
Article:
Ogundipe, O., Zhai, B. orcid.org/0000-0003-1635-1406, Kurt, Z. orcid.org/0000-0003-3186-8091 et al. (1 more author) (2025) Explainable colon cancer stage prediction with multimodal biodata through the attention-based transformer and squeeze-excitation framework. Current Bioinformatics. ISSN 1574-8936

https://doi.org/10.2174/0115748936309582240907160359

Explainable colon cancer stage prediction with multimodal biodata through the attention-based transformer and squeeze-excitation framework

University of Northumbria,
UK

Abstract

The heterogeneity in tumours poses significant challenges to the accurate prediction of cancer stages, necessitating the expertise of highly trained medical professionals for diagnosis. Over the past decade, the integration of deep learning into medical diagnostics, particularly for predicting cancer stages, has been hindered by the black-box nature of these algorithms, which complicates the interpretation of their decision-making processes. This study seeks to mitigate these issues by leveraging the complementary attributes found within functional genomics datasets (including mRNA, miRNA, and DNA methylation) and stained histopathology images. We introduced the Extended Squeeze-and-Excitation Multiheaded Attention (ESEMA) model, designed to harness these modalities. This model efficiently integrates and enhances the multimodal features, capturing biologically pertinent patterns that improve both the accuracy and interpretability of cancer stage predictions. Our findings demonstrate that the explainable classifier utilised the salient features of the multimodal data to achieve an area under the curve (AUC) of 0.9985, significantly surpassing the baseline AUCs of 0.8676 for images and 0.995 for genomic data. Furthermore, the extracted genomics features were the most relevant for cancer stage prediction, suggesting that these identified genes are promising targets for further clinical investigation.

Keywords: Multimodal, Explainable, genomics, H&E Images, Attention model, features-relevance.

1. INTRODUCTION

Colon cancer is a disease which occurs within the colon or rectum of the sizeable abdominal intestine [1]. Correct estimation or prediction of cancer stages is directly related to the patient's prognosis, survival outlook, and treatment management [2–4]. Studies have shown that multimodal data leads to improving the performance of the predictive models for cancer stages, advancing precision oncology, improving survival risk groups, and answering complex cancer questions [5–7]. To understand the complex biology of colon cancer, different fluids and tissues within the body have been studied to comprehend its occurrence properly [8–11]. Some of the data types studied for cancer diagnosis, prognosis, and drug discovery include messenger Ribonucleic Acid (mRNA), micro ribonucleic acid (miRNA), Deoxyribonucleic acid (DNA) Methylation, Copy Number Alteration (CNA), somatic mutation, single-stranded genomics structures, computed tomography images and Hematoxylin and Eosin (H&E) stained histopathological images, among others [12–15]. Several studies involving laboratory experiments have identified numerous potential oncogenes or tumour suppressors as biomarkers for early cancer detection and diagnosis. Other studies have revealed genetic alterations that drive the initiation and progression of colon polyps to colorectal cancer (CRC), while some have identified tumour antigens and designed genomic vaccines [14,16–21].

Dimensionality, explainability, and insufficient annotated data are among the major challenges stalling the adoption of deep neural models into medical assistive support systems [22–27]. Available data for cancer stage prediction exists in high-dimensional space; for example, DNA methylation data contains tens of thousands of features [11,28,29].

Hence, multimodal data will inhabit a larger high-dimensional space, leading to model overfitting, a high risk of multicollinearity, and poor generalisation. These challenges underline the need for novel dimensionality reduction methods to acquire effective low-dimensional data with relevant patterns. In contrast, explainability [24,30,31] focuses on understanding how millions of parameters work simultaneously to arrive at a decision. Explainability can also be defined as the mathematical quantification of how influential or 'relevant' features are to a certain prediction. Different deep network models has been proposed to resolved those challenges [21,32,33].

According to the International Agency for Research on Cancer (IARC): Cancer Tomorrow, the estimated future incidence rate of colon cancer between the years 2022 and 2045 for the 185 countries under consideration will increase by 60.72%[1]. This is an eye-opener that more needs to be done to develop robust computational methods for timely and accurate cancer stage prediction, thereby increasing the survival rate of cancer patients. We hypothesise that integrating extracted features from histopathological images and genomics, comprising DNA methylation, miRNA, and mRNA, will provide a balanced biological pattern for effective cancer stage prediction. First, histopathological images have revealed the tumour environment, including invasive growth patterns and various mutational changes [34–36]. Second, the expression of miRNA, mRNA, and DNA methylation[37–39] has been identified as a key driver in the pathogenesis of diseases in cancer. While the miRNA [7,40], a non-coding RNA, inhibits gene expression at the transcriptional and post-transcriptional levels, DNA methylation plays a crucial role in transcriptional activities and acts as a gene suppressor. The mRNAs [41–43], protein-coding RNA, and single-stranded ribonucleic acid molecules transcribed from a DNA strand carry genetic information that guides protein synthesis. Integrating features from these four data types (i.e., images, miRNA, mRNA, DNA methylation) is expected to improve diagnostic performance. In this research, we propose a novel model incorporating multimodal features to analyse the interrelations between genomic data and histopathological image features in a condensed feature space, aiming to enhance the predictive performance for colon cancer stages. The proposed model, the ESEMA (Extended Squeeze-and-Excitation Multiheaded Attention) network, incorporates dual submodules designed to capture the complementary features from genomics and pathological perspectives. Moreover, an in-depth investigation into the critical genomic features is conducted by examining the response of neurons within the bottleneck layers of the network. The structure of the proposed ESEMA model is illustrated in Fig 1. The overarching objective of developing the ESEMA model is to augment the efficacy and interpretability of the cancer staging prediction model. This enhancement is achieved through the dual mechanisms of channel-wise recalibration of input features via a squeeze-and-excitation block and the establishment of global dependencies among histopathology image patches facilitated by attention mechanisms.

During the study, our attention-based colon cancer diagnostic classifiers achieved the listed novelties:

- Using integrated complementary features from genomics and histopathological images for colon cancer stage prediction.
- Proposing the ESEMA model for analysing interrelations between genomics and histopathological images for colon cancer stage prediction.

[1] https://gco.iarc.fr/tomorrow/en/about

- Generate individual scores/weights describing the relevance of each transformational feature in predicting the colon cancer stages and identifying likely relevant cancer-linked genes at different cancer stages.

The remaining sections of this study are as follows: Related works are the basis for Section 2. The materials and techniques are covered in Section 3. The results are presented in Section 4, and the discussion and conclusion are covered in Sections 5 and 6, respectively.

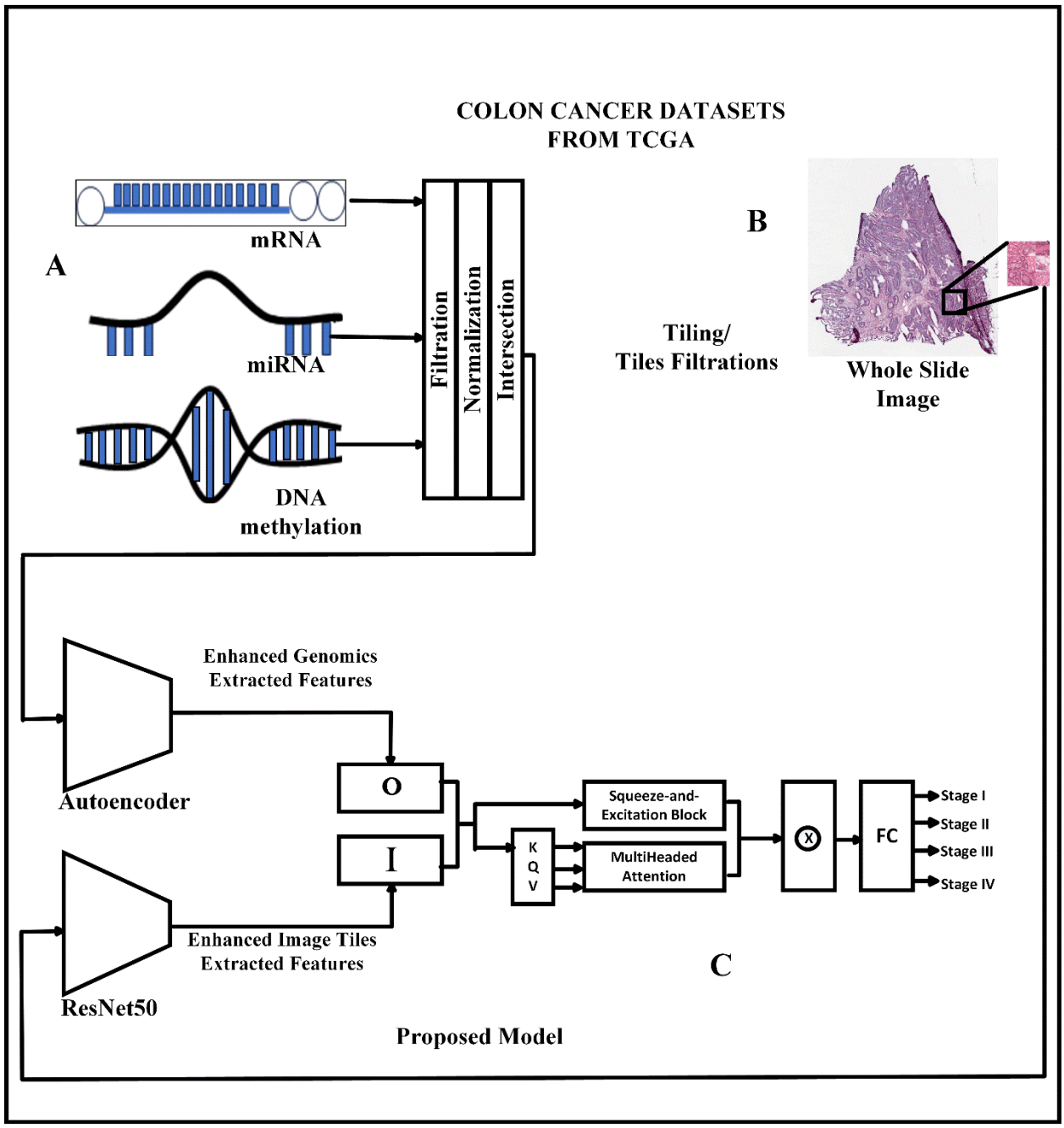

Fig. 1 The proposed Extended Squeeze-and-Excitation with Multiheaded Attention (ESEMA) framework consists of (A) the preprocessing steps, the extraction and fusion of features from DNA-Methylation, miRNA, and mRNA, and

(B) histopathological images. (C) The combined architecture of the Squeeze-and-Excitation block and the Multiheaded Attention.

2. RELATED WORKS

2.1 Pathological methods of cancer staging.

Cancer staging quantifies the progression of cancer in a patient. The stage of cancer helps medical experts determine the best treatment option. The American Joint Committee on Cancer (AJCC) staging system is the most common and often called a pathological Tumour-Nodes-Metastasis (TNM) system. This system combines three major factors to determine the stage of colon cancer. The first factor is the tumour size (T), which defines how extensively cancer has grown into the wall of the colon. The second factor is whether the cancer cells have spread to the nearby lymph nodes (N), while the third factor, metastasis (M), determines whether cancer has spread to distant lymph nodes or body organs. Also, there is clinical colon cancer staging, which focuses on physical examination, biopsies, and imaging tests. Hence, in this study, the datasets used for computational prediction of the TNM system of colon cancer staging are multimodal, comprising pathological images and genomics datasets.

2.2 Deep Neural Network (DNN) in cancer stage prediction.

DNN has been used extensively in cancer-related research [36,44–53] specifically in metastasis diagnosis, in which cancer has spread to other parts or organs of the body. This finding assists in developing appropriate treatment plans for cancer patients [54,55]. Also, it is used in the study of oncogenesis or tumorigenesis-the study of cancer genesis or evolution through which normal cells develop into cancer cells, or the examination of the causes of a cell's abnormal transformations resulting from different biological processes, among which are genetics, epigenetics, and cellular division. Harnessing complementary features from data has been shown to enhance researchers' understanding of cancer metastasis and tumorigenesis [56–60]. This knowledge will be suitable for enhancing and improving our understanding of cancer stage prediction and identification of relevant features on samples for precision oncology. This will eventually translate into an increase in survival outcomes [61,62] and the discovery of new biological molecules for new drug targets [63–67].

