Evaluating Clinical BERT for Multiclass Pathology Report Classification with Interpretability

Umay Kulsoom[1,2], Malika Bendechache[1,2], and Frank G. Glavin[1]

[1] School of Computer Science, University of Galway, Ireland
[2] ADAPT Research Centre, University of Galway, Ireland

Abstract. Pathology reports are essential documents physicians use to establish a diagnosis and formulate a treatment plan for a specific health condition or disease. The significance of these reports is particularly pronounced in the context of cancer. The accurate classification of these reports is essential for optimising clinical decision-making, ensuring timely interventions, and maintaining high-quality patient care. In this work, we present two key contributions to improve the classification of pathology reports. First, we fine-tuned the Bio+Clinical BERT-based model for a multiclass classification approach that accurately distinguishes between 32 cancer tissues. Second, we have integrated explainability by using LIME to examine the interpretability of the BERT-based model's decisions and identified the domain-specific features that influence the classification results. We have demonstrated that high-performance transformer models can maintain transparency in clinical settings. Our interpretable framework enables pathologists to assess model outputs against established diagnostic criteria, facilitating the responsible integration of clinical language processing systems into clinical workflows.

Keywords: BERT, NLP, Pathology Reports, Interpretability, Text Classification, LIME.

1 Introduction

Cancer pathology reports are essential for accurate diagnosis and personalised treatment planning, serving as the gold standard in clinical oncology. Pathology reports provide comprehensive descriptions of the specimens analysed by pathologists, serving as a vital source of medical information for diagnosis, treatment planning, and prognostic assessment [12]. Integration of Artificial Intelligence (AI) has transformed the field of pathology with improvements in diagnosis precision, personalised treatment plans, and streamlining administrative tasks [27]. The efficient classification of these reports into relevant diagnostic categories has been a challenge in modern healthcare informatics, particularly as medical institutions transition toward fully digital workflows.

Machine Learning (ML) has become integral to clinical applications, enabling researchers to uncover patterns in free-text clinical records and develop predictive models for improved Clinical Decision Support System (CDSS)[33]. Traditional ML methods rely on extensive feature engineering, which might lead to errors, however, Deep Learning (DL) techniques have streamlined this process by automatically learning data representations [25]. Among the advancements in DL, one of the latest language models based on attention mechanisms, Bidirectional Encoder Representations from Transformers (BERT) [4], has emerged as a powerful tool in Natural Language Processing (NLP), offering improved performance in both multi-label and multiclass text classification [32]. Multiclass document classification involves assigning documents to multiple predefined categories/classes based on their textual content, enabling automated and efficient information organisation [10]. The complexity of medical texts, characterised by varying lengths, mixed data type text, and the inclusion of medical jargon, is challenging for effective classification models [9]. A single pathology report may include content spanning multiple diagnostic categories

David C. Wyld et al. (Eds): SAIM, SNLP, ACSIT – 2025
pp. 35-49, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.150703

[5], requiring sophisticated computational approaches for accurate and interpretable classification. BERT's ability to understand the complexity of medical language, including terminology and context, positions it as a promising candidate for developing multiclass classifiers [31]. It has the potential to facilitate the accurate categorisation of pathology reports into distinct diagnostic categories, thereby reducing the burden on pathologists.

While the efficacy of BERT in multiclass document classification is evident, the local interpretability and explainability of its predictions remain critical concerns [26]. The "black box" nature of medical AI systems can lead to errors in document classification that are difficult to identify and understand. If these errors go undetected, they may pose significant risks to patients [29]. Explainable classifications establish trust, ensure clinical validity, and facilitate quality control. This need for explainability is particularly acute in pathology, where classification decisions directly impact patient care and treatment decisions [2].

In this paper, we proposed a fine-tuned Bio+Clinical BERT-based model for multiclass classification of pathology reports into 32 cancer types using data from The Cancer Genome Atlas (TCGA) portal [11]. In addition, we have also integrated explainability by using LIME [20] without compromising classification performance. This integration offers valuable insights into the model's decision-making process, thereby promoting the responsible application of NLP in the healthcare domain.

The structure of this paper is as follows. Section 2 reviews the relevant literature, while Section 3 outlines the methodology used in this research. The results and key findings are discussed in Section 4, along with the study's limitations and future directions. Finally, Section 5 concludes the paper.

2 Related Work

With the exponential increase in the volume of clinical data, the need for efficient and accurate document classification systems has become increasingly critical [19]. Recent developments in NLP, especially with the development of models such as BERT, have demonstrated remarkable performance in accurate and automated classification, outperforming traditional and sequence models [30]. BERT-based models have been improved for multi-label and multi-class text classification using attention mechanisms and Long Short-Term Memory (LSTM) to process BERT-generated information [28]. In the earlier stages, challenges arise when applying BERT to long clinical texts, where simpler architectures like hierarchical self-attention networks sometimes outperform BERT-based approaches [7]. However, BERT has been fine-tuned for natural language inference in clinical trials, focusing on semantic representation and faithful reasoning [6]. To address imbalanced clinical text classification, the Multi-label Classification of Imbalanced Clinical Text (MCICIT) model combined BioBERT [13] with a novel graph convolutional network approach, achieving improved F1 scores on clinical datasets [8]. The Knowledge Graph Enhanced BERT for Multi-Type Medical Text Classification (KG-MTT-BERT model) integrates medical knowledge graphs with BERT to handle complex, multi-type medical texts, outperforming baselines in diagnosis-related group classification [9]. These studies demonstrate the ongoing evolution and effectiveness of BERT-based models in clinical text classification tasks.

With the demand for transparent and interpretable AI systems, Explainable AI (XAI) in BERT for clinical document classification has been explored. Gradient-based methods called integrated gradients were investigated with fine-tuned BERT for explainable medical image and neuroradiology protocol assignment [24,23]. To address trustworthi-

ness concerns, explainability techniques like Local Interpretable Model-Agnostic Explanations (LIME) [20] and SHapley Additive exPlanations (SHAP) [14] have been applied to Bio+Clinical BERT, improving the understanding of model strengths and weaknesses [22]. A framework for evaluating explanation quality using infidelity and local Lipschitz metrics has been proposed, allowing assessment of the trade-off between predictive performance and explanation quality across various model types, including BERT variants [17]. These studies highlight the importance of balancing performance and explainability in clinical text classification tasks, particularly when dealing with domain-specific terminology.

