-gram features and a BERT-based model on various combinations of clinical documents. These results should be interpreted not as evidence of a finalized decision support system, but as an initial validation of the dataset's utility in supporting automatic classification tasks in the context of LBP treatment. To provide a meaningful reference point for model performance, we include the results of a simple majority class predictor, which assigns all test instances to the most frequent class observed in the training set for each fold. This baseline yields macro-averaged F_1-scores in the range of 22–30%, establishing a minimal threshold that highlights the added value of learning-based approaches. Our analysis emphasizes comparative insights across different input configurations and modeling strategies.

5.1. Classification Approach

SVM with TF-IDF N-grams Table 2 compares the macro-F_1 performance obtained with unigram, bigram, and trigram TF-IDF vectors. The bigram configuration attains the highest score, **71.34 ± 6.05%**, improving upon unigrams (68.31 ± 5.57%) and trigrams (68.83 ± 8.14%) while exceeding the majority-class baseline of 22% by almost 50 percentage points. The advantage of bigrams is most pronounced when the full set of reports (Visit, X-ray, and MRI) is concatenated, indicating that short multi–word expressions such as *"discopatia lombare"* encapsulate diagnostic nuance that unigrams cannot capture. In contrast, for single-source inputs the benefit is attenuated: unigrams remain preferable for isolated X-ray reports (60.24% vs 54.20%), suggesting that imaging lexicons are adequately represented by individual tokens.

BERT Table 3 shows the fine-tuned `bert-base-italian-xxl-uncased` model results. The model reaches a maximum macro-F_1 of 55.24 ± 9.37% when the clinical visit note precedes the X-ray report (Visit→X-ray), again outperforming the baseline but trailing the best bigram SVM combination by roughly 16 percentage points. Performance varies with document order: reversing the sequence (X-ray→Visit) lowers the score to 53.51 ± 7.95%, and the inclusion of MRI text frequently degrades results. These fluctuations confirm the order sensitivity anticipated in Section 4.1

and underscore that, under the limited data regime of this study, contextual embeddings do not yet capitalise on MRI radiological terminology as efficiently as lexical features.

5.2. Document Combination Analysis

SVM Consistent with the experimental design of Section 4.2, eight input configurations were evaluated using the *n*-gram representation. Among single documents, the clinical visit note achieves the highest macro-F_1 ($69.75 \pm 6.18\%$ for the trigram representation), whereas the MRI report is the *only* configuration that underperforms the majority-class baseline, reaching just $29.71 \pm 4.54\%$. Pairing X-ray with the visit note yields a substantial gain to $68.18 \pm 4.41\%$, and adding MRI further increases performance to the overall peak of $\mathbf{71.34 \pm 6.05\%}$ for the bigram representation. By contrast, the combination X-ray+MRI, which excludes the narrative Visit note, attains only $47.81 \pm 6.04\%$ macro-F_1. This sharp drop, together with the sub-baseline score of the MRI alone, underscores how indispensable free-text clinical observations are for differentiating low-back pain treatments. Beyond classification performance, we also sought to enhance the interpretability of the best-performing model (SVM with TF–IDF bigrams on all reports) through qualitative analysis of its learned features. Each weight reflects the discriminative power of a lexical bigram for a given treatment class. In Appendix B, we present the most informative medical expressions associated with each class, emphasizing how specific terms are strongly linked to particular treatment decisions.

BERT The document-level ranking mirrors that of the SVM but at lower absolute values. The sequence Visit→X-ray tops the list ($55.24 \pm 9.37\%$), followed by X-ray→Visit ($53.51 \pm 7.95\%$) and MRI→X-ray ($52.21 \pm 7.54\%$). Configurations that concatenate all three reports might exceed the 512-token limit and achieve no more than 51%. Despite these constraints, every BERT variant surpasses the baseline, confirming that contextual representations contain useful decision cues even when suboptimal ordering or length truncation is necessary.

6. Discussion

Our comparative evaluation of traditional machine learning and transformer-based approaches for classifying LBP treatments yields several key insights into how NLP models behave across different types of clinical documentation.

In particular, SVM models leveraging TF–IDF representations consistently outperformed BERT-based models across multiple experimental settings, especially when applied to radiological reports (MRI and X-Ray). These reports are typically concise, standardized, and lexically redundant, making them well-suited to models that exploit explicit lexical features. SVMs, in particular, benefit from frequent term patterns and domain-specific collocations captured through n-gram vectorization.

