Leveraging Foundation Models for Enhanced Detection of Colorectal Cancer Biomarkers in Small Datasets

Craig Myles[1], In Hwa Um[2], David J Harrison[2,3], and David Harris-Birtill[1]

[1] School of Computer Science, University of St Andrews, St Andrews KY16 9SX, UK
cggm1@st-andrews.ac.uk
[2] School of Medicine, University of St Andrews, St Andrews KY16 9TF, UK
[3] NHS Lothian Pathology, Division of Laboratory Medicine, Royal Infirmary of Edinburgh, Edinburgh EH16 4SA, UK

Abstract. Colorectal cancer is the second leading cause of cancer death worldwide. Its high incidence and mortality rate highlight the critical role of advanced diagnostics and early detection methods. Advancements in computational pathology can significantly enhance diagnostic precision and treatment personalisation, ultimately improving patient outcomes. Hospitals and labs globally are transitioning toward routine whole slide image (WSI) digitisation. This digitisation process generates large volumes of data, offering an opportunity to enhance diagnostic capabilities through the use of machine learning techniques such as weakly supervised learning and self supervised learning (SSL). This study evaluates the performance of state-of-the-art self-supervised learning (SSL) feature extractor foundation models—CTransPath, Phikon, and UNI—against a pretrained ResNet-50, which serves as a benchmark. Our Transformer network analyses these feature vectors, focusing on their efficacy in predicting key colorectal cancer biomarkers within a small dataset containing 423 WSIs with only 8% of cases exhibiting mismatch repair (MMR) deficiency. The CTransPath model achieved the highest validation AUROC of 0.9466 for MMR classification but exhibited a test AUROC of 0.6880, demonstrating significant variability. In contrast, the UNI model demonstrated greater consistency and robustness, achieving a test AUROC of 0.7136, which additionally represents a 6.3% improvement over ResNet-50's test AUROC of 0.6709. The results highlight the feasibility of using advanced machine learning models with smaller, sparsely annotated datasets, though the variability noted in some models underscores the challenges at the edge of data scarcity. Code and experimental framework available at https://github.com/CraigMyles/SurGen-CRC-Arena.

Keywords: digital pathology · machine learning · transformer · deep learning · slide-level classification · mismatch repair (MMR) · BRAF mutation · RAS mutation · survival prediction

1 Introduction

Colorectal cancer is the third most common cancer globally, and is the second leading cause of cancer related death worldwide[2]. The clinical management of colorectal cancer heavily relies on accurately identifying diagnostic biomarkers such as mismatch repair (MMR) protein status and microsatellite instability (MSI), pivotal in predicting immunotherapy efficacy and diagnosing Lynch syndrome[22]. However, traditional methods of assessing these biomarkers involve time-consuming and expensive immunohistochemistry or molecular tests. The advancements in digital pathology and whole slide imaging (WSI) position biomarkers as prime targets for machine learning, which holds the potential to dramatically enhance their detection and interpretation.

The shift toward routine digitisation of pathology slides has led to an unprecedented exponential increase in the quantity and quality of whole slide image (WSI) data, with hospitals and healthcare providers generating substantial quantities of data daily [7,24,26]. These images, rich in histological detail, are invaluable for diagnostic purposes. However, their utility for automated analysis is often limited by the absence of detailed annotations. Pathologists can generate these annotations, which may delineate regions of interest or classify histopathological features, essential for training and evaluating automated diagnostic systems.

Self-supervised learning (SSL), a variant of unsupervised learning, offers a solution to this limitation by training models to recognise patterns and features in data without the need for manually annotated slide, patch, or patient labels. In digital pathology, generating such labels typically requires significant expertise from pathologists or the execution of time-consuming and costly immunohistochemistry (IHC) assays to identify specific biomarkers. Given the rapid production of WSI data, such labelling is increasingly impractical for processing large cohorts efficiently.

SSL techniques can be utilised to develop foundational models in digital pathology that generate rich, self-learned feature vector embeddings from whole slide images (WSIs). These embeddings capture intricate patterns and details across various tissue types, enhancing the accuracy and efficiency of cancer diagnostics and prognostics without requiring expert labels. For example, UNI[4] is trained on over one hundred million WSI patches and produces feature vectors of 1024 dimensions. These foundational models serve as robust bases for downstream tasks such as cancer detection or biomarker prediction by leveraging their deep understanding of histological patterns. This approach not only reduces reliance on extensive manual annotations and costly lab tests but also could streamline the diagnostic process, facilitating more personalised and timely treatment strategies that ultimately improve patient outcomes.

