BMC Medical Informatics and Decision Making

RESEARCH

Open Access

# Uncertainty-aware automatic TNM staging classification for [$^{18}$F] Fluorodeoxyglucose PET-CT reports for lung cancer utilising transformer-based language models and multi-task learning

Stephen H. Barlow[1*], Sugama Chicklore[1,2], Yulan He[3,4,5], Sebastien Ourselin[1], Thomas Wagner[6], Anna Barnes[1,7†] and Gary J.R. Cook[1,2†]

## Abstract

**Background** [$^{18}$F] Fluorodeoxyglucose (FDG) PET-CT is a clinical imaging modality widely used in diagnosing and staging lung cancer. The clinical findings of PET-CT studies are contained within free text reports, which can currently only be categorised by experts manually reading them. Pre-trained transformer-based language models (PLMs) have shown success in extracting complex linguistic features from text. Accordingly, we developed a multi-task 'TNMu' classifier to classify the presence/absence of tumour, node, metastasis ('TNM') findings (as defined by The Eight Edition of TNM Staging for Lung Cancer). This is combined with an uncertainty classification task ('u') to account for studies with ambiguous TNM status.

**Methods** 2498 reports were annotated by a nuclear medicine physician and split into train, validation, and test datasets. For additional evaluation an external dataset ($n=461$ reports) was created, and annotated by two nuclear medicine physicians with agreement reached on all examples. We trained and evaluated eleven publicly available PLMs to determine which is most effective for PET-CT reports, and compared multi-task, single task and traditional machine learning approaches.

**Results** We find that a multi-task approach with GatorTron as PLM achieves the best performance, with an overall accuracy (all four tasks correct) of 84% and a Hamming loss of 0.05 on the internal test dataset, and 79% and 0.07 on the external test dataset. Performance on the individual TNM tasks approached expert performance with macro average F1 scores of 0.91, 0.95 and 0.90 respectively on external data. For uncertainty an F1 of 0.77 is achieved.

**Conclusions** Our 'TNMu' classifier successfully extracts TNM staging information from internal and external PET-CT reports. We concluded that multi-task approaches result in the best performance, and better computational

†Anna Barnes and Gary J.R. Cook contributed equally to this work.

*Correspondence:
Stephen H. Barlow
stephen.barlow@kcl.ac.uk

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

efficiency over single task PLM approaches. We believe these models can improve PET-CT services by assisting in auditing, creating research cohorts, and developing decision support systems. Our approach to handling uncertainty represents a novel first step but has room for further refinement.

**Keywords** Natural language processing, Deep learning, Electronic health records, Pretrained language models, Transformer, Medical imaging, PET-CT, Radiology

## Background

[$^{18}$F] Fluorodeoxyglucose (FDG) positron emission tomography (PET), combined with computed tomography (CT) as PET-CT imaging, is widely used in determining malignancy in lung nodules and staging known lung cancer [1–3]. The clinical findings of PET-CT images are recorded in a free text report [4], and accordingly the information in these reports is difficult to extract at scale. They also contain specialist terminology and language unique to nuclear medicine which make it difficult for those without domain expertise to understand, potentially including referring oncologists and surgeons [5]. TNM (tumour, node, metastasis) staging [6] is the most widely used staging system in lung cancer to guide clinical management decisions [7] and determining how PET-CT findings relate to this staging system is essential for determining the correct treatment approach for a patient.

Natural language processing (NLP) provides methods which retrieve structured information from unstructured text and has been used in a variety of healthcare applications [8]. The significance of the written report in PET-CT offers opportunities to improve service using NLP techniques. NLP approaches in radiology across all modalities are broadly split between rule-based [9–12] and machine learning based methods [13–15]. More recently, pre-trained language models (PLMs) utilising the transformer architecture [16] have become the basis of many approaches [14, 15, 17–21]. Sykes et al. [20] found rule-based methods remain effective for biomedical tasks but time consuming to develop and potentially inflexible to external datasets. There has been less work focusing on PET-CT specifically, but recently the pre-train and fine-tune transfer learning paradigm using PLMs (demonstrated in [22]) has become prevalent for PET-CT NLP classification tasks such as sentence-level anatomy classification [19], classifying lymphoma Deauville scores [23], and distinguishing lung cancer reports from other cancers [24].

Multi-task learning techniques have also become prevalent in NLP literature [25, 26]. By training a model for multiple tasks parameter-efficiency can be increased [27], and overfitting potentially reduced due to shared knowledge between tasks [28]. These characteristics are beneficial in healthcare where computational resources can be scarce, and confidence in continued performance is important. In PET-CT Eyuboglu et al. [29] explored its benefits for image classification, but its use in classifying reports is currently underexplored.

There is some existing NLP work extracting lung cancer staging information from both PET-CT and CT reports. For PET-CT: Park et al. [30] extracted the presence and location of metastasis using convolutional neural networks and LSTM-based networks, and Nobel et al. [31] adapted a rule-based algorithm from [32] to extract T and N stage information. Neither provided extensive external validation on PET-CT reports. This is likely due to difficulties obtaining external data due to the additional burden of satisfying ethical and data protection requirements. Demonstrating a model's efficacy on external data allows for greater understanding of performance pitfalls, and potentially allows other centres to use the model with confidence. Nobel et al. [31] also discussed the shortcomings of a rule-based approach to TN classification and how machine learning could help. For CT imaging, several approaches were developed for the "RR-TNM subtask of the NTCIR-17 MedNLP-SC shared task" using Japanese-language CT reports [33–36]. As part of this challenge Fukushima et al. [34] demonstrated the potential of fine-tuning PLMs for TNM staging classification. It should be noted that despite similarities PET-CT reports use different language from CT reports, as they primarily involve discussion of normal/abnormal tracer uptake, which is not relevant to CT reports. Accordingly, there is an opportunity to explore how PLMs can extract lung cancer staging information from PET-CT reports specifically.

TNM staging affects the treatment approach for lung cancer patients. For example, any 'M' positive finding would represent stage IV cancer (using the numeric system [37]) and is considered 'advanced' [38] cancer. Accordingly, this could significantly change the patient's treatment plan if not known previously. It would therefore be useful to automatically identify the presence or absence of 'T', 'N' and 'M' findings in lung cancer PET-CT reports for staging, clinical alerts, and future research and audit. This process can currently only be performed reliably by having experts manually read reports. Related to this, uncertainty and ambiguity are factors in PET-CT reports [39–41]. We found previous NLP work in radiology either excluded such reports [21], or seemingly insisted the annotator decide either way [15, 30, 31]. It would be an advantage to capture this information rather than exclude or quantise it.

Accordingly, this study seeks to develop, create, and evaluate a transformer-based multi-task TNMu (Tumour, Node, Metastasis, uncertainty) classifier for FDG PET-CT lung cancer reports. It will then be evaluated on an external PET-CT dataset from another hospital with different reporting practices. Finally, we will compare its performance against human experts and analyse the results.

## Methods

### Clinical data

#### *Data acquisition*

MIMIC III [42] is the only public dataset with free text PET-CT report data. The newest records in this dataset are from 2012 and we found ~1200 FDG PET-CT reports in total, of which ~400 are for lung cancer. Due to the rapid development of PET-CT over the last decade, and the need for more reports for training and testing, we constructed a new dataset from King's College London & Guy's and St Thomas' PET Centre (internal) with additional external test data from the Royal Free Hospital (external). The data use and collection was approved by UK Research Ethics Committee (UK IRAS 228790) as part of Guy's Cancer Cohort (ref: 18/NW/0297) [43].

We included FDG PET-CT reports with clinical indications that suspected lung cancer or investigated confirmed lung cancer. Figure 1 shows how 2601 internal reports dated between January 2013 and December 2022 were sampled from a larger corpus ($n$=60796) for annotation. The ICD10 code 'C34' (malignant neoplasm of bronchus and lung) was used to identify lung cancer cases from the larger corpus. It should be noted that for the internal data source this returned both confirmed and unconfirmed lung cancer scans. As PET-CT reporting practices have developed over time we chose to represent a decade of practices in the training data of the model to encourage better generalisation to external data. During the annotation process 103 non-lung cancer reports were removed. The external data did not have ICD10 codes so 511 scans were selected for annotation using regular