In contrast, BERT showed stronger performance on less structured, semantically dense documents such as clinical visit notes. These notes are written in natural language, often include temporal and referential elements, and require a deeper semantic understanding to accurately interpret. Despite being the least represented document type across all treatment classes, visit notes boosted performance when used alone or in combination with other sources. This indicates their high semantic informativeness and BERT's ability to leverage contextual cues and long-range dependencies.

For a sample of each report type, see Appendix A.

Interestingly, although BERT underperformed compared to SVM in nearly all configurations, its strengths became more evident when visit notes were incorporated into multidocument setups. The best-performing configuration among all SVM experiments was the integration of all three document types. This reinforces the idea that each source contributes distinct and valuable information: X-rays provide succinct structural summaries, MRIs add detailed anatomical insights (especially relevant for surgical decision-making), and visit notes contribute clinical reasoning and narrative depth. The integration of these heterogeneous data sources allows the model to capture a more comprehensive clinical picture, ultimately improving classification accuracy.

BERT was consistently outperformed by SVM across nearly all configurations. A likely explanation lies in the underrepresentation of visit notes within the dataset. Although visit notes are semantically rich, their greatest impact on classification performance becomes evident when they are combined with radiological sources. One of the most notable findings from this dataset is that the integration of all three document types yielded the best-performing configuration in all SVM experiments. This outcome underscores the complementary nature of the information encoded in these documents: X-rays provide concise structural descriptions, MRIs offer detailed anatomical insights (especially valuable for surgical planning), and visit notes contribute clinical reasoning and contextual narrative. The fusion of these heterogeneous inputs enables the model to capture multiple dimensions of the clinical scenario, ultimately leading to improved classification accuracy.

It should be noted that, given the real-world nature of this dataset, not all document combinations are directly comparable due to the differing numbers of available documents across treatment categories. While this vari-

ability accurately reflects actual clinical practice, caution is warranted in interpreting comparative model performances, particularly when smaller document subsets may limit the generalizability of results.

6.1. Clinical Implications

Although MRI is routinely regarded as the most informative examination for surgical planning in low-back pain, its impact in our study was limited by availability: surgical cases accounted for only 34 of 176 patients and contained proportionally fewer MRI reports than the other treatment groups. This scarcity translated into weak stand-alone performance - an SVM trained on MRI text alone fell below the majority-class baseline (macro-F_1 29.7 ± 4.5 %) and, even when coupled with X-ray, remained inferior to the X-ray + visit-note configuration. Clinically, these results indicate that the proposed decision-support tool already offers actionable triage guidance in contexts where MRI access is delayed, while underscoring the need to enrich the dataset with additional surgical MRIs, through prospective collection, to reduce the risk of under-referral for patients who would ultimately benefit from operative management.

7. Conclusions

The results of this study underscore the clinical relevance and future potential of our curated dataset as a foundation for developing NLP-based decision support tools in the context of low back pain. By aligning structured radiology reports with semantically rich clinical narratives and treatment labels drawn from real-world care trajectories, the dataset captures a heterogeneous and realistic cross-section of diagnostic information, reflective of everyday clinical reasoning.

Despite its limited size, the dataset reveals meaningful interactions between document types and model performance. Notably, while magnetic resonance imaging is routinely regarded as the most informative modality for surgical planning, its impact in our study was constrained by availability: only 34 out of 176 patients were classified under the surgical group, and this subset contained proportionally fewer MRI reports than the others. This imbalance translated into weak stand-alone performance.

These results suggest that the proposed dataset already supports the development of decision-support tools capable of offering actionable triage guidance, even in contexts where MRI access is limited or delayed. At the same time, the findings highlight a clear direction for future dataset enrichment: increasing the number of surgical MRIs, either through prospective data collection or active-learning-guided sampling, will be essential to reduce the risk of under-referral for patients who may ultimately require surgical intervention.

In future works, we will explore other models capable of handling longer input sequences, such as recent large language models, allowing us to include the full content of all three documents (MRI, X-ray, and visit notes) without truncation.

We further plan to expand the dataset through the collection of additional clinical cases. Once validated, the extended corpus will be released to foster reproducibility and enable further research. We will also perform systematic hyperparameter optimization on the extended dataset to further improve model performance.