2.3 Explainable AI (Artificial Intelligence)

The UK government, the European Commission, and other world governing bodies are currently adopting a regulatory framework to address the risks and trust issues and improve the explainability and other challenges posed by the adoption of DNN applications in areas considered to be of high or unacceptable risk [68–71]. In critical AI decision-making, explainability focuses on comprehending the model's expected impact and biases. It is anchored on the system's accuracy, fairness, and transparency in its decision-making approach. Experts and legal regulations require and expect DNN models to give reasons for or justify their decision-making process. The earliest works of building explainable functions into the DNN framework involved the introduction of causality concepts and LIME (Local Interpretable Model-agnostics Explanations) [72,73] while the frontier research of building explainable algorithms uses attention techniques [74–82].

In complex and high-risk decision-making applications such as cancer diagnostics, prognostics, and therapeutic discovery, the improved explainability will encourage more adoption of AI in effective clinical decision-making and support [83–86]. Recently, researchers developed an AI Multi-omics Multi-cohort Assessment (MOMA), which was designed to offer transparent reasoning by generating correlations that can be interpreted using existing pathological terms. Likewise, the AlphaFold explainable project has resolved a decade of challenges in generating the protein 3D structure required for drug development [87–90]. Various frontiers of learning to infuse knowledge and reasoning into neural network models are categorised as explainable artificial intelligence (XAI). Some of the most promising results are built on methodologies such as neuro-symbolic, fuzzy, Bayesian and action influence graphs [91–96].

Explainable neural networks have been applied in many medically related diagnostics to discern how the system arrived at its decision-making [69,73,83,97]. DNN explainable tools that we consider in this research are in areas of performance enhancement, feature importance identification and visualising internal system dynamics. We compute feature importance on the multimodal input feature by estimating the weight relevance of the input features in predicting the stage of the given cancer and construct an attention map - a visualisation tool which gives a visual representation of the internal workings of a given function.

2.4 Aggregating different Attention Learning Frameworks

In this study, we augmented two different attention-based models [98–101]. The first model, a multiheaded attention block, used a weighted computation method on the integrated latent features to capture long-range dependency within a sample. The second model is a squeeze-and-excitation block [102] that compresses the spatial dimension, aggregates relevant features, and determines relevance along the channel metrics. Aggregating the two attention methods leads to a robust method for effective cancer-stage prediction.

2.5 Cancer research with unimodal or multimodal data

Several works have demonstrated the power of harnessing complementary representations of features in multimodal datasets for improved precision in cancer oncology [56,103,104]. A substantial challenge encountered in cancer research when employing unimodal data alongside DNNs is the reduced accuracy arising from their limited predictive capabilities [105–107]. We took measures to resolve this difficulty by integrating features from four different types of data. We integrate extracted feature representation from multimodal data to improve predictive output. In this study, we experiment and analyse predictive accuracy based on unimodal and multimodal data. We extract relevant features from sample genomics data comprising mRNA, miRNA, and DNA Methylation and integrate them with salient extracted features from H&E histopathological images of the same individuals. Since the genomics and images of a sample carry separate and complementary features of the same biological disease, we perform various experiments and analyses to verify the validity of our hypothesis that the integrated features will provide a robust multimodal statistically prognostic feature and hence give an improved prediction compared with when only unimodal data are used. Our ESEMA model achieves new state-of-the-art performance for colon cancer stage prediction compared to other state-of-the-art models, as shown in section 4.2.1.

3. MATERIALS AND METHODS

3.1 Dataset

The data used in this study was downloaded from The Cancer Genome Atlas (TCGA) with the TCGA-Assembler function. The dataset with four modalities with colon cancer characteristics was downloaded. The datasets are DNA methylation, miRNA, mRNA, Hematoxylin, and Eosin (H&E) stained histopathological images. The dataset has an unequal number of samples across the four different data types: 328 samples exist in mRNA expression, 261 samples in miRNA expression, 353 samples of DNA methylation, and 463 samples of histopathological images, respectively. Since the study was designed to extract and integrate representational features on the same samples with available data across the four datasets, we used the 177 overlapping samples that are found across all four data types, as summarised in Table 1 and Table 2, showing training and the testing sample size across each stage.

Table 1. Sample and Feature size of genomics data per subject.

Biological data	No. of Overlapping Samples	No. of Features
mRNA	177	16,377
miRNA	177	420
DNA methylation	177	20,129

Table 2. Sample Size Distribution across Whole Slide Images per stage

Stages	Sample Size	Training Sample Size (80%)	Testing Sample Size (20%)	Number of Tiles
I	30	24	6	27,358
II	67	54	13	47,397
III	54	43	11	21,492
IV	26	21	5	15,914
Total	177	142	35	112,161

3.2 Methodology

Let $\mathcal{X} = [\mathrm{X}_{t1}, \mathrm{X}_{t2}, \dots, X_{td}]$ represent a total of d histopathological image tiles from a subject t, and let $\mathcal{Z} = Z_t$ denote the concatenated genomics data (micro-RNA, DNA methylation and mRNA) sequenced from the same subject, both used for predicting the cancer stage y_t. Let the true function, which maps the sample data to the subject's cancer stage, be given as $f: (X, Z) \rightarrow Y$ and defined as:

$$y = f(X, Z) \tag{1}$$

The DNN learns an estimated function $f^*: X, Z \rightarrow Y$, which is so close to the true function f given as:

$$\widehat{y_t} = f^*(X_{td}, Z_t; \theta), \tag{2}$$

Where θ is a set of parameter variables used in estimating $\hat{y}$.

The loss function estimated over a minibatch of size b, using stochastic gradient descent computation is given as:

$$L(XZ, y; \theta) = -\frac{1}{b}\sum_{t=1}^{b} CE(f^*(X_{td}, Z_t), y_t; \theta) \quad (3)$$

$$L(X, y; \theta) = -\frac{1}{b}\sum_{t=1}^{b} y_t * \log \hat{y}_t \quad (4)$$

The CE stands for cross-entropy loss.

3.2.1 Feature Extraction from histopathological whole slide image tile

Features are extracted from image tiles using a ResNet50 pre-trained model. Histopathological feature extraction follows a two-step fine-tuning process using the transfer learning approach. In step 1, we freeze the convolutional base of the pre-trained model, meaning all the trainable layers are set to non-trainable. The model's original classification (output) layer is then replaced with a new classification layer comprising four neurons. This step focuses on training only the new classifier layer with an Adam optimiser at a learning rate of 10^{-4} (40 iterations; batch size = 32).

The second step involves fine-tuning or adapting the weight of the updated model in Step 1 to the histopathological images. This is achieved by selectively unfreezing the uppermost four layers of the model, which are hypothesised to contain more sophisticated representations of the histopathological feature map. We extracted the histopathological feature $I_{oj} \in \mathbb{R}^{2048}$ from a tile image X_{oj} using the fine-tuned ResNet50 encoder.

3.2.2 Feature extractions from genomics data

Before extracting features from genomics data, we separately scale and normalise features in miRNA, mRNA and DNA Methylation and the resulting output features are concatenated to obtain a new multimodal feature. The resulting multimodal genomics dataset with 36,926 features was then used as an input into an autoencoder (AE) neural network. The encoder of AE is then used to encode the raw features with less dimensionality through the bottleneck layer. The encoder and the decoder of sparse AE for the extraction of genomics features were trained to generate a robust feature representation. The AE leverages a scalable hyperparameter optimisation framework that searches the model space for the best hyperparameter values. The output is conjectured to have complementary features for effective cancer stage prediction and stratification purposes, as proposed in [108]. Hence, after the AE was trained, 50 salient genomics features were extracted from the encoder, which were denoted as $O_t \in \mathbb{R}^{50}$.

3.2.3 Reconstruction of the most relevant genes

To understand the significance of gene expression, we employ a masking strategy on the feature representations in the bottleneck layer, aiming to reconstruct the original genetic features. This involves evaluating genes based on the

minimal reconstruction loss. Following this, the feature response values of the bottleneck layer are sorted, and thresholding is applied to selectively mask and retain the feature values (keeping 40, 30, and 20, respectively), as shown in Fig. 2. Subsequently, we compute the intersection of the reconstructed genes (the top 10 genomic feature values with the least reconstruction error) across these varying thresholds to ascertain the relevance of specific genes at different stages of cancer progression. This methodology provides insight into the critical genetic components that are influential across varying cancer stages.

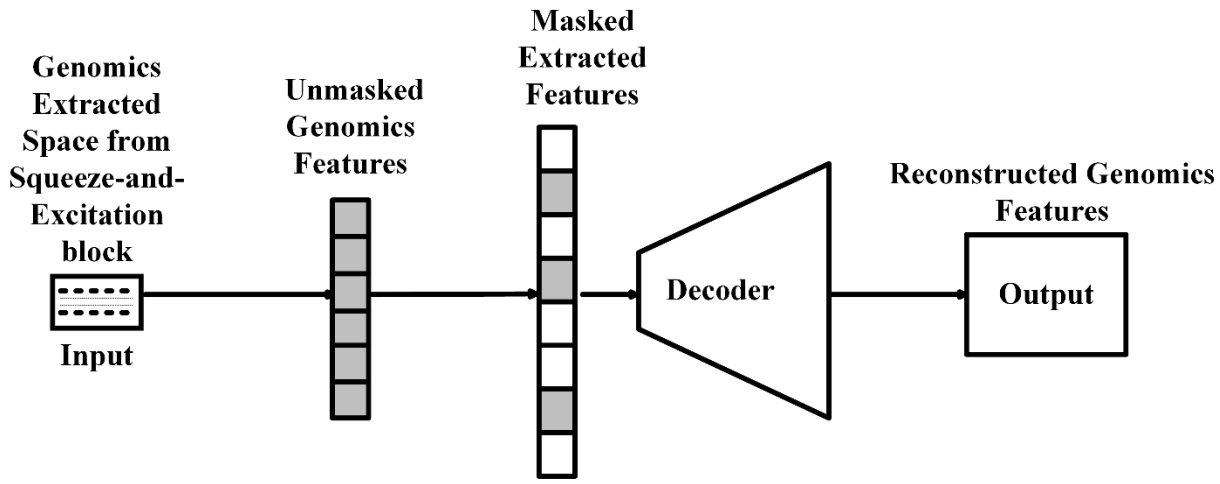

Fig.2 shows the filter genomics latent space from the squeeze and excitation layers. Three separate masking settings of 40%, 60%, and 80% were applied to the genomic latent space, and each masking was decoded and reconstructed into a space of the same dimension n as the initial genomic space before feature extraction. We then reassembled the feature space produced by each masked latent space. Using the reconstruction errors (squared error), we performed a one-to-one mapping of features between the reconstructed and original inputs. We enumerate the index and value of the square error of the features between the real and reconstructed genomic data and sort and return the features in ascending order of square error. The set of most expressed genes is obtained by intersecting the set of features among the three masking operations.

3.3 Experimental Setup

The experiment was conducted to compare the predictive accuracy of our proposed ESEMA model with the other two models, which are the baseline method and multiheaded attention. Each category of three input datasets (images, genomics and fused features) is trained and validated on our proposed model and the stated other two models. The computational efficiency of the proposed model compared with the two other models is tabulated as shown in Tables 3 and 4. Table 5 shows the number of model parameters for each model, and Table 6 compares our proposed model with the existing state-of-the-art models.

3.3.1 Experiment 1: The Baseline Model

The baseline model comprises the input layer, the dense layer, the pooling layer, and the classification layer. The three sets of input features were first trained and analysed using the baseline model to examine the effectiveness of the output of the extracted features when trained on a baseline model. The results of the performance of the three models trained on the fused input features are explained in section 4.3 under the result presentation.

After the initial data preprocessing exercise in Fig. 1A, features are extracted from histopathological images and the genomics data using ResNet50 and sparse autoencoder (AE), as shown in Fig. 1B, respectively. A fine-tuning/transfer learning method on ResNet50 was adopted to extract features from histopathological images. Features extracted from histopathological images, genomics, and their fusion are separately used to train a baseline network and two attention models for explainable cancer stage prediction and feature relevance identification. All experiments were conducted using the TensorFlow deep learning framework.

3.3.2 Experiment 2: The Multi-headed Attention (MA)

We observed that the baseline models performed averagely in fitting the genomics and fused input. Still, there needs to be a clear boundary or difference in the predictive accuracy of genomics from the multimodal features. Also, the need to adopt algorithms with embedded explainable functions calls for another classifier with such features. Hence, the choice of the Multiheaded Attention (MA) framework.

