frontiers | Frontiers in Radiology

TYPE Original Research
PUBLISHED 24 October 2025
DOI 10.3389/fradi.2025.1672364



OPEN ACCESS

EDITED BY
Yanwu Xu,
Baidu, China

REVIEWED BY
Kwang-Hyun Uhm,
Korea University, Republic of Korea
Kang Liu,
Xidian University, China

*CORRESPONDENCE
Theo Di Piazza
✉ theo.dipiazza@creatis.insa-lyon.fr

RECEIVED 24 July 2025
ACCEPTED 30 September 2025
PUBLISHED 24 October 2025

CITATION
Di Piazza T, Lazarus C, Nempont O and Boussel L (2025) Integrating clinical indications and patient demographics for multilabel abnormality classification and automated report generation in 3D chest CT scans.
Front. Radiol. 5:1672364.
doi: 10.3389/fradi.2025.1672364



# Integrating clinical indications and patient demographics for multilabel abnormality classification and automated report generation in 3D chest CT scans

Theo Di Piazza[1,2]*, Carole Lazarus[3], Olivier Nempont[3] and Loic Boussel[1,2]

[1]UCBL, INSA Lyon, CNRS, INSERM, CREATIS UMR 5220, U1294, Villeurbanne, France, [2]Department of Radiology, Croux-Rousse Hospital, Hospices Civils de Lyon, Lyon, France, [3]Philips Clinical Informatics, Innovation Paris, Paris, France

The increasing number of computed tomography (CT) scan examinations and the time-intensive nature of manual analysis necessitate efficient automated methods to assist radiologists in managing their increasing workload. While deep learning approaches primarily classify abnormalities from three-dimensional (3D) CT images, radiologists also incorporate clinical indications and patient demographics, such as age and sex, for diagnosis. This study aims to enhance multilabel abnormality classification and automated report generation by integrating imaging and non-imaging data. We propose a multimodal deep learning model that combines 3D chest CT scans, clinical information reports, patient age, and sex to improve diagnostic accuracy. Our method extracts visual features from 3D volumes using a visual encoder, textual features from clinical indications via a pretrained language model, and demographic features through a lightweight feedforward neural network. These extracted features are projected into a shared representation space, concatenated, and processed by a projection head to predict abnormalities. For the multilabel classification task, incorporating clinical indications and patient demographics into an existing visual encoder, called CT-Net, improves the F1 score to 51.58, representing a $+\Delta 6.13\%$ increase over CT-Net alone. For the automated report generation task, we extend two existing methods, CT2Rep and CT-AGRG, by integrating clinical indications and demographic data. This integration enhances Clinical Efficacy metrics, yielding an F1 score improvement of $+\Delta 14.78\%$ for the CT2Rep extension and $+\Delta 6.69\%$ for the CT-AGRG extension. Our findings suggest that incorporating patient demographics and clinical information into deep learning frameworks can significantly improve automated CT scan analysis. This approach has the potential to enhance radiological workflows and facilitate more comprehensive and accurate abnormality detection in clinical practice.

KEYWORDS

**abnormality classification, report generation, multimodal, 3D CT scans, clinical indications, patient demographics**

# 1 Introduction

Three-dimensional computed tomography (3D CT) scans have become essential tools in medical imaging [1], offering unparalleled insights into anatomical structures and pathological conditions. This type of medical image is critical for identifying diseases such as pleural effusion [2], lung cancer [3], and cardiomegaly [4]. Given the rapidly growing number of scans to analyze [5] and the increasing demand for specialized radiological expertise in many healthcare systems [6, 7], automating abnormality classification has emerged as an active research area [8–10] to enhance radiologist efficiency. The interpretation of 3D CT scans presents a time-intensive challenge, exacerbated by the heterogeneous nature of observed anomalies. Some anomalies, such as lung nodules [11], can be very small, requiring careful attention from radiologists to avoid missing them. Hence, depending on the patient demographics [12] and clinical indications [13], radiologists may dedicate more time to specific anatomical regions that could potentially present anomalies. As illustrated in Table 1, clinical indications consists of a brief paragraph written by the radiologist before the examination, describing the patient's condition, reason for the visit, and any suspected pathologies that might be revealed during the examination.

Inspired by the workflow of radiologists, we propose a multimodal end-to-end model that integrate clinical indications, patient age, and sex to predict chest pathologies [14]. As shown in Figure 1, our approach extends state-of-the-art methods relying on 3D CT scans to the integration of textual data corresponding to clinical indications, along with utilizing structured data such as patient age and sex. These data have a significant impact on the prevalence of a pathology [15, 16]. We separate feature extraction from each modality using individual modules and then aggregate all these extracted features to predict anomalies. As illustrated in Figure 3, we extend our experimental results by leveraging this multimodal encoder to enhance existing automated report generation methods [17, 18]. Our contributions are as follows:

- We introduce a supervised multimodal method for multilabel classification, capable of taking the 3D CT scan, clinical indications, age, and sex as input.
- We evaluate the model on a public dataset and add an ablation study to demonstrate the importance of each module.
- We extend our experimental results by integrating clinical indications and patient demographics into the automated report generation task.

TABLE 1 Examples of patient demographics (sex and age) and clinical indications from the CT-RATE dataset [10].

| ID | Sex | Age | Clinical indications |
|---|---|---|---|
| Patient 1 | F | 64 | Shortness of breath |
| Patient 2 | F | 42 | Suspicion of lung cancer |
| Patient 3 | M | 50 | Hematological malignancy fever chest pain |
| Patient 4 | M | 37 | Patient with multiple myeloma, focus of infection |

# 2 Related work

## 2.1 Supervised abnormality classification

In the domain of abnormality classification in medical imaging [21], significant research has been conducted on 2D imaging [22, 23] across various modalities such as magnetic resonance imaging (MRI) [24, 25], x-rays [26–29], and skin images [30]. In the field of x-ray imaging, the publicly available MIMIC-CXR dataset [31], comprising 2D radiographs and associated clinical reports, has facilitated the development of various supervised approaches for abnormality detection [32–35] and classification [36]. While some methods focus on a single abnormality or a specific anatomical region [37, 38], others adopt a more comprehensive approach by aiming to simultaneously detect or classify multiple anomalies [39–41] using deep learning models. However, new challenges emerged with 3D imaging and the use of CT or 3D MRI. These modalities introduce novel challenges stemming from the scarcity of publicly available datasets in this domain [10], the high-dimensional nature of the data, and the significant computational demands. Prior work [9, 42] adopted traditional convolutional neural network (CNN) architectures. Recent advances have adopted transformer-based architectures [43] for volumetric data analysis. ViViT [44], an extension of the Vision Transformer [45] originally designed for video understanding [46], has demonstrated strong representational capacity and has since been adapted for a range of CT-based tasks, including radiology report generation [18] and synthetic volume generation [19], with the introduction of CT-ViT [19], which has already demonstrated its effectiveness for various tasks such as report generation [18] and abnormality classification [10].

## 2.2 Multimodal fusion

In machine learning, multimodal fusion [47] has played a pivotal role in advancing classification tasks across various research domains [48–51]. By integrating information from multiple data sources or modalities [52–54], such as combining images from different imaging techniques (e.g., MRI, CT, PET) or fusing imaging data with clinical records [55] or biological information [56], multimodal approaches offer significant advantages. They not only enhance the discriminative capability of classification models but also provide resilience against the inherent variability in single-modal datasets [47, 57, 58]. Feature extraction from each modality is typically performed using a module per modality [59] and then aggregated with a fusion module [60, 61]. The fusion of features across modalities can be achieved through simple concatenation [48], by leveraging self-attention mechanisms [59], or via cross-modality attention modules[62]. Regarding specific work on 3D CT scans, CT2RepLong [18] automatically generates a medical report from the volume and imaging report of the previous medical report of the same patient. This fusion between visual and textual features is achieved through a cross-attention module.

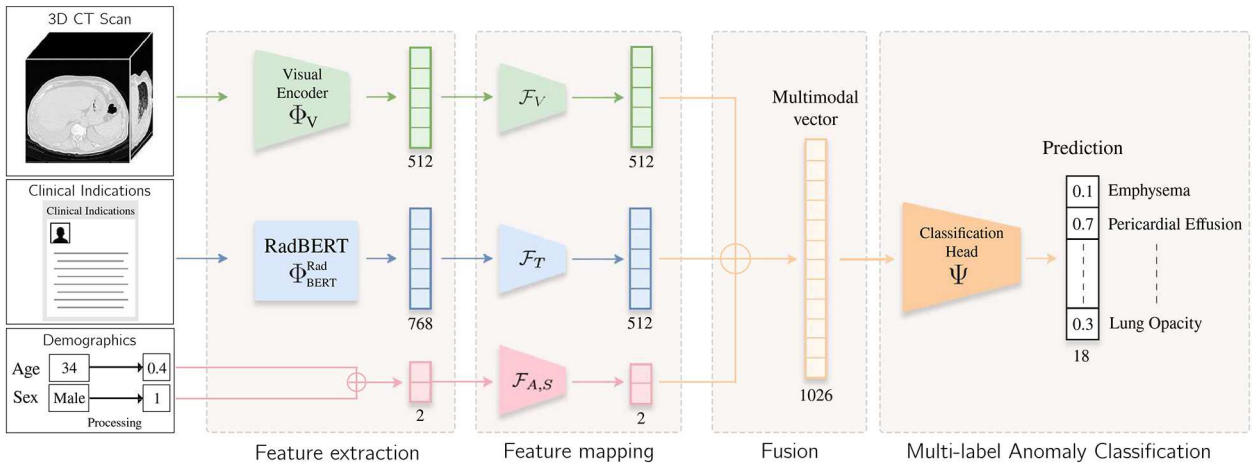


FIGURE 1
Overview of the method. The input volume is processed by a visual extractor $\Phi_V$ [either CT-Net [9] or CT-ViT [19]] and $\mathcal{F}_V$, which generates a visual embedding. Clinical indication is processed by RadBERT [20], yielding a token-level embedding. The [CLS] token is fed into a lightweight MLP $\mathcal{F}_T$ to project textual and visual features into a common latent space. Patient age and sex information are processed by another lightweight MLP $\mathcal{F}_{A,S}$. These vectors are concatenated, and the resulting vector is passed to a classification head $\Psi$, which predicts an abnormality score for each label.