Furthermore, these foundational models present the opportunity to utilise smaller, specialised datasets effectively, significantly reducing the traditional barriers associated with extensive data requirements. By leveraging the sophisticated feature vectors derived from foundational models, it may be possible to lessen the need for vast annotated datasets, making machine learning applications more feasible for smaller labs or studies focusing on less common conditions. Our research specifically explores whether these models can be used in conjunction with Transformer networks[28] to accurately classify MMR status within a dataset consisting of 423 hematoxylin and eosin (H&E) stained WSIs, within which only 8% of cases (n=32) are positive, challenging the convention that extensive data is a prerequisite for effective machine learning applications. The effectiveness of these models is evaluated through rigorous testing of their ability, in conjunction with Transformers, to handle sparse labels in our small dataset. While the definitive capability of these models to handle sparse labels in small datasets is still being explored, this approach could significantly democratise the use of advanced machine learning techniques, making them accessible even to studies with limited resources or specialised needs.

2 Literature Review

Feature Extractors in Digital Pathology. The transition to digital and computational pathology has significantly increased the need for machine learning (ML) tools capable of processing large datasets from WSIs[7]. Feature extractors can be leveraged in this context, as they convert raw histological data into structured feature vectors, for efficient ML training. This section evaluates prominent feature extractors such as CTransPath[31], Phikon[11], UNI[4], and ResNet-50[14]. Due to the nature of whole slide images being very large, and in some cases in excess of 100,000×100,000 pixels, feature extractors play an important role in in the computational pathology pipeline.

The Constrained Attention Multiple Instance Learning (CLAM) model, which incorporates a modified ResNet-50 for feature extraction, demonstrates the adaptability of these tools to digital pathology[21]. By transforming patches sized 256×256 into a 1024-dimensional feature embedding, CLAM effectively utilises clustering within its multiple instance learning framework for weakly supervised slide-level classification. This methodology shows significant promise in enhancing diagnostic accuracy without extensive manual annotation.

Similarly, adaptations of previously trained models such as the Inception-v3 network have been reconfigured for use in histology. Removing the final classification layer of the Inception-v3 model allows it to serve as a powerful feature extractor, configured to produce a 715-dimensional feature vector from each image tile extracted from WSIs[16].

While these ImageNet[27] trained models were not originally designed to process histopathological data, they demonstrate a remarkable ability to adapt to this domain[21]. Nonetheless, significant differences exist between natural images and histopathological images. Objects and structures typical in ImageNet datasets differ vastly in texture and structure from the cellular formations found in pathology slides. These discrepancies pose challenges for direct application, as models pretrained on natural images may be limited in their ability to accurately interpret the unique textural and semantic features inherent in histopathological data.

To mitigate these challenges, transfer learning from large-scale labelled datasets such as ImageNet is often proposed. Though effective in some cases, this strategy is inherently limited due to the substantial domain shift[31]. A more effective approach involves either training from scratch on pathology-specific datasets or employing self-supervised learning methods that do not require manual labels. Such strategies allow for the development of visual representations more aligned with the requirements of histopathological analysis.

Slide-Level Classification and SSL. Slide-level classification in pathology leverages multiple instance learning (MIL), where diagnoses or other predictions are derived from collections of image patches. Self-supervised learning (SSL) enhances model ability by enabling learning from unlabelled data, which is vital given the scarcity of detailed annotations in digital pathology. SSL frameworks like Momentum Contrast (MoCo)[13] and Simple Framework for Contrastive Learning of Visual Representations (SimCLR)[5] have revolutionised this field by enabling robust feature extraction without extensive labelled datasets, demonstrating their utility across various diagnostic tasks[17].

Breakthroughs in SSL and Their Impact. Whilst self supervised visual representation learning is not a new area of research[12], recent breakthroughs in SSL, including the introduction of Momentum Contrast (MoCo)[13] and Simple Contrastive Learning (SimCLR)[5], have considerably advanced various fields by enabling robust feature extraction without the need for extensive labelled datasets. Although originally benchmarked on natural images using datasets such as ImageNet, these techniques have demonstrated exceptional utility across a variety of domains, particularly in computational pathology[31,11,29,9,4,17,20].

Contrastive learning most commonly operates by training models to distinguish between pairs of similar and dissimilar data points, generated through various augmentations, using a loss function that minimises the distance between similar pairs while maximising the distance between dissimilar ones in the feature vector space[13,5]. These techniques leverage data augmentations to generate multiple views of the same image, effectively enhancing the model's ability to learn discriminative features from unlabelled data. This mechanism, combined with the use of augmentations, facilitates the extraction of more generalised and robust features, suitable for application in complex diagnostic tasks in the healthcare domain. Such advancements in SSL have led to a paradigm shift in computational pathology pipelines, enhancing diagnostic accuracy and streamlining processing, thereby marking a significant milestone in the application of artificial intelligence in healthcare.

A comparative analysis of various feature extractors reveals their distinct capabilities in handling specific diagnostic tasks, from MMR to RAS mutation classification. While ResNet-50 serves as a valuable benchmark due to its extensive use in image recognition, SSL-based models like UNI and CTransPath offer promising alternatives that might better capture the features within medical imaging due to their training on pathology-specific datasets.