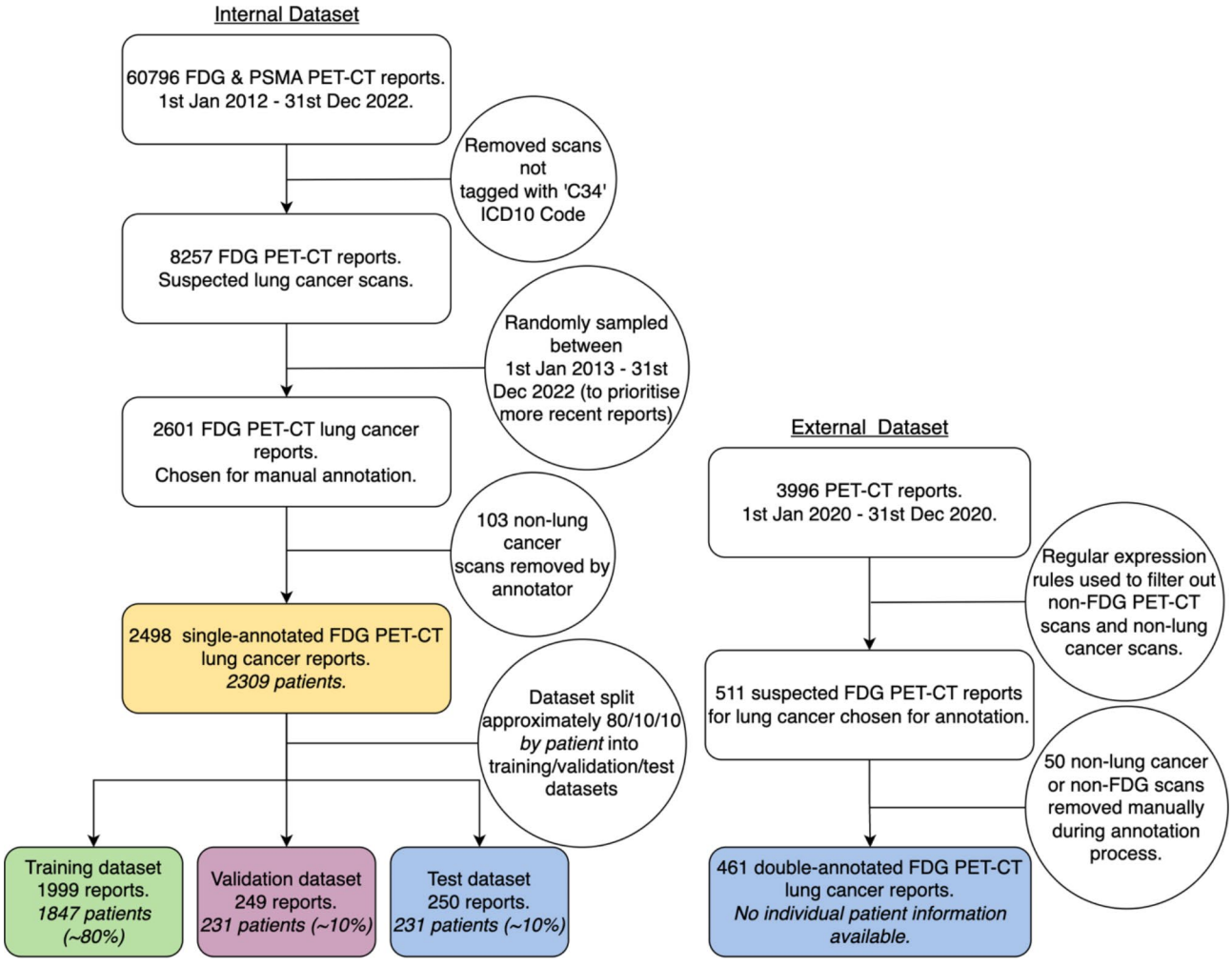


**Fig. 1** Flowcharts outlining the creation of the four datasets used in this study

expression rules which searched for key terms and acronyms like 'primary lung' and 'nsclc'. 461 external reports were then confirmed by the annotators as FDG PET-CT lung cancer reports and included in the final test dataset (Fig. 1).

For the internal data, anonymous patient IDs were created (and the originals permanently deleted) to create fully anonymised, stratified datasets. No other identifiable patient information was kept, and the Spacy library [44] was used to remove any names (clinicians/admin staff) and dates in the report text. The external data were fully anonymised as no identifiable patient information was provided but were also processed with the Spacy pipeline. Patient stratification was not required for the external data as they were only used for model testing.

#### *Data annotation*

A nuclear medicine consultant (30 years of PET experience) annotated the internal reports at the document level for the presence or absence of any finding that would represent 'T', 'N' or 'M'-positive status according to The Eighth Edition of TNM Staging for Lung Cancer [37]. Two nuclear medicine consultants (30 and 14 years experience) annotated the external data, one being the internal annotator. We found that in some reports the original reporter would be unsure of a finding or unclear in describing it. We assert it is useful for the model to have ambiguous examples in its training and testing data to demonstrate it could work effectively on them in deployment. In our initial exploration of the data the number of uncertain/ambiguous examples varied significantly between the TNM tasks. For example, the 'N' findings were rarely uncertain, but uncertain 'T' and 'M' findings were more common. From this we felt creating individual 'uncertainty' classes for 'T', 'N' and 'M' would hurt performance with little benefit. To solve this, we introduce an 'uncertainty' ('u') task with the following definition:

If the annotator deems any of the TNM findings to be uncertain, indeterminate, or ambiguous then the corresponding 'T', 'N' or 'M' class is positive, and the 'u' class also becomes positive. If there is no uncertainty/ambiguity associated with the TNM findings, then the 'u' class is negative.

Four types of uncertainty became apparent during annotation: Reporter uncertainty (where the original reporter is uncertain of a findings significance), uncertainty from unclear language (where the original report text is unclear), uncertainty by omission (where a crucial detail is missing from the original report), and uncertainty from technical limitations (where limitations of PET-CT imaging restrict a definitive opinion). The phrase "middle lobe and left lower lobe nodules are below the resolution of PET but remain suspicious and require ongoing surveillance as a minimum" provides an example of both reporter uncertainty and technical uncertainty as the reporter is unsure of the significance of lung nodules due to technical limitations of PET-CT imaging. Linguistic clarity and omission were rarer types of uncertainty and usually came about because of unusual clinical contexts involving the whole report. Accordingly, each report was annotated with binary labels for 'T', 'N', 'M' and 'u' findings, preserving uncertainty information which would otherwise be lost via exclusion or quantisation, while allowing for consistent classification performance.

Due to the expense and time-pressures of nuclear medicine physicians we used a single annotator for the internal data, and two annotators for the external test data. If the two annotators had contradictory annotations, then these were resolved via discussion and consensus once all examples had been labelled. This allowed us to test inter-annotator agreement and create a suitable gold standard test set for final model evaluation. This annotation approach was the best use of available resources as deep learning models have been shown to get good performance from imperfect training data [29, 30, 45], but still require evaluation on gold standard data.

#### *Dataset splits*

The annotated internal data were split at patient level into training, validation, and test sets at a ratio of 80:10:10 (Fig. 1). Splitting datasets by patient has been used in similar studies [15, 21, 30], and serves as a simulation of external data by making sure no patient is present in more than one dataset. The external data were exclusively used for final testing to assess the model's ability to generalise to data from another centre.

Table 1 Shows the distribution of classes for each dataset. All tasks have class imbalances with the most severe being the uncertainty task. The tumour task has a class imbalance opposite to the others with mostly positive cases. T0 (primary tumour-negative) labels are common due to PET-CT's use in the diagnosis of lung cancer, where it can help determine if a nodule is malignant or not. Cohen's kappa ($\kappa$) was used to test inter-annotator agreement on the external data and $\kappa = 0.77$ for 'T', 0.94 for 'N', 0.82 for 'M', and 0.38 for 'u'. This represents "substantial agreement" for 'T', "almost perfect agreement" for 'N' and 'M', and "fair agreement" for 'u' according to [46].

### Model architecture and training

#### *Model input*

PET-CT reports are usually split into 'Findings' and 'Impression' sections [39]. The exact section names can vary but keep similar semantics [47]. The 'Impression' section serves as the conclusion of a report as reporting guidelines [1] recommend that the clinical relevance of

**Table 1** Class label distribution for the four datasets used in this study in the format '0/1' where '0' represents the number of negative examples for that class and '1' represents the number of positive examples of that class present in the dataset

| Dataset | Task Name (Absence / Presence) | | | |
|---|---|---|---|---|
| | **Tumour ('T') (0/1)** | **Node ('N') (0/1)** | **Metastasis ('M') (0/1)** | **Uncertainty ('u') (0/1)** |
| Internal Train | 472 / 1527 | 1321 / 678 | 1609 / 398 | 1700 / 299 |
| Internal Validation | 55 / 194 | 160 / 89 | 189 / 60 | 213 / 36 |
| Internal Test | 54 / 196 | 157 / 93 | 194 / 56 | 208 / 42 |
| External Test | 89 / 372 | 285 / 176 | 336 / 125 | 393 / 68 |