## 2.3 Report generation

Image captioning [63] refers to generating textual descriptions from input images, with significant progress made across various application domains [64–66]. In medical imaging, early generation methods [67] were introduced for 2D modalities using public datasets, such as x-rays [31]. The initial approaches, based on encoder–decoder architectures [68], extract a vector representation using a visual encoder (typically a CNN or attention-based model) and then pass it to a decoder module, often relying on attention mechanisms, to generate the report. Recently, the incorporation of relational memory [69], prior knowledge [70], large language models (LLMs) [71], reinforcement learning [72], and guidance-based methods [73] has enhanced the quality of generated reports. Existing methods for x-ray report generation [74] incorporate medical knowledge or prior information, often in the form of textual modalities, to enhance the quality of the generated reports [75–77]. For 3D CT volumes, the CT-RATE public dataset [10] enabled the development of CT2Rep [18], the first end-to-end method for report generation that extracts vector representations from CT-ViT [19] and passes them to a decoder to generate the report. Similar to 2D imaging, integrating LLMs [78] or multiview encoders [79] has shown improvements in report quality. In the 2D x-ray imaging domain, prior works have explored the integration of clinical indications for report generation. For example, SEI and MLRG employ cross-attention mechanisms to combine indication features with multiview or historical case information [80, 81], while Pragmatic LLaMA introduces indications as additional input to a large language model for guiding report generation [82]. These approaches share with our work the idea of leveraging clinical indications to enrich textual output. However, they are designed for 2D chest radiographs, whereas our method targets volumetric 3D CT scans, which present unique challenges in terms of data dimensionality, abnormality diversity, and multimodal fusion. Regarding guided methods for 3D CT scans, CT-AGRG [17] decomposes the task into two steps: first, a visual encoder performs feature extraction and abnormality classification, and second, a GPT-2 model [83] fine-tuned on a medical corpus [84] generates a description for each detected abnormality. In our work, we extend these approaches by integrating clinical indications and patient demographics into CT2Rep (an end-to-end method) and CT-AGRG (a guided method) to improve performance on the report generation task.

# 3 Dataset

We used the CT-RATE public dataset [10], containing 50,188 reconstructed non-contrast 3D chest CT volumes, to train and evaluate our method. For each scan, we had access to age, sex, and 18 distinct types of abnormalities. The pseudo-labels were extracted from radiology reports using a RadBERT classifier [10, 20]. We only retained samples containing clinical indications, resulting in a dataset comprising 16,009 unique patients (24,085 volumes) for the train set, 792 patients (1,551 volumes) for the validation set, and 792 patients (1,531 volumes) for the test set. We ensured there was no overlap of patients between the training, validation, and test sets. Following Draelos et al. and Hamamci et al. (author?) [9, 10], all volumes were either center-cropped or padded to achieve a resolution of $240 \times 480 \times 480$ with in-slice spacings of 0.75 mm and 1.5 mm on the $z$-axis. Hounsfield unit (HU) [85] values were clipped between $-1,000$ and $+200$. Subsequently, we normalized the clipped HU values to the range $[-1, 1]$ to

facilitate network training. The input age was min–max-normalized [86] to the range [0, 1] to ensure proper handling by the neural network. Sex was encoded as a binary variable, with 0 representing female and 1 representing male. Figure 2 illustrates the distribution of patient age and sex, along with the 18 abnormalities.

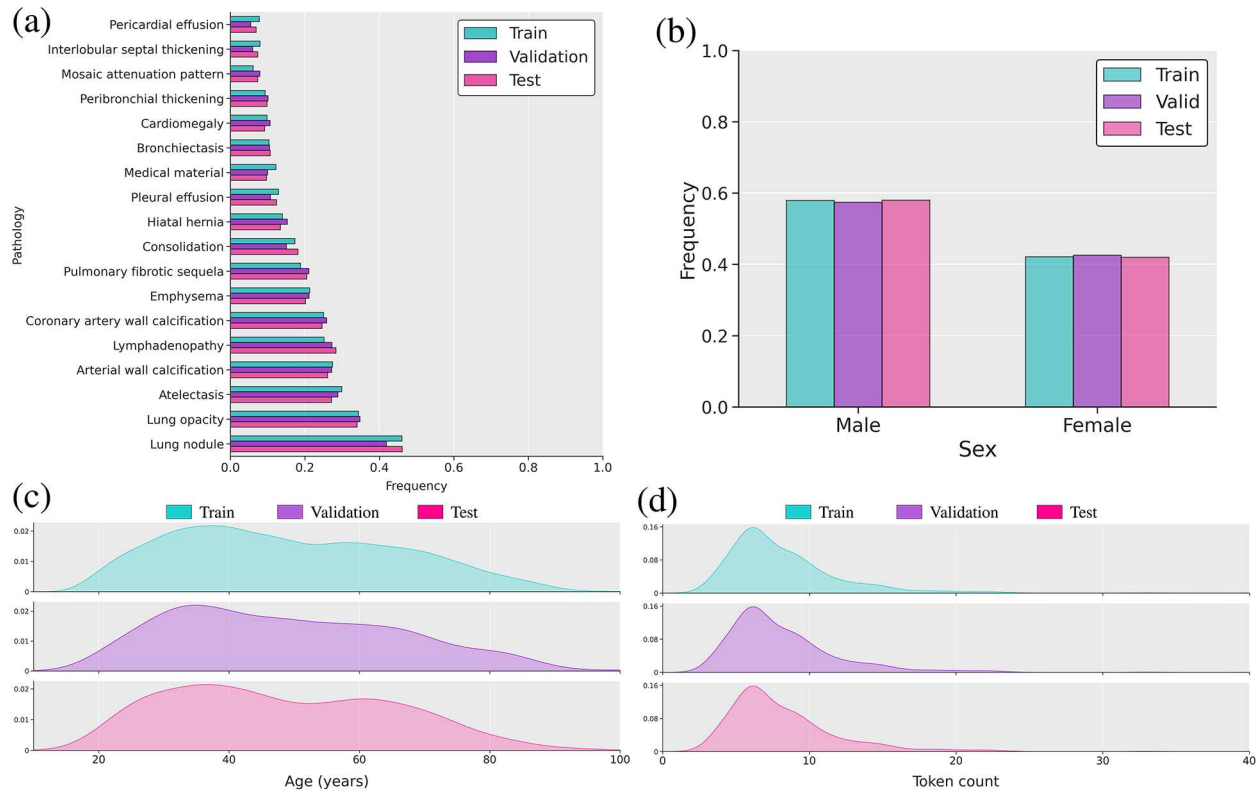


FIGURE 2
Overview of the multimodal dataset. (**a**) Bar plot of label frequency. (**b**) Bar plot of sex frequency. (**c**) Distribution of age in years. (**d**) Distribution plot of reports' lengths based on token count using the RadBERT tokenizer.

