Arends *et al.*
BMC Medical Informatics and Decision Making (2025) 25:115
https://doi.org/10.1186/s12911-025-02897-w

BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Diagnosis extraction from unstructured Dutch echocardiogram reports using span- and document-level characteristic classification

Bauke Arends[1*], Melle Vessies[1], Dirk van Osch[1], Arco Teske[1], Pim van der Harst[1], René van Es[1] and Bram van Es[2]

Abstract

Background Clinical machine learning research and artificial intelligence driven clinical decision support models rely on clinically accurate labels. Manually extracting these labels with the help of clinical specialists is often time-consuming and expensive. This study tests the feasibility of automatic span- and document-level diagnosis extraction from unstructured Dutch echocardiogram reports.

Methods We included 115,692 unstructured echocardiogram reports from the University Medical Center Utrecht, a large university hospital in the Netherlands. A randomly selected subset was manually annotated for the occurrence and severity of eleven commonly described cardiac characteristics. We developed and tested several automatic labelling techniques at both span and document levels, using weighted and macro F1-score, precision, and recall for performance evaluation. We compared the performance of span labelling against document labelling methods, which included both direct document classifiers and indirect document classifiers that rely on span classification results.

Results The SpanCategorizer and MedRoBERTa.nl models outperformed all other span and document classifiers, respectively. The weighted F1-score varied between characteristics, ranging from 0.60 to 0.93 in SpanCategorizer and 0.96 to 0.98 in MedRoBERTa.nl. Direct document classification was superior to indirect document classification using span classifiers. SetFit achieved competitive document classification performance using only 10% of the training data. Utilizing a reduced label set yielded near-perfect document classification results.

Conclusion We recommend using our published SpanCategorizer and MedRoBERTa.nl models for span- and document-level diagnosis extraction from Dutch echocardiography reports. For settings with limited training data, SetFit may be a promising alternative for document classification. Future research should be aimed at training a RoBERTa based span classifier and applying English based models on translated echocardiogram reports.

Keywords Clinical natural language processing, Echocardiogram, Entity classification, Span classification, Document classification

Background

Unstructured electronic health record (EHR) data contains valuable information for a broad spectrum of clinical machine learning applications, including the creation of clinical decision support systems, semi-automated report writing, and cohort identification. The extraction of accurate clinical labels is essential to realize these applications. Relying solely on structured data for this purpose often yields disappointing outcomes, primarily

*Correspondence:
Bauke Arends
b.k.o.arends-4@umcutrecht.nl
[1] Department of Cardiology, University Medical Center Utrecht, Utrecht, The Netherlands
[2] Central Diagnostic Laboratory, University Medical Center Utrecht, Utrecht, The Netherlands

due to two key reasons. Firstly, collecting structured data has only recently gained momentum in clinical practice, leaving a large volume of historical data underutilized. Secondly, the structured data that is collected, may suffer from a lack of precision and reliability [1]. International Classification of Disease (ICD) coding specifically was identified as unreliable for phenotyping EHRs [2–4]. Data annotation is identified as one of the main obstacles in developing clinical natural language processing (NLP) applications [5]. Therefore, utilizing labels extracted from unstructured data has the potential to enhance both data volume and data quality.

Echocardiography, the most commonly performed cardiac imaging diagnostic [6], provides a detailed anatomical and functional description of a wide range of cardiac structures. Data from echocardiography reports are consequently used in many aspects of patient care, as well as many clinical trials. However, the heterogeneous format of the reports, as well as medical text characteristics such as abundant shorthand, domain-specific vocabularies, implicitly assumed knowledge, and spelling and grammar mistakes, make extracting accurate labels challenging. For label extraction, we often resort to automated techniques, because manual extraction by domain experts is both costly and time-consuming.

Previous work on data extraction from echocardiography reports has primarily focused on extracting quantitative measurement values from structured, semi-structured and unstructured parts of the report using rule-based methods [7–10]. Rule-based text-mining systems such as MedTagger [11], Komenti [12], and cTAKES [13] are examples of low-code tools that allow clinicians to develop and apply rules for rule-based text mining. These rule-based methods offer several advantages, as they are transparent, easily modifiable, and do not require large amounts of labelled training data. Furthermore, they can be quite effective despite their simplicity. While their performance can vary based on the developer's expertise and attention to detail, a more specific downside of rule-based methods is their inherent inability to generalize beyond the set of predefined rules.

NLP methods based on machine learning may overcome some of these disadvantages, as they are able to learn rules implicitly from labelled data. In the biomedical field, several open-source systems, such as GATE [14] and cTAKES [13] are available to employ these methods. Additionally, an abundance of model architectures is available for label extraction, including token classification models [15], conditional random fields (CRF) [16], recurrent neural network (RNN) such as long short-term memory (LSTM) [17] and transformers such as BERT [18], support vector machine (SVM) [19] and AutoML methods [20]. However, in the broader field of named entity recognition (NER) in medical imaging reports, there does not seem to be one overall best-performing method [17, 21, 22]. For span identification performance in particular, multiple factors may influence performance, including model architecture and span characteristics such as span frequency, distinctive span boundaries and span length [23].

NER in the medical imaging report domain has mostly been described in English texts [7, 10, 24]. There are limited studies in other languages, such as Dutch [25, 26], German [27], and Spanish [28]. Few publicly available pre-trained Dutch language models exist, and include BERTje [29] and RobBERT [30, 31]. Verkijk and Vossen recently created MedRoBERTa.nl, a version of RoBERTa [32] finetuned on Dutch EHR data [33]. Furthermore, Remy, Demuynck and Demeester developed a multilingual large language model BioLORD-2023M using contrastive learning, which is able to identify biomedical concepts and sentences [34]. To the best of our knowledge, none of these models have been finetuned with the goal of information extraction from Dutch echocardiogram reports.

In this work, we focus on span and document label extraction from unstructured Dutch echocardiogram reports for a wide range of clinical concepts. Both span and document classification can be used to extract labels for downstream machine learning tasks, though they address different levels of information and have distinct applications in clinical workflows. Span classification, for instance, can assist in biological entity linking tasks, whereas document classification is better suited for broader tasks like cohort selection. To capture the most meaningful clinical concepts, we constructed a custom ontology which incorporates most major cardiac abnormalities. We explicitly focused on extracting qualitative labels from unstructured text, as several algorithms exist to extract measurement values from structured and semi-structured data. We evaluated three NLP methods for span-level label extraction, and six NLP methods for document-level label extraction. The best-performing span and document classification models are available on the Huggingface model repository.[1] Additionally, the developed code is publicly available on GitHub.[2]

Methods

This section provides a detailed description of our data and the data annotation process, followed by an overview of our experiments. We employ several neural and statistical methods for extracting span and document-level

[1] https://huggingface.co/UMCUtrecht

[2] https://github.com/umcu/EchoLabeler

labels from Dutch echocardiogram reports. Additional information on model parameters is detailed in Additional file 1.

Data overview

Our dataset consisted of 115,692 unstructured echocardiogram reports collected during routine clinical care from 2003 to 2023, stored in the EHR at University Medical Center Utrecht (UMCU), a large university hospital in the Netherlands. Over this period, there has not been a universal standard for report writing. Reports containing fewer than fifteen characters were excluded, as were reports with fewer than thirty characters that lacked any description of a medical concept. These reports often contained only the phrase *"For the report, see the patient's chart"*.

Data annotation

In a randomly selected subset of the unstructured text portions of these reports, we manually annotated eleven cardiac characteristics, which included left and right ventricular systolic function and dilatation, valvular disease, pericardial effusion, wall motion abnormalities, and diastolic dysfunction. These characteristics were chosen after consultation with two cardiologists, and are the most frequently described cardiac characteristics in Dutch echocardiogram reports. Figure 1 displays an example report including annotations. We assigned mutually exclusive labels for each characteristic to the span/document (Table 1). The hierarchical labeling scheme was selected to align with the terminology used in clinical practice, which varies across different conditions. This resulted in a differing number of labels for each characteristic.

Annotations were checked sample wise by doctors. In cases of uncertainty, cases were jointly reviewed to achieve consensus. Several rounds of training iterations were completed before commencing the annotation task. To streamline the annotation process, each echocardiogram report was annotated for one characteristic at a time, resulting in eleven separate annotation files. For an overview of labelling instructions, see Additional file 2. Prodigy [35] was employed for the annotation task.

Table 1 Span and document label definitions

Label	Description
No label	No statement regarding this characteristic
Normal	Normal function described for this characteristic
Mild	Mildly abnormal function
Moderate	Moderately abnormal function
Severe	Severely abnormal function
Present	Abnormal function, unspecified severity

To ensure an adequate number of labels, we established the following requirements: for each characteristic, a minimum of 5000 documents were annotated, with the same documents used for each characteristic. In addition, to ensure sufficient training data, a minimum of 50 span labels per class, per characteristic were required, resulting in more than 5000 annotated documents for several characteristics (Table 2). Document-level labels were constructed using the span-level labels to provide a summarized representation of the data. Given that multiple span labels could exist within a single document, we aggregated these span labels by selecting the most severe label for each characteristic. This ensures that the document-level labels reflect the most critical information. For comparison we also employed a simplified label scheme with only three possible labels: not mentioned, normal, or present.

Span characteristics are summarized in Table 4. Span length is the average number of tokens per span. Span distinctiveness measures the uniqueness of tokens within

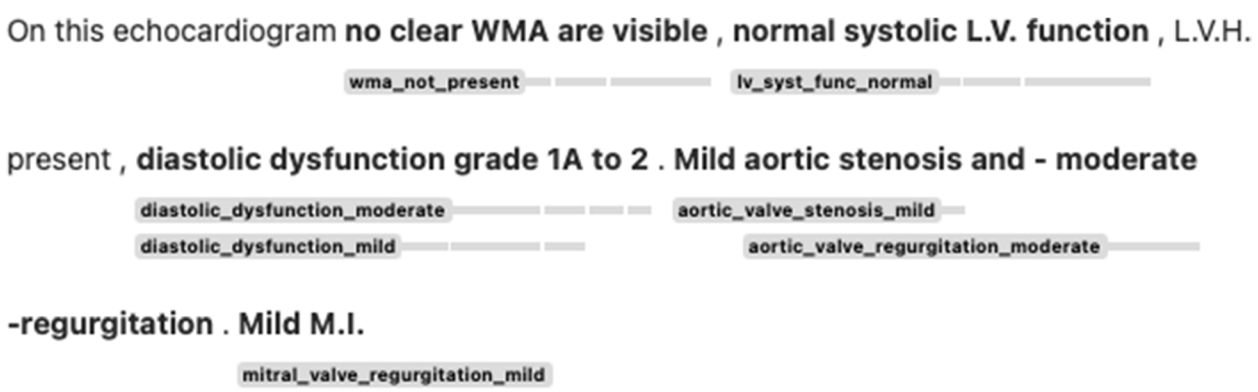

Fig. 1 Example report with manual annotations. For presentation purposes, text has been translated to English

Table 2 Document label counts

Characteristic	Cases	Any label	Normal	Mild	Moderate	Severe	Present
Aortic regurgitation	5615	2403 (42.8%)	1716 (30.6%)	505 (9.0%)	133 (2.4%)	49 (0.9%)	0 (0.0%)
Aortic stenosis	5000	1718 (34.4%)	1461 (29.2%)	108 (2.2%)	68 (1.4%)	81 (1.6%)	0 (0.0%)
Diastolic dysfunction	5000	1526 (30.5%)	521 (10.4%)	632 (12.6%)	243 (4.9%)	130 (2.6%)	0 (0.0%)
Left ventricular dilatation	5000	2402 (48.0%)	1870 (37.4%)	249 (5.0%)	91 (1.8%)	51 (1.0%)	141 (2.8%)
Left ventricular systolic dysfunction	5000	4503 (90.1%)	2881 (57.6%)	879 (17.6%)	378 (7.6%)	365 (7.3%)	0 (0.0%)
Mitral regurgitation	5000	2590 (51.8%)	1605 (32.1%)	733 (14.7%)	187 (3.7%)	65 (1.3%)	0 (0.0%)
Pericardial effusion	8686	1274 (14.7%)	973 (11.2%)	154 (1.8%)	55 (0.6%)	48 (0.6%)	44 (0.5%)
Right ventricular dilatation	8203	2718 (33.1%)	2137 (26.1%)	266 (3.2%)	125 (1.5%)	50 (0.6%)	140 (1.7%)
Right ventricular systolic dysfunction	5000	2462 (49.2%)	1807 (36.1%)	408 (8.2%)	188 (3.8%)	59 (1.2%)	0 (0.0%)
Tricuspid regurgitation	5000	1801 (36.0%)	1333 (26.7%)	262 (5.2%)	140 (2.8%)	66 (1.3%)	0 (0.0%)
Wall motion abnormalities	5000	1224 (24.5%)	389 (7.8%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	835 (16.7%)

spans compared to the entire corpus, using Kullback-Leibler divergence. A higher value indicates that spans contain less frequent tokens, making them stand out from the rest of the text. Similarly, span boundary distinctiveness focuses on the uniqueness of boundary tokens at the start and end of spans. A high boundary distinctiveness suggests well-defined, distinctive span boundaries.