Building a computational pathology pipeline to classify MMR status using CTransPath, Wagner et al. [30] experiment on a cohort of size 13,689, showing that this problem can be tackled with scale. Additionally, they experiment with smaller subsets of data, notably training on 250 cases and achieve an AUROC score of 0.923. However, it is unclear if the label distribution in these smaller datasets is consistent with that of the larger cohort, which could influence the results.

Analysis of Feature Extractors. ResNet-50[14] stands out among these other models as it is the only model among the previously mentioned which is pre-trained on natural images. ResNet, short for Residual Network, is a type of convolutional neural network (CNN) that was introduced to address the vanishing gradient problem in very deep networks. It utilises skip-connections which can bypass layers enabling the direct flow of gradients. This architecture has proven highly effective at image classification and recognition and is regularly used as a benchmark across various domains.

Model and Data Availability. A growing number of SSL foundation models are being developed, though many remain proprietary either with respect to the data used to train and/or the model weights themselves. Open-source models are generally trained on publicly available datasets, which promotes reproducibility and transparency in research. However, models including UNI[4], RudolfV[9], and Virchow[29] are trained on proprietary datasets, posing challenges in validating their effectiveness on unrelated test sets. This review includes an assessment on an independent dataset across a variety of biomarkers to test the generalisability of these models.

Future of SSL in Digital Pathology. As the field of SSL models continues to advance with the introduction of new techniques and expanded training datasets, the potential for clinical application grows. Innovative models such as CONCH[20], which are trained on both vision and language data, significantly enhance interpretability and feature understanding—key aspects for clinical deployment. Future research should concentrate on establishing the robustness and efficacy of these models to meet the rigorous demands of clinical use. This effort should involve rigorous validation across diverse datasets, population groups, and crucially, a variety of downstream diagnostic tasks to ensure broad applicability and reliability in real-world clinical settings. Moreover, the development of more open-source models would encourage continual technological advancements, fostering greater collaboration and potentially accelerating improvements in clinical outcomes and patient care.

3 Method

3.1 Dataset

We utilise the SR386 subset of the SurGen dataset, consisting of 423 patients, for which a single WSI is provided in Carl Zeiss Image (CZI) format. Slides were

Table 1. Comparison of feature extractors detailing the architecture, training dataset size, and computational resources. Note: "M" denotes million and "K" denotes thousand. Training images for CTransPath comprise 15,580,262 patches from 32,220 WSIs at ×20 magnification (1024×1024 pixels). Phikon was trained on 43,374,634 patches from 6,093 WSIs at ×20 magnification (224×224 pixels). ResNet-50 utilised 1,281,167 natural images from ImageNet. UNI was trained on 100,130,900 patches from 100,426 WSIs coined as Mass-100K. Mass-100K is a closed source proprietary dataset consisting of 75,832,905 images at 256×256 pixels at ×20 magnification in addition to 24,297,995 images at 512×512 pixels at ×20 magnification.

Feature Extractor	Training Images	WSI Count	Data Source	Architecture Backbone	Training Regime	Feature Dimension	Number of Parameters	Time	Resource
CTransPath[31]	15.58M patches	32.2K	TCGA[32], PAIP[18]	Swin-T/14[19]+CNN	MoCoV3[6]	768	27,520,038	250h	48 V100s
Phikon[11]	43.37M patches	6.0K	TCGA[32]	ViT-B/16[10]	iBOT[33]	768	85,798,656	1,216h	32 V100s
ResNet-50[14]	1.28M images	n/a	ImageNet[27]	ResNet-50[14]	ResNet-50[14]	2048	25,557,032	n/a	n/a
UNI[4]	100.13M patches	100.4K	Mass-100K[4]	ViT-L/16[10]	DINOv2[25]	1024	303,350,784	32h	32 A100s

hematoxylin-eosin (H&E) stained and digitised using a ZEISS Axioscan 7 at ×40 magnification ($0.1112\mu m$).

Immunohistochemistry (IHC) for MMR proteins were used to identify MLH1, PMS2, MSH2, MSH6 by way of multiplex ligation-dependent probe amplification (MLPA). Due to the sparsity in these labels and all samples belonging to *no MMR loss* (n=391), *MLH1+PMS2 loss* (n=28), and *PMS2* (n=4); These labels were binarised to *no loss*; microsatellite stable (MSS)/ mismatch repair proficient (pMMR) and *loss*; microsatelite instablilty (MSI)/ mismatch repair deficient (dMMR).

Importantly, this dataset has not been used in the training of any foundational models evaluated in this study, ensuring its suitability as an independent test set for downstream classification tasks and for benchmarking foundational models.