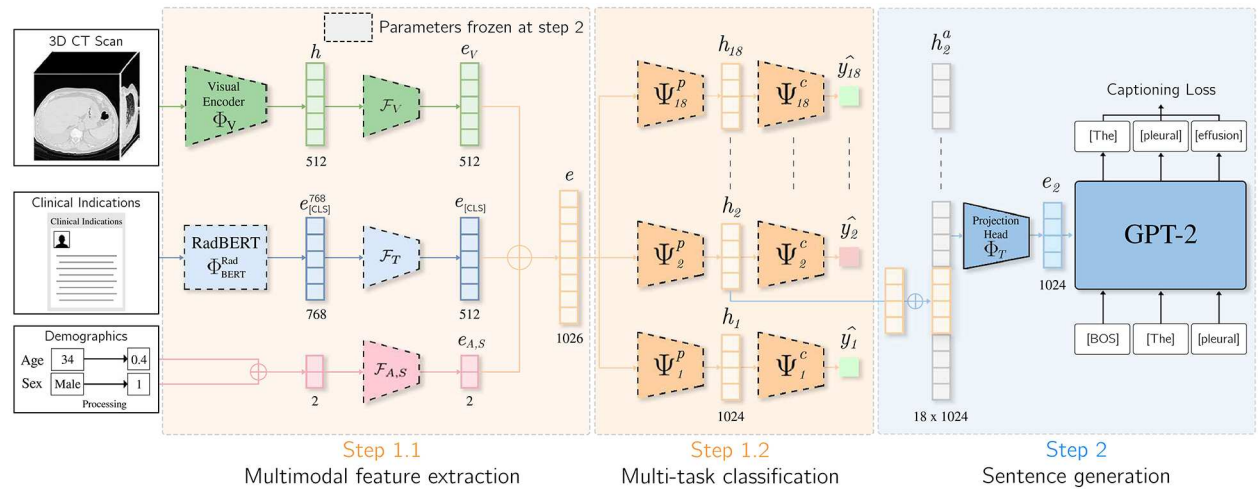


FIGURE 3
Integration of clinical indications and patients demographics for the CT-AGRG method. Features derived from the 3D CT volume, clinical indications, patient age, and sex are aggregated to form vector *e*. This vector is fed into 18 classification heads (one per abnormality). If a classification head predicts an abnormality, the corresponding vector representation is passed to a pretrained GPT-2 model, which generates a textual description of the detected abnormality.

# 4 Methods

As illustrated in Figure 1, our feature extraction module consists of three key components. First, low-level feature extraction is performed independently for each modality, producing modality-specific vector representations. These embeddings are then mapped into a shared feature space using lightweight feedforward networks. Finally, the transformed representations are aggregated via summation to obtain a unified vector representation.

## 4.1 Visual feature extraction

The model receives an input volume $x \in \mathbb{R}^{240\times480\times480}$. This volume is passed to a visual extractor $\Phi_V$, which is either CT-Net [9] or ViViT [19]. To demonstrate the flexibility and generality of our framework across different visual encoders, we conducted experiments using both CT-Net and ViViT. CT-Net consists of 2D ResNet [87] modules followed by a lightweight 3D convolutional network that aggregates the features maps into a compact vector representation [88]. ViViT [44] is a Vision Transformer [45] based on the attention mechanism [43] computed from 3D patches extracted from the initial volume. To ensure a fair evaluation across methods, ViViT is initialized via weight inflation [89] from a 2D ViT [45] pretrained on ImageNet [90], while the 2D ResNet module in CT-Net is directly initialized from a 2D ResNet pretrained on ImageNet. While our contribution focuses on integrating modalities such as clinical indications and demographic information into a visual encoder, we leveraged pretrained weights to facilitate network training, ensuring that model parameters are initialized under comparable conditions. Exploring alternative initialization or pretraining strategies is left for future work. From the initial volume $x$, both CT-Net and CT-ViT yield a vector representation $h \in \mathbb{R}^{512}$. Subsequently, this embedding is passed to a projection head [91] $\mathcal{F}_V$ to obtain $e_V \in \mathbb{R}^{512}$, as defined in Equation 1 such that:

$$e_V = \mathcal{F}_V(h) = (\mathcal{F}_V \circ \Phi_V)(x)\,. \tag{1}$$

## 4.2 Clinical indication feature extraction

To extract embedded tokens from the textual clinical indication report, a pretrained RadBERT [20] model is used. It is a bi-directional neural network, trained on a large radiology report database on a masked language modeling task. From $T$ tokens of the clinical indications report, a single vector representation $e^{768}_{[\mathrm{CLS}]} \in \mathbb{R}^{768}$ is extracted from the Classification [CLS] token [92, 93] outputted by the language model. Working exclusively with the [CLS]-embedded token enables easy projection of textual and visual embeddings into the same-dimensional latent space. Next, $e^{768}_{[\mathrm{CLS}]}$ is passed through a lightweight multilayer perceptron (MLP) $\mathcal{F}_T$ to project the vector representation from textual latent space of dimension 768 to a latent space of dimension 512. The resulting vector $e_{[\mathrm{CLS}]} \in \mathbb{R}^{512}$ is obtained as defined by Equation 2:

$$e_{[\mathrm{CLS}]} = \mathcal{F}_T(e^{768}_{[\mathrm{CLS}]}). \tag{2}$$

## 4.3 Age and sex feature extraction

To handle the normalized age feature $x_A \in [0, 1]$ and the sex feature $x_S \in \{0, 1\}$, a lightweight MLP $\mathcal{F}_{A,S}$, implemented as a linear projection followed by a ReLU activation function, is used to obtain a vector representation $e_{A,S} \in \mathbb{R}^2$, as defined by Equation 3:

$$e_{A,S} = \mathcal{F}_{A,S}(x_A, x_S). \tag{3}$$

## 4.4 Multimodal fusion

The three vector representations associated with different modalities are concatenated [54, 94] into a single vector $e \in \mathbb{R}^{1026}$, such that $e = [e_V,\ e_{[CLS]},\ e_{A,S}]$. A normalization layer [95] is incorporated to ensure stability during training and that the resulting vector $e$ is properly scaled and balanced across its dimensions.

### 4.4.1 Multilabel classification

In the context of abnormality prediction from CT scans, leveraging clinical indications and patient demographics, vector $e$ is given a traditional classification head $\Psi$ to obtain $\hat{y} \in \mathbb{R}^{18}$. As commonly practiced, the model is trained on a multilabel classification task using a binary cross-entropy loss function [96].

### 4.4.2 Report generation

To integrate clinical indications and patient demographics into the report generation task, we extended the CT2Rep [18] and CT-AGRG [17] models by replacing their original visual encoder with our proposed module, which fuses multiple modalities. As illustrated in Figure 3, the decoder responsible for generating the report takes the vector representation $e$ as input. In CT2Rep, the decoder generates the entire report in a single pass from $e$. In contrast, CT-AGRG follows a two-step process: the encoder first predicts the set of abnormalities, and the decoder then generates a detailed description for each predicted abnormality. The models are trained using a next-token prediction objective with binary cross-entropy loss [96]. During inference, the decoder receives only the vector representation $e$ of the input volume and a Beginning Of Sentence [BOS] token to signal the start of the sequence [92]. The report is then generated iteratively, token by token [83].

# 5 Experimental setup

## 5.1 Training details

For the multilabel classification task, the model was trained on 40 epochs on a GPU with 48GB of memory. We used Adam Optimizer [97] with a learning rate of $10^{-4}$ and a batch size of 4. For the report generation experiments, we adopted the same setup as used for CT2Rep [18] and CT-AGRG [17].

## 5.2 Language model

We limited the maximum number of tokens to 40, which is typically found in clinical indication reports [98]. During training, we only fine-tuned the last three layers of RadBERT, with the rest frozen [99].

# 6 Experimental results

This section is organized as follows: we first present quantitative results on the multilabel abnormality classification task with the integration of clinical indications and patient demographics; we then conduct an ablation study to assess the contribution of each module; and finally, we extend our analysis to automatic report generation.

## 6.1 Multilabel classification task

We evaluated the model's performance using commonly used metrics: AUROC, F1 score, precision, recall, and accuracy. We also reported the weighted F1 score, computed by averaging the F1 score of each abnormality, weighted by its occurrence frequency. Because the dataset is dominated by normal findings for most labels (Figure 2), we determined label-specific thresholds on the validation set by maximizing the F1 score [100], as it balances precision and recall [21, 101]. On the test set, we then computed the average of each metric across all labels.

Table 2 demonstrates that incorporating clinical indications and patient demographics significantly improves upon state-of-the-art single-modality methods. Specifically, our model achieves an AUROC of 81.51 ($+\Delta 3.23\%$ over CT-Net) and the highest accuracy of 79.48. However, in an imbalanced multilabel setting, accuracy is primarily driven by correct predictions on abundant classes (especially the normal class) and therefore tends to overestimate overall performance. This also explains why precision (43.93) and recall (65.37) are lower despite high accuracy: even a small number of false positives can markedly reduce precision for rare classes. For this reason, we emphasize the F1 score as a more informative indicator of abnormality detection. Specifically, we achieved an average F1 score of 51.58 [improvements of $+\Delta 6.13\%$ and $+\Delta 16.22\%$ over CT-Net [9] and CT-ViT [10], respectively]. CT-Net with demographics and clinical indications outperforms baseline CT-Net and CT-ViT (paired $t$-test, $p < 0.01$) for all metrics, indicating that incorporating clinical and demographic information enhances classification performance.

Figure 4 details the impact on F1 score for each abnormality when integrating patient demographics and clinical indications, demonstrating that this additional contextual information improves performance for 16 out of 18 anomalies. The largest gains, observed for interlobular septal thickening, consolidation, mosaic attenuation, and lung opacity, suggest that these findings are particularly context-dependent and strongly correlated with clinical factors. While most anomalies benefit from the auxiliary information, a minority, such as bronchiectasis, shows slight performance decreases, possibly because the added inputs may introduce noise for anomalies that already possess distinctive visual signatures. A promising future direction is to develop adaptive integration strategies that selectively incorporate contextual information when it is beneficial.

## 6.2 Ablation study

We conducted a comprehensive ablation study to assess the contributions of the clinical indication feature extractor, each auxiliary input modality, and the fusion module to overall performance.