Data splits

We split the dataset in a training and testing set, allocating 80% and 20%, respectively. We used one train/test split for a simple practical reason: we developed regular expressions for identifying candidate spans, direct labelling and span classification only on the train split. To effectively conduct *N*-fold cross-validation with this method, we would need an independent developer for the regular expression patterns in each fold. This would ensure that new regular expressions are developed solely from the training data and then tested on genuinely unseen test data. However, we found that such an approach would be too time-consuming, leading us to choose a single 80%−20% split instead. We split all 115,692 reports preemptively into either the training or test set. During the labelling process, we randomly sampled documents from the entire corpus until we reached the prespecified requirements. Consequently, due to random sampling, the training and testing splits may not add up to exactly the prespecified percentages. In Table 3, we report the distribution of span-level labels for each data split.

Table 3 Number of characteristics in each dataset

Characteristic	Train	Test
Aortic regurgitation	2108	499
Aortic stenosis	1499	351
Diastolic dysfunction	1293	304
Left ventricular dilatation	2003	466
Left ventricular systolic dysfunction	4212	1035
Mitral regurgitation	2362	540
Pericardial effusion	1048	247
Right ventricular dilatation	2260	552
Right ventricular systolic function	2131	509
Tricuspid regurgitation	1574	380
Wall motion abnormalities	1075	259

Span classification

We present three approaches for span classification. First, we employed a rule-based approach using regular expressions as a baseline method. Second, we used a NER+L extractor, where clinical concept spans are identified and subsequently classified. Finally, we implemented a greedy span classification approach, where all possible spans are classified, and only those exceeding a threshold model certainty are presented. An overview of these models including their advantages and disadvantages, is described in Table 5.

Approximate list lookup

Given a dictionary of lists containing phrases, where each list represents a target label, we built a simple rule-based model. This model uses token-based regular expressions to match phrases in unseen texts, extracting relevant spans based on the dictionary entries. We refer to this method as approximate list lookup (ALL). The main advantage of this approach is its transparency and flexibility, phrases can easily be added or removed to improve performance. The rule-based algorithm was constructed as follows:

Algorithm 1 ***Rule-based look-up algorithm (ALL$_{rule}$)***

```
Data: Document d
Result: Dictionary with labels
for span in spans do
    for label in labels do
        if PhraseMatcher(span, label) is True then
            return label;
        end
    end
end
```

MedCAT

MedCAT is a semi-supervised NER+L extractor that supports bilateral LSTM (biLSTM) and transformer-based span-classifiers [36] (Fig. 2). The benefit is that not all token spans are scanned. However, this requires training the MedCAT model to create a context-database that contains context vectors that are indicative for medical concepts. We performed unsupervised training on the training split and added the spans that were defined during the manual labelling process to MedCAT's vocabulary and context-database. The initial span-detector introduces a selection bias compared to a greedy span-classifier. Consequently, we expected a higher precision but lower recall, as some spans may be missed. We trained a different span classifier for each characteristic where all classifiers were integrated into one MedCAT modelpack. To reduce the occurrence of false negatives, we explicitly added a negative label for each class, set to 1 whenever a class was otherwise unlabelled.

spaCy SpanCategorizer

Similar to MedCAT, spaCy's SpanCategorizer [37] operates in two stages: tokenization and span suggestion, followed by span classification. It employs a rule-based Dutch text tokenizer from spaCy. Unlike MedCAT, SpanCategorizer's default span suggester is greedy, suggesting all n-gram spans within a prespecified range of span lengths (Fig. 3). The range for the n-gram suggester was set at 1–25, due to the expected lengthy sentences describing some of the characteristics. Compared to a pipeline with a stricter span suggester, this setup was expected to yield a higher end-to-end recall but a corresponding lower precision due to an increase in false positives.

The span classification pipeline employs a hybrid architecture, combining multiple neural network components to process and classify spans effectively. Tokens from suggested spans were first embedded using a multi hash embedding function based on a token's lexical attributes, followed by encoding using max-out activation, layer normalization, and residual connections. These encoded representations underwent mean and max pooling before being passed through a hidden layer. Finally, single label classification was performed on these span vectors using a logistic loss function. Each span was classified into one of the labelled classes, or was classified with a negative label (i.e., no label).

Standard components and configuration files were predominantly used to prevent overfitting. However, for some characteristics, >70% of cases had only negative labels (Table 4). To address this imbalance and to prevent our model from solely predicting negative labels, different weights were assigned to negative labels (0.6, 0.8 and 1.0). For each characteristic, models were trained with these weights, and the model yielding the highest weighted F1-score was selected. An overview of model

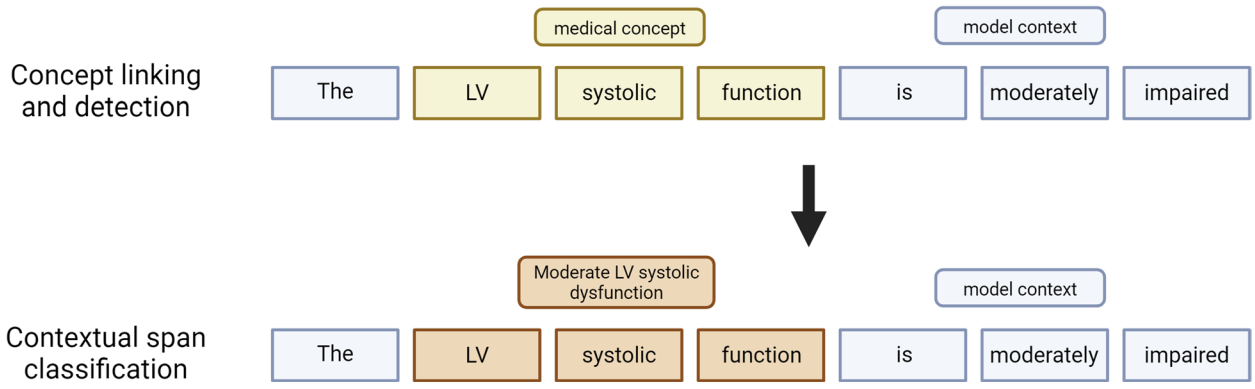

Fig. 2 MedCAT pipeline for identifying and classifying medical concepts

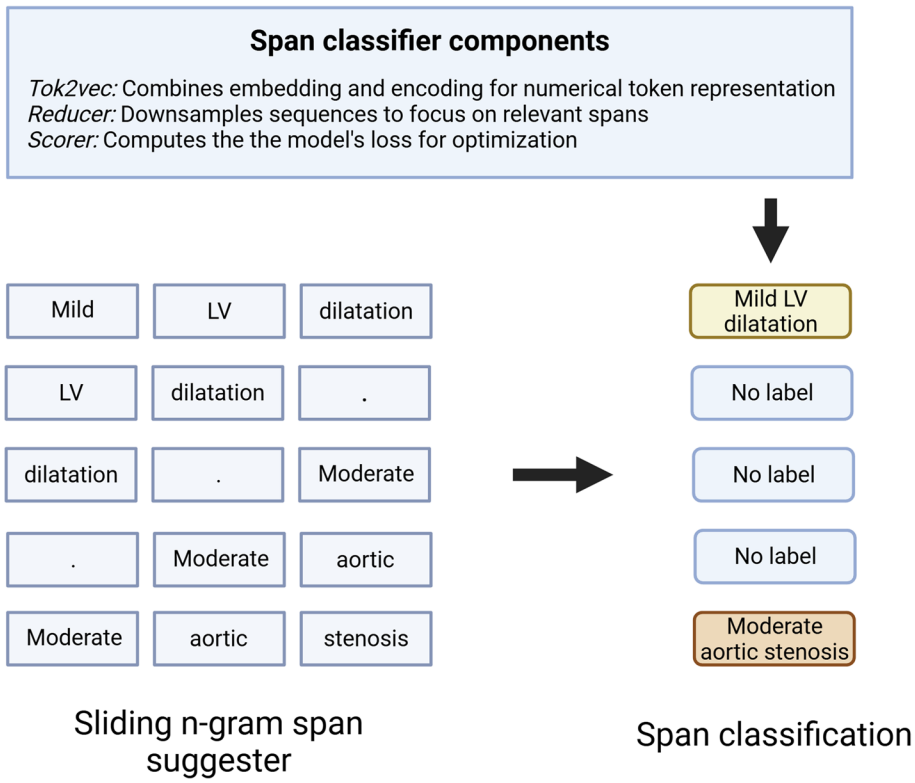

Fig. 3 SpanCat pipeline for iterating and classifying n-gram spans using scanning windows of 1–25 tokens

performance for each negative weight is presented in Additional file 4.

Document classification

For document classification, we used six methods. We implemented two baseline methods: one utilizing a bag-of-words (BOW) approach with medical word embeddings, and the one using indirect document classification via a span-to-document label heuristic, where the best performing span classification method was used to aggregate span-based classifications into document classifications. We also employed SetFit in combination with a pre-trained sentence encoder. Another method involved using RoBERTa, specifically the MedRoBERTa.nl model for this work. Additionally, we applied a RNN model, specifically a bidirectional GRU, and a bidirectional convolutional neural network (CNN). An overview of these models including their advantages and disadvantages, is described in Table 5.

Bag-of-words

Our baseline BOW approach involved several feature extraction steps, detailed in Fig. 4. First, the text was tokenized. Next, we applied term frequency-inverse document frequency (TF-IDF) weighting to each token within a document. We then enriched the features with topic modelling weighting, as described by Bagheri et al. [38]. Additionally, we augmented the features with latent Dirichlet allocation topic probabilities to capture underlying thematic structures. The resulting features were combined with a standard gradient-boosted classifier.

Span classifier heuristic

We selected the best performing span classifier based on its end-to-end performance. Then, we aggregated the span labels into a document label for each characteristic. The process is similar to how we constructed document labels: given a multitude of span labels within one document, we aggregated them by selecting the most severe label per characteristic. This heuristic allowed for more granular analysis by indicating which spans lead to the document classification. However, we expected performance loss due to the increased complexity of span classification.

