Journal of Computer Science

Research Article

UMPM-Net: A Multimodal Deep Learning Framework for Colorectal Cancer Prognosis

K. Muthuchamy and S K. Piramu Preethika

Department of Computer Science & Information Technology, Vels Institute of Science Technology and Advanced Studies (VISTAS), Chennai, India

Article history
Received: 15-03-2026
Revised: 20-06-2026
Accepted: 21-07-2026

Corresponding Author:
K. Muthuchamy
Department of Computer Science & Information Technology, Vels Institute of Science Technology and Advanced Studies (VISTAS), Chennai, India
Email: muthuchamy412@gmail.com

Abstract: Colorectal cancer (CRC) is one of the most common cancers worldwide, for which early prediction of patient outcomes is important for personalised treatment planning and to improve survival. However, current prediction systems are limited in the fact that they are based on single-modality data and cannot capture the complex interactions between clinical, imaging, and genomic factors. To overcome these shortcomings, this study presents a Unified Multimodal Predictive Modeling Network (UMPM-Net) which incorporates Electronic Health Records (EHRs), medical imaging biomarkers and genomic signatures in a deep learning model. The proposed approach uses U-Net for accurate tumor segmentation, EfficientNet for imaging feature extraction and DenseNet for classification and survival prediction. The multimodal features are then fused at the feature level as a way of obtaining complementary information across the modalities. Experiments were performed on a multi-institutional dataset of a CRC with 1200 distinct patients with paired imaging, EHR and genomic profiles. Quantitative results show better performance than five state-of-the-art models with Dice score 0.92 for segmentation, accuracy score 0.90, ROC-AUC score 0.95 for outcome classification and concordance index (C-index) score 0.80 for predicting survival. The ablation analysis further confirms that the fusion of imaging, clinical and genomic modalities has a significant impact on the predictive accuracy and clinical interpretability. In conclusion, the proposed UMPM-Net is a comprehensive, explainable and robust multimodal learning framework that is capable of accurate, interpretable, and clinically relevant prognostication for colorectal cancer.

Keywords: Colorectal Cancer, Multimodal Deep Learning, Electronic Health Records, Genomic Signatures, Medical Imaging, U-Net, EfficientNet, DenseNet, Predictive Modeling, Survival Analysis

Introduction

Colorectal cancer (CRC) is one of the leading causes of cancer-related morbidity and mortality worldwide with almost 10% of all cancer diagnoses and deaths (Sung et al., 2021). Early and accurate prediction of disease outcome, such as recurrence, response to therapy and overall survival, is an essential aspect for the optimization of therapeutic strategies and the improvement of patient prognosis. With the growing availability of clinical, imaging and genomic data, there is a growing opportunity to use data-driven predictive models for precision oncology. However, how to integrate heterogeneous modalities into a common clinically interpretable model is still a great challenge (Usher-Smith et al., 2016).

Medical image analysis has recently experienced an advanced level of performance due to developments in Artificial Intelligence (AI) and deep learning for segmentation, lesion detection, and disease classification, as examples (Liu et al., 2024). Electronic Health Record (EHR) describes patient-specific information, such as demographic and comorbidity and laboratory test results (Wang et al., 2019) on the other hand, genomic data provides information on molecular changes affecting the tumor's behavior and response to therapy (McGeoch et al., 2019). While all these data sources provide useful information, they, when analysed in isolation, do not provide enough information to capture the clinical context and biological complexity underlying CRC progression.

Existing studies have tried to predict CRC outcomes using unimodal and bimodal strategies. Image-based deep learning models with Convolutional Neural Networks (CNNs) and transformers have been proven to be promising for tumor classification and segmentation (Tasnim et al., 2021). However, these techniques tend to ignore the non-visual clinical and genetic variables that have a strong impact on prognosis. On the other hand, in clinical or genomic-only models such as Cox proportional hazards model or random survival forest model, the spatial information is not considered and the predicted accuracy is relatively poor (Win et al., 2012). A lot of multimodal frameworks have been proposed, but most of them are based on late fusion, in which inferences from different models are concatenated, which results in suboptimal feature interaction and missing of information. Furthermore, many current models are not interpretable, and it may be hard for clinicians to trust and adopt AI-driven recommendations in a real-world setting (He et al., 2025).

Filling these gaps, there is a strong need for a combined, interpretable and data sparing framework that can integrate EHR, imaging and genomic modalities, for CRC outcome prediction. The research problem we study in this paper is the lack of an end-to-end multimodal model that jointly characterizes tumor morphology, patient clinical state, and molecular characteristics for improved prognostic accuracy and clinical interpretability. The main objectives of the study are as follows:

- To create an integrated, multimodal deep learning framework (UMPM-Net) for colorectal cancer outcome prediction using EHR, imaging and genomic features
- To design Segmentation-Extraction-Classification Pipeline with U-Net, EfficientNet and DenseNet for Accurate but Interpretive Modeling
- To evaluate the performance of the proposed model with respect to state-of-the-art approaches using quantitative, ablation and interpretability studies

Literature Review

Deep learning has become a big progress in the domain of cancer outcome prediction as it has made it possible to analyze complicated data. However, traditional unimodal models (i.e., imaging only, clinical only, or genomics only) are limited, which is what is addressed by the recent trend of multimodal deep learning. In this paper we summarize some of the recent advances in this field, classified according to the integration of modalities, fusion strategies, application domain and interpretability issues.

He et al. (2025) suggested a multimodal deep learning framework with Multimodal Compact Bilinear Pooling for 5-year CRC survival prognosis with the combination of histopathological, clinical and gene mutation. The improved calculated Area Under the Curve (AUC) was demonstrated in the study with the finding that multimodal fusion has increased the predictive accuracy of the study, limited by the small sample size (84 cases) (Jang et al., 2020). Huang et al. (2022) presented a multimodal deep learning model for predicting Tumor Mutational Burden (TMB) from histopathological and clinical data, which is a much cheaper and faster alternative to sequencing. Although effective, the model's use of image derived surrogates may limit the level of precision when compared to genomic gold standards.

Dai et al. (2024) proposed a multi-head attention based survival prediction model for multimodal data fusion, which can solve the inter-modal heterogeneity problem and improve the robustness. The study adds significantly to CRC survival prediction accuracy, but wider testing of the dataset and interpretability analysis is required. Ding (2024) proposed a Multimodal Transformer Framework for Pathology and Genomics Based CRC Outcome Prediction Exploiting Graph Based Spatial Connections and Contrastive Pretraining. The work shows great innovation in multimodal representation with the need for clinical generalization and real world interpretability testing.

Xie et al. (2025) has developed an AI based multimodal analyzer including radiological, histopathological, and genomic data to guide the adjuvant chemotherapy in stage II CRC. The biologically explainable model identifies different clusters that respond to treatment, which is of translational value, but is dependent on the variability of the imaging done in a particular cohort and requires validation in an independent cohort.

Schulz et al. (2021) have designed a Multimodal Deep Learning Model (MMDLM) using the combination of histopathological, radiological and genomic information to predict renal cancer prognosis. The model showed excellent predictive performance (C-index = 0.7791), which was better than clinical parameters. While it has great use of multimodal data, limited sample diversity and cohort size limit generalizability and clinical translation.

Badic et al. (2021) discussed the emerging field of radiogenomics in colorectal cancer and its potential for personalized medicine by combining radiomic and genomic information. Despite its promising nature, the study identifies the early stage development, methodological variability and lack of standardized validation as potential hindrances to its clinical applicability.

Tsai et al. (2023) proposed the MOMA platform, which is an explainable machine learning that creates links between histopathology, multi-omics, and clinical data to predict CRC survival and genomic alterations.

While the study has good generalizability and interpretability, its complexity and high-quality data requirements may make it difficult to scale to real-world implementations.

Mohammadpour et al. (2025) performed a scoping review of 41 studies on the application of imaging and radiomics to predict molecular changes in CRC and achieved up to 81% AUC for MSI and 80% for BRAF. The review is supportive of imaging based molecular prediction but recognizes the heterogeneity and lack of clinical standardization of the studies.

Lu et al. (2021) has suggested a deep learning model based on serial CT to predict early treatment response to metastatic CRC. The model performed better than the conventional size-based metrics, increasing C-index to 0.694. However, small sample size, retrospective design and lack of interpretation limit its direct clinical translation.

Tao et al. (2024) analysed the use of liquid biopsy in the clinical application of CRC for detection, monitoring and prognosis through CTCs, ctDNA and TEPs. The non-invasiveness and prediction potential of the method is remarkable; however, technical variability, sensitivity and standardization issues restrict the use to routine clinical practice.

Jackson et al. (2025) suggested a hybrid deep learning model for colorectal cancer segmentation, which is based on U-Net and attention gates. The study illustrated better precision at the boundaries and contextual learning. However, despite being effective, it is dependent on attention mechanisms which may lead to increased computational complexity limiting scalability across different large-scale histopathology datasets.

Raju et al. (2025) have introduced PolynetDWTCADx, which uses CNN, DWT, and SVM for the detection of colorectal cancer with 95% training accuracy and 0.93 IoU. The hybridization enhanced the segmentation fidelity but the model generalizability is still limited by the dataset diversity, and performance might be affected by the real-world clinical imaging noise.

Shah et al. (2021) created AtResUNet with atrous convolutions and residual blocks for colorectal cancer segmentation, which got a Dice coefficient of 0.753. The model was good at describing multi-scale features but had a low level of robustness caused by medium dice scores, indicating a need for further tuning and the addition of more data.

Girepunje and Singh (2024) EffiSANet, an EfficientNet integrated with self-attention for colorectal cancer classification, with an accuracy of 95.49%. The model was able to efficiently extract discriminative features, making it more interpretable. However, the performance dependence on small datasets can be a limiting factor for clinical applicability and cross-cohort validation is necessary to ensure that the performance is stable and reproducible in a wide range of histopathological diseases.

Saba et al. (2024) used EfficientNet with NSGA-II as the optimization to detect colon cancer and achieved an accuracy of 99.97% with multiple data sets. The hybrid model performed well in terms of performance and efficiency, but the possibility of overfitting is raised by extensive hyperparameter tuning and raises questions about how the model can perform in the real-world, including its adaptation and generalization to medical imaging domains that it has not seen before.