### 6.2.1 Impact of the clinical indication encoding module

To evaluate the impact of different modules for encoding clinical indications into vector representations, we conducted an ablation study comparing three approaches: a transformer encoder trained from scratch, a BERT language model pretrained on a general corpus, and RadBERT, a BERT-based model pretrained specifically on radiology text. Table 3 and

TABLE 2 Quantitative evaluation of the multilabel classification task on the test set.

| Method | AUROC | Accuracy | F1 score | W. F1 score | Precision | Recall |
|---|---|---|---|---|---|---|
| Random predictions | 49.93 ± 0.51 | 50.11 ± 0.37 | 27.18 ± 0.35 | 33.02 ± 0.39 | 20.24 ± 0.28 | 49.68 ± 0.51 |
| **CT-ViT** [10] | 75.14 ± 0.51 | 73.52 ± 0.57 | 44.38 ± 0.18 | 49.56 ± 0.25 | 35.35 ± 0.51 | 62.42 ± 0.96 |
| **+ clinical ind. + demographics** | 76.09 ± 0.37 | 74.83 ± 0.81 | 45.51 ± 0.24 | 50.97 ± 0.40 | 36.75 ± 0.43 | 63.00 ± 0.98 |
| **CT-Net** [9] | *78.96* ± 0.30 | *78.49* ± 0.55 | *48.60* ± 0.37 | *54.18* ± 0.55 | *42.56* ± 1.01 | *60.15* ± 0.99 |
| **+ clinical ind. + demographics** | **81.51** ± 0.26 | **79.48** ± 0.42 | **51.58** ± 0.54 | **57.60** ± 1.06 | **43.93** ± 0.77 | **65.37** ± 0.88 |

Reported mean and standard deviation metrics were computed over a fivefold cross-validation. The weighted F1 score corresponds to the average of F1 scores for each abnormality, weighted by the frequency of occurrence of the abnormality in the test set. **Best** results are in bold, and *second best* are in italics.

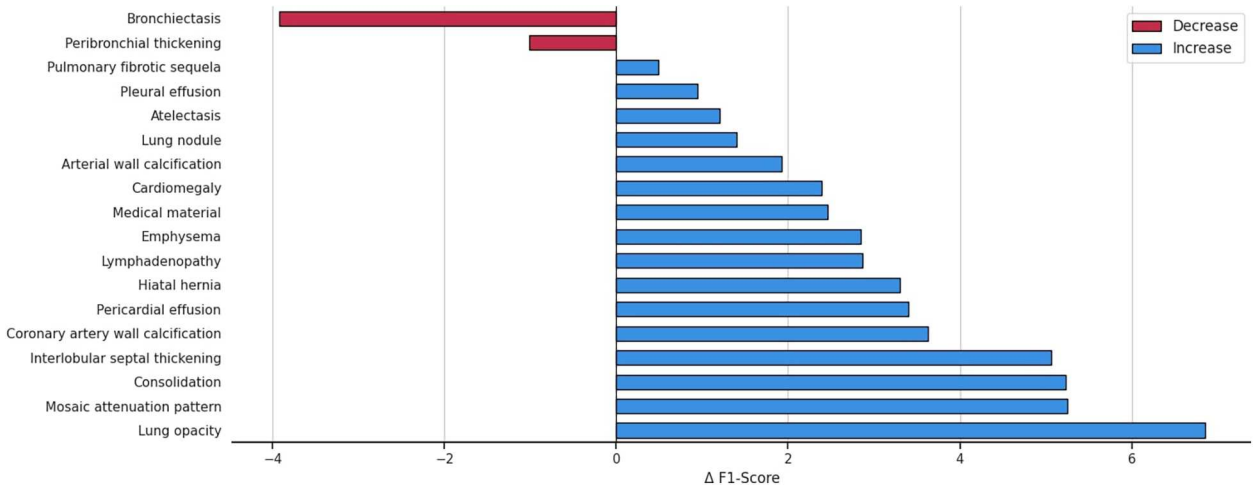


FIGURE 4
Variation in the F1 score across anomalies, highlighting the impact of integrating patient demographics and clinical indications into the multilabel abnormality classification task.

**TABLE 3** Comparative analysis of individual modalities and full integration for multilabel abnormality classification from 3D CT volumes.

| Method | AUROC | Accuracy | F1 score | W. F1 score | Precision |
|---|---|---|---|---|---|
| Random predictions | 49.93 ± 0.51 | 50.11 ± 0.37 | 27.18 ± 0.35 | 33.02 ± 0.39 | 20.24 ± 0.28 |
| Patient demographics | | | | | |
| Age + sex | 62.92 ± 5.46 | 50.83 ± 19.92 | 35.60 ± 4.13 | 43.26 ± 3.69 | 25.71 ± 5.24 |
| Clinical indications | | | | | |
| Transformer encoder [43] | 65.64 ± 0.40 | 64.30 ± 1.13 | 34.87 ± 0.24 | 41.80 ± 0.26 | 26.49 ± 0.63 |
| BERT [92] | 65.96 ± 0.07 | 63.28 ± 0.48 | 35.16 ± 0.29 | 42.00 ± 0.18 | 25.98 ± 0.41 |
| RadBERT [20] | 66.79 ± 0.21 | 65.41 ± 1.07 | 36.38 ± 0.91 | 42.34 ± 0.14 | 27.64 ± 0.64 |
| 3D CT volumes | | | | | |
| CT-Net [9] | *78.96* ± 0.30 | *78.49* ± 0.55 | *48.60* ± 0.37 | *54.18* ± 0.55 | *42.56* ± 1.01 |
| Multimodal fusion | | | | | |
| CT-Net + RadBERT + age + sex | **81.51** ± 0.26 | **79.48** ± 0.42 | **51.58** ± 0.54 | **57.60** ± 1.06 | **43.93** ± 0.77 |

Results are shown for (1) patient demographics only, (2) clinical indications only, (3) visual 3D volumes only, and (4) integration of all inputs.
Bold values show "best results".

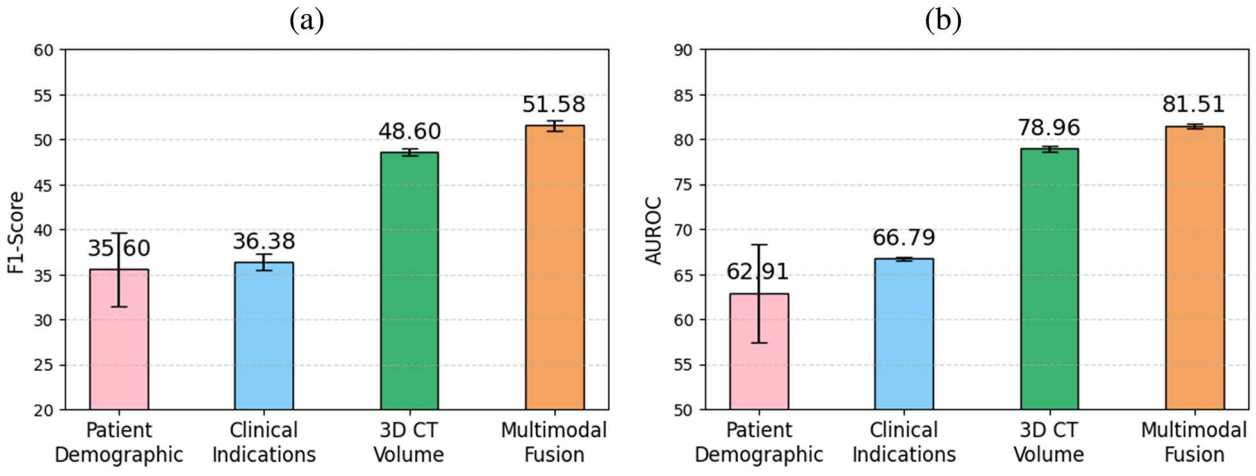


FIGURE 5
Comparison of (a) F1 score and (b) AUROC across clinical indications, patient demographics, 3D CT volume, and multimodal fusion for multilabel abnormality classification. The highest performance is achieved when fusing all modalities, highlighting the benefit of multimodal integration.

Figure 5 report the classification performance achieved when using only clinical indications as input for each of these modules. RadBERT achieved an F1 score of 36.38, representing a $+\Delta 3.47\%$ improvement over general-domain BERT and a $+\Delta 4.33\%$ improvement over the transformer encoder trained from scratch. These results suggest that leveraging a domain-specific pretrained language model facilitates the extraction of more meaningful features from clinical indications, ultimately enhancing classification performance.

### 6.2.2 Impact of auxiliary information

Table 4 presents the ablation study evaluating the incremental impact of incorporating patient demographics and clinical indications as auxiliary inputs alongside the 3D CT Volumes. Adding patient demographics yields an F1 score of 49.79, reflecting a $+\Delta 2.45\%$ improvement over the CT-Net baseline. Incorporating clinical indications results in an F1 score of 50.86, corresponding to a $+\Delta 4.45\%$ gain. For each auxiliary input configuration, a paired $t$-test comparing the F1 score distributions against the baseline yields a $p$-value $< 0.01$, highlighting the statistical significance of the observed performance improvements. Removing CT features led to a consistent drop in performance, indicating that the model does not rely solely on clinical text or metadata.

