

# BIOFUSIONNET: DEEP LEARNING-BASED SURVIVAL RISK STRATIFICATION IN ER+ BREAST CANCER THROUGH MULTIFEATURE AND MULTIMODAL DATA FUSION

Raktim Kumar Mondol<sup>1</sup>, Ewan K.A. Millar<sup>2</sup>, Arcot Sowmya<sup>1</sup>, and Erik Meijering<sup>1,\*</sup>

<sup>1</sup>School of Computer Science and Engineering, University of New South Wales, Sydney, Australia

<sup>2</sup>Department of Anatomical Pathology, NSW Health Pathology, St. George Hospital

\*Correspondence: erik.meijering@unsw.edu.au

## ABSTRACT

Breast cancer is a significant health concern affecting millions of women worldwide. Accurate survival risk stratification plays a crucial role in guiding personalised treatment decisions and improving patient outcomes. Here we present BioFusionNet, a deep learning framework that fuses image-derived features with genetic and clinical data to obtain a holistic profile and achieve survival risk stratification of ER+ breast cancer patients. We employ multiple self-supervised feature extractors (DINO and MoCoV3) pretrained on histopathological patches to capture detailed image features. These features are then fused by a variational autoencoder and fed to a self-attention network generating patient-level features. A co-dual-cross-attention mechanism combines the histopathological features with genetic data, enabling the model to capture the interplay between them. Additionally, clinical data is incorporated using a feed-forward network, further enhancing predictive performance and achieving comprehensive multimodal feature integration. Furthermore, we introduce a weighted Cox loss function, specifically designed to handle imbalanced survival data, which is a common challenge. Our model achieves a mean concordance index of 0.77 and a time-dependent area under the curve of 0.84, outperforming state-of-the-art methods. It predicts risk (high versus low) with prognostic significance for overall survival in univariate analysis ( $HR=2.99$ , 95% CI: 1.88–4.78,  $p<0.005$ ), and maintains independent significance in multivariate analysis incorporating standard clinicopathological variables ( $HR=2.91$ , 95% CI: 1.80–4.68,  $p<0.005$ ).

**Keywords** Multimodal Fusion · Breast Cancer · Whole Slide Images · Deep Neural Network · Survival Prediction

## 1 INTRODUCTION

Breast cancer poses a significant global health concern, with a high incidence rate and substantial impact on morbidity and mortality [1, 2]. The incidence of breast cancer varies across different regions and populations, with higher rates observed in developed countries [1]. The prevalence of breast cancer in Australia, affecting 1 in 8 women up to the age of 85, is a cause for concern due to its rising incidence rate over the past decade [3]. This trend highlights the crucial need for accurately predicting survival risks to identify high-risk patients who may benefit from more intensive treatment or monitoring, thereby potentially improving outcomes [2].

In breast cancer, the estrogen receptor (ER) status plays a critical role in determining treatment strategies and predicting patient prognosis. ER+ breast cancer, which includes Luminal A and Luminal B subtypes, is distinguished by the presence of ERs on cancer cells, making it responsive to hormonal therapies such as tamoxifen [4]. While Luminal A tumours are usually low-grade with a favorable prognosis ( $Ki-67<14\%$ ), Luminal B tumours are usually higher grade and pose a higher recurrence risk and worse outcome ( $Ki-67\geq14\%$ ) [4, 5]. The inherent heterogeneity of breast cancer poses challenges for prediction of prognosis and treatment decisions, particularly in post-menopausal ER+ breast cancer, with previous studies reporting conflicting results on the survival difference between Luminal A and B metastatic breast cancer patients [6, 7]. A common critical clinical dilemma is the

selection of those early ER+ breast cancer patients at high risk of recurrence who may benefit from the addition of chemotherapy to endocrine therapy. Therefore, accurate survival risk prediction models specifically tailored for ER+ breast cancer are essential for personalised treatment decisions.

Traditional methods for survival risk prediction often rely on clinicopathological risk factors (such as age, tumour size, grade, lymph node metastasis and clinical stage), which may fail to fully capture the complex biology of cancer [4, 7–10]. To address this issue, molecular markers and gene expression profiles have been identified as potential prognostic factors that provide valuable insights into tumour biology and potential therapeutic targets [11–15]. Over the past decade, the integration of genomic testing into treatment decision-making processes has been enhanced by the utilisation of several commercial gene panels, such as Prosigna/PAM50, OncotypeDx and Endopredict among others. In addition, histopathological imaging, which offers in-depth insights into the cellular and tissue characteristics of tumours, plays a vital role in both the diagnosis and prognosis of breast cancer [16]. However, to address the varied nature of the disease effectively, it is essential to consider all available data modalities. Therefore, integrating imaging, genetic and clinicopathological information into a single risk prediction model could potentially enhance risk prognostication in the clinic [9]. In this study, we propose a novel multimodal survival risk prediction model that significantly enhances the prognosis of ER+

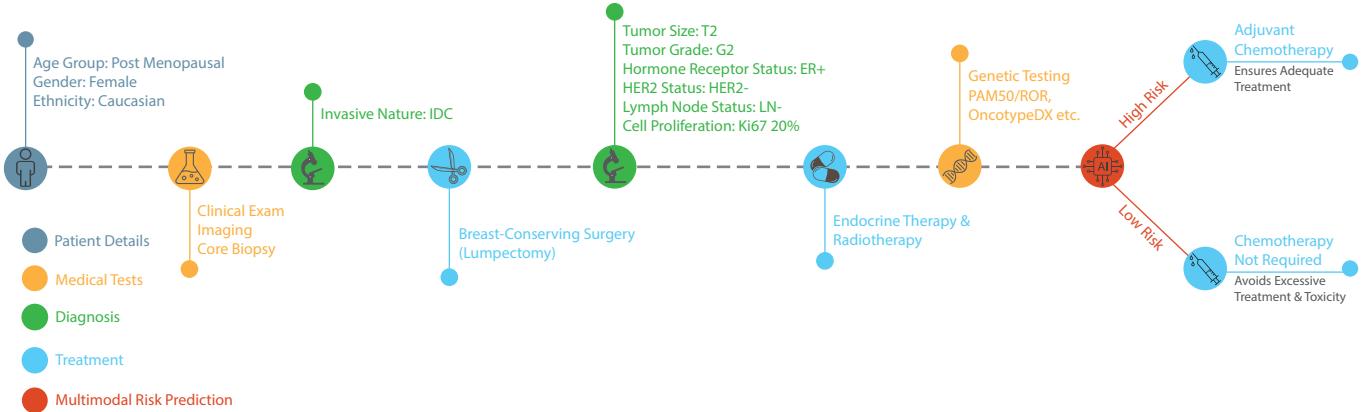


Figure 1: Illustration of the clinical management pathway in treating breast cancer patients. This example concerns a postmenopausal patient who has been diagnosed with breast cancer, specifically invasive ductal carcinoma (IDC). The process begins with an initial diagnosis through clinical examination, imaging and core biopsy. Following this, surgery is performed to completely excise the tumour, and postoperative tumour histopathological classification is performed to assess key factors including tumour size (e.g. T2), grade (e.g. G2), hormone receptor status (e.g. ER+), HER2 status (e.g. HER2-), lymph node status (e.g. LN-), and proliferation index (e.g. Ki67 20%). Subsequent treatments may include hormone therapy and radiotherapy. Additional molecular tests, like genetic testing, are utilised to determine specific cancer molecular subtypes and further assess risk of recurrence. The proposed final step in this pathway is the application of our BioFusionNet model. This model combines tumour characteristics, pathology and genetic testing to determine high and low risk patients, thereby guiding personalised treatment decisions and efficiently preventing both under-treatment and over-treatment. For example, low-risk patients might undergo lumpectomy with hormone therapy and radiotherapy, whereas high-risk patients are advised to have chemotherapy in addition to these treatments. Whilst this pathway mirrors current clinical practice, our study streamlines the integration of all available critical data to derive an automated single risk prediction score.

breast cancer by integrating histopathology images, genetic profiles, and clinical data. By combining these data modalities, our model captures the complex interplay between cellular, molecular, and clinical factors, thereby improving its predictive accuracy. Moreover, we introduce a weighted Cox loss function specifically designed to handle imbalanced survival data, further enhancing the model's performance. The model is thoroughly evaluated using metrics such as the concordance index (C-index) and time-dependent area under the curve (AUC) score, and its performance is compared to existing methods to establish its efficacy. The model's ability to accurately predict survival risks and categorise patients into distinct risk groups has the potential to inform personalised treatment decisions and improve patient outcomes. Finally, we provide explainable analyses to elucidate the influence of different genes and clinical factors on risk prediction, offering valuable insights into the underlying predictive mechanisms. The main contributions of this study are as follows:

1. A novel multimodal survival risk prediction model integrating histopathology images, genetic profiles, and clinical data for ER+ breast cancer prognosis.
2. A weighted Cox loss function designed to handle imbalanced survival data, improving predictive accuracy.
3. Explainable analyses providing insights into the influence of genes and clinical factors on risk prediction.