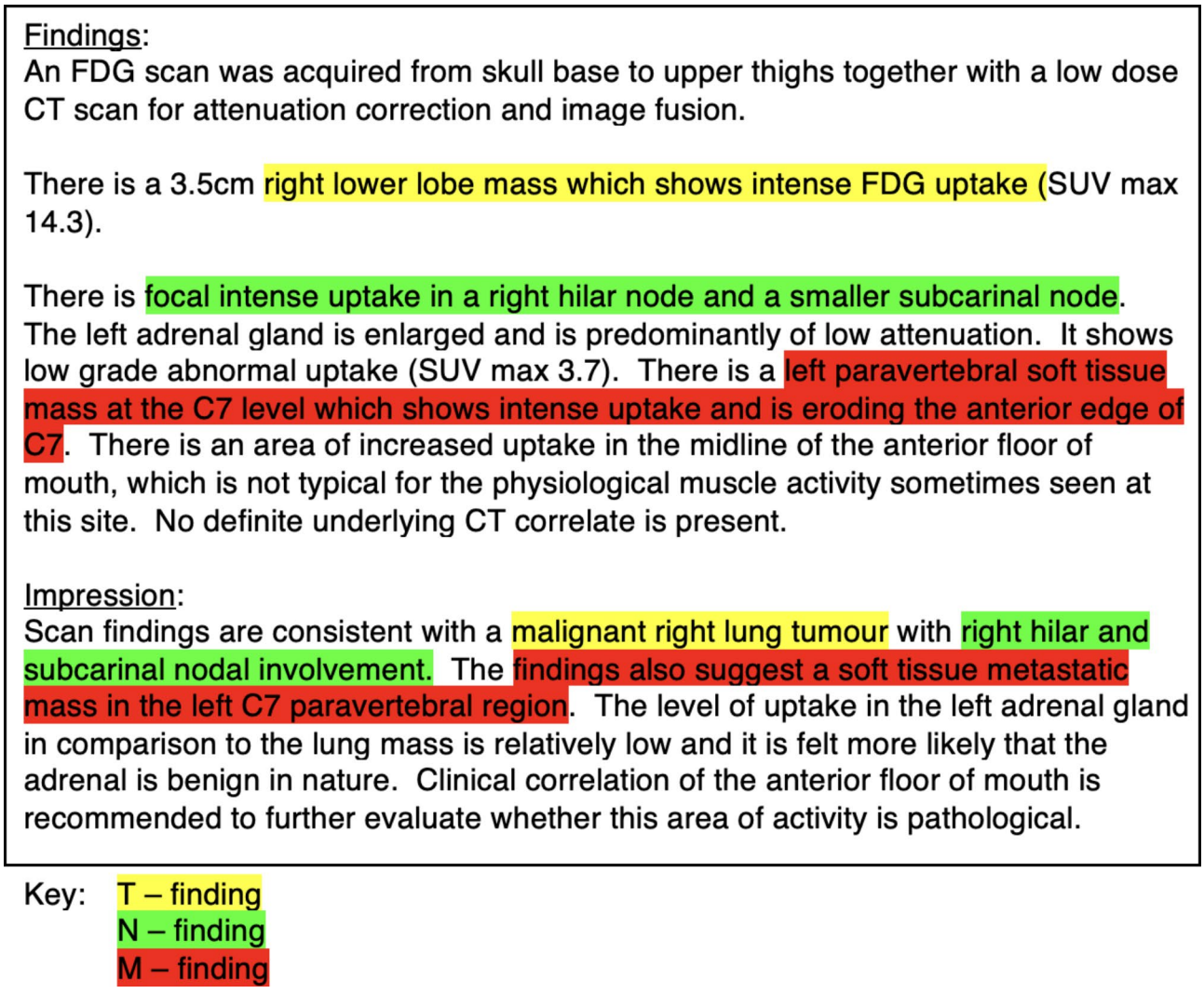


**Fig. 2** Example FDG PET-CT report. Phrases that would correspond to positive T, N or M findings are highlighted in the appropriate colour. This example would be labelled T1 N1 M1 u0 at the document level, as the original reporter has clearly noted positive Tumour, Node, and Metastasis with no ambiguity or uncertainty

any findings is stated here. The 'Findings' section is where the observations of interest on the image are recorded, but not necessarily what they mean for the patient. Previous work has used the 'Impression' section alone [13, 21, 30], the 'Findings' section alone [19], or both [48]. For our work we chose to use the entire report as we found this offered superior performance over the 'Impression' or 'Findings' sections alone. Figure 2 shows an example FDG PET-CT report with sections highlighted that would indicate TNM findings.

Transformer based PLMs use tokenization algorithms such as WordPiece [49] or Byte Pair Encoding (BPE) [50]. We use the appropriate trained tokenizer for each PLM, truncating the start of a report if it exceeded the 512 token limit (common to all models tested) as opposed to the end. This ensures we conserve the 'Impression'

section of a report containing the most clinically relevant information. No other text processing is performed other than that required for the tokenizer (as determined by Huggingface's [51] implementation). We sought to use as little dataset-specific cleaning as possible as this could bias the model towards specific reporting tendencies.

#### *Problem formulation and model architecture*

We used the Transformers [51] and Pytorch [52] libraries to develop the models. TNM staging information can be highly contextual (Fig. 2) so a model that can incorporate contextual information is essential for good performance. PLMs like BERT [22] are pre-trained on a large corpus of text in a self-supervised fashion to gain an understanding of how language works. They can then be fine-tuned for a specific task like document classification. This fine-tuning process is a form of transfer learning [53] and is the paradigm we use in this study.

We formulate the TNMu classification problem as multi-label classification, where each document is assigned four binary labels which are positive ('1') or negative ('0'). PLMs offer the opportunity of using a shared encoder and utilising the benefits of multi-task learning [54, 55], allowing more appropriate features to be learnt that potentially generalise better to new data. This also makes the model more computationally efficient by reducing the number of trainable parameters required for each task. A four neuron classification head is appended to a shared PLM encoder (Fig. 3), and we compare this against training separate binary classifiers for each task with their own encoders. Figure 3 demonstrates the complete pipelines. A report is tokenized, then inputted to a PLM consisting of a series of transformer encoder blocks utilising multi-head self-attention (as detailed in [16]). A dropout [56] regularization layer is then used between the PLM and the classification head (probability set at 0.1) to prevent overfitting. The model finally outputs a number between 0 and 1 for each task and if this number is greater than 0.5, it returns a positive prediction for that task.

We treat the choice of PLM as a hyperparameter and evaluate eleven encoder-only PLMs on the validation set. PLMs encode large amounts of text data via self-supervision, and those used in this study primarily use a masked language modelling objective, where the model attempts to predict the masked word(s) for large quantities of text. Encoder-only PLMs are unable to generate text like auto-regressive and encoder-decoder models (such as GPT [25, 57] and T5 [58]), but benefit from bi-directional context in their predictions, which has proven useful for text classification tasks [22]. These PLMs are trained on corpora with different characteristics, often with a particular focus. These could broadly be described 'general', 'biomedical' and 'clinical'. BERT [22] and RoBERTa [59] use general data sources such as Wikipedia, books, and web crawl data. BioBERT [60] and BioMegatron [61] further pre-train on biomedical literature from PubMed. GatorTron [62], BioClinicalBERT [63] and RadBERT [64] contain actual electronic health records (such as radiology reports) in their training corpora. Generally biomedical

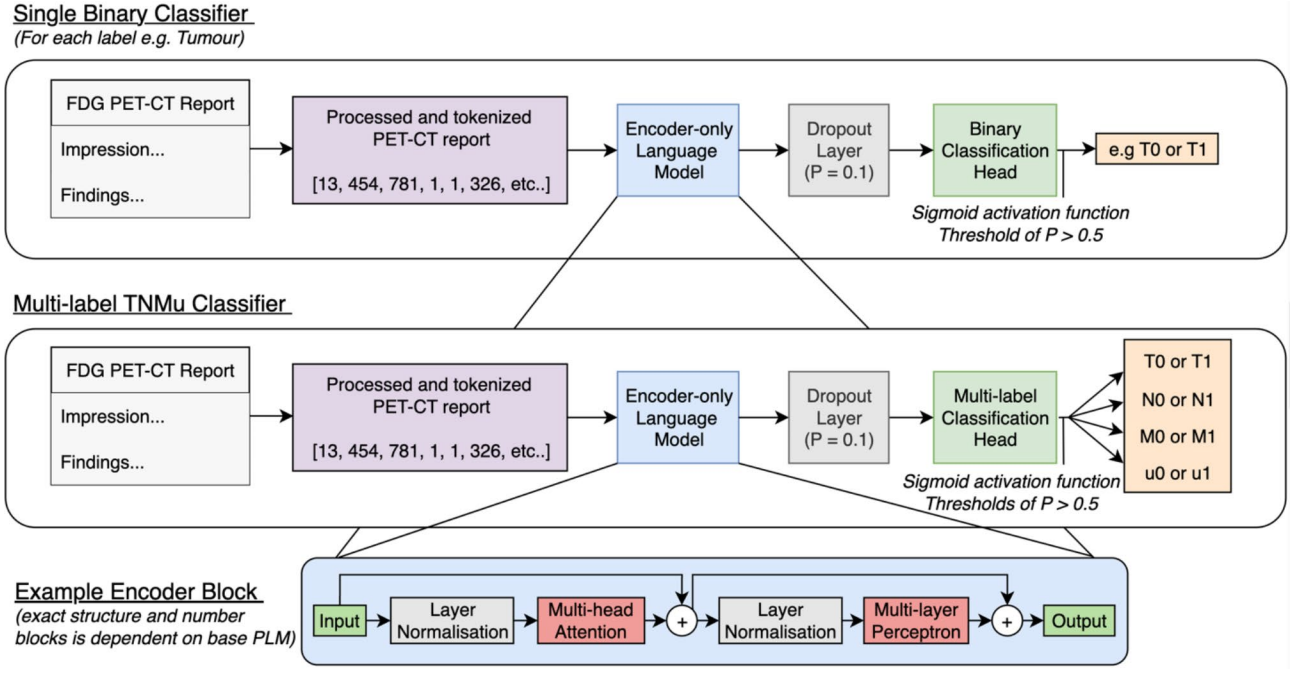


**Fig. 3** Model architecture diagrams displaying the difference between a single task approach for 'T' classification (which would be repeated for 'N', 'M', and 'u' classifications), and a multi-task approach TNMu classifier with a shared PLM encoder

models include general data as well as biomedical data, and clinical models include general, biomedical, and clinical data.