Tissue-type classification using genomic and clinical data is essential in cancer research, as it is vital in facilitating precise diagnosis and personalised treatment strategies. In this context, the work of J. Kefeli et al. [11] stands out, as they achieved remarkable results in binary tissue type classification across 32 different tissues, reporting an Average Area Under the Receiver Operating Characteristic curve (AU-ROC) of 0.992. This high AU-ROC underscores the potential of advanced DL techniques in improving diagnostic capabilities. However, while binary classification provides valuable insights, it does not fully capture the complexities of real-world clinical scenarios, where multiclass tissue type classification is often required. We aim to extend the findings of J. Kefeli et al. [11] by addressing this more clinically relevant challenge. We focus on multiclass tissue type classification, particularly in the context of imbalanced datasets that mimic the distribution of tissue types encountered in actual clinical practice. Such imbalances can pose significant challenges for classification algorithms, potentially leading to biased predictions and suboptimal clinical outcomes. In addition, we made an effort to interpret the clinical decisions made by the BERT model with the help of the LIME XAI technique. LIME is designed to provide local interpretability by generating explanations for individual predictions made by complex models. By perturbing the input data and observing the effects on model predictions [20], LIME identifies the most influential features, such as specific words or phrases, contributing to the model's decisions.

3 Methodology

Our approach uses a BERT-based architecture for the cancer-type multiclass classification of pathology reports. The methodology takes advantage of the powerful contextual understanding capabilities of Bio+Clinical-BERT while incorporating an explainability technique to improve model transparency in medical text classification. The pipeline integrates data preparation, fine-tuning strategies, and explainability methods to create an interpretable and accurate classification system.

3.1 Pathology Reports Dataset

The dataset used in this study consists of pathology reports from '*The Cancer Genome Atlas*' (TCGA) portal [11], a valuable resource for training AI models in text-based cancer research. The dataset comprises 9,523 machine-readable pathology reports spanning 32 cancer types. The pathology reports contain detailed descriptions of tissue samples, including tumour characteristics and diagnostic information. The distribution of tissues across the reports is represented in Figure 1. The abbreviation of each tissue sample, which can be found on the TCGA portal, is provided in Table 1 for the convenience of the readers.

Table 1. 32 Cancer Type Abbreviations

ACC	Adrenocortical Carcinoma	BLCA	Bladder Urothelial Carcinoma
BRCA	Breast Invasive Carcinoma	CESC	Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma
CHOL	Cholangiocarcinoma	COAD	Colon Adenocarcinoma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	ESCA	Esophageal Carcinoma
GBM	Glioblastoma Multiforme	HNSC	Head and Neck Squamous Cell Carcinoma
KICH	Kidney Chromophobe	KIRC	Kidney Renal Clear Cell Carcinoma
KIRP	Kidney Renal Papillary Cell Carcinoma	LGG	Brain Lower Grade Glioma
LIHC	Liver Hepatocellular Carcinoma	LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma	MESO	Mesothelioma
OV	Ovarian Serous Cystadenocarcinoma	PAAD	Pancreatic Adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma	PRAD	Prostate Adenocarcinoma
READ	Rectum Adenocarcinoma	SARC	Sarcoma
SKCM	Skin Cutaneous Melanoma	STAD	Stomach Adenocarcinoma
TGCT	Testicular Germ Cell Tumors	THYM	Thymoma
THCA	Thyroid Carcinoma	UCEC	Uterine Corpus Endometrial Carcinoma
UCS	Uterine Carcinosarcoma	UVM	Uveal Melanoma

3.2 Text Preprocessing and Tokenisation

The TCGA-Reports were already preprocessed by the dataset creators to remove all Protected Health Information (PHI) and are standardised for machine readability. Our additional preprocessing involved removing characters such as double and single quotation marks and contraction replacement to optimise the text for the BERT-based classification model and explainability analysis.

Each pathology report was tokenised using the BERT tokeniser, adhering to its vocabulary and ensuring a uniform sequence length of 512 tokens. The tokenisation process involved splitting text into subword units, which is essential for language processing by the BERT model.

4 Proposed Approach

The foundation of our approach is based on Bio+ClinicalBERT [1], which is initialised from BioBERT. ClinicalBERT is a multi-layer bidirectional transformer encoder that has been specifically fine-tuned for clinical text processing with the capacity to understand and analyse medical language effectively. A pre-trained BERT model consists of 12 transformer layers with 12 attention heads each and a hidden dimension size of 768. We implemented a custom BERTClassifier class using the PyTorch Lightning framework for streamlined training and evaluation. This approach enabled a standardised implementation that facilitates reproducible experimentation across multiple computational environments.

The classification head consists of:

1. A dropout layer with a rate of 0.1 to mitigate overfitting
2. A linear transformation layer that projects the 768-dimensional hidden representation to a 32-dimensional output space
3. A softmax activation function that transforms the model outputs into probability scores across all 32 cancer types.

