Biomedical Physics & Engineering Express

PAPER

KRASFormer: a fully vision transformer-based framework for predicting *KRAS* gene mutations in histopathological images of colorectal cancer

Vivek Kumar Singh[1,2], **Yasmine Makhlouf**[1], **Md Mostafa Kamal Sarker**[3], **Stephanie Craig**[1], **Juvenal Baena**[1], **Christine Greene**[1], **Lee Mason**[1], **Jacqueline A James**[1,4], **Manuel Salto-Tellez**[1,4,5,6], **Paul O'Reilly**[5] and **Perry Maxwell**[1,7]

[1] Precision Medicine Centre of Excellence, Health Sciences Building, The Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Belfast, BT9 7AE, United Kingdom
[2] Centre for Biomarkers and Biotherapeutics, Barts Cancer Institute, Queen Mary University of London, London, EC1M 6BQ, United Kingdom
[3] Institute of Biomedical Engineering, University of Oxford, Oxford, OX3-7DQ, United Kingdom
[4] Regional Molecular Diagnostic Service, Belfast Health and Social Care Trust, Belfast, BT9 7AE, United Kingdom
[5] Sonrai Analytics, Belfast, BT9 7AE, United Kingdom
[6] Cellular Pathology, Belfast Health and Social Care Trust, Belfast City Hospital, Lisburn Road, Belfast BT9 7AB, United Kingdom
[7] Present address: Precision Medicine Centre of Excellence, Health Sciences Building, The Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Belfast, BT9 7AE, United Kingdom.

E-mail: p.maxwell@qub.ac.uk

Keywords: colorectal cancer, vision-transformer, KRAS gene, whole slide image, mutation classification, next generation sequencing

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
29 January 2024

REVISED
27 May 2024

ACCEPTED FOR PUBLICATION
26 June 2024

PUBLISHED
17 July 2024

Abstract

Detecting the Kirsten Rat Sarcoma Virus (*KRAS*) gene mutation is significant for colorectal cancer (CRC) patients. The *KRAS* gene encodes a protein involved in the epidermal growth factor receptor (EGFR) signaling pathway, and mutations in this gene can negatively impact the use of monoclonal antibodies in anti-EGFR therapy and affect treatment decisions. Currently, commonly used methods like next-generation sequencing (NGS) identify *KRAS* mutations but are expensive, time-consuming, and may not be suitable for every cancer patient sample. To address these challenges, we have developed *KRASFormer*, a novel framework that predicts *KRAS* gene mutations from Haematoxylin and Eosin (H & E) stained WSIs that are widely available for most CRC patients. *KRASFormer* consists of two stages: the first stage filters out non-tumor regions and selects only tumour cells using a quality screening mechanism, and the second stage predicts the *KRAS* gene either wildtype' or mutant' using a Vision Transformer-based XCiT method. The XCiT employs cross-covariance attention to capture clinically meaningful long-range representations of textural patterns in tumour tissue and *KRAS* mutant cells. We evaluated the performance of the first stage using an independent CRC-5000 dataset, and the second stage included both The Cancer Genome Atlas colon and rectal cancer (TCGA-CRC-DX) and in-house cohorts. The results of our experiments showed that the XCiT outperformed existing state-of-the-art methods, achieving AUCs for ROC curves of 0.691 and 0.653 on TCGA-CRC-DX and in-house datasets, respectively. Our findings emphasize three key consequences: the potential of using H & E-stained tissue slide images for predicting *KRAS* gene mutations as a cost-effective and time-efficient means for guiding treatment choice with CRC patients; the increase in performance metrics of a Transformer-based model; and the value of the collaboration between pathologists and data scientists in deriving a morphologically meaningful model.

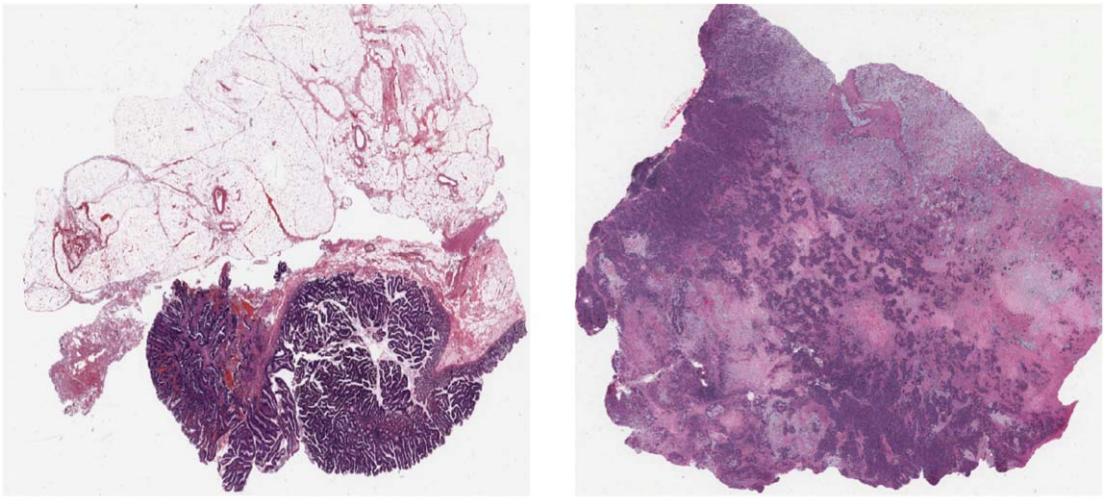


Figure 1. Visualization of a Haematoxylin and Eosin (H & E) stained WSI for *KRAS* gene in CRC. Here (a) and (b) show *KRAS* mutant and wildtype examples, respectively.

1. Introduction

Colorectal cancer (CRC) is the third most common type of cancer [1], projected to cause about 3.2 million new cases and around 1.6 million mortality by 2040 worldwide [2]. Specific mutations may make a tumour resistant to certain treatments, while others may make it more susceptible. By identifying mutations, clinicians can tailor treatment plans to the particular patient, potentially leading to better outcomes [3]. From tissue samples, molecular testing can identify specific mutations that may predict response to specific and targeted therapies [4]. The need for molecular testing of the *KRAS* gene in CRCs has long been established, where mutations are negative predictive markers for anti-EGFR pathway monoclonal antibodies [5]. The prediction of *KRAS* mutations aids in selecting an appropriate treatment for a cancer patient. Technologies for molecular testing such as Next Generation Sequencing (NGS) require minimal tissue samples, typically between 5% and 20% tumour cell content and minimum nucleic acid content for successful testing [6]. Nonetheless, extensive molecular testing is challenging on an enormous scale and is not done at all the labs with limited resources for each patient. In turn, Haematoxylin and Eosin (H & E) stain images are commonly available on CRC tissue biopsy samples. Using H & E stained images to predict cancer mutations is a cost-effective solution that broadens the scope of sample types that can be accurately tested.

In recent years, there has been exceptional success in the field of computer vision tasks, particularly in object classification, image segmentation, and object detection within natural image datasets [7]. Various deep-learning methods have been utilized, including Convolutional Neural Networks (CNNs) and Vision Transformer-based models. These methods have proven to be highly effective in various applications of computer vision tasks by extracting robust feature representation [8, 9]. In particular, CNN layers utilizing digital filters to perform convolution operations allow models to extract spatial features (such as shape, texture, intensity, and margins) and global features that describe the complete image. Additionally, Vision Transformers-based methods are employed in modeling long-range dependencies through its strong capability of a self-attention mechanism that focuses on essential, meaningful features in images while ignoring the irrelevant ones [10]. In the field of medical imaging, these methods have played a crucial role in the disease diagnosis, prognosis, and treatment of several types of cancer, including colon, breast, lung, rectal, prostate, and ovarian cancer [11, 12]. Vision Transformer-based approaches are currently being utilized as an alternative to CNNs for analyzing histopathology images to extract important morphological features that can aid in predicting clinically relevant information [13].

