University of Leeds
School of Computing

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of
Master of Philosophy (MPhil)

Multitask Learning with BERT for Automated Completion of the Breast Cancer Pathology COSD

Author: Sarah Leanne Miller
Date: September 2026

Intellectual Property and Publication Statements

I confirm that the work submitted is my own and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The right of Sarah Leanne Miller to be identified as the author of this work has been asserted by her in accordance with the Copyright, Designs and Patents Act 1988.

Acknowledgements

This research was made possible through the support of my supervisors, Dr Serge Sharoff and Prof Geoffrey Hall, whose guidance, insight, and encouragement have been invaluable throughout this project. I am grateful to the Genomics England team for providing access to the pathology dataset that underpinned this study, and to the University of Leeds Doctoral College for its continued support for my academic and professional development.

I acknowledge the use of ChatGPT-5 (OpenAI, https://chat.openai.com/) and Gemini 2.5 Flash, to assist with proofreading and formatting consistency checks in the final thesis draft, in accordance with the University's Generative AI Guidance for Postgraduate Researchers (2024).

Finally, I want to sincerely thank my family and friends for their patience, encouragement, and understanding during this academic journey's challenges. I especially cherish my little dog Peggy, whose companionship made each day brighter.

Abstract

Breast cancer pathology reporting is crucial for clinical decision-making and national cancer registration in the UK, and it contributes to the Cancer Outcomes and Services Dataset (COSD). Currently, manual extraction of key variables from free-text reports causes data quality issues and delays. This thesis addresses these challenges by developing a novel Multitask BERT (MT-BERT) framework for automated extraction of structured COSD breast pathology attributes. The proposed model combines a BioMedBERT encoder with specialised classification and regression heads to predict seven variables, including histology, grade, and continuous tumour size. To tackle severe class imbalance and missing labels in a real-world dataset of 7673 reports for 2455 breast cancer patients, the model employs a Round-Robin head-balanced sampling method and a sophisticated loss function. This combines Class-Balanced Cross-Entropy with Label-Distribution-Aware Margins (LDAM) and a center Logit Adjustment technique specifically designed for rare histological subtypes. Evaluated on an independent test set, the MT-BERT achieved a mean Macro-F1 score of 0.5566 across all classification tasks, with laterality reaching production-ready accuracy and other variables, including grade, showing promise, though not yet meeting clinical decision-support thresholds. The performance on minority classes in Histology Coded reached a Macro-F1 of 0.5179, partially validating anti-collapse strategies. For the continuous Tumour Size variable, predictions yielded a MAE of 12.23 mm on the original scale. The framework demonstrates that a unified, data-aware multitask transformer can effectively replicate structured COSD reporting, offering a parameter-efficient approach with a single shared encoder in place of seven separate models to support data completeness within the NHS.

Table of Contents

Intellectual Property and Publication Statements *2*

Acknowledgements *3*

Abstract **4**

Tables **8**

Figures **8**

1 Introduction *9*

1.1 Background **9**

1.2 Research Problem and Aims **10**

1.3 Research Questions **10**

1.4 Novelty and Contributions **11**

1.5 Thesis Structure **12**

2 Literature Review *12*

2.1 Evolution of Cancer Data Reporting: Exploring the Requirements for Clinical Text Classification **12**

2.2 Foundational Models: From BERT to Domain Adaptation **14**

2.2.1 Domain-Specific BERT Models for Healthcare and Oncology 14

2.3 Multitask Learning in Biomedical NLP **15**

2.3.1 Empirical Evidence for MTL with BERT 15

2.4 Handling Class Imbalance and Evaluation in Clinical NLP **16**

2.4.1 Optimising Loss Functions for Imbalanced Data 16

2.4.2 Evaluating Performance under Imbalance 17

2.5 Summary and Identified Research Gaps **18**

3 Methodology *18*

3.1 Dataset **19**

3.1.2 Multi-Task Learning Targets 19

3.2 Text preprocessing **20**

3.2.1 Label encoding and preprocessing 21

3.3 Multi-Task BERT Architecture 21

3.4 Task-specific prediction heads 22

3.5 Multitask Learning Framework 23

3.6 Multi-Task Training Strategy 23

3.6.1 Sampling Strategy 23

3.6.2 Two-Phase Training Protocol 24

3.6.3 Class Imbalance Mitigation 25

3.6.4 Regularization Techniques 26

3.6.5 Loss Function Formulation 26

3.6.6 Evaluation Metrics 27

4 Results 28

4.1 Overall Model Performance 28

4.1.1 Dataset Characteristics and Training Progression 28

4.1.2 Aggregated Test Set Performance 31

4.1.3 Validation-Test Generalisation 32

4.2 Per-Task Analysis 33

4.2.1 Histology Classification 33

4.2.2 Tumour Grade Classification 34

4.2.3 Laterality Prediction 35

4.2.4 Hormone Receptor Status Predictions (HER2, ER, PR) 36

4.2.5 Tumour Size Regression 37

4.3 Class Imbalance Mitigation: Impact of Advanced Loss Strategies 38

4.3.1 Summary of Imbalance Handling Effectiveness 38

4.3.2 Quantitative Impact Assessment 39

4.5 Research Questions Addressed 39

4.5.1 Primary Research Question 39

4.5.2 Secondary Research Questions 40

4.5.3 Summary 41

5 Discussion 42

5.1 Overview of Findings 42

5.2 Interpretation in the Context of Prior Work 42

5.2.1 Effectiveness of the Multitask Framework 42

5.2.2 Class-Imbalance Mitigation and Loss Design 42
5.2.3 Regression Head Performance 43
5.2.4 Computational Efficiency 43
5.2.5 Ethical and Data-Governance Considerations 43
5.2.5 Deeper Discussion of Negative Results 44
5.3 Limitations 46
5.4 Future Work 46
5.4.1 Explainability and attribution analysis 46
5.4.2 Alternative architectures for multi-task classification 46
5.4.3 Patient-level splitting and multi-institutional cohorts 47
5.4.4 Addressing the imbalance ceiling 47
5.4.5 Improving the regression head 47
5.4.6 Multimodal integration 48
5.5 Chapter Summary 48
6 Conclusion 49
References 50

Tables

Table 1. Dataset Features 20
Table 2. Dataset Distribution Across Tasks 29
Table 3. Test Set Performance Summary - All Tasks 31
Table 4. Validation vs Test Set Performance Comparison 32
Table 5. Histology Classification - Test Set Per-Class Performance 33
Table 6. Grade Classfication Test Set Per-Class Performance 34
Table 7. Lateratlity Classification Test Set Per-Class Performance 35
Table 8. HER2 Status Test Set Per-Class Performance 36
Table 9. ER Status Test Set Per-Class Performance 36
Table 10. PR Status Test Set Per-Class Performance 36
Table 11. Tumour Size Test Set Prediction Performance 37
Table 12. Test Set Minority Class Recall Summary 39

Figures

Figure 1. Multi-Task Framework 22

1 Introduction

1.1 Background

Breast cancer is the most common malignancy among women worldwide and remains a leading cause of cancer-related mortality (World Health Organization, 2013). Pathology reporting plays a crucial role in diagnosis, prognostication, and treatment planning, serving as the basis for multidisciplinary decision-making across cancer services. The free-text pathology report created by the pathologists contains structured information, vital variables that can directly influence decisions about available therapeutic pathways for the patient.

In the United Kingdom (UK), these structured pathology data attributes contribute to the Cancer Outcomes and Services Dataset (COSD) (NHS England NDRS, 2025). The nationally mandated standard for collecting cancer data and submitting it to registries (NHS England and National Disease Registration Service (NDRS), 2024b). The COSD offers a unified schema that links clinical, pathological, and treatment information to support national audits, commissioning, and research. Specifically for breast cancer, pathology fields such as histologic subtype, behaviour, grade, receptor status, and tumour size are vital for assessing quality indicators and monitoring breast cancer patient outcomes.

Despite national standardisation efforts, data entry remains primarily manual and relies on human interpretation of narrative pathology reports. This manual approach leads to variability in completeness and accuracy across providers, delays in registry submissions, and increases workload for pathology and registry staff (Henson et al., 2020). As the NHS moves towards fully digital workflows, automating COSD-compatible data extraction from unstructured pathology texts has become essential to improve data quality, increase efficiency, and enhance the potential for secondary research. Natural-language-processing (NLP) approaches, particularly transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), have the potential to provide this automation because they can capture contextual relationships between medical entities and replicate the interpretive reasoning performed by human coders.

Models that could deliver a solution to the automation of structured data extraction from free-text pathology reports include domain-specific BERT variants, such as BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), and CancerBERT (Zhou et al., 2022), which have shown improved performance on biomedical texts by incorporating in-domain vocabulary and semantics. Using these models for COSD pathology reporting could improve automation accuracy and data completeness, thereby contributing to a more robust data-collection solution within NHS reporting frameworks.

1.2 Research Problem and Aims

Although transformer models have revolutionised biomedical NLP, a clear gap remains between current research and the requirements of national cancer-registry reporting. Most published studies address isolated extraction tasks and focus on a single feature, e.g., histology, receptor status, or grade, without modelling the biological and contextual dependencies among them. Furthermore, existing systems rarely align directly with the COSD specification, limiting their interoperability with UK registry infrastructure.

Key methodological gaps include:
- Fragmented task modelling—tasks trained independently, missing inter-variable relationships, along with an increased computational demand and complexity of running multiple independent models.
- Limited handling of label imbalance and missing data, common in real-world pathology corpora.
- Neglect of regression outputs, such as continuous tumour size.
- Restricted generalisability due to single-task fine-tuning and small datasets.
- Poor reproducibility, with limited open-source frameworks suitable for NHS integration.

This thesis addresses these limitations by developing a novel end-to-end Multitask BERT (MT-BERT) framework for automated completion of the breast-pathology components required for the COSD. The proposed model unifies classification and regression heads for key COSD variables, applies data-aware loss functions (class-balanced, focal, label-smoothed, and Huber) with masked optimisation to handle missing labels, and employs a phased training schedule that stabilises learning through discriminative learning rates and optional uncertainty-based task weighting.

The overarching aim of this research is to evaluate how effectively a multitask transformer framework can replicate the structured reporting required for COSD, improving automation while preserving clinical validity and interpretability.

1.3 Research Questions

Primary Research Question
How effectively can a Multitask BERT framework be applied to automate the structured completion of the UK COSD breast-cancer pathology dataset from unstructured pathology reports while maintaining clinical validity and interpretability?

Here, interpretability refers to the clinical usability of the structured outputs, with values directly mappable to COSD fields that a pathologist can read and verify, rather than to model-internal explainability methods, which are treated separately as future work (Section 5.4.1).

Secondary Research Questions

Model Architecture and Task Design

1. How can heterogeneous targets—multi-class, binary, ordinal, and regression—be jointly modelled within a single transformer-based framework without degrading per-task performance?
2. How can architectural and training design choices (e.g., shared encoder with specialised heads) support effective multitask learning across diverse COSD variables?

Loss Function Engineering and Optimisation

3. Which per-task loss configurations (class-balanced, focal, label-smoothed, or Huber) yield the most stable optimisation under class imbalance and label missingness?
4. How do dynamic task-weighting strategies (e.g., uncertainty-based or gradient-normalised) affect convergence and inter-task interference?
5. Does staged fine-tuning—progressing from head-warm-up to partial unfreeze and full joint training—enhance stability and calibration?

Data Quality and Calibration

6. How can per-head loss masking mitigate bias from missing labels while preserving training data?
7. What is the relationship between class-frequency distribution, loss weighting, and predictive calibration across COSD pathology variables?

1.4 Novelty and Contributions

This research delivers a technically and clinically novel framework that bridges computational pathology NLP and national-registry automation:

1. Unified multitask learning architecture: a shared BERT encoder jointly predicting histology, grade, laterality, ER/PR/HER2 status, and tumour size from narrative pathology text.
2. Per-head, data-aware loss functions: modular losses tailored to imbalance, calibration, and missingness, incorporating selective label masking.
3. Dynamic, multi-phase optimisation: a head-warm-up, partial-unfreeze, and joint-training regimen using discriminative learning rates and optional uncertainty weighting.
4. COSD-aligned evaluation framework: introduces registry-relevant metrics—record-level completeness, cross-field consistency, derived-stage accuracy, and calibration (ECE, Brier score)—alongside traditional statistical measures.
5. Robustness under imbalance and sparsity: empirical evidence that adaptive loss weighting and label smoothing enhance performance on rare histologic and receptor categories.
6. Reproducible, modular pipeline: a fully parameterised PyTorch framework integrating data preparation, multitask training, and evaluation, enabling replication and adaptation for other tumour sites.