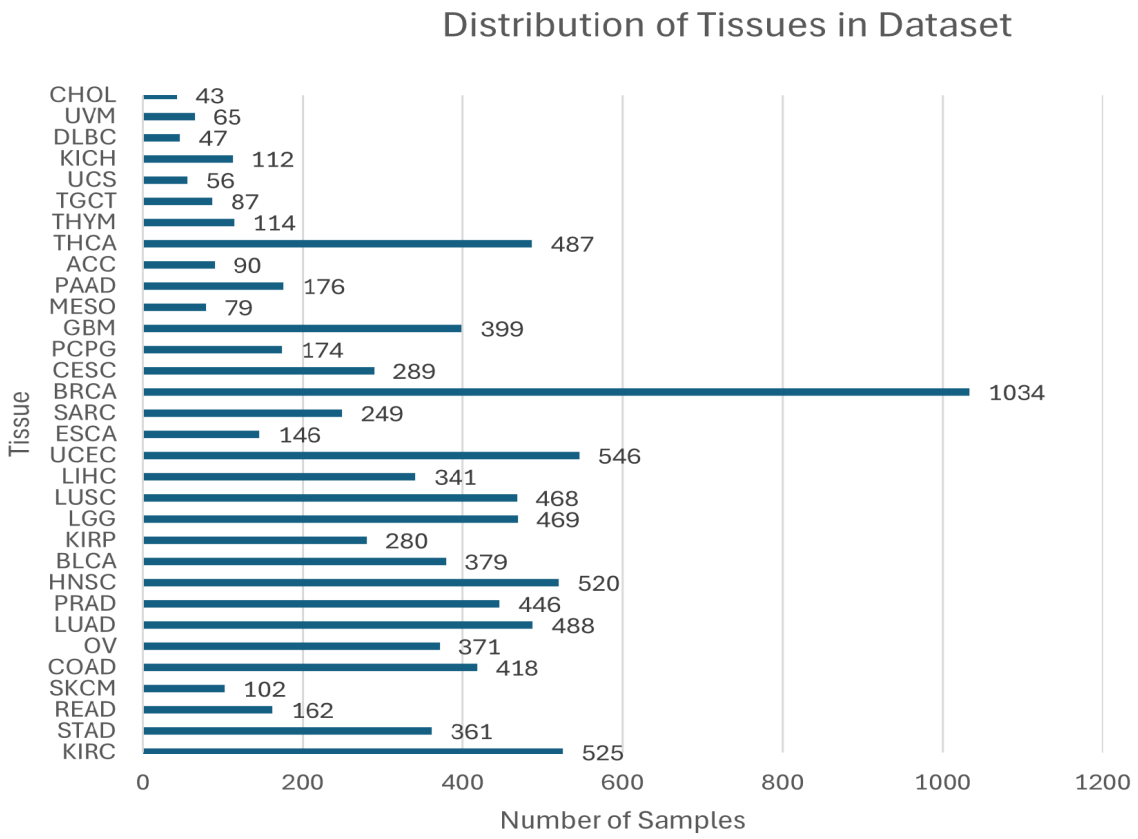

Fig. 1. Distribution of Tissue Subcategories in the TCGA Dataset

The model was configured to address classification, optimising the cross-entropy loss function for the multiclass classification scenario. This configuration played an important role, given the diverse but distinct nature of the 32 cancer types represented in the TCGA pathology reports.

The architecture was specifically designed to balance the competing requirements of classification accuracy, computational efficiency, and model interpretability, making it particularly suited for the challenging task of automated cancer-type classification from complex pathology reports. By maintaining attention weights for explainability while optimising classification performance, our model provides accurate predictions and insights into the reasoning behind those predictions.

An overview of our approach is detailed in Figure 2.

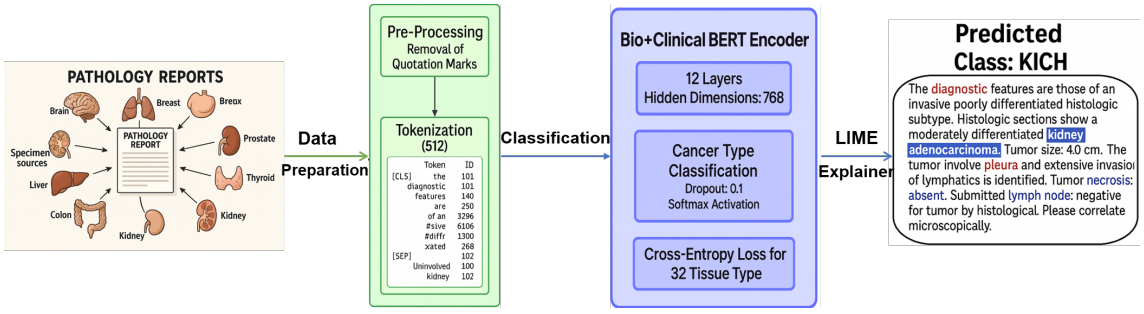

Fig. 2. Overview of the Proposed Approach

4.1 Implementation Details

In our implementation, we used PyTorch 2.3.0, the primary DL framework, which allowed us to develop and train the model flexibly and efficiently. To work with pre-trained BERT models and fine-tune them for our specific classification task, we used Hugging Face Transformers 4.48.3, which enabled the straightforward adaptation of the Bio+ClinicalBERT model.

For cross-validation and model evaluation, scikit-learn 1.5.2 was used, while the computational demands of training transformer-based models were met by CUDA 12.2 for GPU acceleration. Our training was conducted on NVIDIA RTX 6000 Ada Generation GPUs, each with 49GB of memory. This setup enabled us to process the dataset of pathology reports while keeping training times manageable.

4.2 Model Training

To prepare the data for model training, we divided the input dataset into five distinct folds. We use a stratified approach so that the distribution of labels is consistent across each fold to ensure a balanced representation of the 32 tissue types across the folds. For each fold, we used LabelEncoder to convert the categorical labels into a numerical format, making them suitable for model training. A sample of k-fold splitting is illustrated in Figure 3.

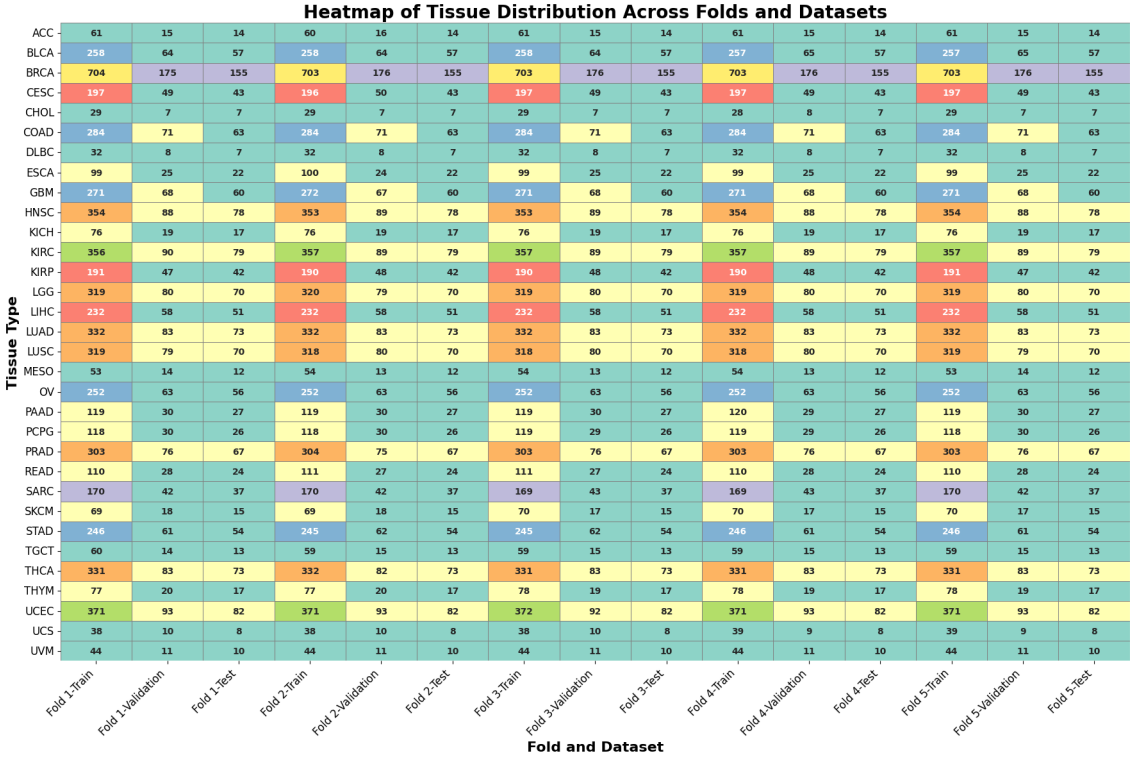

Fig. 3. Sample of K-Fold Splitting

The training process was managed using the Trainer component of PyTorch Lightning, allowing efficient handling of the training workflow. We used GPU acceleration whenever