The multiheaded attention framework [99] is based on self-attention intuition [109]. The attention-based framework has proved to be the state-of-the-art model in many challenging machine learning research areas. In the self-attention mechanism, output is influenced by attention to the relevant input features. We combined the genomics feature O and histopathological feature I to form a join modality representation $J \in \mathbb{R}^{1\times 2548}$, where N is the number of elements and d is the number of feature dimensions. The self-attention defines a query $Q \in \mathbb{R}^{N\times d}$, key $K \in \mathbb{R}^{N\times d}$, and value $V \in \mathbb{R}^{N\ x\ d}$ matrix transformation on the input [109]. The score of the dot-product is computed using the query-key matrix given as:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

Followed by multiheaded distribution computation defined over projections of queries, keys and values given as:

$$MultiHead(Q,K,V) = concat(head_1,\dots,head_h)W^0 \tag{6}$$

Finally, a linear combination of values is computed with the distribution output given by:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{7}$$

Where $W_i^Q \in \mathbb{R}^{d\times d_k}$, $W_i^K \in \mathbb{R}^{d\times d_k}$, $W_i^V \in \mathbb{R}^{d\times d_v}$ and $W^o \in \mathbb{R}^{hd_v\times d}$ are the parameter matrices defined over Q,K,V learned projections and d represents the model embedding dimension and i represents the ith block.

We adopt a second classifier, the multiheaded attention model, having the logic to capture intricate patterns and dependencies in the input data. Analysis shows that the attention model effectively improves prediction performance in unimodal and multimodal features compared with the result obtained in the baseline model. Although the attention classifier now has a tool to examine its decision-making process, the query-key-value mapping transformation is not enough to bring about much-expected differences in the predictive power in the multimodal and genomics features.

3.3.3 Experiment 3: Integrating Extended Squeeze-and-Excitation with Multiheaded Attention (ESEMA).

To address the challenges encountered in experiment 2, we extend the Multiheaded attention model by integrating it with a squeeze-and-excitation block. The input is concurrently processed with SE and MA layers. The output from the two layers is then added, renormalised and processed with the fully connected layer for prediction.

The Squeeze and Excitation (SE) block proposed by J. Hu et al. [75] factor in channel-wise relevance into the deep learning process. SE adaptively recalibrates channel-wise features by modelling interdependency between channels for quality feature representations. The SE block computes and assigns weights to the transformation input vector passed through the block. The block learns to assign greater weight to the most relevant features/channels. The metrics within the block network layers reprocess the feature maps using the learned global knowledge to identify informative features and suppress less relevant features. The spatial dimensions of the input features are compressed during the squeeze parts, and the channels are re-evaluated for relevance. The excitation part within the block decompresses the spatial component to the original dimensions of the input features. The resulting output from the block is robust feature representations of the transformational input features.

We adapt the squeeze-and-excitation method to the extracted fused features from genomics and histopathological images to rank the features in order of relevance. Given $\widetilde{X} \in \mathbb{R}^{H\times W\times C}$ is the latent feature. $\tilde{X}^{*} \in \mathbb{R}^{H\times W\times C}$ is recalibrated by squeezing global spatial information into its channel definition and a weighted relevance descriptor $\boldsymbol{z}$ is generated for $c-$ element (the c-th feature map) as shown in equation (8).

$$\boldsymbol{z}_c = G(\tilde{X}) = \frac{1}{H\times W}\sum_{m=1}^{H}\sum_{n=1}^{W}\tilde{X}_c(m,n) \tag{8}$$

We adopted a simple gating mechanism with a ReLU and a sigmoid function that is realised by forming a bottleneck with two fully-connected (FC) layers to obtain a weight vector $\mathbf{s}$:

$$\mathbf{s} = G_{ex}(z,\mathbf{W}) = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1 z)) \tag{9}$$

defined on z in terms of the activation function $\mathbf{s}$, follow by an excitation function, and the enhanced representational output is given as

$$\tilde{X}^{*}{}_c = G_{scale}(\tilde{X}_c, s_c) \tag{10}$$

Hence, the ESEMA model (F) jointly defined with components from squeeze-and-excitation block equation (10) estimated as E and multiheaded attention equation (7) estimated as M is given as:

$$F(G_{scale} \circ A) = G_{scale}\left(\tilde{X},\mathbf{s}\right) + A((QK^T)V) \tag{11}$$

3.4 Models Training, Validation, and Testing

For an effective training mechanism and robust generalisation fitting, the learning rate was set to (0.001), which is meant for effective controls in the step size during training and to prevent our model from slow or quick convergence. The momentum was set to 0.9, which assists the stochastic gradient descent to accelerate in the right direction and smooth the learning process. Also, weight decay was set to 0.0001, a regularisation term that penalises large weights

and prevents overfitting. The embedding dimension is 64, the batch size is 64, and the number of epochs is 100. The representation resulting from the MA and SE were subjected to layer normalisation and global average pooling. Subsequently, the processed outputs were fed into a fully connected layer for prediction purposes.

Scikit-learn stratified 5-fold cross-validation is applied when splitting data into train/validation/test sets. This helps to account for the imbalance in each class sample and preserve the ratio of samples for each class. Using stratified 5-fold cross-validation splitting, 80% of the dataset is split for model training/validation, while the remaining 20% is set aside for testing the model.

Valuable metrics are computed with AUC - Area under the ROC curve, F1-score, Precision and recall from the standard Scikit-learn library on the three predictive models, and necessary comparisons are made—the precision measure's ability of the model not to misclassify. The recall measures the fraction of instances in a class the model correctly classified from all class instances. AUC – Area under the ROC curve measures the two-dimensional area underneath the ROC curve, a measure of an aggregate performance across all possible classification thresholds, and F1-Score measure the test harmonic mean of the precision and recall estimates, where,

$$Precision = \frac{TP}{TP+FP} \tag{11}$$

$$Recall = \frac{TP}{TP+FN} \tag{12}$$

$$F1\_Score = \frac{2 \times Precision \times Recall}{Precision+Recall} \tag{13}$$

4 RESULTS

4.1 Analysing classifier's fitness with loss functions on the fused input.

We present an analysis of the loss function regarding the behaviour and performance of ESEMA and the other two models during the learning process on the fused input features. Fig. 3 (A-C) shows the plots for the measures of training and validation losses over epochs of the three classifiers for the fused features.

Fig. 3A shows the loss function over the number of iterations plotted for the baseline model. As shown in the plot, we observe that the training and validation loss functions behave differently. Our observation showed high spikes in the validation curve, which indicates that the model underfits the input data. This shows that the baseline framework cannot harness the non-linear signals in the multimodal data for effective prediction.

To resolve the challenges observed in the baseline model, we introduce an eight-headed Multiheaded attention framework. During the learning process, we observed that the validation function is moderately better than what is obtainable in the baseline framework. The spikes and loss are moderately low, and the model shows better convergence than what is observed in the baseline model (Fig. 3B). The challenges here are minimising the distance between the training and validation loss and ensuring smooth convergence in the learning process.

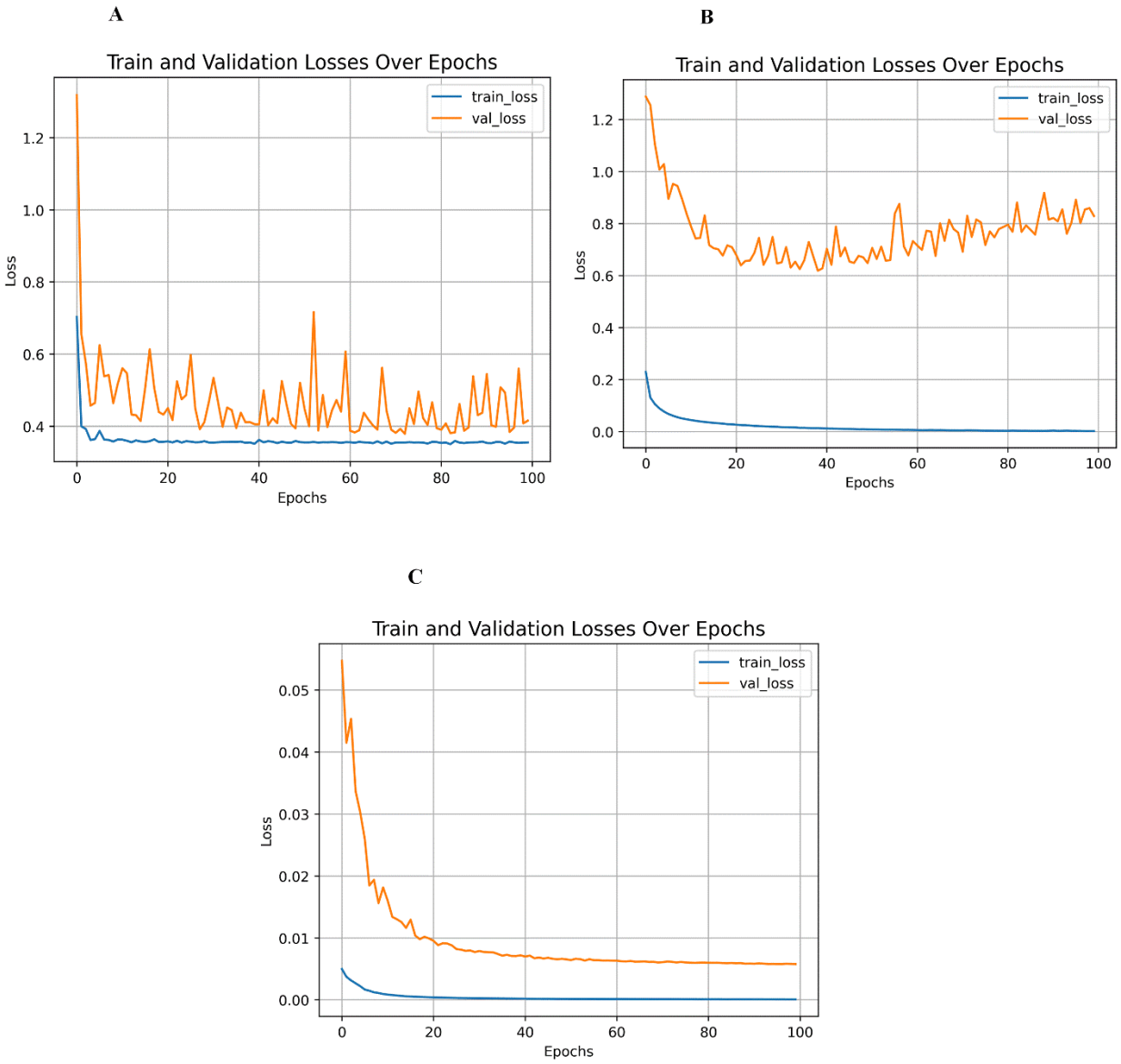

Fig. 3(A-C) shows training and validation loss curves for three models, all trained with the same fused multimodal input. (A) is the loss function for the baseline model, while (B) and (C) are the loss function curves for the MA and ESEMA models, respectively. MA: Multiheaded Attention, ESEMA: Extended Squeeze-and-Excitation with Multiheaded Attention

To resolve the problem posed by using the Multiheaded attention model, we introduce and extend Multiheaded attention with a modified version of the squeeze-and-excitation block. This measure achieves a practical learning experience. In Fig 3C, adding the modified squeeze-and-excitation block enhances the network's representation

capacity, and this is reflected in the loss function, which exhibits improved training stabilisation over time, converges effectively, and generalises well.

4.2 Classifiers performance accuracy with fused features

4.2.1 Evaluating the quality of the model predictions.

To validate the robustness of our predictive models, we calculated multiple metrics during the testing phase. Table 3 shows the robustness of the three models. ESEMA obtained the best predictive accuracy result in all categories of the datasets. The results of the integrated features show that the proposed ESEMA model harnesses the salient and complementary feature representation from the multimodal data, leading to improved model performance and predictive accuracy. Table 3 is the AUC result for the classifiers across the three categories of the input data. We can see that ESEMA gives the most improved predictive accuracy with a difference for the multimodal feature (AUC=0.9985 for fused) compared to the AUC result obtained with the unimodal feature (AUC= 0.9950 for genomics and AUC=0.8676 for images). This proof of concept signifies that multimodal features give the most improved predictive accuracy. Results from other performance metrics also support this justification. Models' performance results for precision, recall, and F1-score measures show the performance improvement when using the multimodal features over unimodal input features, as tabulated in Table 3.

Table 3. Model performance in unimodal and fused data.