#### *Training*

Hyperparameters were selected by experimenting on the internal validation set. Table 2 shows all the combinations tested, with the final choices highlighted. The final choices were determined by the best classification performance. The learning rate was set at 1e-5 using an Adam optimizer with weight decay [65]. This was decayed linearly to zero over a total training time of five epochs with the first 10% of steps used as a warmup period. The loss function used was the binary cross entropy over the four tasks. For each configuration three models were fine-tuned with different random seeds allowing the reporting of the mean and standard deviation of these runs. The batch size was set at 8 as this was the largest possible for the larger models on an Nvidia RTX 3090 GPU. These hyperparameters offered consistently good, representative performance for each PLM when tested over three random seeds.

### Evaluation approach

To evaluate the models, we use metrics that characterise both the overall performance of the classifier, and individual task performance. For overall performance accuracy ($ACC_{TNMu}$) is used to gauge how many reports are classified completely correctly over all four tasks, and we also use an accuracy score which excludes 'u' to focus on the clinical tasks ($ACC_{TNM}$). Hamming loss ($HL_{TNMu}$) is also utilised as it is more lenient, and evaluates the model by penalising individual incorrect classifications, as opposed to an incorrect set of classifications, with a lower score being better.

All four tasks demonstrated class imbalance to varying degrees which needs to be considered in evaluation. Macro average F1 score is used to marry the concerns of precision and recall and has the benefit of treating classes equally regardless of imbalance. $F1_{TNMu}$ serves as the mean of the F1 scores for each task.

To evaluate multi-task learning we compare against the single-task models and a baseline, non-deep learning model which uses TF-IDF (term frequency-inverse document frequency) [66] encodings and a logistic regression [67] classifier for each task (implemented using Scikit-learn [68]). A 'TNM only' multi-task model is also trained to test if the inclusion of 'u' training labels degrades performance.

Finally, we take the best performing multitask model (determined using the average performance across all stated metrics over both test sets) and compare against the two annotators of the external data and an ensemble of the best single task models for each task. This ensemble was created by selecting the single task classifiers with the best macro average F1 scores (averaged over both test sets) for 'T', 'N', 'M', and 'u'. Receiver operator characteristic (ROC), precision-recall curves (PRC), and confusion matrices are used to compare the final model's performance on each task individually.

## Results

Table 3 shows how the different PLMs performed on the validation set. GatorTron significantly outperforms all the other models tested by all metrics and was accordingly chosen as the base model for the TNMu models. The other interesting finding is that the smaller models (~110 million parameters) struggled to classify the 'u' task. The only models to achieve a macro average F1 score over 0.70 for 'u' had at least 340 million parameters.

Table 4 Compares multi-task and single task approaches on the internal and external test dataset over three different training runs. Results of a TF-IDF logistic regression model are also shown to serve as a baseline. For the multi-task approaches we train models including uncertainty labels ('TNMu') and models without ('TNM only') to test if including these labels degrades performance. Unsurprisingly the PLM pipelines dramatically outperform the traditional machine learning baseline. For the PLM pipelines similar performance is observed on the TNM tasks, but a multi-task approach offers significant improvements in 'u' task performance and generalisation, which in turn improves the overall accuracy and hamming loss metrics. The TNM only pipeline shows no significant difference in performance, justifying the use of uncertainty labels. The PLMs generalise well to external data, but a modest drop in performance remains. This is most pronounced on the 'u' task.

The best performing individual multi-task GatorTron model was selected (from the three trained in table 4) as determined by its average performance across all metrics across both test sets. In Table 5 it is compared against an

**Table 2** This table shows the hyperparameter combinations we tested (in all permutations). The chosen values are in underlined bold. These were determined by best classification performance

| Hyperparameter | Values tested (best performing in underlined bold) | | | | | |
|---|---|---|---|---|---|---|
| Learning rate | 9e-6 | **1e-5** | 2e-5 | 3e-5 | 4e-5 | 5e-5 |
| Epochs | 2 | 3 | 4 | **5** | 6 | 7 |
| Batch size | 1 | 2 | 4 | **8** | - | - |
| Dropout probability | 0 | **0.1** | 0.2 | 0.3 | 0.4 | 0.5 |

**Table 3** Comparison of PLMs on the internal validation set. Each model was fine-tuned three times with different random seeds for 5 epochs. The results show the mean and standard deviation for each metric of those training runs. The training corpora focus gives an idea of the corpus the PLM was pre-trained on. Bold text indicates the best result for that metric. All F1 scores are macro averaged

| PLM | Training Corpora Focus | Parameters | $ACC_{TNMu}$ ↑ | $HL_{TNMu}$ ↓ | $F1_T$ ↑ | $F1_N$ ↑ | $F1_M$ ↑ | $F1_u$ ↑ |
|---|---|---|---|---|---|---|---|---|
| BERT (Base) [22] | General | 110 m | 0.63 ± 0.03 | 0.13 ± 0.01 | 0.78 ± 0.04 | 0.91 ± 0.02 | 0.80 ± 0.02 | 0.46 ± 0.00 |
| BERT (Large) [22] | General | 340 m | 0.67 ± 0.02 | 0.12 ± 0.01 | 0.83 ± 0.03 | 0.91 ± 0.01 | 0.81 ± 0.01 | 0.53 ± 0.08 |
| RoBERTa (Base) [59] | General | 125 m | 0.69 ± 0.03 | 0.10 ± 0.01 | 0.83 ± 0.02 | 0.93 ± 0.01 | 0.84 ± 0.01 | 0.56 ± 0.08 |
| RoBERTa (Large) [59] | General | 355 m | 0.77 ± 0.01 | 0.08 ± 0.01 | 0.91 ± 0.01 | 0.93 ± 0.01 | 0.87 ± 0.02 | 0.74 ± 0.01 |
| BioBERT (Base) [60] | Biomedical | 110 m | 0.70 ± 0.01 | 0.10 ± 0.00 | 0.85 ± 0.01 | 0.92 ± 0.01 | 0.85 ± 0.01 | 0.50 ± 0.02 |
| BioBERT (Large) [60] | Biomedical | 340 m | 0.75 ± 0.02 | 0.09 ± 0.01 | 0.91 ± 0.02 | 0.93 ± 0.01 | 0.86 ± 0.02 | 0.70 ± 0.02 |
| BioClinicalBERT [63] | Clinical | 110 m | 0.66 ± 0.01 | 0.12 ± 0.00 | 0.81 ± 0.03 | 0.90 ± 0.01 | 0.78 ± 0.03 | 0.46 ± 0.00 |
| BioMegatron [61] | Biomedical | 345 m | 0.76 ± 0.02 | 0.08 ± 0.01 | 0.90 ± 0.01 | 0.95 ± 0.01 | 0.85 ± 0.01 | 0.73 ± 0.00 |
| RadBERT [64] | Clinical | 110 m | 0.62 ± 0.01 | 0.13 ± 0.01 | 0.78 ± 0.02 | 0.88 ± 0.03 | 0.79 ± 0.03 | 0.46 ± 0.00 |
| RadBERT-RoBERTa-4 m | Clinical | 125 m | 0.71 ± 0.01 | 0.10 ± 0.01 | 0.88 ± 0.02 | 0.93 ± 0.01 | 0.84 ± 0.01 | 0.59 ± 0.02 |
| GatorTron (Base) [62] | Clinical | 345 m | **0.84 ± 0.01** | **0.06 ± 0.01** | **0.96 ± 0.01** | **0.96 ± 0.00** | **0.90 ± 0.01** | **0.81 ± 0.01** |

**Table 4** A comparison of machine learning pipelines including two multi-task approaches using a shared GatorTron PLM encoder (one including and one excluding uncertainty labels in training), an ensemble of finetuned binary classifiers using GatorTron, and a traditional machine learning model using TF-IDF encodings and individual logistic regression classifiers for each binary task. Each approach was trained three times with different random seeds with the mean result and standard deviation reported. For the single task ensembles we calculate the 'TNMu' and 'TNM' metrics using the models trained from that random seed. Bold values represent the best performing pipeline for that metric on each test dataset. All F1 scores are macro averaged