In the past few years, very limited studies have been aimed at the possibility of predicting *KRAS* gene mutations through images of CRC tumours stained with H & E. Figure 1 shows the examples of *KRAS* mutant and wildtype WSI acquired from the TCGA-CRC-DX cohort [14]. Tried to predict CRC mutations from the H & E-stained WSIs. The authors analyzed five genes: *APC*, *KRAS*, *PIK3CA*, *SMAD4*, and *TP53*. They used the 629 CRC patients from The Cancer Genome Atlas (TCGA-COAD and TCGA-READ) and 142 CRC samples from in-house datasets. The authors employed the CNN-based Inception-v3 architecture to predict the gene mutations and obtained mean area under the receiver operating characteristic (AUROC) scores ranging from 0.645 to 0.783 in formalin-fixed paraffin-embedded (FFPE) tissue slides. Another

study by [15] developed a weakly supervised deep learning-based framework to identify molecular pathways and specific gene mutations in CRC patients using H & E images. The framework employed the ResNet-18 model to distinguish between tumour and non-tumour patches. Subsequently, the ResNet-34 model was utilized to predict the probabilities of each patch corresponding to molecular labels such as high or low mutation density, microsatellite instability or stability, chromosomal instability or genomic stability, *CpG* island methylator phenotype (*CIMP*), *BRAF*, *TP53*, and *KRAS*. The authors used 502 WSIs of primary colorectal tumours from 499 patients extracted from the TCGA-CRC-DX dataset. The framework achieved an AUROC score of 0.6 for *KRAS* mutation prediction. While most studies have focused on predicting gene mutations through WSIs, their analysis was constrained to assessing the potential of Vision Transformer-based methods for classifying *KRAS* mutations using a limited H & E dataset.

A recent study [16] proposed an innovative Vision Transformer-based approach for predicting biomarkers from H & E-stained CRC tissue images. The authors combined a pre-trained Vision Transformer encoder and Transformer network to accomplish patch aggregation. The study's primary focus was on managing multi-institutional cohort data, with less attention paid to conducting an in-depth analysis of the clinical or pathological context that could limit the model's effectiveness for use in clinical settings. The authors utilized ten CRC datasets of over 9, 000 patients to predict genetic biomarkers related to microsatellite instability (MSI) and mutations in the *BRAF* and *KRAS* genes. The authors trained their model using four large cohorts to predict *KRAS* gene mutations and obtained an AUROC score of 0.75. However, due to the large cohorts utilized, the authors still need to address their proposed model optimization issues related to details and training hyperparameters of the proposed Vision Transformer-based architecture.

There has been some interest in categorizing gene mutations in CRC from H & E slides [17], but mutation detection from H & E slides are relatively uncommon due to limitations of H & E which is primarily used for morphology, and where tumour heterogeneity is seen. For instance, staining provides information about tissue architecture and cellular morphology but does not explicitly highlight genetic alterations or mutations. It may reveal specific morphological changes that can raise suspicion of an underlying genetic alteration. The tumours are often genetically heterogeneous, containing a mixture of genetic mutations and alterations. The H & E slides may not capture the full extent of this heterogeneity, making it challenging to identify specific mutations and mutational complexities accurately. The availability of H & E-stained slide images of CRC tumours may facilitate employing deep learning models to predict *KRAS* mutation status. Using such resources may be more cost-effective than traditional NGS molecular testing.

This paper introduces a novel framework called *KRASFormer* that predicts *KRAS* gene mutations in patients with CRC. The framework consists of two stages: a quality screening mechanism that selects colorectal tumour tissue by disregarding non-tumour patches and *KRAS* gene mutation prediction using H & E stained WSI, classified as either wildtype ($KRAS_{WT}$) or mutant ($KRAS_{MT}$). The *KRASFormer* framework relies on a Vision Transformer network inspired by the XCiT model [18]. It has three components: a cross-variance attention (XCA) block, a local patch interaction layer (LPI) layer, and a feed-forward network (FFN). The XCA block operates on the feature or channel dimensions of the input patches to capture long-range representation using spatial and global features of the input patches. The LPI and FFN components enhance the extracted features and provide per-patch knowledge through fully connected layers. An extensive ablation experiment was conducted, comparing the performance of CNNs and Vision Transformer-based methods, using three datasets to build both stages of the experiments. The results show that the XCiT model outperforms existing CNNs and other Vision Transformer-based methods in the *KRAS* mutation prediction task. Note that this article discusses the technical and clinical aspects of understanding the prediction of the *KRAS* gene mutation from WSIs. It does not propose a new deep-learning model but explores how the Vision Transformer-based XCiT approach can improve tumour tissue selection and *KRAS* mutation classification. This paper's main contributions are five folds:

- Propose a new framework called *KRASFormer* for classifying *KRAS* gene mutations from H & E stained slide images in CRC patients.
- We develop a two-stage system that selects tumour tissue by disregarding non-tumor tissues types such as stroma, complex stroma, lymph, debris, mucosa, adipose, and background in WSI and predicts the *KRAS* gene mutation.
- Our framework used the XCiT model, combining cross-variance attention as a backbone architecture that works on feature or channel dimensions of the WSI input patches and helps extract clinically meaningful local and global features from H & E stained slides.
- Providing extensive experiments with different sets of ablation analyses using three datasets to demonstrate the effectiveness of pre-trained CNNs and Vision Transformer-based methods pre-trained on ImageNet [19] to predict *KRAS* mutations.
- Building a relationship to bridge the gap between pathologists and data science researchers to

understand better the XCiT model for *KRAS* gene mutation prediction in WSI. Our experimental findings demonstrate that XCiT can efficiently learn the textural patterns of CRC tumour tissue.

This study is organized as follows. Section 2 describes the CRC datasets and the proposed *KRASFormer* framework with its architecture. Section 3 presents the experimental results with a comprehensive ablation analysis. Section 4 is dedicated to discussing both the strengths and the limitations of this study. Finally, section 5 concludes the study while also outlining potential avenues for future research.

2. Material and method

2.1. Datasets

In this study, we have utilized three different H & E datasets, namely CRC-5000 [20], The Cancer Genome Atlas colon and rectal cancer (TCGA-CRC-DX)[8] [21], and samples from the Northern Ireland Biobank (NIB). Out of these, we have employed 133 samples from TCGA-CRC-DX and NIB datasets to develop the *KRAS* mutation dataset in WSIs. Below is the summary of each dataset.

- CRC-5000: This dataset included 10 H & E-stained CRC WSIs that were acquired from the University Medical Center Mannheim, Germany [20]. This dataset comprised eight different tissue types, including tumour epithelium, simple stroma, complex stroma, immune cells, debris, normal mucosal glands, adipose tissue, and background with no tissue information. Each of these tissue types had distinct textural features, and the dataset contained both low-grade and high-grade tumours that were manually labelled by experts. The experts extracted 625 non-overlapping tissue patches with a size of 150×150 pixels. In total, the dataset contained 5000 patch images extracted from the 10 H & E-stained WSI, which were used to train and evaluate the classification model. Figure 2 displays six examples of patch images for each of the eight tissue types. As can be seen from the patches, there is texture variability and stain intensity diversity for each tissue type.
- TCGA-CRC-DX: In this dataset [21], 166 WSIs of CRC have *KRAS* gene mutations extracted from the right site. Due to the subpar quality (i.e., air bubbles, tissue folding, compression artefacts, out-of-focus blur, and pen markings) of the WSIs, only 106 samples with *KRAS* gene information were utilized. Out of these, there are 55 $KRAS_{WT}$ and 51 $KRAS_{MT}$ samples. These samples consist of formalin-fixed paraffin-embedded (FFPE) tissue slides categorized into four stages I, II, III, and IV, with the following distribution: 1 sample (0.094%) for stage I, 15 samples (14.15%) for stage II, 70 samples (66.03%) for stage III, and 20 samples (18.86%) for stage IV.
- NIB: This dataset comprises 27 CRC patients WSIs, 17 of which are wildtype and 10 are mutant, with a share of 12 right-sided, 14 left-sided, and 1 transversal, collected from the Northern Ireland Biobank at the Patrick G Johnston Center for Cancer Research, Queen's University Belfast in the United Kingdom. The digital images provided were WSI of formalin-fixed paraffin-embedded CRC.

[8] https://doi.org/10.7937/TCIA.YZWQ-ZZ63

2.2. WSIs annotation protocol

Figure 3 shows the general pipeline for *KRAS* WSI data preparation. This study had a team of data science researchers and two pathologists with over 20 years of experience in their respective fields. We developed an annotation protocol to identify tumour regions of interest (ROI) containing malignant cells to predict *KRAS* gene mutations in WSIs. In brief, the data science researchers collected the raw WSIs from the TCGA-CRC-DX and NIB cohorts and utilized QuPath v0.2.3 software [22] to create respective projects in '.svs' file format. Afterward, the pathologists accessed the created projects to outline the necessary regions (ROIs) in the raw WSIs to produce accurate annotations.

2.2.1. Annotation selection criteria

We established the standard criteria for selecting annotations on two H & E WSI datasets using QuPath software [22]. Multiple ROIs were identified if necessary; artefacts were avoided; ROIs were adjusted to avoid white pixels if needed; the final result could be filtered by the data scientists using image processing or deep learning techniques. We also defined our exclusion criteria. In this process, the full WSI was excluded if the stain was not a recognizable H & E; there were no tumour areas; artefacts where substantial (more than 30% of the image); image blurring was present, or most of the section was composed of white pixels or red blood cells.