possible to expedite the training process. The model was trained on the training data, validated on a separate validation set, and finally tested on a held-out test set to assess its performance. During tokenisation, we ensured that each report was either truncated or padded to a maximum length of 512 tokens, which is the input size limit for BERT. For each fold in our cross-validation setup, the model was initialised with pre-trained BERT weights to use the linguistic knowledge acquired during pre-training. The fine-tuning process focused on the training portion, which comprised 85% of the available data, allowing the model to adapt to the specific language patterns of the pathology reports. To prevent overfitting, we implemented an early stopping mechanism that monitored validation loss, ensuring that training terminated when the performance plateaued. Additionally, we applied gradient clipping at a maximum norm of 1.0 to maintain training stability by preventing exploding gradients during backpropagation. To enhance computational efficiency without sacrificing model performance, we employed mixed precision training (FP16), which reduced memory usage and accelerated the training process while maintaining numerical precision where critical.

The model was optimised using the AdamW optimiser, which combines the Adam algorithm with proper weight decay regularisation. We used a relatively low learning rate of 2e-5 to fine-tune the pre-trained BERT parameters. This optimisation strategy allowed for effective transfer learning, where the general language understanding capabilities of BERT are preserved while adapting the model to the specific task of cancer-type classification from pathology reports. The weight decay component helped prevent overfitting, which is particularly important given the specialised vocabulary and structure of the pathology reports. The hyperparameter configuration is provided in Table 2.

Table 2. Hyperparameter Configuration

Parameter	**Value**
Batch Size	32
Learning Rate	2e-5 with linear decay
Maximum Sequence Length	512
Training Epochs	10 for each fold
Optimizer	AdamW with weight decay of 0.01
Warmup	10% of total training steps

To improve training, we implemented a linear learning rate scheduler with a warmup phase that spanned the first 10% of the training steps. The model underwent training for a total of 10 epochs, with early stopping criteria based on validation performance to prevent overfitting and ensure that the model generalises well to unseen data.

4.3 Evaluation Metrics

To evaluate the effectiveness of our proposed model, we calculated a comprehensive suite of performance metrics. These included overall accuracy, balanced accuracy, macro-averaged AUROC (Area Under the Receiver Operating Characteristic curve), and per-class AUROC. In addition, a classification report was also generated, providing detailed insights into precision, recall, and F1-score for each class for the model's performance assessment across different tissue types.

Logging and Reporting Throughout the training and evaluation process, we logged all performance metrics using Weights & Biases (wandb), a powerful tool for experiment tracking and visualisation. This logging functionality enabled us to capture a comprehensive collection of metrics for each fold, as well as average metrics across all folds. Such detailed tracking allows for an in-depth analysis of the model's performance, facilitating continuous improvement and refinement of our approach.

4.4 Explainability Integration

Recognising the importance of transparency in clinical applications, we incorporated explainability into our model. Model-specific and model-agnostic methods have been categorised as two main categories of XAI techniques used in medical text classification [16]. Local Interpretable Model-agnostic Explanation (LIME) has been advocated to be more suitable for text-based models [16]. Therefore, we applied LIME to provide insights into how the text classification model arrives at its predictions.

LIME is a model-agnostic technique that approximates the decision boundary of a complex model locally using a simpler, interpretable model [18]. This approach helps identify which features (in this case, words) significantly influence the model's predictions. LIME works by perturbing the input data and observing the corresponding changes in the model's predictions.

A linear model was fitted to these samples, where each sample was weighted based on its proximity to the original instance. The objective function minimises the sum of the loss associated with the complex model and a measure of the complexity of the explanatory model. The coefficients of this linear model represented the contribution of each word to the prediction, effectively approximating the partial derivatives of the prediction function for each feature.

The framework included a prediction function wrapper that accepted raw text as input, processed it through the BERT tokeniser, and generated probability distributions across 32 cancer classes using the softmax function. The LIME explainer created permutations of the original text by randomly removing words and maintained a dataset of these permutations along with their corresponding predictions. In our LIME implementation, we generated 1000 perturbed samples for each text instance being explained. This number of samples provides a robust statistical basis for approximating the local decision boundary of our model. For the scope of application, LIME was applied to a stratified subset of the test set, comprising both correctly and incorrectly classified examples (top 20 of each). This approach allowed us to examine the behaviour of the model in a representative set of cases while maintaining computational efficiency. The selection strategy ensured coverage across different cancer types, enabling us to identify domain-specific patterns in the model's decision-making process for both successful predictions and error cases. A kernel function weights samples based on their similarity to the original text. For each text sample, the framework identified the most significant words contributing to classification decisions, quantifying both the magnitude and direction of each word's influence and calculating an impact percentage representing each word's relative importance.

5 Results and Discussion

To evaluate the performance of our custom Bio+ClinicalBERT-based classifier and to ensure a fair comparison with the baseline work by J. Kefeli et al. [11], we used the Area Under the Receiver Operating Characteristic curve (AU-ROC) as our primary evaluation

metric. The choice of AU-ROC is driven by two main considerations. First, since the baseline study [11] used AU-ROC as their evaluation metric, we tried to maintain consistency by adopting the same standard. Second, given the imbalanced nature of the dataset — consistent with the approach of the original contributors — AU-ROC has been recommended by [21] as a suitable evaluation metric for such scenarios.

Our custom Bio+ClinicalBERT-based classifier demonstrated remarkable performance in distinguishing among all 32 cancer types, achieving an impressive average AU-ROC score of 0.997, as illustrated in Figure 4. This nearly perfect score underscores the model's strong capability to differentiate between various cancer types based on the text extracted from pathology reports. The performance of our model is comparable to that of the binary classification of pathology reports presented in [11], which achieved an average AU-ROC of 0.992. To rigorously evaluate the model's discriminative ability across multiple tissue types, we calculated the AU-ROC scores using a stratified 5-fold cross-validation approach. This methodology ensured reliable performance estimation while mitigating the effects of data variation.