SetFit

Reimers et al. [39] employed Siamese networks with contrastive learning on similar and dissimilar sentences to produce transformer-based encoders that capture semantic information along different axes of similarity,

Table 4 Span characteristics

Characteristic	Severity	No. of spans	Length	SD	BD
Aortic regurgitation	Overall	2607	2.47	2.62	1.25
	Normal	1849	2.48	2.42	1.28
	Mild	562	2.39	3.08	1.05
	Moderate	146	2.68	2.90	1.45
	Severe	50	2.37	4.33	1.80
Aortic stenosis	Overall	1850	2.48	2.60	1.35
	Normal	1582	2.48	2.40	1.31
	Mild	111	2.45	3.54	1.57
	Moderate	73	2.53	3.43	1.71
	Severe	84	2.39	4.33	1.52
Diastolic dysfunction	Overall	1597	4.58	2.42	1.28
	Normal	536	4.26	1.59	1.44
	Mild	665	4.89	2.77	1.06
	Moderate	263	4.73	2.92	1.41
	Severe	133	3.96	3.01	1.46
Left ventricular dilatation	Overall	2469	3.31	2.11	1.46
	Normal	1925	3.42	1.80	1.30
	Mild	256	3.21	2.98	1.88
	Moderate	94	3.15	3.41	2.40
	Severe	52	3.16	3.75	2.78
	Present	142	2.19	3.35	1.75
Left ventricular systolic dysfunction	Overall	5144	4.81	1.41	1.10
	Normal	3113	4.85	1.28	1.01
	Mild	1042	4.77	1.55	1.14
	Moderate	495	4.82	1.53	1.28
	Severe	494	4.64	1.82	1.39
Mitral regurgitation	Overall	2902	2.56	2.56	1.33
	Normal	1793	2.51	2.34	1.34
	Mild	814	2.62	2.90	1.22
	Moderate	228	2.74	2.82	1.46
	Severe	67	2.71	3.31	1.69
Pericardial effusion	Overall	1295	3.65	2.86	1.48
	Normal	987	2.51	2.97	1.47
	Mild	158	5.10	2.51	1.44
	Moderate	55	11.50	2.15	1.70
	Severe	50	12.81	2.21	1.65
	Present	45	3.94	3.26	1.47
Right ventricular dilatation	Overall	2812	3.54	2.06	1.40
	Normal	2195	3.63	1.79	1.30
	Mild	294	3.43	2.88	1.72
	Moderate	132	3.30	3.28	1.91
	Severe	50	3.43	3.52	1.97
	Present	141	2.51	2.89	1.64
Right ventricular systolic dysfunction	Overall	2640	4.34	1.70	1.37
	Normal	1932	4.29	1.65	1.37
	Mild	445	4.63	1.74	1.31
	Moderate	199	3.98	2.05	1.44
	Severe	64	5.11	2.07	1.51

Table 4 (continued)

Characteristic	Severity	No. of spans	Length	SD	BD
Tricuspid regurgitation	Overall	1954	2.47	2.84	1.48
	Normal	1422	2.48	2.57	1.49
	Mild	294	2.46	3.34	1.40
	Moderate	165	2.37	3.72	1.50
	Severe	73	2.53	4.01	1.56
Wall motion abnormalities	Overall	1334	3.81	2.33	1.10
	Normal	421	3.42	2.38	1.09
	Present	913	4.00	2.31	1.10

Abbreviations: SD span distinctiveness, *BD* span boundary distinctiveness

Table 5 Overview of all used models, including important advantages and disadvantages

Model	Advantages	Disadvantages
Span classification		
Approximate list lookup	Transparency, flexibility, fast, easy to implement	Time-consuming, operator-dependent, cannot generalize beyond provided list
MedCAT (biLSTM)	Can extract medical concepts and their relationships, leveraging knowledge from existing ontologies	Can have limited adaptability to new terms
SpanCategorizer	Uses pooling to make the model more robust, optimized for span classification	More complex model design may require tuning for optimal results
Document classification		
Bag-of-words	Works well with sparse data, simple, easy to implement	Ignores word order and context, which can lead to loss of information
Span classifier heuristic	Allows span-level analysis of results	Suboptimal performance due to the increased complexity of span classification
SetFit	Effective learning from limited data due to few-shot learning	May require hyperparameter tuning to yield optimal results
MedRoBERTa.nl	Pre-trained on Dutch medical text, provides a strong starting point. Capable of capturing context	Requires significant computational resources, may need further domain-specific adaptations
Bidirectional GRU	Captures context backward and forward	May require extensive training to avoid overfitting
Bidirectional CNN	Effective at extracting local patterns and features	May struggle with long-range dependencies

Abbreviations: CNN convolutional neural network, *GRU* gated recurrent unit, *LSTM* long-term short memory unit, *MedCAT* medical concept annotation tool

such as polarity and temporality. Tunstall et al. [40] expanded on this approach with SetFit, a few-shot classification method that fine-tunes a pre-trained sentence encoder via contrastive learning based on label pairs. This fine-tuned encoder then supports a classification head, as shown in Fig. 5.

For our work, we used BioLORD-2023M developed by Remy et al. [34], a multilingual sentence encoder designed to discriminate between medical concepts using existing ontologies, and a sentence encoder that was trained on top of the RobBERTv2 model (see [30, 41]), a Dutch language model without specific domain knowledge. For the classification head we used a $\mu-$SVM model, a technique also applied by Beliveau et al. [42], who reported varying performance among state-of-the-art classification models.

In training, we used SetFit with 500 randomly sampled documents, which constitutes approximately 10% of the total training dataset. The document set was transformed into pairs of positive-negative labeled samples, leading to a quadratic increase in the number of sample pairs. We limited the number of samples to 500 due to practical constraints and the observation that model performance was not improving with increasing sample count. The sentence encoder demonstrating the highest performance is highlighted in the Results section.

MedRoBERTa.nl

MedRoBERTa.nl is based on RoBERTa, a variant of BERT, originally developed by Devlin et al. [43], which itself is rooted in the transformer architecture (see Vaswani et al. [44]). BERT leverages the transformer's bidirectional

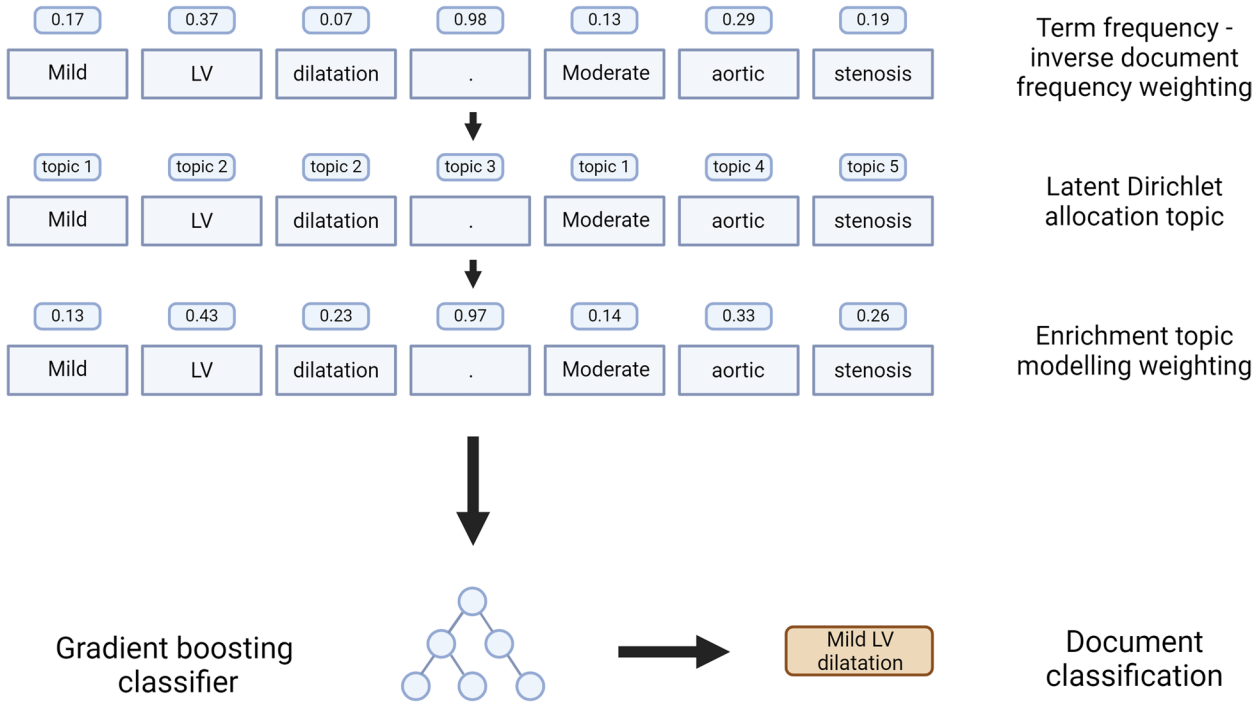

Fig. 4 BOW pipeline involving tokenization, TF-IDF weighting, topic modelling and classification using a gradient-boosted classifier

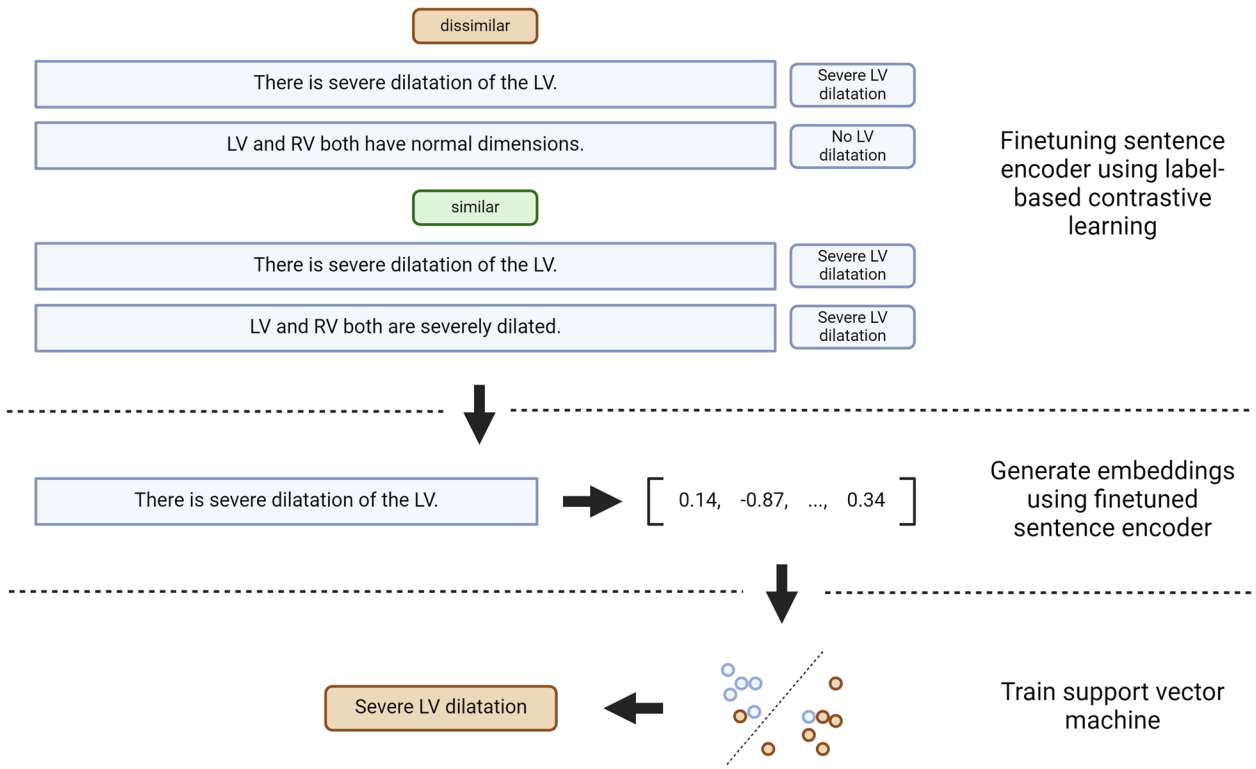

Fig. 5 SetFit pipeline: fine-tuning the sentence encoder with label-based contrastive learning, followed by classification

self-attention mechanism to capture contextual dependencies across words in a document. RoBERTa optimizes this approach by adjusting training objectives and hyperparameters, achieving enhanced performance on several NLP benchmarks. This optimization makes RoBERTa particularly effective for tasks such as classification and span identification, where the model identifies specific sections (spans) of text that correspond to relevant entities or phrases.