2.3. KRASFormer framework

Figure 4 provides an overview of our proposed KRASFormer framework, which consists of two stages for tumour tissue selection and *KRAS* mutation prediction. Stage I was developed to establish a quality screening pipeline that enriches for malignant epithelial cells. A Vision Transformer-based XCiT model was utilized to classify eight types of tissue patches extracted from the WSI. Stage II involves the classification of WSIs for *KRAS* mutations status as either $KRAS_{WT}$ or $KRAS_{MT}$ using the TCGA-CRC-DX and NIB cohorts. QuPath software, v0.2.3, was used to extract non-overlapping patches from the $40 \times$ magnification WSI, where experienced pathologists

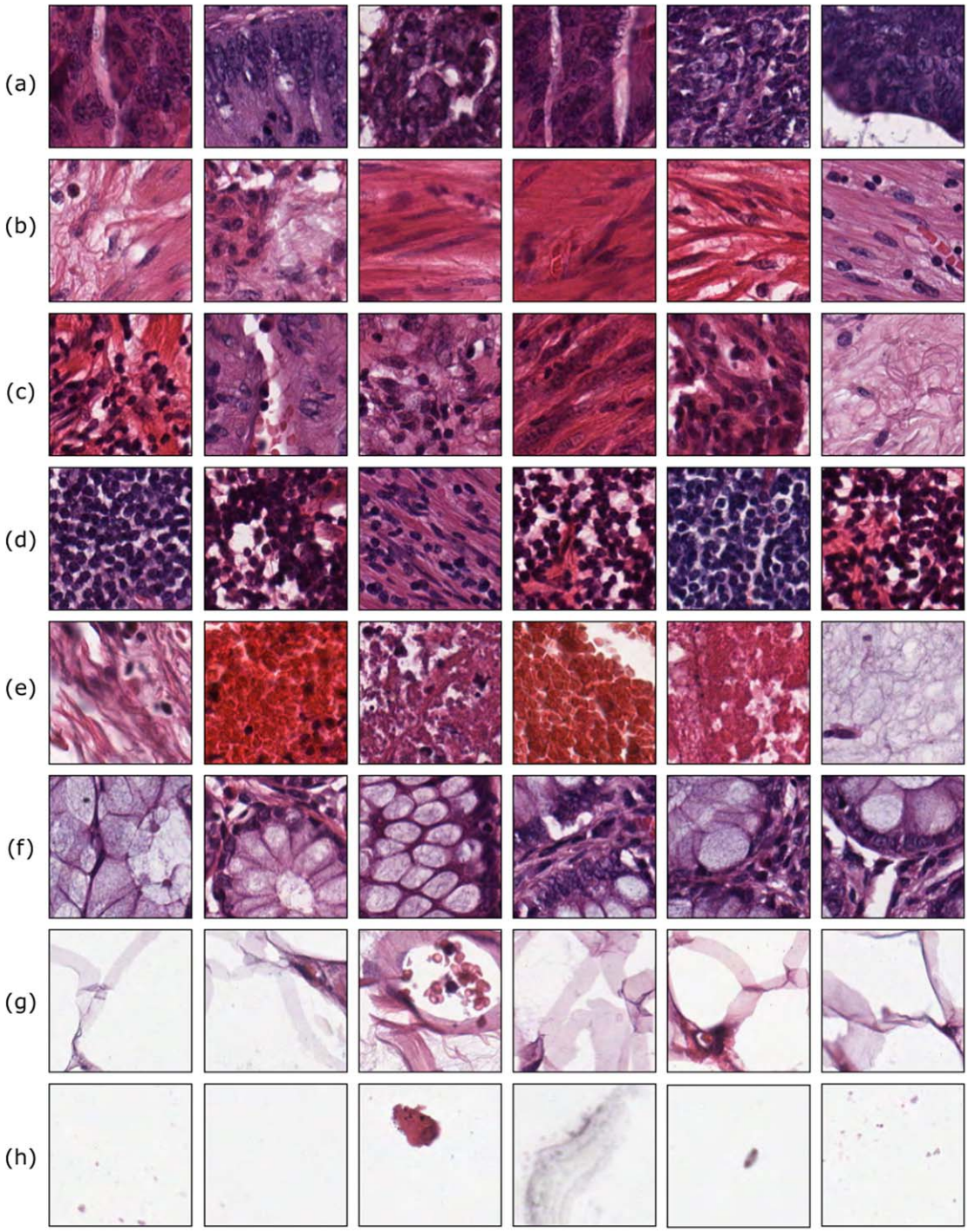


Figure 2. Overview of six patch images for each of eight tissue types in CRC. Note that (a) tumour epithelium, (b) stroma, (c) complex stroma, (d) immune cell, (e) debris and mucus, (f) normal mucosal glands, (g) adipose tissue, and (h) background with no tissue.

annotated the crucial ROI. The QuPath software generated these patches in .png' file format. The patch images were extracted with a fixed size of 256 × 256. Figure 5 illustrates the general description of the XCiT layer [18]. It incorporates three main components: cross-covariance attention layer (XCA), local patch interaction (LPI), and feed-forward network (FFN). During training, we resized the patches using bilinear interpolation to 224 × 224 pixels to serve as input for the model. Later, it used a patch size measuring 16 × 16, an embedded dimension of 768, 16 heads, a multilayer perceptron (MLP) ratio of 4, and two class attention layers. More details about each XCiT component are explained in the below subsections.

2.3.1. Cross-covariance attention

To attain a more profound understanding of cross-covariance attention [18], our initial focus was on the self-attention mechanism utilized in the Vision Transformer. This mechanism lets the model focus on crucial and relevant features within images while ignoring irrelevant information like non-tumour tissue in WSIs.

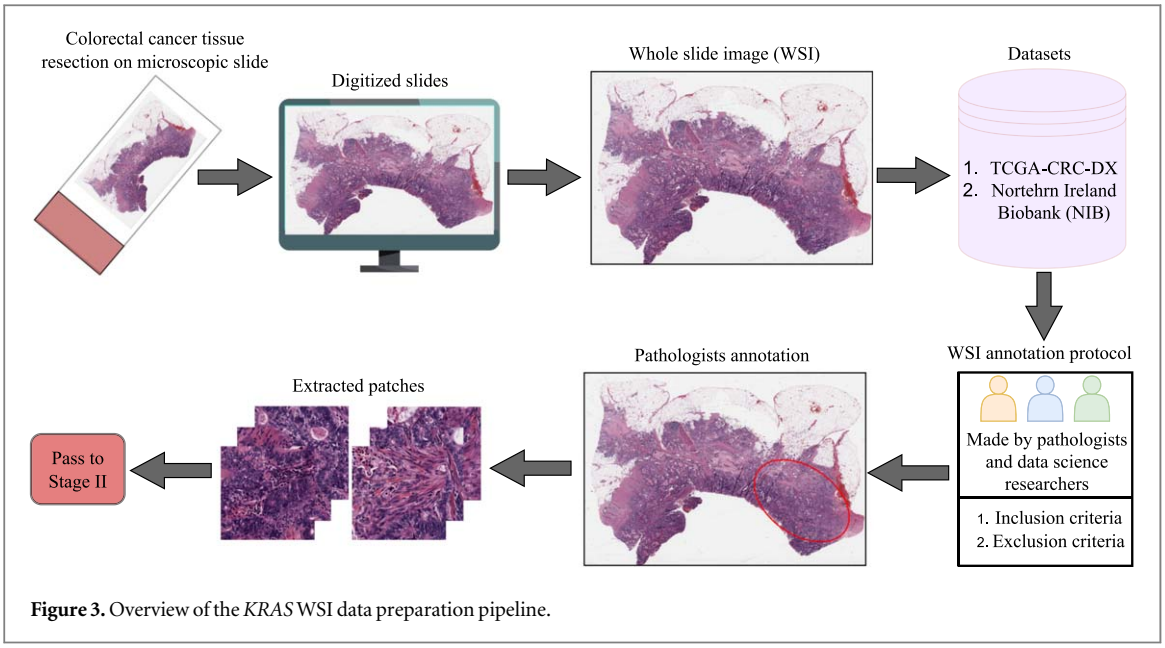


Figure 3. Overview of the *KRAS* WSI data preparation pipeline.