Model	Data Type	AUC	Precision	Recall	F1_Score
Baseline	Images	0.7800	0.5291	0.5312	0.5240
	Genomics	0.9700	0.7996	0.9160	0.8401
	fused	0.9700	0.8000	0.9244	0.8431
Multiheaded Attention (MA)	Images	0.7906	0.5756	0.5512	0.5546
	Genomics	0.9932	0.9572	0.9510	0.9533
	Fused	0.9838	0.9423	0.9418	0.9417
Extended- Squeeze-and-Excitation Multiheaded Attention (ESEMA)	Images	0.8676	0.6607	0.6443	0.6485
	Genomics	0.9950	0.9593	0.9514	0.9542
	Fused	0.9985	0.9887	0.9864	0.9875

Table 4 shows the model performance and the predictive accuracy for the three classifiers by data types in the four cancer stages. Regarding model performance, our proposed model ESEMA shows the best performance across all the performance metrics. Also, by data type, the multimodal fused input dataset shows an improved predictive accuracy across all stages per model, and ESEMA shows where this difference is most significant compared with the other models.

Table 4. Predictive accuracy across stages per model in unimodal and fused data

Model	Data Type	Cancer Stages	AUC	Precision	Recall	F1_Score
Multiheaded Attention (MA)	Images	1	0.8396	0.5928	0.6315	0.6088
		2	0.7974	0.6491	0.7270	0.6847
		3	0.7626	0.5136	0.3807	0.4277
		4	0.7626	0.5467	0.4656	0.4973
	Genomics	1	0.9970	0.9898	0.9368	0.9619
		2	0.9938	0.9594	0.9841	0.9716
		3	0.9953	0.9503	0.9622	0.9558
		4	0.9867	0.9293	0.9207	0.9239
	Fused	1	0.9902	0.9534	0.9475	0.9503
		2	0.9834	0.9579	0.9586	0.9583
		3	0.9763	0.9153	0.9220	0.9184
		4	0.9852	0.9424	0.9389	0.9397
Extended Squeeze-and-Excitation Multiheaded Attention (ESEMA)	Images	1	0.9056	0.7490	0.7019	0.7212
		2	0.8749	0.7351	0.8048	0.7676
		3	0.8230	0.5681	0.4979	0.5274
		4	0.8667	0.5905	0.5724	0.5776
	Genomics	1	0.9942	0.9753	0.9501	0.9623
		2	0.9971	0.9535	0.9817	0.9672
		3	0.9936	0.9729	0.9291	0.9487
		4	0.9949	0.9354	0.9447	0.9385
	Fused	1	0.9995	0.9907	0.9931	0.9919
		2	0.9990	0.9913	0.9950	0.9931
		3	0.9983	0.9834	0.9834	0.9834
		4	0.9970	0.9895	0.9742	0.9817

Table 5. The number of network parameters for baseline, multi-head attention, and ESEMA concerns three sets of input datasets.

Network	Parameters (Fused) (M)	Parameters (Genomic) (K)	Parameters (Image) (M)
Baseline	2.15	1.10	2.15
Multiheaded Attention (MA)	17.21	416.29	16.81
ESEMA	17.95	416.69	17.50

The suggested ESEMA model, which combines Multiheaded Attention and Squeeze-and-Excitation processes, has more trainable parameters (17.95 million) than either the baseline model (2.15 million) or the solo Multiheaded Attention network (17.21 million), as shown in Table 5. This implies that the ESEMA model will need additional computer resources for training and inference. To improve computational efficiency for a real-world deployment, systems with GPU support would be advantageous due to their capacity to accelerate calculations concurrently. Furthermore, methods like quantisation and pruning might be used on the ESEMA model to minimise computational complexity while maintaining performance. Quantisation reduces the precision of weights and activations, whereas pruning eliminates unneeded connections or parameters. Implementing these strategies reduces the computing needs of the ESEMA model, making it more suitable for deployment in resource-constrained contexts.

Table 6. Performance comparison of ESEMA with the state-of-the-art model.

Model	AUC
Arya et al.[110]	0.9300
Olatunji et al.[111]	0.8100
Guo et al. [112]	0.9710
Zhang et al. [113]	0.9322
Proposed ESEMA	0.**9985**

The proposed Extended Squeeze-and-Excitation Multiheaded Attention (ESEMA) model outperforms all other strategies, with an AUC of 99.85%. This indicates ESEMA's superior prediction capacity for accurately detecting cancer stages compared to earlier approaches. As shown in Table 6, while Arya et al. and Zhang et al. perform well with 0.9300 and 0.9322 AUC, respectively, ESEMA exceeds them, as do Guo et al., who achieve a commendable 0.9710 AUC. ESEMA's high accuracy demonstrates the utility of its improved attention mechanism and design in gathering complex patterns in data, making it a good option for prediction decisions.

4.2.2 Analysing the Confusion-Matrix (CM).

The confusion matrix is an essential predictive model with the functionality to give insight into the model that gets confused and shows an overall model performance. Fig.4 shows the confusion matrix computed from our proposed ESEMA classifier on the three categories of input features. The diagonal of the matrix gives the number of cases accurately predicted for each class. Table 7 compares the number of accurately predicted cases in MA vs ESEMA. Results show that ESEMA has 6.5% improved accuracy over MA on integrated features, 2.61% improved predictive accuracy with genomics-only features, and 11.57% improved predictive accuracy with image-only features.

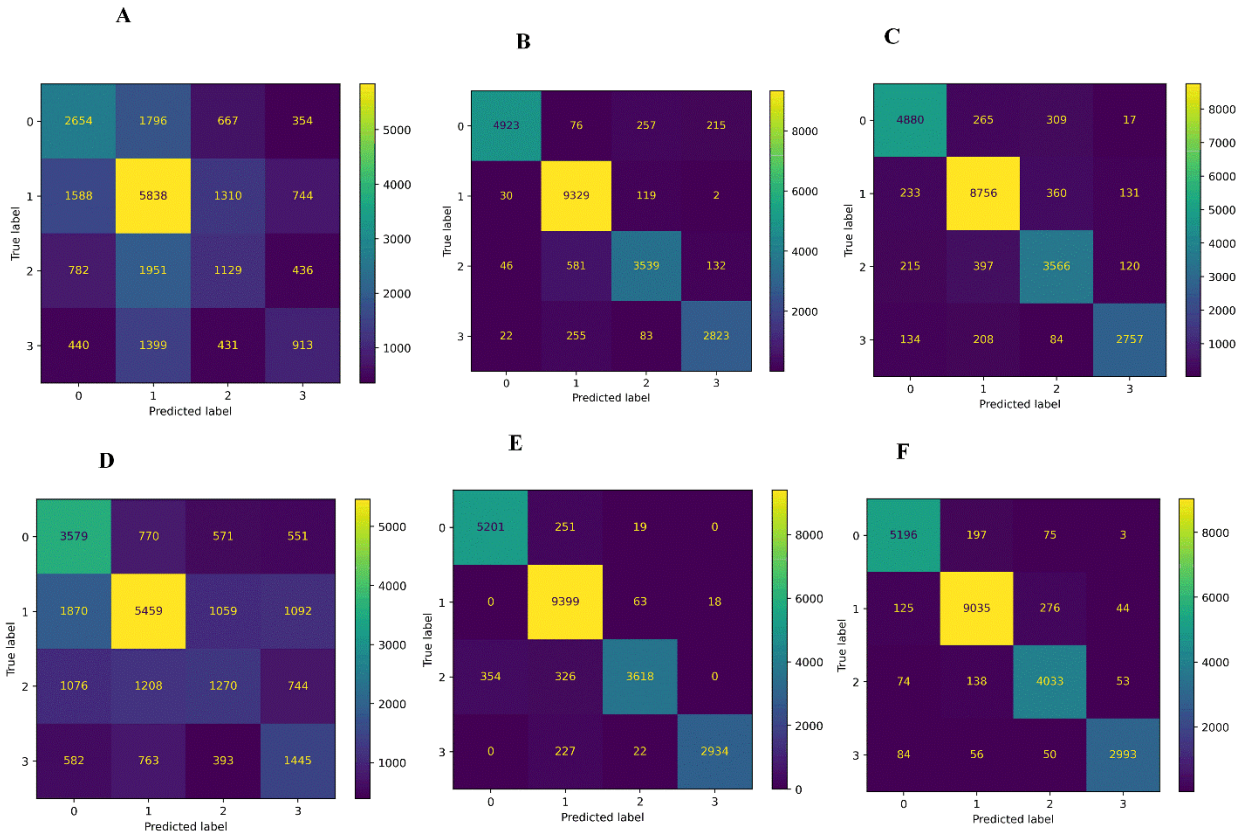

Fig. 4 Confusion matrix (CM) for Multiheaded and ESEMA classifiers. (A)-(C) are CMs for Image, genomics and fused features with the MA model, respectively, and (D)-(F) are CMs for the three feature sets with the ESEMA model

Table 7. Confusion Matrix summary regarding the number of samples accurately predicted by the two classifiers (MA vs ESEMA) for three different inputs.

	MA (Cases Correctly Predicted)			ESEMA (Cases Correctly Predicted)		
Stages	**Images**	**Genomics**	**Fused**	**Images**	**Genomics**	**Fused**
1	2654	4923	4880	3579	5201	5196
2	5838	9329	8756	5459	9399	9035
3	1129	3539	3566	1270	3618	4033
4	913	2823	2757	1445	2934	2993

4.2.3 Explainable - Investigating the model attention heads representations.

In this study, we approach explainability using two different computational utilities – mean attention distance and heatmap visualisation metrics [109,114,115]. The attention score is calculated as a dot product between key-query vectors of the projected input. Fig. 5 shows the graph distribution of attention score over the attention heads in the block generated during the forward pass of the input in inference mode through the ESEMA model. We can see the

spread of the attention heads within the transformer block over the input features in the plot. Also, in Fig.5, we can see attention heads 0, 1, and 7 attending to local abstraction activations and attention heads 3 and 4 attending to global abstraction activations.

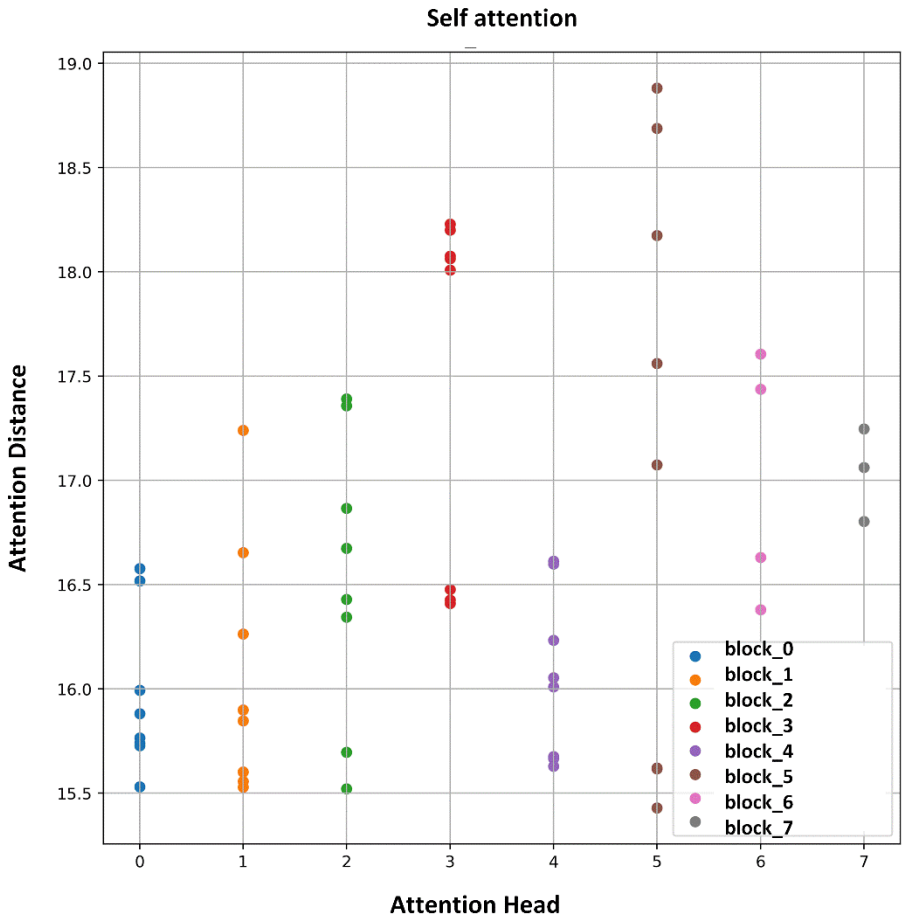

Fig. 5. Attention heads with global and local learning of salient features for cancer stage prediction.