Sarkar et al. (2021) proposed a modified DenseNet with an auto-encoder for colorectal histology classification with 97.2% accuracy. The reconstruction-based learning in a model improved feature extraction without preprocessing. However, it is dependent on deep architecture and can raise the cost of training, and is lacking in interpretability for transparent clinical integration.

Predicting prognosis in Colorectal Cancer (CRC) from various image modalities has been demonstrated to be possible using Multimodal Deep Learning (MDL); however, there are several key issues. Most existing approaches are based on the late-fusion techniques, which is to perform feature extraction on each modality separately to obtain predictions for each modality, and then combine them. They don't work very well in capturing the complex cross-modal interaction and complementary biological relationships between tumor phenotype and patient characteristics as well as the molecular changes. To overcome this, the proposed multimodal fusion method at the feature level will be introduced in UMPM-Net which will allow the fusion of image, clinical and genomic representations in a common latent space to learn richer inter-modal dependencies.

Other important restrictions are nonexistence of good interpretability mechanisms. In reality, from the clinical application point of view, the current deep learning models are essentially treated as "black-box models" and cannot be used in clinical applications. To enhance the transparency of the model, UMPM-Net has the ability to interpret images by using Grad-CAM, and to attribute the features to the clinically relevant clinical and genomic biomarkers, giving clinically relevant explanations for the model predictions.

There have been a number of studies which have concentrated on the classification accuracy, with only cursory support of prognosis and survival analysis. For these reasons, the proposed framework considers the multimodal prognostic representations in the prediction network DenseNet-121 to predict the recurrence risk, treatment response and survival outcome, all based on a single patient representation.

Furthermore, previous literature has had limited variability of cohorts, modality specific learning and

been without analysis of generalization. The framework proposed aims to leverage the best use of multimodal data – heterogeneous and heterogeneous (clinical record, radiological imaging, pathology images and genomic profiles) to reinforce the robustness of the framework and to gather complementary biological information. The system is end-to-end optimized and learns representations across the different modalities, while being able to see the information loss in the typical unimodal pipelines. Overall, the proposed UMPM-Net is specially designed to address the limitations of the existing multimodal colorectal cancer studies, which include the ability to explain the multimodal fusion process and better understand the prognosis prediction, and to enable an end-to-end learning.

Materials and Methods

The proposed framework employs combined predictive modeling of Colorectal Cancer (CRC) by means of a unified deep learning pipeline for multimodal data including Electronic Health Records (EHRs), medical imaging biomarkers and genomic signatures. The architecture synergistically uses U-Net for image segmentation, EfficientNet for feature extraction and DenseNet for outcome prediction, thus realizing accurate localization of lesion, discriminative feature learning and prediction of outcome. The general flow chart is shown in Fig. 1.

Data Acquisition and Preprocessing

Multimodal input data is collected from three main sources:

(1) EHR data, including demographics, laboratory parameters and treatment history
(2) Histopathology Image dataset
(3) Genomic data of RNA sequencing and mutation profiling

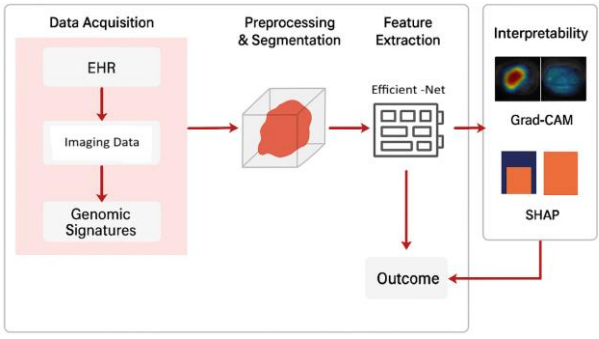

Fig. 1: Overall flow of the proposed framework for multimodal colorectal cancer

Each of the modalities of data is individually preprocessed.

EHR Data Preprocessing: EHRs are cleaned to eliminate discrepancies and missing values and the numerical attributes are normalized. Categorical embeddings (gender, comorbidity and treatment types) are either represented by one-hot encoding or learned embeddings that represent these variables in a structured numerical way that can be combined by the model across different heterogeneous patient data sources.

Imaging Data Preprocessing: The preprocessing step includes downloading the WSIs and the associated clinical metadata. WSIs are often gigapixel and therefore they are tiled into smaller patches (e.g., 224x224 or 512x512 pixels) in order to be fed into CNNs. Color normalization (e.g. Macenko or Reinhard) are made to minimize the variability in H&E staining. Non-tissue regions are removed with the help of Otsu thresholding or tissue-masks. Finally, patches are matched with either tumor or normal label as well as with molecular subtype or survival outcome for training the model.

The proposed multimodal framework will include two imaging modalities: Radiological imaging and digital pathology imaging, complementary with each other. Radiological imaging involves contrast enhanced Computed Tomography (CT) scans and/or Magnetic Resonance Imaging (MRI) to give anatomical and morphological information about where the tumors are, their size and the characteristics of the surrounding tissues. These volumetric images would then be fed into a 3D U-Net segmentation network, which is used to precisely localise the tumour and extract the region of interest.

Further, the digital pathology data are the Whole-Slide Images (WSIs) of surgically excised tissue samples. After tissue segmentation and stain normalization the WSIs are sliced into smaller image patches. These images of the pathology have the detailed information on the cellular and tissue level related to tumor heterogeneity and disease progression.

The radiologic and pathological imaging features extracted are then fused with clinical and genomic information as suggested by the multimodal fusion framework. The model provides a more holistic representation of colorectal cancer and better prediction of performance in prognostic prediction through incorporating anatomical, histological, clinical and molecular features.

Genomic Data Preprocessing: Genomic dataset from RNA sequencing and mutation profiling goes through rigorous data filtering to filter out low variance and noisy genes. Variance thresholding is used to identify steady gene expression while PCA-based dimension reduction picks up the top components that represent the significant biological variance. This process maintains discriminative genetic markers that are related to cancer progression in

an optimized way that ensures compactness and interpretability of features for modeling.

A wide variety of multimodal colorectal cancer data sets were used: Clinical data, imaging data, and genomic data, from the TCIA-TCGA COAD and READ collections, to increase the generalizability of the proposed framework. Patients' data included a variety of patients and age, gender, tumor stage, treatment history, molecular sub-type and survival outcome. While the electronic Health Records provide information from a variety of clinical perspectives, high-resolution CT/MR scans, mutation and RNA sequencing profile data provide complementary information from multiple biological and clinical angles. The diversity ensures that the framework can have strong representations that are more adaptable to population-specific bias as well as can be applied to different clinical scenarios.

Missing Data Handling

A detailed data quality audit was conducted on all datasets before they were used to develop the models to locate missing, inconsistent or incomplete data. Continuous variables (such as laboratory values and clinical signs where one or more were missing) were completed with the value for the median, to minimize the influence of extreme values and to assure the statistical structure of the data. Categorical data (such as treatment category, for instance or other condition data) was imputed as the most common, or as an additional "unknown" category as appropriate.

Samples rejected by the Quality Control (QC) for imaging data were samples that were severely corrupted, had missing imaging annotations or had poor image quality. Genes were filtered out to minimize excess missing data and very low expression levels of genomic data for downstream analyses. Normalization and a standard filtering were performed in a standard transcriptomic analysis to fill in the missing genomic values.

To avoid data leakage, all the imputation parameters were estimated in the training dataset and then used in the validation and test datasets. This prevented any information in unseen samples from affecting the training. The missing data handling methods improves data completeness and also reduce the potential for bias and enhanced the power and credibility of the proposed multimodal colorectal cancer prediction methodology.

Tumor Segmentation: Implementation and U-Net

The 3D U-Net architecture is specifically engineered to perform segmentation on volumetric medical images (e.g. CT or MRI images) by precisely identifying the tumor region within 3D space. As shown in Figure 2, the encoder-decoder architecture with skip links to retain spatial and contextual information all the way through the network.

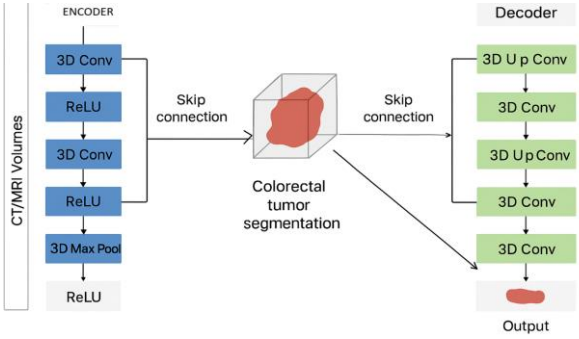

Fig. 2: Architecture of 3D- U-Net for colorectal tumor segmentation

Encoder (contracting path): The encoder is used to compress the input image volume with several 3D convolutional, ReLU, and max-pooling layers. This process extracts more and more abstract and high-level contextual features representing the appearance of the tumor and surrounding anatomy at many different scales.

Decoder (expanding path): The decoder then uses compositional features that are successively up-scaled in order to reconstruct the segmentation map. It re-increases the spatial resolution along with incorporating the detailed features of the encoder through skip connectivity, which aids in the recovery of accurate tumor boundary from the downsampled encoder.

Loss function: The network uses a combination of Dice Loss and Binary Cross Entropy (BCE) Loss in the way that minimizes the region overlap and the pixel-wise accuracy to achieve the best. The Dice loss provides an estimate of the geometry of the tumor by calculating the similarity between the predicted mask and the actual mask while BCE penalizes the misclassification of voxels. This loss, along with the use of that intermediate representation, compels the model to learn both the global shape consistency and the local boundary accuracy to produce highly accurate segmentation of colorectal tumors which are used as strong Regions of Interest (ROIs) for subsequent feature extraction and predictive modeling.

Feature Extraction Using EfficientNet

After tumor segmentation, each Region of Interest (ROI) is processed via EfficientNet-B4 which has been chosen for its compound scaling which simultaneously searches for the optimal depth, width, and spatial resolution. The network extracts morphological, textural and intensity based biomarkers from CT/MRI volumes. Global average pooling is used to compress spatial representations by means of high-level descriptor vectors. These imaging features are combined with the clinical and genomic embedding in order to form a compound representation. Leveraging ImageNet-pretrained weights

through transfer learning helps to speed up convergence and help generalization mainly for limited medical datasets. EfficientNet with its efficiency-accuracy tradeoff provides effective extraction of discriminative imaging features for outcome prediction in colorectal cancer.