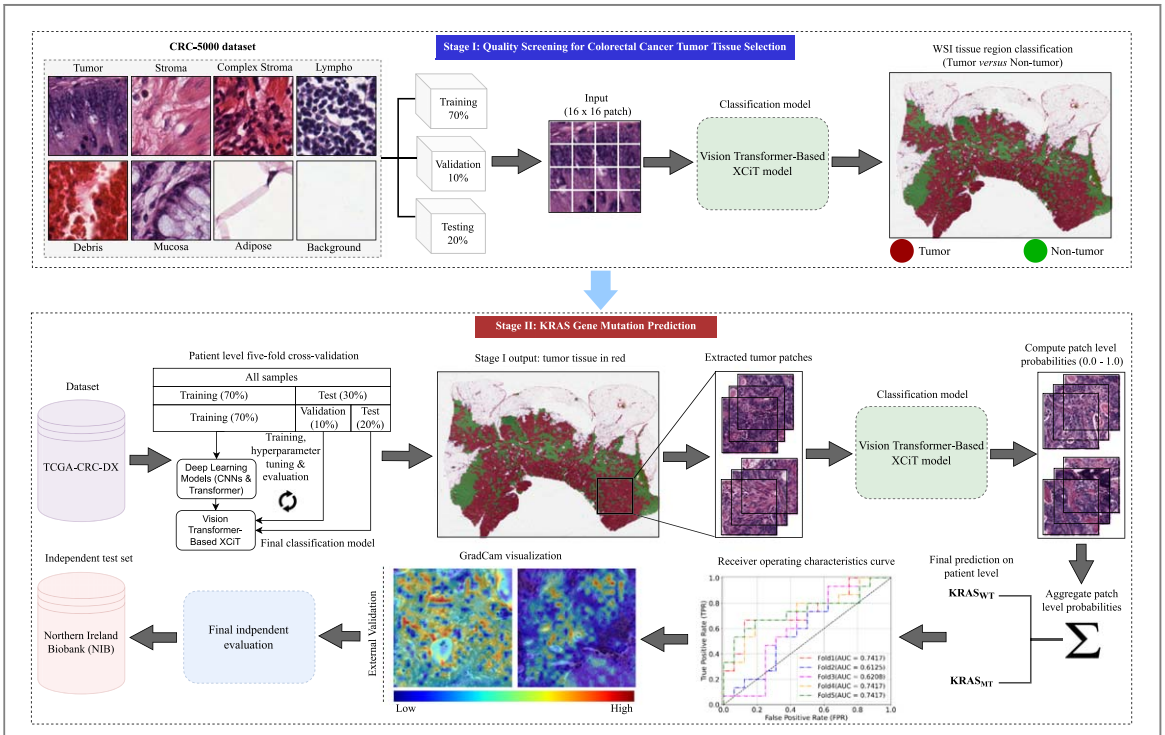


Figure 4. Illustration of the integrated *KRASFormer* framework, combining stages I and II. In stage I, the XCiT model is employed as a robust feature extractor, initially identifying tumour regions while disregarding non-tumour regions. In Stage II, the TCGA-CRC-DX dataset is first patient-wise split into five-fold cross-validation where deep learning-based CNN and Vision Transformer models are trained and evaluated separately. During this process, stage I help to select the tumour tissue region by ignoring the other tissue types. Then, the selected tumour patches undergo computation of patch-level probabilities for the presence of the *KRAS* gene. These probabilities are subsequently combined to assess the performance at the patient level by classifying them into *KRAS* wildtype and mutant categories. The current framework provides feature learning explainability using GradCam when the model pays more attention to the specific region highlighted in red and less represented by blue. Stage II *KRAS* gene mutation classification performance is independently externally validated on the Northern Ireland Biobank (NIB) dataset.

The attention mechanism calculates a weighted sum of all the features obtained from given patch images. The Transformer model is then trained to learn the weights allocated to each extracted feature, which are used to estimate the attention coefficient.

The XCA is generally described as a layer that uses cross-covariance attention to compute attention along the features or channels dimension instead of the token dimension. This can be represented as follows [18]:

$$XCA_{Attn}(Q, K, V) = V\mathcal{A}_{\mathrm{XCA}}(K, Q) \qquad (1)$$

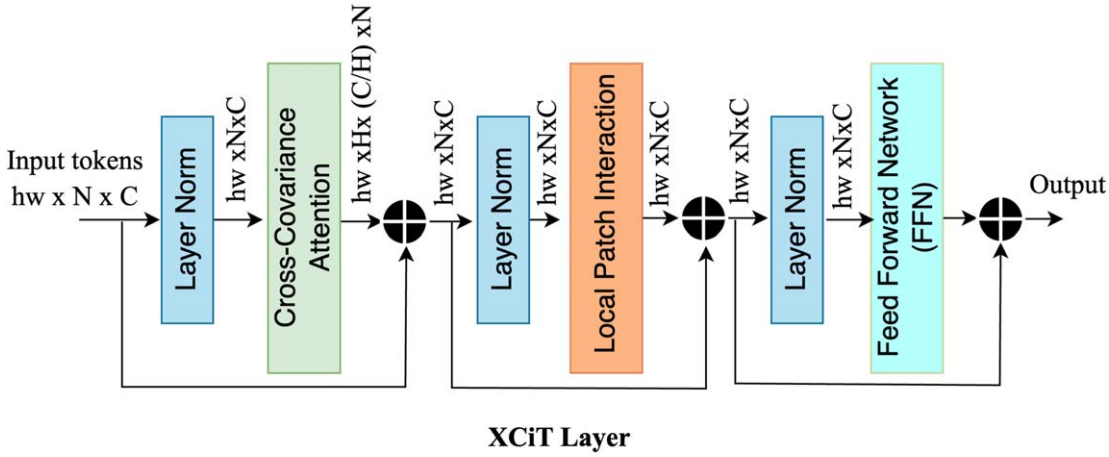


Figure 5. Overview of the XCiT layer that includes layer normalization layers, cross-covariance attention, local patch interaction, and a feed-forward network. Here, *hw*, *H* , *N* , and *C* refer to the height, width, head, token length, and channels respectively.

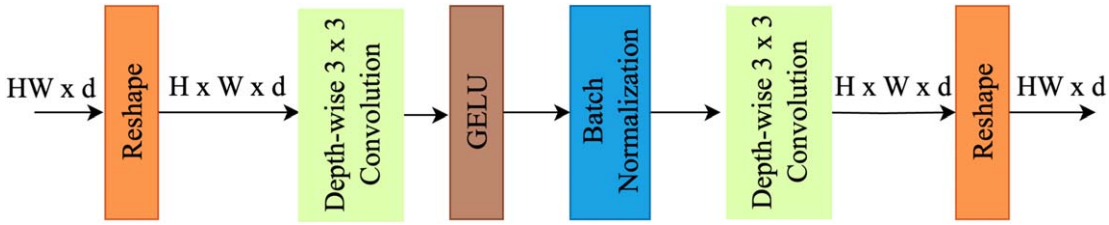


Figure 6. Overview of the local patch interaction layer that includes two depth-wise convolutional layers, a non-linear GELU activation function, and a batch normalization layer.

Where $\mathcal{A}_{\mathrm{XCA}}(K, Q)$ is the $Softmax(\hat{K}^{\top}\hat{Q}/\tau)$ that generate the attention scores, and τ correspond to a learnable temperature that provides better model training. For every patch embedding, three linear projections are used to acquire three vectors: Query (Q), Keys (K), and Values (V). It is essential to mention that the estimation of attention weights $\mathcal{A}$ relies on the cross-covariance matrix.

2.3.2. Local patch interaction

Figure 6 shows the general architecture of LPI layer. The local associations between pixels in input images that the model can capture may be limited by the need for a more direct connection between tokens in XCA. To address this, the LPI layer comes into play, combining information between tokens in the input sequence. The attention layers from the attention mechanism are typically utilized to merge this information. However, the XCA attention layer takes information integration between features or channels in the input sequence a step further, similar to CNNs. This results in better outcomes and allows for the capture of local spatial features. The convolutional layers within the LPI block have a kernel size of 3×3, and they are implemented using two depth-wise convolutions separated by batch normalization and a nonlinear Gaussian Error Linear Unit (GELU) activation function.

2.3.3. Feed forward network

By utilizing the point-wise FFN layer with a single hidden layer containing four-dimensional hidden units, the XCA block enables seamless interaction between all features. This feature proves to be particularly useful, especially in situations where there are no feature relations in the LPI block.