### 6.2.3 Impact of the fusion module

Our ablation study results related to the fusion module, presented in Table 5, indicate that concatenating features yields the highest AUROC and F1 score increase seen through the integration of clinical indications. Specifically, we obtained an F1 score of 50.86, demonstrating a $+\Delta 0.97$ improvement over sum and a $+\Delta 2.12\%$ improvement over cross-modality attention. This suggests that, in our specific setting, direct concatenation provides a strong signal without the overhead of more complex interaction modeling where the modalities may have relatively low complexity. While more expressive mechanisms such as cross-attention mechanisms demonstrate robust performances in large-scale multimodal learning, we found that in our setting, where the dataset is relatively modest, a simpler fusion provides more robust performance, requiring fewer parameters to fully benefit from modalities effectively. We evaluated an alternative fusion strategy where clinical indications and demographic features are combined into a prompt for the BioMistral LLM [102]. As presented in Table 6, independent concatenation of modality-specific embeddings demonstrates a $+\Delta 0.89\%$ improvement in the F1 score and a $+\Delta 1.05\%$ improvement in AUROC over the LLM-based prompt fusion. We attribute this improvement to the robustness of simpler fusion in the context of our relatively small dataset. While prompt-based fusion offers more expressive modeling, it may require larger datasets to fully realize its benefits, highlighting the importance of matching fusion complexity to dataset scale.

## 6.3 Report generation task

We extend our experiments to the task of automated report generation by integrating clinical information for two methods: CT2Rep [18], which generates the entire report in a single pass, and CT-AGRG [17], which first predicts abnormalities and then generates a description for each detected abnormality. Once the report is generated, we evaluate its quality using two sets of metrics: natural language generation (NLG) metrics and clinical efficacy (CE) metrics [69, 73]. NLG metrics assess the similarity between the generated text and the ground truth. We used BLEU-1 [103], which compares the overlapping 1-grams

TABLE 4 Ablation study on the contribution of auxiliary information for multilabel abnormality classification from 3D CT volumes.

| 3D CT volumes | Patient demographics | Clinical indications | AUROC | F1 score | Paired $t$-test $p$-value | Training time | Inference time |
|---|---|---|---|---|---|---|---|
| ✓ | | | 78.96 ± 0.30 | 48.60 ± 0.37 | – | 15.09 ± 1.55 | 4.16 ± 0.68 |
| ✓ | ✓ | | 80.51 ± 0.49 | 49.79 ± 0.69 | $< 0.01$ | 16.05 ± 1.06 | 4.11 ± 0.32 |
| ✓ | | ✓ | *81.00* ± 0.42 | *50.86* ± 0.36 | $< 0.01$ | 111.30 ± 3.33 | 76.19 ± 0.79 |
| ✓ | ✓ | ✓ | **81.51** ± 0.26 | **51.58** ± 0.54 | $< 0.01$ | 112.96 ± 4.38 | 76.76 ± 0.97 |

We report performance using (1) visual encoder alone, (2) integration of patient demographics, (3) integration of clinical indications, and (4) integration of both. The paired $t$-test $p$-value column reports the statistical significance of the F1 score improvements compared to the baseline using only 3D CT volumes. Training time and inference time indicate the average time per sample (in ms) for forward and backward passes (training) and for inference, respectively.
Bold values show "best results".

TABLE 5 Impact of the aggregation module between features extracted by a visual encoder from the 3D CT volumes and those extracted by RadBERT from clinical indications.

| Method | AUROC | Accuracy | F1 score | W. F1 score | Precision |
|---|---|---|---|---|---|
| Random predictions | 49.93 ± 0.51 | 50.11 ± 0.37 | 27.18 ± 0.35 | 33.02 ± 0.39 | 20.24 ± 0.28 |
| CT-Net [9] | 78.96 ± 0.30 | 78.49 ± 0.55 | 48.60 ± 0.37 | 54.18 ± 0.55 | 42.56 ± 1.01 |
| Methods below utilize clinical indications | | | | | |
| With self-attention | 80.35 ± 0.18 | 78.67 ± 0.78 | 49.21 ± 0.57 | 55.35 ± 0.48 | 42.23 ± 0.59 |
| With cross-attention | 80.36 ± 0.41 | 78.11 ± 1.02 | 49.84 ± 0.31 | 55.44 ± 0.27 | 42.29 ± 0.91 |
| With sum | *80.99* ± 0.28 | *78.89* ± 0.65 | *50.37* ± 0.50 | *56.08* ± 0.36 | *43.05* ± 0.25 |
| With concatenation | **81.00** ± 0.42 | **79.80** ± 0.37 | **50.86** ± 0.36 | **56.58** ± 0.20 | **43.84** ± 0.32 |

Bold values show "best results".

between the reference and the prediction. ROUGE evaluates recall-oriented metrics, like overlap and precision, between *n*-grams. BERTScore [104], an embedding-based metric, measures the cosine similarity of vector representations of the embedded tokens between the reference and the generated text. Clinical efficacy metrics evaluate the clinical accuracy of generated reports. We extracted abnormality mentions as one-hot vectors using a RadBERT language model classifier [20], which was originally used for CT-RATE [10] label annotation. These predictions are then compared to ground-truth labels using standard multilabel classification metrics, such as the F1 score. In addition, we report the CRG score [105], a recently proposed distribution-aware metric for radiology report generation. Unlike conventional metrics, CRG focuses exclusively on clinically relevant abnormalities explicitly described in the reference report, while also accounting for class imbalance.

As shown in Table 7, the integration of clinical indications and patient demographics significantly enhances both NLG and CE metrics. For the CT2Rep model, incorporating these additional data results in a BLEU-1 score of 0.342, reflecting a $+\Delta 7.55\%$ increase compared to the baseline model, and an F1 score of 36.57, which corresponds to a $+\Delta 14.78\%$ improvement over the baseline. Similarly, for the CT-AGRG guided method, the inclusion of clinical indication and patient demographic information leads to a performance boost, achieving a recall of 60.43 ($+\Delta 12.32\%$ increase) and an F1 score of 48.30 ($+\Delta 6.69\%$ increase) over the original model. For each method, we performed a paired *t*-test comparing the F1 score obtained with and without the integration of clinical indications and patient demographics. The resulting *p*-values are all strictly below 0.01, indicating statistically significant improvements in the quality of the generated reports. In addition, Figure 6 allows us to identify which anomalies benefit most from richer multimodal inputs, making performance gains more clinically interpretable and highlighting where report generation is most reliable. Figure 7 illustrates two examples of report generation compared to the ground truth, emphasizing that our

TABLE 6 Comparison of fusion strategies for incorporating clinical indications and demographic information in 3D chest CT abnormality classification.

| Demographics and indications | Fusion strategy | AUROC | Accuracy | F1 score | W. F1 score | Recall |
|---|---|---|---|---|---|---|
| ✗ | | 78.96 ± 0.30 | 78.49 ± 0.55 | 48.60 ± 0.37 | 54.18 ± 0.55 | 60.15 ± 1.39 |
| ✓ | Prompt-based | 80.66 ± 0.45 | *79.68* ± 0.47 | 51.12 ± 0.51 | 56.78 ± 0.37 | 63.50 ± 1.28 |
| ✓ | Modality-specific | *81.51* ± 0.26 | 79.48 ± 0.42 | *51.58* ± 0.54 | *57.60* ± 1.06 | *65.37* ± 1.53 |

We evaluated modality-specific embeddings for clinical indications, patient age, and sex, concatenated with CT features, against prompt-based fusion, where the same information is integrated into a BioMistral-7B input prompt. *Best results* are underlined.

TABLE 7 Quantitative evaluation of the report generation task.

| Method | NLG metrics | | | CE metrics | | |
|---|---|---|---|---|---|---|
| | BLEU-1 | $ROUGE_L$ | BERT | CRG | Recall | F1 score |
| Random predictions | – | – | – | 0.397 ± 0.004 | 50.92 ± 1.48 | 27.18 ± 0.35 |
| CT2Rep [18] | 0.318 ± 0.007 | 0.236 ± 0.006 | 0.863 ± 0.002 | 0.417 ± 0.009 | 32.56 ± 2.62 | 31.86 ± 1.74 |
| **+ clinical indications + demographics** | 0.342 ± 0.006 | 0.259 ± 0.003 | **0.871** ± 0.001 | 0.430 ± 0.004 | 36.22 ± 0.81 | 36.57 ± 0.73 |
| CT-AGRG [17] | *0.386* ± 0.011 | *0.265* ± 0.001 | 0.863 ± 0.001 | *0.488* ± 0.004 | *53.80* ± 0.05 | *45.27* ± 0.19 |
| **+ clinical indications + demographics** | **0.395** ± 0.006 | **0.268** ± 0.003 | 0.867 ± 0.001 | **0.509** ± 0.001 | **60.43** ± 0.11 | **48.30** ± 0.05 |

We used NLG metrics and CE metrics with CRG, recall, and F1 score.
Bold values shows "best results".