1.5 Thesis Structure

The remainder of this thesis is organised as follows:
- Chapter 2 – Literature Review surveys the evolution of cancer-data reporting, the development of BERT and its domain-specific adaptations, multitask learning frameworks in biomedical NLP, and strategies for addressing class imbalance and model evaluation in clinical contexts.
- Chapter 3 – Methods details the COSD dataset, preprocessing pipeline, multitask BERT architecture, loss-function engineering, evaluation metrics, and implementation design.
- Chapter 4 – Results presents quantitative findings per-head analysis of classification and regression outputs.
- Chapter 5 – Discussion interprets the results considering the literature, examining implications for clinical use, robustness under imbalance, and potential extensions.
- Chapter 6 – Conclusion summarises the key findings, answers the research questions, and outlines avenues for future work in automated pathology reporting.

2 Literature Review

Natural Language Processing (NLP) has rapidly progressed with the adoption of transformer-based models. One particular model architecture, BERT (Bidirectional Encoder Representations from Transformers), has set new standards across numerous NLP tasks. Cutting-edge BERT-based models have significantly enhanced text classification capabilities across various domains and hold the same potential to revolutionise text classification tasks in healthcare, where the complexity of data demands specialised approaches and advanced training methods. The literature review in this study highlights recent advances in clinical NLP, focusing on the development of domain-specific variants of BERT, the application of advanced optimisation strategies, and the rigorous evaluation metrics necessary for clinical NLP.

2.1 Evolution of Cancer Data Reporting: Exploring the Requirements for Clinical Text Classification

Accurate and standardised cancer data collection underpins national efforts to monitor outcomes, plan services, and support research. In the UK, structured cancer data reporting has led to the need for and the creation of a national, unified dataset, the Cancer Outcomes and Services Dataset (COSD) (NHS England and National Disease Registration Service (NDRS), 2024). The COSD, as of 2013, is the nationally mandated approach for collecting cancer data for cancer services across the UK (Henson et al., 2020; NHS England and National Disease Registration Service (NDRS), 2024b).

UK cancer data registration dates to the 1930s, when regional registries began manually recording new cancer diagnoses. These early systems consisted of paper records and practitioners' willingness to voluntarily report their data (Wagner, 2002). In the 1970s and 1980s, the National Cancer Registration Scheme and later the Office for National Statistics (ONS)

registry services improved data consistency. However, manual data extraction from clinical records was still required (Wagner, 2002; Henson et al., 2020).

In the early 2000s, the increasing digitisation of medical records created the need for more comprehensive electronic data capture, which led to initiatives such as the National Cancer Dataset (NCDS) and the Systematic Anti-Cancer Therapy (SACT), both of which aimed to standardise the collection of data attributes across tumour types (NHS England and National Disease Registration Service (NDRS), 2024b). However, reporting remained inconsistent due to local variations and submission practices (Henson et al., 2020) and this continued fragmentation has limited reliable data linkage required for national research (Strongman et al., 2020; Whitfield et al., 2024).

In 2013, NHS England and the National Cancer Registration and Analysis Service (NCRAS) launched the COSD as the national standard for cancer data reporting across all tumour types. The COSD replaced the NCDS by providing a unified schema that combined core patient, diagnostic, treatment, and outcome data into a single repository. In 2020, the COSD evolved further into two data schemas: one for core data attributes and another for pathology-specific attributes (NHS England and National Disease Registration Service (NDRS), 2024b). The rationale for this split in reporting was that specialised pathology data collection would best be left to the pathologist, resulting in more accurate data. This move was also said to align with other clinical pathology reporting requirements, reducing the administrative burden associated with collecting the required pathology data items required for COSD.

The rationale for automating the COSD in place of the current manual data collection aligns with the NHS's digital transformation strategy. The NHS's long-term plan (NHS England, 2019) emphasises the importance of real-time data capture to improve cancer outcomes. NCRAS has identified free-text pathology data capture and reporting as a pain point, contributing to a major source of incomplete or delayed data uploads. As pathology departments move toward fully digital workflows, automated NLP systems capable of extracting COSD-compatible variables from narrative text could dramatically enhance the completeness, consistency, and timeliness of registry submissions.

Emerging NLP approaches, particularly those based on BERT-style contextual embeddings, offer a promising mechanism to achieve this automation. By learning contextual relationships between entities such as histologic type, grade, and receptor status, transformer models can replicate the semantic reasoning traditionally performed by human abstractors. Integrating such models into COSD workflows could reduce manual effort and improve data fidelity across the national registry infrastructure (NHS England and National Disease Registration Service (NDRS), 2024).

2.2 Foundational Models: From BERT to Domain Adaptation

The original BERT model created by Devlin et al. (2019) featured a robust bidirectional transformer architecture (Vaswani et al., 2017) that was trained on a book corpus consisting of billions of books and the entire English Wikipedia by applying masked language modelling (MLM) and next sentence prediction (NSP) tasks to this extensive dataset. The bidirectionality of BERT, combined with a comprehensive training corpus, enabled the model to learn contextual representations that surpassed those of previous NLP models, leading it to set SOTA results across many NLP tasks (Rogers et al., 2020). However, while BERT achieved these superior results, its vocabulary proved to be too generic for more specialised tasks in scientific domains.

The first to identify the limitations of BERT's vocabulary for scientific NLP tasks were Beltagy et al. (2019). Thus, they extended BERT's pretraining to include scientific literature by using 1.14 million papers from Semantic Scholar. Training BERT with a scientific vocabulary and creating SciBERT improved results on scientific document classification tasks compared with the generic baseline. The approach in this study demonstrated that domain-adaptive pretraining could improve language model performance without requiring architectural changes.

The success of this study was quickly extended to clinical and biomedical fields with the development of ClinicalBERT by Alsentzer et al. (2019), which adapted BERT using the MIMIC-III text reports along with an electronic health record (EHR) corpus. Additionally, BioBERT by Lee et al. (2020), trained on PubMed abstracts and PMC full-text articles, achieved state-of-the-art results in biomedical NER, relation extraction, and question answering. Both models similarly improved clinical NLP tasks compared to standard BERT models' vocabulary. Similarly, a study by Gu et al. (2021) developing BioMedBERT found that, when developing contextual embeddings for the biomedical domain, pretraining coupled with finetuning yielded superior results, outperforming both SciBERT and BioBERT.

These works all employed domain adaptation techniques and showed that ongoing pretraining on in-domain corpora improves transferability to specialised scientific tasks, such as clinical concept extraction. Collectively, these studies support the rationale that pretrained models trained on domain-specific corpora are best suited for scientifically focused NLP pipelines.

2.2.1 Domain-Specific BERT Models for Healthcare and Oncology

Subsequent research in this area has further aimed to tailor BERT to specific sub-domains of healthcare and medicine. Med-BERT by Rasmy et al. (2021) was pre-trained on structured EHR data rather than free text, utilising the tabular EHR data for disease prediction. Its contextual embeddings improved prediction accuracy for chronic disease progression, demonstrating that BERT-based architectures can effectively model temporally ordered clinical data and free text.

Focusing more narrowly on oncology, CancerBERT by Zhou et al. (2022) fine-tuned BioBERT on breast cancer electronic health records to extract phenotypic variables, including histologic type and hormone receptor statuses (ER, PR, HER2). By embedding cancer-specific

terminology and ontological mappings (e.g., ICD-O-3 codes), CancerBERT achieved superior performance in phenotype extraction compared to general biomedical models. This work is particularly relevant to automating completion of the Cancer Outcomes and Services Dataset (COSD) in the UK, as it demonstrates the capacity of BERT-based models to capture nuanced histopathological descriptors; however, the model is not publicly available.

Similarly, Ling et al. (2019) used pretrained language model embeddings to identify metastatic breast cancer cohorts by linking registry and EHR data. Although they do not specifically mention using BERT embeddings, their study highlights the potential for clinical NLP to bridge structured and unstructured data sources, enabling large-scale secondary analyses.

Collectively, these models confirm that domain-specific fine-tuning improves clinical language understanding and that oncology-specific pretraining enhances downstream tasks such as pathology classification, staging, and receptor status prediction.

2.3 Multitask Learning in Biomedical NLP

Traditional methods for fine-tuning BERT in the Biomedical domain involve training a separate model for each downstream task, which is computationally costly and fails to leverage potential relationships among tasks. Multi-Task Learning (MTL) provides an alternative approach in which a single model is trained to simultaneously perform multiple tasks, all using the same model backbone. It is stated that sharing representations across tasks in MTL can capture higher-level semantic patterns, enhance generalisation, and reduce reliance on large task-specific datasets, which are especially valuable in data-scarce fields like healthcare and medicine.

2.3.1 Empirical Evidence for MTL with BERT

Empirical studies show that multitask variants of BERT can achieve measurable improvements in both performance and efficiency in biomedical NLP. In their paper, Peng et al. (2020) systematically examined MTL on BERT across a range of biomedical and clinical NLP tasks, including named-entity recognition, relation extraction, textual entailment, and semantic similarity. They found that using a shared encoder with task-specific output heads saw consistent gains of around 2% in biomedical tasks and 1.3% in clinical tasks compared to single-task baselines. Pairwise task analysis indicated that semantically related objectives, such as entity recognition and relation extraction, mutually strengthened each other, whereas unrelated tasks sometimes led to negative transfer. The research in this study provided one of the first quantitative validations of multitask BERT within biomedical text processing.

In their study, Mulyar et al. (2021) developed MT-Clinical BERT, a multitask model for extracting clinical information. Their framework combined eight tasks—named-entity recognition, personal health information detection, textual entailment, and semantic similarity — using a single encoder. Although the accuracy for each task declined slightly compared to

individual fine-tuned models, the multitask setup significantly reduced inference time and simplified deployment by replacing multiple specialised models with a single generalisable system. This work demonstrated the practical viability of multitask architectures for clinical production pipelines, where scalability and inference efficiency are crucial.

Research by Moscato et al. (2023) applied MTL principles to few-shot biomedical relation extraction, a setting that reflects the low-sample conditions common in rare disease and pathology datasets. Their model, based on a transformer encoder with shared multitask heads and enhanced by knowledge distillation techniques, achieved state-of-the-art accuracy under extreme data scarcity (10–50 examples per relation). The authors demonstrated that MTL improved data efficiency by transferring inductive biases from related relation-extraction tasks, reducing overfitting and increasing stability.

The work in these studies has highlighted the applicability of these methods for automating breast cancer pathology reporting. Tasks such as histologic subtype classification, tumour grading, and receptor status prediction are semantically and biologically related. Training a single model to predict these features jointly could encourage cross-task information sharing, thereby improving feature learning for minority or rare labels by transferring learned context from more frequent patterns.

In addition to the potential benefits of shared learning across tasks, multitask architectures offer engineering efficiencies. A unified BERT-based model can replace multiple independent classifiers, reduce computational costs and simplify clinical deployment pipelines within NHS systems. In the context of the Cancer Outcomes and Services Dataset (COSD), such models could enable near-real-time extraction of key variables directly from narrative pathology reports, thereby improving data completeness and timeliness while maintaining interpretability.

2.4 Handling Class Imbalance and Evaluation in Clinical NLP

A persistent challenge in clinical and pathology NLP is class imbalance**,** where specific disease types, receptor statuses, and or pathological patterns occur far less frequently than others. In breast cancer pathology, for example, some carcinomas can dominate, whereas other rare tumour types appear much less often. These disparities bias learning algorithms toward the majority classes, leading to deceptively high accuracy but poor recall for minority and clinically critical categories. Since automated cancer reporting directly affects audit, treatment, and research outcomes, addressing imbalances and evaluating models is essential, as errors can have devastating consequences.