4.2.4 Attention heatmaps – cancer stages activation heatmap with fused features.

We examine and highlight what the classifier attends to by superimposing the attention maps over the fused input features space. We randomly chose three blocks from a total of eight blocks (each block has eight attention heads) in our classifier to generate the attention heat maps and analyse insightful regions. Fig. 6 shows the attention heatmap from randomly selected blocks 1, 5, and 8. We observed different and insightful heatmap activations on each of the attention heads within each block. We can see that the region of interest being attended to within the attention heads differs across each block. The activation within block 1 is global, while block 5 comprises a mixture of local and global attention; also, block 8, which is the innermost block, is more of local attention. These are explainable and interpretable tools that medical experts can use to improve precision cancer treatment design and management.

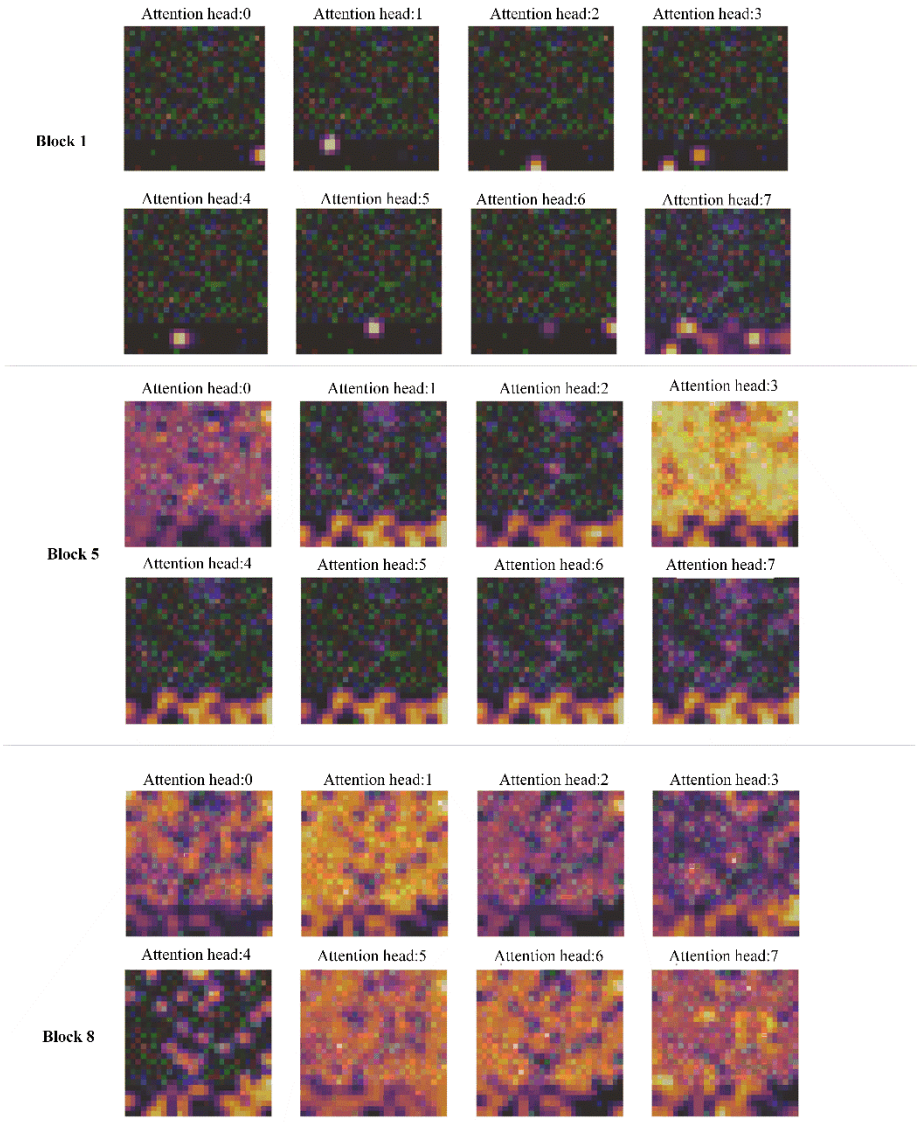

Fig. 6. Attention heatmap showing activation regions across eight attentions in three different transformer blocks (Block 1, 5, and 8), the core mechanism and process in learning contextual dependencies within the feature space. For visualisation purposes, the attention heatmap is obtained using 2048 image features and 652 genomic features.

4.2.5 Feature Relevance

Since cancer is highly heterogeneous and given that two cancer samples in the same stage can be driven by different biomarkers, we implemented a feature relevance adaptive function and adaptively attached relevance to features within the input that might be useful to experts for further therapeutic discovery design, prognosis management, or for other survival risk stratification.

As shown in Fig. 7, we explore the relevance of each feature in the input space. We compute attention weight over each feature, which signifies the contribution of each feature in the input to determine the cancer stage of a sample patient. This feature relevance computation was made possible using the state-of-the-art squeeze-and-excitation block

[75]. This represents another explainable and interpretable tool that medical experts could use to explore how the model arrives at its decision. The identified relevant features could assist the colon cancer pathologist/specialist in building trust in the neural network model, obtain insights for further laboratory testing and screening, and decide the stage for borderline cases.

The graphs in Fig. 7 indicate how important each extracted feature is in predicting the cancer stages. The extracted features close to the origin are the most relevant, while those far off can be said to be less relevant. Different features are attributed to the model for its decision-making and provide an informative direction for further analysis by the expert. The relevance of the genomics extracted features in arriving at sample cancer stages are identified by the orange lines within each graph, Fig. 7 (A-D)

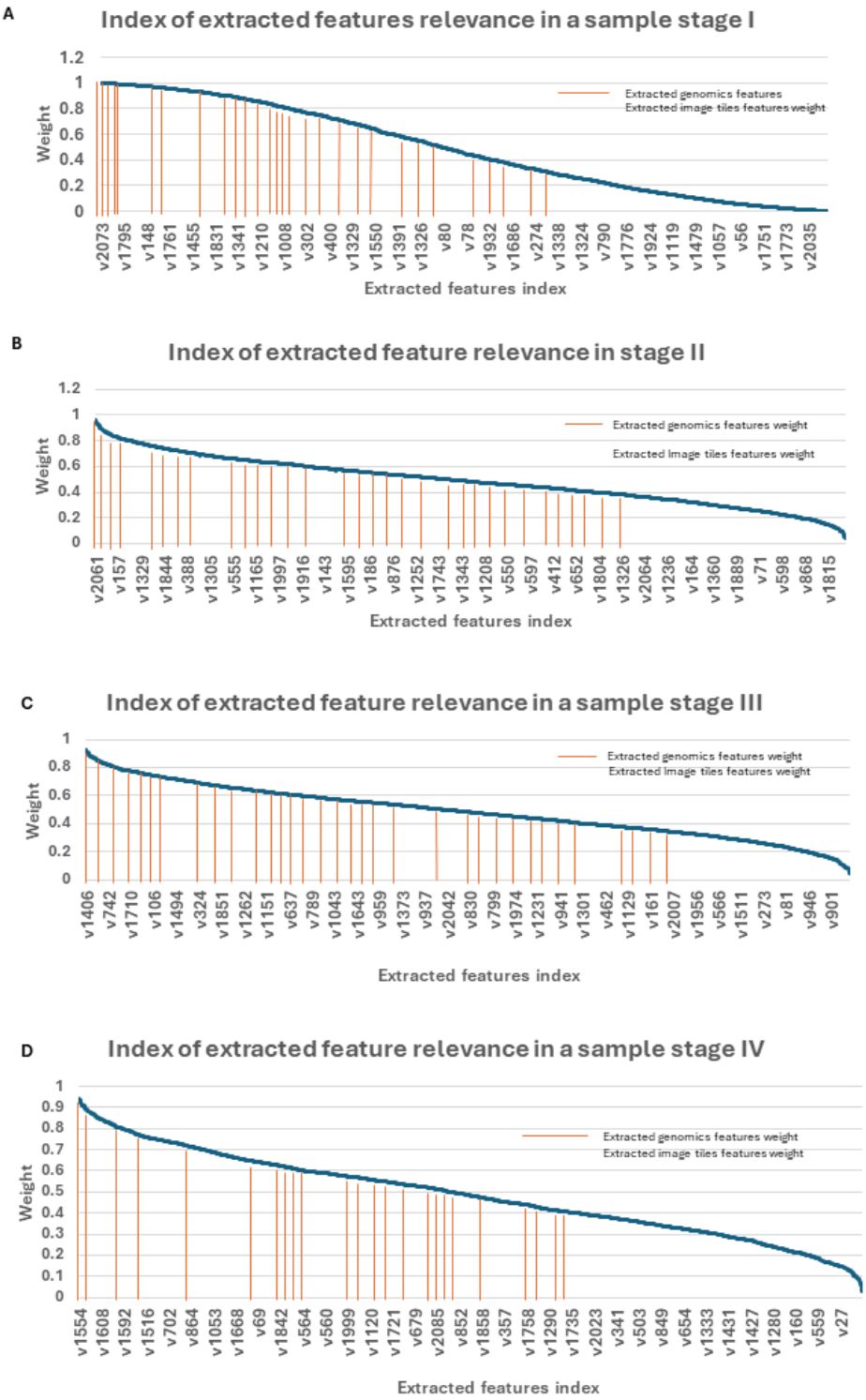

Fig.7(A-D) shows attention weights for input features predicting cancer stages. (A) shows the order of features considered relevant for predicting stage I in colon cancer. (B) – (D) are relevant features for predicting stages II, III and IV cancers, respectively. The orange lines represent the genomics features extracted in order of relevance.

As shown across the four graphs, different features are identified as relevant in predicting a typical cancer stage. This is in accordance with the heterogeneousness and complexity of cancer's diseases. Across the four cancer stages, the top important features are from the genomics representational features, as indicated by orange vertical lines in Fig. 7.

4.2.6 Mapping the genomic reconstructed space into the original genomics features space.

Given that genomics characteristics are frequently reported as the most relevant features in predicting cancer stages, we rebuild a feature space with a comparable dimension to that of the input genomic space from the latent feature space. Next, we calculated the Square Error (SE) between the functional genomic space and the decoded reconstructed space features' values, as described in section 3.2.2. The result of the intersection operation on the set of features returned by the three masking settings is shown in Table 8. The top 10 most relevant genes in the four stages are first highlighted in Fig 7. Furthermore, Fig. 8(A-D) shows the snapshot of the intersection of the reconstructed genes (the top 10 genetic values with the least reconstruction error)

Some of the genes shown in Table 8 have previously been associated with colon cancer, whilst some others are novel biomarkers with no prior association with colon cancer. Previous research has demonstrated a link between some genetic characteristics and colon and other cancer prognostic or diagnostic outcomes.

Table 8. The top 10 most relevant genes in different stages are estimated by our proposed model.

No	Stage 1	Stage 2	Stage 3	Stage 4
1	BEX1	ACTL8	BEST4	AMPH
2	CACNA1I	AMPH	C8G	ASTE1
3	CD14	AQP8	CA7	CACNA1E
4	COL2A1	ASTE1	CCDC141	CACNB1
5	DBH	BAAT	CLCA2	CCDC141
6	DHRS2	BEST4	DEFA5	CCDC181
7	DIRAS3	CACNA1E	DEFA6	CHGB
8	DNAH11	CNGA3	DHRS2	CHRNB2
9	EPHB6	CYP3A4	DIRAS3	CLCA2
10	GLYATL2	ELOVL2	FMNL3	CNGA3

David et al.[116] linked ASTE1 to tumour-infiltrating lymphocytes in colorectal malignancies. Rachael et al.[117] found that CACNA1E amplification and overexpression are associated with relapse in favourable histology of Wilms' tumours. Xiao-Shun et al.[118] found that BEST4 expression is elevated in clinical CRC samples and that high levels of expression are associated with advanced TNM (tumour, lymph nodes, distant metastasis) stage, LNM (lymph node metastasis), and poor survival. XIAOHANG et al.[119] showed that CLCA accessory mRNA expression has predictive significance in colon cancer. Guang-Zhen et al.[120] established a relationship between CA7 and expression in colorectal carcinoma, and CYP3A4 was identified as an anti-colorectal cancer (CRC) target of formononetin (FN) by Zhang et al. [121] Furthermore, Qinglin et al. [122] connected FMNL3 to overexpression in tumour tissues and

predicted an immuno-hot phenotype in pancreatic cancer. Alexander et al. [123] report that COL2A1 has reoccurring alterations in chondrosarcoma. Hao Yu et al.[124] found that EphB6 deficiency in intestinal neurons enhances tumour formation in colorectal cancer.