In addition, we computed the standard deviation across folds for each tissue type to measure model stability. To effectively communicate the distribution of AU-ROC performance across tissue types, we developed a hierarchical visualisation approach that categorised tissues into performance tiers. The perfect performance tier (AU-ROC = 1.0) includes tissue types where the model achieved perfect discrimination across all folds with zero standard deviation. The Excellent Performance tier (AU-ROC$\geq$0.99) comprises tissue types with near-perfect discrimination, while the Very Good tier (AU-ROC$\geq$0.97) includes those with strong discriminative performance. Tissue types that still have room for improvement fall into the Good Performance tier (AU-ROC $<$ 0.97). This tiered visualisation approach effectively manages the presentation of high-dimensional classification results, emphasising meaningful performance distinctions rather than negligible numerical differences, and allows for the immediate identification of exceptional performers and potential areas for improvement.

We not only evaluated the model's performance using AU-ROC, as done by [11], but also did a comprehensive performance analysis, as detailed below.

5.1 Cross-Validation Performance Analysis

Figure 5 presents the aggregate confusion matrix across all 5 folds of our cross-validation experiment for each of the 32 tissue types. The overall model achieved a mean accuracy of 0.97 across all folds, with a balanced accuracy of 0.966, indicating robust performance despite the class imbalance in the dataset.

The model's performance remained consistent across all five folds, with the individual fold accuracies ranging from 0.964 to 0.972. This consistency indicates that the model's learning ability is reliable and not overly affected by data splitting. Such stability establishes the strength and efficacy of our approach in contrast to Binary Classification [11].

5.2 Performance Consistency

The performance metrics for a multiclass classification model are detailed in Table 3. Each row represents a distinct cancer tissue evaluated for three standard metrics, i.e. Precision, Recall, and F1-score. Tissue types like BRCA, CHOL, LIHC, MESO, PCPG, PRAD, TGCT, and UVM achieved perfect scores across all three metrics. Most classes have F1

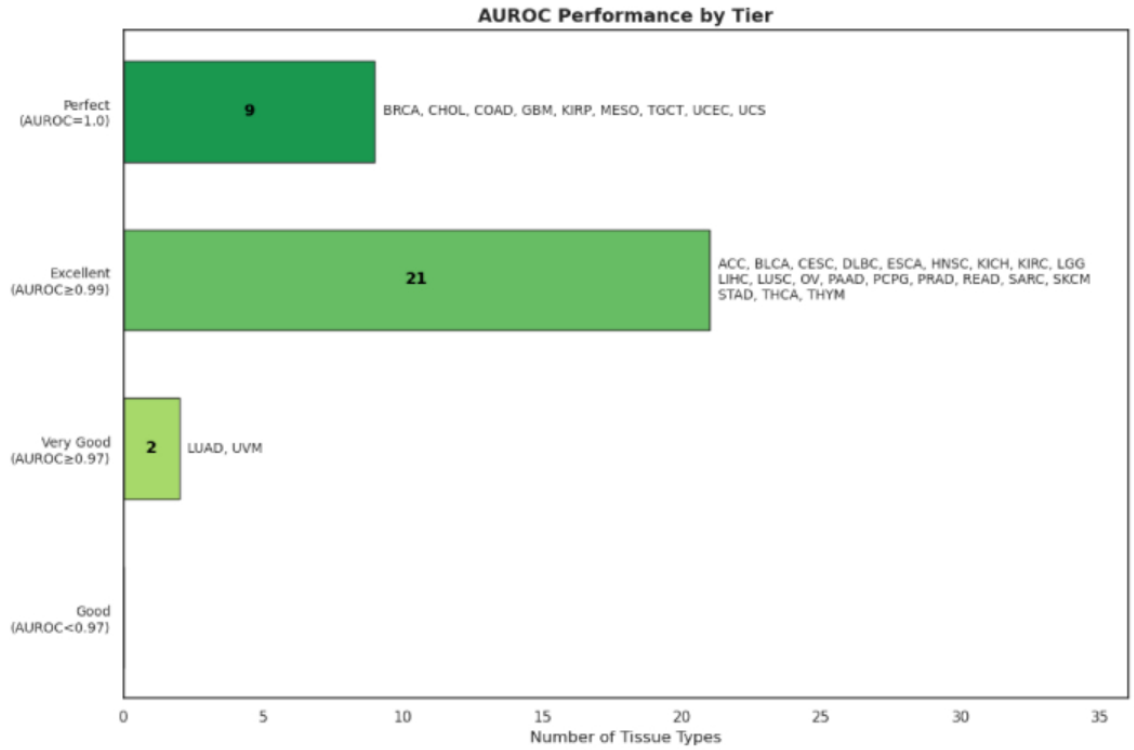

Fig. 4. AUROC Performance Across Folds

scores above 0.9, indicating strong overall performance. Only a few classes have shown notably lower performance.

While the overall results are promising with our custom-built Bio+ClinicalBERT classifier, we acknowledge that certain classes, such as READ and UCS, demonstrated lower performance, suggesting the need for further fine-tuning to enhance classification accuracy for these specific tissue types. The high performance across most classes suggests that the model effectively captures the distinctive features and patterns in the pathology reports for accurate multiclass classification.

5.3 Explainability Results

We implemented a visualisation method that highlights the original text using colour-coded words, where the intensity of the background colour signifies the importance of each feature, in this case, a word. Figure 6 presents a sample of such visualisation, with a blue background indicating words that have a positive influence and a red background representing words that contribute negatively.

Our analysis of the LIME explainability covered correctly and incorrectly classified samples, thereby providing a dual perspective on the model's performance. Figure 6 represents the word-level LIME explanation and highlights how specific medical terms relate to the model's predictions. For example, the term "Chromphobe" plays a crucial role in classifying the outcome as 'KICH', with a contribution value higher than other words in the text, indicating its importance in distinguishing KICH from KIRP, another form of kidney cancer. [For readers without a medical background: Chromophobe refers to chromophobe renal cell carcinoma, which is the same cancer represented by KICH, but distinct from KIRP [15,3]]. A similar analysis was carried out for the incorrectly classified instances, too.