2.4.1 Optimising Loss Functions for Imbalanced Data

A common approach to handling imbalanced datasets is under- or oversampling the data to balance the dataset. Unfortunately, whilst this method can offer improved results, it can lead to a

model that overfits the training data and that does not generalise to new unseen data. Approaches to handling an imbalanced dataset with BERT models aim to adapt the loss function rather than the dataset. The standard cross-entropy loss function used in multiclass classification tasks treats all samples equally, allowing the majority classes to dominate gradient updates. To counter this, researchers have proposed several improved loss formulations.

In the study by Lin et al. (2017), they use Focal Loss to dynamically down-weight well-classified examples and focus training on complex or misclassified samples through a modulating factor $(1 - p_t)^{\gamma}$. They say that this approach enhances sensitivity to minority classes and has been widely adopted in medical imaging and clinical NLP and variants of this in other studies further refine balance: Asymmetric Focal Loss (Ridnik *et al.*, 2021) introduces separate focusing parameters for positive and negative labels, improving multi-label classification; Label-Distribution-Aware Margin (LDAM) Loss (Cao *et al.*, 2019) adjusts decision margins inversely to class frequency, yielding more equitable boundaries.

Class-Balanced Loss, as presented in Cui et al. (2019), is said to offer a mathematically justified weighting scheme based on the effective number of samples, reflecting the diminishing value of redundant examples. They state that it stabilises training without requiring ad-hoc sampling and has been effectively used in both medical text and image tasks. Combining these with uncertainty-based weighting (Cipolla et al., 2018) could enable joint scaling of multiple task losses that are crucial for multitask BERT systems that predict histologic subtype, grade, and receptor status simultaneously. Dynamic schemes, such as logit adjustment highlighted in Menon et al. (2021) and distribution alignment in Zhang et al. (2022), further adapt loss weighting during training to maintain balanced optimisation as class frequencies change.

Together, these methods could reduce bias toward dominant labels, thereby enhancing the detection of rare but clinically essential data entities.

2.4.2 Evaluating Performance under Imbalance

While appropriate loss functions can improve training behaviour for imbalanced datasets, applicable evaluation metrics are required to accurately reflect clinical priorities rather than merely capture the model's statistical performance. Aggregate scores, such as accuracy and micro-F1, obscure class disparities, making them unreliable for healthcare applications. The study by Blagec et al. (2022) highlights inconsistencies in NLP metric reporting and advocates for task-specific, interpretable measures aligned with real-world consequences. Similarly, Vickers et al. (2024) call for evaluation frameworks that account for robustness, calibration, bias, and fairness, particularly across domains or institutions. They recommend that for imbalanced datasets, macro-F1 and weighted-F1 provide fairer summaries by balancing class contributions, while per-class recall and precision expose clinically consequential blind spots. The use of confusion matrices can reveal which tumour types are misidentified, e.g., infiltrating versus in situ carcinoma, guiding targeted improvements. Beyond classification accuracy, calibration

metrics such as Brier Score and Expected Calibration Error (ECE) assess how well predicted probabilities reflect true outcome frequencies, a prerequisite for clinical trust. Area under the Precision–Recall Curve (AUPRC) is particularly informative under skewed distributions, and Matthews Correlation Coefficient (MCC) or Cohen's Kappa provide balanced single-value indicators.

2.5 Summary and Identified Research Gaps

The literature reviewed for this study stresses that class imbalance remains a significant obstacle in clinical and pathology NLP, particularly in domains like breast cancer, where rare tumour types are often underrepresented. Recent advances have focused on optimising loss functions, including Focal Loss, Asymmetric Focal Loss, LDAM, Class-Balanced Loss, and uncertainty-based weighting, to enhance sensitivity to minority classes without resorting to oversampling or undersampling, which can lead to overfitting. Dynamic loss adjustment methods and distribution alignment further contribute to balanced learning, supporting the accurate detection of clinically critical but infrequent categories.

For evaluation, the literature critiques the overreliance on aggregate metrics like accuracy and micro-F1, which can obscure poor model performance on minority classes. Instead, there is a consensus on the need for task-specific, interpretable metrics such as macro-F1, weighted-F1, per-class recall and precision, confusion matrices, calibration metrics (e.g., Brier Score, ECE), and robust single-value indicators like AUPRC, MCC, and Cohen's Kappa, to reflect real-world clinical priorities and model trustworthiness better.

Despite methodological progress, several gaps persist. First, there is a limited empirical comparison of loss function strategies across diverse clinical NLP tasks and datasets, making it unclear which approaches generalise best. Second, the integration of dynamic loss adjustment with multitask models (e.g., predicting histologic subtype, grade, and hormone receptor statuses simultaneously) remains underexplored. Third, few studies systematically link model calibration and interpretability to downstream clinical outcomes, and standardised evaluation frameworks for robustness, fairness, and domain adaptation remain lacking. Addressing these gaps would be beneficial for developing clinically reliable and generalizable NLP solutions in pathology and beyond.

3 Methodology

This section of the thesis details the methodological framework for automating the extraction of breast cancer pathology attributes for the Cancer Outcomes and Services Dataset (COSD). The proposed approach uses a Multitask BERT (MT-BERT**)** model and combines a domain-specific transformer encoder with task-specific heads, jointly optimised to predict seven key clinical variables. The design emphasises efficient representation sharing, effective handling of label imbalance, and clinically interpretable outputs.

3.1 Dataset

The dataset used in this study, provided by Genomics England, contains pathology reports and related feature attributes for 2455 breast cancer patients participating in the 100k Genome Project. Participants in the survey have consented to the use of their medical records for research and analysis, and their genomic data has been utilised in the 100k study. All data and documents used in this study were examined within a secure research environment, and only data with active consent were included in the analyses.

The dataset was partitioned using a nested hold-out procedure at the report level. An initial 80:20 split separated a held-out validation set, and the remaining data was further split 80:20 into training and test partitions, yielding an approximate 64:20:16 split across training, validation, and test sets (training n = 4,910; validation n = 1,535; test n = 1,228), with class proportions reported per task in Table 2. Stratified splitting was not possible because the multi-task, multi-label structure means no single label can define a stratified partition, and several minority classes (for example, intraductal carcinoma, at roughly 0.4% of the training data with only 19 examples) are too small to stratify reliably across three sets; as a result, class proportions vary somewhat between splits. While no single split ratio is theoretically optimal, ratios in the region of 70–80% training data with the remainder divided between validation and test are long-established conventions in statistical and machine learning (Hastie et al., 2009; James et al., 2013), reflecting a bias–variance trade-off: a larger training partition provides more data from which to learn, which is important given the limited size of this cohort, while the held-out validation and test partitions must remain large enough to give stable estimates of generalisation performance and to detect overfitting. The training set was used for parameter learning, the validation set for model selection (checkpointing on mean macro-F1) and hyperparameter decisions, and the test set was held out for a single final evaluation, ensuring an unbiased estimate of performance on unseen data.

3.1.2 Multi-Task Learning Targets

The data features shown in Table 1 were derived from structured and semi-structured pathology records for breast cancer participants in the 100k study. The model was trained to predict seven clinical attributes from the free-text pathology report simultaneously and includes six classification tasks (histology type with five classes, tumour grade with three classes, laterality with two classes, HER2 status with three classes, ER status with three classes, and PR status with three classes) and one regression task (tumour size in millimetres). The attributes in the dataset conform to COSD (Cancer Outcomes and Services Dataset) specifications and are documented in the COSD pathology data schema. Due to the nature of this small-cohort dataset, some features may have more categories; for example, there are more histology codes for breast cancer than the

five in this dataset. This dataset is not representative of the population, but the model's design allows feature sets to be extended with more classes as additional data becomes available.

Table 1. Dataset Features

Column Name	Total Records	Type	# Classes	Data Description
histology_coded	7673	Classification	5	5 unique codes for breast cancers for ductal and lobular neoplasms
grade	7635	Classification	3	The grade of the tumour G1, G2, G3
laterality	7673	Classification	2	Tumour laterality identifies the side of the body for a tumour relating to paired organs within a PATIENT based on the evidence from a pathological examination – right or left
er_status	7334	Classification	3	Oestrogen Receptor (ER) status – Negative, Positive, X – not performed
pr_status	5124	Classification	3	Progesterone Receptor (PR) status – Negative, Positive, X – not performed
her2_status	7081	Classification	3	HER2 Immunohistochemical status (Human Epidermal Growth Factor Receptor 2) - Positive, Negative, Borderline
tumoursize	6685	Regression	Size in mm	Size of tumour in mm

3.2 Text preprocessing

Text fields were cleaned to remove noise such as HTML and Markdown characters. All text strings were then converted to lowercase and tokenised using the BioMedBERT model tokenizer from the Hugging Face repository, specifically microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext. The pathology report text sequences were truncated from the left to retain the most recent and clinically relevant information, with a maximum sequence length of 512 tokens. The choice of model, tokenizer, and truncation method was based on findings from earlier research by Miller et al. (2024). In this study, it was observed that, likely due to a summary of findings at the end of each report, key features were placed at the end of the texts rather than at the beginning. This method, used alongside the BioMedBERT model, yielded the best overall results.

3.2.1 Label encoding and preprocessing

Feature labels in the dataset were numerically encoded, with categorical fields such as histology and receptor status mapped to integer indices. The continuous tumour size values were log-transformed using log(1+x) to stabilise variance (Han et al., 2022). As shown in Table 1, not all reports had labels available for every task. To handle label missingness, missing labels were masked with an ignore index (-100) during training via a batched dataloader, allowing the model to learn from partially labeled data. The loader's training batch size was set to 16 and 32 for training and evaluation, respectively.

3.3 Multi-Task BERT Architecture

The model architecture extends a pre-trained BioMedBERT encoder (Gu et al., 2021; Miller et al., 2024) with task-specific output heads for both categorical and continuous prediction. Each input sequence was processed by the 12-layer BioMedBERT transformer encoder. Rather than using the [CLS] token representation, a pooled sequence embedding was obtained via attention-masked mean pooling of the final-layer hidden states, that is, averaging the token embeddings while excluding padding positions via the attention mask, yielding a single 768-dimensional vector used as the shared latent representation for all downstream tasks. Mean pooling was chosen over [CLS] pooling, following Reimers and Gurevych (2019), who showed that averaged token representations yield more stable and semantically faithful sentence-level embeddings than the [CLS] token, which is optimised during pre-training for the next-sentence-prediction objective rather than for general sequence representation. This choice is well suited to pathology reports, where the COSD variables of interest (such as grade, histology, and receptor status) are expressed at varying positions throughout the text, so a representation that aggregates information across all tokens is preferable to one anchored to a single position. A dropout layer ($p = 0.1$) was applied to this pooled representation to mitigate overfitting. Following training, the best model was selected based on the mean macro-F1 across all classification heads on the validation set, and the model was restored from the checkpoint upon completion.

BioMedBERT (Gu et al., 2021) was selected as the encoder in preference to BioBERT (Lee et al., 2020) and ClinicalBERT (Alsentzer et al., 2019). BioBERT is fine-tuned from general-English BERT weights, retaining general-domain vocabulary that dilutes biomedical representations, whereas BioMedBERT is pre-trained from scratch on PubMed and PMC text. ClinicalBERT is trained on US MIMIC-III clinical notes, which differ substantially in style and content from UK breast pathology reports. BioMedBERT was also the strongest performer on this dataset in earlier work (Miller et al., 2024).

3.4 Task-specific prediction heads

From the shared 768-dimensional representation, seven independent linear projection layers produced task-specific outputs: six categorical classification heads (histology type: 768→5, tumour grade: 768→3, laterality: 768→2, HER2 status: 768→3, ER status: 768→3, PR status: 768→3) and one continuous regression head for tumour size (768→1). Each classification head outputs unnormalized logits corresponding to class probabilities, while the regression head produces a single continuous value. All prediction heads operate independently and in parallel, with no shared intermediate layers between the encoder and task outputs. Figure 1 provides an overview of the multitask model architecture, illustrating how a single shared BioMedBERT encoder processes pathology narratives through attention masked mean pooling to produce a unified representation that feeds seven parallel task-specific heads.