## 2 BACKGROUND

Cancer risk prediction is of paramount importance due to its potential for guiding personalised screening, prevention and treatment strategies, ultimately leading to improved patient out-

comes and reduced mortality (Fig. 1). This approach is especially vital in the context of ER+ breast cancer, where accurate risk prediction is essential for identifying individuals at higher recurrence risk. These patients may benefit from more aggressive treatments like chemotherapy [17–20]. On the other hand, risk models are equally critical in recognising lower-risk patients, potentially sparing them from unnecessary treatments and their side effects [21]. Online algorithms such as Predict<sup>1</sup> and Adjuvant<sup>2</sup> are used clinically to estimate the risk of recurrence and the benefit of adding chemotherapy to endocrine therapy, which is a major treatment dilemma.

Developments in deep learning, such as the Cox proportional hazards deep neural network, have revolutionised cancer research by improving survival data modeling and treatment recommendation systems [22]. Multimodal data fusion, which combines information from diverse sources such as imaging, genomics and clinical data, has gained attention in cancer research due to its ability to provide a comprehensive understanding of the disease and improve predictive outcomes [23–25]. Cross-attention transformer mechanisms that integrate histopathological images and genomic data capture complementary information from different modalities, leading to improved survival prediction [26].

In this evolving landscape, models such as MultiDeepCox-SC [27], MCAT [28], MultiSurv[29], HFBSurv [30], Pathomic Fusion [31], and TransSurv [32] exemplify significant progress in multimodal analysis. These models harness unique strategies to integrate diverse data types for enhanced survival prediction. MultiDeepCox-SC combines histopathological image-derived

<sup>1</sup><https://breast.predict.nhs.uk/tool>

<sup>2</sup><https://oncoassist.com/adjuvant-tools/>

risk scores with clinical data and gene expression through the Cox proportional hazards (CoxPH) model, providing a detailed risk assessment. MCAT utilises a genomic-guided co-attention layer to map relationships between whole slide images and genomic features, enhancing interpretability in computational pathology. MultiSurv simplifies the integration of multimodal data by merging feature vectors into a unified representation for predicting survival probabilities. In contrast, HFBSurv applies an hierarchical framework with attentional factorised bilinear modules, systematically processing information from lower to higher complexity levels. Pathomic Fusion adopts a gating-based attention mechanism, effectively filtering out noise to focus on salient features across modalities. Finally, TransSurv employs cross-attention transformers to combine histopathological and genomic data, capturing complementary information that significantly improves predictive accuracy.

Despite these advances, the integration of multimodal data, marked by its inherent heterogeneity and dimensional variability, remains a significant challenge, highlighting the need for advanced methodologies to effectively integrate and leverage diverse data sources for robust survival risk assessment [33–35]. To address this, BioFusionNet introduces a unique strategy by combining self-supervised learning models, specifically DINO and MoCoV3, for feature extraction. This method is further enhanced by a co-dual-cross-attention mechanism for effective multimodal fusion. The co-attention component facilitates synchronised learning from multiple data types by highlighting mutually informative features, whereas the dual cross-attention mechanism provides a deeper interaction layer, allowing for more effective integration of these features. Additionally, BioFusionNet mitigates issues of data imbalance through the implementation of a weighted Cox loss function. This comprehensive approach to multimodal fusion marks a significant advancement in predicting survival risks for ER+ breast cancer, addressing notable gaps in current research.

## 2.1 Data Collection

Our study used hematoxylin-and-eosin-stained (H&E) formalin-fixed paraffin-embedded (FFPE) digital slides from The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA). Whole-slide images (WSIs) from the TCGA-BRCA data collection were downloaded from the [GDC Portal](#) (accessed 25 August 2023). In this work, we chose a subset of 249 cases from the TCGA-BRCA dataset, in order to maintain the proportions of Luminal A and B subtypes as well as the proportion of survival events within each subtype. Among the 249 cases, 83 had survival events, while the remaining cases were censored. The survival events were distributed as follows: 54 events in the Luminal A subtype and 29 events in the Luminal B subtype. For each subtype, we selected cases such that the total was about three times the number of survival events, resulting in 149 Luminal A cases and 100 Luminal B cases. By preserving these proportions, we obtained a subset representative of key characteristics in the full dataset. Additionally, we obtained transcriptome-wide RNA-sequencing data representing mRNA expression levels for a total of 20,438 genes in the reference genome from the TCGA dataset. These data were processed using RNA-sequencing by expectation maximisation (RSEM) and were downloaded from the [cBioPortal](#) platform[36]. This dataset included a range of clinical information for each patient, such as tumour grade, tumour size, lymph node status,

age at diagnosis and molecular subtypes. Overall, patients who had WSIs, RNA-sequencing and clinical data available were included in the study.

## 2.2 Data Preparation

### 2.2.1 Slide Annotation

An expert breast pathologist manually annotated the selected slides using QuPath [37]. The annotation was performed for localisation of the tumour outline, excluding any necrosis but including stroma and tumour infiltrating lymphocytes (TILs). The pathologist was blinded to any molecular or clinical features during annotation.

### 2.2.2 Image Data Preparation

The images were first downsampled to 0.25  $\mu\text{m}/\text{pixel}$ , corresponding to approximately 40 $\times$  magnification. The annotated tumour regions were processed semi-automatically with QuPath to create 224 $\times$ 224-pixel patches, resulting in approximately 500 nonoverlapping patches per sample. To address staining inconsistencies, vector-based colour normalisation was applied [38].

### 2.2.3 RNA-Sequencing Data Preparation

From the extensive set of 20,438 genes, we selected genes featured in various commercial assays, namely Oncotype DX, Mammaprint, Prosigna (PAM50), EndoPredict, BCI (Breast Cancer Index), and Mammostrat [39–44], as these are the most relevant genetic markers to our study's objectives. This resulted in a subset of 138 genes. The RNA-sequencing data obtained from the TCGA dataset had already undergone processing using RSEM, and no further normalization was applied to the gene expression values.

### 2.2.4 Clinical Data Preparation

From the clinical data, we selected variables based on their established relevance in breast cancer prognosis and treatment outcomes [45–48]. Specifically, we included tumour grade (categorised as grade 1&2 versus grade 3), tumour size ( $>20 \text{ mm}$  versus  $\leq 20 \text{ mm}$ ), patient age ( $>55$  versus  $\leq 55$ ) and lymph node status (positive versus negative). The decision to binarise the clinical data was made to facilitate CoxPH analysis. Binarisation also simplifies both univariate and multivariate hazard analyses and makes it easier to understand how each clinical factor affects the chance of survival.

## 2.3 Proposed Model

The proposed deep learning model, which we call BioFusionNet, is an innovative feature extraction and multimodal fusion framework designed to leverage and integrate diverse data types, including histology images, genomic features and clinical data for enhanced cancer outcome prediction (Figs. 2 and 3). The essence of BioFusionNet lies in its capability to fuse these data modalities into a cohesive tensor representation, effectively capturing both bimodal and trimodal interactions. This approach is aimed at surpassing the performance of traditional unimodal and existing multimodal representations in survival risk prediction.