Data Availability The data used to support the findings of this study, including the SR386 subset of the SurGen dataset, can be accessed at https://github.com/CraigMyles/SurGen-CRC-Arena.

Ethics Board Approval This research and use of data has been approved by the University of St Andrews School of Computer Science Ethics Committee (Approval Code: CS16224). All patient data were anonymised prior to our access, ensuring compliance with data protection regulations and maintaining the confidentiality of patient information.

Data Structuring and Stratification To ensure a robust and valid analysis, the dataset was carefully structured into training, validation, and test sets, with splits stratified by age, sex, MMR status, BRAF mutation, RAS mutation, and five-year survival. This stratification ensures that each set is representative of the overall dataset, maintaining consistency across key demographic and clinical characteristics. These variables were chosen based on their potential influence on treatment outcomes and prognosis in colorectal cancer. For a detailed breakdown of the data distribution across the training, validation, and test sets, see Table 2.

Table 2. Breakdown of *SurGen SR386* Cohort data distribution for train, validate, and test sets. Each patient has precisely one associated whole slide image.

Category	Total	Train	Validate	Test
Origin	Scotland	Scotland	Scotland	Scotland
WSI file format	CZI	CZI	CZI	CZI
Magnification	×40	×40	×40	×40
Microns per pixel (pixel width)	$0.1112\mu m$	$0.1112\mu m$	$0.1112\mu m$	$0.1112\mu m$
Number of patients	423 (100%)	255 (60%)	84 (20%)	84 (20%)
Mean age at diagnosis (std. dev.)	67.89 (±11.97)	67.98 (±12.12)	67.71 (±11.40)	67.80 (±12.20)
Male, n (%)	228 (54%)	138 (54.1%)	46 (54.7%)	44 (52.3%)
Female, n (%)	195 (46.0%)	117 (45.8%)	38 (45.2%)	40 (47.6%)
MSS/pMMR, n (%)	391 (92%)	235 (92%)	78 (93%)	78 (93%)
MSI/dMMR, n (%)	32 (8%)	20 (8%)	6 (7%)	6 (7%)
Five year survival (true), n (%)	159 (38%)	100 (39%)	30 (36%)	29 (35%)
Five year survival (false), n (%)	264 (62%)	155 (61%)	54 (64%)	55 (65%)
RAS mutation, n (%)	158 (37%)	97 (38%)	31 (37%)	30 (36%)
RAS wild type, n (%)	265 (63%)	158 (62%)	53 (63%)	54 (64%)
BRAF mutation, n (%)	47 (11.1%)	29 (11.4%)	9 (10.7%)	9 (10.7%)
BRAF wild type, n (%)	375 (88.6%)	225 (88.2%)	75 (89.2%)	75 (89.2%)
BRAF fail, n (%)	1 (0.2%)	1 (0.4%)	0 (0%)	0 (0%)

3.2 Pipeline

We implement background subtraction methods from [21] which utilises threshholding and morphological operations in order to segment tissue regions and remove holes across each WSI in the dataset. These are stored background and non-background regions which are then tessellated across. We extract non-overlapping tissue patches of size 224×224 at ×20 magnification or 1.0 microns per pixel (MPP).

These extracted patches are then fed into the respective feature embedding models in a frozen state, see table 1 for overview feature extractors. We deployed CTransPath, Phikon, ResNet-50, and UNI to provide an effective comparison feature extractors. For the case of ResNet-50, the model is implemented as normal but with the final fully-connected classification layer removed. This effort effectively generates four additional datasets which are compressed and computed by these foundation models.

These specialist feature vector embeddings are then benchmarked against one another with the use of a Transformer Network, which has been shown to be an effective method for computational pathology biomarker prediction[30].

The extracted feature vectors serve as input for a Transformer Network. The network is configured to operate with a batch size of one to manage compute limitations effectively. This setup ensures that all available patches can be processed in memory without reaching hardware limits. The performance of each feature extractor model is evaluated against others by benchmarking their outputs using the Transformer Network. This comparative analysis helps determine which model provides the most informative features for accurate biomarker prediction. The specifics of the Transformer's configuration, such as the number of encoder layers and attention heads, are optimised through a grid search, details of which are summarised in Table 3. See figure 1 for a pipeline overview.

Table 3. Grid search hyperparameters for optimising the transformer-based model. This table includes both structural parameters that define the model's architecture and adjustable hyperparameters subject to optimisation. *Model dimension* (512) specifies the size of the input and output layers. *Feedforward dimension* (2048) indicates the size of the inner layer of the feedforward networks within each transformer block. *Transformer Heads* (4) enhance the model's ability to attend to different aspects of the input data simultaneously through multiple attention mechanisms. *Encoder layers* (6) describe the number of sequential layers within the model, affecting depth and complexity. The table also lists ranges for *Activation function*, *Dropout*, and *Learning rate* to explore their impact on model performance, along with different *Feature extractors* used to assess comparative effectiveness.