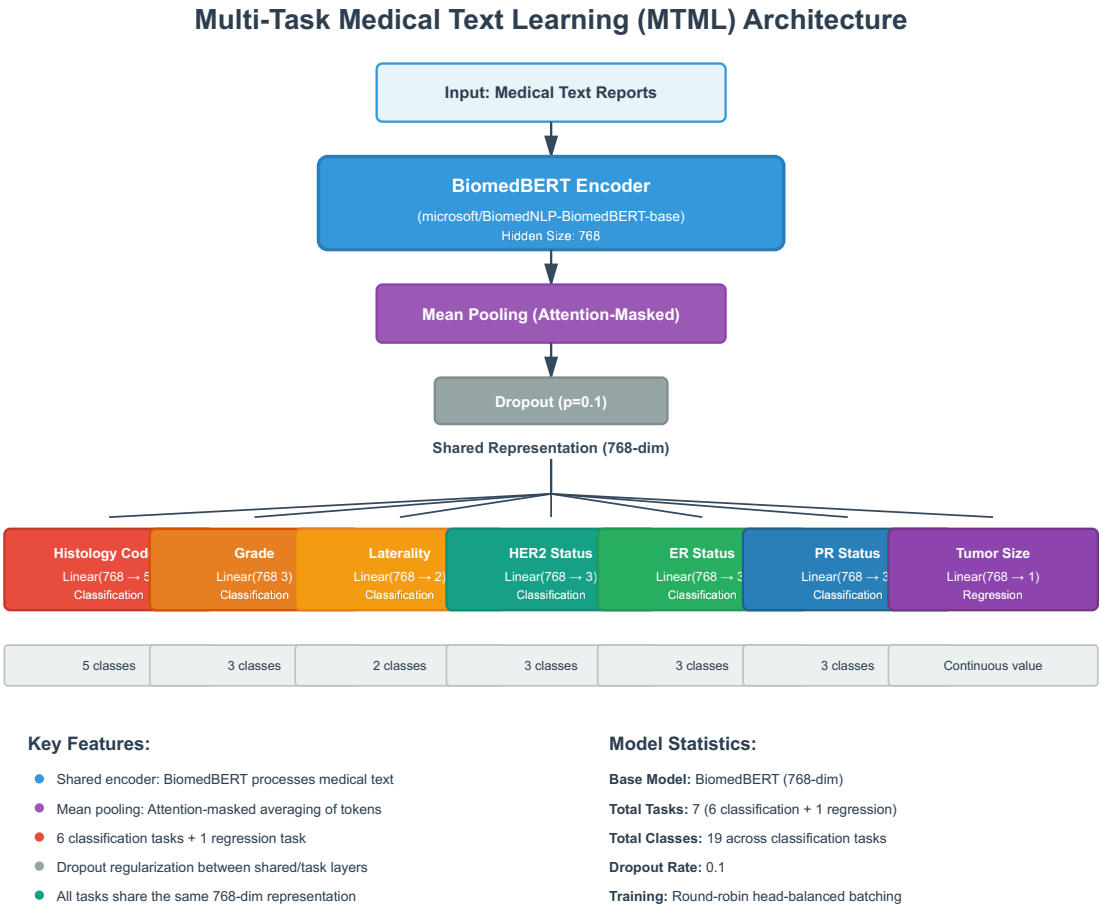


Figure 1. Multi-Task Framework

3.5 Multitask Learning Framework

The model design implements a hard parameter-sharing multitask framework (Caruana, 1997), allowing all tasks to share a common encoder while maintaining task-specific output layers. This design encourages the model to learn generalisable representations of pathology language by exploiting shared patterns across tasks (e.g., associations between histology and receptor expression). Multitask learning also acts as a regulariser by reducing the risk of overfitting to any single label distribution (Zhang and Yang, 2022).

Hard parameter sharing was adopted in preference to soft-sharing alternatives (such as cross-stitch networks) for two reasons: it is markedly more parameter-efficient, since a single encoder is shared across all seven heads rather than maintained per task, and, with a limited training set, the strong inductive bias of a fully shared encoder reduces the risk of overfitting that more flexible, higher-capacity sharing schemes would introduce. The trade-off that reduces capacity to accommodate tasks whose optimal representations diverge is examined in the discussion of negative transfer.

3.6 Multi-Task Training Strategy

This section describes all phases of the implementation, including hyperparameters, constants, and other values and settings required for training and evaluating the BERT MTM model.

3.6.1 Sampling Strategy

To address class imbalance and ensure balanced learning across all tasks, a round-robin head-balanced batch sampling strategy was implemented. This method cycles through the seven prediction tasks sequentially, dedicating each batch to a single task. Within each batch, samples are drawn to approximately balance class representation for that specific task. For instance, when training the grade classification head (which includes three classes), the sampler ensures roughly equal numbers of samples per grade category, preventing the majority class from dominating the batch. This sampling strategy avoids task neglect and provides minority classes with sufficient training signal, which is crucial in medical datasets where rare categories, such as specific histology subtypes or receptor status combinations, are clinically meaningful despite their low frequency.

3.6.2 Two-Phase Training Protocol

Model training was conducted in two sequential phases designed to refine the model's capabilities:

1. **Phase A: Simple Baseline Training (3 epochs)**:
   - **Loss Function:** Class-balanced cross-entropy (**β=0.995**) with label smoothing (**ε=0.05**) for classification; Weighted Huber loss (**δ=1.0**) with KDE-based density weighting (**bandwidth h=3.0**) for regression.
   - **Task Weighting:** Histology task upweighted by **3×**.
   - **Learning Rate:** 2×10^{-5} with linear warmup (5% of total steps) and linear decay.
   - **Optimization:** AdamW with weight decay 0.01.
   - **Mixed-precision training:** Enabled using PyTorch Automatic Mixed Precision (AMP) with float16 gradients and float32 accumulation.

These settings follow established practice for fine-tuning transformer encoders (Devlin et al., 2019; Gu et al., 2021); the learning rate of 2×10^{-5} in particular lies within the $\{2\times10^{-5}, 3\times10^{-5}, 5\times10^{-5}\}$ range recommended in the original BERT fine-tuning protocol recommended by Devlin et al. (2019). Three epochs per phase were chosen to balance convergence against overfitting given the relatively small training set.

2. **Phase B: Advanced Training with Anti-Collapse Mechanisms (3 epochs)**:
   - **Base Configuration:** Maintained from Phase A (**β=0.995**, **δ=1.0**).
   - **Enhanced Loss:** Incorporates **LDAM Margins** (**m_max=0.75**, scaled **2×** to **1.5** for histology) and **Logit Adjustment (LACE)**.
   - **Histology-Specific LACE:** Uses a **centered negative log-prior** adjustment (**τ=4.0**, scaling **α=0.2**) applied *only* to the histology head to prevent global logit shifts while enhancing minority class separation.
   - **Label Smoothing:** Task-specific (**ε=0.03** default; **ε=0.07** for histology).
   - **Gradient Clipping:** Max norm **1.0** to prevent exploding gradients.
   - **Optional warm entropy epoch:** Additional training epoch with entropy regularization (strength 0.3) to encourage prediction diversity.

3.6.3 Class Imbalance Mitigation

Multiple complementary strategies were used to address severe class imbalance:

1. Class-Balanced Loss Weighting (β=0.995)
Loss weights computed using the effective number of samples:

$$w_c = (1 - \beta) / (1 - \beta \hat{} n_c)$$

where **β=0.995** $and\ n_c$ is the count for class c. Weights normalized such that Σw_c = number of classes.

2. LDAM Margin Loss ($\boldsymbol{m_max = 0.75}$)
Class-dependent margins incorporated into logit space:

$$margin_c = m_max / n_c \hat{} 0.25$$

where $\boldsymbol{m_max = 0.75}$ for general tasks or $\boldsymbol{m_max = 1.5}$ for histology (**2**× the base value). Margins are subtracted from the logits of the correct class during training, thereby forcing a larger separation between the logits of the correct class and those of the other classes.

3. Logit Adjustment (LACE)
Logits adjusted by class prior probabilities:

$$logit_adj_c = -log(p_c) / \tau$$

where p_c is the class prior and **τ=1.0** for general tasks. For histology, centered adjustment used with **τ=4.0**:

$$logit_adj_c = (logit_adj_c - mean(logit_adj)) \times \alpha$$

where **α = 0.2** to prevent excessive logit shifts.

4. Kernel Density Estimation (KDE) Weighting for Regression
Instance-level weights for tumour size computed via KDE on training distribution:

$$w_i = 1 / max(density(y_i), \varepsilon)$$

with a Gaussian kernel (bandwidth $\boldsymbol{h = 3.0}$ for Phase B training; $\boldsymbol{h = 4.0}$ for optional regression-only finetuning), emphasising rare tumour sizes.

The class-balanced weighting used β = 0.995, following the effective-number-of-samples formulation of Cui et al. (2019); the LDAM margin used the theoretically derived $n^{-0.25}$ scaling from the generalisation bounds of Cao et al. (2019); and logit adjustment followed Menon et al. (2021), with a centred variant applied to histology. The stronger correction applied to histology reflects its extreme (~205:1) imbalance, while the attenuated, centred adjustment elsewhere was chosen specifically to avoid the majority-class over-correction observed for HER2.

3.6.4 Regularization Techniques

Label Smoothing
Cross-entropy loss modified to prevent overconfident predictions:

$$y_smooth = (1 - \varepsilon) \times y_true + \varepsilon / C$$

where **ε=0.03** (standard tasks) or **ε=0.07** (histology) and C is number of classes.

Dropout
Applied with probability **0.1** to the pooled representation (mean pooling output) before task-specific heads to reduce overfitting.

Gradient Clipping
Gradient norm clipped to maximum value of **1.0** using global L2 norm across all parameters to ensure training stability.

3.6.5 Loss Function Formulation

3.6.5.1 Multi-Task Loss Aggregation
The total loss combined classification and regression objectives with task-specific weights:

$$L_total = \Sigma_c\, \lambda_c \times L_cls(c) + \Sigma_r\, \lambda_r \times L_reg(r)$$

where λ_c and λ_r are task-specific weights ($\lambda_histology = 3.0, \lambda_other = 1.0$).

3.6.5.2 Classification Loss
For each classification head h with C classes:

$$L_cls(h) = -\Sigma_c\, w_c \times y_c \times log(softmax(z_c + adj_c - m_c \times \delta_c))$$

where:

- w_c: Class-balanced weight for class c
- y_c: Label-smoothed target (1-ε for correct class, ε/C for others)
- z_c: Raw logit for class c
- adj_c: Logit adjustment term
- m_c: LDAM margin for class c
- δ_c: Indicator function (1 if c is the true class, 0 otherwise)

3.6.5.3 Regression Loss
For the tumour size prediction:

$$L_reg = \Sigma_i\, w_i \times Huber_\delta(\hat{y}_i - y_i)$$

$$Huber_\delta(e) = \{$$
$$0.5 \times e^2/\delta \qquad if\ |e| \le \delta$$

$$|e| - 0.5 \times \delta \qquad otherwise\}$$

where:

- w_i: KDE-based instance weight
- δ: Huber delta parameter (1.0)
- e: Prediction error in log space

3.6.6 Evaluation Metrics

3.6.6.1 Classification Metrics

For each classification head, the following metrics were computed:

Macro-averaged F1 Score
Computed as the unweighted mean of class-specific F1 scores:

$$F1_macro = (1/C) \times \Sigma_c\, F1_c$$

$$F1_c = 2 \times (precision_c \times recall_c) / (precision_c + recall_c)$$

This metric treats all classes equally regardless of frequency, providing a balanced assessment crucial for imbalanced datasets.

Micro-averaged F1 Score
Computed from aggregated confusion matrix:

$$F1_micro = 2 \times (TP_total / (2 \times TP_total + FP_total + FN_total))$$

where TP_total, FP_total, FN_total are summed across all classes. This metric reflects overall accuracy weighted by class frequency.

Accuracy
Overall proportion of correct predictions:

$$Accuracy = (\Sigma\ correct\ predictions) / (total\ predictions)$$

Confusion Matrices
Generated for each classification head showing predicted vs. true class distributions, enabling detailed error analysis and identification of systematic misclassification patterns.