MedRoBERTa.nl, developed by Verkijk and Vossen [33], is a Dutch RoBERTa model trained specifically on clinical notes from the Amsterdam University Medical Center. With 125 million parameters, it is currently the only Dutch clinical language model available, and has demonstrated to be well suited for supervised finetuning on clinical data by Van Es et al. [26].

The MedRoBERTa.nl model processes input by first tokenizing the clinical text and embedding it as a sequence of tokens, including positional embeddings to capture word order within the text (Fig. 6). This tokenized input format enables MedRoBERTa.nl to analyze the text bidirectionally, preserving the clinical context for each token in relation to others in the sequence. In this study, MedRoBERTa.nl was finetuned on Dutch clinical text over three epochs, during which all model weights were updated.

Recurrent neural networks

RNNs, including LSTMs, gated recurrent units (GRUs), and quasi-recurrent neural networks (QRNNs), are commonly used for both span and document classification tasks due to its design for sequential data. We selected bidirectional GRU (biGRU) for its ability to capture bidirectional context (left-to-right and right-to-left) in token sequences, which is important for understanding the full sequence context. Unlike standard RNNs, which suffer from vanishing gradients and struggle with long-term dependencies, GRUs and LSTMs handle these dependencies much better. GRUs, have a simpler architecture than LSTMs, using only two gates (update and reset) instead of three, resulting in fewer training times while maintaining comparable performance (Fig. 7).

Convolutional neural networks

CNNs are another powerful type of neural network, often used for span and document classification tasks. In our study, we utilised a bidirectional variant of CNN, which processes text sequences in both forward and backward directions. This bidirectional approach helps capture context from both ends of the sequence, similar to bidirectional GRUs. CNNs excel at capturing local patterns in data, making them well-suited for text, where n-grams or small phrases can be crucial for understanding context. Unlike RNNs, CNNs can process data in parallel, significantly speeding up the training process.

The primary advantage of CNNs is their ability to efficiently capture spatial hierarchies in data through convolutional and pooling layers (Fig. 8). However, they may struggle with maintaining long-range dependencies compared to RNNs like biGRU. Despite this, bidirectional CNNs are computationally efficient and less sensitive to hyperparameter tuning, making them a practical choice for many text classification tasks.

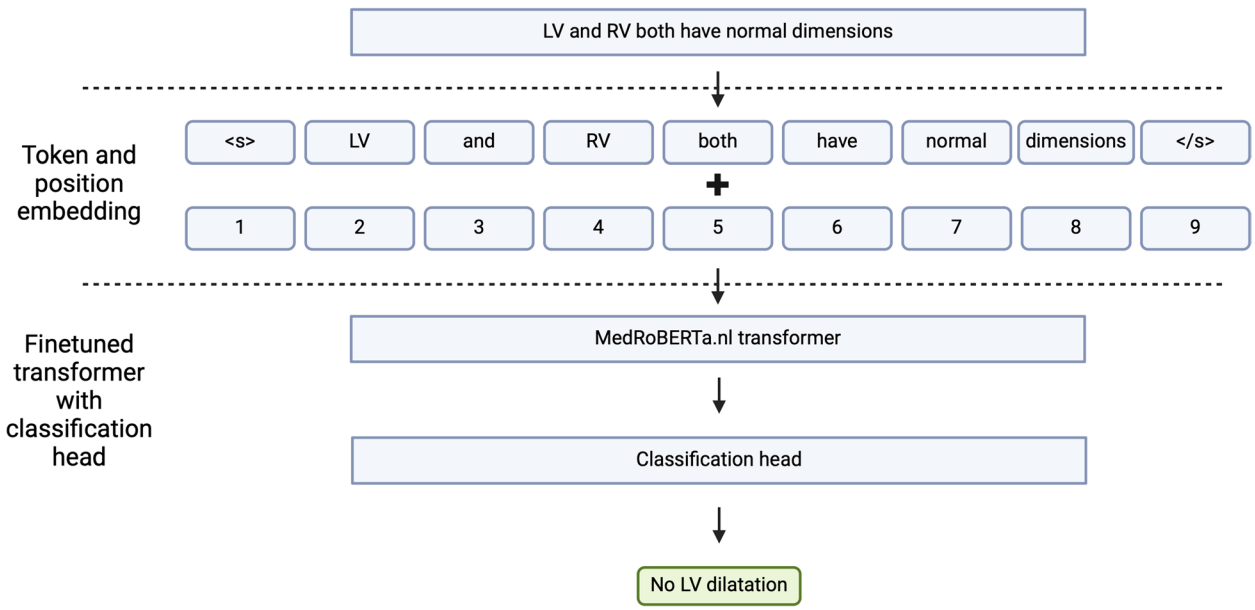

Fig. 6 MedRoBERTa.nl pipeline: input is first tokenized and embedded, then processed through the transformer layers for classification

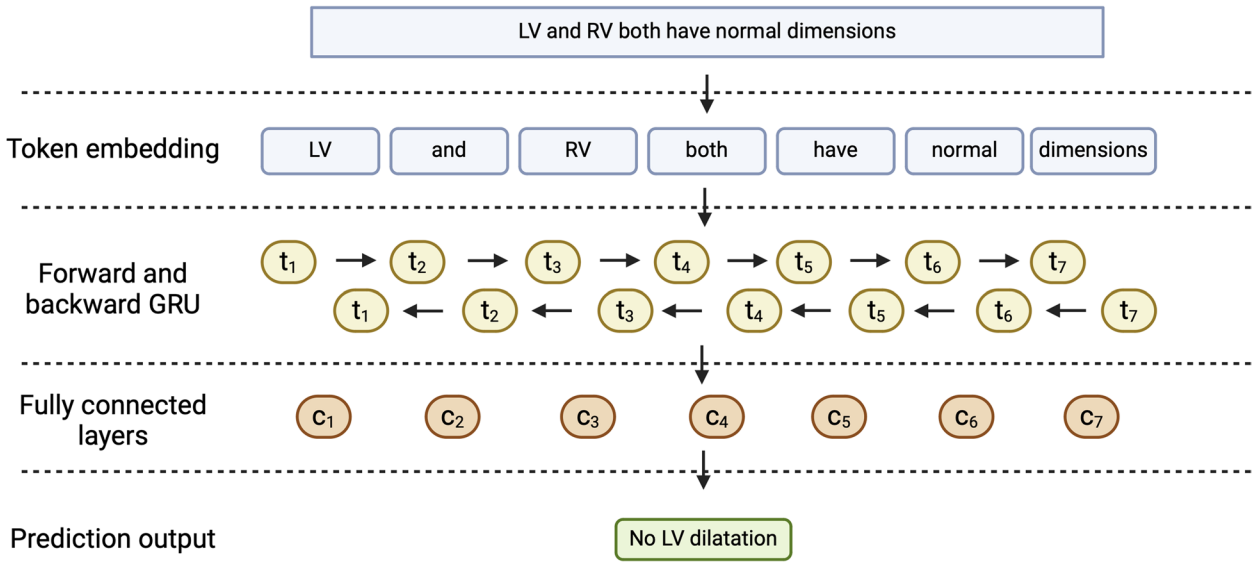

Fig. 7 biGRU pipeline: The input sentence is tokenized and embedded, passing through a bidirectional GRU layer to capture contextual information in both directions. The fully connected layer uses this information to produce a final classification output

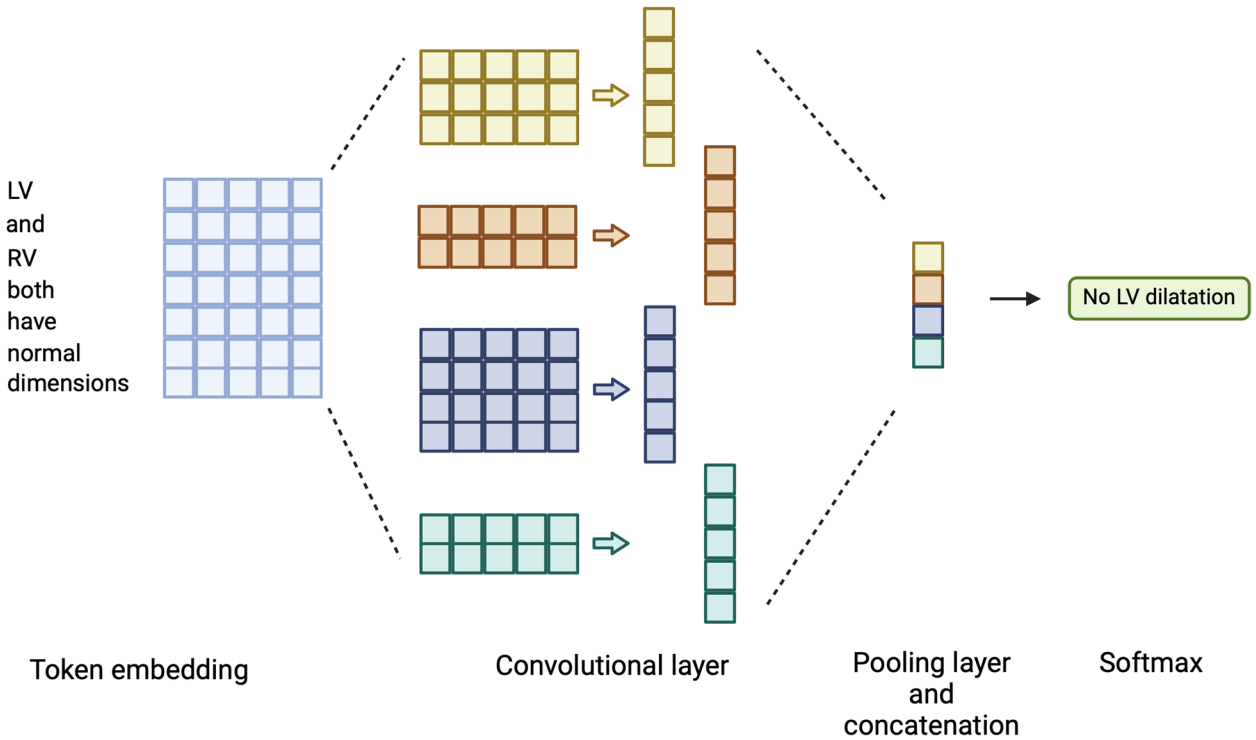

Fig. 8 CNN pipeline: Text is tokenized and embedded, then passed through multiple convolutional layers with varying filter sizes. Outputs from each convolutional layer are pooled and concatenated before passing through a final softmax layer

Performance evaluation

Span classification involves two distinct tasks, identifying spans and subsequently classifying them. Therefore, our performance evaluation included two aspects. We assessed span identification performance using a token-based coverage expressed using the Jaccard index. For span classification, we evaluated assuming the correct spans are identified. Additionally, we measured

end-to-end performance, which combines both span identification and span classification. For both span and document classification, we reported weighted and macro precision, recall, and F1-score.

Finally, in clinical practice it is important to consider the number of false labels, i.e., the number of spans that are falsely labelled with *any* class value (other than "*no label*" or "*normal*"). We present the rate of false labelling relative to the total number of identified spans for our span classification task.

Results

This section provides the performance scores on the span and document-level label extraction tasks.