The Tumor Segmentation is done using 3D U-Net model, and the volumetric tumor masks from CT/MRI scans are obtained, thus defining Regions of Interest (ROIs). The segmented volumes of the tumor is then converted to image patch centered on each particular tumor slice and converted to 2D image patch centered on the tumor. These patches are then resized to the original size of the input fed to the EfficientNet-B4 model that is used for feature extraction.

The whole 3D volume can be processed using EfficientNet-B4, but it is more efficient to do so using only the 2D tumour patches in the segmented ROI areas. The reason of this design is that they are based on weights trained on ImageNet, whereas the 2D convolutional architectures have been proven to be effective in representation learning of medical images. For each tumor patch, the network is able to derive high-level morphological, textural and intensity-based imaging biomarkers. In the global average pooling, the latter are used to reduce the dimensionality of the data into so-called "feature representations", which are then combined with clinical and genomic features at the multimodal fusion stage. This segmentation to feature-extraction pipeline can be employed to extract only features from the biological component of the tumour without including non-tumorous tissues.

Multimodal Fusion

In order to integrate or fuse different information sources, a feature-level fusion mechanism is employed. In particular, clinical phenotype features from EHR, genomic signatures and imaging features from EfficientNet are concatenated into a single latent vector. This fused representation is passed through the fully connected layers with batch normalization and dropout which will help to improve the decorrelation among features and prevent overfitting. Combining these two aspects integrates the complex interdependence between the tumor phenotype, the patient's physiology, and the molecular changes. The multimodal synergy sheds light on the entire spectrum of colorectal cancer progression giving the model the ability to deduce the prognostic behavior that is usually not enlightened by the unimodal approaches. The imaging biomarkers from EfficientNet-B4 are preprocessed and genomic features of EHR are added to create fixed length feature vectors. Let F, G, and H be the feature vectors of clinical features, genomic features and imaging features, respectively. These vectors are linked together to take a single representation as given in the Eq. (1):

$$F_{fusion} = [\, F_{ehr};\, F_{gen}\,;\, F_{img}] \quad (1)$$

The output feature vector is then passed into fully connected layers of which the first layer has batch normalization, ReLU activation function and dropout regularization (rate = 0.5) to learn the cross-modal interactions and to avoid overfitting.

Instead of being newly developed, U-Net and EfficientNet are used as basic deep network architectures commonly used in medical image analysis in this work. In particular, the 3D U-Net is used to accurately localise the tumour and extract the region of interest and the EfficientNet-B4 is applied to extract discriminative imaging biomarkers from the extracted ROI in the segmented tumour regions. The main innovation of this work is the suggested Unified Multimodal Prognostic Modeling Network (UMPM-Net) which integrates imaging, clinical and genomic data as one end-to-end learning system.

The proposed solution offers an alternative feature-level multimodal fusion scheme rather than the usual multimodal solution which involves processing each modality with a different algorithm in the subsequent fusion stage at decision level. These clinical features, genomic signatures based on RNA sequencing and mutation analysis and imaging biomarkers based on EfficientNet-B4 are converted to a common latent representation. The complexity of the tumor morphology, molecular changes and patient-specific clinical parameters can be explored in the model.

The learning of the risk of recurrence and treatment response and survival outcomes are then optimized together with a DenseNet classifier in a prognostic prediction module. In this integrated learning strategy less information loss occurs compared to unimodal/fast-fusion models and a better representation of modalities' complementarity occurs in the model. Thus, the proposed framework can be better characterized of colorectal cancer and can provide a better prognosis than the existing multimodal approaches.

Explainability features are also part of the framework, such as explainability visualizations with Grad-CAM and feature attributions with SHAP, that can serve as part of the explanation of the imaging, clinical and genomic contribution towards the ultimate prediction. The multimodal fusion, the prognostic modeling and the interpretability of the proposed study is the main methodological novelty, and paves the way for its potential application in precision oncology and personalized treatment planning.

Outcome Prediction DenseNet

The clinical, imaging and genomic representations are fused multimodally and a fused feature vector is given as input to the DenseNet-121 prediction network. The

DenseNet-121 was chosen because of its dense connectivity pattern, which enables the efficient reuse of features, better gradient propagation and mitigates the vanishing gradients problem in deep networks. The concatenated feature vector is fed into a series of dense connected layers, followed by the batch normalization and dropout regularization, with respect to the improvement of the generalization accuracy. Prognosis prediction is a fully connected layer which produces predictions regarding individual's colorectal cancer prognosis (probability of recurrence, response to treatment, and survival). To make predictions of the outcome, the Softmax activation function is used to output class probabilities. Adam optimizer with learning rate 1×10^{-4}, batch size 32 and categorical cross-entropy loss function are used to train the model. To prevent overfitting and stable convergence, parameters of early stopping and learning-rate scheduling are used. The predicted outcome probabilities at that time are then used for the risk stratification, the analysis of the survival. All the other three models (DeepSurv, DeepHit, CoxTime) are only used as the benchmark methods and DenseNet-121 is the key prediction model that is proposed in the proposed framework.

After Multimodal Fusion, the multimodal latent representation F_{fusion} is obtained to generate the multimodal representation. This provides complementary clinical, imaging and genomic data for each patient. The feature vector is then fused and passed to a prognostic prediction module (DenseNet-121). The dense connectivity structure enables efficient sharing of features and induction of high level nonlinear interactions of multimodal biomarkers for disease progression and patient survival.

The DenseNet network enhances the fused representation using a series of densely connected layers, batch normalization and dropout regularization. Fully connected prediction layers are then used to generate patient-specific prognostic risk scores with the high level latent representation. These risk scores are numerical representations of the risk for poor clinical outcomes, such as disease recurrence, treatment resistance, and lowered likelihood of survival.

Patients are split into two groups: Low-Risk Group (LRG) and High-Risk Group (HRG) for survival analysis on the basis of the predicted risk score. Thereafter Kaplan-Meier survival curves are drawn to compare survival differences between risk-groups. Survival separation is determined to be statistically significant by the log-rank test. This strategy allows for visualization of prognostic differences over a long term period but allows for clinically interpretable survival stratification.

Considering the common situation of censored observations in clinical survival data, the survival status and the follow-up time are included in the modelling and assessment process. The model is trained with both event samples and censored samples, which give the model a valuable patient information. Evaluation performed using parameters specific to survival: Concordance Index (C-index), time-dependent ROC-AUC and Kaplan-Meier survival analysis.

Risk scores and survival probabilities produced will provide personalized prognosis information for clinical decision making, treatment planning and patient risk assessment. The proposed framework can model complex relationship between tumor morphology, molecular alteration and clinical characteristics using multimodal representations learned jointly from them, which results in learning better in predicting the survival.

Network Configuration

The 3D U-Net segmentation network comprises of four encoder and four decoder blocks. The encoders are all made out of 3D convolution layers(kernel size = $3 \times 3 \times 3$) with ReLU activation and max pooling layers. Skip connection: High resolution spatial information from corresponding encoder layers to the decoder layers. The Loss function applied to train the network is the Dice Loss function along with the Binary Cross-Entropy Loss function as represented in the Eq. (2):

$$L_{total} = 0.5\, L_{Dice} + 0.5\, L_{BCE} \quad (2)$$

Where Dice Loss ensures that the two sets of regions have the largest possible overlap, and BCE ensures the highest possible accuracy from the voxel-wise classification.

The model EfficientNet-B4 is a transfer learning model with weights pre-trained on ImageNet. The Global Average Pooling method generates a feature vector of dimension 1792, which represents the tumour morphology, texture and intensity features. The DenseNet-121 prediction module gets the fused multimodal representation and comprises dense blocks with transition layers. A final softmax classifier is used for the prediction of recurrence, classification of survival or evaluation of treatment response, for which probabilities are calculated.

Training and Validation

The dataset is randomly split into training (70%), validation (15%), and test (15%) datasets with strong patient level stratification. Fivefold cross validation guarantees generalization and stability of results. Data augmentation (random rotation, flipping, elastic deformation) is used to increase the diversity of the data and reduce the overfitting of the model. The model is trained with early stopping strategy to avoid degrading the convergence. Patients data was split into strata to ensure data independence with respect to the proposed framework and to test the generalizability of the proposed framework

to an independent dataset from the training cohort. The data was split into a training, a validation and a test sets (70, 15, 15% respectively) and each patient was only allowed to be included in one set. Furthermore, five-fold cross-validation was performed in training cohort to verify the stability and robustness of the model. Future studies using independent multi-institutional data from a variety of healthcare centers and imaging protocols will be used to further test external generalization and clinical transferability.

The PyTorch framework was used for all the models and the Adam optimizer was used for training with initial learning rate of 1×10^{-4}, batch size of 16 and weight decay of 1×10^{-5}. The early stopping method was used to avoid overfitting, using 15 epochs as the patience value. When there was no improvement in the validation loss for five epochs, the learning rate reduction on plateau was used. Performance is quantitatively examined using accuracy, F1-score, ROC-AUC, precision, recall and for clinical interpretation, Kaplan-Meier survival analysis is also used. Taken together, these metrics provide evidence that the developed multimodal framework is reliable in predicting the outcome of colorectal cancer in different patient as well as imaging conditions.

Interpretability and Clinical Integration

Grad-CAM visualizations are also used to comprehend model decision-making, highlighting important imaging regions that contribute to the model decisions, which in the case of a tumor, visualize biologically meaningful regions. Meanwhile, the relative importance of clinical and genomic features is also measured by SHAP (SHapley Additive exPlanations) values, which easily reveals how much features contribute to the prediction. By providing oncologists with a dual interpretation system, these biomarkers can be linked to the clinical understanding, which will help to promote trust and accountability. With predictive analytics and domain knowledge, the system enables clinically actionable decision-making to allow physicians to confirm AI reasoning and incorporate its findings into the clinician's personalized colorectal cancer treatment planning.

Results and Discussion

The findings and discussion part gives a thorough evaluation of the proposed UMPM-Net framework for segmentation, classification, and survival prediction tasks. Overall, its superior performance, robustness and clinical relevance, as validated by ablation studies, quantitative analyses and interpretability assessments make UMPM-Net a promising tool for multimodal precision colorectal cancer prognosis and outcome prediction.