### 2.3.1 Feature Extraction Using DINO and MoCoV3

Histopathological images, rich in phenotypic information, are pivotal for understanding cancer pathology. BioFusionNet

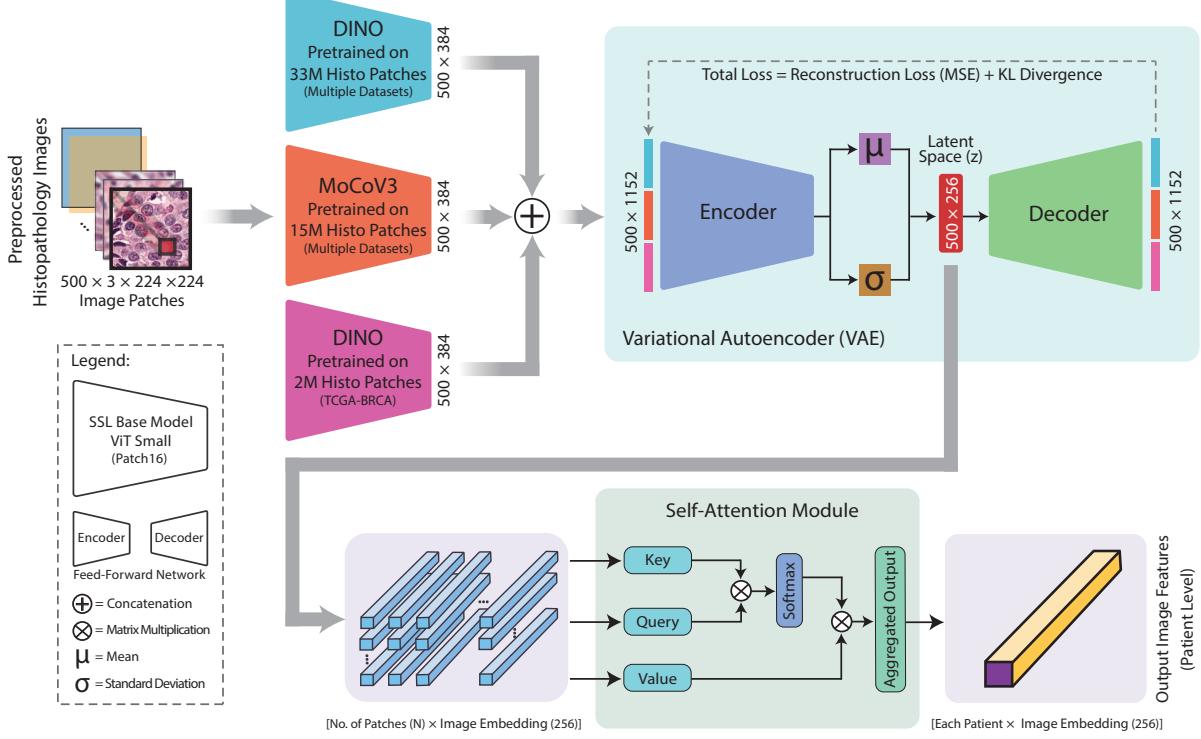


Figure 2: BioFusionNet Stage 1: The proposed model integrates self-supervised image feature extraction methods, namely DINO and MoCoV3, pretrained on three distinct datasets. Features are concatenated and fed to a Variational AutoEncoder (VAE). Subsequently, the latent space of the VAE is utilised to feed a self-attention network, which aggregates patch-level features into a comprehensive patient-level representation.

utilises two advanced self-supervised learning models, DINO (self-DIstillation with NO labels) and MoCoV3 (Momentum Contrast version 3), both based on the Vision Transformer (ViT) architecture, to extract morphological features from histology images crucial for identifying cancer-related patterns.

**DINO:** The DINO framework employs a dual-network architecture, consisting of a student and a teacher network, both being ViTs. The student network learns by attempting to replicate the output of the teacher network, which in turn is an exponential moving average of the student's parameters. The core process involves generating multiple augmented views ( $I_1, I_2, \dots, I_k$ ) of a given input image ( $I$ ), which are then processed by these networks. The resultant feature vectors from the student ( $F_s$ ) and teacher ( $F_t$ ) networks are utilised to compute the distillation loss as follows:

$$\mathcal{L}_D = \sum_{i=1}^k \text{CrossEntropy}\left(F_s(I_i), \text{Softmax}\left(\frac{F_t(I_i)}{\tau}\right)\right), \quad (1)$$

where  $\tau$  represents the temperature scaling parameter. Notably, DINO is pretrained on a broad range of datasets, including BACH, CRC, MHIST, PatchCamelyon and CoNSeP, comprising 33 million patches (DINO33M), and 2 million patches from TCGA-BRCA (DINO2M) [49, 50]. The output functions for DINO33M and DINO2M, denoted as  $f_{\text{DINO33M}}(x)$  and  $f_{\text{DINO2M}}(x)$  respectively, convert an input image  $x$  of size  $224 \times 224 \times 3$  into a  $1 \times 384$  feature vector. We utilised these two models for feature extraction: DINO33M trained on diverse datasets providing a broad perspective and enabling the model to

recognise a wide array of general histopathological features, and DINO2M specifically trained on breast cancer data for more specialised and precise feature extraction relevant to breast cancer pathology.

**MoCoV3:** The MoCoV3 framework, which incorporates ViTs, represents a significant advancement in self-supervised learning through its adoption of the momentum contrast (MoCo) approach [51]. At the heart of this framework lies the contrastive learning mechanism, designed to differentiate between positive and negative pairs, thereby enhancing the model's feature learning capabilities. MoCoV3's architecture is defined by two main components: a query encoder that processes the current batch of images, and a key encoder updated via a momentum-based moving average of the query encoder's parameters:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q, \quad (2)$$

where  $\theta_k$  and  $\theta_q$  are the parameters of the key and query encoders respectively and  $m$  is the momentum coefficient. This enables the key encoder to maintain a queue of encoded keys representing previously seen images, thus enhancing the model's ability to maximise agreement between differently augmented views of the same image (positive pairs) and minimise similarity with other images (negative pairs). The framework uses the InfoNCE loss [52]:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(q \cdot k^+ / \tau)}{\sum_{i=0}^K \exp(q \cdot k_i / \tau)}, \quad (3)$$

where  $q$  and  $k^+$  are the query and positive key feature vectors,  $k_i$  are the negative key vectors,  $K$  is the number of negative

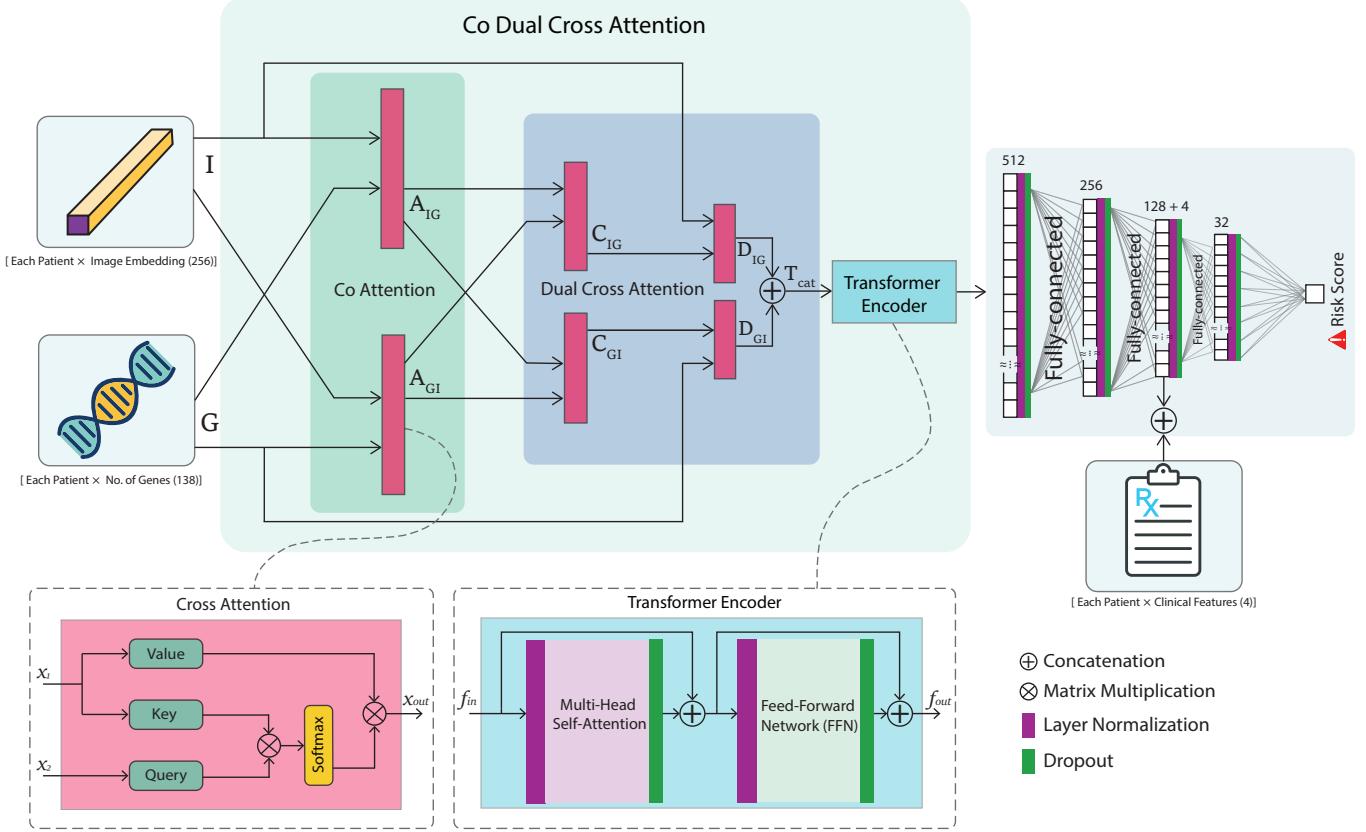


Figure 3: BioFusionNet Stage 2: The proposed model fuses image embeddings generated from Stage 1 with genetic data through a co-dual-cross-attention mechanism. This fusion is subsequently combined with clinical data using a feed-forward network (FFN), leading to the generation of the final risk score output.

keys and  $\tau$  is the temperature parameter. MoCoV3 has been pretrained on an extensive collection of 15 million histology patches from over 30 thousand WSIs derived from the TCGA and Pathology AI Platform (PAIP) datasets, encompassing a wide variety of cancer types and histological features. This endows the model with a robust and versatile capability to extract meaningful features from histopathological data [53]. Similar to DINO, the output function of MoCoV3,  $f_{\text{MoCoV3}}(x)$ , transforms an input image  $x$  of dimensions  $224 \times 224 \times 3$  into a  $1 \times 384$  feature vector.