Span classification

Table 6 shows that for most characteristics, SpanCategorizer achieved the highest weighted and macro F1-scores for the span classification task. However, ALL_{rule} performed particularly well in classifying valvular disorders. This high performance may be attributed to these disorders being often described with very short, distinct phrases (Table 4). Conversely, the remaining characteristics are typically described using longer, less distinctive spans, where SpanCategorizer demonstrated a better performance. MedCAT demonstrated a lower precision and recall in the span classification task. These results may be due to an imperfect span suggestion. This hypothesis is supported by Tables 7 and 8, which illustrate a high performance in span classification when the exact spans containing a label are suggested, but a low Jaccard-index when comparing MedCAT's end-to-end predicted spans containing a label with the ground truth. In addition, Table 9 details that MedCAT has a high percentage of false positive span labels, leading to a reduced precision. This indicates that using MedCAT combined with a greedy span suggester could improve results even further.

Document classification

Results for the document classification task are presented in Tables 10 and 11. From these tables, MedRoBERTa.nl outperforms all other models on weighted and macro F1-score, precision, and recall. Indirect document classification using span classifiers resulted in a suboptimal performance, highlighting the added value of direct document classification models. BOW, our second baseline approach, performed quite well considering that we did not perform feature processing except TF-IDF and lemmatisation. An explanation might be that, because we are dealing with short staccato notes, containing little elaborations, and primarily containing statements of facts. Another reason may be that the number of negations is limited in echocardiogram reports. We also applied a document averaging of clinical word embeddings, but this was not favorable with respect to BOW with TF-IDF.

For MedRoBERTa, we applied a de-abbreviation step to investigate whether the presence of several abbreviations, combined with the relative brevity of the notes, would undermine the model's performance. MedRoBERTa is competitive with methods like biLSTMs, especially in the case of larger contexts. However, we did not observe an improvement over the original texts. This could be due to the already high performance without de-abbreviation. For both biGRU and CNN models, the use of de-abbreviations also did not impact the performance favorably.

Additionally, we experimented with using pre-trained word vectors concatenated with the original trainable

Table 6 Semantic end-to-end performance of span classification methods

	SpanCategorizer			MetaCAT			ALL_{rule}		
Characteristic	**F1**	**recall**	**precision**	**F1**	**recall**	**precision**	**F1**	**recall**	**precision**
Aortic regurgitation	0.90 (0.67)	0.85 (0.62)	**0.94** (0.73)	0.49 (0.46)	0.54 (0.50)	0.50 (0.46)	**0.92** (**0.89**)	**0.90** (**0.87**)	**0.94** (**0.92**)
Aortic stenosis	0.82 (0.74)	0.79 (0.67)	**0.86** (**0.85**)	0.45 (0.38)	0.46 (0.51)	0.43 (0.40)	**0.83** (**0.75**)	**0.84** (**0.75**)	0.83 (0.77)
Diastolic dysfunction	**0.87** (**0.83**)	**0.85** (**0.81**)	**0.90** (**0.86**)	0.55 (0.66)	0.69 (0.66)	0.60 (0.65)	0.60 (0.60)	0.60 (0.59)	0.61 (0.61)
Left ventricular dilatation	**0.84** (**0.89**)	**0.82** (0.85)	**0.85** (**0.93**)	0.57 (0.65)	0.32 (0.46)	0.40 (0.53)	0.75 (0.85)	0.81 (**0.86**)	0.70 (0.86)
Left ventricular systolic dysfunction	**0.77** (**0.42**)	**0.75** (0.41)	**0.79** (**0.43**)	0.33 (0.24)	0.69 (**0.49**)	0.44 (0.32)	0.21 (0.22)	0.16 (0.19)	0.33 (0.31)
Mitral regurgitation	**0.93** (0.71)	0.90 (0.69)	**0.97** (0.72)	0.63 (0.76)	0.59 (0.60)	0.61 (0.66)	0.92 (**0.91**)	**0.91** (**0.89**)	0.93 (**0.92**)
Pericardial effusion	**0.79** (0.28)	**0.70** (0.25)	**0.89** (**0.32**)	0.66 (**0.35**)	0.60 (**0.26**)	0.62 (0.29)	0.74 (0.21)	0.62 (0.19)	0.93 (0.26)
Right ventricular dilatation	**0.90** (0.72)	**0.88** (0.71)	**0.93** (0.74)	0.26 (0.44)	0.23 (0.33)	0.25 (0.37)	0.77 (**0.85**)	0.80 (**0.83**)	0.75 (**0.88**)
Right ventricular systolic dysfunction	**0.89** (**0.64**)	**0.88** (**0.66**)	**0.90** (0.62)	0.61 (0.60)	0.68 (0.51)	0.64 (0.54)	0.53 (0.48)	0.37 (0.33)	0.96 (**0.95**)
Tricuspid regurgitation	0.90 (**0.83**)	0.88 (0.81)	**0.93** (**0.85**)	0.38 (0.40)	0.51 (0.58)	0.41 (0.44)	**0.92** (**0.83**)	**0.93** (**0.91**)	0.91 (0.82)
Wall motion abnormalities	**0.60** (**0.60**)	**0.61** (**0.63**)	**0.59** (**0.59**)	0.24 (0.24)	0.51 (0.52)	0.32 (0.32)	0.16 (0.24)	0.18 (0.25)	0.15 (0.23)

Weighted and macro (in brackets) scores. The highest performance for each characteristic is denoted in bold

Table 7 Semantic performance of span classification methods, assuming matching spans

	SpanCategorizer			MetaCAT			ALL_{rule}		
Characteristic	**F1**	**recall**	**precision**	**F1**	**recall**	**precision**	**F1**	**recall**	**precision**
Aortic regurgitation	0.91 (0.54)	0.85 (0.50)	0.97 (0.58)	**0.98** (0.92)	0.98 (0.89)	0.98 (0.94)	0.91 (**0.89**)	0.86 (0.87)	0.96 (**0.92**)
Aortic stenosis	0.88 (0.63)	0.79 (0.53)	**1.00** (0.78)	**0.97** (**0.84**)	**0.98** (**0.79**)	0.97 (**0.91**)	0.84 (0.75)	0.83 (0.75)	0.85 (0.77)
Diastolic dysfunction	0.91 (0.70)	0.85 (0.64)	**0.98** (0.77)	**0.98** (**0.94**)	**0.98** (**0.95**)	0.98 (0.93)	0.60 (0.60)	0.58 (0.59)	0.62 (0.61)
Left ventricular dilatation	0.90 (0.76)	0.82 (0.71)	**1.00** (0.83)	**0.97** (**0.89**)	**0.97** (**0.9**)	0.97 (**0.88**)	0.82 (0.85)	0.79 (0.86)	0.85 (0.86)
Left ventricular systolic dysfunction	0.84 (0.41)	0.75 (0.36)	**0.97** (0.49)	**0.95** (**0.64**)	**0.95** (**0.65**)	0.95 (**0.63**)	0.20 (0.22)	0.14 (0.18)	0.37 (0.31)
Mitral regurgitation	0.93 (0.57)	0.90 (0.55)	0.96 (0.59)	**0.98** (**0.94**)	**0.98** (**0.93**)	**0.98** (**0.95**)	0.90 (0.91)	0.86 (0.89)	0.94 (0.92)
Pericardial effusion	0.76 (0.24)	0.70 (0.21)	0.85 (0.30)	**0.97** (**0.56**)	**0.97** (**0.53**)	**0.97** (**0.66**)	0.74 (0.21)	0.62 (0.19)	0.93 (0.26)
Right ventricular dilatation	0.93 (0.62)	0.88 (0.59)	**0.98** (0.65)	**0.98** (**0.92**)	**0.98** (**0.94**)	0.98 (**0.91**)	0.78 (0.85)	0.76 (0.83)	0.80 (0.88)
Right ventricular systolic dysfunction	0.91 (0.54)	0.88 (0.53)	0.94 (0.55)	**0.97** (**0.9**)	**0.97** (**0.85**)	**0.97** (**0.95**)	0.52 (0.48)	0.36 (0.33)	0.97 (**0.95**)
Tricuspid regurgitation	0.92 (0.68)	0.88 (0.65)	**0.97** (0.74)	**0.98** (**0.9**)	**0.98** (**0.89**)	**0.98** (**0.91**)	0.90 (**0.83**)	0.89 (**0.84**)	0.91 (**0.82**)
Wall motion abnormalities	0.75 (0.51)	0.61 (0.42)	**0.99** (0.66)	**0.98** (**0.96**)	**0.98** (**0.96**)	0.98 (**0.96**)	0.16 (0.24)	0.18 (0.25)	0.15 (0.23)

Weighted and macro (in brackets) scores. The highest performance for each characteristic is denoted in bold

Table 8 Jaccard-index of span classification methods

Characteristic	**SpanCategorizer**	**MetaCAT**	**ALL_{rule}**
Aortic regurgitation	0.96	0.56	**0.99**
Aortic stenosis	**0.98**	0.47	0.96
Diastolic dysfunction	**0.98**	0.78	0.85
Left ventricular dilatation	**0.96**	0.47	**0.96**
Left ventricular systolic dysfunction	**0.95**	0.74	0.84
Mitral regurgitation	0.99	0.64	0.99
Pericardial effusion	**0.96**	0.76	**0.96**
Right ventricular dilatation	**0.99**	0.32	0.93
Right ventricular systolic dysfunction	**0.99**	0.75	**0.99**
Tricuspid regurgitation	**0.99**	0.57	**0.99**
Wall motion abnormalities	**0.88**	0.55	0.74

The highest performance for each characteristic is denoted in bold

Table 9 Fraction of false positive span labels

Characteristic	**SpanCategorizer**	**MetaCAT**	**ALL_{rule}**
Aortic regurgitation	<0.01	0.13	<0.01
Aortic stenosis	0.01	0.12	<0.01
Diastolic dysfunction	0.03	0.10	0.04
Left ventricular dilatation	0.01	0.05	0.01
Left ventricular systolic dysfunction	0.01	0.76	0.02
Mitral regurgitation	0.01	0.12	0.01
Pericardial effusion	<0.01	0.04	<0.01
Right ventricular dilatation	0.02	0.11	0.01
Right ventricular systolic dysfunction	0.03	0.08	<0.01
Tricuspid regurgitation	<0.01	0.13	0.01
Wall motion abnormalities	0.04	0.22	0.08

embedding layer for the CNN and biGRU models. We did not see a significant improvement in performance, but the added embeddings did incur increased computational cost. The benefit of such pre-trained embeddings might be more noticeable with smaller training sizes, adding contextual information that the model might not learn from a small dataset alone. We also experimented with stacked dilated CNNs and TextCNN, again with no noticeable performance improvement while incurring increased computational cost.

SetFit performed well considering that we used about 10% of the samples resulting in about 12,000 contrastive examples. The sentence embeddings based on the BioLord2023 model are notably worse than the sentence embeddings based on the more generic RobBERTv2 model (Additional file 3). This can be explained by the fact that the SBERT model for RobBERTv2 was trained on a broad semantic range or sentences whereas BioLORD used the LORD training that seeks to maximize difference between medical concept definitions and i.e. is more suitable for named-entity-recognition tasks.

Retraining all models on a reduced label set improves performance markedly (Table 12). Using a further reduced label set only including the presence or absence of a mention of an characteristic yielded near-perfect results. This approach can be particularly useful in practical applications where high precision is required, and resources for manual data labelling are limited.