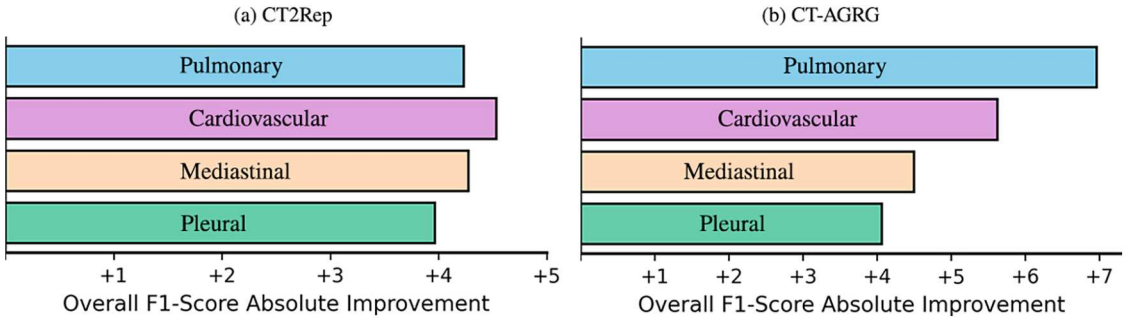


FIGURE 6
F1 score improvements across four abnormality groups, when incorporating demographic and clinical indication information into 3D CT volume report generation.

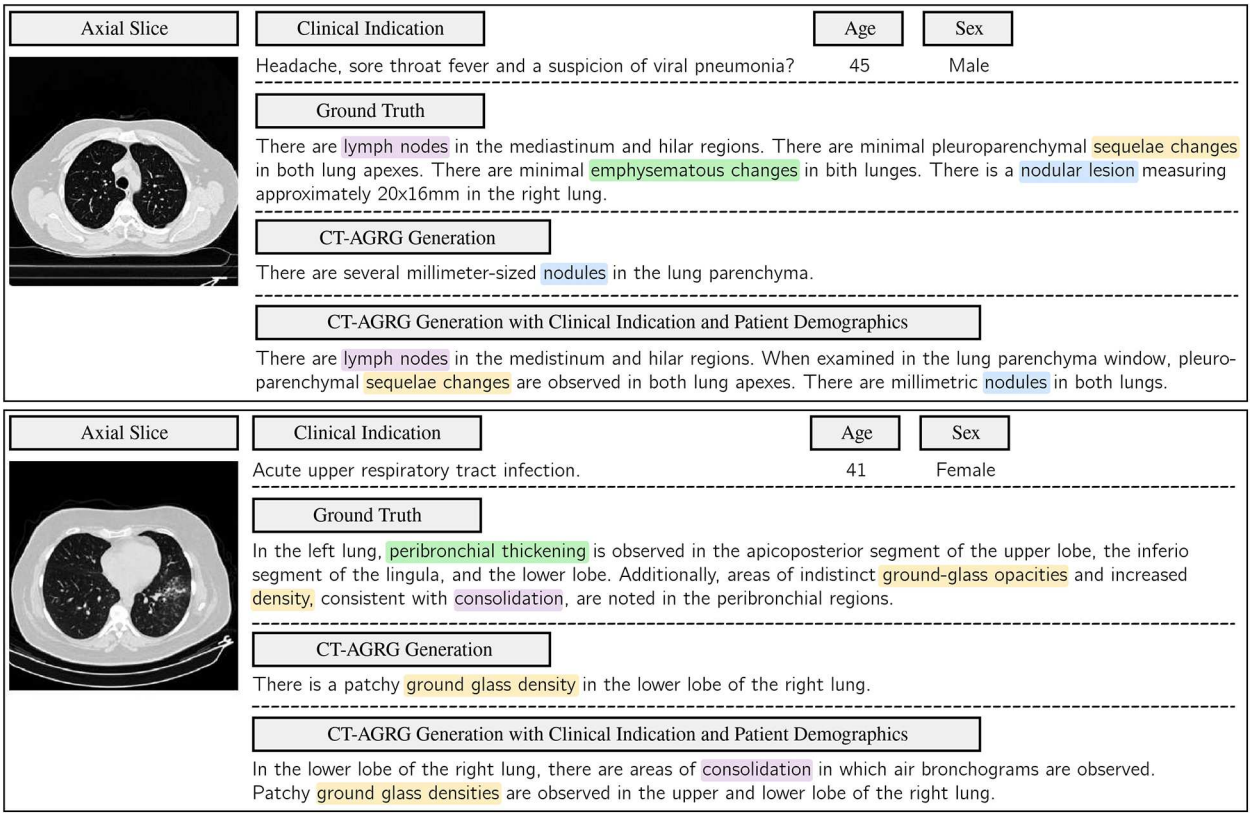


FIGURE 7
Comparison of ground-truth labels with the report generated by the CT-AGRG model with and without the integration of clinical indications and patient demographics. For each of the two CT-RATE test set examples, we present an axial slice, clinical indications, demographic information, ground truth, and the generated report. Clinical relevance is highlighted using color-coded annotations.

method produces reports with a structure and terminology closely resembling those written by radiologists.

## 7 Conclusion and discussion

In this paper, we present a simple and effective method capable of integrating various sources of information to classify multiple anomalies from chest 3D CT scans, available clinical indications, and age and sex features. We also integrate these information sources for report generation, demonstrating their ability to enhance model performance across various tasks related to 3D CT scans. Furthermore, our experiments validate the effectiveness of each module and the use of a pretrained language model for clinical indication feature extraction. Due to the limited availability of multimodal publicly accessible 3D chest CT datasets, our findings are based solely on the CT-RATE dataset. While this provides a solid foundation for initial validation, reliance on a single dataset may introduce biases related to language patterns, labeling conventions, or demographic representations. Moreover, the demographic features considered in this study (age and sex) remain limited. Future work should therefore aim to include external validation on independent datasets and explore richer metadata to better assess model generalizability and robustness. To enhance multimodal representation of a patient, future work could incorporate additional modalities, such as longitudinal patient data, richer demographic features, or similarity-based retrieval of reports and volumes, to further strengthen multimodal fusion.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author/s.

## Ethics statement

The studies involving humans obtained ethical approval from the Clinical Research Ethics Committee at Istanbul Medipol University (E-10840098-772.02-6841, 27/10/2023) for open-sourcing the CT-RATE dataset. Please refer to the publicly available CT-RATE dataset released by the University of Zurich with Istanbul Medipol University. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was

not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

## Author contributions

TDP: Writing – original draft, Writing – review & editing. CL: Writing – review & editing. ON: Writing – review & editing. LB: Writing – review & editing.

## Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note



## References

1. Singh SP, Wang L, Gupta S, Goli H, Padmanabhan P, Gulyás B. 3D deep learning on medical images: a review. *Sensors (Basel)*. (2020) 20:5097. doi: 10.3390/s20185097

2. Jany B, Welte T. Pleural effusion in adults—etiology, diagnosis, and treatment. *Deutsches Ärzteblatt Int*. (2019) 116:377–86. doi: 10.3238/arztebl.2019.0377

3. Dela Cruz CS, Tanoue LT, Matthay RA. Lung cancer: epidemiology, etiology, and prevention. *Clin Chest Med*. (2011) 32:605–44. doi: 10.1016/j.ccm.2011.09.001

4. Amin H, Siddiqui WJ. Cardiomegaly. In: *StatPearls*. Treasure Island, FL: StatPearls Publishing (2024). Available online at: https://www.ncbi.nlm.nih.gov/books/NBK542296/ (Accessed June 25, 2024).

5. Goergen SK, Pool FJ, Turner TJ, Grimm JE, Appleyard MN, Crock C, et al. Evidence-based guideline for the written radiology report: methods, recommendations and implementation challenges. *J Med Imaging Radiat Oncol*. (2013) 57:1–7. doi: 10.1111/jmiro.2013.57.issue-1

6. Bastawrous S, Carney B. Improving patient safety: avoiding unread imaging exams in the national VA enterprise electronic health record. *J Digit Imaging*. (2017) 30:309–13. doi: 10.1007/s10278-016-9937-2

7. Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ (Clin Res Ed.)*. (2017) 359:j4683. doi: 10.1136/bmj.j4683

8. Djahnine A, Jupin-Delevaux E, Nempont O, Si-Mohamed SA, Craighero F, Cottin V, et al. Weakly-supervised learning-based pathology detection and localization in 3D chest CT scans. *Med Phys*. (2024) 51:8272–82. doi: 10.1002/mp.v51.11

9. Draelos RL, Dov D, Mazurowski MA, Lo JY, Henao R, Rubin GD, et al. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Med Image Anal*. (2021) 67:101857. doi: 10.1016/j.media.2020.101857

10. Hamamci IE, Er S, Almas F, Simsek AG, Esirgun SN, Dogan I, et al. Data from: A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities (2024). doi: 10.48550/arXiv.2403.17834

11. Shen W, Zhou M, Yang F, Yang C, Tian J. Multi-scale convolutional neural networks for lung nodule classification. *Inf Process Med Imaging: Proc Conf*. (2015) 24:588–99. doi: 10.1007/978-3-319-19992-4_46

12. Raveh D, Gratch L, Yinnon AM, Sonnenblick M. Demographic and clinical characteristics of patients admitted to medical departments. *J Eval Clin Pract*. (2005) 11:33–44. doi: 10.1111/jep.2005.11.issue-1

13. Hattori S, Yokota H, Takada T, Horikoshi T, Takishima H, Mikami W, et al. Impact of clinical information on CT diagnosis by radiologist and subsequent clinical management by physician in acute abdominal pain. *Eur Radiol*. (2021) 31:5454–63. doi: 10.1007/s00330-021-07700-8

14. Di Piazza T, Lazarus C, Nempont O, Boussel L. *Leveraging Clinical Indications and Demographics to Improve Multi-Label Abnormality Classification in 3D Chest CT Scans*. San Diego, CA: SPIE (2025).