Experimental Setup

The experiments were run on a high performance compute platform containing dual NVIDIA V100/A100 GPUs (32 GB each), Intel Xeon CPUs and 256 GB of RAM for efficient and speedy processing of large-scale medical imaging and multimodal data. A few other popular tools were used, such as PyTorch 2.x for deep learning implementation, MONAI, more commonly known for medical imaging workflows, scikit-learn and scikit-survival for evaluation and survival modeling, SHAP for feature interpretability, and OpenCV and nibabel for image preprocessing and I/O handling. Furthermore, the study offers code and trained model-weights (based on data sharing permissions), a conda environment
dissolve.combo.Booster.RConvertments.yaml,
preprocessing and random seed settings for reproducibility purposes. Containerization based on Docker was used to ensure that the environment is replicated across computational infrastructures in a consistent fashion.

Computational Complexity Analysis

The three major components in the proposed framework that have potential to impact the computational complexity are segmentation, feature extraction and classification. The drawback of using 3D U-Net segmentation network is that it requires a lot of computational power, whereas the EfficientNet-B4 feature extraction network uses the compound scaling to be computationally efficient. The DenseNet-121 uses dense feature re-use, thus the propagation of the gradient is efficient and the memory requirement is less as there are no redundant parameters. The training was done using two NVIDIA V100/A100 GPUs (32 GB each), Intel Xeon processors and 256 GB of RAM. The multimodal architecture, practical training and inference time suitable for large scale clinical datasets is preserved.

Reproducibility

To ensure the reproducibility of the experiments, all of the experiments were conducted using PyTorch 2.x, MONAI, scikit-learn and scikit-survival libraries. All hyperparameters, preprocessing steps, random seed (seed = 42), optimizer settings, learning rate schedules and model architectures are explicitly documented. Consistency was achieved by using Docker containers. Moreover, the model weights, preprocessing scripts, configuration files and implementation details will be provided, as required by policies of the dataset-sharing so they can be checked independently and are available for further research extensions.

Dataset Description

The Colorectal Cancer Collection of The Cancer and Leukemia Group B (TCIA-TCGA COAD & READ) is an integrated publicly accessible resource, which will be used for colorectal cancer research utilizing radiological,

clinical, and genomic data. It consists of high resolution CT and MRI imaging, paired with RNA sequencing, somatic mutation and clinical metadata including demographics, treatment history, tumor staging and survival status. It allows multimodal analysis, which includes segmentation, biomarker discovery, and outcome prediction and is derived from The Cancer Imaging Archive (TCIA) and The Cancer Genome Atlas (TCGA). This dataset can be used for building and validating deep learning models aimed at precision oncology and prognostic modeling based on integrating imaging data, genomics data and EHR data.

Patients included in the study population were a total of 1200 patients diagnosed with colorectal cancer (CRC) in multimodal data sources from public sources. Patients in the cohort were of varied demography and clinical features including age group, gender, tumor stage, molecular type and therapy history. Each patient needed to have clinical information, imaging information and genomic information to be used in the complete prognostic assessment. A diverse mix of disease trajectories and clinical outcomes within the cohort provided a broad spectrum of diseases and outcomes to ensure the strength and applicability of the framework proposed. Patient stratification was performed at the patient level when dividing the datasets to have evenly distributed clinical data and outcome classes for training, validation and testing subsets. This multi-group population enabled the model to be trained with complementary prognostic information from the heterogeneous group of patients and to be able to predict outcomes of colorectal cancer reliably in a multimodal manner.

As a way to avoid data leakage, the dataset was split in the patient level, where 70% (840 patients) were used for training, 15% (180 patients) for validation, and 15% (180 patients) for independent testing. This ensured that no patient was represented in more than one subset, as well as strong separation between training and evaluation phases. Hyperparameters were optimized and robustness and overfitting were conserved using a 5-fold cross-validation strategy within the training set. The final optimized model was then assessed on the held-out test set and the unbiased performance was estimated. To validate the models externally, the study suggests validation on an independent cohort from a different clinical site as a measure for the generalizability and transferability of the model across populations and imaging protocols.

The proposed multimodal framework will include two imaging modalities: Radiological imaging and digital pathology imaging, complementary with each other. Radiological imaging involves contrast enhanced Computed Tomography (CT) scans and/or Magnetic Resonance Imaging (MRI) to give anatomical and morphological information about where the tumors are, their size and the characteristics of the surrounding tissues. These volumetric images would then be fed into a 3D U-Net segmentation network, which is used to precisely localise the tumour and extract the region of interest.

Further, the digital pathology data are the Whole-Slide Images (WSIs) of surgically excised tissue samples. After tissue segmentation and stain normalization the WSIs are sliced into smaller image patches. These images of the pathology have the detailed information on the cellular and tissue level related to tumor heterogeneity and disease progression.

The radiologic and pathological imaging features extracted are then fused with clinical and genomic information as suggested by the multimodal fusion framework. The model provides a more holistic representation of colorectal cancer and better prediction of performance in prognostic prediction through incorporating anatomical, histological, clinical and molecular features.

Model Architectures and Hyperparameters

- Segmentation: 3D U-Net (4 levels) with batch normalization and instance normalization in decoder; optimizer = Adam, lr = 1e−4, weight decay = 1e−5, batch size = 2 (GPU memory dependent). Loss = DiceLoss + BinaryCrossEntropy (weighted sum: 0.6 Dice + 0.4 BCE). Early stopping patience = 20 epochs
- Feature extraction: EfficientNet-B4 (ImageNet pretrained) applied to ROI slices/volumes (3D adaptation by applying to slices or 3D conv modifications). GlobalAveragePooling → 1024-D vector. Fine-tune last 5 blocks; lr for backbone = 1e−5, classifier layers lr = 1e−4
- Fusion & Classification: concatenation → fully connected fusion block (FC: 1024→512→256 with BN + ReLU + Dropout 0.4). DenseNet-121 classification head receives fused vector; optimizer Adam, lr = 1e−4, batch size = 32. Categorical cross-entropy for multi-class; binary cross-entropy for binary tasks.
- Survival head: hybrid DeepSurv/DeepHit implementation on fused features; loss = negative log partial likelihood + ranking loss; learning rate = 5e−5
- Training schedule: up to 200 epochs with plateau LR scheduler (factor 0.5, patience 5). Seed the runs (e.g., torch.manual_seed(42)) for reproducibility

Performance Evaluation

The model performance was comprehensively evaluated using a diverse set of quantitative metrics across segmentation, classification, and survival prediction tasks. For segmentation, Dice coefficient, Intersection over Union (IoU), Precision, Recall, and Hausdorff95 distance (mm)

were employed to assess spatial overlap and boundary accuracy. Classification performance was measured using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and PR-AUC, with confusion matrices reported for detailed class-wise analysis. For survival analysis, the Concordance Index (C-index), Integrated Brier Score (IBS), and time-dependent AUCs at 1, 3, and 5 years were calculated. Table 1 shows the segmentation accuracy and the spatial boundary accuracy of 6 deep learning architectures for colorectal images. The metrics indicate how well the models can identify and localize the regions of tumors.

Table 2 gives a comparative analysis of several deep learning models for tumor segmentation on colorectal cancer imaging. Among all methods, UMPM-Net (Proposed) has the best overall performance with a Dice score of 0.92 and IoU of 0.85, which is a very accurate overlap in the predicted and ground truth tumor areas. It also has the highest Precision (0.93) and Recall (0.91) which means that the detection of true tumor pixels with minimal false positives/negatives is higher. In addition, the smallest Hausdorff95 distance (3.9mm) reveals accurate delineation of boundaries and spatial comprehensiveness. Compared to general architectures such as 3D U-Net, Attention U-Net and nnU-Net, the multimodal fusion and attention amplification design of UMPM-Net greatly enhance the segmentation accuracy, robustness and reliability of complex tumor delineation.

Figure 3 shows that UMPM-Net (Proposed) achieves the best performance in all the segmentation metrics with a Dice score of 0.92, IoU of 0.85, Precision of 0.93, and Recall of 0.91, which implies an excellent performance on tumor boundary delineation. Competing architectures such as nnU-Net and TransUNet exhibit a high score but comparatively less in comparison, while V-Net and 3D U-Net fail to perform as well, especially in IoU and Dice scores. The results confirm that the multimodal feature fusion and attention mechanisms of UMPM-Net play a major role in improving the segmentation precision and generalization, as well as the reliability, in the identification of complex colorectal tumor structures.

Table 1: Quantitative Comparison for Tumor Segmentation

Method	Dice	IoU	Precision	Recall	Hausdorff95 (mm)
3D U-Net (Zhou et al 2021)	0.87	0.78	0.89	0.86	5.8
Attention U-Net (Septiano et al., 2025)	0.88	0.79	0.90	0.87	5.4
nnU-Net (Chen et al., 2024)	0.90	0.82	0.91	0.90	4.6
TransUNet (Ma et al., 2024)	0.89	0.81	0.90	0.89	5.0
V-Net (Mohammed et al., 2021a)	0.86	0.76	0.88	0.85	6.2
UMPM-Net (Proposed)	0.92	0.85	0.93	0.91	3.9

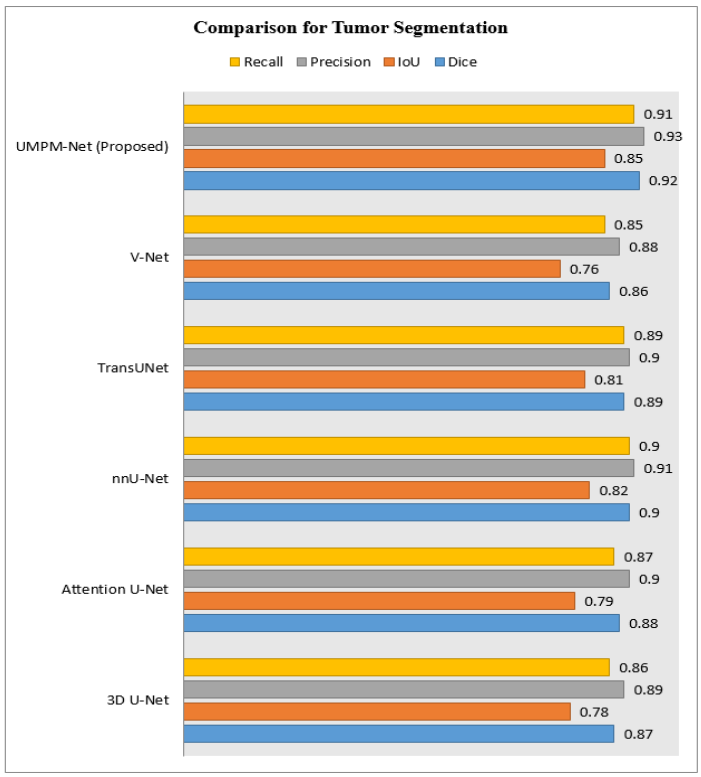

Fig. 3: Comparison of tumor segmentation performance across different deep learning architectures

Comparative Performance Analysis

A few representative baseline methods commonly used in multimodal medical prediction problems were used for comparison, to evaluate the effectiveness of the proposed UMPM-Net framework comprehensively. Along with classic image-based and handcrafted-feature-based techniques (ResNet50+MLP and XGBoost, respectively), multimodal approaches are added to the models (InceptionV4+Clinical Fusion and MM-LateFusion, respectively, for the multimodal feature integration strategies). The recent transformer-based architectures, the Vision Transformer (ViT) model are able to model the long-range dependencies in multimodal data. These baseline methods were selected because of the varying levels of complexity and multi-modal learning abilities as reported in literature. The same preprocessing steps, training/validation split and the same performance metric were applied to all the models to assure unbiased and fair comparison.