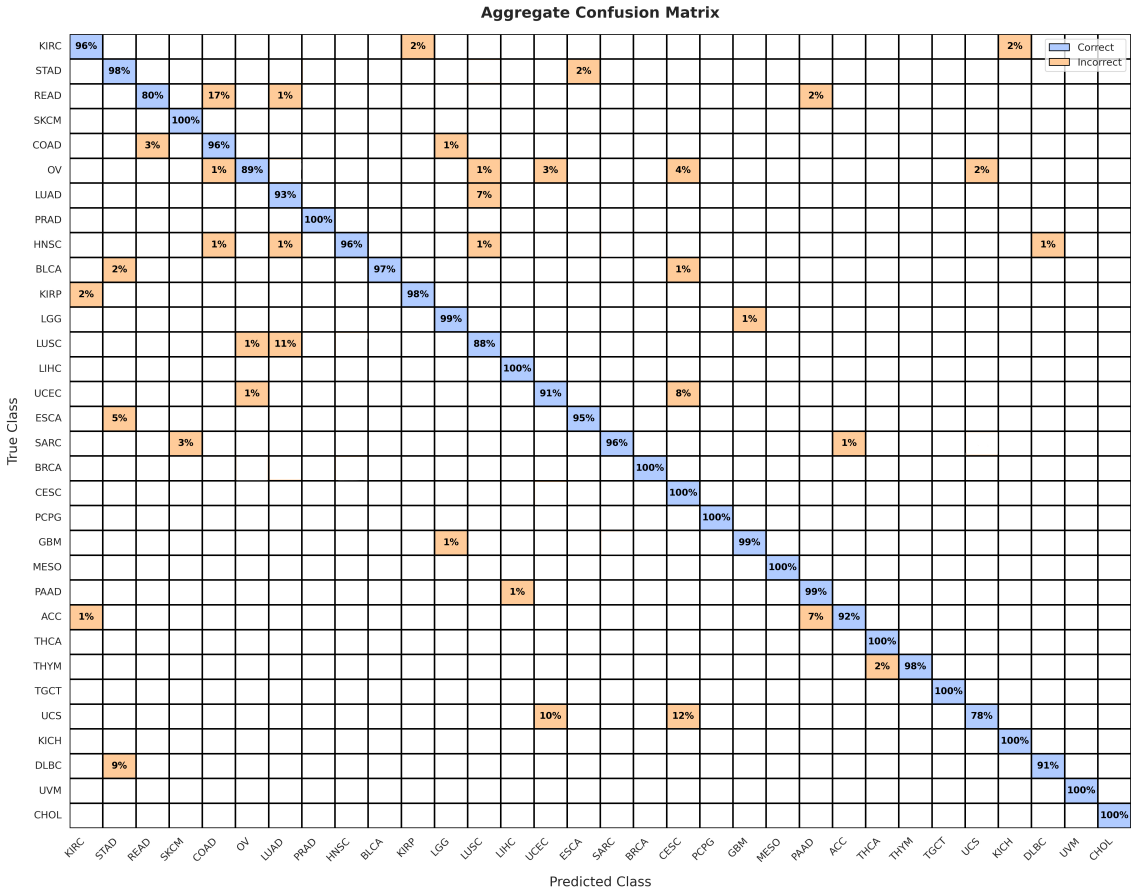

Fig. 5. Aggregate Confusion Matrix

The individual contributions of terms were visualised through a bar plot, which provided insights into the Bio+ClinicalBERT model's decision-making. Figure 7 represents the bar plot of one of the samples with the top contributing words plotted only. The term "chromophobe" appeared twice, each instance exhibiting different contribution values. This duplication highlights the contextual understanding of these pre-trained models in medical AI. The same term carries varying significance depending on where and how it appears in the clinical text. Unlike traditional approaches, such as the Bag-of-Words, the vector representation of each instance of a word is based on the surrounding text, sentence structure, and position within a document.

Similarly, terms such as "kidney" and "renal" contributed positively, while words like "perinephric" and "small" contributed negatively towards classification. These differen-

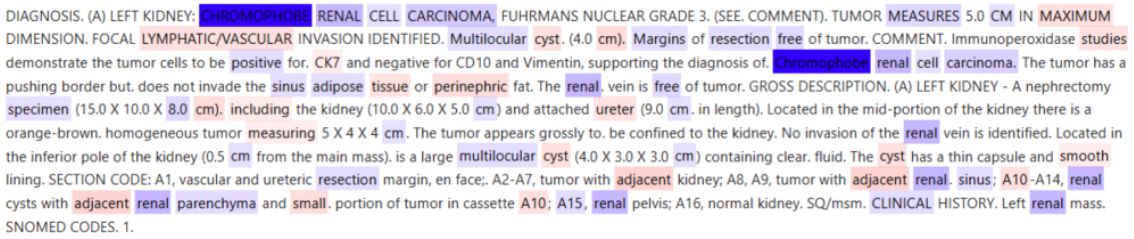

Fig. 6. Explainability of Correctly Classified Class - **KICH**

Table 3. Performance Metrics for Each Tissue Class

Class	**Prec.**	**Rec.**	**F1**	**Class**	**Prec.**	**Rec.**	**F1**
KIRC	0.99	0.96	0.97	SARC	0.98	0.96	0.97
STAD	0.95	0.98	0.97	BRCA	1.00	1.000	1.00
READ	0.90	0.80	0.85	CESC	0.80	1.00	0.89
SKCM	0.94	1.00	0.97	PCPG	1.00	1.00	1.00
COAD	0.92	0.96	0.94	GPM	0.99	0.99	0.99
OV	0.98	0.89	0.93	MESO	1.00	1.00	1.00
LUAD	0.89	0.93	0.91	PAAD	0.94	0.99	0.97
PRAD	1.00	1.00	1.00	ACC	0.97	0.91	0.94
HNSC	0.99	0.96	0.98	THCA	0.99	1.00	1.00
BLCA	1.00	0.98	0.99	THYM	1.00	0.98	0.99
KIRP	0.97	0.98	0.97	TGCT	1.00	1.00	1.00
LGG	0.99	0.99	0.99	UCS	0.82	0.78	0.79
LUSC	0.91	0.88	0.90	KICH	0.90	1.00	0.95
LIHC	1.00	1.00	1.00	DLBC	0.86	0.91	0.89
UCEC	0.96	0.91	0.94	UVM	1.00	1.00	1.00
ESCA	0.96	0.95	0.96	CHOL	1.00	1.00	1.00

tial contributions represent how contextual embedding effectively identifies the semantic complexity inherent in clinical language, where the same term can convey distinct meanings and weights depending on its context. This pattern of word contributions reflects the decision-making processes of pathologists, who consider and give importance to the diagnostic findings based on the section of the report where they appear. Moreover, it also gives a direction to explore further and compare the explainability techniques for the development of transparent frameworks that facilitate the integration of AI tools into clinical practice.