2.4. Cost function

In this study, we assigned higher weights to minority classes, which contained fewer patches in each cross-validation fold, the Weighted Cross Entropy (WCE) method prioritizes their impact on the loss function, resulting in a more balanced training of the model. This strategy efficiently tackles the difficulties presented by imbalanced class distributions, improving the model's ability to generalize across all classes. It can be expressed as follows:

$$\mathcal{L}^{WL}(x, y) = -w_{y_n} \log \frac{\exp(x_{n,y_n})}{\sum_{c=1}^{C} \exp(x_{n,c})} \qquad (2)$$

where *x*, *y*, *w*, and *C* are the input, target, weight, and number of classes, respectively.

Table 1. The hyperparameters used for stages I and II.

Parameter	Stage	
	I	II
input size	224 × 224	224 × 224
learning rate	0.0002	0.0002
optimizer	Adam	Adam
batch size	4	4
epoch	100	100
loss function	WCE	WCE
data augmentation	Yes	Yes
number of classes	8	2

Table 2. Comparison of the XCiT results against existing methods in classifying the tumour versus non-tumour tissues in H & E on the CRC-5000 dataset. Note that ↑ indicates higher is best highlighted in bold.

Method	Accuracy (↑)
Kather *et al* [20]	87.40
ARA-CNN [25]	92.44
Ohata *et al* [26]	92.08
Zeid *et al* [27]	94.75
Proposed	**96.75**

3. Experimental setup and results

3.1. Training and evaluation

A stratified five-fold cross-validation technique was used during the training and testing phases to produce robust classification results from the model. We split our dataset into the patient-wise setting with a ratio of 70%, 10%, and 20% for training validation and testing, respectively. The XCiT utilized a WCE loss function with a batch size of 4 patch images to minimize the error accurately. The model was trained with the Adam optimizer with a learning rate of 0.0002 and 100 epochs. To validate and evaluate stage II performance, patient-level predictions were calculated by averaging the patch-level probabilities generated by the model.

$$KRAS_{MT} = \frac{1}{P_{total}} \sum P_i \tag{3}$$

Where P_i is the mutation score of slide patch i and P_{total} is the total number of patches present in each patient.

During both stages of the experiment, we utilized data augmentation techniques to enhance the diversity and heterogeneity of the tissue images used in model training. This involved horizontal flipping with a probability of 0.5, scaling, rotation by 30 degrees, and illumination changes. The model parameters were saved based on the highest achieved classification accuracy on the validation set. We employed a similar experimental setup for all methods to ensure a fair comparison between our XCiT model and other methods. Table 1 summarizes the hyperparameters used in the experiment.

3.2. Computational setup

The experiments in this study were carried out on PyTorch, utilizing the National High-Performance Computing (HPC) Kelvin-2 administered by Queen's University Belfast (QUB). All models were trained on a 32 GB GPU memory equipped with CUDA version 11.2. The model evaluation, however, was executed on a local workstation with a Linux operating system and an RTX2080 Ti GPU (11GB).

3.3. Evaluation metrics

We used five evaluation metrics to determine the effectiveness of our proposed approach and its comparison with other methods. These metrics include accuracy, precision, recall, and F1-score [23].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{7}$$

During the evaluation process, we considered the four terms: TP, TN, FP, and FN. These terms refer to true positives, true negatives, false positives, and false negatives. Additionally, we utilized Area Under the AUROC as a metric to assess the model's ability to differentiate between classes [24]. A higher value of AUROC, closer to 1, indicates a better classification performance by the model.

3.4. Results

3.4.1. Stage I: Colorectal tumour tissue classification

Table 2 demonstrates the tissue classification results yielded via the XCiT compared with recently published works, including [20, 25, 26], and [27] using the CRC-5000 dataset.

The experimental results show that the XCiT achieved the most promising classification results with 2% improvement than second-best [27]. The cross-covariance attention facilitates linear computation against different tissue region structures aggregating local and global features in given patches. From the visual inspection and tissue morphology, the patch images of the eight classes were different in their textural patterns. This suggests that the XCiT learned to recognize meaningful textural and morphological features, encouraging multiclass classification accuracies. To classify the eight tissue types, [20] employed a set of classical machine learning-based techniques that contained the first benchmark accuracy of 87.4%. They assigned the 1-nearest neighbour, linear support vector

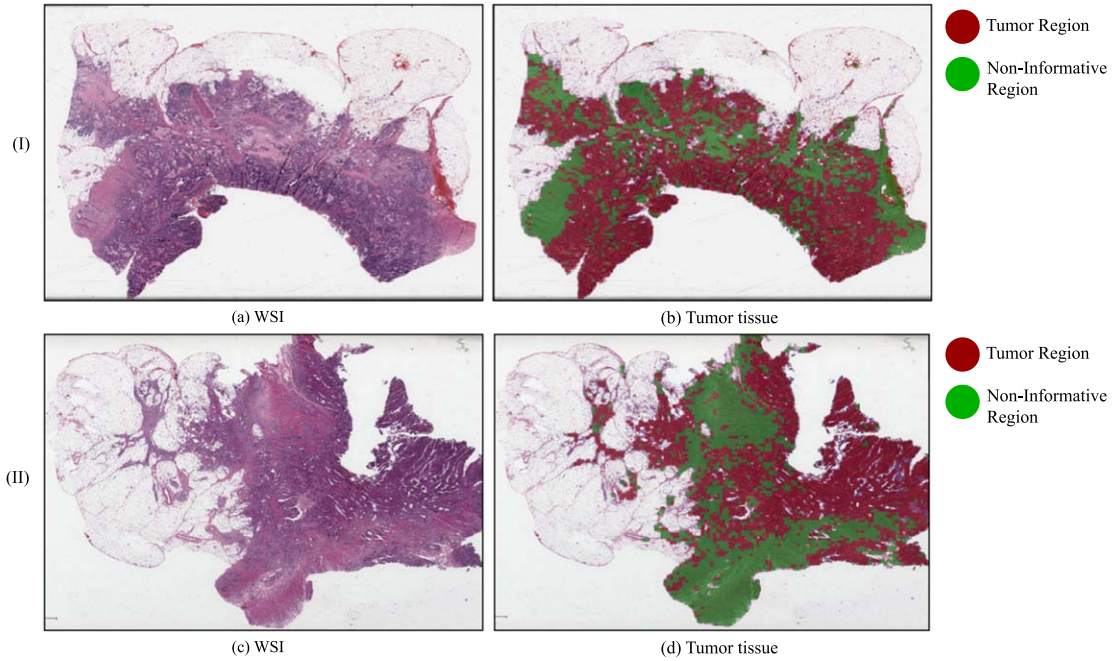


Figure 7. Classification of tumour tissues versus non-tumour regions using the XCiT in Stage I, validated on TCGA-CRC-DX cohort. Pathologists visually examined the patches and found that the majority were accurately classified as tumour tissues. Note that (a) and (c) show examples of whole slide images (WSIs), while (b) and (d) display patch-level predictions. Tumour patches are highlighted in red, while the remaining classes are depicted in green.

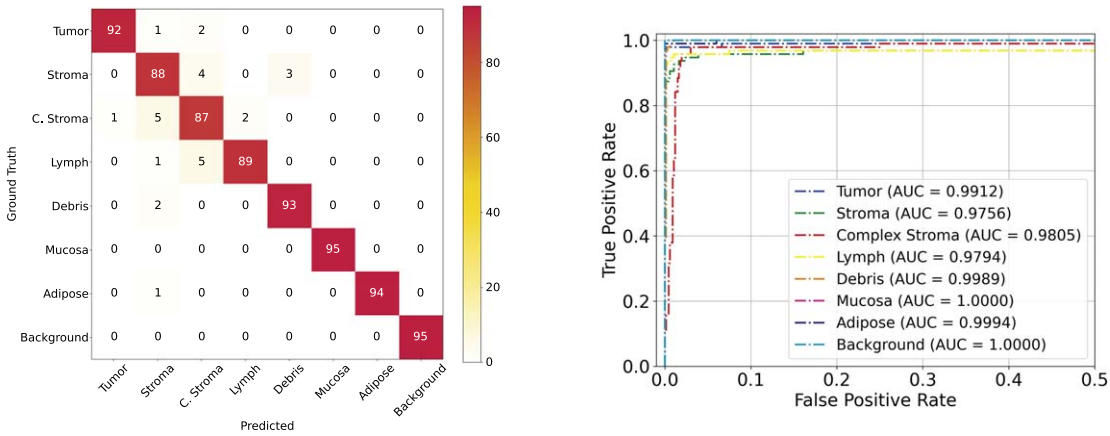
machine (SVM), radial basis functions, and decision tree classifiers to classify tumour and non-tumor patch images. Notably, this method included the hand-crafted feature extractor such as local binary pattern (LBP), grey-level co-occurrence matrix (GLCM), and Gabor filters to extract the features of patch images. The XCiT model exceeded the [25] with a substantial margin of 4% where they leveraged the bayesian convolutional neural network (ARA-CNN). In addition, [26] obtained 2% lower result against the XCiT based solely on the transfer learning approach. Conclusively, the XCiT extracted the small details through small patches with a cross-covariance attention mechanism. The image featured details were mixed between the channels, which provided an improvement in classification.