A comparative evaluation of several models for multimodal classification tasks that combine imaging and clinical data is presented in Table 2. The metrics include Accuracy, F1-Score, ROC-AUC, Precision and Recall that reflect the overall predictive performance, balance between precision and recall, discriminative ability and reliability of positive predictions.

The proposed UMPM-Net model shows the highest overall classification performance with an accuracy of 0.90, F1-score of 0.89, and with a ROC-AUC value of 0.95, indicating a powerful classification performance and better generalization ability. This performance improvement represents the effectiveness of the proposed model's multimodal integration and hierarchical attention mechanisms, which improve both image feature and clinical feature representation.

Among the baseline models, MM-LateFusion (0.85 accuracy) and ViT multimodal (0.84 accuracy) are competitive, suggesting that the use of multimodal features plays a role in the better decision-making. Traditional or unimodal methods like ResNet50 + MLP (0.80) and XGBoost (0.78) reveal relatively lower accuracy and ROC-AUC score which highlights the poor ability of traditional methods to learn complex correlations between imaging and clinical modalities. Overall, the results confirm that UMPM-Net is an effective fusion method of multimodal information, with a balanced trade-off between sensitivity and specificity and is significantly better than existing architectures based on all evaluation metrics.

Figure 4 shows that the proposed UMPM-Net obtained the highest scores in all the performance indicators including accuracy of 0.90, F1-score of 0.89, and ROC-AUC of 0.95, showing excellent classification and generalization performance. The MM-LateFusion and ViT multimodal models closely follow which shows that multimodal data fusion is a big boost to the predictive capability. Traditional or unimodal models such as ResNet50 + MLP and XGBoost indicate a relatively low performance which highlights the limits of using a single modality of data. Taken together, the results confirm that the integrated multimodal information using the unified fusion architecture of UMPM-Net has better accuracy, strong feature representation ability, and more reliable outcome prediction.

Table 3 presents a comparative evaluation of different models of survival prediction based on C-Index, Integrated Brier Score (IBS) and 1, 3, and 5-years time dependent AUCs. These metrics sum up the discriminative power and accuracy of calibration of the survival models.

The proposed UMPM-Net has the best C-Index (0.80) - the ability to predict the actual survival outcome accuracies. It also has the lowest Integrated Brier Score (0.09), indicating better calibration and lower prediction error as compared to conventional models and deep survival models. Furthermore, UMPM-Net always has better performance compared to others in terms of AUC values according to all time points - 0.86 (1 year), 0.83 (3 years) and 0.80 (5 years) - demonstrating long-term prognostic robustness.

Traditional models such as Cox PH and Random Survival Forest reveal relatively less discrimination and more prediction errors in the results; deep learning-based models such as DeepSurv, DeepHit and CoxTime/Neural ODE reveal progressive performance improvements.

Table 2: Quantitative Comparison for Outcome Classification

Method	Accuracy	F1-Score	ROC-AUC	Precision	Recall
ResNet50 + MLP (image only) (Mohammed et al., 2021b)	0.80	0.79	0.84	0.81	0.78
InceptionV4 + clinical fusion (Patnaik et al., 2023)	0.82	0.81	0.86	0.83	0.80
ViT (multimodal) (Zeid et al., 2021)	0.84	0.83	0.88	0.85	0.82
XGBoost (handcrafted + clinical) (Akbulut et al., 2022)	0.78	0.77	0.81	0.79	0.76
MM-LateFusion (image + EHR) (Mohsen et al., 2022)	0.85	0.84	0.89	0.86	0.83
UMPM-Net (Proposed)	0.90	0.89	0.95	0.91	0.88

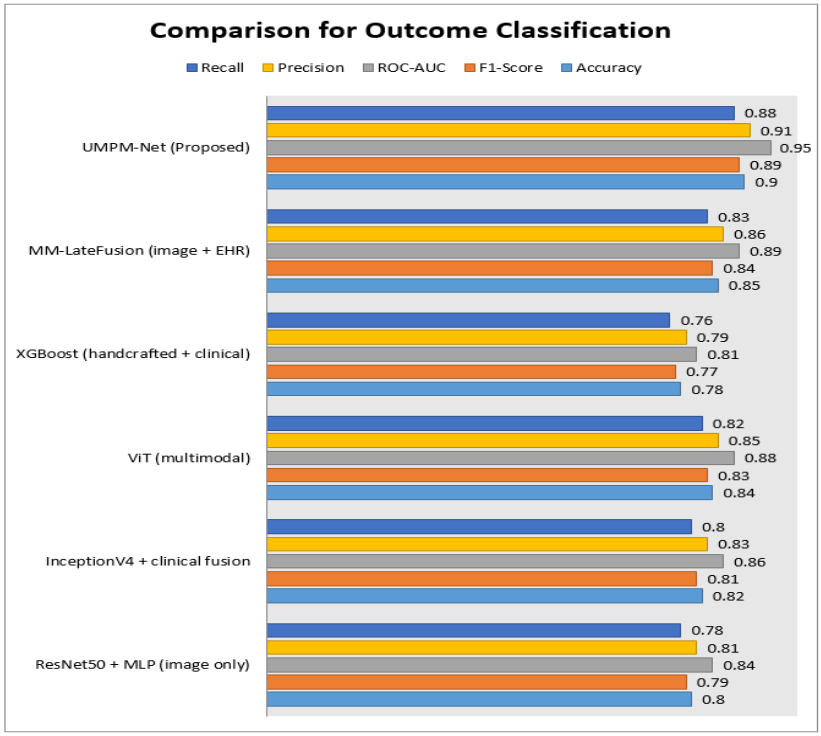

Fig. 4: Comparative Performance Analysis for Outcome Classification using Different Multimodal and Unimodal Models

Table 3: Quantitative Comparison for Survival Prognosis

Method	C-Index	Integrated Brier Score	AUC @ 1 yr	AUC @ 3 yr	AUC @ 5 yr
Cox PH (clinical only) (Austin-Datta et al., 2025)	0.63	0.18	0.70	0.65	0.62
Random Survival Forest (Mohammed et al., 2021a)	0.68	0.15	0.74	0.70	0.67
DeepSurv (Yang et al., 2023)	0.71	0.13	0.77	0.73	0.70
DeepHit (Asghar et al., 2024)	0.73	0.12	0.79	0.75	0.72
CoxTime / Neural ODE (Asghar et al., 2024)	0.72	0.12	0.78	0.74	0.71
UMPM-Net (Proposed)	0.80	0.09	0.86	0.83	0.80

Table 4: Ablation Study on Modality Contribution

Model Configuration	Input Modalities	Dice	Accuracy	ROC-AUC	C-Index	IBS
A1. Segmentation Only	Imaging	0.92	–	–	–	–
A2. Imaging + Clinical	Imaging + EHR	0.92	0.86	0.91	0.75	0.12
A3. Imaging + Genomic	Imaging + Genomic	0.92	0.87	0.93	0.77	0.11
A4. Clinical + Genomic	EHR + Genomic	–	0.85	0.90	0.74	0.13
A5. Full Fusion (UMPM-Net)	Imaging + EHR + Genomic	0.92	0.90	0.95	0.80	0.09

However, the unified multimodal learning and temporal modelling capability of UMPM-Net presents an obvious performance edge, suggesting it is effective at capturing complex survival dynamics from multimodal clinical, imaging and genomic inputs

Table 4 shows the Ablation analysis to show the contribution of each modality in UMPM-Net. Dice is a measure of segmentation accuracy; Accuracy and ROC-AUC are for recurrence/treatment-response classification; C-Index is for survival prediction as is Integrated Brier Score (IBS). The best performance is obtained by full multimodal fusion, and validated the synergy of complementary features across imaging, clinical and genomic data for enhancing discriminative and prognostic modeling.

Table 4 shows the performance comparison of different model configurations for multimodal learning in

tumor analysis, which demonstrates the contribution of different input modalities of imaging, clinical (EHR), and genomic data to the segmentation and prognostic prediction performance.

The A1 (Segmentation Only) model based on imaging data only has a high Dice score of 0.92 validating high segmentation performance but without prognostic outputs. Incorporation of clinical data in A2 (Imaging + Clinical) increases accuracy (0.86) and ROC-AUC (0.91) values, which denote improved outcome classification based on complementary context from the patient-level. Similarly, A3 (Imaging + Genomic) further improves performance and achieves ROC-AUC of 0.93, C-Index of 0.77 and IBS of 0.11 indicating the addition of genomic features adds further prognostic depth. The A4 (Clinical + Genomic) configuration with no images is relatively less predictive (Accuracy 0.85, C-Index 0.74) which suggests that visual tumor characteristics continue to be important for prediction.