| Dataset | Pipeline | $ACC_{TNMu}$ ↑ | $ACC_{TNM}$ ↑ | $HL_{TNMu}$ ↓ | $F1_{TNMu}$ ↑ | $F1_T$ ↑ | $F1_N$ ↑ | $F1_M$ ↑ | $F1_u$ ↑ |
|---|---|---|---|---|---|---|---|---|---|
| ***Internal Test*** | Multi-task (TNMu) | **0.84 ± 0.01** | **0.86 ± 0.00** | **0.05 ± 0.00** | **0.92 ± 0.00** | 0.93 ± 0.00 | 0.94 ± 0.00 | **0.92 ± 0.01** | **0.87 ± 0.00** |
| | Multi-task (TNM only) | N/a | 0.85 ± 0.01 | N/a | N/a | 0.94 ± 0.01 | 0.95 ± 0.01 | 0.89 ± 0.01 | N/a |
| | Single task | 0.80 ± 0.02 | **0.86 ± 0.00** | 0.06 ± 0.00 | 0.91 ± 0.01 | **0.95 ± 0.00** | **0.96 ± 0.00** | 0.89 ± 0.02 | 0.85 ± 0.02 |
| | TF-IDF + Logistic Regression | 0.50 ± 0.00 | 0.60 ± 0.00 | 0.16 ± 0.00 | 0.66 ± 0.00 | 0.69 ± 0.00 | 0.81 ± 0.00 | 0.69 ± 0.00 | 0.45 ± 0.00 |
| ***External Test*** | Multi-task (TNMu) | **0.78 ± 0.01** | **0.83 ± 0.01** | **0.07 ± 0.00** | **0.88 ± 0.01** | **0.89 ± 0.02** | **0.95 ± 0.01** | 0.89 ± 0.01 | **0.77 ± 0.00** |
| | Multi-task (TNM only) | N/a | **0.83 ± 0.02** | N/a | N/a | 0.88 ± 0.01 | **0.95 ± 0.00** | **0.91 ± 0.02** | N/a |
| | Single task | 0.73 ± 0.00 | 0.82 ± 0.00 | 0.08 ± 0.00 | 0.85 ± 0.01 | 0.88 ± 0.00 | **0.95 ± 0.00** | 0.90 ± 0.01 | 0.68 ± 0.02 |
| | TF-IDF + Logistic Regression | 0.52 ± 0.00 | 0.61 ± 0.00 | 0.16 ± 0.00 | 0.64 ± 0.00 | 0.49 ± 0.00 | 0.85 ± 0.00 | 0.76 ± 0.00 | 0.46 ± 0.00 |

**Table 5** The external test set is used to compare the best performing multi-task model, an ensemble of the four best performing single task classifiers (all determined by average performance across all metrics on both internal and external test datasets), and the two expert annotators. Bold values represent which AI model pipeline performed best. All F1 scores are macro averaged

| | $ACC_{TNMu}$ ↑ | $ACC_{TNM}$ ↑ | $HL_{TNMu}$ ↓ | $F1_{TNMu}$ ↑ | $F1_T$ ↑ | $F1_N$ ↑ | $F1_M$ ↑ | $F1_u$ ↑ |
|---|---|---|---|---|---|---|---|---|
| Multi-task | **0.79** | **0.84** | **0.07** | **0.89** | **0.91** | **0.95** | 0.90 | **0.78** |
| Single task | 0.74 | **0.84** | 0.08 | 0.87 | 0.89 | **0.95** | **0.92** | 0.70 |
| Annotator 1 | 0.90 | 0.93 | 0.04 | 0.94 | 0.95 | 0.99 | 0.96 | 0.84 |
| Annotator 2 | 0.89 | 0.93 | 0.04 | 0.93 | 0.94 | 0.99 | 0.95 | 0.83 |

**Table 6** Cohen's Kappa is used to compare the best multi-task model's inter-annotator agreement with each expert annotator (before agreement process) on the external test set

| | Tumour (κ) | Node (κ) | Metastasis (κ) | Uncertainty (κ) |
|---|---|---|---|---|
| *Annotator 1* | **0.78** | 0.88 | **0.81** | **0.56** |
| *Annotator 2* | 0.75 | **0.89** | 0.75 | 0.46 |

ensemble of the best performing single task models for each task, and both expert annotators on the external data. The multi-task model performs better on average than the ensemble, and approaches expert performance on the individual tasks. These errors compound on the overall metrics however, to create a gap in aggregate performance against the experts.

As the training data were annotated by a single annotator, we test whether this manifests in annotator-bias on the external test set using Cohen's Kappa. If the agreement scores are significantly higher with the annotator who labelled both sets (when compared with agreement to the other annotator) we could argue there is a degree of bias. Table 6 shows that the discrepancy is not uniform with the tumour and node tasks displaying minimal

differences, however the metastasis and uncertainty tasks do display greater agreement with the training data annotator suggesting a degree of bias towards their judgement. The lower agreement scores for uncertainty reflect that this task had lower labelling agreement between annotators initially.

Figure 4 shows receiver operating characteristic (ROC), precision-recall curves, and the corresponding areas under each curve for the best performing multi-task

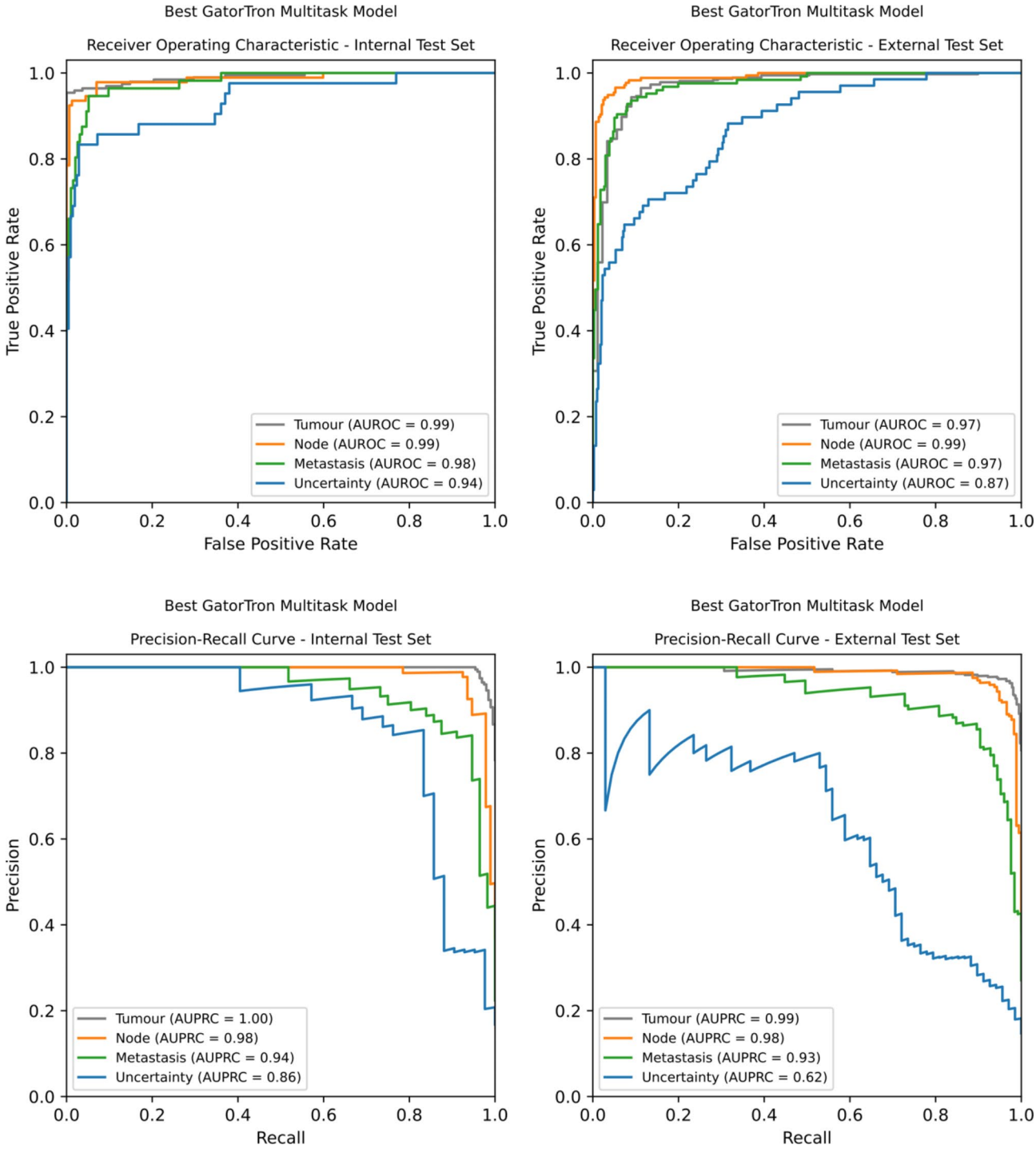


**Fig. 4** Receiver operating characteristic (ROC) and precision-recall curves for the best multi-task GatorTron model on both internal and external test sets. The 'T' precision-recall curve is included for completeness, but due to the class distribution being skewed towards the positive label it overstates performance and is likely not a suitable performance metric

model on the internal and external test sets. The TNM tasks show similar patterns on both datasets, but the uncertainty task drops in performance on the external data. The tumour task precision-recall curve is included for completeness, but due to it having predominantly positive labels, its performance is likely overstated with precision-recall curve metrics.

Figure 5 shows confusion matrices for the model on the external dataset. The model predominantly makes correct classifications but the errors trend towards false-positives for the 'T' and 'N' tasks, and false-negatives on 'M' and 'u'.