Table 8 shows the insights gained from this investigation, including:

- Certain Stage-Specific Gene Expression: Some genes show stage-specific expression patterns, indicating their possible roles in cancer development and progression phases. For example, BEX1, CACNA1I, CD14, COL2A1, and DBH are only expressed at certain phases, implying that they are involved in the related biological processes or pathways. Previous research connected BEX1, CACNA1I, and DBH genes to expression in various cancers, such as the brain, neurodevelopmental disorders, and lungs, with only CD14 correlating with poor clinical outcomes in CRC. The CD14 molecule shows promise as a target for colorectal cancer immunotherapy.
- Gene Co-Expression: The table shows DIRAS3, AMPH, ASTE1, CACNA1E, and CACNB1 as co-expressed genes, which occurs when many genes are expressed simultaneously at the same stage. This co-expression might imply functional links or common regulatory mechanisms among these genes. Bildik et al. [125] found that DIRAS3 is downregulated in malignancies of the ovary, breast, lung, prostate, colon, brain, and thyroid. AMPH (Amphiphysin) - According to Protein Atlas, AMPH is a gene associated with glioma and is used as a predictive biomarker in pancreatic cancer, but it has no known association with colon cancer. A prior study by Zhou et al. [126] found that higher CACNB1 (Calcium Voltage-Gated Channel Auxiliary Subunit Beta 1) expression in CRC is related to poor patient survival, although their biological activities in CRC are unclear. While ASTE1 (Asteroid Homolog 1) is connected with lymphocytes in colorectal cancers, amplification and overexpression of CACNA1E are associated with recurrence in Wilms' tumours with good histological characteristics. Understanding these co-expression patterns can give insight into the molecular pathways that drive cancer growth.
- Potential Biomarkers: Genes with constant expression throughout many stages, such as DIRAS3, AMPH, ASTE1, CACNA1E, and CACNB1, may be used as biomarkers for cancer progression. Their consistent expression patterns indicate their importance in cancer biology and possible use as diagnostic or prognostic markers.

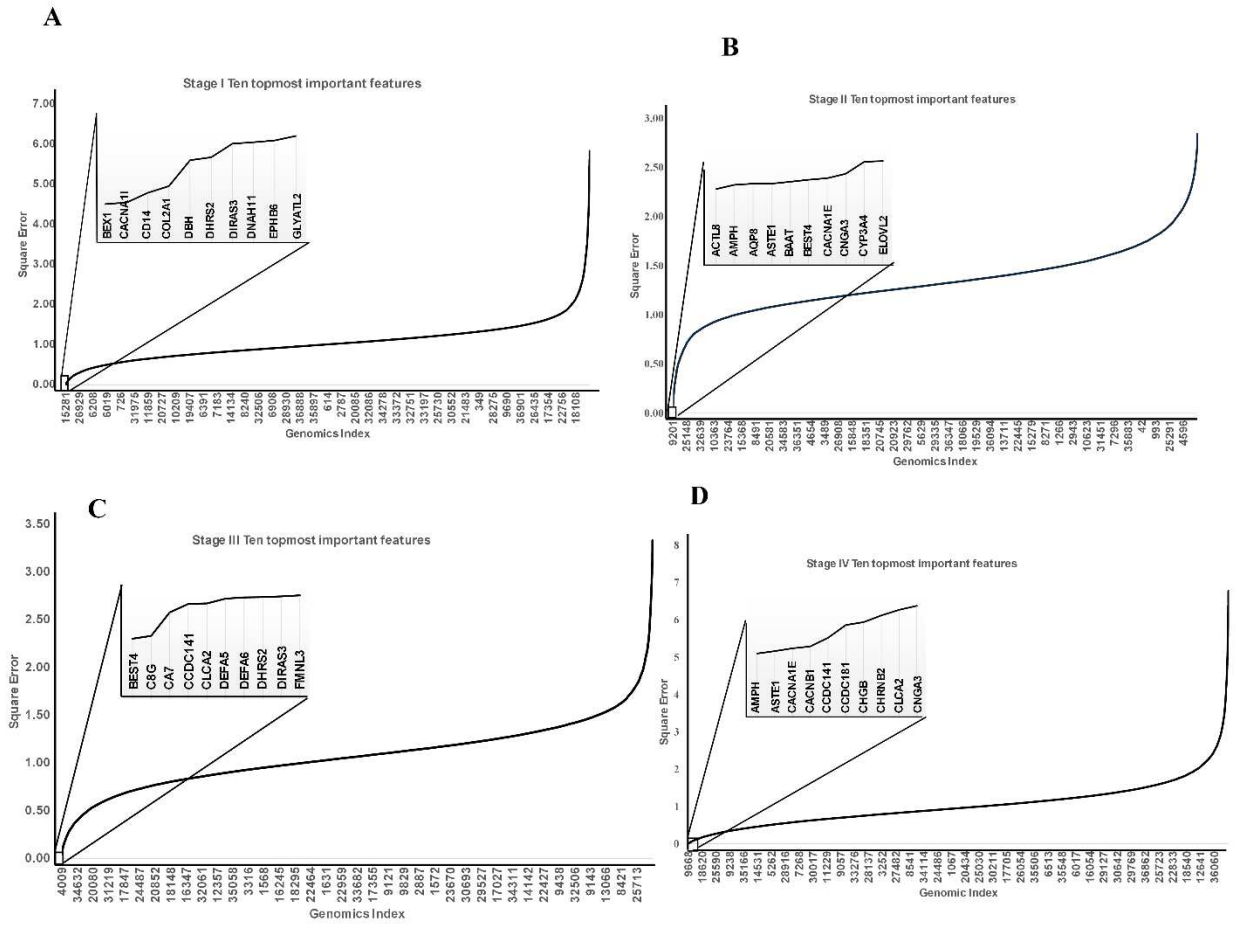

Fig. 8 Snapshot of the intersection of the reconstructed genes (the top 10 relevant gene features that have the least reconstruction error) for (A) Stage I, (B) Stage II, (C) Stage III, and (D) Stage IV.

5. DISCUSSION

Our proposed model achieved improved performance in colon cancer stage prediction based on multimodal fused features, justifying our proof-of-concept that multimodal input features perform better compared with unimodal input. Also, explainable and interpretation tools are used to clarify how the classifier arrived at its decision to predict the cancer stages.

This study examines the application of a deep neural network in predicting colon cancer stages, utilising multiple data sources. The focus extends beyond mere accuracy to include an exploration of candidate genes relevant for cancer stage classification through the proposed masked reconstruction autoencoder approach. This dual strategy seeks to not only enhance colon cancer detection but also to discover novel biomarkers for future research. Our proposed methods can be applied to study other types of cancers. Though the introduction of the attention mechanism resulted in improved prediction and stratification accuracy, some shortcomings were still observed. The first identified shortcoming to be considered in our future research is the need to include more diverse datasets such as somatic

mutation (SM), Copy number variation (CNV), Reverse Phase Protein Array (RPPA), Insertion – Deletions (InDels) of bases in genes, Structural Variants (SVs), Single-Nucleotide Variants (SNVs), metabolome and proteome to obtain more salient complementary features within the multimodal datasets integration for improved predictive performance. The future work should consider the development of novel algorithms to investigate the interactions between features across multiple datasets, using graph-like deep neural network

When deployed in a clinical real-application environment, the two major challenges of this combined architecture are principally related to high model parameterisation and memory requirement management. The computational efficiency of the model can be optimised through knowledge distillation [127], the use of model quantisation [128–130] or layer fusion [131].

Similar to this, the current limitations in evaluating the effectiveness of the combined architecture are limited samples that existed in the current open-access datasets for the model training, the lack of external datasets for effective model validation, and imbalanced classes across different cancer stages. Evaluation methods can be improved through data generation using generative models. The most relevant features in the study are extracted features from genomics datasets. In our future study, we hope to adopt the feature selection method and directly identify expressed genes.

6. CONCLUSION

Examining model performance reveals diverse patterns across different architectures and data types. Starting with the basic model, our model performs moderately overall, with genomic data regularly beating image data and fused data improving marginally over separate modalities. Moving to the Multiheaded Attention (MA) paradigm improves performance significantly across all data categories, with genetic data continuing to dominate the relevance. The Extended Squeeze-and-Excitation Multiheaded Attention (ESEMA) model outperforms MA, demonstrating the effectiveness of its improved attention mechanism. Again, genomic data is the most predictive, with fused data constantly outperforming other modalities.

The findings underline the need to combine multimodal data with enhanced attention processes to attain the best-predicted performance in cancer staging tasks.

References

1. Wulczyn E, Steiner DF, Moran M, Plass M, Reihs R, Tan F, et al. Interpretable survival prediction for colorectal cancer using deep learning. NPJ Digit Med [Internet]. 2021;4(1):1–13. Available from: http://dx.doi.org/10.1038/s41746-021-00427-2

2. Vale-Silva LA, Rohr K. Long-term cancer survival prediction using multimodal deep learning. Sci Rep [Internet]. 2021;11(1):1–12. Available from: https://doi.org/10.1038/s41598-021-92799-4

3. Liu P, Li L, Yu C, Fei S. Two Staged Prediction of Gastric Cancer Patient's Survival Via Machine Learning Techniques. 2020;105–16.

4. Gupta S, Kalaivani S, Rajasundaram A, Ameta GK, Oleiwi AK, Dugbakie BN. Prediction Performance of Deep Learning for Colon Cancer Survival Prediction on SEER Data. Biomed Res Int. 2022;2022.

5. Tan K, Huang W, Liu X, Hu J, Dong S. A multi-modal fusion framework based on multi-task correlation learning for cancer prognosis prediction. Artif Intell Med [Internet]. 2022;126(January):102260. Available from: https://doi.org/10.1016/j.artmed.2022.102260

6. Cheng J, Zhang J, Han Y, Wang X, Ye X, Meng Y, et al. Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis. Cancer Res. 2017;77(21):e91–100.

7. He J, Wu F, Han Z, Hu M, Lin W, Li Y. Biomarkers (mRNAs and Non-Coding RNAs) for the Diagnosis and Prognosis of Colorectal Cancer – From the Body Fluid to Tissue Level. 2021;11(April):1–14.

8. Franco EF, Rana P, Cruz A, Calderón V V., Azevedo V, Ramos RTJ, et al. Performance comparison of deep learning autoencoders for cancer subtype detection using multi-omics data. Vol. 13, Cancers. 2021.

9. Haslam A, Kim MS, Prasad V. Overall survival for oncology drugs approved for genomic indications. Eur J Cancer [Internet]. 2022;160:175–9. Available from: https://doi.org/10.1016/j.ejca.2021.10.028

10. Sun D, Li A, Tang B, Wang M. Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. Comput Methods Programs Biomed. 2018;161:45–53.

11. Id LL, Id QM, Weng C, Id QL, Wang T, Id YW. PLOS COMPUTATIONAL BIOLOGY Explainable deep transfer learning model for disease risk prediction using high- dimensional genomic data. 2022;1–23. Available from: http://dx.doi.org/10.1371/journal.pcbi.1010328

12. Schneider L, Laiouar-Pedari S, Kuntz S, Krieghoff-Henning E, Hekler A, Kather JN, et al. Integration of deep learning-based image analysis and genomic data in cancer pathology: A systematic review. Eur J Cancer. 2022;160:80–91.

13. Qiao Y, Zhao L, Luo C, Luo Y, Wu Y, Li S, et al. Multi-modality artificial intelligence in digital pathology. Brief Bioinform. 2022;23(6):1–17.

14. Biomarker C. BBA - Reviews on Cancer DNA methylation-based diagnostic , prognostic , and predictive biomarkers in colorectal cancer. 2022;1877(February).

15. Lima LJG, Morais AHF, Valentim RAM, Barreto EJSS. miRNAs as biomarkers for early cancer detection and their application in the development of new diagnostic tools. Biomed Eng Online. 2021;1–20.

16. Zitt M, Zitt M. DNA methylation in colorectal cancer – Impact on screening and therapy monitoring modalities ? 2007;23:51–71.

17. Fan J, Li J, Guo S, Tao C, Zhang H, Wang W, et al. Genome-wide DNA methylation profiles of low- and high-grade adenoma reveals potential biomarkers for early detection of colorectal carcinoma. 2020;1–13.