Acknowledgments

Authors were supported by two projects: 1) the European Union - Next Generation EU - NRRP M6C2 - Investment 2.1 Enhancement and strengthening of biomedical research in the NHS, project n. PNRR-MAD-2022-12376692_VADALA' - CUP F83C22002470001. 2) the European Union under the Horizon Europe Programme through the Innovative Health Initiative Joint Undertaking (IHI JU) – Project GRACE (Project number: 101194778, Project name: bridGing gaps in caRdiAC health managEment). 3) the European Union - Next Generation EU - NRRP M6C2 - Investment 2.1 Enhancement and strengthening of biomedical research in the NHS - Project PNRR-MR1- 2022-12376635 - "Early Detection of Rare Inherited Retinal Dystrophies and Cardiac Amyloidosis enhanced by Artificial Intelligence: the impact on the patient's pathway in Campania Region" (CUP: C83C22001540007)

References

[1] A. Wu, L. March, X. Zheng, J. Huang, X. Wang, J. Zhao, F. M. Blyth, E. Smith, R. Buchbinder, D. Hoy, Global low back pain prevalence and years lived with disability from 1990 to 2017: estimates from the global burden of disease study 2017, Annals of translational medicine 8 (2020) 299.

[2] T. Zhou, D. Salman, A. H. McGregor, Recent clinical practice guidelines for the management of low back pain: a global comparison, BMC musculoskeletal disorders 25 (2024) 344.

[3] O. Airaksinen, J. I. Brox, C. Cedraschi, J. Hildebrandt, J. Klaber-Moffett, F. Kovacs, A. F. Mannion, S. Reis, J. Staal, H. Ursin, et al., European guidelines for the management of chronic nonspecific low back pain, European spine journal 15 (2006) s192.

[4] H.-J. Kong, Managing unstructured big data in

healthcare system, Healthcare informatics research 25 (2019) 1–2.
[5] J. Liang, Y. Li, Z. Zhang, D. Shen, J. Xu, X. Zheng, T. Wang, B. Tang, J. Lei, J. Zhang, Adoption of electronic health records (ehrs) in china during the past 10 years: consecutive survey data analysis and comparison of sino-american challenges and experiences, Journal of medical Internet research 23 (2021) e24813.
[6] L. Bacco, F. Russo, L. Ambrosio, F. D'Antoni, L. Vollero, G. Vadalà, F. Dell'Orletta, M. Merone, R. Papalia, V. Denaro, Natural language processing in low back pain and spine diseases: a systematic review, Frontiers in Surgery 9 (2022) 957085.
[7] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, S. Horng, Mimic-cxr: A large publicly available database of labeled chest radiographs, 2019. URL: https://physionet.org/content/mimic-cxr/2.0.0/. doi:10.13026/cr8q-rw49, rRID:SCR_007345.
[8] A. Johnson, T. Pollard, S. Horng, L. A. Celi, R. Mark, Mimic-iv-note: Deidentified free-text clinical notes (version 2.2), PhysioNet (2023). URL: https://doi.org/10.13026/1n74-ne17. doi:10.13026/1n74-ne17, rRID:SCR_007345.
[9] F. A. Matsuoka, H. N. Onaga, Classifying domains, benchmarking gpt-4, a portuguese dataset for medical ai q&a, bioRxiv (2024) 2024–12.
[10] V. Basile, C. Bosco, M. Fell, V. Patti, R. Varvara, et al., Italian nlp for everyone: Resources and models from evalita to the european language grid, in: 2022 Language Resources and Evaluation Conference, LREC 2022, European Language Resources Association (ELRA), 2022, pp. 174–180.
[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.
[13] N. Sheehan, Magnetic resonance imaging for low back pain: indications and limitations, Postgraduate medical journal 86 (2010) 374–378.
[14] R. U. Din, X. Cheng, H. Yang, Diagnostic role of magnetic resonance imaging in low back pain caused by vertebral endplate degeneration, Journal of Magnetic Resonance Imaging 55 (2022) 755–771.
[15] L. Bacco, A. Cimino, L. Paulon, M. Merone, F. Dell'Orletta, A machine learning approach for sentiment analysis for italian reviews in healthcare, in: CEUR Workshop Proceedings, volume 2769, CEUR-WS, 2020.
[16] R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, M. Esposito, A novel covid-19 data set and an effective deep learning approach for the de-identification of italian medical records, Ieee Access 9 (2021) 19097–19110.
[17] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune bert for text classification?, in: China national conference on Chinese computational linguistics, Springer, 2019, pp. 194–206.

A. Sample Reports

We present three representative reports that illustrate distinct documentation styles: the MRI and X-ray findings are conveyed with technical details, whereas the clinical evaluation is presented as a concise narrative.