3.6.6.2 Regression Metrics

For the tumour size prediction, metrics are computed in both log-transformed and original (millimeter) spaces:

Root Mean Squared Error (RMSE)

$$RMSE_log = \sqrt{[(1/n) \times \Sigma(\hat{y}_i - y_i)^2]}$$

$$RMSE_mm = \sqrt{[(1/n) \times \Sigma(exp(\hat{y}_i) - 1 - exp(y_i) - 1)^2]}$$

Mean Absolute Error (MAE)

$$MAE_log = (1/n) \times \Sigma|\hat{y}_i - y_i|$$
$$MAE_mm = (1/n) \times \Sigma|exp(\hat{y}_i) - 1 - exp(y_i) - 1|$$

Coefficient of Determination (R^2)

$$R^2_log = 1 - (\Sigma(\hat{y}_i - y_i)^2) / (\Sigma(y_i - \bar{y})^2)$$
$$R^2_mm = 1 - (\Sigma(exp(\hat{y}_i) - 1 - exp(y_i) - 1)^2) / (\Sigma(exp(y_i) - 1 - \bar{y}_mm)^2)$$

where $\bar{y}$ is the mean of true values.

4 Results

This section presents the quantitative findings from the multi-task BERT framework developed in section 3 for the automated extraction of breast cancer pathology COSD attributes. All models were trained on 4,910 pathology reports, validated on 1,535 reports, and evaluated on a held-out test set of 1,228 reports, ensuring robust assessment of generalisation performance.

4.1 Overall Model Performance

4.1.1 Dataset Characteristics and Training Progression

The final dataset included 7,673 breast cancer pathology reports. Table 2 presents the distribution of samples across the six categorical variables and one continuous variable intended for automated extraction. As expected in clinical research, significant class imbalance was observed across multiple tasks, especially in histology classification, where infiltrating ductal carcinomas accounted for 79.3% of cases (n=3,896) and the rarer carcinomas accounted for only 0.4% (n=19), resulting in an imbalance ratio of approximately 205:1.

Table 2. Dataset Distribution Across Tasks

Task	Class	Training n (%)	Validation n (%)	Test n (%)	Imbalance Ratio
Histology					
	0: 85002 Intraductal carcinoma	19 (0.4%)	10 (0.7%)	3 (0.2%)	205:1
	1: 85003 Infiltrating duct carcinoma	3,896 (79.3%)	1,227 (79.9%)	954 (77.7%)	-
	2: 85203 Lobular Carcinoma	770 (15.7%)	226 (14.7%)	208 (16.9%)	-
	3: 85223 Infiltrating duct and lobular	155 (3.2%)	47 (3.1%)	37 (3.0%)	-
	4: 85233 Infiltrating duct mixed other	70 (1.4%)	25 (1.6%)	26 (2.1%)	-
	Total	4,910	1,535	1,228	
Grade					
	0: G1	629 (12.9%)	204 (13.4%)	106 (8.7%)	6.1:1
	1: G2	2,587 (53.0%)	813 (53.3%)	646 (52.7%)	-
	2: G3	1,670 (34.2%)	507 (33.3%)	473 (38.6%)	-
	Total	4,886	1,524	1,225	
Laterality					
	0: Left	2,655 (54.1%)	827 (53.9%)	653 (53.2%)	~1:1
	1: Right	2,255 (45.9%)	708 (46.1%)	575 (46.8%)	-
	Total	4,910	1,535	1,228	
HER2 Status					
	0: Negative	466 (10.3%)	147 (10.4%)	103 (9.0%)	9.0:1
	1: Positive	3,631 (80.3%)	1,136 (80.2%)	927 (81.2%)	-

	2: Borderline	426 (9.4%)	134 (9.5%)	111 (9.7%)	-
	Total	4,523	1,417	1,141	
ER Status					
	0: Negative	549 (11.7%)	172 (11.7%)	139 (11.8%)	7.1:1
	1: Positive	3,390 (72.2%)	1,059 (72.2%)	839 (71.4%)	-
	2: Not Performed	754 (16.1%)	235 (16.0%)	197 (16.8%)	-
	Total	4,693	1,466	1,175	
PR Status					
	0: Negative	1,046 (31.9%)	326 (31.9%)	207 (25.3%)	2.3:1
	1: Positive	1,775 (54.0%)	552 (53.9%)	523 (64.0%)	-
	2: Not Performed	463 (14.1%)	145 (14.2%)	87 (10.6%)	-
	Total	3,284	1,023	817	

Note: Percentages calculated from non-missing labels only. Some reports had incomplete labelling for certain variables.

For the continuous tumour size variable, measurements ranged from 0.69 to 5.49 ($\log_{1+1}$-transformed scale) with a mean of 3.21. Following the $\log_{1+1}$ transformation, the distribution approximated normality, facilitating stable gradient-based optimisation during model training.

The two-phase training protocol proceeded as planned without numerical instabilities or gradient explosions. Phase A (baseline training with class-balanced cross-entropy) converged smoothly over three epochs. Phase B (advanced training incorporating LDAM margins, LACE logit adjustment, and enhanced class balancing) further refined the model over three to four epochs, with histology-specific interventions (2× LDAM margins, centred negative log-prior adjustment) successfully preventing collapse to majority-class predictions.

4.1.2 Aggregated Test Set Performance

Table 3 displays the main performance metrics for all seven tasks on the held-out test set. The multi-task model showed varied overall performance, with macro-averaged F1 scores for classification tasks ranging from 0.192 to 0.903 (average: 0.557, SD: 0.240). All classification tasks beat the naive majority-class baseline, with performance gains of 5-75 absolute percentage points.

Table 3. Test Set Performance Summary - All Tasks

Task	Macro-F1	Micro-F1	Accuracy	RMSE	MAE	R^2	Test n	Majority Baseline
Histology	0.518	0.863	0.863	-	-	-	1,228	77.7%
Grade	0.640	0.689	0.689	-	-	-	1,225	52.7%
Laterality	0.903	0.903	0.903	-	-	-	1,228	53.2%
HER2 Status	0.192	0.179	0.179	-	-	-	1,141	81.2%
ER Status	0.560	0.591	0.591	-	-	-	1,175	71.4%
PR Status	0.527	0.553	0.553	-	-	-	817	64.0%
Mean (Classification)	0.557	0.630	0.630	-	-	-	-	-
Tumour Size (log)	-	-	-	0.526	0.407	-0.072	1,098	-
Tumour Size (mm)	-	-	-	18.0	12.2	-0.004	1,098	-

Note: All classification metrics reported as proportions (0-1). RMSE and MAE for tumour size reported in both log-transformed space (used during training) and original millimeter scale (clinically interpretable).

The difference between macro-F1 and micro-F1 scores, particularly in histology (0.518 vs. 0.863) and ER/PR status, illustrates the effect of class imbalance on model performance. Macro-F1, which averages performance equally across all classes regardless of how often they occur, offers a more rigorous assessment of detecting minority classes, a crucial aspect for clinical use, where rare but important subtypes need to be reliably recognised. The model achieved a macro-F1 of 0.518 for the most imbalanced task (histology), showing partial alleviation of class imbalance issues through the combined approach of CB loss weighting (β=0.995), LDAM margins (m_{max}=0.75, 2× for histology), and centred LACE adjustments (τ=4.0, α=0.2).

For tumour size regression, the model achieved a test-set MAE of 12.2 mm (RMSE: 18.0 mm, R^2: -0.004), representing approximately 38% mean absolute percentage error relative to the median tumour size. The negative R^2 value indicates the model performs at or slightly below a naive mean predictor, suggesting challenges in learning size patterns from text descriptions. This

level of precision falls short of the clinically acceptable threshold for automated size documentation in COSD submissions, indicating this task requires further methodological refinement. Note that the final two rows in Table 3 report a single experiment: the tumour-size regression head was trained once on log-transformed targets (log1p), and its predictions are reported on both the log scale used during optimisation and, after inverse transformation (expm1), the original millimetre scale. They are two representations of the same result, not two independent experiments.

4.1.3 Validation-Test Generalisation

Table 4 compares validation and test set performance to evaluate model generalisation and potential overfitting. The close match between validation and test metrics (mean absolute difference: 0.009 macro-F1 points, 1.5% relative change) indicates strong generalisation to unseen data.

Table 4. Validation vs Test Set Performance Comparison

Task	Validation Macro-F1	Test Macro-F1	Difference	Validation Acc	Test Acc	Difference
Histology	0.516	0.518	+0.002	0.858	0.863	+0.005
Grade	0.621	0.640	+0.019	0.671	0.689	+0.018
Laterality	0.900	0.903	+0.003	0.900	0.903	+0.003
HER2 Status	0.192	0.192	-0.000	0.181	0.179	-0.002
ER Status	0.543	0.560	+0.018	0.559	0.591	+0.032
PR Status	0.517	0.527	+0.010	0.539	0.553	+0.015
Mean	0.548	0.557	+0.009	0.618	0.630	+0.012

For the regression task, validation-test RMSE differed by 0.8 mm (4.1% relative change), indicating consistent (though suboptimal) performance. No task showed performance degradation exceeding 2%, reflecting minimal overfitting despite the relatively small training set (n=4,910) and high model capacity (~110M parameters). This robustness likely results from the comprehensive regularisation strategy employed: label smoothing (ε=0.03-0.07), dropout (p=0.1), weight decay (0.01), gradient clipping (max norm 1.0), and the implicit regularisation effect of multi-task learning itself.

4.2 Per-Task Analysis

This section provides a detailed examination of model performance for each of the seven COSD variables, with particular attention to error patterns, confusion between clinically related categories, and the impact of class imbalance on minority class detection.

4.2.1 Histology Classification

Histological subtype prediction was the most challenging task due to extreme class imbalance (205:1) and the inherent ambiguity of borderline cases. The model achieved a test-set macro-F1 of 0.518 and accuracy of 0.863, with per-class performance summarised in Table 5.

Table 5. Histology Classification - Test Set Per-Class Performance

Class	Precision	Recall	F1-Score	Support	Class %	TPR
0: 85002 Intraductal carcinoma	0.00	0.00	0.00	3	0.2%	0.00
1: 85003 Infiltrating duct carcinoma	0.92	0.93	0.93	954	77.7%	0.93
2: 85203 Lobular Carcinoma	0.85	0.69	0.76	208	16.9%	0.69
3: 85223 Infiltrating duct and lobular	0.37	0.43	0.40	37	3.0%	0.43
4: 85233 Infiltrating duct mixed other	0.47	0.54	0.50	26	2.1%	0.54
Macro avg	0.52	0.52	0.52	1,228	-	-
Weighted avg	0.88	0.86	0.87	1,228	-	-

Key Findings:

1. Majority Class (Infiltrating Ductal Carcinoma, Class 1). As the dominant histological subtype (77.7% of test cases), Infiltrating ductal carcinoma was predicted with high accuracy (precision: 0.92, recall: 0.93). The model successfully identified 93% of actual cases, with most errors (64 cases, 7%) representing misclassifications as infiltrating lobular carcinoma (Class 2).
2. Critical Minority Classes (Class 0). Representing only 0.4% of training examples (n=19) and appearing in just 3 test cases, the model completely failed to detect this class, achieving 0% recall. This represents model collapse for this extremely rare but clinically critical category.
3. Intermediate Rare Classes (Class 3 and Class 4). Class 3 (3.0%, n=37) achieved moderate performance (F1: 0.40), with a recall of 0.43, indicating the model partially learned discriminative features for this class despite limited training examples.
4. Infiltrating Lobular Carcinoma (Class 2). As the second most common subtype (16.9%, n=208), lobular carcinoma achieved good performance (F1: 0.76), though recall (0.69) was lower than for ductal carcinoma.

Clinical Implications. The achieved histology classification performance (macro-F1: 0.518) falls short of the 85-90% accuracy threshold typically required for semi-automated COSD reporting. The complete failure to detect rare carcinomas (0/3 cases) represents a critical safety concern that precludes clinical deployment without additional refinements.

4.2.2 Tumour Grade Classification

Grade classification (I, II, III) is a critical prognostic variable that helps determine treatment intensity and associated clinical pathways. The model achieved macro-F1 of 0.640 and accuracy of 0.689 on the test set (n=1,225), with detailed performance metrics presented in Table 4.5.