This approach facilitated an understanding of how the BERT model processes pathology-specific vocabulary, identifies key diagnostic terms that drive cancer classification, and compares classification patterns between correctly and incorrectly classified reports. It also validated whether the model prioritises clinically relevant features. The "black box" nature of BERT is addressed with a mathematical approximation of feature importance in a format that is interpretable to human users. Such interpretability is essential for building trust in clinical AI applications and facilitating error analysis.

5.4 Limitations and Future work

The TCGA pathology reports dataset offers a substantial corpus for training our classification model, however, the distribution of samples across the 32 cancer types is not uniform, potentially leading to classification bias favouring more prevalent cancer types. We did not balance the dataset before classification to do a fair comparison of results. One limitation of our approach is the absence of an external validation dataset. Cross-institutional variations in pathology reporting conventions, terminology preferences, and formatting styles may challenge the model's generalisability beyond the TCGA framework. In the next step, we will attempt to acquire a similar dataset of pathology reports that is not otherwise publicly available, to the best of our knowledge.

The BERT-based architecture limits the input sequence length to 512 tokens, which necessitates the truncation of longer pathology reports, potentially resulting in the loss of clinically relevant information. In future work, we plan to explore the impact of different token size limits and investigate both left and right truncation strategies. Moreover, we will explore alternative architectures such as Clinical BigBird or Clinical Longformer models, specifically designed to handle longer text sequences.

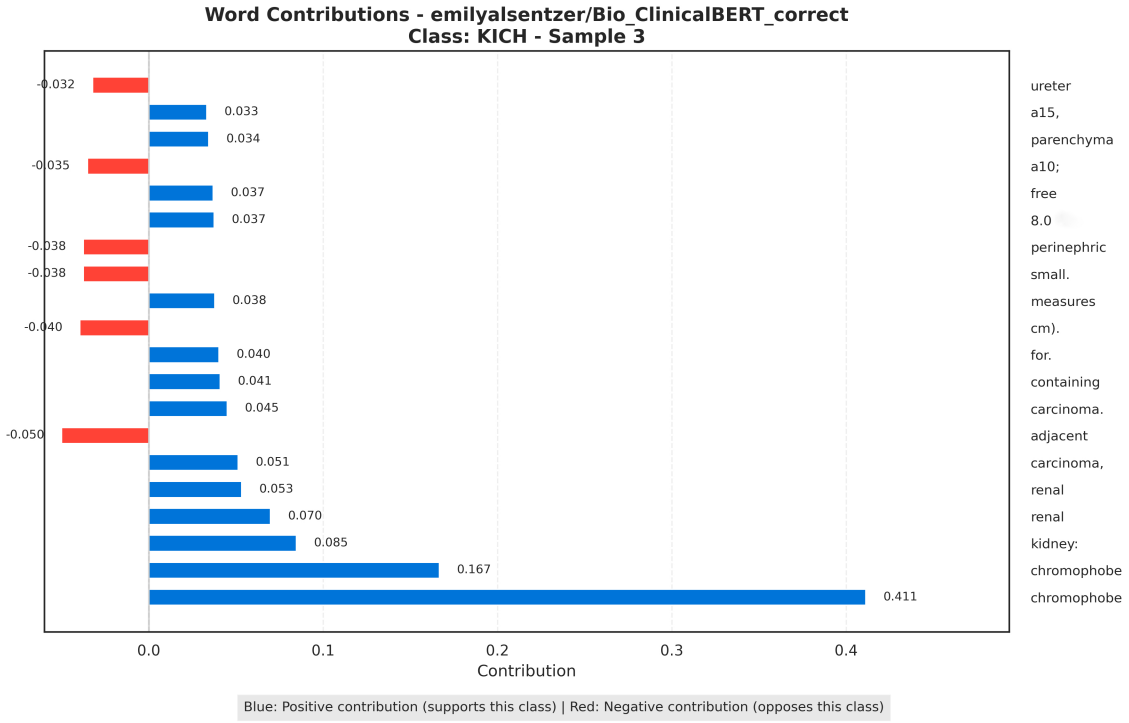

Fig. 7. Words Contribution for Correctly Classified Class- **KICH**

6 Conclusion

In our work, we have demonstrated the effectiveness of our custom-built Bio+Clinical-BERT classifier for the multiclass classification of 32 cancer tissue types. The stability of our results establishes the reliability of the model's learning capabilities and highlights the strength of our approach compared to traditional binary classification methods. However, we observed that certain cancer tissue types, specifically READ, UCS, and LUAD, performed relatively poorly. It appears to be Clinical-BERT's tendency to confuse these cancer types with others, such as misclassifying READ as COAD or LUAD as LUAS, particularly when the prevalence of these types is significantly imbalanced. In future, including clinical notes for these low-prevalence cancer types can be considered while exploring different models and tokenisers to improve performance. Another interesting area of research would be to study the impact of tokeniser size and the effect of truncation on individual classification. By addressing these challenges, we can further improve the accuracy and reliability of cancer tissue classification, ultimately contributing to better diagnostic outcomes in clinical practice.

Acknowledgement. This research was funded by Higher Education Authority (HEA) Ireland and supported by Research Ireland under Grant No 13/RC/2106_P2 (ADAPT). For Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission

References

1. Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clin-*

ical Natural Language Processing Workshop, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