## Discussion

In this work we have developed a deep learning NLP model which classifies the TNM status of FDG PET-CT reports for lung cancer, and whether uncertainty is associated with the findings. The best performing TNMu classifier was a multi-task model utilising GatorTron [62]. This model classified 84% of internal reports, and 79% of external reports correctly across all four tasks. There was a 5% ($HL_{TNMu} = 0.05$) chance of it classifying an individual internal TNMu class incorrectly, and a 7% ($HL_{TNMu} = 0.07$) chance of incorrectly classifying an external TNMu class. We also evaluate the model against the experts' classifications on the external data (before consensus) as humans also make errors in classifying documents. Despite not matching expert performance, the model approaches it, particularly on the clinical TNM tasks, and is capable of categorising large numbers of reports in a fraction of the time, which would be particularly useful for audit purposes, and creating research cohorts. We believe this classification performance makes the model applicable for both primary uses of PET-CT for lung

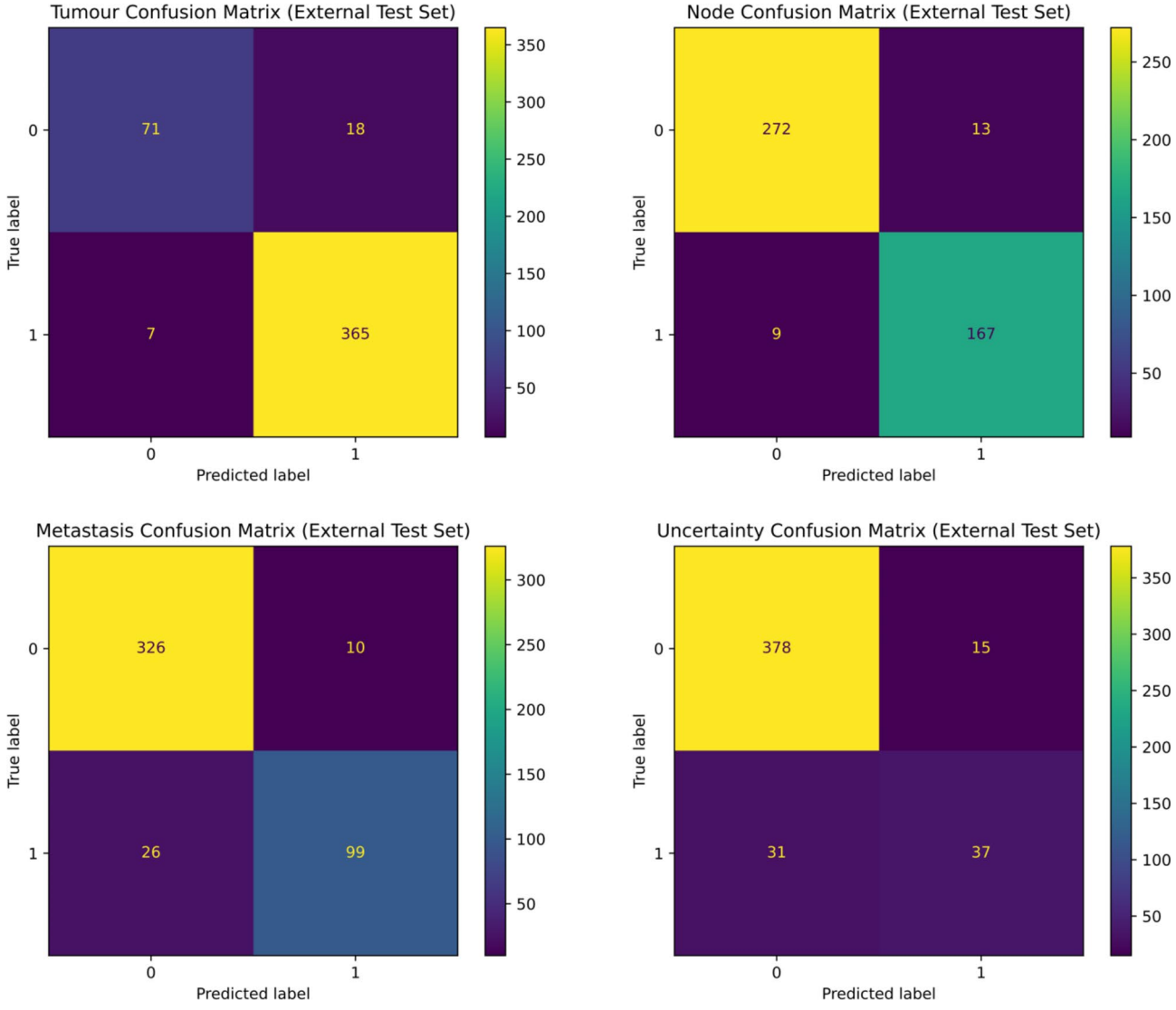


**Fig. 5** Confusion matrices for each task using the best performing multi-task GatorTron model on the external test set

cancer patients. Being able to distinguish between T-positive and T-negative assists the use case where a lung nodule is being characterised for malignancy, potentially before formal diagnosis, and the binary classifications of 'N' and 'M' status assist staging known lung cancers, where positive findings can significantly impact treatment pathways for patients going forward. A last potential use is to use these report models as a noisy labeller for PET-CT images themselves to create larger datasets for a TNM image classifier. This approach was used successfully for brain MRI in [15, 69].

The main performance difference between the two datasets concerns the 'u' task. With this task removed from consideration the internal accuracy over the TNM tasks becomes 86% and only drops to 84% for external data. This suggests good generalisation for the clinical tasks and is encouraging as the external reporting practices are different from those used internally. The external reporters more frequently use an "anatomic" approach for the 'findings' section (e.g. report findings in the head and neck first, then thorax etc… [70]), as opposed to a "priority" approach prevalent internally (where the most clinically relevant findings are stated first in order of T, N and then M [70]). This suggests transformer-based NLP models can be robust to different reporting styles for TNM classification. This change in style could also offer a reason for why the 'u' task generalises less well. Uncertainty is not as formally defined as the TNM tasks, so it could be more dependent on individual human judgements, which could be affected by reporting style choices. This may also explain why the inter-annotator agreement for the 'u' task was significantly lower than for the other tasks on the external test set. Despite this limitation it was interesting that a multi-task model was able to learn more generalisable features for uncertainty than a single task approach, and we assert capturing this information noisily is better than removing it. We found no other similar work which attempted to handle uncertainty outside of removing ambiguous reports [21], and most did not mention how uncertainty/ambiguity was handled, suggesting quantisation into the positive or negative categories. Uncertainty and ambiguity in reports is a potential concern in the wider field of radiology reporting [40, 41] and with further research a model could be developed as a teaching, or warning tool if report text is deemed to be too ambiguous. It is also worth noting that the discrepancy in performance between the 'TNM' and 'u' tasks would not have been as clear without the emphasis put on external evaluation, nor the improvements from multi-task learning. This confirms that external validation of deep learning models is useful for finer-grain analysis of model performance.

Looking at the TNM tasks individually it was observed that 'N' was the easiest for both the models and experts to classify across every experiment. We speculate this is likely this is due to being more self-evident than 'T' or 'M'. Lymph nodes tend to be either abnormal or normal, as determined by increased uptake of the FDG tracer, and the language used in reports seems to reflect this. The 'T' task may depend on contextual information that is not be explicitly stated in every report, and 'M' findings can be associated with multiple organs, resulting in a wider range of descriptors. It is interesting to note that the experts' classification decisions against the gold standard are also imperfect. Their individual classification performance and inter-annotator agreement statistics follow a similar pattern to the models where 'N' status is more successfully determined than 'T' and 'M'. This suggests they are harder to ascertain from reading PET-CT reports than 'N' findings.

Developing the model has provided insights into important parts of the methodology. We found the choice of PLM to be the most crucial component of developing these systems. Other literature [60, 61] has shown that medical text benefits from specialised PLMs, but FDG PET-CT reports seem to benefit from further specialisation. GatorTron dramatically outperformed all the other PLMs tested, and we suspect this is due to the pre-training data. It was pre-trained on >90 billion words of de-identified EHRs from the University of Florida [62]. We speculate its superior performance stems from being the only PLM found that contains contemporary PET-CT reports in its pre-training corpus. Tan et al. [21] also found GatorTron to be the best performing PLM in classifying CT radiology reports, but we note the performance difference was much smaller than witnessed in this study. It is also interesting that RadBERT, which is specialised for radiology tasks [64], and pre-trained on radiology reports, did not perform as well in comparison. This suggests that PET-CT reporting contains language that is distinct from other imaging modalities. A final feature of the PLMs tested was that smaller models (~110 million parameters) struggled to classify the uncertainty task, whereas larger models (>340 million parameters) were able to make a reliable distinction.

Multi-task learning was also found to provide benefits to performance and computational (and therefore also energy) efficiency, as has previously been described in other work [25–27, 54, 55]. Using a single 345 million GatorTron encoder out-performs individual classifiers for TNMu classification. This was true even when creating an ensemble from the best performing individual classifiers. This approach also reduces both the time and computation required for both training and inference, as our multi-task approach adds only three trainable parameters to the model (~0.00000087%). All this suggests there is no downside to experimenting with multi-task approaches, yet many potential upsides. It is possible that

reformulating other medical classification tasks as multi-label problems may provide similar benefits, although this is likely to be task dependent.