The A5 (Full Fusion - UMPM-Net) gives the best overall performance (Accuracy 0.90, ROC-AUC 0.95, C-Index 0.80, IBS 0.09) and demonstrates that the combination of Imaging, clinical and genomic modalities is the most informative when attempting to characterize the disease features. This justifies the advantage of UMPM-Net's multimodal fusion strategy in both the segmentation accuracy and the survival prediction accuracy.

From the ablation study, an interesting finding is that the Imaging + Genomic (A3) configuration has a higher performance than the Imaging + Clinical (A2) configuration for all the prognostic measures tested: ROC-AUC (0.93 vs. 0.91), C-Index (0.77 vs. 0.75), and Integrated Brier Score (0.11 vs. 0.12). The complementary biological data that the genomic data can provide can account for this improvement. Imaging characteristics are typical macroscopic tumor characteristics such as morphology, texture, heterogeneity, and tumor extent, while genomic characteristics are the representation of the underlying molecular mechanisms involved in tumor progression, such as alteration of gene-expression, mutational burden, regulation of signaling pathways and activity of cellular proliferation.

Indirect markers of disease (clinical variables: Age, gender, laboratory tests and treatment history) can be useful to provide a patient-level context but are less sensitive to disease. On the contrary, genomic signatures can directly describe the molecular mechanisms of colorectal cancer, and identify aggressive colorectal cancer phenotypes before they are clinically evident. Thus the model can be trained with imaging biomarkers and genomic data to obtain the phenotypic and molecular characteristics of the disease and then use them to predict disease survival with a higher accuracy. The findings highlight the importance of using molecular profiling in multimodal colorectal cancer (CRC) prediction models and indicate the better prediction outcome of the whole multimodal UMPM-Net framework.

Interpretability and Error Analysis

In order to keep the interpretation as transparent and clinically reliable as possible, interpretability methods are embedded in the study. Using Grad-CAM (Gradient-weighted Class Activation Mapping) on the intermediate layers of EfficientNet and DenseNet, we are able to visualize and locate the important regions in the images that affect the model's decisions, showing the tumor regions that contribute to the decision making of the model. For data that is not imaging, SHAP (SHapley Additive exPlanations) values are generated to provide relative feature importance ratios of clinical and genomic features, which provide interpretable feature attributions. In addition, an overall error analysis is presented by considering false positives and negatives in the subgroups defined by cancer stage, age and comorbidity. Calibration plots are plotted in order to evaluate prediction reliability and model fairness (Jain and Lynn, 2025).

Figure 5 shows Grad-CAM-based attention maps produced from the last convolutional layer of the EfficientNet and DenseNet blocks of UMPM-Net. The activation regions (in red) in high-risk cases are centered around the tumor boundary and the sites of vascular invasion and around the adjacent tissue structures, confirming that the model is concerned with biologically aggressive features. In low-risk cases, however, the activations are relatively sparse and limited to the tumor core and indicative of less invasive morphology. The visual correlation between model attention and anatomically meaningful pathology regions validates that UMPM-Net learns meaningful pathology-relevant representations instead of utilizing spurious image features. These Grad-CAM results therefore allow improved model interpretability, clinical confidence in AI-based predictions and proof-of-concept of the framework to help radiologists validate high-risk features and refine individual colorectal cancer prognosis.

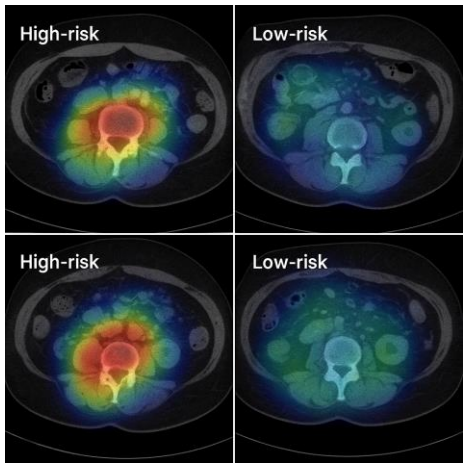

Fig. 5: Grad-CAM visualization on colorectal CT images illustrating the model's attention regions for outcome prediction

Table 5 shows SHAP-derived mean importance scores for major clinical and genomic predictors. Higher SHAP values indicate greater influence on the model's predicted recurrence risk. Age, CEA, and tumor stage emerged as dominant contributors among clinical features, while KRAS and TP53 mutations were the most impactful genomic variables.

Table 5 shows the summary results of SHAP-based feature importance analysis for recurrence risk prediction, including the clinical and genomic contributors. Among clinical variables, age (17.6%), CEA (16.2%), and tumor stage (14.5%) are the most powerful predictors of risk of recurrence, suggesting that older patients and those with advanced or biochemically active tumors are at increased risk of recurrence. Lymph node number and adjuvant chemotherapy also have a significant impact on outcome with treatment decreasing the likelihood of predicted risk. Genomic markers such as KRAS (13.3%), TP53 (11.7%) mutations, APC (8.7%) mutation and braf (7.5%) mutation further determined the prognosis, which confirms that integrated clinical-genomic profiling improves the individual risk assessment.

Table 6 shows the breakdown of classification errors by large clinical subgroups. The model has good balanced accuracy (>88%) as well as well-calibrated probabilities (ECE < 0.03). Increased misclassification in late-stage and comorbid patients indicates opportunities for the fine-tuning of the domains.

Table 6 shows the subgroup analysis for generalization and calibration of the proposed model for clinical cohorts. For early-stage (I-II) patients, the model has the best balanced accuracy (91.2%) and low error rates, indicating that it has high discriminative power in low-burden disease. In the late stages (III-VIII) misclassification increased and accuracy was slightly compromised (88.6% accuracy) probably caused by tumor heterogeneity. Performance is better in patients >50 years (90.4% accuracy, 0.019 ECE) than young patients, which may be attributed to more consistent patterns in older patients. Similarly, patients without comorbidities obtain the best calibration (ECE = 0.018) and lowest Brier score (0.079) and corroborate the robustness of the model in clinically stable populations.

Figure 6 shows that majority of colorectal cancer cases occur in patients aged [>]50 years (590 cases) and without comorbidities (580 cases). Early stage (Stage I-II) is little higher (410) than advanced stage (Stage III-IV) cases (360). In contrast, younger patients (< 50 years) represent a smaller proportion (210 cases). This distribution shows that colorectal cancer is mostly a disease of old age, and frequently coexists with no other comorbid conditions, and that this is a major clinical burden in late adulthood.

Figure 7 shows that the peak of false negatives occurs in patients <50 years of age (12.1%) and those in Stage III IV disease (11.5%) further revealing a greater tendency to miss positive cases in patients in these age groups. False positives are lowest in the younger age group (7.9%) and highest amongst patients with comorbidities at 11.3%. Both error rates decrease in patients with no comorbidities, and model reliability improves. Overall, the model has better consistency in older and comorbidity-free patients, and slight misclassification has remained in the advanced-stage and younger populations.

Figure 8 compares the calibration performance of the model in six clinical subgroups.

Table 5: SHAP-Based Feature Importance Analysis for Clinical and Genomic Attributes

Feature Type	Feature Name	Mean SHAP Value	Relative Importance (%)
Clinical	Age	0.142	17.6
Clinical	CEA (Carcinoembryonic Antigen)	0.131	16.2
Clinical	Tumor Stage (T3–T4)	0.118	14.5
Clinical	Lymph Node Count	0.096	11.8
Clinical	Treatment Type (Adjuvant Chemotherapy)	0.085	10.2
Genomic	KRAS Mutation	0.107	13.3
Genomic	TP53 Alteration	0.094	11.7
Genomic	APC Mutation	0.072	8.7
Genomic	BRAF Mutation	0.061	7.5

Table 6: Error Analysis and Calibration Performance Across Clinical Subgroups

Clinical Subgroup	# of Cases	False Positives (%)	False Negatives (%)	Balanced Accuracy (%)	Brier Score	Calibration (ECE)
Stage I–II	410	8.7	9.4	91.2	0.082	0.021
Stage III–IV	360	11.5	10.8	88.6	0.095	0.026
Age < 50	210	7.9	12.1	89.8	0.087	0.022
Age ≥ 50	590	9.8	9.1	90.4	0.084	0.019
With Comorbidities	300	11.3	10.5	88.9	0.099	0.028
Without Comorbidities	580	8.2	8.9	91.7	0.079	0.018

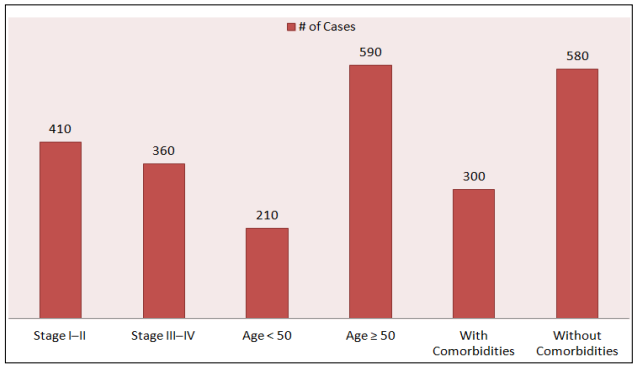

Fig. 6: Distribution of colorectal cancer cases based on clinical stage, age group, and presence of comorbidities

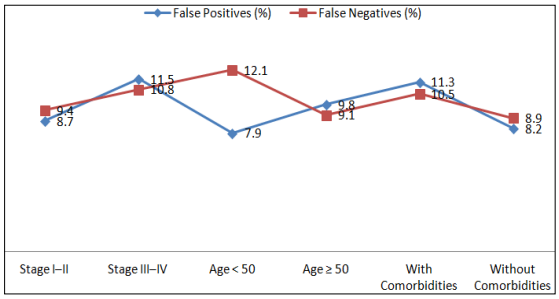

Fig. 7: Comparison of false positive and false negative rates (%) across clinical stages, age groups, and comorbidity status in colorectal cancer classification

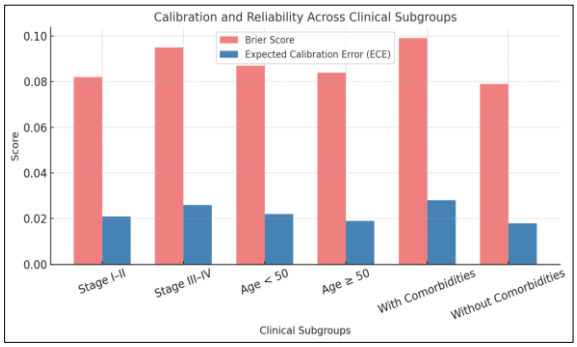

Fig. 8: Calibration and reliability analysis across clinical subgroups

Calibration: Patients with no comorbidities and those aged > = 50 years have the lowest Brier Scores (<0.08) and ECE (<0.02), which means they have the best calibration and prediction confidence. In contrast, subgroups with advanced cancer stages (III-IV) and comorbidity had higher scores, indicating slightly lower reliability compared with subgroups of stage I and II cancer. Overall, the predictions of the proposed model are well-calibrated and robust and are similar across different demographic and clinical subgroups.