Discussion

This study aimed to explore and compare various NLP methods for extracting clinical labels from unstructured Dutch echocardiogram reports. We developed and evaluated several approaches for both span- and

Table 10 Semantic performance of document classification methods

	BOW			SetFit (RobBERT)			MedRoBERTa.nl			biGRU			CNN		
Characteristic	**F1**	**recall**	**precision**	**F1**	**recall**	**precision**	**F1**	**recall**	**precision**	**F1**	**recall**	**precision**	**F1**	**recall**	**precision**
Aortic regurgitation	0.90 (0.74)	0.90 (0.65)	0.90 (0.84)	0.93 (0.86)	0.93 (0.88)	0.93 (0.84)	**0.96** (**0.93**)	0.95 (**0.90**)	**0.96** (**0.96**)	0.94 (0.89)	0.94 (0.88)	0.93 (0.90)	0.95 (0.89)	0.95 (0.85)	0.95 (0.95)
Aortic stenosis	0.95 (0.77)	0.93 (0.72)	0.93 (0.89)	0.91 (0.82)	0.91 (0.93)	0.91 (0.75)	**0.96** (0.89)	**0.95** (0.91)	**0.96** (0.88)	0.94 (0.88)	0.94 (0.87)	0.94 (0.89)	0.94 (**0.91**)	0.94 (**0.93**)	0.94 (0.89)
Diastolic dysfunction	0.93 (0.82)	0.93 (0.80)	0.93 (0.84)	0.95 (0.91)	0.95 (**0.97**)	0.95 (0.87)	**0.97** (**0.95**)	**0.97** (0.96)	**0.98** (**0.94**)	0.93 (0.84)	0.93 (0.82)	0.93 (0.86)	0.94 (0.93)	0.93 (0.93)	0.94 (0.92)
Left ventricular dilatation	0.86 (0.56)	0.86 (0.51)	0.85 (0.63)	0.95 (0.91)	0.95 (**0.95**)	0.95 (0.87)	**0.96** (**0.95**)	**0.96** (**0.95**)	**0.97** (0.95)	0.94 (0.90)	0.93 (0.87)	0.94 (**0.96**)	0.94 (0.93)	0.93 (0.93)	0.94 (0.92)
Left ventricular systolic dysfunction	0.89 (0.82)	0.89 (0.80)	0.89 (0.84)	0.95 (0.91)	0.95 (0.92)	0.95 (0.89)	**0.97** (**0.93**)	**0.96** (**0.92**)	**0.97** (**0.95**)	0.93 (0.89)	0.93 (0.89)	0.93 (0.91)	0.95 (0.92)	0.95 (0.91)	0.95 (0.92)
Mitral regurgitation	0.88 (0.68)	0.88 (0.65)	0.88 (0.74)	0.94 (0.88)	0.94 (0.93)	0.94 (0.85)	**0.96** (0.92)	**0.96** (**0.94**)	**0.97** (0.90)	0.92 (0.87)	0.92 (0.85)	0.93 (0.88)	0.94 (0.92)	0.94 (0.92)	0.94 (0.93)
Pericardial effusion	0.95 (0.42)	0.95 (0.40)	0.94 (0.48)	0.92 (0.51)	0.92 (0.60)	0.92 (0.49)	**0.98** (**0.81**)	**0.98** (**0.80**)	0.97 (0.84)	0.97 (0.75)	0.96 (0.69)	**0.98** (**0.86**)	0.97 (0.63)	0.97 (0.60)	0.97 (0.69)
Right ventricular dilatation	0.86 (0.54)	0.87 (0.49)	0.86 (0.68)	0.92 (0.81)	0.92 (0.92)	0.92 (0.74)	**0.96** (**0.95**)	**0.96** (**0.96**)	**0.96** (**0.94**)	0.93 (0.91)	0.93 (0.90)	0.94 (0.93)	0.94 (0.89)	0.94 (0.88)	0.94(0.89)
Right ventricular systolic dysfunction	0.89 (0.66)	0.89 (0.64)	0.89 (0.75)	0.94 (0.89)	0.94 (0.93)	0.94 (0.85)	**0.96** (**0.93**)	**0.96** (**0.94**)	**0.96** (**0.92**)	0.92 (0.76)	0.91 (0.77)	0.92 (0.75)	0.91 (0.78)	0.90 (0.75)	0.92 (0.84)
Tricuspid regurgitation	0.90 (0.64)	0.90 (0.63)	0.90 (0.79)	0.92 (0.83)	0.92 (0.86)	0.92 (0.80)	0.96 (0.92)	0.96 (0.92)	0.96 (0.92)	0.95 (0.92)	0.95 (0.91)	0.95 (0.94)	0.96 (0.95)	0.96 (**0.95**)	0.96 (0.94)
Wall motion abnormalities	0.95 (0.90)	0.95 (0.87)	0.95 (0.93)	0.95 (0.92)	0.95 (0.93)	0.95 (0.91)	**0.97** (**0.95**)	**0.97** (**0.95**)	**0.97** (**0.96**)	0.95 (0.92)	0.95 (0.90)	0.95 (0.93)	0.96 (0.94)	0.96 (0.92)	0.96 (**0.96**)

Weighted and macro (in brackets) scores. The highest performance for each characteristic is denoted in bold

Table 11 Semantic performance of span → document classification heuristics

	SpanCategorizer			MedCAT			ALL_{rule}		
Characteristic	**F1**	**recall**	**precision**	**F1**	**recall**	**precision**	**F1**	**recall**	**precision**
Aortic regurgitation	**0.95** (0.74)	**0.95** (0.71)	**0.95** (0.77)	0.6 (0.46)	0.58 (0.56)	0.63 (0.42)	**0.95** (**0.91**)	**0.95** (**0.88**)	**0.95** (**0.95**)
Aortic stenosis	0.94 (0.83)	0.94 (0.76)	0.94 (0.93)	0.64 (0.48)	0.62 (0.66)	0.67 (0.42)	**0.95** (**0.9**)	**0.95** (**0.86**)	**0.95** (**0.96**)
Diastolic dysfunction	**0.94** (**0.87**)	**0.94** (**0.84**)	**0.94** (0.91)	0.75 (0.67)	0.74 (0.77)	0.81 (0.62)	0.93 (0.82)	0.93 (0.76)	0.93 (**0.93**)
Left ventricular dilatation	0.91 (0.59)	0.91 (0.59)	0.91 (0.6)	0.65 (0.54)	0.67 (0.58)	0.69 (0.54)	**0.94** (**0.91**)	**0.94** (**0.89**)	**0.94** (**0.94**)
Left ventricular systolic dysfunction	**0.92** (**0.88**)	**0.92** (**0.89**)	**0.92** (**0.89**)	0.87 (0.79)	0.86 (0.8)	0.88 (0.81)	0.33 (0.37)	0.33 (0.41)	0.33 (0.75)
Mitral regurgitation	**0.96** (**0.92**)	**0.96** (**0.9**)	**0.96** (**0.95**)	0.67 (0.64)	0.67 (0.65)	0.68 (0.64)	0.94 (**0.92**)	0.94 (**0.9**)	0.94 (0.94)
Pericardial effusion	0.95 (0.32)	0.95 (0.32)	0.95 (0.32)	0.87 (**0.48**)	0.85 (**0.55**)	0.9 (**0.6**)	**0.96** (0.37)	**0.96** (0.46)	**0.96** (0.36)
Right ventricular dilatation	**0.93** (0.72)	**0.93** (0.68)	**0.93** (0.78)	0.64 (0.43)	0.63 (0.49)	0.67 (0.44)	0.9 (**0.83**)	0.9 (**0.8**)	0.9 (**0.88**)
Right ventricular systolic dysfunction	**0.94** (**0.72**)	**0.94** (**0.75**)	**0.94** (0.7)	0.78 (0.63)	0.78 (0.67)	0.8 (0.6)	0.72 (0.55)	0.72 (0.47)	0.72 (**0.89**)
Tricuspid regurgitation	**0.96** (0.92)	**0.96** (0.9)	**0.96** (0.96)	0.6 (0.48)	0.57 (0.69)	0.68 (0.43)	**0.96** (**0.97**)	**0.96** (**0.96**)	**0.96** (**0.98**)
Wall motion abnormalities	**0.95** (**0.92**)	**0.95** (0.9)	**0.95** (**0.96**)	0.55 (0.45)	0.52 (0.61)	0.77 (0.48)	**0.95** (**0.92**)	**0.95** (**0.93**)	**0.95** (0.91)

Weighted and macro (in brackets) scores. The highest performance for each characteristic is denoted in bold

document-level label extraction on an internal test set, demonstrating high performance in identifying eleven commonly described cardiac characteristics, including left and right ventricular systolic dysfunction, left and right ventricular dilatation, diastolic dysfunction, aortic stenosis, aortic regurgitation, mitral regurgitation, tricuspid regurgitation, pericardial effusion, and wall motion abnormalities. The main findings indicate that SpanCategorizer consistently outperformed other models in span-level classification tasks, achieving weighted F1-scores ranging from 0.60 to 0.93 across these characteristics, while MedRoBERTa.nl excelled in document-level classification with a weighted F1-score exceeding 0.96 for all characteristics.

In this study, we observed a variation in results of different span classification approaches. The baseline approach, using regular expressions, achieved a high performance for some characteristics but performed poorly for others. These outcomes are likely linked to span length, frequency, and distinctiveness [23]. Our most poorly performing characteristics - left ventricular systolic dysfunction, pericardial effusion, and wall motion abnormalities - have larger span lengths, and lower span frequencies. Macro performance is particularly impacted by the 'severe' classes, which have a low span frequency and high span length, which have both been previously linked to worse performance [23].

Although no other Dutch studies have focused on information extraction from echocardiogram reports, comparisons can be made to studies conducted on English-language echocardiogram data. Most of these studies address both continuous and discrete measurement extraction, whereas our study uniquely focuses solely on discrete measurement extraction. This distinction makes direct comparisons challenging. However, when focusing on the extraction of discrete measurements, our methods demonstrate competitive [7] or superior [10, 43] performance. For instance, F1-scores reported in [7] range between 0.93 and 0.94, whereas our document classification approach using MedRoBERTa.nl achieves F1-scores exceeding 0.96 across all cardiac characteristics. These findings highlight the effectiveness of our methods for discrete cardiac label extraction, particularly in a non-English setting.

The MedCAT approach has a very high overall precision but lacked recall due to imperfect span suggestions. Therefore, for medical applications, it may be more effective to use a greedy span-classifier as the primary span suggestion method, with a NER+L extraction serving as an augmentation tool to extract additional features. Alternatively, to make the MedCAT model more robust, we should consider using fuzzy matching with varying proximities, using tools like clinlp [46], instead of adding possible spans directly from the training phase of the labelling process in Prodigy. Another approach, given the results from the document classification task, could involve training a RoBERTa-based or CNN/biGRU span classifier, using either a MedCAT or greedy span suggester. Additionally, a joint entity/relation extraction model could be constructed [43]. However, these approaches are outside the scope of the current paper and require significantly higher computational cost.