The main limitation of this study was how much expert annotated data we could accrue. Annotation is the most expensive part of the project, as experienced nuclear medicine physicians are required to make the judgements we are trying to model. Accordingly, we used a single annotator on the internal data. This means the model will potentially be biased towards that annotator's judgement, and annotation mistakes made during that process could affect its classifications. We attempted to explore how much bias is encoded into the model by evaluating Cohen's Kappa against the training data annotator, and the second annotator on the external data. Interestingly, there seems to be negligible bias on 'T' and 'N' tasks, but more on the 'M and 'u,' potentially because these tasks require more personal experience in the decision process. By evaluating the model on external gold standard data labelled by two annotators with a consensus process, and by quantifying the potential level of annotator bias in the model, we hope to mitigate most of the concerns a single annotator training dataset might create. Another limitation is that inter-annotator agreement on the 'u' task ($\kappa=0.38$) is lower than would be ideal for a clinical task. This is likely due to the amount of personal experience that enters in to determining a class with no formal definition like the TNM tasks. As the 'u' task does not directly affect clinical decision making, and we feel there is a need to address uncertain or ambiguous reporting, this work represents a first step while conceding that more refinement is required. We also note that we did not have access to certain patient demographic information (e.g. race, gender) from either hospital. Accordingly, we cannot exclude the possibility that the model performs differently on certain demographics and cannot report detailed demographic information about the datasets used in this study. Finally, these models are specifically trained and tested on FDG PET-CT reports for confirmed or suspected lung cancer. TNM staging is defined in relation to specific cancers so performance on other cancers or other imaging modalities cannot be guaranteed.

For future work we are interested in exploring multi-modal techniques, potentially combining NLP approaches with structured data such as radiomic features, which have been shown to have good predictive value [71, 72]. Multi-modal models have been applied to text and images for CT classification tasks [73], showing promising performance, but less work has been done utilising radiomics in a multi-modal (and PET-CT) context.

## Conclusions

We created a multi-task transformer-based NLP model which successfully classifies lung cancer FDG PET-CT radiology reports for the presence or absence of tumour, node, metastasis findings and whether the report contains uncertain or ambiguous findings for these. We successfully demonstrate it performs on a dataset from another hospital with a different reporting style. We believe this has the potential to assist the creation of research cohorts, the development clinical alert systems for previously unknown findings, and to assist auditing. The uncertainty/ambiguity classification represents a novel first step, but further refinement is needed.

#### Acknowledgements

Not applicable.

#### Author contributions

A.B., G.C., S.B. and S.O. conceived and planned the project. S.B. developed Python software, carried out experiments and prepared the first draft of the manuscript. Data analysis and interpretation was performed by S.B., A.B. and G.C. S.B. and T.W. retrieved, anonymised and pre-processed clinical data. G.C. and S.C. performed expert annotation of PET-CT reports. Y.H. provided technical insight, assisted with experiment design, and verified methods used. A.B., G.C. and S.O. provided academic supervision. All authors read and approved the final manuscript.

#### Funding

The authors acknowledge financial support from: EPSRC Research Council, part of the EPSRC DTP, Grant Ref: [EP/T517963/1], the Cancer Research UK National Cancer Imaging Translational Accelerator (C1519/A28682), and the Wellcome/Engineering and Physical Sciences Research Council Centre for Medical Engineering at King's College London (WT 203148/Z/16/Z).

#### Data availability

The data presented in this study are available on request from the corresponding author. Patient report data used in this study are not publicly available for ethical reasons. Reference Python code as used to develop and evaluate models is available at https://github.com/stephenhbarlow/TNM-classification. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

## Declarations

#### Ethics approval and consent to participate

The data use and collection was approved by United Kingdom Research Ethics Committee (UK IRAS 228790) as part of Guy's Cancer Cohort (ref: 18/NW/0297) [43]. The Guy's Cancer Cohort committee waived the need for individual consent to participate as the study uses retrospective, anonymised data.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare no competing interests.

#### Author details

[1]School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK
[2]King's College London and Guy's and St. Thomas' PET Centre, St. Thomas' Hospital, London, UK
[3]Department of Informatics, King's College London, London, UK
[4]Department of Computer Science, University of Warwick, Coventry, UK
[5]Alan Turing Institute, London, UK
[6]Department of Nuclear Medicine, Royal Free Hospital, London, UK

[7]King's Technology Evaluation Centre (KiTEC), School of Biomedical Engineering & Imaging Science, King's College London, London, UK

Received: 22 June 2024 / Accepted: 9 December 2024
Published online: 18 December 2024

**References**

1. Hofman MS, Hicks RJ. How we read oncologic FDG PET/CT. Cancer Imaging. 2016;16:35.
2. Sheikhbahaei S, Mena E, Yanamadala A, Reddy S, Solnes LB, Wachsmann J, Subramaniam RM. The value of FDG PET/CT in Treatment Response Assessment, Follow-Up, and Surveillance of Lung Cancer. AJR Am J Roentgenol. 2017;208:420–33.
3. Farsad M. FDG PET/CT in the staging of Lung Cancer. Curr Radiopharm. 2020;13:195–203.
4. Bijan B, Melendres G, Nguyen T. Fundamentals of oncologic PET-CT reporting. Mol Imaging Radionucl Ther. 2013;22:1–2.
5. Patel Z, Schroeder JA, Bunch PM, Evans JK, Steber CR, Johnson AG, Farris JC, Hughes RT. Discordance between Oncology Clinician–Perceived and Radiologist-intended meaning of the Postradiotherapy Positron Emission Tomography/Computed Tomography Freeform Report for Head and Neck Cancer. JAMA Otolaryngol Head Neck Surg. 2022;148:927–34.
6. Brierley J, Gospodarowicz MK, Wittekind C. Union for International Cancer C: TNM classification of malignant tumours. Chichester, West Sussex, UK: Wiley Blackwell; 2017.
7. Carter BW, Lichtenberger JP III, Benveniste MK, De Groot PM, Wu CC, Erasmus JJ, Truong MT. Revisions to the TNM staging of lung cancer: rationale, significance, and clinical application. Radiographics. 2018;38:374–91.
8. Zhou B, Yang G, Shi Z, Ma S. Natural Language Processing for Smart Healthcare. IEEE Rev Biomed Eng. 2022;1–17.
9. Sippo DA, Warden GI, Andriole KP, Lacson R, Ikuta I, Birdwell RL, Khorasani R. Automated extraction of BI-RADS Final Assessment categories from Radiology Reports with Natural Language Processing. J Digit Imaging. 2013;26:989–94.
10. Elkin PL, Froehling D, Wahner-Roedler D, Trusko B, Welsh G, Ma H, Asatryan AX, Tokars JI, Rosenbloom ST, Brown SH. NLP-based identification of pneumonia cases from free-text radiological reports. AMIA Annu Symp Proc. 2008:172–176.
11. Navitski A, Goyal P, Ahsanuddin S, Zheng S, Joffe E. Automated identification of lymphoma involving the bone from PET/CT reports using natural language processing and adaptive learning. J Clin Oncol. 2020;38:e19201–19201.
12. Yim WW, Kwan SW, Johnson G, Yetisgen M. Classification of hepatocellular carcinoma stages from free-text clinical and radiology reports. AMIA Annu Symp Proc. 2017:1858–1867.
13. Chen MC, Ball RL, Yang L, Moradzadeh N, Chapman BE, Larson DB, Langlotz CP, Amrhein TJ, Lungren MP. Deep learning to Classify Radiology Free-text reports. Radiology. 2018;286:845–52.
14. Batch KE, Yue J, Darcovich A, Lupton K, Liu CC, Woodlock DP, El Amine MAK, Causa-Andrieu PI, Gazit L, Nguyen GH, Zulkernine F, Do RKG, Simpson AL. Developing a Cancer Digital Twin: supervised metastases Detection from Consecutive Structured Radiology reports. Front Artif Intell. 2022;5.
15. Wood DA, Kafiabadi S, Al Busaidi A, Guilhem EL, Lynch J, Townend MK, Montvila A, Kiik M, Siddiqui J, Gadapa N, Benger MD, Mazumder A, Barker G, Ourselin S, Cole JH, Booth TC. Deep learning to automate the labelling of head MRI datasets for computer vision applications. Eur Radiol. 2022;32:725–36.
16. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Adv Neural Inf Process Syst. 2017;30.
17. Datta S, Roberts K. Fine-grained spatial information extraction in radiology as two-turn question answering. Int J Med Inf. 2022;158:104628.
18. Zaman S, Petri C, Vimalesvaran K, Howard J, Bharath A, Francis D, Peters NS, Cole GD, Linton N. Automatic diagnosis labeling of Cardiovascular MRI by using Semisupervised Natural Language Processing of text reports. Radiol Artif Intell. 2022;4:e210085.
19. Nishigaki D, Suzuki Y, Wataya T, Kita K, Yamagata K, Sato J, Kido S, Tomiyama N. BERT-based transfer learning in sentence-level anatomic classification of Free-text Radiology Reports. Radiol Artif Intell. 2023;5:e220097.
20. Sykes D, Grivas A, Grover C, Tobin R, Sudlow C, Whiteley W, McIntosh A, Whalley H, Alex B. Comparison of rule-based and neural network models for negation detection in radiology reports. Nat Lang Eng. 2021;27:203–24.
21. Tan R, Lin Q, Low GH, Lin R, Goh TC, Chang CCE, Lee FF, Chan WY, Tan WC, Tey HJ, Leong FL, Tan HQ, Nei WL, Chay WY, Tai DWM, Lai GGY, Cheng LT, Wong FY, Chua MCH, Chua MLK, Tan DSW, Thng CH, Tan IBH, Ng HT. Inferring cancer disease response from radiology reports using large language models with data augmentation and prompting. J Am Med Inf Assoc. 2023;30:1657–64.
22. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. Preprint at arXiv:181004805.
23. Huemann Z, Lee C, Hu J, Cho SY, Bradshaw TJ. Domain-adapted large Language models for Classifying Nuclear Medicine reports. Radiol Artif Intell. 2023;5:e220281.
24. Mithun S, Jha AK, Sherkhane UB, Jaiswar V, Purandare NC, Rangarajan V, Dekker A, Puts S, Bermejo I, Wee L. Development and validation of deep learning and BERT models for classification of lung cancer radiology reports. Inf Med Unlocked. 2023;40:101294.
25. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog. 2019;1:9.
26. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural Language Processing (almost) from scratch. J Mach Learn Res. 2011;12:2493–537.
27. Pilault J, Elhattami A, Pal C. Conditionally adaptive multi-task learning: Improving transfer learning in nlp using fewer parameters & less data. 2020. Preprint at arXiv:200909139.
28. Chen S, Zhang Y, Yang Q. Multi-task learning in natural language processing: An overview. 2021. Preprint at arXiv:210909138.
29. Eyuboglu S, Angus G, Patel BN, Pareek A, Davidzon G, Long J, Dunnmon J, Lungren MP. Multi-task weak supervision enables anatomically-resolved abnormality detection in whole-body FDG-PET/CT. Nat Commun. 2021;12:1880.
30. Park HJ, Park N, Lee JH, Choi MG, Ryu J-S, Song M, Choi C-M. Automated extraction of information of lung cancer staging from unstructured reports of PET-CT interpretation: natural language processing with deep-learning. BMC Med Inf Decis Mak. 2022;22:229.
31. Nobel JM, Puts S, Krdzalic J, Zegers KML, Lobbes MBI, Robben F, Dekker SG. Natural Language Processing Algorithm used for staging Pulmonary Oncology from Free-text Radiological reports: including PET-CT and validation towards clinical use. J Imaging Inf Med. 2024;37:3–12.
32. 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 Form Res. 2023;7:e38125.
33. Nakamura Y, Hanaoka S, Yada S, Wakamiya S, Aramaki E. NTCIR-17 MedNLP-SC Radiology Report Subtask overview: dataset and solutions for automated Lung Cancer Staging. Proc NTCIR. 2023;–17:17:145–51.
34. Fukushima T, Otsuki Y, Yada S, Wakamiya S, Aramaki E. NAISTSOCRR at the NTCIR-17 MedNLP-SC Radiology Report Subtask. Proc NTCIR. 2023;–17:17:163–6.
35. Nishio M, Matsuo H, Matsunaga T, Fujimoto K, Rohanian M, Nooralahzadeh F, Rinaldi F, Krauthammer M. Zero-shot classification of TNM staging for Japanese radiology report using ChatGPT at RR-TNM subtask of NTCIR-17 MedNLP-SC. NTCIR-17. 2023;17:155–62.
36. Fujimoto K, Nishio M, Tanaka C, Rohanian M, Nooralahzadeh F, Krauthammer M, Rinaldi F. Classification of cancer TNM stage from Japanese radiology report using on-premise LLM at NTCIR-17 MedNLP-SC RR- TNM subtask. NTCIR-17. 2023;17:200–7.
37. Lababede O, Meziane MA. The Eighth Edition of TNM staging of Lung Cancer: Reference Chart and diagrams. Oncologist. 2018;23:844–8.
38. Oudkerk M, Liu S, Heuvelmans MA, Walter JE, Field JK. Lung cancer LDCT screening and mortality reduction — evidence, pitfalls and future perspectives. Nat Rev Clin Oncol. 2012;18:135–51.
39. Niederkohr RD, Greenspan BS, Prior JO, Schöder H, Seltzer MA, Zukotynski KA, Rohren EM. Reporting Guidance for oncologic 18F-FDG PET/CT imaging. J Nucl Med. 2013;54:756–61.
40. Pencharz D, Wagner T. Actionable reporting versus unwanted advice in PET-CT reports. Clin Radiol. 2023;78:666–70.
41. Audi S, Pencharz D, Wagner T. Behind the hedges: how to convey uncertainty in imaging reports. Clin Radiol. 2021;76:84–7.
42. Johnson AEW, Pollard TJ, Shen L, Lehman L-wH, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035.