15. Noordzij M, Dekker FW, Zoccali C, Jager KJ. Measures of disease frequency: prevalence and incidence. *Nephron Clin Pract*. (2010) 115:c17–20. doi: 10.1159/000286345

16. Tenny S, Hoffman MR. Prevalence. In: *StatPearls*. Treasure Island, FL: StatPearls Publishing (2024). Available online at: https://www.ncbi.nlm.nih.gov/books/NBK430867/ (Accessed July 31, 2024).

17. Di Piazza T, Lazarus C, Nempont O, Boussel L. CT-AGRG: Automated abnormality-guided report generation from 3D chest CT volumes. In: *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*. Houston, TX: IEEE (2025). p. 1–5.

18. Hamamci IE, Er S, Menze B. Data from: CT2Rep: automated radiology report generation for 3D medical imaging (2024). arXiv:2403.06801 [cs, eess].

19. Hamamci IE, Er S, Simsar E, Sekuboyina A, Prabhakar C, Tezcan A, et al. Data from: GenerateCT: text-conditional generation of 3D chest CT volumes (2023). arXiv:2305.16037 [cs].

20. Yan A, McAuley J, Lu X, Du J, Chang EY, Gentili A, et al. RadBERT: adapting transformer-based language models to radiology. *Radiol Artif Intell*. (2022) 4:e210258. doi: 10.1148/ryai.210258

21. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. (2017) 42:60–88. doi: 10.1016/j.media.2017.07.005

22. Aljuaid A, Anwar M. Survey of supervised learning for medical image processing. *SN Comput Sci*. (2022) 3:292. doi: 10.1007/s42979-022-01166-1

23. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVF, editor. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu: IEEE (2017). p. 3462–71.

24. Tang Y-X, Tang Y-B, Peng Y, Yan K, Bagheri M, Redd BA, et al. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *npj Digit Med*. (2020) 3:1–8. doi: 10.1038/s41746-020-0273-z

25. Zhou J, Luo L, Dou Q, Chen H, Chen C, Li G, et al. Weakly supervised 3D deep learning for breast cancer classification and localization of the lesions in MR images. *J Magn Reson Imaging*. (2019) 50:1144–51. doi: 10.1002/jmri.v50.4

26. Hamamci IE, Er S, Simsar E, Yuksel AE, Gultekin S, Ozdemir SD, et al. Data from: DENTEX: an abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic x-rays (2023). doi: 10.48550/arXiv.2305.19112

27. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. Data from: CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison (2019). doi: 10.48550/arXiv.1901.07031

28. Nguyen HQ, Lam K, Le LT, Pham HH, Tran DQ, Nguyen DB, et al. Data from: VinDr-CXR: an open dataset of chest x-rays with radiologist's annotations (2022). doi: 10.48550/arXiv.2012.15029

29. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. Data from: CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning (2017). doi: 10.48550/arXiv.1711.05225

30. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. (2017) 542:115–8. doi: 10.1038/nature21056

31. Johnson AEW, Pollard T, Mark R, Berkowitz S, Horng S. Data from: The MIMIC-CXR database (2019). doi: 10.13026/C2JT1Q

32. Dubey AK, Young MT, Stanley C, Lunga D, Hinkle J. Data from: Computer-aided abnormality detection in chest radiographs in a clinical setting via domain-adaptation (2020). doi: 10.48550/arXiv.2012.10564

33. Li J, Fu G, Chen Y, Li P, Liu B, Pei Y, et al. A multi-label classification model for full slice brain computerised tomography image. *BMC Bioinform*. (2020) 21:200. doi: 10.1186/s12859-020-3503-0

34. Yu K, Ghosh S, Liu Z, Deible C, Batmanghelich K. Anatomy-guided weakly-supervised abnormality localization in chest x-rays. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S, editors. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Cham, Switzerland: Springer Nature (2022). p. 658–68.

35. Zhao X, Wang X, Xia W, Zhang R, Jian J, Zhang J, et al. 3D multi-scale, multi-task, and multi-label deep learning for prediction of lymph node metastasis in T1 lung adenocarcinoma patients' CT images. *Comput Med Imaging Graph*. (2021) 93:101987. doi: 10.1016/j.compmedimag.2021.101987

36. Pooch EHP, Ballester P, Barros RC. Can we trust deep learning based diagnosis? The impact of domain shift in chest radiograph classification. In: Petersen J, San José Estépar R, Schmidt-Richberg A, Gerard S, Lassen-Schmidt B, Jacobs C, Beichel R, Mori K, editors. *Thoracic Image Analysis*. Cham, Switzerland: Springer International Publishing (2020). p. 74–83.

37. Ausawalaithong W, Marukatat S, Thirach A, Wilaiprasitporn T. Data from: Automatic lung cancer prediction from chest x-ray images using deep learning approach (2018). arXiv: 1808.10858 [eess].

38. Jaiswal AK, Tiwari P, Kumar S, Gupta D, Khanna A, Rodrigues JJPC. Identifying pneumonia in chest x-rays: a deep learning approach. *Measurement*. (2019) 145:511–8. doi: 10.1016/j.measurement.2019.05.076

39. Albahli S, Rauf HT, Algosaibi A, Balas VE. AI-driven deep CNN approach for multi-label pathology classification using chest x-rays. *PeerJ Comput Sci*. (2021) 7: e495. doi: 10.7717/peerj-cs.495

40. Ge Z, Mahapatra D, Sedai S, Garnavi R, Chakravorty R. Data from: Chest x-rays classification: a multi-label and fine-grained problem (2018). doi: 10.48550/arXiv.1807.07247

41. Ibrahim DM, Elshennawy NM, Sarhan AM. Deep-chest: multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases. *Comput Biol Med*. (2021) 132:104348. doi: 10.1016/j.compbiomed.2021.104348

42. Kim K, Oh SJ, Lee JH, Chung MJ. 3D unsupervised anomaly detection through virtual multi-view projection and reconstruction: clinical validation on low-dose chest computed tomography. *Expert Syst Appl*. (2024) 236:121165. doi: 10.1016/j.eswa.2023.121165

43. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Data from: Attention is all you need (2023). doi: 10.48550/arXiv.1706.03762

44. Arnab A, Dehghani M, Heigold G, Sun C, Lucic M, Schmid C. ViViT: a video vision transformer. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE (2021). p. 6816–26.

45. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. Data from: An image is worth $16 \times 16$ words: transformers for image recognition at scale (2021). doi: 10.48550/arXiv.2010.11929

46. Villegas R, Babaeizadeh M, Kindermans P-J, Moraldo H, Zhang H, Saffar MT, et al. Data from: Phenaki: variable length video generation from open domain textual description (2022). arXiv: 2210.02399 [cs].

47. Baltrusaitis T, Ahuja C, Morency L-P. Multimodal machine learning: a survey and taxonomy. *IEEE Trans Pattern Anal Mach Intell*. (2019) 41:423–43. doi: 10.1109/TPAMI.2018.2798607

48. Huang S-C, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digit Med*. (2020) 3:136. doi: 10.1038/s41746-020-00341-z

49. Lee SI, Yoo SJ. Multimodal deep learning for finance: integrating and forecasting international stock markets. *J Supercomput*. (2020) 76:8294–312. doi: 10.1007/s11227-019-03101-3

50. Lichtenwalter D, Burggräf P, Wagner J, Weißer T. Deep multimodal learning for manufacturing problem solving. *Procedia CIRP*. (2021) 99:615–20. doi: 10.1016/j.procir.2021.03.083

51. Sarraf S, Noori M. *Multimodal Deep Learning Approach for Event Detection in Sports Using Amazon SageMaker | AWS Machine Learning Blog* (2021).

52. Antropova N, Huynh BQ, Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Med Phys*. (2017) 44:5162–71. doi: 10.1002/mp.2017.44.issue-10

53. Suk H-I, Lee S-W, Shen D. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage*. (2014) 101:569–82. doi: 10.1016/j.neuroimage.2014.06.077

54. Tan K, Huang W, Liu X, Hu J, Dong S. A multi-modal fusion framework based on multi-task correlation learning for cancer prognosis prediction. *Artif Intell Med*. (2022) 126:102260. doi: 10.1016/j.artmed.2022.102260

55. Hsieh C, Nobre IB, Sousa SC, Ouyang C, Brereton M, Nascimento JC, et al. MDF-Net for abnormality detection by fusing x-rays with clinical data. *Sci Rep*. (2023) 13:15873. doi: 10.1038/s41598-023-41463-0

56. Joo Y, Namgung E, Jeong H, Kang I, Kim J, Oh S, et al. Brain age prediction using combined deep convolutional neural network and multi-layer perceptron algorithms. *Sci Rep*. (2023) 13:22388. doi: 10.1038/s41598-023-49514-2

57. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. Multimodal deep learning. In: *Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11*. Madison, WI: Omnipress (2011). p. 689–96.