Figure 7 shows two examples of WSI and how the XCiT model selected the patch-level tumour tissue while disregarding the non-tumour regions, highlighted in red and green, respectively. Notably, the trained stage I model was applied to stage II for selecting tumour tissue in the TCGA-CRC-DX and NIB cohorts. Since patch-level multiple tissue regions ground truth is unavailable, experienced pathologists manually reviewed the WSIs of the TCGA-CRC-DX and NIB cohorts. Their visual inspection found that the model correctly identified more than 80% of the tumour tissue patches in each patient.

The results of the XCiT model's classification of CRC tissue into eight classes are presented in the confusion matrix shown in figure 8(a). The model successfully classified the majority of tissue patches, with only a few patches of misclassification. The distinct textural patterns of these eight tissue types helped the model learn more accurate feature representations. However, it can be challenging to determine the patch class in some cases due to the presence of various tissue types with different textural or morphological information. Despite this, the model classified most patches as tumours due to the distinct textural pattern. There were three patches of misclassification of the tumour as stroma and complex stroma, and stroma was misclassified as complex stroma or debris four and three times, respectively. The remaining classes (mucosa, adipose, and background) had a low error rate of only 2%, while the lymphocytic class had a slightly higher error rate of 6% with misclassifications to the stroma and complex stroma categories. The results of the AUROC score, shown in figure 8(b), confirm the high classification performance of the XCiT, with all classes having a score of more than 98%. Therefore, the more visually simple the tissue type, the more successful the model was in identifying the correct tissue. With complexity inevitably came errors in classification, with the Transformer model confusing complex stroma with malignancy. Figure 9 shows that the t-SNE visualization obtained by the XCiT utilizes an 8-dimensional class vector in its final layer embedding. This vector is then transformed into 2 dimensions. The resulting plot demonstrates that each patch sample class is easily distinguishable and forms its own distinct cluster.

3.4.2. Stage II: A KRAS *mutation prediction on TCGA-CRC-DX cohort*

The quantitative results of *KRAS* mutation prediction in H & E can be seen in table 3. It showcases the



(a) Confusion matrix (b) AUROC curve

Figure 8. The XCiT model classification results for the eight tissue regions are depicted in the confusion matrix and AUROC shown in (a) and (b), respectively. .

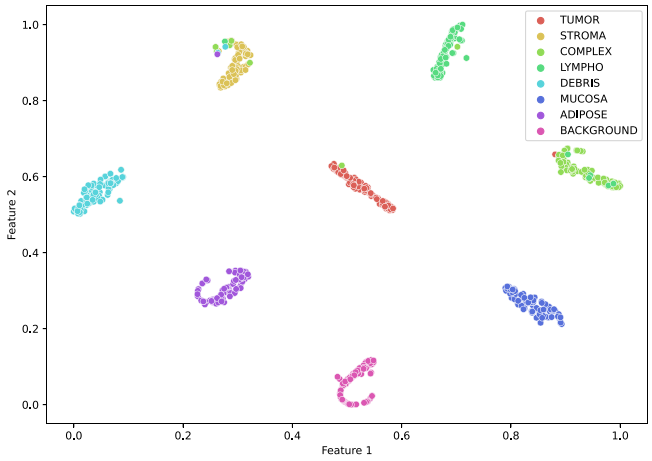


Figure 9. t-SNE visualization for eight-class colorectal tissue region classification from CRC-5000 dataset.

Table 3. A quantitative comparison of the XCiT model results for predicting *KRAS* mutations in the TCGA-CRC-DX cohort is provided with existing CNN and Vision Transformer-based methods. The results are reported at the patient level and obtained by averaging five-fold cross-validation scores along with the standard deviation. The best significant results are in bold.

CNN-Based	Models	Metrics				
		Accuracy (↑)	Precision (↑)	Recall (↑)	F1-Score (↑)	AUROC (↑)
	ResNet-50	61.25 ± 11.47	62.77 ± 13.62	60.04 ± 16.86	59.92 ± 12.58	64.27 ± 13.59
	ResNext	64.51 ± 8.90	66.14 ± 7.84	62.66 ± 19.13	61.54 ± 11.63	67.05 ± 10.80
	EfficientNet-B7	57.63 ± 16.95	58.24 ± 17.77	56.82 ± 20.35	57.68 ± 14.27	61.43 ± 14.11
	EfficientNetV2	59.91 ± 13.31	59.82 ± 14.38	57.33 ± 14.96	57.83 ± 13.06	63.01 ± 12.19
	MobileNetV3	56.66 ± 17.73	57.03 ± 19.60	55.14 ± 18.21	57.34 ± 15.89	59.17 ± 15.52
Vision Transformer-Based	ViT	60.45 ± 11.93	61.83 ± 10.20	59.61 ± 13.44	60.25 ± 12.50	61.94 ± 13.21
	Swim	63.63 ± 11.27	64.72 ± 9.93	62.14 ± 10.85	63.66 ± 8.89	65.48 ± 9.19
	BEiT	59.99 ± 9.91	60.65 ± 9.65	48.0 ± 19.50	51.90 ± 16.52	61.58 ± 14.30
	ResMLP	60.72 ± 12.36	62.08 ± 11.23	58.69 ± 14.55	60.07 ± 11.81	63.44 ± 11.73
	Proposed	**67.09 ± 8.06**	**68.80 ± 5.52**	**69.33 ± 6.79**	**67.35 ± 6.03**	**69.16 ± 6.12**

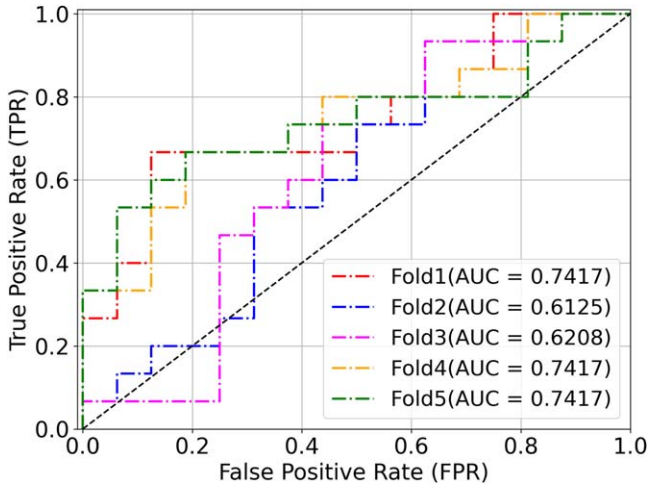


Figure 10. Illustration of AUROC using stratified five-fold cross-validation on the TCGA-CRC-DX cohort. Folds 1, 4, and 5 achieved similar scores. However, Fold 2 and 3 obtained the AUROC scores of 0.612 and 0.62%.

performance of state-of-the-art CNN-based and Transformer-based models that have been fine-tuned for WSI prediction tasks on the patient level. CNN-based models include ResNet-50 [28], ResNext-101 [29], EfficientNet-B7 [30], EfficientNet-V2 [31], and MobileNetV2 [32]. The Transformer-based models comprise of recently developed ViT [33], Swim [34], BEiT [35], ResMLP [36], and XCiT [18] networks, all pre-trained on ImageNet.

The ResNext-101 model demonstrated the highest performance among all CNN-based models, with a classification accuracy of 64.51% and an AUROC score of 67.05%. This model used the "split-transform-merge" technique, which enabled better discrimination between wildtype and mutant classes than other methods. ResNext-101 could extract rich features from the input due to its increased cardinality, and it had a lower standard deviation rate than other CNN-based methods. The ResNet-50 obtained the second-best results in terms of accuracy, precision, recall, F1-score, and AUROC scores with values of 61.25%, 62.77%, 60.04%, 59.92%, and 64.27%, respectively. In comparison, EfficientNet-B7, EfficientNet-V2, and MobileNetV3 performed similarly and did not significantly contribute to the *KRAS* mutation prediction task. The MobileNetV3, a lightweight deep CNN that uses depthwise separable convolutions, had the lowest classification scores among all methods due to its limited capacity to extract meaningful features.