### 2.3.2 Unimodal Feature Integration

Extracted features from DINO33M, DINO2M and MoCoV3 are concatenated to form a comprehensive  $1 \times 1152$  feature vector  $f_{\text{cat}}(x) = f_{\text{DINO33M}}(x) \oplus f_{\text{DINO2m}}(x) \oplus f_{\text{MoCoV3}}(x)$  (Fig. 2). Following this, we employ a VAE to encode the integrated image features into a 256-dimensional feature in latent space. The latent feature vector is then passed through a self-attention model and aggregated using sum pooling to generate patient-level features.

### 2.3.3 Feature Fusion Using Variational Autoencoding

Our VAE consists of an encoder and a decoder. The encoder function  $f_{\text{enc}}$  maps the concatenated feature vector  $f_{\text{cat}}(x)$  to the latent space by generating the mean  $\mu$  and standard deviation  $\sigma$

of the latent representation:

$$(\mu, \sigma) = f_{\text{enc}}(f_{\text{cat}}(x)). \quad (4)$$

With 500 patches per patient, the latent space is structured as a matrix of size  $500 \times 256$ . This is achieved by sampling  $z$  using the reparameterization trick:

$$z = \mu + \sigma \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad z \in \mathbb{R}^{500 \times 256}. \quad (5)$$

The decoder  $f_{\text{dec}}$  then attempts to reconstruct the input from the latent variable  $z$ :

$$\hat{x} = f_{\text{dec}}(z). \quad (6)$$

The VAE is optimised using a loss function that combines mean squared error (MSE) for reconstruction accuracy and Kullback-Leibler (KL) divergence for distribution regularisation:

$$\mathcal{L}_{\text{VAE}} = \text{MSE}(\hat{x}, x) + \beta \cdot \text{KL}(\mathcal{N}(\mu, \sigma^2) \parallel \mathcal{N}(0, I)), \quad (7)$$

where  $\beta$  balances the reconstruction and regularisation terms. The MSE term ensures that the reconstructed image closely resembles the original input, while the KL divergence term encourages the latent distribution to approximate a standard normal distribution. This enables BioFusionNet to effectively blend the features from the different self-supervised models, enhancing the overall feature representation.

### 2.3.4 Patch-to-Patient Aggregation

To aggregate the patch-level features from the latent space of the VAE into a comprehensive patient-level representation capturing the interdependencies among image patches, our model uses self-attention (Fig. 2). The self-attention module computes a weighted sum of the key ( $K$ ), query ( $Q$ ), and value ( $V$ ) vectors:

$$y_i = \sum_{j=1}^{500} \text{Softmax}(s(Q_i, K_j)) V_j, \quad (8)$$

where  $s(Q_i, K_j)$  is the attention function that determines the relevance between each query and key pair. This offers two significant benefits. First, by leveraging the VAE’s latent vectors, the model focusses on the most pertinent image features, resulting in a more precise feature representation. Second, by considering all latent representations, the self-attention mechanism contextualises each patch within the broader histopathology of the patient. This aggregation process effectively combines the patch-level features of all 500 patches, taking into account their relevance and interdependencies, to generate a comprehensive patient-level representation.

### 2.3.5 Multimodal Fusion Using Co Dual Cross Attention

To integrate patient-level image embeddings with genetic feature data, our model uses a co-dual-cross-attention mechanism (Fig. 3). This is achieved using a complex architecture comprising co-attention and dual-cross-attention modules.

The co-attention module applies linear transformations to the image embeddings  $I$  and genetic features  $G$ , yielding their respective query ( $Q$ ), key ( $K$ ), and value ( $V$ ) vectors, and computes co-attention scores scaled by  $\sqrt{d_k}$ , where  $d_k$  is the dimensionality of the keys:

$$A_{IG} = \text{Softmax}\left(Q_I K_G^T / \sqrt{d_k}\right) V_G, \quad (9)$$

$$A_{GI} = \text{Softmax}\left(Q_G K_I^T / \sqrt{d_k}\right) V_I, \quad (10)$$

where  $A_{IG}$  represents the attention from images to genetic features, and  $A_{GI}$  the attention from genetic features to images. The Softmax function normalises these scores, facilitating an effective weighting of feature importance in the fusion process. This bidirectional attention prepares the ground for more complex interactions in the subsequent stage of the co-dual-cross-attention mechanism. Subsequently, the dual-cross-attention module further refines the integration of the image and genetic features in two distinct stages. The first stage concerns the interaction between the co-attended image features ( $A_{IG}$ ) and co-attended genetic features ( $A_{GI}$ ). The cross attention is computed as:

$$C_{IG} = \text{Softmax}\left(A_{IG}(A_{GI})^T / \sqrt{d_k}\right) A_{IG}, \quad (11)$$

$$C_{GI} = \text{Softmax}\left(A_{GI}(A_{IG})^T / \sqrt{d_k}\right) A_{GI}, \quad (12)$$

where  $C_{IG}$  represents the cross-attention output when image features attend to genetic features, and  $C_{GI}$  the reverse. This stage is crucial for enhancing each modality by integrating contextually relevant information from the other. In the second stage, the outputs from the first stage are further refined by reapplying them to their respective original features. This enhances the depth of the multimodal integration:

$$D_{IG} = \text{Softmax}\left(C_{IG} I^T / \sqrt{d_k}\right) I, \quad (13)$$

$$D_{GI} = \text{Softmax}\left(C_{GI} G^T / \sqrt{d_k}\right) G, \quad (14)$$

---

### Algorithm 1 Weighted Cox Loss

---

**Require:**  $r$ : risks (log hazard ratios),  $e$ : events,  $w$ : weights  
**Ensure:**  $\mathcal{L}_{\text{WCox}}$ : negative log likelihood loss

- 1: Compute  $E_w = \sum_{i=1}^N w_i e_i$  as total weighted events
- 2: Sort samples by descending  $r$  and align  $e, w$
- 3: **for**  $i \in 1, \dots, N$  **do** ▷  $N$  is the number of samples
- 4:     Compute hazard ratio  $h_i = \exp(r_i)$
- 5: **end for**
- 6: Init weighted cumulative hazard  $H_{w_0} = 0$
- 7: **for**  $i \in 1, \dots, N$  **do**
- 8:     Update cumulative sum  $H_{w_i} = H_{w_{i-1}} + w_i h_i$
- 9: **end for**
- 10: **for**  $i \in 1, \dots, N$  **do**
- 11:     Compute uncensored log likelihood  
         $u_i = w_i(r_i - \log(H_{w_i}))$
- 12: **end for**
- 13: Compute  $c = u \odot e$  ▷  $\odot$  is the element-wise product
- 14: Compute loss  $\mathcal{L}_{\text{WCox}} = -\frac{1}{E_w} \sum_{i=1}^N c_i$
- 15: **return**  $\mathcal{L}_{\text{WCox}}$