18. Xue VW, Cheung MT, Chan PT, Luk LLY, Lee VH, Au TC, et al. Non-invasive Potential Circulating mRNA Markers for Colorectal Adenoma Using Targeted Sequencing. Sci Rep. 2019;(February):1–10.

19. He J, Wu F, Han Z, Hu M, Lin W, Li Y. Biomarkers (mRNAs and Non-Coding RNAs) for the Diagnosis and Prognosis of Colorectal Cancer – From the Body Fluid to Tissue Level. 2021;11(April):1–14.

20. Liu C, Papukashvili D, Dong Y, Wang X, Hu X, Yang N, et al. Identification of Tumor Antigens and Design of mRNA Vaccine for Colorectal Cancer Based on the Immune Subtype. 2022;9(January):1–12.

21. Kwak MS, Lee HH, Yang JM, Cha JM, Jeon JW, Yoon JY, et al. Deep Convolutional Neural Network-Based Lymph Node Metastasis Prediction for Colon Cancer Using Histopathological Images. Front Oncol. 2021;10(January):1–9.

22. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions [Internet]. Vol. 8, Journal of Big Data. Springer International Publishing; 2021. Available from: https://doi.org/10.1186/s40537-021-00444-8

23. D'Amour A, Heller K, Moldovan D, Adlam B, Alipanahi B, Beutel A, et al. Underspecification Presents Challenges for Credibility in Modern Machine Learning. 2020; Available from: http://arxiv.org/abs/2011.03395

24. van der Velden BHM. Explainable AI: current status and future potential. Eur Radiol [Internet]. 2023;4–6. Available from: https://doi.org/10.1007/s00330-023-10121-4

25. Xie N, Ras G, van Gerven M, Doran D. Explainable Deep Learning: A Field Guide for the Uninitiated. 2020; Available from: http://arxiv.org/abs/2004.14545

26. Larasati R. Explainable AI for Breast Cancer Diagnosis: Application and User's Understandability Perception. International Conference on Electrical, Computer, and Energy Technologies, ICECET 2022. 2022;(June):1–6.

27. Massafra R, Fanizzi A, Amoroso N, Bove S, Comes MC, Pomarico D, et al. Analyzing breast cancer invasive disease event classification through explainable artificial intelligence. Front Med (Lausanne). 2023;10.

28. Tan J, Ung M, Cheng C, Greene CS. Extraction From Genome-Wide Assays of Breast Cancer. Pac Symp Biocomput [Internet]. 2015;132–43. Available from: http://europepmc.org/backend/ptpmcrender.fcgi?accid=PMC4299935&blobtype=pdf

29. Fan J, Li J, Guo S, Tao C, Zhang H, Wang W, et al. Genome-wide DNA methylation profiles of low- and high-grade adenoma reveals potential biomarkers for early detection of colorectal carcinoma. 2020;1–13.

30. Amoroso N, Quarto S, La Rocca M, Tangaro S, Monaco A, Bellotti R. An eXplainability Artificial Intelligence approach to brain connectivity in Alzheimer's disease. Front Aging Neurosci. 2023;15.

31. Phillips PJ, Hahn CA, Fontana PC, Broniatowski DA, Przybocki MA, Hahn CA, et al. Four Principles of Explainable Artificial Intelligence: Draft NISTIR 8312. National Institute of Standards and Technology Interagency or Internal Report [Internet]. 2020;(August). Available from: https://doi.org/10.6028/NIST.IR.8312-draft

32. Abdelhafiz D, Yang C, Ammar R, Nabavi S. Deep convolutional neural networks for mammography: Advances, challenges and applications. Vol. 20, BMC Bioinformatics. 2019.

33. Kitaguchi D, Takeshita N, Matsuzaki H, Igaki T, Hasegawa H, Ito M. Development and Validation of a 3-Dimensional Convolutional Neural Network for Automatic Surgical Skill Assessment Based on Spatiotemporal Video Analysis. JAMA Netw Open. 2021;4(8):1–10.

34. Schmauch B, Romagnoni A, Pronier E, Saillard C, Maillé P, Calderaro J, et al. expression of tumours from whole slide images. Nat Commun [Internet]. (2020):1–15. Available from: http://dx.doi.org/10.1038/s41467-020-17678-4

35. Ertosun MG, Rubin DL. Automated Grading of Gliomas using Deep Learning in Digital Pathology Images: A modular approach with ensemble of convolutional neural networks. AMIA Annu Symp Proc. 2015;2015:1899–908.

36. Sena P, Fioresi R, Faglioni F, Losi L, Faglioni G, Roncucci L. Deep learning techniques for detecting preneoplastic and neoplastic lesions in human colorectal histological images. Oncol Lett. 2019;18(6):6101–7.

37. Zitt M, Zitt M. DNA methylation in colorectal cancer – Impact on screening and therapy monitoring modalities ? 2007;23:51–71.

38. Bach S, Paulis I, Sluiter NR, Tibbesma M, Martin I, Wiel MA Van De. Detection of colorectal cancer in urine using DNA methylation analysis. Sci Rep [Internet]. 2021;1–11. Available from: https://doi.org/10.1038/s41598-021-81900-6

39. Xu W, Xu M, Wang L, Zhou W, Xiang R, Shi Y, et al. Integrative analysis of DNA methylation and gene expression identified cervical cancer-specific diagnostic biomarkers. Signal Transduct Target Ther [Internet]. 2019;4(1). Available from: http://dx.doi.org/10.1038/s41392-019-0081-6

40. Learning D. Prediction of Long Non-Coding RNAs Based on. 2019;

41. Liu C, Papukashvili D, Dong Y, Wang X, Hu X, Yang N, et al. Identi fi cation of Tumor Antigens and Design of mRNA Vaccine for Colorectal Cancer Based on the Immune Subtype. 2022;9(January):1–12.

42. Xue VW, Cheung MT, Chan PT, Luk LLY, Lee VH, Au TC, et al. Non-invasive Potential Circulating mRNA Markers for Colorectal Adenoma Using Targeted Sequencing. Sci Rep [Internet]. 2019;(February):1–10. Available from: http://dx.doi.org/10.1038/s41598-019-49445-x

43. Herring E, Tremblay É, Mcfadden N, Kanaoka S, Beaulieu J françois. Multitarget Stool mRNA Test for Detecting Colorectal Cancer Lesions Including Advanced Adenomas. 2021;1–12.

44. Feng L, Liu Z, Li C, Li Z, Lou X, Shao L, et al. Development and validation of a radiopathomics model to predict pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: a multicentre observational study. Lancet Digit Health. 2022;4(1):e8–17.

45. MAZAKI J, KATSUMATA K, OHNO Y, UDO R, TAGO T, KASAHARA K, et al. A novel predictive model for anastomotic leakage in colorectal cancer using auto-artificial intelligence. Anticancer Res. 2021;41(11):5821–5.

46. Kitaguchi D, Takeshita N, Matsuzaki H, Igaki T, Hasegawa H, Ito M. Development and Validation of a 3-Dimensional Convolutional Neural Network for Automatic Surgical Skill Assessment Based on Spatiotemporal Video Analysis. JAMA Netw Open. 2021;4(8):1–10.

47. Masum S, Hopgood A, Stefan S, Flashman K, Khan J. Data analytics and artificial intelligence in predicting length of stay, readmission, and mortality: a population-based study of surgical management of colorectal cancer. Discover Oncology. 2022;13(1).

48. Abraham JP, Magee D, Cremolini C, Antoniotti C, Halbert DD, Xiu J, et al. Clinical Validation of a Machine-learning–derived Signature Predictive of Outcomes from First-line Oxaliplatin-based Chemotherapy in Advanced Colorectal Cancer. Clinical Cancer Research. 2021;27(4):1174–83.

49. Bilal M, Raza SEA, Azam A, Graham S, Ilyas M, Cree IA, et al. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. Lancet Digit Health. 2021;3(12):e763–72.

50. Kang Y, Kim YJ, Park S, Ro G, Hong C, Jang H, et al. Development and operation of a digital platform for sharing pathology image data. BMC Med Inform Decis Mak. 2021;21(1):1–8.

51. Misumi Y, Nonaka K, Takeuchi M, Kamitani Y, Uechi Y, Watanabe M, et al. Comparison of the Ability of Artificial-Intelligence-Based Computer-Aided Detection (CAD) Systems and Endoscopists to Detect Colorectal Neoplastic Lesions on Endoscopy Video. J Clin Med. 2023;12(14).

52. Wei JW, Suriawinata AA, Vaickus LJ, Ren B, Liu X, Lisovsky M, et al. Evaluation of a Deep Neural Network for Automated Classification of Colorectal Polyps on Histopathologic Slides. JAMA Netw Open. 2020;3(4):1–11.

53. Theodosi A, Ouzounis S, Kostopoulos S, Glotsos D, Kalatzis I, Tzelepi V, et al. Design of a hybrid deep learning system for discriminating between low- and high-grade colorectal cancer lesions, using microscopy images of IHC stained for AIB1 expression biopsy material. Mach Vis Appl [Internet]. 2021;32(3):1–17. Available from: https://doi.org/10.1007/s00138-021-01184-8

54. Filipów S. Blood Circulating miRNAs as Cancer Biomarkers for Diagnosis and Surgical Treatment Response. 2019;10(March):1–7.

55. Aikemu B, Xue P, Hong H, Jia H, Wang C, Li S, et al. Artificial Intelligence in Decision-Making for Colorectal Cancer Treatment Strategy: An Observational Study of Implementing Watson for Oncology in a 250-Case Cohort. Front Oncol. 2021;10(February):1–8.

56. Huo Y, Guo Y, Wang J, Xue H, Feng Y, Chen W, et al. Integrating multi-modal information to detect spatial domains of spatial transcriptomics by graph attention network. Journal of Genetics and Genomics [Internet]. 2023;50(9):720–33. Available from: https://doi.org/10.1016/j.jgg.2023.06.005

57. Wang C, Lye X, Kaalia R, Kumar P, Rajapakse JC. Deep learning and multi-omics approach to predict drug responses in cancer. BMC Bioinformatics [Internet]. 2021;22:1–15. Available from: https://doi.org/10.1186/s12859-022-04964-9

58. Pang J, Liang B, Ding R, Yan Q, Chen R, Xu J. A denoised multi-omics integration framework for cancer subtype classification and survival prediction. Brief Bioinform. 2023;24(5):1–12.

59. Franco EF, Rana P, Cruz A, Calderón V V., Azevedo V, Ramos RTJ, et al. Performance comparison of deep learning autoencoders for cancer subtype detection using multi-omics data. Vol. 13, Cancers. 2021.

60. Gong P, Cheng L, Zhang Z, Meng A, Li E, Chen J, et al. Multi-omics integration method based on attention deep learning network for biomedical data classification. Comput Methods Programs Biomed [Internet]. 2023;231:107377. Available from: https://doi.org/10.1016/j.cmpb.2023.107377

61. Chaudhary K, Poirion OB, Lu L, Garmire LX. Survival in Liver Cancer. 2019;24(6):1248–59.

62. Jiang A, Liu N, Zhao R, Liu S, Gao H, Wang J, et al. Construction and Validation of a Novel Nomogram to Predict the Overall Survival of Patients With Combined Small Cell Lung Cancer: A Surveillance, Epidemiology, and End Results Population-Based Study. Cancer Control. 2021;28:1–12.

63. Aikemu B, Xue P, Hong H, Jia H, Wang C, Li S, et al. Artificial Intelligence in Decision-Making for Colorectal Cancer Treatment Strategy: An Observational Study of Implementing Watson for Oncology in a 250-Case Cohort. Front Oncol. 2021;10(February):1–8.

64. Wu QY, Liu SL, Sun P, Li Y, Liu GW, Liu SS, et al. Establishment and clinical application value of an automatic diagnosis platform for rectal cancer T-staging based on a deep neural network. Chin Med J (Engl). 2021;134(7):821–8.

65. DeSilvio T, Antunes JT, Bera K, Chirra P, Le H, Liska D, et al. Region-specific deep learning models for accurate segmentation of rectal structures on post-chemoradiation T2w MRI: a multi-institutional, multi-reader study. Front Med (Lausanne). 2023;10(May):1–11.

66. Wang H, Dai C, Wen Y, Wang X, Liu W, He S, et al. GADRP: graph convolutional networks and autoencoders for cancer drug response prediction. Brief Bioinform. 2023;24(1):1–12.

67. Quraish R ul, Hirahata T, Quraish A ul, ul Quraish S. An Overview: Genetic Tumor Markers for Early Detection and Current Gene Therapy Strategies. Vol. 22, Cancer Informatics. SAGE Publications Ltd; 2023.