2. Bernie Croal. The communication of critical and unexpected pathology results. *The Royal College of Pathologists*, 2017.
3. Caleb F Davis, Christopher J Ricketts, Min Wang, Lixing Yang, Andrew D Cherniack, Hui Shen, Christian Buhay, Hyojin Kang, Sang Cheol Kim, Catherine C Fahey, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer cell*, 26(3):319–330, 2014.
4. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
5. C Eloy, P Seegers, E Bazyleva, and F Fraggetta. The 1 million words pathology report or the challenge of a reproducible and meaningful message. *ESMO Real World Data and Digital Oncology*, 4:100044, 2024.
6. Anass Fahfouh, Abdessamad Benlahbib, Jamal Riffi, and Hamid Tairi. Usmba-nlp at semeval-2024 task 2: Safe biomedical natural language inference for clinical trials using bert. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 432–436, 2024.
7. Shang Gao, Mohammed Alawad, M Todd Young, John Gounley, Noah Schaefferkoetter, Hong Jun Yoon, Xiao-Cheng Wu, Eric B Durbin, Jennifer Doherty, Antoinette Stroup, et al. Limitations of transformers on clinical text classification. *IEEE journal of biomedical and health informatics*, 25(9):3596–3607, 2021.
8. Yao He, Qingyu Xiong, Cai Ke, Yaqiang Wang, Zhengyi Yang, Hualing Yi, and Qilin Fan. Mcict: Graph convolutional network-based end-to-end model for multi-label classification of imbalanced clinical text. *Biomedical Signal Processing and Control*, 91:105873, 2024.
9. Yong He, Cheng Wang, Shun Zhang, Nan Li, Zhaorong Li, and Zhenyu Zeng. Kg-mtt-bert: Knowledge graph enhanced bert for multi-type medical text classification. *arXiv preprint arXiv:2210.03970*, 2022.
10. WH Inmon, Daniel Linstedt, and Mary Levins. *Data Architecture: A Primer for the Data Scientist: A Primer for the Data Scientist*. Academic Press, 2019.
11. Jenna Kefeli and Nicholas Tatonetti. Tcga-reports: A machine-readable pathology report resource for benchmarking text-based ai models. *Patterns*, 5(3), 2024.
12. Jeongeun Lee, Hyun-Je Song, Eunsil Yoon, Seong-Bae Park, Sung-Hye Park, Jeong-Wook Seo, Peom Park, and Jinwook Choi. Automated extraction of biomarker information from pathology reports. *BMC medical informatics and decision making*, 18:1–11, 2018.
13. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
14. Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.
15. Yueji Luo, Danna Chen, and Xiao-Liang Xing. Comprehensive analyses revealed eight immune related signatures correlated with aberrant methylations as prognosis and diagnosis biomarkers for kidney renal papillary cell carcinoma. *Clinical Genitourinary Cancer*, 21(5):537–545, 2023.
16. Ibrahim Alaa Eddine Madi, Akram Redjdal, Jacques Bouaud, and Brigitte Seroussi. Exploring explainable ai techniques for text classification in healthcare: A scoping review. *Digital Health and Informatics Innovations for Sustainable Health Care Systems*, pages 846–850, 2024.
17. Mitchell Naylor, Christi French, Samantha Terker, and Uday Kamath. Quantifying explainability in nlp and analyzing algorithms for performance-explainability tradeoff. *arXiv e-prints*, pages arXiv–2107, 2021.
18. Sai Ram Aditya Parisineni and Mayukha Pal. Enhancing trust and interpretability of complex machine learning models using local interpretable model agnostic shap explanations. *International Journal of Data Science and Analytics*, 18(4):457–466, 2024.
19. Anjani Kumar Rai, Upendra Singh Aswal, Suresh Kumar Muthuvel, Akhil Sankhyan, S Lakshmana Chari, and A Kakoli Rao. Clinical text classification in healthcare: Leveraging bert for nlp. In *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)*, volume 1, pages 1–7. IEEE, 2023.
20. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.
21. Eve Richardson, Raphael Trevizani, Jason A Greenbaum, Hannah Carter, Morten Nielsen, and Bjoern Peters. The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns*, 5(6), 2024.

22. Aditi Sankaranarayanan, Diya Shetty, Kirti Chetwani, and BR Shambhavi. Exploring bioclinical bert's nlp capabilities with explainability techniques. In *2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS)*, pages 1–6. IEEE, 2024.
23. Salmonn Talebi, Elizabeth Tong, Anna Li, Ghiam Yamin, Greg Zaharchuk, and Mohammad RK Mofrad. Exploring the performance and explainability of fine-tuned bert models for neuroradiology protocol assignment. *BMC Medical Informatics and Decision Making*, 24(1):40, 2024.
24. Salmonn Talebi, Elizabeth Tong, and Mohammad RK Mofrad. Exploring the performance and explainability of bert for medical image protocol assignment. *medRxiv*, pages 2023–04, 2023.
25. Mohammad Mustafa Taye. Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers*, 12(5):91, 2023.
26. Multi-Modal Team. Evaluating text pre-processing strategies for clinical document classification with bert. 2024.
27. Alon Vigdorovits, Maria Magdalena Köteles, Gheorghe-Emilian Olteanu, and Ovidiu Pop. Breaking barriers: Ai's influence on pathology and oncology in resource-scarce medical systems. *Cancers*, 15(23):5692, 2023.
28. Haojia Wu, Xinfeng Ye, and Sathiamoorthy Manoharan. Enhancing multi-class text classification with bert-based models. In *2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6. IEEE, 2023.
29. Hanhui Xu and Kyle Michael James Shuttleworth. Medical artificial intelligence and the black box problem: a view based on the ethical principle of "do no harm". *Intelligent Medicine*, 4(1):52–57, 2024.
30. Myriam Youssef, Mervat Abu-Elkheir, and Maggie Mashaly. Automating clinical document classification: Ai solutions for enhanced healthcare decision support. In *2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, pages 345–348. IEEE, 2024.
31. Ken G Zeng, Tarun Dutt, Jan Witowski, GV Kranthi Kiran, Frank Yeung, Michelle Kim, Jesi Kim, Mitchell Pleasure, Christopher Moczulski, L Julian Lechuga Lopez, et al. Improving information extraction from pathology reports using named entity recognition. *Research Square*, pages rs–3, 2023.
32. Xu Zhang, Zejie Liu, Yanzheng Xiang, and Deyu Zhou. Complicate then simplify: a novel way to explore pre-trained models for text classification. In *Proceedings of the 29th international conference on computational linguistics*, pages 1136–1145, 2022.
33. Yinyuan Zhang, Ricardo Henao, Zhe Gan, Yitong Li, and Lawrence Carin. Multi-label learning from medical plain text with convolutional residual models. In *Machine Learning for Healthcare Conference*, pages 280–294. PMLR, 2018.

Authors

U. Kulsoom is pursuing a PhD in Natural Language Processing at the School of Computer Science, University of Galway, Ireland

Dr. M. Bendechache is an Assistant Professor at the School of Computer Science, University of Galway, and a SFI Funded Investigator with the SFI ADAPT and Lero research centres. She holds a PhD in data analytics from the Insight Centre for Data Analytics at University College Dublin, Ireland. Her expertise spans the areas of Big Data Analytics, Machine Learning, Computer Vision in Healthcare, and AI & Data Governance.

Dr. F.G. Glavin is an Assistant Professor in the School of Computer Science, University of Galway. He holds a PhD in Artificial Intelligence and an MSc by research in Machine Learning from the University of Galway, Ireland. He is the Programme Director for the university's Advanced MSc in Computer Science (Data Analytics). His research interests include Machine Learning, Computer Science Education, Game AI, Reinforcement Learning, Dynamic Difficult Adjustment, and Computer Vision.