The imaging reports describe features such as lumbar disc degeneration, spondylolisthesis, and preserved vertebral alignment. In contrast, the consult note summarizes patient history, describes symptoms, and reports physical examination findings, before referencing the imaging results. Thus, the narrative note provides clinical context, while the radiological reports contribute detailed anatomical and pathological descriptions.

B. SVM's Lexical Feature Analysis

To improve the interpretability of our best-performing SVM classifier, trained with TF–IDF bigrams on the full set of clinical documents, we analyzed the feature weights learned by the model from the top-performing fold of the 5-fold cross-validation. These weights indicate the contribution of each lexical bigram to treatment classification, highlighting expressions with clear clinical significance.

We manually prioritized domain-specific expressions (e.g., anatomical or pathological descriptors)from the top 50 lexical features (unigrams and brigrams) ranked by coefficient value for each treatment class, over generic tokens (e.g., grado, presenza), which, despite their assigned weights, lack standalone diagnostic value. The most informative medically relevant features identified by the model for each treatment class—Conservative, Regenerative, and Surgical—are reported in Table 5, along with their associated weights and frequencies in the training and test sets. Importantly, the selected inspected features exhibit meaningful clinical relevance, effectively capturing diagnostic and pathological indicators that inform therapeutic decision-making.

Specifically, conservative treatment is associated with clinically less invasive descriptors such as *sostanzial-*

Italian	English
MRI: *Sostanzialmente conservata la fisiologica lordosi lombare; lieve deviazione sinistro-convessa del rachide lombare a fulcro L3-L4. Discopatia degenerativa a livello L4-L5 ed L5-S1; in particolare:* *• a livello L4-L5 si osserva protrusione discale ad ampio raggio che occupa bilateralmente il pavimento dei forami neurali e, a destra entra in contatto con il tratto preforaminale della radice L5 destra; si associa a tale livello alterazione dell'intensità di segnale dei contrapposti versanti intersomatici tipo Modic 2–3.* *• a livello L5-S1 è presente protrusione discale ad ampio raggio che non entra in conflitto con le radici nervose adiacenti.* *Conservata la morfologia delle restanti unità disco-somatiche. Non ci sono alterazioni focali ossee nei segmenti scheletrici esaminati. Canale vertebrale di dimensioni nella norma. Nella norma l'intensità di segnale del cono midollare, posizionato a livello D12. Conservato il trofismo dei muscoli para-vertebrali al passaggio lombo-sacrale. Cisti aracnoidee sacrali a livello S1-S2, del diametro massimo di 3 cm.*	**MRI:** *Essentially preserved physiological lumbar lordosis; slight left-convex deviation of the lumbar spine with apex at L3-L4. Degenerative disc disease at L4-L5 and L5-S1; specifically:* *• at L4-L5, a broad-based disc protrusion is observed, bilaterally occupying the floor of the neural foramina and, on the right, contacting the preforaminal tract of the right L5 root; associated with a mild signal intensity alteration of the opposing endplates (Modic type 2–3).* *• at L5-S1, a broad-based disc protrusion is present, which does not impinge on adjacent nerve roots.* *Morphology of the remaining disc–vertebral units is preserved. No focal bone abnormalities in the examined skeletal segments. Vertebral canal dimensions are within normal limits. Signal intensity of the conus medullaris is normal, positioned at D12. Paravertebral muscle trophism at the lumbosacral junction is preserved. Sacral arachnoid cysts at S1-S2 level, with a maximum diameter of 3 cm.*
X-Ray: *Sostanzialmente conservata la fisiologica lordosi lombare. Non evidenti alterazioni ossee radiograficamente apprezzabili nei segmenti ossei in esame. Normoallineati i muri somatici posteriori sia in proiezione LL standard che in massima estensione; disallineamento dei muri somatici posteriori con spondilolistesi anteriore L4-L5 di grado 1 in massima flessione, come segno di instabilità articolare a tale livello. Lieve riduzione in altezza dello spazio intersomatico L4-L5, come segno di discopatia degenerativa. Tono calcico conservato.*	**X-Ray:** *Essentially preserved physiological lumbar lordosis. No radiographically appreciable bone abnormalities in the examined osseous segments. Posterior vertebral walls are normally aligned in both standard LL projection and maximum extension; misalignment of the posterior vertebral walls with Grade I anterior spondylolisthesis at L4-L5 in maximum flexion, indicating articular instability at that level. Mild reduction in intervertebral space height at L4-L5, indicating degenerative disc disease. Preserved bone density.*
Visit: *APR: n.d.r. APP: Il paziente riferisce lombalgia da diversi anni, esacerbata durante attività sportiva. NRS colonna lombosacrale 6/10. Ha praticato FKT con temporaneo beneficio. Il dolore è maggiormente lateralizzato a sinistra a livello del rachide lombosacrale. Non episodi di sciatalgia. La sintomatologia inficia il riposo notturno, ma non si altera con la manovra di Valsalva. Presenta limitazione della flesso-estensione del rachide lombosacrale. Porta in visione RMN colonna LS (11/09/2020) che mostra discopatia L4-L5 ed L5-S1 in presenza di alterazione degenerativo-infiammatoria dei piatti vertebrali contrapposti e dell'osso subcondrale a livello L4-L5 in fase acuta del tipo Modic 1. EO: Dolore in iperestensione del rachide lombosacrale ed inclinazione laterale. Ipercifosi dorsale. Marcata contrattura paravertebrale. Dolore all'articolazione sacro-iliaca SX. Deambulazione possibile in taligrado e digitigrado. Lasègue bilaterale. Non deficit di TA, EPA ed ECD. Diagnosi: Discopatia L4-L5 ed L5-S1 in presenza di alterazione degenerativo-infiammatoria dei piatti vertebrali contrapposti e dell'osso subcondrale a livello L4-L5 in fase acuta del tipo Modic 1.*	**Visit:** *APR: no relevant medical history recorded. APP: The patient reports low back pain for several years, exacerbated during sports activity. NRS lumbosacral score 6/10. He underwent physiokinetic therapy with temporary relief. Pain is predominantly lateralized to the left at the lumbosacral spine. No episodes of sciatica. Symptoms disrupt sleep but do not change with the Valsalva maneuver. Presents with limitation of flexion-extension of the lumbosacral spine. Brings MRI of LS spine (11/09/2020) showing discopathy at L4-L5 and L5-S1 with degenerative-inflammatory changes of the opposing vertebral endplates and subchondral bone at L4-L5 in acute Modic 1 phase. EO: Pain on hyperextension of the lumbosacral spine and lateral bending. Thoracic hyperkyphosis. Marked paravertebral muscle contracture. Pain at the left sacroiliac joint. Ambulation possible on heels and toes. Bilateral Lasègue's sign. No deficits in TA, EPA, and ECD. Diagnosis: Discopathy at L4-L5 and L5-S1 with degenerative-inflammatory changes of the opposing vertebral endplates and subchondral bone at L4-L5 in acute Modic 1 phase.*
Età: 45	**Age:** 45
Sesso: M	**Sex:** M