43. Moss C, Haire A, Cahill F, Enting D, Hughes S, Smith D, Sawyer E, Davies A, Zylstra J, Haire K, Rigg A, Van Hemelrijck M. Guy's cancer cohort – real world evidence for cancer pathways. BMC Cancer. 2020;20:187.
44. Honnibal M, Montani I. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. GitHub, 2017.
45. Rolnick D, Veit A, Belongie S, Shavit N. Deep learning is robust to massive label noise. 2017. Preprint at arXiv:170510694.
46. Landis JR, Koch GG. The measurement of Observer Agreement for Categorical Data. Biometrics. 1977;33:159–74.
47. Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH. Automatic identification of critical follow-up recommendation sentences in radiology reports. AMIA Annu Symp Proc. 2011:1593–1602.
48. Bradshaw T, Cho S. Evaluation of large language models in natural language processing of PET/CT free-text reports. J Nucl Med. 2021;62:1188–1188.
49. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K. Google's neural machine translation system: Bridging the gap between human and machine translation. 2016. Preprint at arXiv:160908144.
50. Gage P. A new algorithm for data compression. C Users J. 1994;12:23–38.
51. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Brew J. HuggingFace's Transformers: State-of-the-art Natural Language Processing. 2019. Preprint at ArXiv abs/1910.03771.
52. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L. Pytorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst. 2019;32.
53. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q. A comprehensive survey on transfer learning. Proc IEEE. 2020;109:43–76.
54. Zhang Y, Yang Q. A Survey on Multi-task Learning. IEEE Trans Knowl Data Eng. 2022;34:5586–609.
55. Zhang Y, Yang Q. An overview of multi-task learning. Natl Sci Rev. 2017;5:30–43.
56. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15:1929–58.
57. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901.
58. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21:5485–551.
59. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. Roberta: A robustly optimized bert pretraining approach. 2019 Preprint at arXiv:190711692.
60. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019;36:1234–40.
61. Shin H-C, Zhang Y, Bakhturina E, Puri R, Patwary M, Shoeybi M, Mani R. BioMegatron: Larger biomedical domain language model. 2020. Preprint at arXiv:201006060.
62. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, Compas C, Martin C, Costa AB, Flores MG, Zhang Y, Magoc T, Harle CA, Lipori G, Mitchell DA, Hogan WR, Shenkman EA, Bian J, Wu Y. A large language model for electronic health records. NPJ Digit Med. 2022;5:194.
63. Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, McDermott MBA. Publicly Available Clinical BERT Embeddings. 2019. Preprint at ArXiv abs/1904.03323.
64. Yan A, McAuley J, Lu X, Du J, Chang EY, Gentili A, Hsu C-N. RadBERT: adapting transformer-based language models to radiology. Radiol Artif Intell. 2022;4:e210258.
65. Kingma DP, Ba J. Adam: a method for stochastic optimization. In Proc. 3rd International Conference on Learning Representations (ICLR), 2015.
66. Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. J Doc. 1972;28:11–21.
67. Berkson J. Application of the logistic function to Bio-assay. J Am Stat Assoc. 1944;39:357–65.
68. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
69. Wood DA, Kafiabadi S, Busaidi AA, Guilhem E, Montvila A, Lynch J, Townend M, Agarwal S, Mazumder A, Barker GJ, Ourselin S, Cole JH, Booth TC. Deep learning models for triaging hospital head MRI examinations. Med Image Anal. 2022;78:102391.
70. Rohren EM. Positron Emission Tomography-computed tomography reporting in Radiation Therapy Planning and Response Assessment. Semin Ultrasound CT MR. 2010;31:516–29.
71. Alongi P, Stefano A, Comelli Al, Spataro A, Formica G, Laudicella R, Lanzafame H, Panasiti F, Longo C, Midiri F, Benfante V, La Grutta L, Burger IA, Bartolotta TV, Baldari S, Lagalla R, Midiri M, Russo G. Artificial Intelligence Applications on Restaging [18F]FDG PET/CT in metastatic colorectal Cancer: a preliminary Report of Morpho-Functional Radiomics classification for prediction of Disease Outcome. Appl Sci. 2022;12:2941.
72. Lovinfosse P, Polus M, Van Daele D, Martinive P, Daenen F, Hatt M, Visvikis D, Koopmansch B, Lambert F, Coimbra C, Seidel L, Albert A, Delvenne P, Hustinx R. FDG PET/CT radiomics for predicting the outcome of locally advanced rectal cancer. Eur J Nucl Med Mol Imaging. 2018;45:365–75.
73. Wang L, Zhang C, Li J. A hybrid CNN-Transformer Model for Predicting N staging and survival in Non-small Cell Lung Cancer patients based on CT-Scan. Tomography. 2024;10:1676–93.


## Publisher's note