Parameter	Values	Justification
Epoch	200	Value determined after initial experimentation showed some models did not converge within 50 epochs. ResNet-50 in particular was seen to take longer to converge across runs, contradicting [3].
Model Dimension	512	Typical for transformer networks, and aligns with [30].
Feedforward Dimension	2048	The feedforward dimension of 2048, typically four times the model dimension, allows for a more expansive representation in the feedforward layers of the transformer.
Transformer Heads	4	Whilst [30] use 8 heads, we reduce this slightly to find a balance with the number of encoder layers.
Encoder Layers	6	Increase number of encoder layers compared to [30] in a conscientious effort to account for the larger feature vectors generated by UNI (1024-dim), and ResNet-50 (2048-dim), and any additional complexities contained by these which the model may be able to pick up on.
Activation Function	relu, gelu	The inclusion of both ReLU and GELU activation functions in the grid search provides options for non-linearity. ReLU is traditionally used for its simplicity and effectiveness in avoiding vanishing gradient issues, whereas GELU, as a smoother alternative, can potentially offer better performance and convergence properties as indicated by recent research[15][8].
Dropout	0.05, 0.1, 0.15, 0.2, 0.25, 0.5	Whilst [30] does not report utilising dropout in their experiments, it is an effective tool for regularisation and can improve model generalisability[1][8].
Learning Rate	1e-04, 1e-05, 1e-06, 1e-07	Covers ranges seen in [30] and covers values below and above.
Feature Extractor	CTransPath, Phikon, ResNet-50, UNI	Multiple models are evaluated to determine which offers the most useful features for enhancing transformer performance on downstream tasks.

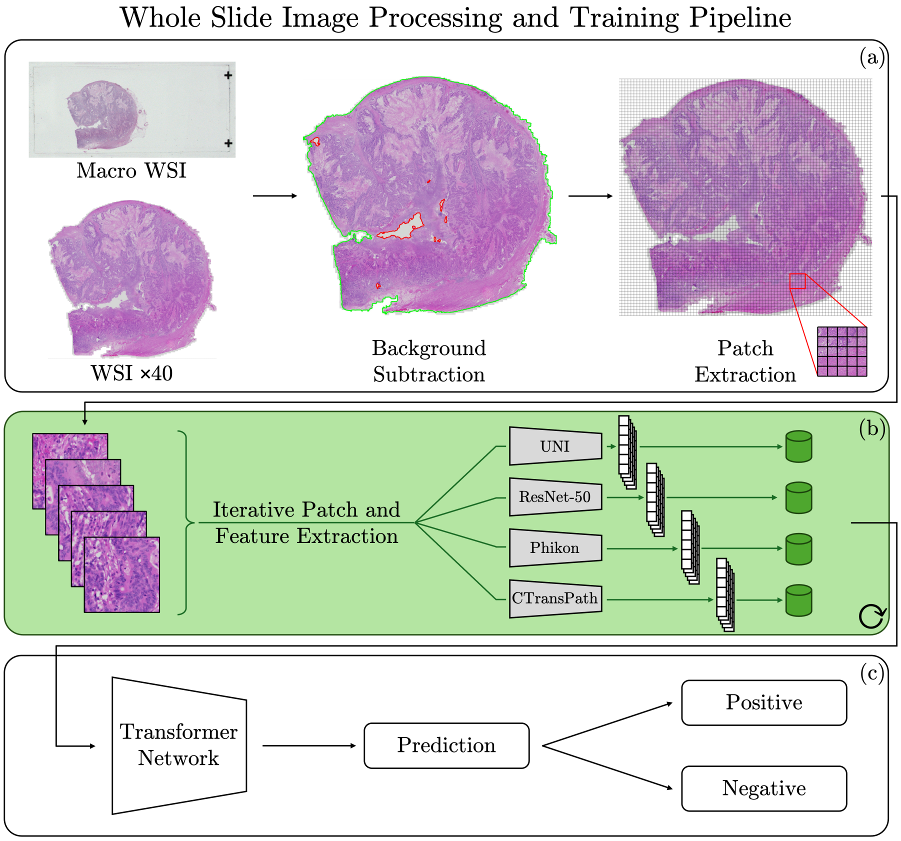

Fig. 1. Preprocessing and Transformer Pipeline – Overview of the development and processing pipeline with (a) Background removal and patch extraction, highlighting tissue (in green) and voids (in red); (b) Feature extraction on tissue patches, iterating across all 224×224 non-overlapping patches at ×20 magnification. Feature vectors for each WSI are stored to disk to enable fast and iterative Transformer training later in the pipeline; (c) Transformer network trained to predict binary biomarker labels.