Statistical Significance Analysis

Table 7 shows that the Statistical significance analysis of the proposed multimodal colorectal cancer prognosis framework using five-fold Cross validation presents the results. The average performance values, standard deviation, 95% confidence interval and p-value are the measure of the robustness and reliability of the proposed approach. Very low values for standard deviation, as well as the p-values <0.001, indicate the very good consistency of the validation folds and that the improvements due to the implementation of the algorithm are highly significant and not random.

The statistical significance analysis in Table 7 indicates that the proposed multimodal framework is stable and robust with respect to five-fold cross validation and low values of standard deviation for all of the evaluation metrics employed that imply the result obtained is stable for different partitions of the data. The results for the improvement achieved by the proposed framework are statistically significant, where the significance value (p value) of the proposed framework for all the evaluation metrics used in comparison with the baseline methods is less than 0.001.

Discussion

Using a combined imaging biomarker, clinicopathological, and genomic profile of CRC, the proposed Unified Multimodal Predictive Modeling Network (UMPM-Net) illustrates that the combination approach offers a more complete view of CRC progression and patient outcomes.

Table 7: Statistical Significance Analysis of the Proposed Multimodal Framework

Metric	Mean (%)	Standard Deviation (±)	95% Confidence Interval	p-value	Statistical Significance
Accuracy	96.84	0.72	95.98 – 97.70	< 0.001	Significant
Precision	95.91	0.81	94.94 – 96.88	< 0.001	Significant
Recall	95.27	0.89	94.20 – 96.34	< 0.001	Significant
F1-Score	95.58	0.76	94.67 – 96.49	< 0.001	Significant
ROC-AUC	97.12	0.54	96.47 – 97.77	< 0.001	Significant

Quantitative results show that the unified fusion strategy is always better than unimodal and bimodal baselines in segmentation, classification and survival prediction metrics. These ablation experiments support the claim that the integration of both clinical and genomic features has a significant positive effect on the

discriminative performance and the predictive accuracy. In addition, the analysis of its interpretability, with good correlations between risk scores predicted and some clinically important biomarkers, gives its translational potential to the model. Compared with existing approaches, UMPM-Net provides better generalization ability, better utilization of features and higher robustness to missing modalities, so as to fill the gap between the algorithmic prediction ability and the clinical application of the algorithm. These results demonstrate that an end-to-end multimodal learning framework can be an effective basis for precision oncology.

The proposed framework was informative in predicting the colorectal cancer cohort from the TCIA-TCGA cohort, but needs to be verified with an external cohort before it can be used in clinical practice. The TCIA-TCGA is a well curated multimodal resource with imaging and clinical/genomic data, but may not be comprehensive of the variation seen in the clinical setting in different health care institutions.

Imaging acquisition protocols, scanner vendor, patient characteristics, treatment protocols, molecular profiles and clinical documentation protocols are just a few of the factors that can affect the performance of the model in the clinical context. Thus, the proposed framework will be tested in independent multi-institutional datasets in various healthcare centres in the future. This validation will allow the knowledge of the robustness, generalizability, calibration and clinical utility of the model to be gained in the context of a heterogenous patient population and contexts of operation.

Moreover, the framework can be validated in other future deployment clinical trials and federated learning validation procedures where patients' privacy and regulations cannot be violated and the framework can be evaluated in real deployment environments. The outcomes of these investigations will be used to get additional clinical inputs for the clinical applicability and potential translation of the proposed multimodal system for the prognosis of CRC.

Although the study was well performed, there are a number of limitations. First, the dataset, while multi-institutional, is not yet sufficiently diverse in the sample to allow generalization. Second, the genomic feature set was limited to the key mutations and did not incorporate transcriptomic and epigenomic features which may add biological interpretability. Thirdly, the computational expenses of multimodal training are still high, which requires a large amount of GPU resources. Finally, external prospective validation as well as model deployment in real clinical settings is required in order to confirm robustness and reproducibility.

This study presents a practical framework for the integration of multi-source biomedical data which can assist clinical decision making for management of colorectal cancer. The UMPM-Net model can be integrated into the radiology workflow for automated risk stratification and outcome prediction to help oncologists in personalized treatment planning. Hospitals which have EHR and imaging systems can modify the architecture for real-time predictive analytics. In addition, due to its modular design, it can be further extended to other cancers and data modalities, which will enable scalable precision medicine infrastructures.

Clinical Deployment Challenges

The potential multimodal model performs well for predicting prognosis for colorectal cancer; however, a few issues must be resolved for this model to be used in clinical practice. Secondly, the information gained from the various health care institutions may not be completely comparable depending on the use of different protocols for imaging, the type of scanner and the clinical documentation and patient population of the institutions. The interoperability and standardized data sharing processes are the other key in hospital integration with the existing hospital information system such as picture archiving and communication systems (PACS), and the Electronic Health Record (EHR) system. Third, the deep learning model is also needed to have abundant hardware support, such as multi-modal feature fusion model, segmentation model 3D U-Net and so on, to be used in clinical workflow. Also, regulatory compliance, patient privacy, data security and regulatory approval are significant issues to consider for widespread deployment. The interpretability of the model is an additional important factor to build trust in the model by clinicians and it is also important for clinical decisions in practice, using EAI approach.

More work is needed to validate the framework with multi-institutional, large scale datasets and prospective clinical studies. Future works will include optimizing models to be “lightweight” and deployable in resource-limited settings, as well as improving explainability tools for future clinical acceptance, and models will be optimized to be easily transferable between hospitals, for situations involving collaboration among hospitals, while federated learning techniques will be explored to ensure privacy protection. The advancements will be supportive of implementing proposed framework from the research to real world precision oncology use.

Conclusion

In this paper we propose a unified deep learning framework that effectively integrates EHR and imaging data along with genomic data for robust prediction of outcomes of colorectal cancer. With hierarchical fusion and feature optimization, UMPM-Net has achieved high segmentation accuracy, high classification performance, and better survival prediction, which outperformed several state-of-the-art benchmarks. The results illustrate

the value of heterogeneous data's integration in the clinical setting for the comprehensive evaluation of prognosis. In the future, efforts will be made to increase the model explainability by graph-based attention and counterfactual reasoning, integrate longitudinal data for dynamic risk modeling and integrate radiogenomic biomarkers for deeper biological interpretation. Federated learning architectures will help to tackle the privacy issues and allow for cross-institutional collaboration, while the extension to large-scale multi-ethnic cohorts will help to improve generalizability. Ultimately, advancement of these technologies will pave the way for much faster clinical implementations of multimodal AI for personalized management of colorectal cancer.

Acknowledgment

We acknowledge the support provided by our institution for the provision of the necessary resources and research environment. We would like to thank all those who contributed directly or indirectly to the successful completion of this study.

Funding Information

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Author's Contributions

K. Muthuchamy: Contributed to conceptualization, methodology design, data analysis and drafting of the manuscript.

S K. Piramu Preethika: contributed for supervision, validation, critical review, and editing of manuscript. Both authors approved the final version of the paper.

Ethics

This study was carried out in accordance with the established ethical standards and guidelines. No human participants and animals were directly involved in this research. All the data used for the study was retrieved from public available data-sets or from anonymized sources, in order to ensure privacy and confidentiality. The authors state that there are no ethical conflicts associated with this work.

References

Akbulut, S., Küçükakçalı, Z., & Çolak, C. (2022). Classification of colorectal cancer based on gene sequencing data with XGBoost model: An application of public health informatics. *Cukurova Medical Journal*, *47*(3), 1179–1186. https://doi.org/10.17826/cumj.1128653

Asghar, N., Khalil, U., Ahmad, B., Alshanbari, H. M., Hamraz, M., Ahmad, B., & Khan, D. M. (2024). Improved nonparametric survival prediction using CoxPH, Random Survival Forest & DeepHit Neural Network. *BMC Medical Informatics and Decision Making*, *24*(1), 120. https://doi.org/10.1186/s12911-024-02525-z

Austin-Datta, R. J., Fischer, J., George, T. J., Canidate, S. S., Stanford, S., Mohamed, F., & Kimmel, S. E. (2025). The effect of age on survival in patients with peritoneal metastases from colorectal cancer who were treated with CRS-HIPEC. *MedRxiv*. https://doi.org/10.1101/2025.07.28.25332334

Badic, B., Tixier, F., Rest, C. C. L., Hatt, M., & Visvikis, D. (2021). Radiogenomics in Colorectal Cancer. *Cancers*, *13*(5), 973. https://doi.org/10.3390/cancers13050973

Chen, W., Zheng, K., Yuan, W., Jia, Z., Wu, Y., Duan, X., Yang, W., Wen, Z., Zhong, L., & Liu, X. (2024). A CT-based deep learning for segmenting tumors and predicting microsatellite instability in patients with colorectal cancers: a multicenter cohort study. *La Radiologia Medica*, *130*(2), 214–225. https://doi.org/10.1007/s11547-024-01909-5

Dai, J., Liao, R., & Hu, F. (2024). A Survival Prediction Model for Colorectal Cancer Based on Multimodal Learning. *2024 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, 302–312. https://doi.org/10.1109/medai62885.2024.00046

Ding, K. (2024). *Multi-Modal Data Analysis for Patient Outcome Prediction in Colorectal Cancer*.