Presently, Vision Transformer-based approaches have gained tremendous success in image recognition tasks. Table 3 demonstrates the results of *KRAS* mutation prediction by the latest Transformer-based methods. The XCiT model exceeded the performance of all other CNN and Transformer methods, achieving accuracy and AUROC scores of 67.09% and 69.12%, respectively. Figure 10 presents the AUROC for each cross-validated fold. The Swim Transformer was the second best performer, with an AUROC of 65.48%, thanks to its shifted windows method, which restricts self-attention computations to non-overlapping local tissue patches and enables cross-window connections. However, it scored 4%–6% lower than the XCiT on all metrics. Meanwhile, ViT, BEiT, and ResMLP produced comparable results and did not significantly contribute to solving this challenging *KRAS* prediction task. Figure 11 shows examples of the XCiT prediction overlaid on the raw TCGA slide images to examine the patch-wise prediction. We randomly chose four examples corresponding to the $KRAS_{MT}$ and $KRAS_{WT}$ categories. $KRAS_{MT}$ examples I and II show a very high confidence rate of more than 90% in correctly predicting the mutation in H & E. Nevertheless, $KRAS_{WT}$ patches also predicted well and obtained lower mutant scores.

3.4.3. Stage II: KRAS *mutation prediction on external independent NIB cohort*

We assessed the robustness of the XCiT using the NIB cohort as an independent test to classify the digital slide images into $KRAS_{MT}$ and $KRAS_{WT}$ classes. Figure 12 (a) presents the classification results using 27 H & E samples. We achieved an accuracy, precision, recall, and F1-score of 70.37%, 84.61%, 80.0%, 66.67%, and 65.29%, respectively. The model performed significantly better than the TCGA-CRC-DX cohort due to the good quality of slide images containing lower imaging artifacts. Figure 12 (b) provides the AUROC curve with a slightly lower outcome of 65.21% than the TCGA-CRC-DX cohort. Based on the experimental findings, we found that the model has the potential to predict the mutation in H & E but requires an additional data set of samples to verify the outcome for diagnostic purposes in the clinical setting.

4. Discussion and limitation

The study presents a promising approach to predict *KRAS* mutations in CRC tissue samples using H & E

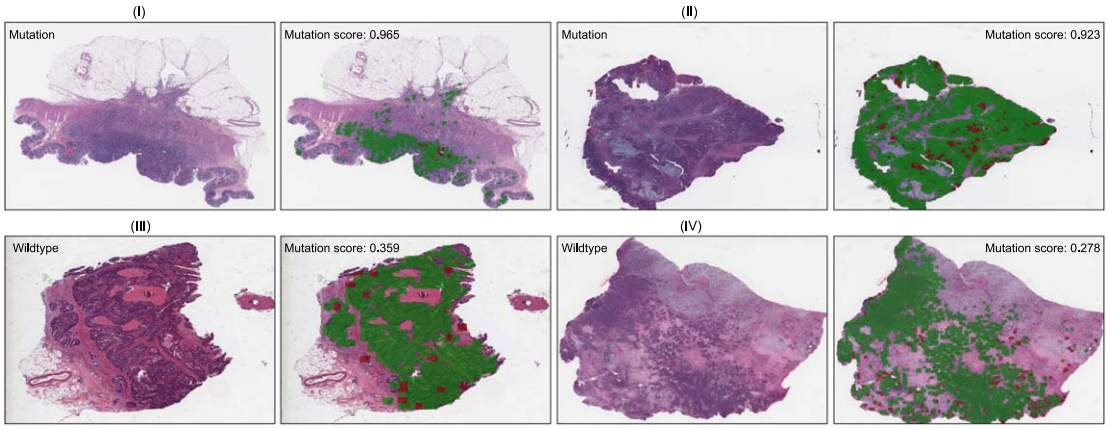


Figure 11. Examples of patch-wise prediction visualization using the XCiT overlaid on the raw TCGA slide images. The randomly selected two examples belong to the mutant and wild-type categories. Based on the classification, illustrate the mutation score calculated by averaging the probabilities of all the patches for each example. The correctly predicted patches are shown in green, and the incorrectly predicted ones are shown in red.

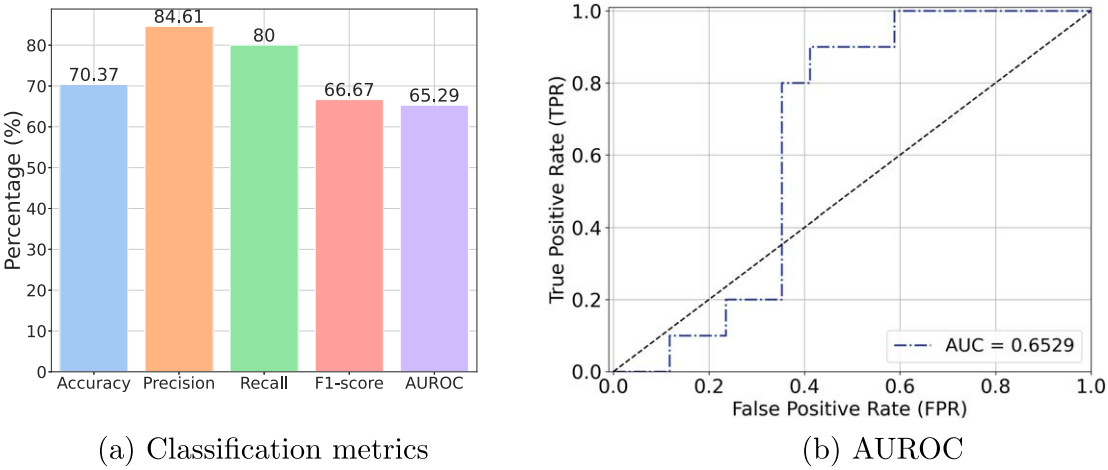


Figure 12. Illustration of the classification results with five metrics and area under the curve (AUROC) shown in (**a**) and (**b**), respectively, to predict the *KRAS* mutation achieved using the XCiT.

stained images. The study utilized a collaborative approach between pathologists and data scientists to develop an automated solution to predict *KRAS* mutations. We conducted several ablation experiments to gain insight into the clinical context of the problem. We found that eliminating unwanted tissue types, such as stroma, complex stroma, adipose, debris, and background information, was crucial for the model's feature learning capabilities and final classification performance. Further, the textural pattern identification between $KRAS_{WT}$ and $KRAS_{MT}$ was complex in H & E images because some cell features shared similar morphological properties, leading to incorrect predictions. The study demonstrated that the Vision Transformer-based XCiT method could extract better feature representation than traditional convolutional neural networks (CNNs).

Furthermore, the study found that selecting the standard patch size of 256×256 was crucial in capturing the relationship of spatial features through cross-covariance attention mechanisms. Although the study aimed to demonstrate the potential for the model as a cost-effective alternative to expensive molecular testing, further training using the principles outlined in the study is necessary for potential clinical application. The model's clinical utility will depend on its ability to accurately predict *KRAS* mutations in a wide range of sample types as well as the cost-effectiveness of implementing the model in clinical practice.

Earlier, Jang *et al* [14] attempted to predict CRC mutations from H & E-stained WSIs. They utilized the CNN-based Inception-v3 architecture to predict *KRAS* gene mutations in H & E and achieved AUROC scores ranging from 0.645 to 0.783 on the TCGA-CRC-DX and in-house datasets, respectively. Bilal *et al* [15] designed a weakly supervised framework to identify molecular pathways and specific gene mutations in CRC patients using H & E images. They employed the

ResNet-34 model, which achieved an AUROC score of 0.60 on the TCGA-CRC-DX cohorts. While most studies have focused on predicting gene mutations through WSIs, their analysis was limited to assessing the potential of Vision Transformer-based methods for classifying KRAS mutations. Our proposed method achieved improved mutation prediction using the Vision Transformer-based XCiT model, with a 0.69 AUROC score on the TCGA-CRC-DX cohort.

There are some limitations associated with this study. Specifically, the TCGA-CRC-DX cohort does not accurately represent real-world scenarios and fails to encompass the diverse range of morphologies that pathologists commonly encounter daily. With the limited number of samples, it is challenging to ensure a full variation of histopathological morphology. These limitations suggest that improving the model's performance may necessitate the inclusion of supplementary data sources from diverse groups, which could prove beneficial when employing advanced deep learning techniques.