---

where  $D_{IG}$  and  $D_{GI}$  denote the refined cross-attention outputs, further enhancing the original image and genetic data with additional contextual insights. By sequentially processing the attended features, the model achieves a richer and more contextually informed representation of the fused image and genetic data, suited for complex tasks like risk prediction. The concatenated output  $T_{\text{cat}} = D_{IG} \oplus D_{GI}$  is fed to a Transformer Encoder, which employs multiple layers of self-attention and a feed-forward network (FFN) to achieve a deeper assimilation and transformation of the fused features. The Transformer Encoder consists of  $N$  identical layers, each containing two sublayers: a multihead self-attention mechanism and a position-wise fully connected FFN. In the multihead self-attention sublayer, the input is first linearly projected into a number of different subspaces equal to the number of attention heads. The self-attention mechanism is then applied to each subspace independently, allowing the model to jointly attend to information from different representation subspaces at different positions. The outputs from all heads are then concatenated and linearly projected back to the original dimension. The second sublayer is a position-wise FFN, which consists of two linear transformations with a ReLU activation in between:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (15)$$

where  $W_1$ ,  $W_2$ ,  $b_1$ , and  $b_2$  are learnable parameters. A residual connection is employed around each of the two sublayers, followed by layer normalisation. The output of the Transformer Encoder represents the deeply integrated features from both the imaging and genetic modalities and is further processed by four fully-connected layers, the third of which also integrates the clinical information (Fig. 3). The integration of clinical variables at this stage is crucial due to their dimensionality and characteristics. Clinical data, comprising only four features, may be overshadowed by the higher-dimensional features from histopathological images and genetic profiles if introduced earlier in the model. By integrating these clinical variables in a later layer of the network, closer to the output, their impact is more effectively mapped onto the model’s survival risk prediction. This approach, referred to as ‘late fusion’, allows the clinical variables to have a more significant influence on the

Table 1: Performance comparison of multimodal and unimodal models for cancer risk prediction using C-index.

Fold	Imaging+Genetic+Clinical		Imaging+Genetic	Imaging	Clinical		Genetic	
	BioFusionNet	BioFusionNet	BioFusionNet	CoxPH	MLP	CoxPH	MLP	
1	0.78	0.73	0.58	0.52	0.66	0.62	0.66	
2	0.71	0.72	0.69	0.42	0.64	0.55	0.73	
3	0.72	0.69	0.61	0.59	0.69	0.51	0.66	
4	0.81	0.75	0.70	0.64	0.70	0.62	0.75	
5	0.82	0.65	0.69	0.72	0.66	0.63	0.67	
Mean ± Std	0.77 ± 0.05	0.71 ± 0.04	0.65 ± 0.05	0.58 ± 0.11	0.67 ± 0.02	0.59 ± 0.05	0.69 ± 0.04	

final output compared to ‘early fusion’, where clinical variables are integrated in earlier layers of the network.

This multimodal integration results in a holistic representation of both the phenotypic and genotypic information of ER+ breast cancer. Finally, the network employs a linear output layer that predicts the survival risk score.

#### 2.4 Proposed Loss Function

For training BioFusionNet, we propose a novel loss function termed the weighted Cox loss (computed by Algorithm 1), which is tailored to address the challenges of imbalanced survival data (a common issue in survival analysis):

$$\mathcal{L}_{\text{WCox}} = -\frac{1}{\sum_{i=1}^N w_i e_i} \sum_{i=1}^N w_i e_i (r_i - \log(H_{w_i})), \quad (16)$$

where  $e_i$  denotes the event occurrence,  $w_i$  the assigned weight,  $r_i$  the log hazard ratio,  $H_{w_i}$  the weighted cumulative hazard and  $N$  the number of samples. Unlike the traditional Cox proportional hazards loss ( $\mathcal{L}_{\text{Cox}}$ ) [22], the proposed loss uses weighting to mitigate the effects of uneven distribution of events within the dataset. In this work we used  $w_i = 3$ , considering that censored data (denoted as ‘0’) is almost three times as prevalent as event data (denoted as ‘1’), and thus the sensitivity of the loss function to the latter should be enhanced accordingly, mitigating the bias towards censored data.

#### 2.5 Model Training

The training of BioFusionNet is divided into two distinct stages as follows (Figs. 2 and 3):

**Stage 1: Feature Extraction:** The self-supervised pretrained models DINO33M, MoCoV3 and DINO2M extract features from histopathology image patches, which are concatenated and then fed into a VAE to produce embeddings, which in turn are processed by a self-attention module to produce a patient-level feature vector (Fig. 2). The VAE was optimised using AdamW with a learning rate of 0.0001 and a batch size of 12. The employed loss function is a combination of MSE and KL divergence, targetting the construction of an advanced latent space to generate detailed patient-level features.

**Stage 2: Risk Prediction:** The proposed co-dual-cross-attention mechanism, followed by multiple FFNs and a final output node that uses a linear activation function, predicts the patient-level risk from the image-based, genetic and clinical information (Fig. 3). Here, training was performed using the proposed weighted Cox loss ( $\mathcal{L}_{\text{WCox}}$ ), which was optimised using the Adam algorithm with a learning rate of 0.001 and a batch size of 12. To mitigate overfitting, an early stopping mechanism based

on the validation loss was implemented. This involved halting the training process after a patience period of 10 epochs if no improvement was observed.

This two-stage training approach for BioFusionNet was motivated by two key factors. First, the computational complexity of training sophisticated models, especially those incorporating elements like multiple feature extractors, VAE, and various attention blocks, can be significantly high. An end-to-end training process often requires increased memory usage and even more data, which may not be feasible always. By adopting a two-stage training process, we effectively manage this complexity, breaking down the training into more manageable parts. Second, a modular design facilitates conducting ablation experiments. By training in stages, we can isolate the effects of specific modules or features, systematically analysing their impact and optimising their configuration. This modular approach enhances our ability to refine and improve the learning architecture.

Both training stages used a dataset comprising 199 training samples and 50 validation samples within a five-fold cross-validation framework. The trained model predicts a continuous risk score for every patient within each validation fold. For survival analysis, we employed the median risk score  $\theta_{\text{opt}}$ , derived from each training set, as a threshold to classify patients in the validation set into two categories: high risk (risk score  $> \theta_{\text{opt}}$ ) and low risk (risk score  $< \theta_{\text{opt}}$ ). We note that  $\theta_{\text{opt}}$  is not a user-defined parameter, but its value depends on and is automatically calculated from the specific training set used in each fold, and thus varies depending on the dataset. To achieve optimal performance, we used Optuna<sup>3</sup> for hyperparameter optimisation, tuning key parameters including learning rate, weight decay, number of neurons, number of layers, and dropout for each model.

#### 2.6 Evaluation Metrics

To quantitatively evaluate survival risk score prediction, we employed the concordance index (C-index) and the area under the curve (AUC) as our primary metrics. The C-index assesses the concordance between predicted survival times and observed outcomes, especially in the presence of censored data:

$$\text{C-index} = \frac{\sum_{i=1}^n \sum_{j=1}^n \mathcal{I}(y_i < y_j, \delta_i = 1) \mathcal{I}(\hat{f}(x_i) < \hat{f}(x_j))}{\sum_{i=1}^n \sum_{j=1}^n \mathcal{I}(y_i < y_j, \delta_i = 1)}. \quad (17)$$

where  $n$  is the number of patients,  $y_i$  and  $y_j$  denote the observed survival times,  $\delta_i$  indicates whether the event was observed (not censored),  $\hat{f}(x_i)$  represents the predicted risk for the  $i$ th patient,

<sup>3</sup><https://optuna.org/>

Table 2: Performance comparison of multimodal fusion methods for cancer risk prediction.

Method	Fold	C-index		AUC	
		Value	Mean ± Std	Value	Mean ± Std
MultiSurv [29, 54, 55]	1	0.71		0.74	
	2	0.59		0.52	
	3	0.60	0.63 ± 0.07	0.61	0.63 ± 0.09
	4	0.69		0.70	
	5	0.54		0.57	
MultiDeepCox-SC [27]	1	0.71		0.68	
	2	0.68		0.66	
	3	0.55	0.60 ± 0.08	0.49	0.58 ± 0.08
	4	0.58		0.57	
	5	0.50		0.50	
HFBSurv [30]	1	0.58		0.51	
	2	0.47		0.51	
	3	0.56	0.54 ± 0.07	0.45	0.49 ± 0.04
	4	0.45		0.44	
	5	0.62		0.53	
PathomicFusion [31]	1	0.63		0.68	
	2	0.43		0.10	
	3	0.56	0.52 ± 0.08	0.54	0.47 ± 0.23
	4	0.50		0.63	
	5	0.46		0.38	
MCAT [28]	1	0.71		0.70	
	2	0.69		0.67	
	3	0.64	0.70 ± 0.04	0.65	0.71 ± 0.04
	4	0.70		0.69	
	5	0.76		0.72	
TransSurv [32][26]	1	0.70		0.68	
	2	0.61		0.60	
	3	0.69	0.69 ± 0.04	0.65	0.66 ± 0.04
	4	0.69		0.67	
	5	0.74		0.72	
BioFusionNet (Proposed)	1	0.78		0.82	
	2	0.71		0.93	
	3	0.72	0.77 ± 0.05	0.79	0.84 ± 0.05
	4	0.81		0.81	
	5	0.82		0.83	

and  $\mathcal{I}$  is the indicator function that returns 1 when its condition is true and 0 otherwise. The time-dependent AUC offers a dynamic view of the model accuracy over time  $t$  and incorporates weights  $\omega_i$ , and is calculated using the following formula:

$$\text{AUC}(t) = \frac{\sum_{i=1}^n \sum_{j=1}^n \omega_i \mathcal{I}(y_i \leq t) \mathcal{I}(y_j > t) \mathcal{I}(\hat{f}(x_j) \leq \hat{f}(x_i))}{\sum_{i=1}^n \omega_i \mathcal{I}(y_i \leq t) \sum_{i=1}^n \mathcal{I}(y_i > t)}. \quad (18)$$

Both the C-index and AUC values range from 0 to 1, with higher values indicating better performance. In addition to these metrics, we also compared the computational properties of the models in terms of the number of trainable parameters, memory usage, and floating-point operations (FLOPS).