68. van der Velden BHM. Explainable AI: current status and future potential. Eur Radiol. 2023;4–6.

69. Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, et al. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion. 2023;99(April):101805.

70. Phillips PJ, Hahn CA, Fontana PC, Broniatowski DA, Przybocki MA, Hahn CA, et al. Four Principles of Explainable Artificial Intelligence: Draft NISTIR 8312. National Institute of Standards and Technology Interagency or Internal Report. 2020;(August).

71. Key facts UNESCO ' s the Ethics of Artificial Intelligence. 2021;(November).

72. Ribeiro MT, Singh S, Guestrin C. Model-Agnostic Interpretability of Machine Learning. 2016;(Whi).

73. Larasati R. Explainable AI for Breast Cancer Diagnosis: Application and User's Understandability Perception. International Conference on Electrical, Computer, and Energy Technologies, ICECET 2022. 2022;(June):1–6.

74. Xie N, Ras G, van Gerven M, Doran D. Explainable Deep Learning: A Field Guide for the Uninitiated. 2020;

75. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-Excitation Networks. :1–13.

76. Li Z, Chen X, Zhang X, Jiang R, Chen S. Latent feature extraction with a prior-based self-attention framework for spatial transcriptomics. Genome Res. 2023;33(10):1757–73.

77. Kang M, Lee S, Lee D, Kim S. Learning Cell-Type-Specific Gene Regulation Mechanisms by Multi-Attention Based Deep Learning With Regulatory Latent Space. Front Genet. 2020;11(September):1–11.

78. Yang Y, Zhou M, Fang Q, Shen H Bin. AnnoFly: Annotating Drosophila embryonic images based on an attention-enhanced RNN model. Bioinformatics. 2019;35(16):2834–42.

79. Park S, Koh Y, Jeon H, Kim H, Yeo Y, Kang J. Enhancing the interpretability of transcription factor binding site prediction using attention mechanism. Sci Rep. 2020;10(1):1–10.

80. Yang M, Huang L, Huang H, Tang H, Zhang N, Yang H, et al. Integrating convolution and self-Attention improves language model of human genome for interpreting non-coding regions at base-resolution. Nucleic Acids Res. 2022;50(14):E81.

81. Gan Y, Huang X, Guo W, Yan C, Zou G. Predicting synergistic anticancer drug combination based on low-rank global attention mechanism and bilinear predictor. Bioinformatics. 2023;39(10).

82. Choi SR, Lee M. Transformer Architecture and Attention Mechanisms in Genome Data Analysis: A Comprehensive Review. Biology (Basel). 2023;12(7).

83. Amoroso N, Quarto S, La Rocca M, Tangaro S, Monaco A, Bellotti R. An eXplainability Artificial Intelligence approach to brain connectivity in Alzheimer's disease. Front Aging Neurosci. 2023;15.

84. Comes MC, Arezzo F, Cormio G, Bove S, Calabrese A, Fanizzi A, et al. An explainable machine learning ensemble model to predict the risk of ovarian cancer in BRCA-mutated patients undergoing risk-reducing salpingo-oophorectomy. Front Oncol. 2023;13(July):1–11.

85. Prelaj A, Galli EG, Miskovic V, Pesenti M, Viscardi G, Pedica B, et al. Real-world data to build explainable trustworthy artificial intelligence models for prediction of immunotherapy efficacy in NSCLC patients. Front Oncol. 2023;12(January):1–14.

86. Massafra R, Fanizzi A, Amoroso N, Bove S, Comes MC, Pomarico D, et al. Analyzing breast cancer invasive disease event classification through explainable artificial intelligence. Front Med (Lausanne). 2023;10.

87. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug 26;596(7873):583–9.

88. Pearce R, Zhang Y. Deep learning techniques have significantly impacted protein structure prediction and protein design. Vol. 68, Current Opinion in Structural Biology. Elsevier Ltd; 2021. p. 194–207.

89. Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, et al. MGnify: The microbiome analysis resource in 2020. Nucleic Acids Res. 2020 Jan 1;48(D1):D570–8.

90. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020 Jan 30;577(7792):706–10.

91. Zuo C, Dai H, Chen L. Deep cross-omics cycle attention model for joint analysis of single-cell multi-omics data. Bioinformatics. 2021;37(22):4091–9.

92. Li H, Ma T, Hao M, Guo W, Gu J, Zhang X, et al. Decoding functional cell–cell communication events by multi-view graph learning on spatial transcriptomics. Brief Bioinform. 2023;24(6):1–13.

93. Benhammou Y, Achchab B, Herrera F, Tabik S. BreakHis based breast cancer automatic diagnosis using deep learning: Taxonomy, survey and insights. Neurocomputing. 2020;375(October):9–24.

94. Guo H, Lv X, Li Y, Li M. Attention-based GCN integrates multi-omics data for breast cancer subtype classification and patient-specific gene marker identification. Brief Funct Genomics. 2023;22(5):463–74.

95. Wang C, Lye X, Kaalia R, Kumar P, Rajapakse JC. Deep learning and multi-omics approach to predict drug responses in cancer. BMC Bioinformatics. 2021;22:1–15.

96. Pang J, Liang B, Ding R, Yan Q, Chen R, Xu J. A denoised multi-omics integration framework for cancer subtype classification and survival prediction. Brief Bioinform. 2023;24(5):1–12.

97. Singh A, Sengupta S, Lakshminarayanan V. Explainable deep learning models in medical image analysis. :1–18.

98. Lee J, Lee Y, Kim J, Kosiorek AR, Choi S, Teh YW. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. 2018;

99. Guo M hao, Liu Z ning, Mu T jiang, Hu S min, Member S. Beyond Self-attention : External Attention using Two Linear Layers for Visual Tasks. 2015;14(8):1–11.

100. Zuo C, Zhang Y, Cao C, Feng J, Jiao M, Chen L. Elucidating tumor heterogeneity from spatially resolved transcriptomics data by multi-view graph collaborative learning. Nat Commun. 2022;13(1):1–14.

101. Kaczmarek E, Jamzad A, Imtiaz T, Nanayakkara J, Renwick N, Mousavi P. Multi-Omic Graph Transformers for Cancer Classification and Interpretation. Pac Symp Biocomput. 2022;27:373–84.

102. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-Excitation Networks. :1–13.

103. Jeong D, Koo B, Oh M, Kim TB, Kim S. GOAT: Gene-level biomarker discovery from multi-Omics data using graph ATtention neural network for eosinophilic asthma subtype. Bioinformatics. 2023;39(10).

104. Ouyang D, Liang Y, Li L, Ai N, Lu S, Yu M, et al. Integration of multi-omics data using adaptive graph learning and attention mechanism for patient classification and biomarker identification. Comput Biol Med. 2023;164(March).

105. Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: A review. Brief Bioinform. 2022;23(2):1–15.

106. Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP. Harnessing multimodal data integration to advance precision oncology. Nat Rev Cancer. 2022;22(2):114–26.

107. Vale-Silva LA, Rohr K. Long-term cancer survival prediction using multimodal deep learning. Sci Rep. 2021;11(1):1–12.

108. Lopez Pinaya WH, Vieira S, Garcia-Dias R, Mechelli A. Autoencoders. Machine Learning: Methods and Applications to Brain Disorders. 2019;193–208.

109. Škrlj B, Daeroski S, Lavrač N, Petković M. Feature importance estimation with self-attention networks. Frontiers in Artificial Intelligence and Applications. 2020;325:1491–8.

110. Arya N, Saha S. Multi-Modal Classification for Human Breast Cancer Prognosis Prediction: Proposal of Deep-Learning Based Stacked Ensemble Model. IEEE/ACM Trans Comput Biol Bioinform. 2022;19(2):1032–41.

111. Olatunji I, Cui F. Multimodal AI for prediction of distant metastasis in carcinoma patients. Frontiers in Bioinformatics. 2023;3.

112. Guo W, Liang W, Deng Q, Zou X. A Multimodal Affinity Fusion Network for Predicting the Survival of Breast Cancer Patients. Front Genet. 2021 Aug 20;12.

113. Zhang Y, Li A, He J, Wang M. A Novel MKL Method for GBM Prognosis Prediction by Integrating Histopathological Image and Multi-Omics Data. IEEE J Biomed Health Inform. 2020 Jan 1;24(1):171–9.

114. Wulczyn E, Steiner DF, Moran M, Plass M, Reihs R, Tan F, et al. Interpretable survival prediction for colorectal cancer using deep learning. NPJ Digit Med. 2021;4(1):1–13.

115. Bora K, Bhuyan MK, Kasugai K, Mallik S, Zhao Z. Computational learning of features for automated colonic polyp classification. Sci Rep. 2021;11(1):1–16.

116. Tougeron D, Fauquembergue E, Rouquette A, Le Pessot F, Sesboüé R, Laurent M, et al. Tumor-infiltrating lymphocytes in colorectal cancers with microsatellite instability are correlated with the number and spectrum of frameshift mutations. Modern Pathology. 2009 Sep;22(9):1186–95.

117. Natrajan R, Little SE, Reis-Filho JS, Hing L, Messahel B, Grundy PE, et al. Amplification and overexpression of CACNA1E correlates with relapse in favorable histology Wilms' tumors. Clinical Cancer Research. 2006 Dec 15;12(24):7284–93.

118. He XS, Ye WL, Zhang YJ, Yang XQ, Liu F, Wang JR, et al. Oncogenic potential of BEST4 in colorectal cancer via activation of PI3K/Akt signaling. Oncogene. 2022 Feb 18;41(8):1166–77.

119. Pan X, Wang Q, Xu C, Yan L, Pang S, Gan J. Prognostic value of chloride channel accessory mRNA expression in colon cancer. Oncol Lett. 2019 Sep 1;18(3):2967–76.

120. Yang GZ, Hu L, Cai J, Chen HY, Zhang Y, Feng D, et al. Prognostic value of carbonic anhydrase VII expression in colorectal carcinoma. BMC Cancer. 2015 Apr 1;15(1).

121. Zhang L, Gong Y, Wang S, Gao F. Anti-colorectal cancer mechanisms of formononetin identified by network pharmacological approach. Medical Science Monitor. 2019 Oct 14;25:7709–14.

122. Zhang Q, Nie H, Pan J, Xu H, Zhan Q. FMNL3 is Overexpressed in Tumor Tissues and Predicts an Immuno-Hot Phenotype in Pancreatic Cancer. Int J Gen Med. 2022;15:8285–98.

123. Brodsky AS, Khurana J, Guo KS, Wu EY, Yang D, Siddique AS, et al. Somatic mutations in collagens are associated with a distinct tumor environment and overall survival in gastric cancer. BMC Cancer. 2022 Dec 1;22(1).

124. Yu H, Qin XK, Yin KW, Li ZM, Ni E De, Yang JM, et al. EphB6 deficiency in intestinal neurons promotes tumor growth in colorectal cancer by neurotransmitter GABA signaling. Carcinogenesis. 2023 Aug 1;44(8–9):682–94.

125. Bildik G, Liang X, Sutton MN, Bast RC, Lu Z. DIRAS3: An Imprinted Tumor Suppressor Gene that Regulates RAS and PI3K-driven Cancer Growth, Motility, Autophagy, and Tumor Dormancy. Vol. 21, Molecular Cancer Therapeutics. American Association for Cancer Research Inc.; 2022. p. 25–37.

126. Zhou L, Yu Y, Wen R, Zheng K, Jiang S, Zhu X, et al. Development and Validation of an 8-Gene Signature to Improve Survival Prediction of Colorectal Cancer. Front Oncol. 2022 May 10;12.

127. Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network. 2015 Mar 9; Available from: http://arxiv.org/abs/1503.02531

128. Banner R, Hubara I, Soudry D. Scalable Methods for 8-bit Training of Neural Networks.

129. Chmiel B, Ben-Uri L, Shkolnik M, Hoffer E, Banner R, Soudry D. Neural gradients are near-lognormal: improved quantized and sparse training. 2020 Jun 15; Available from: http://arxiv.org/abs/2006.08173

130. Gholami A, Kim S, Dong Z, Yao Z, Mahoney MW, Keutzer K. A Survey of Quantization Methods for Efficient Neural Network Inference. 2021 Mar 25; Available from: http://arxiv.org/abs/2103.13630

131. O' Neill J, Steeg G Ver, Galstyan A. Compressing Deep Neural Networks via Layer Fusion [Internet]. Available from: www.aaai.org