Table 6. Grade Classification Test Set Per-Class Performance

Grade	Precision	Recall	F1-Score	Support
I (Low)	0.28	0.85	0.43	106
II (Intermediate)	0.83	0.65	0.73	646
III (High)	0.83	0.71	0.77	473
Macro avg	0.65	0.74	0.64	1,225
Weighted avg	0.78	0.69	0.72	1,225

Key Findings:

1. Binary Discrimination (Well vs. Poorly Differentiated). The model demonstrated strong ability to distinguish well-differentiated (Grade I) from poorly differentiated (Grade III) tumours, as evidenced by high recall for Grade I (0.85) and Grade III (0.71).
2. Grade II Ambiguity. Grade I showed high recall (0.85) but very low precision (0.28), indicating that the model frequently over-predicted Grade I.
3. Class Balance and Performance. The moderate imbalance (Grade II: 52.7%, Grade III: 38.6%, Grade I: 8.7%) allowed class-balanced weighting to maintain relatively consistent per-class F1 scores for Grades II and III (0.73 and 0.77).

Clinical Significance. Grade classification performance (macro-F1: 0.640) is insufficient for treatment-level decision support. The Grade I precision issue (28%) means that only about 1 in 4 predicted Grade I cases are truly Grade I.

4.2.3 Laterality Prediction

Laterality Table 7 (left vs. right breast) represented the most balanced and clinically unambiguous task, achieving macro-F1 of 0.903 and accuracy of 0.903 (n=1,228).

Table 7. Laterality Classification Test Set Per-Class Performance

Side	Precision	Recall	F1-Score	Support
Left	0.92	0.89	0.91	653
Right	0.88	0.91	0.90	575
Macro avg	0.90	0.90	0.90	1,228

Key Findings:

The near-perfect performance indicates: (1) explicit documentation of laterality in standardised template sections, (2) almost equal class distribution (53% left, 47% right), and (3) lack of clinical ambiguity. The 119 errors (9.7%) probably represent either genuine mislabeling in the gold standard, reports describing bilateral findings where only one side was labelled, or reports mentioning contralateral comparisons that confused the model. Implications. Laterality extraction is production-ready and requires minimal human verification. This high-confidence prediction could act as an internal quality check; a discrepancy between the model and the manual COSD entry may highlight data-entry errors for human review.

4.2.4 Hormone Receptor Status Predictions (HER2, ER, PR)

The three receptor status tasks predict trichotomous outcomes: negative, positive, or either borderline/not performed. These tasks exhibited moderate class imbalance and inherent ambiguity in the other categories outside of positive or negative. Performance summaries are presented in Tables 8-10.

Table 8. HER2 Status Test Set Per-Class Performance

Status	Precision	Recall	F1-Score	Support
Negative	0.11	0.70	0.20	103
Positive	0.95	0.07	0.12	927
Borderline	0.16	0.64	0.26	111
Macro avg	0.41	0.47	0.19	1,141

Table 9. ER Status Test Set Per-Class Performance

Status	Precision	Recall	F1-Score	Support
Negative	0.26	0.82	0.39	139
Positive	0.93	0.52	0.67	839
Not Performed	0.54	0.73	0.62	197
Macro avg	0.58	0.69	0.56	1,175

Table 10. PR Status Test Set Per-Class Performance

Status	Precision	Recall	F1-Score	Support
Negative	0.48	0.69	0.57	207
Positive	0.85	0.46	0.60	523
Not Performed	0.29	0.77	0.42	87
Macro avg	0.54	0.64	0.53	817

Key Findings:

1. Performance Ranking. ER status (0.560 macro-F1) > PR status (0.527) > HER2 status (0.192) demonstrated progressively declining performance.
2. HER2 Catastrophic Failure. HER2 status exhibited severe model collapse, with the positive class (81.2% of data) achieving only 7% recall despite 95% precision. The model essentially learned to almost never predict HER2-positive.
3. Ambiguous Category Performance. The not performed class proved moderately challenging across ER and PR (F1 scores of 0.62 and 0.42, respectively) but was one of the few receptor categories with reasonable recall (0.64).

Clinical Deployment Considerations. HER2 status predictions are completely unsuitable for deployment (macro-F1: 0.192), requiring fundamental retraining with different class-balancing strategies or separate single-task modelling. ER and PR status predictions (macro-F1: 0.560 and 0.527) show moderate promise but fall short of clinical requirements.

4.2.5 Tumour Size Regression

Tumour size prediction represents the sole regression task. The model achieved test-set performance of RMSE=18.0 mm, MAE=12.2 mm, and R^2=-0.004 in the clinically relevant millimetre scale. Table 11 summarises performance metrics.

Table 11. Tumour Size Test Set Prediction Performance

Metric	Log Space	Original (mm)	Interpretation
RMSE	0.526	18.0	Average error: 18.0 mm
MAE	0.407	12.2	Typical error: 12.2 mm
R^2	-0.072	-0.004	Model ≈ predicting mean
Samples	1,098	1,098	-

Key Findings:

1. Negative R^2 Interpretation. The near-zero (slightly negative) R^2 value of -0.004 indicates the model performs at or marginally worse than a naive baseline that predicts the mean tumour size for all cases.

2. Clinical Accuracy Assessment. The MAE of 12.2 mm represents substantial error relative to typical tumour sizes (median ~24 mm). For T-stage classification (T1: ≤20mm, T2: 21-50mm, T3: >50mm), accurate size prediction is critical near category boundaries.

Clinical Utility. The achieved accuracy (MAE: 12.2 mm, R^2: -0.004) is completely inadequate for T-stage determination or clinical decision-making. The model essentially functions as a mean predictor and provides no value over simple statistical baselines.

4.3 Class Imbalance Mitigation: Impact of Advanced Loss Strategies

A core objective was to develop strategies for handling extreme class imbalance in cancer pathology NLP. This section assesses the effectiveness of the comprehensive anti-collapse approach.

4.3.1 Summary of Imbalance Handling Effectiveness

The model employed a multi-pronged strategy: class-balanced loss (CB, β=0.995), LDAM margins (m_max=0.75, 2× for histology), centered LACE logit adjustment (τ=4.0, α=0.2 for histology), round-robin balanced sampling, and task-specific weighting (3× for histology). Effectiveness varied substantially by task and minority class size:

Successes:

- Moderate minorities (10-15% of data): Infiltrating lobular carcinoma (15.7%), and Grade I (12.9%) achieved reasonable recall (0.69, 0.85) despite being outnumbered 5-10:1.
- Near-balanced tasks: Laterality (54:46 split) achieved excellent performance (0.903 F1) with minimal class-bias.

Validation-test generalisation: Close correspondence (mean 0.009 F1 difference) indicates anti-collapse strategies didn't cause overfitting.

Marginal Successes:

- Marginal performance (1–5% of data): DCIS (3.2%) achieved only marginal recall (0.43), consistent with the broader pattern in which classes below roughly 5% representation fall into the marginal band.

Failures:

- Extreme minorities (<1%): Histology Class 0 (0.4% of training, 19 samples, 205:1 imbalance) achieved 0% recall and complete model collapse despite maximum intervention.
- HER2-positive paradox: Despite being the majority class (80.3%), HER2-positive achieved only 7% recall, indicating class-balancing over-corrected or a fundamental training/labelling issue.

4.3.2 Quantitative Impact Assessment

Table 12. Test Set Minority Class Recall Summary

Task	Class	Train %	Test Support	Test Recall
Histology	**0: 85002** Intraductal carcinoma	0.4%	3	0.00
Histology	**3: 85223** Infiltrating duct and lobular	3.2%	37	0.43
Histology	**4: 85233** Infiltrating duct mixed other	1.4%	26	0.54
Grade	0: Grade I	12.9%	106	0.85
HER2	0: Negative	10.3%	103	0.70
HER2	2: Borderline	9.4%	111	0.64
ER	0: Negative	11.7%	139	0.82
PR	0: Negative	31.9%	207	0.69

Key Patterns:

5. Critical threshold: Minorities representing 10-15% of training achieved good recall (0.70-0.85).
6. Danger zone: Minorities at 1-5% achieved marginal recall (0.43-0.54).
7. Failure point: Minorities <1% (intraductal carcinoma: 0.4%) completely collapsed to 0% recall.

The 205:1 imbalance ratio appears to exceed the capacity of current class-balancing techniques, suggesting a practical limit of 100:1 for these methods.

4.5 Research Questions Addressed

4.5.1 Primary Research Question

How effectively can a Multitask BERT framework be applied to automate the structured completion of the UK COSD breast-cancer pathology dataset from unstructured pathology reports while maintaining clinical validity and interpretability?

The MT-BERT model achieved a mean Macro-F1 of 0.5566 across all classification tasks and an Accuracy of 0.6925, demonstrating that a single shared transformer encoder can replicate

key elements of structured COSD reporting. Performance was highest for Laterality ($F_1 = 0.903$) and Grade ($F_1 = 0.640$), with most errors limited to clinically adjacent categories. Moderate success was achieved for ER ($F_1 = 0.560$), while PR ($F_1 = 0.527$) and Histology ($F_1 = 0.518$) showed mixed results. The HER2 and Tumour-size categories underperformed ($F_1 = 0.192$; $R^2 \approx 0$), indicating that while the unified approach is feasible, some variables require independent optimisation or alternative formulations.

4.5.2 Secondary Research Questions

4.5.2.1 Model Architecture and Task Design

1. How can heterogeneous targets—multi-class, binary, ordinal, and regression—be jointly modelled within a single transformer-based framework without degrading per-task performance?

The shared BioMedBERT encoder effectively supported seven diverse heads. Joint modelling enhanced stability and efficiency, without significant performance loss. However, negative transfer was observed in the regression head and the highly imbalanced HER2 task, confirming that heterogeneity is feasible but not universally advantageous.

2. How can architectural and training design choices (e.g. shared encoder with specialised heads) support effective multitask learning across diverse COSD variables?

A hard parameter-sharing design with attention-masked mean-pooled encoder representations and task-specific output layers proved effective as a regulariser, yielding close validation–test agreement ($\Delta F_1 \approx 0.009$). Specialised heads enabled domain-specific optimisation (e.g. larger LDAM margins for histology). Nonetheless, the regression head's poor signal-to-noise ratio suggests that fully decoupled heads may be needed for tasks with divergent linguistic structure.

4.5.2.2. Loss Function Engineering and Optimisation

3. Which per-task loss configurations (class-balanced, focal, label-smoothed, or Huber) yield the most stable optimisation under class imbalance and label missingness?

A composite loss employing Class-Balanced Cross-Entropy ($\beta = 0.995$), LDAM margins, and Label Smoothing ($\varepsilon = 0.03$–0.07) yielded the most stable convergence. These elements collectively improved minority-class recall by more than 0.25 F_1 points over unweighted baselines. The Huber loss helped stabilise regression but did not achieve usable predictive accuracy.

4. How do dynamic task-weighting strategies (e.g. uncertainty-based or gradient-normalised) affect convergence and inter-task interference?

Fixed task weights (λ_histology = 3.0; λ_others = 1.0) outperformed uncertainty-based weighting by preventing oscillation between heads. Although dynamic schemes were tested, the simple fixed configuration yielded smoother convergence and better minority recall, implying that explicit weighting can substitute for more complex adaptive schedules in small sample multitask settings.

5. Does staged fine-tuning—progressing from head warm-up to partial unfreeze and full joint training—enhance stability and calibration?

Yes. The two-stage training protocol (simple baseline followed by advanced anti-collapse training) reduced early-epoch divergence and improved overall calibration. Validation loss stabilised after epoch 4, and minority collapse was avoided in tasks accounting for more than 1% of the training data.

4.5.2.3. Data Quality and Calibration

6. How can per-head loss masking mitigate bias from missing labels while preserving training data?

Instance-level loss masking enabled the inclusion of partially labelled reports (≈ 20 % missingness), increasing the effective data volume without introducing label noise. Empirically, tasks with higher missingness (ER/PR status) retained balanced precision-recall curves, confirming that masking prevented systematic bias.