### 3 EXPERIMENTAL RESULTS

#### 3.1 Comparison Across Modalities

The effectiveness of BioFusionNet in cancer risk prediction was evaluated by comparing its C-index performance across different modality configurations. The results (Table 1) show that the mean performance in the cross-validation experiments consistently increased from using only imaging data, to using imaging and genetic data, to combining imaging, genetic and clinical data. Two traditional methods, specifically CoxPH and the MultiLayer Perceptron (MLP), were also evaluated and showed no

Table 3: Comparison of computational properties.

Method	Parameters	Size	FLOPS
MultiSurv	1.79 M	6.99 MB	2.67 G
MultiDeepCox-SC	0.71 M	2.72 MB	1.18 G
HFBSurv	0.79 M	3.08 MB	2.65 G
PathomicFusion	102.28 M	397.12 MB	3.87 G
MCAT	2.06 M	8.06 MB	2.67 G
TransSurv	1.97 M	7.72 MB	2.67 G
BioFusionNet (Proposed)	4.37 M	17.20 MB	12.11 G

consistent advantage when using only clinical or only genetic data, and their performance was inferior to that of BioFusionNet. This suggests that BioFusionNet’s architecture and training process allow it to effectively extract and integrate complementary information from the three different modalities, leading to improved predictive performance, although it remains elusive exactly how the information in each modality complements the information in the other.

#### 3.2 Comparison With State-of-the-Art Fusion Methods

The performance of BioFusionNet was compared with several state-of-the-art multimodal fusion methods for cancer risk prediction in terms of both C-index and AUC (Table 2). For this experiment we included methods using concatenation (MultiSurv), Cox proportional hazards with image-derived risk scores (MultiDeepCox-SC), hierarchical attention (HFBSurv), gating attention (PathomicFusion), co-attention (MCAT) and cross attention (TransSurv). To ensure fair comparison, we used the same hyperparameter optimisation framework (Optuna) for all models. The AUC was calculated using average values over 5-year and 10-year periods. The proposed model consistently outperformed all these previous methods, showing substantial improvements in both metrics. However, BioFusionNet is computationally demanding, requiring 12.11 G (Giga) FLOPS, the highest among the compared models, although it is more memory efficient than PathomicFusion (Table 3).

#### 3.3 Evaluation of Loss Functions

The performance of two different loss functions was compared using the C-index across five cross-validation folds for BioFusionNet and MoCoV3 (Table 4). The results show that the mean performance improved for both methods when using the weighted Cox loss ( $\mathcal{L}_{\text{WCox}}$ ) proposed in this paper, compared to the traditional Cox loss ( $\mathcal{L}_{\text{Cox}}$ ).

#### 3.4 Univariate and Multivariate Hazard Analysis

A comprehensive hazard analysis was conducted to evaluate the overall survival (OS) in the TCGA dataset of ER+ patients. Both univariate and multivariate analyses were performed (Table 5). The analysis encompassed various parameters, including tumour grade, tumour size, age, lymph node (LN) status, subtype and the risk predictions made by BioFusionNet. In the multivariate analysis, positive LN status was associated with a hazard ratio (HR) of 1.87 (95% CI: 1.32–2.64), demonstrating a significant effect on survival ( $p < 0.005$ ). Additionally, patients over the age of 55 had a HR of 1.77 (95% CI: 1.07–2.91), also showing a significant impact on survival ( $p = 0.03$ ). However, no significant associations were found between tumour grade, size, or

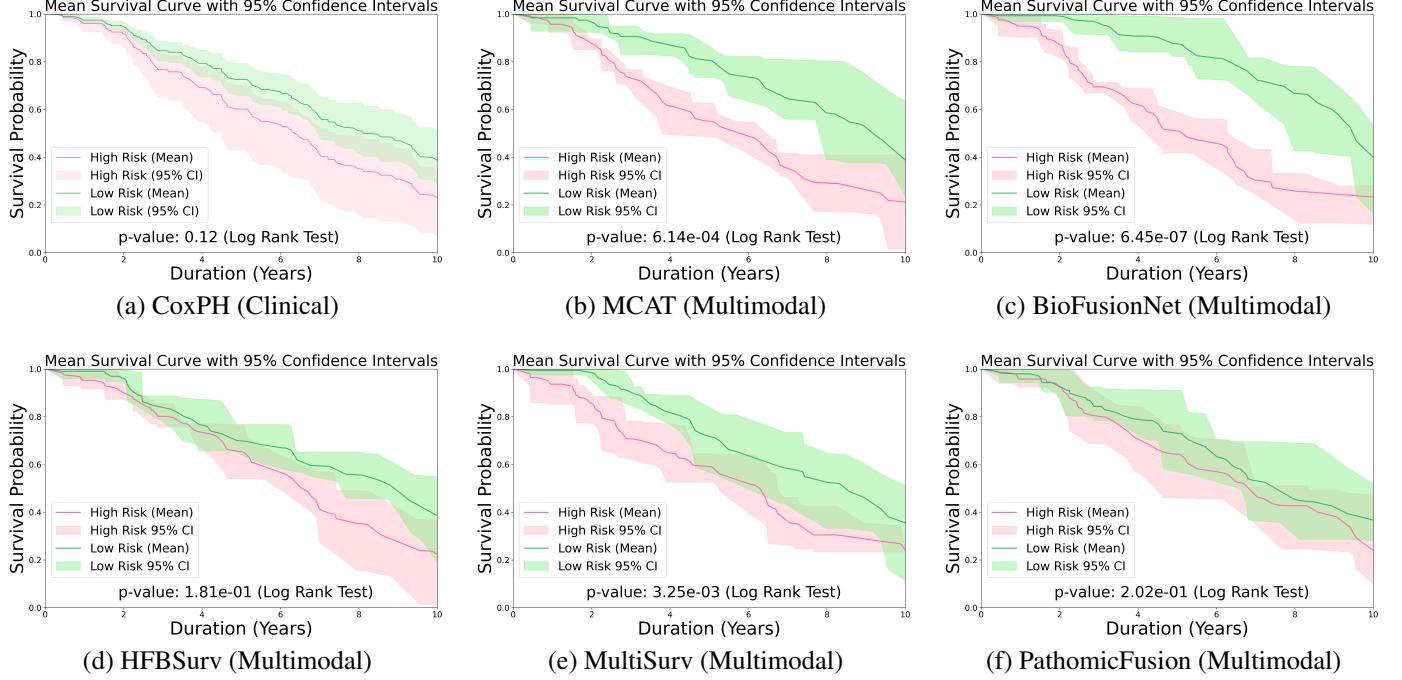


Figure 4: Performance comparison of BioFusionNet and other methods using Kaplan-Meier survival curves.

Table 4: Performance comparison of loss functions for cancer risk prediction using two different methods.

Loss	Method	C-index (Fold)					Mean $\pm$ Std
		1	2	3	4	5	
$\mathcal{L}_{\text{Cox}}$	BioFusionNet	0.69	0.54	0.59	0.75	0.80	$0.67 \pm 0.10$
	MoCoV3	0.66	0.57	0.57	0.78	0.80	$0.67 \pm 0.11$
$\mathcal{L}_{\text{WCox}}$ (Proposed)	BioFusionNet	0.78	0.71	0.72	0.81	0.82	$0.77 \pm 0.05$
	MoCoV3	0.70	0.66	0.66	0.77	0.72	$0.70 \pm 0.04$

subtype and survival outcomes in this analysis. Notably, the BioFusionNet-predicted risk group (high vs. low) demonstrated a significant correlation with OS, with a HR of 2.91 (95% CI: 1.80–4.68) ( $p < 0.005$ ). Univariate analysis indicated that tumour grade, size, age, and subtype were not statistically significant, whereas LN status (HR of 1.84, 95% CI: 1.33–2.55,  $p < 0.005$ ) and BioFusionNet risk group (HR of 2.99, 95% CI: 1.88–4.78,  $p < 0.005$ ) were significant predictors of survival. We note that the LN status had 51 missing values, which were imputed using a fixed value of 2. Kaplan-Meier survival analysis further supported the results, showing a significant difference in survival probabilities between the high- and low-risk groups as predicted by BioFusionNet (log-rank test  $p = 6.45e-7$ ) (Fig. 4c).