Compute Resources. We utilised a NVIDIA DGX-1 with 8 V100-32GB GPUs and 503GiB RAM. Incurring a total of 1775.5 GPU hours across tile feature extraction and model training.

4 Results

The hyperparameter optimisation process and its impact on model performance are illustrated in Figure 2, which displays a parallel coordinates plot for various parameters against the validation area under the receiver operating characteristic (AUROC) score.

Table 4 compares the best-performing validation AUROC results from transformer models trained using feature embeddings from CTransPath, UNI, Phikon, and ResNet-50. The model employing CTransPath feature vectors achieved the highest validation AUROC of 0.9466 for mismatch repair (MMR) classification. In light of this, CTransPath feature vectors were further utilised for downstream tasks.

Additional performance metrics for models trained with CTransPath feature embeddings across various tasks are detailed in Table 5.

Lastly, Figure 3 depicts the Receiver Operating Characteristic (ROC) curves for the best-performing model configurations across all feature vectors on the test set, providing a clear depiction of the models' predictive accuracy in a clinical context.

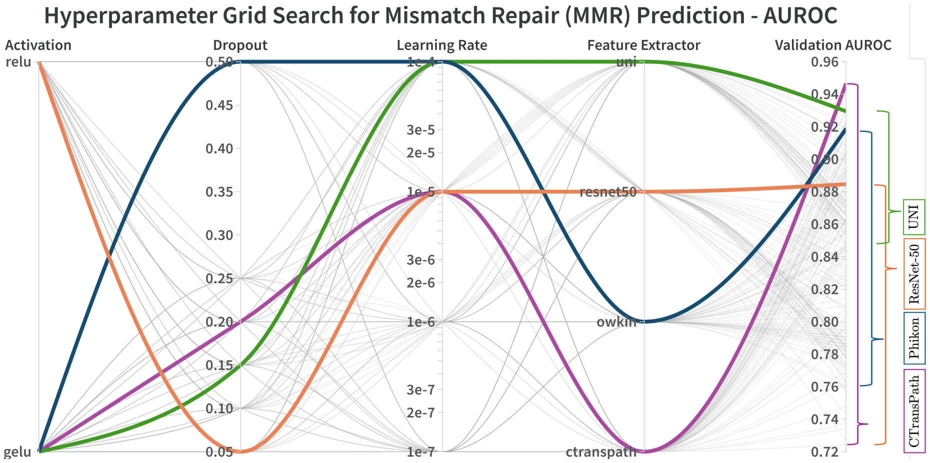

Fig. 2. Parallel coordinates plot illustrating the optimisation of hyperparameters for mismatch repair (MMR) prediction, showing their impact on the validation AUROC score. Each line represents an individual run. Highlighted runs represent best performing hyperparameter combination for each feature extractor respectively. Range bars highlight the variance in model performance across runs by feature extractor. This visualisation aids in identifying the most effective hyperparameter combinations for model performance.

5 Discussion

The comparison between the UNI and CTransPath models highlights an intriguing aspect of model selection for digital pathology. The UNI model has shown consistent robustness, with validation AUROC scores generally ranging between 0.9295 and 0.8483, as illustrated in the parallel coordinates plot (Figure 2). This

Table 4. Comparative table displaying the best validation AUROC scores alongside test AUROC scores for mismatch repair (MMR) prediction, achieved by Transformer models trained with feature embeddings from CTransPath, UNI, Phikon, and ResNet-50.LR = Learning Rate

Feature Extractor	Task	Tests AUROC	Val AUROC	Epoch	LR	Dropout	Activation
CTransPath	MMR	0.6880	0.9466	33	1e-05	0.2	gelu
UNI	MMR	0.7136	0.9295	2	1e-04	0.15	gelu
Phikon	MMR	0.7136	0.9188	7	1e-04	0.5	gelu
ResNet-50	MMR	0.6709	0.8846	77	1e-05	0.05	relu

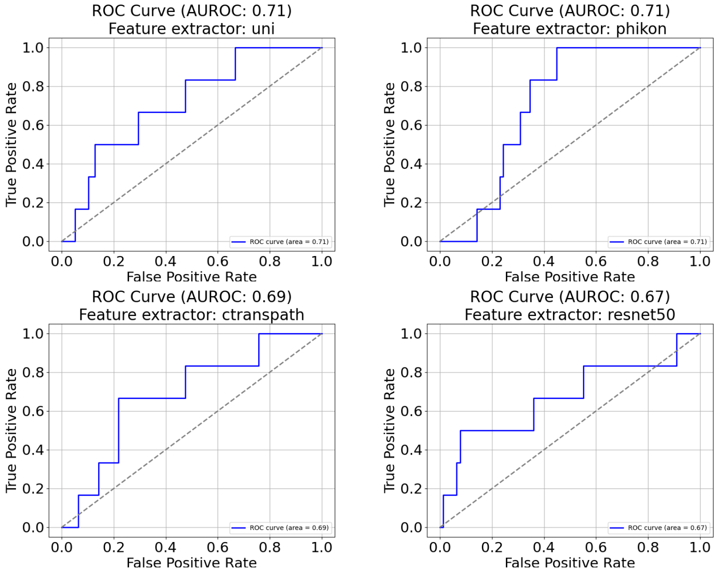

Fig. 3. Receiver Operating Characteristic (ROC) curves for the optimal hyperparameter settings across four feature extractor models tested on the test set, illustrating their predictive performance and diagnostic accuracy.