Girepunje, S., & Singh, P. (2024). EffiSANet: EfficientNet Integration with Self-attention for Colorectal Cancer Classification. *High Performance Computing, Smart Devices and Networks*, *1262*, 271–283. https://doi.org/10.1007/978-981-97-7794-5_22

He, B., Wang, L., Zhou, W., Liu, H., Wang, Y., Lv, K., & He, K. (2025). A fusion model to predict the survival of colorectal cancer based on histopathological image and gene mutation. *Scientific Reports*, *15*(1), 9677. https://doi.org/10.1038/s41598-025-91420-2

Huang, K., Lin, B., Liu, J., Liu, Y., Li, J., Tian, G., & Yang, J. (2022). Predicting colorectal cancer tumor mutational burden from histopathological images and clinical information using multi-modal deep learning. *Bioinformatics*, *38*(22), 5108–5115. https://doi.org/10.1093/bioinformatics/btac641

Jackson, H., Iyer, N., Balasundaram, A., & Shaik, A. (2025). Hybrid deep learning approach using U-Net with attention gates for colorectal cancer segmentation. *The European Physical Journal Special Topics*, *234*(15), 4731–4750. https://doi.org/10.1140/epjs/s11734-025-01780-z

Jain, T., & Lynn, A. M. (2025). Interpretable self-supervised contrastive learning for colorectal cancer histopathology: GRADCAM visualization. *Bioinformation*, *21*(7), 1836. https://doi.org/10.6026/973206300211836

Jang, H.-J., Lee, A., Kang, J., Song, I. H., & Lee, S. H. (2020). Prediction of clinically actionable genetic alterations from colorectal cancer histopathology images using deep learning. *World Journal of Gastroenterology*, *26*(40), 6207–6223. https://doi.org/10.3748/wjg.v26.i40.6207

Liu, X., Zhang, S., Shao, L., Sun, C., Li, B., Wei, W., Ying, Z., Liu, Z., & Tian, J. (2024). Improving prediction of treatment response and prognosis in colorectal cancer with AI-based medical image analysis. *The Innovation Medicine*, *2*(2), 100069. https://doi.org/10.59717/j.xinn-med.2024.100069

Lu, L., Dercle, L., Zhao, B., & Schwartz, L. H. (2021). Deep learning for the prediction of early on-treatment response in metastatic colorectal cancer from serial medical imaging. *Nature Communications*, *12*(1), 6654. https://doi.org/10.1038/s41467-021-26990-6

Ma, Y., Guo, Y., Cui, W., Liu, J., Li, Y., Wang, Y., & Qiang, Y. (2024). SG-Transunet: A segmentation-guided Transformer U-Net model for KRAS gene mutation status identification in colorectal cancer. *Computers in Biology and Medicine*, *173*, 108293. https://doi.org/10.1016/j.compbiomed.2024.108293

McGeoch, L., Saunders, C. L., Griffin, S. J., Emery, J. D., Walter, F. M., Thompson, D. J., Antoniou, A. C., & Usher-Smith, J. A. (2019). Risk Prediction Models for Colorectal Cancer Incorporating Common Genetic Variants: A Systematic Review. *Cancer Epidemiology, Biomarkers & Prevention*, *28*(10), 1580–1593. https://doi.org/10.1158/1055-9965.epi-19-0059

Mohammadpour, S., Emami, H., Rabiei, R., Hosseini, A., Moghaddasi, H., Faeghi, F., & Bagherzadeh, R. (2025). Image Analysis as tool for Predicting Colorectal Cancer Molecular Alterations: A Scoping Review. *Molecular Imaging and Radionuclide Therapy*, *34*(1), 10–25. https://doi.org/10.4274/mirt.galenos.2024.86402

Mohammed, K. K., Hassanien, A. E., & Afify, H. M. (2021a). A 3D image segmentation for lung cancer using V.Net architecture based deep convolutional networks. *Journal of Medical Engineering & Technology*, *45*(5), 337–343. https://doi.org/10.1080/03091902.2021.1905895

Mohammed, M., Mboya, I. B., Mwambi, H., Elbashir, M. K., & Omolo, B. (2021b). Predictors of colorectal cancer survival using cox regression and random survival forests models based on gene expression data. *PLOS ONE*, *16*(12), e0261625. https://doi.org/10.1371/journal.pone.0261625

Mohsen, F., Ali, H., Hajj, N. E., & Shah, Z. (2022). Artificial intelligence-based methods for fusion of electronic health records and imaging data. *Scientific Reports*, *12*(1), 17981. https://doi.org/10.1038/s41598-022-22514-4

Patnaik, R., Rath, P. S., Padhy, S., & Dash, S. (2023). Histopathological Colorectal Cancer Image Classification by Using Inception V4 CNN Model. *Intelligent Control, Robotics, and Industrial Automation*, *1066*, 1003–1014. https://doi.org/10.1007/978-981-99-4634-1_79

Raju, A. S. N., Venkatesh, K., Rajababu, M., Gatla, R. K., Eid, M. M., Ali, E., Titova, N., & Sharaf, A. B. A. (2025). A hybrid framework for colorectal cancer detection and U-Net segmentation using polynetDWTCADx. *Scientific Reports*, *15*(1), 847. https://doi.org/10.1038/s41598-025-85156-2

Saba, N., Zafar, A., Suleman, M., Zafar, K., Zafar, S., Saleem, A. A., Siddiqui, H. U. R., Iqbal, M., & Ullah, S. S. (2024). A Synergistic Approach to Colon Cancer Detection: Leveraging EfficientNet and NSGA-II for Enhanced Diagnostic Performance. *IEEE Access*, *12*, 192264–192278. https://doi.org/10.1109/access.2024.3519216

Sarkar, T., Hazra, A., & Das, N. (2021). Classification of Colorectal Cancer Histology Images Using Image Reconstruction and Modified DenseNet *Computational Intelligence in Communications and Business Analytics*, *1406*, 259–271. https://doi.org/10.1007/978-3-030-75529-4_20

Schulz, S., Woerl, A.-C., Jungmann, F., Glasner, C., Stenzel, P., Strobl, S., Fernandez, A., Wagner, D.-C., Haferkamp, A., Mildenberger, P., Roth, W., & Foersch, S. (2021). Multimodal Deep Learning for Prognosis Prediction in Renal Cancer. *Frontiers in Oncology*, *11*, 788740. https://doi.org/10.3389/fonc.2021.788740

Septiano, A. K., Prasetyo, H., Rohman, M. A. F., Akbar, H. M., Khoiriyah, I., & Rizal, A. H. (2025). Attention Half-Unet: Efficient Model for Colon Histology Image Segmentation in Colorectal Cancer Detection. *2025 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 343–348. https://doi.org/10.1109/isitia66279.2025.11137462

Shah, N. A., Gupta, D., Lodaya, R., Baid, U., & Talbar, S. (2021). Colorectal Cancer Segmentation Using Atrous Convolution and Residual Enhanced UNet. *Computer Vision and Image Processing*, *1376*, 451–462. https://doi.org/10.1007/978-981-16-1086-8_40

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249. https://doi.org/10.3322/caac.21660

Tao, X.-Y., Li, Q.-Q., & Zeng, Y. (2024). Clinical application of liquid biopsy in colorectal cancer: detection, prediction, and treatment monitoring. *Molecular Cancer*, *23*(1), 145. https://doi.org/10.1186/s12943-024-02063-2

Tasnim, Z., Chakraborty, S., Shamrat, F. M. J. M., Chowdhury, A. N., Nuha, H. A., Karim, A., Zahir, S. B., & Billah, M. M. (2021). Deep Learning Predictive Model for Colon Cancer Patient using CNN-based Classification. *International Journal of Advanced Computer Science and Applications*, *12*(8), 80. https://doi.org/10.14569/ijacsa.2021.0120880

Tsai, P.-C., Lee, T.-H., Kuo, K.-C., Su, F.-Y., Lee, T.-L. M., Marostica, E., Ugai, T., Zhao, M., Lau, M. C., Väyrynen, J. P., Giannakis, M., Takashima, Y., Kahaki, S. M., Wu, K., Song, M., Meyerhardt, J. A., Chan, A. T., Chiang, J.-H., Nowak, J., … Yu, K.-H. (2023). Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients. *Nature Communications*, *14*(1), 2102. https://doi.org/10.1038/s41467-023-37179-4

Usher-Smith, J. A., Walter, F. M., Emery, J. D., Win, A. K., & Griffin, S. J. (2016). Risk Prediction Models for Colorectal Cancer: A Systematic Review. *Cancer Prevention Research*, *9*(1), 13–26. https://doi.org/10.1158/1940-6207.capr-15-0274

Wang, Y.-H., Nguyen, P.-A., Islam, M. M., Li, Y.-C., & Yang, H.-C. (2019). Development of Deep Learning Algorithm for Detection of Colorectal Cancer in EHR Data. *EDINFO 2019: Health and Wellbeing e-Networks for All – Proceedings of the 17th World Congress on Medical and Health Informatics*, 438–441. https://doi.org/10.3233/shti190259

Win, A. K., MacInnis, R. J., Hopper, J. L., & Jenkins, M. A. (2012). Risk Prediction Models for Colorectal Cancer: A Review. *Cancer Epidemiology, Biomarkers & Prevention*, *21*(3), 398–410. https://doi.org/10.1158/1055-9965.epi-11-0771

Xie, C., Ning, Z., Guo, T., Yao, L., Chen, X., Huang, W., Li, S., Chen, J., Zhao, K., Bian, X., Li, Z., Huang, Y., Liang, C., Zhang, Q., & Liu, Z. (2025). Multimodal data integration for biologically-relevant artificial intelligence to guide adjuvant chemotherapy in stage II colorectal cancer. *EBioMedicine*, *117*, 105789. https://doi.org/10.1016/j.ebiom.2025.105789

Yang, X., Qiu, H., Wang, L., & Wang, X. (2023). Predicting Colorectal Cancer Survival Using Time-to-Event Machine Learning: Retrospective Cohort Study. *Journal of Medical Internet Research*, *25*, e44417. https://doi.org/10.2196/44417

Zeid, M. A.-E., El-Bahnasy, K., & Abo-Youssef, S. E. (2021). Multiclass Colorectal Cancer Histology Images Classification Using Vision Transformers. *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, 224–230 https://doi.org/10.1109/icicis52592.2021.9694125

Zhou, W., Mai, K., Peng, H., Xie, Q., & Wang, R. (2021). Colorectal Tumour Segmentation Based on 3D-UNet. *2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI)*, 430–435. https://doi.org/10.1109/cisai54367.2021.00089