For document classification, the MedRoBERTa.nl model demonstrated the best overall performance. This aligns with previous findings, which highlight the additional value of BERT-based models in cases involving infrequently occurring spans [23]. We did not attempt

Table 12 Semantic performance of document classification methods for simplified label scheme (No label, Normal, and Present)

	BOW			SetFit (RobBERT)			MedRoBERTa.nl			biGRU			CNN		
Characteristic	**F1**	**recall**	**precision**	**F1**	**recall**	**precision**	**F1**	**recall**	**precision**	**F1**	**recall**	**precision**	**F1**	**recall**	**precision**
Aortic regurgitation	0.92 (0.89)	0.92 (0.88)	0.92 (0.89)	0.94 (0.93)	0.94 (0.94)	0.94 (0.91)	0.97 (0.97)	0.97 (0.97)	0.97 (0.97)	0.96 (0.95)	0.96 (0.95)	0.96 (0.96)	0.96 (0.96)	0.96 (0.95)	0.96 (0.96)
Aortic stenosis	0.94 (0.89)	0.94 (0.88)	0.94 (0.90)	0.91 (0.86)	0.91 (0.94)	0.91 (0.82)	0.95 (0.93)	0.95 (0.95)	0.95 (0.93)	0.95 (0.93)	0.95 (0.94)	0.95 (0.92)	0.96 (0.94)	0.96 (0.95)	0.96 (0.95)
Diastolic dysfunction	0.94 (0.91)	0.94 (0.90)	0.94 (0.91)	0.95 (0.92)	0.95 (**0.97**)	0.95 (0.89)	**0.97** (**0.96**)	**0.97** (**0.97**)	**0.97** (**0.96**)	0.96 (0.94)	0.96 (0.95)	0.96 (0.92)	0.96 (0.95)	0.96 (0.95)	**0.97** (0.95)
Left ventricular dilatation	0.88 (0.82)	0.88 (0.81)	0.88 (0.84)	0.95 (0.94)	0.95 (0.96)	0.95 (0.93)	**0.96** (0.94)	**0.96** (**0.95**)	**0.96** (0.94)	0.95 (0.94)	0.95 (0.94)	0.95 (0.94)	**0.96** (**0.95**)	**0.96** (**0.95**)	**0.96** (**0.95**)
Left ventricular systolic dysfunction	0.92 (0.90)	0.92 (0.89)	0.92 (0.90)	0.96 (0.94)	0.96 (**0.95**)	0.96 (0.93)	**0.97** (**0.95**)	**0.97** (**0.95**)	**0.97** (**0.95**)	0.96 (0.94)	0.96 (0.94)	**0.97** (0.94)	0.96 (0.94)	0.96 (0.93)	0.96 (0.94)
Mitral regurgitation	0.90 (0.88)	0.90 (0.88)	0.90 (0.89)	0.94 (0.94)	0.94 (0.95)	0.94 (0.93)	**0.97** (**0.97**)	**0.97** (**0.97**)	**0.97** (0.96)	0.95 (0.94)	0.94 (0.95)	0.95 (0.94)	0.96 (0.95)	0.96 (0.96)	0.96 (0.95)
Pericardial effusion	0.96 (0.84)	0.97 (0.82)	0.96 (0.88)	0.95 (0.85)	0.95 (0.93)	0.95 (0.79)	**0.99** (**0.95**)	**0.99** (**0.96**)	**0.99** (0.94)	0.98 (0.94)	0.98 (0.95)	0.98 (0.94)	0.98 (0.94)	0.98 (0.94)	0.98 (**0.95**)
Right ventricular dilatation	0.87 (0.79)	0.88 (0.77)	0.87 (0.81)	0.91 (0.86)	0.91 (0.91)	0.91 (0.83)	**0.95** (**0.93**)	**0.95** (**0.95**)	**0.96** (0.92)	0.94 (0.92)	0.94 (0.92)	0.94 (0.92)	**0.95** (0.92)	**0.95** (0.93)	0.95 (0.92)
Right ventricular systolic dysfunction	0.91 (0.86)	0.90 (0.85)	0.90 (0.87)	0.93 (0.91)	0.93 (0.94)	0.93 (0.90)	**0.97** (**0.95**)	**0.97** (**0.94**)	**0.97** (**0.96**)	0.94 (0.92)	0.94 (0.91)	0.94 (0.93)	0.94 (0.92)	0.94 (0.92)	0.95 (0.93)
Tricuspid regurgitation	0.93 (0.90)	0.93 (0.90)	0.93 (0.89)	0.93 (0.91)	0.93 (0.93)	0.93 (0.89)	**0.97** (**0.97**)	**0.97** (0.96)	**0.97** (**0.97**)	0.96 (0.95)	0.96 (0.95)	0.96 (0.95)	**0.97** (0.96)	**0.97** (0.96)	**0.97** (0.95)
Wall motion abnormalities	0.94 (0.90)	0.94 (0.85)	0.94 (0.92)	0.95 (0.92)	0.95 (0.93)	0.95 (0.91)	**0.97** (**0.95**)	**0.97** (**0.94**)	**0.97** (**0.96**)	0.96 (0.93)	0.96 (0.90)	0.96 (**0.96**)	**0.97** (0.94)	**0.97** (0.93)	**0.97** (**0.96**)

Weighted and macro (in brackets) scores. The highest performance for each characteristic is denoted in bold

to train a BERT-based model from scratch due to the limited number of available documents. Previous studies have shown that pre-training on a small corpus yields suboptimal results, whereas models with general domain pre-training, such as MedRoBERTa.nl, achieve highly competitive results without requiring domain-specific feature engineering [48–53]. The biGRU and CNN models demonstrated a competitive performance, especially considering their significantly lower computational cost. Alternatives like TextCNN or hierarchical architectures such as Hierarchical Attention Networks might perform better with longer contexts, such as discharge summaries [54, 55].

The BOW approach, while effective considering its simplicity, could have been extended with more sophisticated weighting mechanisms, such as incorporating negation estimation, part-of-speech tagging, and dependency parsing. These additions could have improved the contextual understanding of the text, potentially leading to better document classification. However, such extensions would require significantly more complex feature engineering and computational resources, which were beyond the scope of this study.

Regarding the SetFit method, three remarks can be made. First, training a new sentence transformer from scratch based on the MedRoBERTa.nl model might yield better results than using the arithmetic mean. Second, the BioLORD-2023M model is constrastively trained to discriminate between medical span-level concepts, rather than explicitly between semantic differences. Third, we achieved performance close to the best-performing method using only 10% of the data. Therefore, this approach may be most suitable given the resources required for manual data labelling.

The class distribution in our dataset reflects real-world practice, with over 75% of documents lacking a label for at least one characteristic, and a small percentage containing moderate or severe labels. While this distribution poses challenges for model performance, particularly in terms of macro scores, it also highlights the need for models to perform well under realistic clinical conditions. Expanding the dataset was not feasible due to the extensive manual labeling process, which already took several months. An alternative approach to enhance model performance could involve utilizing English BERT-based models on translated texts, as suggested by Muizelaar et al. [51].

We employed a single train/test split for our experiments, which, while practical, could introduce certain limitations. One potential concern is the risk of overfitting to the specific data in the training set, particularly when using handcrafted features like regular expressions. This might result in models that perform well on the test set but may not generalize as effectively to new, unseen data. Ideally, a cross-validation approach would provide a more comprehensive evaluation by averaging performance across multiple splits, thereby reducing the variance and offering a more robust assessment of model performance. However, given the infeasibility of developing regular expressions for each fold, our approach represents a pragmatic balance between practical constraints and methodological rigor. The use of a single split also means that our performance estimates may be somewhat optimistic, as they are tied to the specific characteristics of the selected test set. This is particularly relevant for our span classification tasks, where the performance varied significantly across different span types. In future work, incorporating cross-validation or a more extensive test set could help mitigate these limitations, providing a clearer picture of how well these models might perform in broader clinical applications.

Our findings suggest distinct use cases for span and document classification within clinical practice. Span classification, while adding a layer of explainability by highlighting specific spans that contribute to a particular label, exhibit too much variability in performance to be reliably used in clinical settings. This inconsistency, especially across different characteristics, limits its utility for direct clinical application at this stage. In contrast, document classification demonstrated significantly better and more consistent performance, making it a more viable option for integration into clinical workflows. This approach could be effectively used for tasks such as constructing patient cohorts for research or automating parts of the diagnostic pipeline. Additionally, we observed that reducing the number of labels significantly improved the performance of document classification models. This reduced label model might be employed to flag cases that require more detailed review, either by activating a clinician's attention or by supporting active labeling in research settings, such as using Prodigy. This approach not only enhances model accuracy but also provides a practical pathway for implementing NLP tools in clinical environments where efficiency and precision are essential.

Conclusions

This study addresses the need for information retrieval on Dutch medical data, specifically focusing on extracting span- and document-level labels from unstructured echocardiogram reports in Dutch. By evaluating both neural and statistical NLP methods, we provide a comprehensive baselien for structured information retrieval in a domain with limited pre-trained resources. Our results demonstrate high performance in identifying eleven cardiac characteristics, with the SpanCategorizerachieving

weighted F1-scores ranging from 0.60 to 0.93 for span classification, and the MedRoBERTa.nl model surpassing a weighted F1-score of 0.96 document-level classification.

This comparison of out-of-the-box models highlights that MedRoBERTa.nl and SpanCategorizer are effective tools for document and span-level diagnosis extraction in the Dutch clinical context. These models, publicly available through HuggingFace, provide a practical and accessible starting point for individuals aiming to implement NLP-based tools without extensive customization or hyperparameter tuning.

Future work may include validation in external institutions, ensemble modelling, or the extension to other cardiac characteristics. In case of a limited amount of data, SetFit may be a suitable alternative for document classification.

Abbreviations

ALL	Approximate list lookup
biGRU	Bidirectional GRU
biLSTM	Bilateral LSTM
BOW	Bag-of-words
CNN	Convolutional neural network
CRF	Conditional random fields
EHR	Electronic health record
GRU	Gated recurrent unit
ICD	International Classification of Disease
LSTM	Long short-term memory
NER	Named entity recognition
NLP	Natural language processing
QRNN	Quasi-recurrent neural network
RNN	Recurrent neural network
SVM	Support vector machine
TF-IDF	Term frequency-inverse document frequency
UMCU	University Medical Center Utrecht

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-02897-w.

Additional file 1. Model parameters. Provides additional information on the model parameters, settings, and supporting packages used to develop the span and document classification model.

Additional file 2. Labelling instructions. Provides additional information on the instructions given to the manual labellers.

Additional file 3. SetFit model performance. Provides two additional table on the performance of both models used for SetFit. Supplementary Table 1 provides a description of the performance of both SetFit models for the main document classification task. Supplementary Table 2 provides the same description for the reduced label scheme.

Additional file 4. SpanCategorizer model performance per negative label weighting. Provides one additional table on the performance of SpanCategorizer for each negative label weighting used.

Acknowledgements
The authors thank Celina Berkhoff for her contribution to the manual labelling process. Figures were created with BioRender.com.

Authors' contributions
B.A., M.V., B.E., and R.E. were responsible for conceptualisation. Data curation and manual labeling was performed by B.A. and M.V.. Methodology was developed by B.E., B.A. and M.V.. P.H. and A.T. were responsible for project supervision and administration. B.E., B.A., and M.V. performed experiments and validated the results. A.T. and D.O. gave clinical input on endpoint definitions. The original draft of this manuscript was written by B.A. and B.E.. All authors read and approved the final manuscript.

Funding
The work received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101057849 (DataTools4Heart project).
The collaboration project is co-funded by PPP Allowance awarded by Health~Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships.
This work is part of the project MyDigiTwin with project number 628.011.213 of the research programme "COMMIT2DATA - Big Data & Health" which is partly financed by the Dutch Research Council (NWO).

Data availability
The datasets generated and/or analysed during the current study are not publicly available due to potential privacy-sensitive information, but are available from the corresponding author upon reasonable request and local institutional approval. Research code is publicly available on GitHub.

Declarations

Ethics approval and consent to participate
The UMCU quality assurance research officer confirmed under project number 22U-0292 that this study does not fall under the scope of the Dutch Medical Research Involving Human Subjects Act (WMO) and therefore does not require approval from an accredited medical ethics committee. The study was performed compliant with local legislation and regulations. All patient data were deidentified in compliance with the European Union General Data Protection Regulation, and as a result, written informed consent was not required by the UMCU ethical committee.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 15 August 2024 Accepted: 28 January 2025
Published online: 07 March 2025

References

1. Misset B, Nakache D, Vesin A, Darmon M, Garrouste-Orgeas M, Mourvillier B, et al. Reliability of Diagnostic Coding in Intensive Care Patients. Crit Care. 2008;12(4):R95. https://doi.org/10.1186/cc6969.
2. de Hond TA, Niemantsverdriet MS, van Solinge WW, Oosterheert JJ, Haitjema S, Kaasjager KA. Sepsis labels defined by claims-based methods are ill-suited for training machine learning algorithms. Clin Microbiol Infect. 2022;28:1170–1.
3. Anderson HD, Pace WD, Brandt E, Nielsen RD, Allen RR, Libby AM, et al. Monitoring suicidal patients in primary care using electronic health records. J Am Board Fam Med. 2015;28(1):65–71.
4. Perlis R, Iosifescu D, Castro V, Murphy S, Gainer V, Minnier J, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychol Med. 2012;42(1):41–50.
5. Spasic I, Nenadic G. Clinical Text Data in Machine Learning: Systematic Review. JMIR Med Inform. 2020;8(3):e17984. https://doi.org/10.2196/17984.
6. Pearlman AS, Ryan T, Picard MH, Douglas PS. Evolving Trends in the Use of Echocardiography. J Am Coll Cardiol. 2007;49(23):2283–91. https://doi.org/10.1016/j.jacc.2007.02.048.
7. Nath C, Albaghdadi MS, Jonnalagadda SR. A Natural Language Processing Tool for Large-Scale Data Extraction from Echocardiography Reports.