Table 4
Sample clinical report comparison for a patient receiving conservative treatment.

Lexical Bigram	SVM Weight by Treatment Class			Frequency	
	Conservative	Regenerative	Surgical	Train	Test
With Polarity Inversion					
ernia	**+0.332**	–	**-0.302**	47	23
discopatia	**+0.587**	–	**-0.418**	101	24
muri somatici	**-0.376**	**+0.424**	–	43	6
proiezioni dinamiche	**-0.374**	**+0.363**	–	39	6
spondilolistesi	–	**-0.517**	**+0.698**	38	10
stenosi	–	**-0.533**	**+0.688**	37	7
Other High-Weight Bigrams					
sostanzialmente conservati	**+0.445**	–	–	9	4
protrusione discale	–	**+0.331**	–	72	14
antero listesi	–	–	**+0.389**	11	1

Table 5
Medical lexical features with the highest SVM weights per treatment class. The symbol '-' indicates the absence of the feature for the given class.

mente conservati and *degenerazioni artrosiche*. Regenerative treatments, meanwhile, are characterized by medically pertinent terms like *muri somatici* and *proiezioni dinamiche*. Finally, surgical treatment features expressions indicative of more severe pathology, including *spondilolistesi* and *stenosi*, both frequently occurring in the training data and receiving high positive weights (0.698 and 0.688, respectively).

Notably, our analysis highlighted polarity inversion phenomena, whereby certain clinically relevant terms (e.g., *spondilolistesi*, *ernia*) showed positive weights in one treatment class and negative weights in another. This underlines the context-sensitive nature of their clinical interpretation.

Furthermore, it is worth emphasizing that feature frequency alone does not fully explain clinical importance: even relatively infrequent terms can receive high model weights if they demonstrate strong discriminative power. For example, antero listesi appeared only 11 times in the training set yet emerged as one of the top-ranked surgical features, confirming the model's capability to identify clinically informative lexical indicators.