Table 5. Table comparing the performance of Transformer models trained on CTransPath feature embeddings for different diagnostic tasks. This table presents the validation and test AUROC scores, demonstrating the models' predictive accuracy across tasks such as Mismatch Repair (MMR), BRAF Mutation, 5-Year Survival, and RAS Mutation. LR = Learning Rate.

Feature Extractor	Task	Test AUROC	Val AUROC	Epoch	LR	Dropout	Activation
CTransPath	MMR	0.6880	0.9466	32	1e-05	0.2	gelu
CTransPath	BRAF_M	0.4667	0.8074	39	1e-07	0.25	relu
CTransPath	5Y_SUR	0.5724	0.7574	179	1e-07	0.1	relu
CTransPath	RAS_M	0.4031	0.6954	87	1e-04	0.25	gelu

consistency suggests a higher reliability for clinical applications where variability in performance could lead to significantly different diagnostic outcomes.

Conversely, while the CTransPath model achieved the highest validation AUROC of 0.9466, it also showed a surprising breadth in performance during training, with some configurations dipping as low as 0.7244 AUROC. This variability may pose a risk in clinical settings where consistent performance is crucial. The best-performing configuration of CTransPath, however, was utilised for further downstream tasks due to its peak performance, as detailed in Table 5.

Additionally, a critical examination of benchmark results from other studies, such as those by Wagner et al. [30], raises questions about direct comparisons. Whilst they achieved an AUROC of 0.9 with 250 samples for MMR status prediction, their report lacks details on the distribution of labels in these subsets, which makes it challenging to assess how well their findings relate to ours.

This analysis also revealed that ResNet-50, although traditionally used for natural image processing, performed adequately and sometimes surpassed other models under specific configurations. This adaptability suggests that even traditional architectures can be competitive when optimally tuned, though care must be taken to manage potential overfitting, a common issue with ResNet-50 that could be mitigated with strategies like early stopping.

Overall, the findings underscore the importance of not only considering the highest accuracy or AUROC scores when selecting models but also assessing their performance consistency and reliability. Future work will focus on refining these models further, optimising their configurations for broader tasks within digital pathology, and validating their effectiveness across different diagnostic challenges. This approach will help ensure that the models not only achieve high accuracy but are also robust and reliable enough for clinical application.

5.1 Future work

Scope for future work includes implementing and investigating methods such as classification thresholding as well as techniques such as cosine annealing for learning rates as well as gradient accumulation which could potentially improve model convergence and generalisation capabilities on smaller whole slide image datasets. Adjustments to the transformer network architecture, such as modifying the number of layers and attention heads, could also be investigated to optimise processing and improve learning efficiency for varied histopathological features. Future foundational models may also improve upon the metrics achieved in this study.

Additionally, improving diagnostic accuracy and model generalisability will be crucial. This could involve implementing attention-based saliency mapping [23] to provide deeper insights into the model's decision-making processes, thereby enhancing interpretability and clinical relevance.

6 Conclusion

This research has explored the capabilities of state-of-the-art self-supervised learning models for digital pathology, particularly focusing on their ability to classify mismatch repair (MMR) status from whole slide images (WSIs) with a sparse label distribution. The models investigated—CTransPath, Phikon, and UNI—demonstrate significant potential to enhance the accuracy of diagnostic predictions, with the UNI model showing exceptional promise due to its robustness and stability as evidenced by its superior performance over the conventional, ImageNet trained, ResNet-50.

The findings reveal that while achieving high AUROC scores on validation datasets, such as CTransPath's peak of 0.9466, the generalisation of these models can be limited in scenarios where labels are sparse or have an significant imbalance in distribution. The UNI model, however, with a test AUROC of 0.7136, suggests a promising avenue for clinical use, showing an improvement of 6.3% over ResNet-50's performance, which underscores the viability of SSL models trained on large-scale datasets to handle real-world complexities in diagnostic tasks.