PLoS ONE. 2016;11(4):e0153749. https://doi.org/10.1371/journal.pone.0153749.
8. Szeker S, Fogarassy G, Vathy-Fogarassy A. A General Text Mining Method to Extract Echocardiography Measurement Results from Echocardiography Documents. Artif Intell Med. 2023;143:102584. https://doi.org/10.1016/j.artmed.2023.102584.
9. Kaspar M, Morbach C, Fette G, Ertl M, Seidlmayer LK, Krebs J, et al. Information Extraction from Echocardiography Reports for a Clinical Follow-up Study-Comparison of Extracted Variables Intended for General Use in a Data Warehouse with Those Intended Specifically for the Study. Methods Inf Med. 2019;58(4):140–50. https://doi.org/10.1055/s-0039-3402069.
10. Patterson O, Freiberg M, Skanderson M, Fodeh S, Brandt C, DuVall S. Unlocking Echocardiogram Measurements for Heart Disease Research through Natural Language Processing. BMC Cardiovasc Disord. 2017;17(151). https://doi.org/10.1186/s12872-017-0580-8.
11. Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, Jonnalagadda SR, et al. An information extraction framework for cohort identification using electronic health records. AMIA Summits Transl Sci Proc. 2013;2013:149.
12. Slater LT, Bradlow W, Hoehndorf R, Motti DF, Ball S, Gkoutos GV. Komenti: a semantic text mining framework. bioRxiv. 2020;2020–08.
13. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17(5):507–13. https://doi.org/10.1136/jamia.2009.001560.
14. Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. PLoS Comput Biol. 2013;9(2):e1002854. https://doi.org/10.1371/journal.pcbi.1002854.
15. Saha SK, Sarkar S, Mitra P. Feature Selection Techniques for Maximum Entropy Based Biomedical Named Entity Recognition. J Biomed Inform. 2009;42:905–11. https://doi.org/10.1016/j.jbi.2008.12.012.
16. Ruan W, Lee WS. Recognising Named Entity of Medical Imaging Procedures in Clinical Notes. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2018.
17. Qin L, Xu X, Ding L, Li Z, Li J. Identifying Diagnosis Evidence of Cardiogenic Stroke from Chinese Echocardiograph Reports. BMC Med Inform Decis Mak. 2020;20(126). https://doi.org/10.1186/s12911-020-1106-3.
18. Richter-Pechanski P, Geis NA, Kiriakou C, Schwab DM, Dieterich C. Automatic Extraction of 12 Cardiovascular Concepts from German Discharge Letters Using Pre-Trained Language Models. Digit Health. 2021;7:1–10. https://doi.org/10.1177/20552076211057662.
19. Wang Z, Shah AD, Tate AR, Denaxas S, Shawe-Taylor J, Hemingway H. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. PLoS ONE. 2012;7(1):e30412.
20. Mustafa A, Rahimi Azghadi M. Automated machine learning for healthcare and clinical notes analysis. Computers. 2021;10(2):24.
21. Sugimoto K, Takeda T, Oh JH, Wada S, Konishi S, Yamahata A, et al. Extracting Clinical Terms from Radiology Reports with Deep Learning. J Biomed Inform. 2021;116. https://doi.org/10.1016/j.jbi.2021.103729.
22. Garcia-Largo MAMC, Segura-Bedmar I. Extracting Information from Radiology Reports by Natural Language Processing and Deep Learning. In: Conference and Labs of the Evaluation Forum. 2021.
23. Papay S, Klinger R, Padó S. Dissecting span identification tasks with performance prediction. http://arxiv.org/abs/2010.02587. Accessed 3 July 2024.
24. Navarro DF, Ijaz K, Rezazadegan D, Rahimi-Ardabili H, Dras M, Coiera E, et al. Clinical Named Entity Recognition and Relation Extraction Using Natural Language Processing of Medical Free Text: A Systematic Review. Int J Med Informa. 2023;177:105–22. https://doi.org/10.1016/j.ijmedinf.2023.105122.
25. Puts S, Nobel M, Zegers C, Bermejo I, Robben S, Dekker A. How Natural Language Processing Can Aid With Pulmonary Oncology Tumor Node Metastasis Staging From Free-Text Radiology Reports: Algorithm Development and Validation. JMIR Formative Res. 2023;7:e38125. https://doi.org/10.2196/38125.
26. Es B, Reteig LC, Tan SC, Schraagen M, Hemker MM, Arends SRS, et al. Negation detection in dutch clinical texts: an evaluation of rule-based and machine learning methods. http://arxiv.org/abs/2209.00470. Accessed 1 May 2024.
27. Jantscher M, Gunzer F, Kern R, Hassler E, Tschauner S, Reishofer G. Information Extraction from German Radiological Reports for General Clinical Text and Language Understanding. Sci Rep. 2023;13(2353). https://doi.org/10.1038/s41598-023-29323-3.
28. Ahumada R, Dunstan J, Rojas M, Peñafiel S, Paredes I, Báez P. Automatic Detection of Distant Metastasis Mentions in Radiology Reports in Spanish. JCO Clin Cancer Inform. 2024. https://doi.org/10.1200/CCI.23.00130.
29. Vries W, Cranenburgh A, Bisazza A, Caselli T, Noord G, Nissim M. BERTje: a Dutch BERT model. http://arxiv.org/abs/1912.09582. Accessed 14 June 2024.
30. Delobelle P, Winters T, Berendt B. RobBERT: a Dutch RoBERTa-based Language Model. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. 2020. pp. 3255–65.
31. Delobelle P, Remy F. RobBERT-2023: Keeping Dutch Language Models Up-To-Date at a Lower Cost Thanks to Model Conversion. Comput Linguist Neth J. 2024;13:193–203.
32. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a Robustly Optimized BERT Pretraining Approach. http://arxiv.org/abs/1907.11692. Accessed 25 Apr 2024.
33. Verkijk S, Vossen P. MedRoBERTa.nl: A Language Model for Dutch Electronic Health Records. Comput Linguist Neth J. 2021;11:141–59.
34. Remy F, Demuynck K, Demeester T. BioLORD-2023: Semantic Textual Representations Fusing Large Language Models and Clinical Knowledge Graph Insights. J Am Med Inform Assoc. 2024;00:1–12. https://doi.org/10.1093/jamia/ocae029.
35. Montani I, Honnibal M. Prodigy: A modern and scriptable annotation tool for creating training data for machine learning models. Explosion. https://prodi.gy/.
36. Kraljevic Z, Saerle T, Shek A, Roguski L, Noor K, Bean D, et al. Multi-Domain Clinical Natural Language Processing with MedCAT: The Medical Concept Annotation Toolkit. Artif Intell Med. 2021;117. https://doi.org/10.1016/j.artmed.2021.102083.
37. Honnibal M, Montani I, Van Landeghem S, Boyd A. spaCy: Industrial-strength Natural Language Processing in Python. 2020. https://doi.org/10.5281/zenodo.1212303.
38. Bagheri A, Sammani A, Van Der Heijden PGM, Asselbergs FW, Oberski DL. ETM: Enrichment by Topic Modeling for Automated Clinical Sentence Classification to Detect Patients' Disease History. J Intell Inf Syst. 2020;55(2):329–49. https://doi.org/10.1007/s10844-020-00605-w.
39. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. CoRR. 2019;abs/1908.10084. http://arxiv.org/abs/1908.10084. Accessed 19 May 2024.
40. Tunstall L, Reimers N, Jo UES, Bates L, Korat D, Wasserblat M, et al. Efficient few-shot learning without prompts. arXiv. http://arxiv.org/abs/2209.11055. Accessed 10 July 2024.
41. Delobelle P, Winters T, Berendt B. RobBERTje: A Distilled Dutch BERT Model. Comput Linguist Neth J. 2021;11:125–40.
42. Beliveau V, Kaas H, Prener M, Ladefoged C, Elliott D, Knudsen GM, et al. Classification of Medical Text in Small and Imbalanced Datasets in a Non-English Language. In: Medical Imaging with Deep Learning. 2024.
43. Devlin J, Chang MW, Lee K, Toutanova K. 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). Minneapolis: Association for Computational Linguistics; 2019. pp. 4171–86. https://doi.org/10.18653/v1/N19-1423.
44. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc.; 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
45. Dong T, Sunderland N, Nightingale A, Fudulu DP, Chan J, Zhai B, et al. Development and Evaluation of a Natural Language Processing System for Curating a Trans-Thoracic Echocardiogram (TTE) Database. Bioengineering. 2023;10(11):1307. https://doi.org/10.3390/bioengineering10111307.
46. Menger V, van Es B, Snackey M. umcu/clinlp: v0.9.0. Zenodo. 2024. https://doi.org/10.5281/zenodo.12706016.

47. Eberts M, Ulges A. Span-Based Joint Entity and Relation Extraction with Transformer Pre-Training. In: European Conference on Artificial Intelligence. 2020.
48. Garcia-Pablos A, Perez N, Cuadros M. Sensitive data detection and classification in Spanish clinical text: experiments with BERT. arXiv. http://arxiv.org/abs/2003.03106. Accessed 17 June 2024.
49. Madabushi HT, Kochkina E, Castelle M. Cost-Sensitive BERT for generalisable sentence classification with imbalanced data. arXiv. http://arxiv.org/abs/2003.11563. Accessed 17 June 2024.
50. Limsopatham N. Effectively Leveraging BERT for Legal Document Classification. In: Proceedings of the Natural Legal Language Processing Workshop 2021. Association for Computational Linguistics; 2021. pp. 210–6. https://doi.org/10.18653/v1/2021.nllp-1.22.
51. Muizelaar H, Haas M, Van Dortmont K, Van Der Putten P, Spruit M. Extracting patient lifestyle characteristics from Dutch clinical text with BERT models. BMC Med Inform Decis Mak. 2024;24(1):151. https://doi.org/10.1186/s12911-024-02557-5.
52. Rietberg MT, Nguyen VB, Geerdink J, Vijlbrief O, Seifert C. Accurate and Reliable Classification of Unstructured Reports on Their Diagnostic Goal Using BERT Models. Diagnostics. 2023;13(7):1251. https://doi.org/10.3390/diagnostics13071251.
53. Yogarajan V, Montiel J, Smith T, Pfahringer B. Transformers for Multi-label Classification of Medical Text: An Empirical Comparison. In: Tucker A, Henriques Abreu P, Cardoso J, Pereira Rodrigues P, Riaño D, editors. Artificial Intelligence in Medicine. vol. 12721. Springer International Publishing; 2021. pp. 114–23. Series Title: Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-030-77211-6_12.
54. Kim Y. Convolutional Neural Networks for Sentence Classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics; 2014. pp. 1746–51. https://doi.org/10.3115/v1/D14-1181.
55. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical Attention Networks for Document Classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics; 2016. pp. 1480–9. https://doi.org/10.18653/v1/N16-1174.

Publisher's Note