### 3.5 Ablation Study

We also evaluated the performance of various versions of BioFusionNet for ER+ breast cancer risk stratification. The results (Table 6) show that the base model, BioFusionNet-B0 (MoCoV3, ViT Small) with  $\mathcal{L}_{\text{WCox}}$ , achieved the lowest C-index, and incorporating single cross-attention (SCA), dual cross-attention (DCA), or co-attention (CoA) yielded slight improvements, as did BioFusionNet-B1 (DINO33M, ViT Small)

and BioFusionNet-B2 (DINO2M, ViT Small), both with DCA and  $\mathcal{L}_{\text{WCox}}$ . Combining BioFusionNet-B0, B1, and B2 with just  $\mathcal{L}_{\text{WCox}}$  also resulted in slightly better performance than the base model, as did the inclusion of SCA and VAE. More substantial improvements of the combined model were obtained with the inclusion of DCA or CoA instead of SCA. The best performance was achieved by the combined model using VAE, CoA, DCA and late clinical data fusion ( $C\text{-Fusion}_{\text{Late}}$ ) with  $\mathcal{L}_{\text{WCox}}$ , which clearly outperformed the same model using early clinical data fusion ( $C\text{-Fusion}_{\text{Early}}$ ) or  $\mathcal{L}_{\text{Cox}}$ .

### 3.6 Interpretability of BioFusionNet

BioFusionNet utilises a self-attention mechanism to analyse histopathological image patches, identifying regions of high and low attention within both high-risk and low-risk patient profiles. Visual inspection of the results (Fig. 5) reveals that regions with high attention contain distinct cellular patterns crucial for synthesising features from patch-level to patient-level, whereas areas of low attention typically exhibit less cellular atypia. This shows the model capacity to pinpoint clinically relevant features within tissue morphology. Additionally, SHAP analysis (Fig. 6) reveals the influence of individual genes on the model predictions, ranked from high to low, providing interpretability of the risk assessment process. From this analysis, gene SLC39A6 (an estrogen regulated Zinc transporter protein with a role in epithelial to mesenchymal transition (EMT)) was identified as the most important predictor, with high expression levels producing high SHAP values, indicating positive impact on the model’s cancer risk prediction. Other influential genes include ERBB2 (the gene for HER2), ESR1 (the gene for ER), with low expression levels producing high SHAP values therefore positive impact on the model. Moreover, the distribution of SHAP values for clinical features (Fig. 7) indicates that higher values of clinical parameters—such as positive LN status, higher tumour grade,

Table 5: Univariate and multivariate analysis for overall survival (OS) in the TCGA dataset of ER+ patients.

Parameter	Risk Group Cutoff	#Patients/Group	Multivariate (n=249)			Univariate (n=249)		
			HR	95% CI	p	HR	95% CI	p
Tumour Grade	3 vs. 1 & 2	64 vs. 185	0.83	0.46–1.49	0.54	1.13	0.67–1.91	0.65
Tumour Size	>20 vs. ≤20 (mm)	167 vs. 82	1.45	0.88–2.37	0.14	1.37	0.86–2.19	0.19
Age	>55 vs. ≤55	159 vs. 90	1.77	1.07–2.91	0.03	1.47	0.91–2.36	0.11
LN Status*	pos. vs. neg.	110 vs. 88	1.87	1.32–2.64	<0.005	1.84	1.33–2.55	<0.005
Subtype	lum B vs. A	100 vs. 149	1.43	0.88–2.34	0.15	1.38	0.88–2.18	0.16
BioFusionNet	high vs. low	132 vs. 117	2.91	1.80–4.68	<0.005	2.99	1.88–4.78	<0.005

\*LN Status had 51 missing values which were imputed with a fixed value for both multivariate and univariate analysis.

Table 6: Ablation study of BioFusionNet

Model	C-index
BioFusionNet-B0 (MoCoV3, ViT Small) + C-Fusion <sub>late</sub> + $\mathcal{L}_{\text{Wcox}}$	0.65 ± 0.05
BioFusionNet-B0 (MoCoV3, ViT Small) + SCA + C-Fusion <sub>late</sub> + $\mathcal{L}_{\text{Wcox}}$	0.69 ± 0.04
BioFusionNet-B0 (MoCoV3, ViT Small) + DCA + C-Fusion <sub>late</sub> + $\mathcal{L}_{\text{Wcox}}$	0.70 ± 0.03
BioFusionNet-B0 (MoCoV3, ViT Small) + CoA + C-Fusion <sub>late</sub> + $\mathcal{L}_{\text{Wcox}}$	0.70 ± 0.04
BioFusionNet-B1 (DINO33M, ViT Small) + DCA + C-Fusion <sub>late</sub> + $\mathcal{L}_{\text{Wcox}}$	0.68 ± 0.02
BioFusionNet-B2 (DINO2M, ViT Small) + DCA + C-Fusion <sub>late</sub> + $\mathcal{L}_{\text{Wcox}}$	0.67 ± 0.03
BioFusionNet-Concat(B0+B1+B2) + C-Fusion <sub>late</sub> + $\mathcal{L}_{\text{Wcox}}$	0.67 ± 0.04
BioFusionNet-Concat(B0+B1+B2) + SCA + C-Fusion <sub>late</sub> + $\mathcal{L}_{\text{Wcox}}$	0.69 ± 0.03
BioFusionNet-Concat(B0+B1+B2) + VAE + SCA + C-Fusion <sub>late</sub> + $\mathcal{L}_{\text{Wcox}}$	0.68 ± 0.04
BioFusionNet-Concat(B0+B1+B2) + VAE + DCA + C-Fusion <sub>late</sub> + $\mathcal{L}_{\text{Wcox}}$	0.75 ± 0.04
BioFusionNet-Concat(B0+B1+B2) + VAE + CoA + C-Fusion <sub>late</sub> + $\mathcal{L}_{\text{Wcox}}$	0.70 ± 0.03
BioFusionNet-Concat(B0+B1+B2) + VAE + CoA + DCA + Clinic-F <sub>late</sub> + $\mathcal{L}_{\text{cox}}$	0.67 ± 0.10
BioFusionNet-Concat(B0+B1+B2) + VAE + CoA + DCA + C-Fusion <sub>Early</sub> + $\mathcal{L}_{\text{Wcox}}$	0.74 ± 0.04
BioFusionNet-Concat(B0+B1+B2) + VAE + CoA + DCA + C-Fusion <sub>late</sub> + $\mathcal{L}_{\text{Wcox}}$	0.77 ± 0.03

increased tumour size, and postmenopausal age group—tend to have a positive impact on the model’s output. In this context, a ‘positive impact’ implies that the model associates these values with a higher likelihood of predicting patients at high risk.

## 4 DISCUSSION AND CONCLUSION

As demonstrated by the experimental results, the proposed BioFusionNet is highly effective for cancer risk prediction, showing superior performance compared to alternative approaches. Clearly, the multimodal fusion of imaging, genetic and clinical data allows the model to achieve substantially higher C-index scores compared to unimodal and dual-modal configurations, as well as compared to the traditional unimodal CoxPH and MLP. Furthermore, BioFusionNet outperforms existing multimodal fusion methods such as MultiSurv, HFBSurv, PathomicFusion, MCAT and TransSurv, and achieves the highest mean C-index ( $0.77 \pm 0.05$ ) and AUC ( $0.84 \pm 0.05$ ). Partly, the superior performance of BioFusionNet is due to the introduction of the proposed weighted Cox loss function instead of using the traditional Cox loss. Univariate and multivariate analyses showed the significant impact of age and BioFusionNet predictions on survival outcomes, while other clinical parameters such as tumour grade, size, lymph node status and subtype did not exhibit a significant correlation with survival outcomes. Kaplan-Meier analysis revealed a distinct separation in survival probabilities between the high-risk and low-risk groups identified by BioFusionNet. In addition, the results of the ablation experiment confirmed the importance of attention mechanisms in improving prediction accuracy, with the combined model configuration utilising VAE, CoA and DCA and the weighted Cox loss showing the highest performance.