5. Conclusion

We developed a cost-effective and time-efficient *KRASFormer* framework, where the use of a Vision Transformer-based XCiT model provides a viable alternative to CNNs for identifying *KRAS* gene mutation from H & E stained WSIs. This complete framework was split into two separate stages. The first stage only selected the tumour region, and the second stage was crucial in precisely identifying the *KRAS* gene mutation. We employed the potential of the XCiT model to extract the most relevant clinically meaningful features from the WSIs. Our experimental findings indicated promising results that the model could identify *KRAS* mutant patterns in two cohorts. The XCiT models possess the potential to function as screening tools for the prediction of *KRAS* gene mutations after further clinical validation. It is premature to conclude that the XCiT model can supplant conventional techniques like NGS. As the model's evaluation is restricted to a limited dataset, further confirmation in clinical validation studies is required to develop a clinical application that allows cancer patient treatment stratification.

Acknowledgments

The Northern Ireland Biobank has received funds from the HSC Research and Development Division of the Public Health Agency in Northern Ireland, Cancer Research UK, and the Friends of the Cancer Centre. The Precision Medicine Centre of Excellence has received funding from Invest Northern Ireland, Cancer Research UK, the Health and Social Care Research and Development Division of the Public Health Agency in Northern Ireland, the Sean Crummey Memorial Fund, and the Tom Simms Memorial Fund.

TCGA-CRC-DX data used in this publication were generated by the National Cancer Institute Clinical Proteomic tumour Analysis Consortium (CPTAC).

We are grateful for using the computing resources from the Northern Ireland High-Performance Computing (NI-HPC) service funded by EPSRC (EP/T022175).

Data availability statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

Funding

This work is supported by the PathLAKE consortium (www.pathlake.org accessed on 20 October 2023) Ref no.: 104 689. PathLAKE is one of a network of five Centres of Excellence in digital pathology and medical imaging supported by a 50 m investment from the Data to Early Diagnosis and Precision Medicine strand of the Industrial Strategy Challenge Fund, managed and delivered by UK Research and Innovation (UKRI).

Conflict of interest

Manuel Salto-Tellez is a scientific advisor to Mindpeak and Sonrai Analytics and has received honoraria recently from BMS, MSD, Roche, Sanofi, and Incyte. He has received grant support from Phillips, Roche, MSD, and Akoya. None of these disclosures are related to this work. All other authors declare no conflict of interest.

Institutional review board statement

This study made secondary use of image data generated under previously approved Northern Ireland Biobank projects (NIB11-009 and NIB15-0168). The Northern Ireland Biobank is a HTA Licenced Research Tissue Bank with generic ethical approval from The Office for Research Ethics Committees Northern Ireland (ORECNI REF 21/NI/0019).

Ethics approval

PathLAKE 19/SC/0363.

Epi700 CRC cohort: NIB19-310. CD3: (REC: 10/NIR02/53) and (REC:11/NI/0013).

Consent to participate

Samples and images used in this study were provided by the Northern Ireland Biobank [29] under NIB19/

310. The Northern Ireland Biobank is a HTA Licenced Research Tissue Bank with generic ethical approval from The Office of Research Ethics Committees Northern Ireland (ORECNIREF 21/NI/0019) and can confer ethical approval for projects and in particular, for the current study.

Consent for publication

All authors have read and approved the manuscript submitted.

ORCID iDs

Vivek Kumar Singh https://orcid.org/0000-0002-8259-7087
Perry Maxwell https://orcid.org/0000-0002-4117-1805

References

[1] Siegel R L, Miller K D, Goding Sauer A, Fedewa S A, Butterly L F, Anderson J C, Cercek A, Smith R A and Jemal A 2020 *CA: A Cancer Journal for Clinicians* **70** 145–64
[2] Morgan E, Arnold M, Gini A, Lorenzoni V, Cabasag C, Laversanne M, Vignat J, Ferlay J, Murphy N and Bray F 2023 *Gut* **72** 338–44
[3] Salto-Tellez M 2012 Overview of molecular tests and personalized cancer medicine *Principles of Molecular Diagnostics and Personalized Cancer Medicine* (Wolters Kluwer Health Adis (ESP)) pp 196–205
[4] Alam M R, Seo K J, Abdul-Ghafar J, Yim K, Lee S H, Jang H J, Jung C K and Chong Y 2023 Recent application of artificial intelligence on histopathologic image-based prediction of gene mutation in solid cancers *Brief. Bioinform.* **24** bbad151
[5] Lievre A *et al* 2006 *Cancer Research* **66** 3992–5
[6] Southwood M *et al* 2020 *J. Pathol.: Clinical Research* **6** 40–54
[7] Jiao L and Zhao J 2019 *IEEE Access* **7** 172231172231–63
[8] Moutik O, Sekkat H, Tigani S, Chehri A, Saadane R, Tchakoucht T A and Paul A 2023 *Sensors* **23** 734
[9] Khan S, Naseer M, Hayat M, Zamir S W, Khan F S and Shah M 2022 *ACM Computing Surveys (CSUR)* **54** 1–41
[10] Wu L, Guo S, Ding Y, Wang J, Xu W, Xu R Y and Zhang J 2022 Demystify self-attention in vision transformers from a semantic perspective: analysis and application arXiv preprint arXiv:2211.08543
[11] Yao X, Wang X, Wang S H and Zhang Y D 2020 A comprehensive survey on convolutional neural network in medical image analysis *Multimedia Tools Appl.* **81** 41361–405
[12] Shamshad F, Khan S, Zamir S W, Khan M H, Hayat M, Khan F S and Fu H 2022 Transformers in medical imaging: a survey *Medical Image Analysis* **88** 102802
[13] Deininger L, Stimpel B, Yuce A, Abbasi-Sureshjani S, Schönenberger S, Ocampo P, Korski K and Gaire F 2022 arXiv preprint arXiv:2206.00389
[14] Jang H J, Lee A, Kang J, Song I H and Lee S H 2020 *World Journal of Gastroenterology* **26** 6207
[15] Bilal M, Raza S E A, Azam A, Graham S, Ilyas M, Cree I A, Snead D, Minhas F and Rajpoot N M 2021 *The Lancet Digit. Health.* **3** e763–72
[16] Wagner S J *et al* 2023 Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study *Cancer Cell* **41** 1650–61
[17] Jiang Y, Chan C K, Chan R C, Wang X, Wong N, To K F, Ng S S, Lau J Y and Poon C C 2022 *IEEE Open J. Eng. in Med. Biol.* **3** 115–23
[18] Ali A *et al* 2021 *Advances in Neural Information Processing Systems* **34** 20014–27
[19] Deng J, Dong W, Socher R, Li L J, Li K and Fei-Fei L 2009 Imagenet: a large-scale hierarchical image database *2009 IEEE Conference on Computer Vision and Pattern Recognition (Ieee)* pp 248–55
[20] Kather J N, Weis C A, Bianconi F, Melchers S M, Schad L R, Gaiser T, Marx A and Zöllner F G 2016 *Sci. Rep.* **6** 1–11
[21] Clark K *et al* 2013 *Journal of Digital Imaging* **26** 1045–57
[22] Bankhead P *et al* 2017 *Sci. Rep.* **7** 1–7
[23] Hossin M and Sulaiman M N 2015 *International Journal of Data Mining & Knowledge Management Process* **5** 1
[24] Fawcett T 2006 *Pattern Recognit. Lett.* **27** 861–74
[25] Raczkowski L, Mozejko M, Zambonelli J and Szczurek E 2019 *Sci. Rep.* **9** 14347
[26] Ohata E F *et al* 2021 *The Journal of Supercomputing* **77** 9494–519
[27] Zeid M A E, El-Bahnasy K and Abo-Youssef S 2021 Multiclass colorectal cancer histology images classification using vision transformers *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS) (IEEE)* pp 224–30
[28] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 770–8
[29] Xie S, Girshick R, Dollár P, Tu Z and He K 2017 Aggregated residual transformations for deep neural networks *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 1492–500
[30] Tan M and Le Q 2019 Efficientnet: Rethinking model scaling for convolutional neural networks *International Conference on Machine Learning (PMLR)* pp 6105–14
[31] Tan M and Le Q 2021 Efficientnetv2: Smaller models and faster training *International Conference on Machine Learning (PMLR)* pp 10096–106
[32] Sandler M, Howard A, Zhu M, Zhmoginov A and Chen L C 2018 Mobilenetv2: inverted residuals and linear bottlenecks *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 4510–20
[33] Dosovitskiy A *et al* 2020 *arXiv preprint* arXiv:2010.11929
[34] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S and Guo B 2021 Swin transformer: hierarchical vision transformer using shifted windows *Proceedings of the IEEE/CVF International Conference on Computer Vision* pp 10012–22
[35] Bao H, Dong L, Piao S and Wei F 2021 *arXiv preprint* arXiv:2106.08254
[36] Touvron H *et al* 2022 ResMLP: feedforward networks for image classification with data-efficient training *IEEE Trans. Pattern Anal. Mach. Intell.* **45** 5314–21