This study has demonstrated the efficacy of state-of-the-art self-supervised learning models in classifying mismatch repair status from whole slide images, its findings also underscore a broader applicability. Specifically, the techniques and methodologies refined through this research hold promise for tackling diagnostic challenges associated with rarer diseases and conditions characterised by smaller datasets. The ability to effectively utilise sparse data could significantly enhance diagnostic accuracy and patient outcomes in less common pathologies, thereby extending the benefits of advanced computational techniques to a wider range of clinical scenarios. Such potential highlights the importance of continuing to advance and adapt machine learning strategies to meet diverse and critical healthcare needs, ultimately driving forward the impact of computational pathology in personalised medicine.

Acknowledgements CM is supported by NHS Lothian. The authors would like to thank NHS Lothan for providing tissue specimen. This work is supported in part by the Industrial Centre for AI Research in Digital Diagnostics (iCAIRD) which is funded by Innovate UK on behalf of UK Research and Innovation (UKRI) (project number 104690).

References

1. Baldi, P., Sadowski, P.J.: Understanding dropout. Advances in neural information processing systems **26** (2013)
2. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **68**(6), 394–424 (2018)

3. Campanella, G., Kwan, R., Fluder, E., Zeng, J., Stock, A., Veremis, B., Polydorides, A.D., Hedvat, C., Schoenfeld, A., Vanderbilt, C., et al.: Computational pathology at health system scale–self-supervised foundation models from three billion images. arXiv preprint arXiv:2310.07033 (2023)
4. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine pp. 1–13 (2024)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
6. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9640–9649 (2021)
7. Dawson, H.: Digital pathology–rising to the challenge. Frontiers in medicine **9**, 888896 (2022)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Dippel, J., Feulner, B., Winterhoff, T., Schallenberg, S., Dernbach, G., Kunft, A., Tietz, S., Jurmeister, P., Horst, D., Ruff, L., et al.: Rudolfv: A foundation model by pathologists for pathologists. arXiv preprint arXiv:2401.04079 (2024)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Mac Kain, A., Saillard, C., Schiratti, J.B.: Scaling self-supervised learning for histopathology with masked image modeling. medRxiv pp. 2023–07 (2023)
12. Goyal, P., Mahajan, D., Gupta, A., Misra, I.: Scaling and benchmarking self-supervised visual representation learning. In: Proceedings of the ieee/cvf International Conference on computer vision. pp. 6391–6400 (2019)
13. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
16. Iizuka, O., Kanavati, F., Kato, K., Rambeau, M., Arihiro, K., Tsuneki, M.: Deep learning models for histopathological classification of gastric and colonic epithelial tumours. Scientific reports **10**(1), 1504 (2020)
17. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3344–3354 (2023)
18. Kim, Y.J., Jang, H., Lee, K., Park, S., Min, S.G., Hong, C., Park, J.H., Lee, K., Kim, J., Hong, W., et al.: Paip 2019: Liver cancer segmentation challenge. Medical image analysis **67**, 101854 (2021)
19. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings

of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)

20. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine pp. 1–12 (2024)
21. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)
22. McCarthy, A.J., Capo-Chichi, J.M., Spence, T., Grenier, S., Stockley, T., Kamel-Reid, S., Serra, S., Sabatini, P., Chetty, R.: Heterogenous loss of mismatch repair (mmr) protein expression: a challenge for immunohistochemical interpretation and microsatellite instability (msi) evaluation. The Journal of Pathology: Clinical Research **5**(2), 115–129 (2019)
23. Mohammadi, M., Cooper, J., Arandelović, O., Fell, C., Morrison, D., Syed, S., Konanahalli, P., Bell, S., Bryson, G., Harrison, D.J., et al.: Weakly supervised learning and interpretability for endometrial whole slide image diagnosis. Experimental Biology and Medicine **247**(22), 2025–2037 (2022)
24. Montezuma, D., Monteiro, A., Fraga, J., Ribeiro, L., Gonçalves, S., Tavares, A., Monteiro, J., Macedo-Pinto, I.: Digital pathology implementation in private practice: specific challenges and opportunities. Diagnostics **12**(2), 529 (2022)
25. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
26. Retamero, J.A., Aneiros-Fernandez, J., Del Moral, R.G.: Complete digital pathology for routine histopathology diagnosis in a multicenter hospital network. Archives of pathology & laboratory medicine **144**(2), 221–228 (2020)
27. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
29. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Liu, S., Mathieu, P., van Eck, A., Lee, D., Viret, J., et al.: Virchow: A million-slide digital pathology foundation model. arXiv preprint arXiv:2309.07778 (2023)
30. Wagner, S.J., Reisenbüchler, D., West, N.P., Niehues, J.M., Zhu, J., Foersch, S., Veldhuizen, G.P., Quirke, P., Grabsch, H.I., van den Brandt, P.A., et al.: Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. Cancer Cell **41**(9), 1650–1661 (2023)
31. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis **81**, 102559 (2022)
32. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M.: The cancer genome atlas pan-cancer analysis project. Nature genetics **45**(10), 1113–1120 (2013)
33. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)