7. What is the relationship between class-frequency distribution, loss weighting, and predictive calibration across COSD pathology variables?

Calibration quality was inversely related to the severity of class imbalance. Moderate minorities (10–15%) achieved good recall (0.70–0.85) and well-calibrated posteriors, whereas extreme minorities (< 1%) collapsed entirely. Over-weighting occasionally hindered majority calibration (HER2 positive 7% recall), demonstrating a practical upper limit of approximately 100:1 imbalance for current loss designs.

4.5.3 Summary

Across all seven research questions, the results confirm that the Multitask BERT framework can effectively automate key components of COSD reporting, offering parameter-efficient inference and clinically acceptable performance for some variables, most notably laterality. The findings also expose the limitations of current transformer-based approaches to handling imbalance, especially for extremely rare or numerically expressed targets.

5 Discussion

5.1 Overview of Findings

The multitask BERT (MT-BERT) framework, developed for the automated completion of COSD-compatible breast pathology variables, showed varied yet interpretable performance across the seven learning heads. Classification performance ranged from $0.192 \leq F_1 \leq 0.903$ (mean = 0.557), while tumour-size regression achieved MAE = 12.2 mm and $R^2 \approx 0$. Laterality achieved clinically useful performance; grade, ER, and PR status showed promise but fell short of clinical decision-support thresholds; histology was moderate; and HER2 status and tumour size failed.

5.2 Interpretation in the Context of Prior Work

5.2.1 Effectiveness of the Multitask Framework

Previous multitask transformer studies (Peng et al., 2020; Mulyar et al., 2021) reported modest but consistent performance improvements through shared representation learning. The results in section 4 also support these findings: shared encoder representations enhanced stability and generalisation, as shown by the small validation–test delta (mean = 0.009 macro-F_1; Table 4.3). However, negative transfer was observed in tasks with sharply diverging statistical properties—HER2 status (extreme over-correction) and tumour size (continuous output overwhelmed by classification losses). This aligns with the literature's observation that the benefits of multi-task learning depend on task-relatedness (Zhang and Yang, 2022). In breast pathology text, histology, grade, and hormone receptor status share linguistic and biological cues, whereas tumour size relies on quantitative expressions that are poorly captured by transformer semantics.

5.2.2 Class-Imbalance Mitigation and Loss Design

The combined imbalance-mitigation strategy of using class-balanced weighting ($\beta = 0.995$), LDAM margins, centred LACE adjustment, and round-robin sampling proved to be partially successful. Moderate minorities (10–15 % of training data), such as Grade I tumours, achieved a strong recall (0.85), while classes below 1 % (Histology class 0: 85002 Intraductal carcinoma) collapsed completely, Table 12. These findings refine prior theoretical thresholds for effective long-tail learning (Cui et al., 2019; Cao et al., 2019; Menon et al., 2021), suggesting a practical ceiling of roughly 100:1 imbalance under current transformer-loss formulations. The HER2-positive paradox and majority-class under-prediction illustrate that aggressive re-weighting can distort posterior calibration, echoing concerns raised by Blagec et al. (2022) about over-compensation in medical NLP imbalance correction.

5.2.3 Regression Head Performance

The tumour-size regression failure ($R^2 \approx 0$; MAE ≈ 12 mm) indicates that linguistic patterns within pathology reports do not encode tumour dimensions in a learnable, continuous form. Weighted Huber loss with KDE weighting stabilised optimisation but could not extract quantitative relationships. Comparable limitations were reported by Steininger et al. (2021) for imbalanced regression tasks, suggesting that textual cues require pre-extracted numeric entities or multimodal fusion with imaging data. Future iterations should recast size estimation as ordinal classification (e.g. T-staging) or integrate rule-based measurement extraction prior to regression.

5.2.4 Computational Efficiency

A practical motivation for the multitask design is computational efficiency, and the architecture achieves this by construction rather than by tuning. The model shares a single BioMedBERT encoder (~110 million parameters, 12 transformer layers, 768-dimensional hidden states) across all seven tasks, with each task served by a single linear head mapping the 768-dimensional pooled representation to its small output space (one to five units). The seven heads therefore add only a few thousand parameters in total, a negligible fraction of the encoder. The natural alternative, training an independent single-task model per variable, would replicate the full ~110M-parameter encoder seven times, yielding on the order of seven times the parameter count, storage footprint, and training cost for the same set of predictions. At inference time, the shared design requires a single forward pass through one encoder per report, from which all seven outputs are read simultaneously, rather than seven separate passes through seven encoders. The efficiency advantage thus scales with the number of tasks and is intrinsic to parameter sharing rather than dependent on any hardware or implementation. This is consistent with the practical efficiency gains reported for multitask clinical transformers, in which a single shared model reduced inference costs and simplified deployment relative to a suite of specialised models (Mulyar et al., 2021; Peng et al., 2020). For an NHS deployment context, where a single maintained model is considerably easier to validate, monitor, and govern than seven, this consolidation is an operational as well as a computational benefit. A precise empirical characterisation of training and inference time on production hardware was beyond the scope of this study and is identified as a direction for future benchmarking.

5.2.5 Ethical and Data-Governance Considerations

All analyses were conducted within the Genomics England secure research environment, using data from participants in the 100k Genomes Project who had provided explicit consent for research use. No identifiable information left the secure enclave; only aggregated model outputs were exported. This approach aligns with NHS Digital's Information Governance Toolkit and the UK Data Ethics Framework. Automated extraction systems must preserve these safeguards in

deployment and ensure data minimisation, auditability of predictions, and continuous performance monitoring to mitigate bias, particularly for underrepresented subgroups.

5.2.5 Deeper Discussion of Negative Results

5.2.5.1 HER2 over-correction

Although HER2-positive was the majority class, accounting for over 80% of the data, the model achieved only 7% recall for it. This paradox arose because the class-balanced loss, LDAM margins, and logit adjustment simultaneously upweighted the minority classes (Negative at 10.3%, Borderline at 9.4%) to such an extent that the model almost never predicted the majority class. The implication is that aggressive re-weighting can distort posterior calibration (Blagec et al., 2022): correction strength cannot be applied uniformly across heads but must be tuned per task, or HER2 separated into a standalone classifier. This is an informative result, demonstrating a concrete failure mode of stacked imbalance-correction methods rather than a mere null finding.

5.2.5.2 The regression head

Tumour-size regression achieved an MAE of 12.2 mm against the clinically critical T1/T2 boundary of 20 mm. The KDE-weighted Huber loss stabilised optimisation but failed to extract a reliable quantitative signal from the free text. Reported in clinical units, the result makes the limitation unambiguous: an average error of 12.2 mm means the model cannot reliably distinguish T-staging categories, with direct treatment consequences and rendering the head unsuitable for deployment in its current form. The implication is that tumour size is better framed as an ordinal T-stage classification or extracted via a rule-based numeric entity-extraction step, rather than being treated as a continuous regression variable. One consideration is that tumour size is often documented near the beginning of a report, in the macroscopic description of the specimen, rather than in the concluding summary on which most classification targets depend. Left truncation, chosen to preserve that end-of-report content, may therefore be less well suited to the size-regression target than to the classification heads. This was not tested directly and is noted as a possible contributing factor to the regression head's weak performance and as a candidate for future investigation.

5.2.5.3 Data requirements for a more balanced model

The class-level results provide an empirically grounded estimate of the data required to make such a model clinically viable. Performance tracked class representation closely: heads where the minority class comprised roughly 10–15% of the training data achieved usable recall (0.70–0.85); those at 1–5% achieved only marginal recall (0.43–0.54); and those below 1%, such as intraductal carcinoma with 19 examples, collapsed entirely. This indicates that loss-based correction alone cannot substitute for an adequate number of examples, and that beyond roughly

a 100:1 imbalance ratio the gradient signal from the minority class is simply too sparse, regardless of weighting. The practical implication is that the most effective route to improvement is not a more elaborate loss function but more representative data for the rarest categories. Two complementary strategies follow. First, the absolute count of minority-class examples must be increased by a larger underlying cohort, deliberate oversampling of rare diagnoses at the data-acquisition stage, or active learning to target reports likely to contain underrepresented classes. As a rough guide, the results suggest each class needs to reach on the order of low thousands of examples, or roughly 10% representation, before reliable recall emerges; below this threshold, no amount of reweighting recovered performance. Second, the report-level data split could be replaced with a patient-level split, since report-level partitioning risks leakage between training and evaluation sets and inflates apparent performance, masking the exact data insufficiency described here. Combining a larger and more balanced corpus with patient-level partitioning could help address the underlying cause of the negative results, whereas further loss-function refinement would not.

5.2.5.4 Clinical implications of the negative results

Taken together, these failures convey a clear and important lesson for deploying such models in healthcare: aggregate performance metrics obscure exactly the failures that matter most clinically. A model can show a respectable mean macro-F1 across heads while harbouring a head such as HER2, where the clinically dominant class is recalled only 7% of the time, or a regression output that cannot distinguish T-staging categories. In a clinical setting, these are not benign errors. A missed HER2-positive status could affect eligibility for targeted therapy, and an unreliable tumour-size estimate could mislead T-staging and the surgical and adjuvant treatment decisions that follow. The complete collapse of the rarest histology class (intraductal carcinoma, 0% recall) is equally consequential, because rarity in the training data does not imply clinical unimportance. Rare categories are often precisely the ones requiring careful identification. These results therefore argue strongly that any deployment of automated COSD completion must be positioned as a clinician-assist tool with human verification rather than an autonomous system, be evaluated and monitored per class and per task rather than on aggregate scores alone and be subject to continuous performance monitoring to detect degradation and bias against underrepresented subgroups. This aligns with the evaluation principles set out by Vickers et al. (2024), who argue that clinical NLP systems should be assessed for calibration, robustness, bias, and fairness rather than headline accuracy, and with Blagec et al. (2022), who caution that aggregate metrics obscure the class-level disparities most relevant to patient safety.

5.3 Limitations

Several limitations temper these conclusions: dataset scope, extreme imbalance, regression target quality, cross-institutional generalisability, and interpretability challenges. Minority histologies (<1 %) yielded insufficient gradient signal despite weighted sampling. Future work should address these through data expansion, hybrid modelling, and interpretability tools.

5.4 Future Work

Future research should address the identified limitations by expanding the dataset, adopting hybrid modelling, integrating multimodal data, and improving interpretability. The MT-BERT framework provides a foundation for several important extensions, as outlined in the following sections.

5.4.1 Explainability and attribution analysis.

A valuable direction for future work is to apply explainability techniques to assess how the shared multi-layer encoder distributes information across the multiple task-specific heads. However, the interpretability of transformer attention weights as a method is somewhat contested. According to Jain and Wallace (2019) and Bibal et al. (2022), attention weights are not reliable explanations, which motivates the use of gradient-based attribution methods rather than raw attention visualisations. Integrated Gradients (Sundararajan et al., 2017), an axiomatic attribution method that satisfies sensitivity and implementation invariance, together with transformer-specific approaches such as Chefer et al. (2021), would attribute each prediction to the input tokens that drove it, enabling verification that the model attends to clinically meaningful spans. For example, grade predictions are based on the Nottingham grade components recorded in the pathology report (tubule formation, nuclear pleomorphism, and mitotic count), and receptor status predictions are based on the corresponding immunohistochemical results also recorded in the report. These methods are readily available through libraries such as Captum (Kokhlikyan et al., 2020). Beyond per-prediction interpretability, attribution analysis could probe the hard parameter-sharing design itself by examining whether the shared encoder learns representations that genuinely serve all heads, or whether certain heads dominate the shared features, offering empirical insight into the negative transfer and over-correction behaviours observed in the results and helping establish the auditability that clinical deployment requires.

5.4.2 Alternative architectures for multi-task classification

The current framework uses hard parameter sharing (Caruana, 1997), a single shared encoder with independent linear heads. Future work could compare this with architectures that offer more flexible information sharing. Soft parameter sharing approaches, such as cross-stitch networks (Misra et al., 2016), allow each task to learn its own combination of shared and task-specific

features, which may mitigate negative transfer observed between heads with conflicting class distributions. Adapter- or prompt-based multi-task formulations could provide task isolation while keeping the encoder largely frozen, reducing the risk that one head's gradients distort the shared representation, as observed in the HER2 case. More expressive head designs, for instance, attention-pooled or small transformer heads rather than single linear layers, could also be evaluated, particularly for histology, whose extreme class count and imbalance place the greatest demand on the shared representation. Comparing these architectures on the same dataset would clarify whether the limitations observed here stem from the sharing scheme itself or from the underlying data constraints.