58. Wang W, Tran D, Feiszli M. Data from: What makes training multi-modal classification networks hard? (2020). doi: 10.48550/arXiv.1905.12681

59. Costanzino A, Ramirez PZ, Lisanti G, Di Stefano L. Data from: Multimodal industrial anomaly detection by crossmodal feature mapping (2023). doi: 10.48550/arXiv.2312.04521

60. Wei X, Zhang T, Li Y, Zhang Y, Wu F. Multi-modality cross attention network for image and sentence matching. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA: IEEE (2020). p. 10938–47.

61. Zhu Q, Mathai TS, Mukherjee P, Peng Y, Summers RM, Lu Z. Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports (2023). arXiv:2306.08749v2.

62. Deng R, Shaikh N, Shannon G, Nie Y. Data from: Cross-modality attention-based multimodal fusion for non-small cell lung cancer (NSCLC) patient survival prediction (2024). arXiv: 2308.09831 [eess].

63. Mathur P. A survey on various deep learning models for automatic image captioning. *J Phys: Conf Ser*. (2021) 1950:012045. doi: 10.1088/1742-6596/1950/1/012045

64. Gurari D, Li Q, Stangl AJ, Guo A, Lin C, Grauman K, et al. Data from: VizWiz grand challenge: answering visual questions from blind people (2018). doi: 10.48550/arXiv.1802.08218

65. Vinyals O, Toshev A, Bengio S, Erhan D. Data from: Show and tell: a neural image caption generator (2015). doi: 10.48550/arXiv.1411.4555.

66. Xue H, Zhang C, Liu C, Wu F, Jin X. Data from: Multi-task prompt words learning for social media content generation (2024). doi: 10.48550/arXiv.2407.07771

67. Kougia V, Pavlopoulos J, Androutsopoulos I. Data from: A survey on biomedical image captioning (2019). arXiv:1905.13302 [cs] version: 1.

68. Jing B, Xie P, Xing E. On the automatic generation of medical imaging reports. In: Gurevych I, Miyao Y, editors. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics (2018). p. 2577–86.

69. Chen Z, Song Y, Chang T-H, Wan X. Data from: Generating radiology reports via memory-driven transformer. *arXiv. arXiv:2010.16056 [cs]* (2022). doi: 10.48550/arXiv.2010.16056

70. Liu F, Wu X, Ge S, Fan W, Zou Y. Data from: Exploring and distilling posterior and prior knowledge for radiology report generation (2021). doi: 10.48550/arXiv.2106.06963

71. Wang Z, Liu L, Wang L, Zhou L. Data from: R2GenGPT: radiology report generation with frozen LLMs (2023). doi: 10.48550/arXiv.2309.09812

72. Han W, Kim C, Ju D, Shim Y, Hwang SJ. Data from: Advancing text-driven chest x-ray generation with policy-based reinforcement learning (2024). doi: 10.48550/arXiv.2403.06516

73. Tanida T, Müller P, Kaissis G, Rueckert D. Interactive and explainable region-guided radiology report generation. In: CVF, editor. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC: IEEE (2023). p. 7433–42.

74. Sloan P, Clatworthy P, Simpson E, Mirmehdi M. Automated radiology report generation: a review of recent advances. *IEEE Rev Biomed Eng*. (2025) 18:368–87. doi: 10.1109/RBME.2024.3408456

75. Gajbhiye GO, Nandedkar AV, Faye I. Translating medical image to radiological report: adaptive multilevel multi-attention approach. *Comput Methods Programs Biomed*. (2022) 221:106853. doi: 10.1016/j.cmpb.2022.106853

76. Li J, Li S, Hu Y, Tao H. Data from: A self-guided framework for radiology report generation (2022). doi: 10.48550/arXiv.2206.09378

77. Nishino T, Miura Y, Taniguchi T, Ohkuma T, Suzuki Y, Kido S, et al. Factual accuracy is not enough: planning consistent description order for radiology

report generation. In: Goldberg Y, Kozareva Z, Zhang Y, editors. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics (2022). p. 7123–38.

78. Liu C, Wan Z, Wang Y, Shen H, Wang H, Zheng K, et al. Data from: Benchmarking and boosting radiology report generation for 3D high-resolution medical images (2024). doi: 10.48550/arXiv.2406.07146

79. Deng X, He X, Bao J, Zhou Y, Cai S, Cai C, et al. Data from: MvKeTR: chest CT report generation with multi-view perception and knowledge enhancement (2025). doi: 10.48550/arXiv.2411.18309

80. Liu K, Ma Z, Kang X, Li Y, Xie K, Jiao Z, et al. Data from: Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation (2025). arXiv: 2502.20056 [cs].

81. Liu K, Ma Z, Kang X, Zhong Z, Jiao Z, Baird G, et al. Data from: Structural entities extraction and patient indications incorporation for chest x-ray report generation (2024). arXiv: 2405.14905 [eess].

82. Nguyen D, Chen C, He H, Tan C. Data from: Pragmatic radiology report generation (2023). arXiv: 2311.17154 [cs].

83. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models Are Unsupervised Multitask Learners (2019).

84. Papanikolaou Y, Pierleoni A. Data from: DARE: data augmented relation extraction with GPT-2 (2020). doi: 10.48550/arXiv.2004.13845

85. DenOtter TD, Schubert J. Hounsfield unit. In: *StatPearls*. Treasure Island, FL: StatPearls Publishing (2024). Available online at: https://www.ncbi.nlm.nih.gov/books/NBK547721/ (Accessed June 06, 2024).

86. Patro SK, Sahu KK. Normalization: a preprocessing stage. *IARJSET*. (2015) 20–2. doi: 10.17148/IARJSET

87. He K, Zhang X, Ren S, Sun J. Data from: Deep residual learning for image recognition (2015). doi: 10.48550/arXiv.1512.03385

88. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Data from: Learning spatiotemporal features with 3D convolutional networks (2015). doi: 10.48550/arXiv.1412.0767.

89. Zhang Y, Huang S-C, Zhou Z, Lungren MP, Yeung S. Adapting pre-trained vision transformers from 2D to 3D through weight inflation improves medical image segmentation (2023). arXiv:2302.04303 [cs].

90. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: CVF, editor. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE. (2009). p. 248–55.

91. Chen T, Kornblith S, Norouzi M, Hinton G. Data from: A simple framework for contrastive learning of visual representations (2020). doi: 10.48550/arXiv.2002.05709

92. Devlin J, Chang M-W, Lee K, Toutanova K. Data from: BERT: pre-training of deep bidirectional transformers for language understanding (2019). doi: 10.48550/arXiv.1810.04805

93. Sun C, Qiu X, Xu Y, Huang X. Data from: How to fine-tune BERT for text classification? (2020). doi: 10.48550/arXiv.1905.05583

94. Gao J, Li P, Chen Z, Zhang J. A survey on deep learning for multimodal data fusion. *Neural Comput*. (2020) 32:829–64. doi: 10.1162/neco_a_01273

95. Ba JL, Kiros JR, Hinton GE. Data from: Layer normalization (2016). doi: 10.48550/arXiv.1607.06450

96. Good IJ. Rational decisions. *J R Stat Soc Ser B (Methodol)*. (1952) 14:107–14. doi: 10.1111/j.2517-6161.1952.tb00104.x

97. Kingma DP, Ba J. Data from: Adam: A method for stochastic optimization (2017). doi: 10.48550/arXiv.1412.6980

98. Levy M, Jacoby A, Goldberg Y. Data from: Same task, more tokens: the impact of input length on the reasoning performance of large language models (2024). doi: 10.48550/arXiv.2402.14848

99. Lee J, Tang R, Lin JJ. What would Elsa do? freezing layers during transformer fine-tuning. *arXiv* [Preprint]. *arXiv:1911.03090* [cs.CL] (2019). doi: 10.48550/arXiv.1911.03090

100. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. *Sci Rep*. (2024) 14:6086. doi: 10.1038/s41598-024-56706-x

101. Powers DMW. Data from: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation (2020). doi: 10.48550/arXiv.2010.16061.

102. Labrak Y, Bazoge A, Morin E, Gourraud P-A, Rouvier M, Dufour R. Data from: BioMistral: a collection of open-source pretrained large language models for medical domains (2024). arXiv: 2402.10373 [cs].

103. Papineni K, Roukos S, Ward T, Zhu W-J. Bleu: a method for automatic evaluation of machine translation. In: Isabelle P, Charniak E, Lin D, editors. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA: Association for Computational Linguistics (2002). p. 311–8.

104. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. Data from: BERTScore: evaluating text generation with BERT (2020). doi: 10.48550/arXiv.1904.09675

105. Hamamci IE, Er S, Shit S, Reynaud H, Kainz B, Menze B. Data from: CRG score: a distribution-aware clinical metric for radiology report generation (2025). arXiv: 2505.17167 [cs].