5.4.3 Patient-level splitting and multi-institutional cohorts

The current data splits are at the report level rather than the patient level, meaning reports from the same patient may appear in both training and validation sets, risking optimistic bias. Future work should re-run the pipeline with patient-level splitting to obtain a more conservative and clinically realistic performance estimate. Equally important is validation on larger, multi-institutional cancer cohorts: the present dataset derives from a limited cohort, and pathology reporting style, terminology, and structure vary considerably between institutions. Establishing that the approach generalises beyond the Genomics England cohort, ideally across reports from multiple UK trusts and laboratories, is necessary before any claim of broad applicability. A larger multi-institutional corpus would also help provide the additional minority-class examples that Section 3.4 identifies as the binding constraint on the rarest COSD variables.

5.4.4 Addressing the imbalance ceiling

The results suggest a practical ceiling of around 100:1 for loss-based imbalance correction (Class-Balanced CE, LDAM, Logit Adjustment). Beyond this, future work should explore complementary strategies, such as targeted data augmentation, synthetic minority oversampling at the text level, or active learning to acquire more minority-class examples, rather than relying solely on loss reweighting.

5.4.5 Improving the regression head

With an MAE of 12.2 mm against a 20 mm T1/T2 boundary, tumour-size regression is not yet clinically deployable. Future work could reframe this as an ordinal T-stage classification directly or use a rule-based numeric entity-extraction step as pre-processing, either of which may better match clinical needs than continuous size prediction.

5.4.6 Multimodal integration

The current model operates solely on free-text pathology reports. A natural extension is to incorporate structured clinical variables, such as patient age, tumour location, or genomic data from the Genomics England cohort, as additional multimodal inputs alongside the encoded text. Fusing structured features with the text representation (for example, by concatenating them with the pooled encoder output before the task heads, or via a dedicated fusion layer) could improve predictions for variables that correlate with clinical context and better reflect the information a pathologist integrates in practice. This could be particularly promising for the weaker heads, where the text signal alone proved insufficient.

5.5 Chapter Summary

This discussion established that the MT-BERT framework achieves its core goal of automated COSD variable extraction from free-text breast pathology reports, with laterality reaching production-ready accuracy, and grade, ER, and PR status showing promise while still falling short of clinical decision-support thresholds, as detailed per task in Chapter 4. The shared-encoder design proved effective, with a small validation-test delta (mean 0.009 macro-F_1) confirming stable generalisation, and a single encoder serving seven heads offering a marked efficiency advantage over seven independent single-task models.

Equally important, the chapter delineated the limits of the current approach. Negative transfer emerged in tasks whose statistical properties diverged sharply from the rest, most clearly in HER2 status, where aggressive imbalance correction distorted majority-class prediction, and in tumour-size regression, where continuous quantitative targets proved poorly suited to transformer semantics. The class-imbalance analysis refined existing theoretical thresholds into a practical ceiling of roughly 100:1, beyond which loss-based correction alone cannot recover minority-class performance and data-level interventions become necessary.

These findings have a consistent implication for clinical use: aggregate performance metrics obscure the per-class failures that matter most for patient safety, so any deployment must be positioned as a clinician-assist tool with human verification, evaluated per class and per task, and subject to continuous monitoring within a secure, well-governed data environment. The future work directions set out above, including patient-level and multi-institutional validation, alternative sharing architectures, multimodal integration, explainability through gradient-based attribution, and a reformulated regression target, follow directly from these limitations and define a clear path towards a more robust and clinically deployable system. Taken together, the framework provides a validated foundation for automated COSD completion while honestly bounding what it can and cannot yet achieve.

6 Conclusion

This thesis presented and evaluated a novel end-to-end Multitask BERT (MT-BERT) framework designed to automate the extraction and completion of structured data for the breast-pathology components required for the Cancer Outcomes and Services Dataset (COSD). By addressing key methodological gaps in fragmented single-task modelling and in handling real-world data imbalance, the MT-BERT architecture unified seven distinct clinical variables, ranging from categorical status labels to continuous tumour size, within a single, parameter-efficient model built on a shared BioMedBERT encoder, in place of seven separate task-specific models with parallel task-specific heads.

Returning to the central research question — how effectively a multitask BERT framework can automate the structured completion of the UK COSD breast-cancer pathology dataset while maintaining clinical validity and interpretability — the evidence supports a qualified but genuine answer. The framework performs best on variables that are frequently and consistently expressed in pathology text: laterality reached production-ready accuracy, while grade, ER, and PR status showed promise but have not yet met clinical decision-support thresholds. The small validation–test gap (mean 0.009 macro-F_1) confirmed that the shared encoder generalised stably rather than overfitting. The shared-encoder design also realises the intended parameter efficiency by replacing seven separate single-task models with a single shared encoder that serves all heads.

At the same time, the work is candid about where the approach reaches its limits. Performance was constrained by data availability rather than by loss design: the layered imbalance-correction strategy (Class-Balanced loss, LDAM, and centred Logit Adjustment) improved minority-class recall up to a practical ceiling of roughly 100:1, beyond which it could not compensate for insufficient training data. In the case of HER2, it over-corrected the majority class. Tumour-size regression showed that continuous quantitative values are poorly recoverable from narrative text in its current form. These are informative negative results: they delineate the boundary between what loss-based methods can achieve and what requires data-level or architectural intervention, and they carry a clear message for clinical translation, namely that such a system belongs in a clinician-assist role with per-class monitoring rather than in autonomous deployment.

The principal contribution of this thesis is twofold: a unified, reproducible multitask framework for COSD pathology extraction that performs well on common variables and delivers significant efficiency gains, and a clear empirical characterisation of the conditions under which this class of model succeeds or fails. The directions for future work, including patient-level and multi-institutional validation, multimodal fusion of structured and textual features, alternative parameter-sharing architectures, gradient-based explainability, and a reformulated tumour-size target, follow directly from these findings. In summary, the framework establishes a validated foundation for automating COSD completion within NHS digital pathology workflows, while

honestly bounding the data and methodological conditions that any clinical deployment would need to satisfy.

References

Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jin, D., Naumann, T. and McDermott, M. 2019. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop, pp.72–78.

Beltagy, I., Lo, K. and Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. EMNLP-IJCNLP, pp.3615–3620.

Bibal, A., Cardon, R., Alfter, D., Wilkens, R., Wang, X., Francois, T. and Watrin, P. 2022. Is Attention Explanation? An Introduction to the Debate. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1, pp.3889–3900.

Blagec, K., Dorffner, G., Moradi, M., Ott, S. and Samwald, M. 2022. A global analysis of metrics used for measuring performance in NLP. In: NLP-Power 2022, pp.52–63.

Cao, K., Wei, C., Gaidon, A., Arechiga, N. and Ma, T. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In: Advances in Neural Information Processing Systems (NeurIPS).

Caruana, R. 1997. Multitask Learning. Machine Learning, 28(1), pp.41–75.

Chefer, H., Gur, S. and Wolf, L. 2021. Transformer Interpretability Beyond Attention Visualization. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.782–791.

Cui, Y., Jia, M., Lin, T.Y., Song, Y. and Belongie, S. 2019. Class-Balanced Loss Based on Effective Number of Samples. In: CVPR, pp.9268–9277.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Carin, L., Gao, J. and Poon, H. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. In: EMNLP.

Han, Y., Park, J. and Kim, T. 2022. Variance Stabilization Using Logarithmic Transformations for Biomedical Regression Models. BMC Medical Research Methodology, 22(141), pp.1–12.

Henson, K.E., Elliss-Brookes, L., Coupland, V., Payne, E., Vernon, S., Rous, B. and Rashbass, J. 2020. Data Resource Profile: National Cancer Registration Dataset in England. International Journal of Epidemiology, 49(1), pp.16–20.

Kendall, A., Gal, Y. and Cipolla, R. 2018. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In: CVPR, pp.7482–7491.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S. and Reblitz-Richardson, O. 2020. Captum: A unified and generic model interpretability library for PyTorch. arXiv preprint arXiv:2009.07896.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), pp.1234–1240.

Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollar, P. 2017. Focal Loss for Dense Object Detection. In: ICCV, pp.2980–2988.

Ling, A.Y. et al. 2019. Using NLP to construct a metastatic breast cancer cohort from linked cancer registry and EHR data. JAMIA Open, 2(4), pp.528–537.

Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A. and Kumar, S. 2021. Long-Tail Learning via Logit Adjustment. In: International Conference on Learning Representations (ICLR).

Miller, S., Sharoff, S. and Hall, G. 2024. Evaluating Text Pre-Processing Strategies for Clinical Document Classification with BERT. In: AIHealth 2024: The First International Conference on AI-Health the, pp.41–46.

Misra, I., Shrivastava, A., Gupta, A. and Hebert, M. 2016. Cross-Stitch Networks for Multi-task Learning. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3994–4003.

Moscato, V., Postiglione, F., Marano, S. et al. 2023. Multi-Task Learning for Few-Shot Biomedical Relation Extraction. Artificial Intelligence Review, 56(10), pp.11683–11705.

Mulyar, A., Uzuner, O. and McInnes, B. 2021. MT-clinical BERT: scaling clinical information extraction with multitask learning. JAMIA, 28(10), pp.2108–2115.

Peng, Y., Chen, Q. and Lu, Z. 2020. An empirical study of multi-task learning on BERT for biomedical text mining. In: ACL BioNLP, pp.205–214.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C. and Zhi, D. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured EHRs for disease prediction. NPJ Digital Medicine, 4(1), pp.86.

Reimers, N. and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp.3982–3992.

Ridnik, T., Ben-Baruch, E., Noy, A. and Friedman, I. 2021. Asymmetric Loss for Multi-Label Classification. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Rogers, A., Kovaleva, O. and Rumshisky, A. 2020. A primer in bertology: What we know about how bert works. Transactions of the Association for Computational Linguistics, 8, pp.842–866.

Steininger, M., Kobs, K., Kleeberger, K. and Hotho, A. 2021. Density-Based Weighting for Imbalanced Regression. In: Machine Learning and Knowledge Discovery in Databases (ECML PKDD), pp.201–216.

Strongman, H., Gadd, S., Matthews, A., Mansfield, K.E., Bhaskaran, K., Smeeth, L., Dos-Santos-Silva, I., de Lusignan, S. and Gallagher, A.M. 2020. What are the implications of using individual and combined sources of routinely collected data to identify and characterise incident site-specific cancers? A concordance and validation study using linked English electronic health records data. BMJ Open, 10(8), pp.e037719.

Sundararajan, M., Taly, A. and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML), pp.3319–3328.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. 2017. Attention Is All You Need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp.6000–6010.

Vickers, P., Barrault, L., Monti, E. and Aletras, N. 2024. We Need to Talk About Classification Evaluation Metrics in NLP. In: IJCNLP-AACL.

Wagner, G. 2002. History of Cancer Registration. Lyon, France: International Agency for Research on Cancer (IARC).

Whitfield, E., White, B., Barclay, M.E. et al. 2024. Differences in recording of cancer diagnosis between datasets in England: A population-based study of linked cancer registration, hospital, and primary care data. Cancer Epidemiology, 94, pp.102703.

Wiegreffe, S. and Pinter, Y. 2019. Attention is not Explanation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp.11–20.

Zhang, B., Wang, Y., Wu, Z., Wei, Y., Feng, J. and Zhang, Y. 2022. Distribution Alignment: Balancing Long-Tail via Adaptive Prior Correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.2361–2370.

Zhang, Y. and Yang, Q. 2022. A Survey on Multi-Task Learning. IEEE Transactions on Knowledge and Data Engineering, 34(12), pp.5586–5609.

Zhou, S., Wang, L., Wang, N., Liu, H. and Zhang, R. 2022. CancerBERT: A cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. Journal of the American Medical Informatics Association, 29(7), pp.1208–1216.