BioFusionNet also presents a significant advancement in the interpretation of histopathological images, leveraging a self-attention mechanism to distinguish critical regions in patient profiles. A key contribution is the model’s ability to align high-attention areas with distinct cellular patterns, crucial for transitioning from patch-level to patient-level analysis, thereby enhancing the diagnostic process. SHAP analysis amplifies this by clarifying the influence of specific genes and clinical features on the model’s predictions. We observed that elevated SLC39A6 gene expression correlates with a high-risk prediction in ER+ breast cancers, where previous studies have shown conflicting findings, associating high SLC39A6 levels with good prognosis [56, 57], while others associated it with increased proliferation and lymph node involvement [58–60]. Similarly, our model identified high ESR1 expression as indicative of low risk, aligning with literature that associates ESR1 positivity with enhanced responsiveness to endocrine therapy and, consequently, a better prognosis in ER+ breast cancer patients [61]. In contrast, our analysis revealed an unexpected association between ERBB2 overexpression and a favorable prognosis in ER+ breast tumours, contrasting with the established view that ERBB2 overexpression indicates a poor prognosis [62, 63]. As our study was specifically tailored to analyze ER+ samples, excluding the HER2-enriched subtype (known for its high ERBB2 expression and aggressiveness) likely influenced the findings. Moreover, the SHAP analysis for clinical factors (LN positivity, higher tumour grade and size, and postmenopausal age) significantly influences our model’s ability to identify patients at increased risk, highlighting the critical role of these factors in breast cancer prognosis. Our analysis provides a transparent understanding of how each gene and clinical feature contributes to the model’s predictions, providing actionable insights for clinical decision-making. While the risk assessment process mirrors current clinical practice, BioFusionNet streamlines the integration of all available data (patient features, tumour features and molecular features) to derive an automated single risk prediction score as a potential clinical oncology tool of the future.

While insightful, this study has certain limitations. We primarily opted for OS as the key outcome measure, instead of disease-free survival (DFS). This choice was made because DFS presented challenges such as a lower rate of events and a higher degree of data censorship, which could have limited the depth of the analysis. While OS is a feasible choice, it potentially overlooks critical insights into early-stage disease progression, typically highlighted by DFS. Moreover, the study’s reliance on specific datasets such as TCGA for ER+ patients may affect the broad applicability of our findings. Another shortcoming of this study

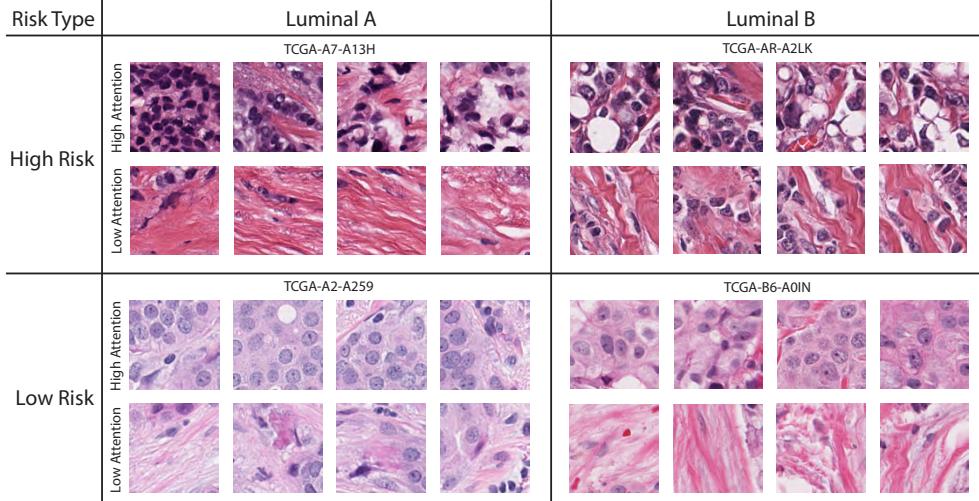


Figure 5: Visualisation of model-derived attention regions and associated risk types in Luminal A and Luminal B breast cancer patients. The figure presents raw histopathological image patches processed with BioFusionNet, which identifies areas of high and low attention, subsequently categorising patients into high and low risk.

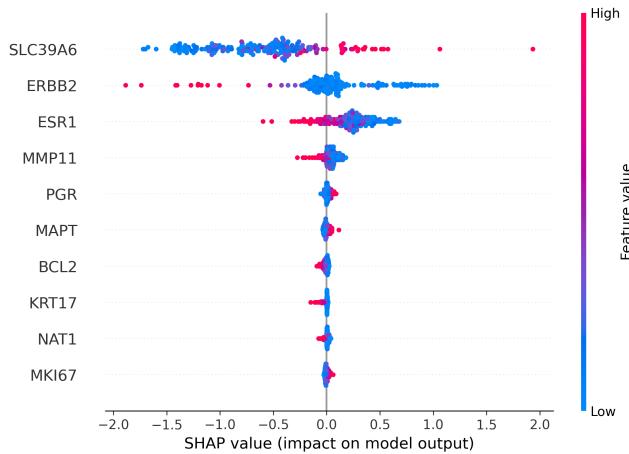


Figure 6: SHAP analysis of genetic features. The x-axis represents the SHAP value; colour intensity indicates gene expression level. The plot is sorted vertically by the features' overall importance.

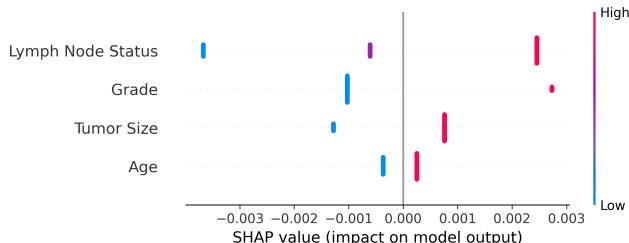


Figure 7: SHAP value distribution of clinical features. In this distribution, higher clinical values showing positive impact on the model, as indicated by its SHAP values.

is the inherent limitations of the clinical data, which, during univariate analysis, identified tumour size, grade, and age as in-

significant while only LN Status emerged as significant. Despite its limitations, the effectiveness of deep learning algorithms in analysing this clinical data arises from their ability to uncover complex patterns and interactions within dataset. Future research should therefore aim to validate these findings across a wider range of datasets to bolster the model's generalisability. Incorporating organ-level data, such as mammograms, could further enhance the predictive accuracy of our model. Additionally, extending the application of BioFusionNet to other cancer types and clinical scenarios could yield more comprehensive insights, making the research more universally relevant and applicable. Finally, a limitation of our model is that it is computationally demanding, primarily due to its extensive use of attention mechanisms. Whether the better performance justifies the higher computational cost depends on user needs and resources. Further research may provide ways to reduce the computational requirements of the model while retaining its high performance.

## DATA AVAILABILITY

TCGA image data and clinical data are publicly available at <https://portal.gdc.cancer.gov/>.

## CODE AVAILABILITY

Our work is fully reproducible and source code is publicly available on GitHub at <https://github.com/raktim-mondol/BioFusionNet>.

## ACKNOWLEDGEMENT

This research was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government. Additionally, data preprocessing was performed using the computational cluster Katana, which is supported by Research Technology Services at UNSW Sydney.

## COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by The Cancer Genome Atlas Breast Cancer (TCGA-BRCA) dataset, accessible through the National Cancer Institute's [Genomic Data Commons \(GDC\)](#) portal. Ethical approval was not required for this study, in accordance with the [ethical policies](#) set forth by The Cancer Genome Atlas program.

## REFERENCES

- [1] Y. B. Shvetsov, L. R. Wilkens, K. K. White, M. Chong, A. Buyum, G. Badowski, R. T. L. Guerrero, and R. Novotny, "Prediction of breast cancer risk among women of the Mariana Islands: the BRISK retrospective case-control study," *BMJ Open*, vol. 12, no. 12, p. e061205, 2022.
- [2] W. Guo, W. Liang, Q. Deng, and X. Zou, "A multimodal affinity fusion network for predicting the survival of breast cancer patients," *Frontiers in Genetics*, vol. 12, p. 709027, 2021.
- [3] M. D. J. Peters, I. Ramsey, K. Kennedy, G. Sharplin, and M. Eckert, "Culturally safe, high-quality breast cancer screening for transgender people: a scoping review protocol," *Journal of Advanced Nursing*, vol. 78, no. 1, pp. 276–281, 2021.
- [4] C. Shuai, F. Yuan, Y. Liu, C. Wang, J. Wang, and H. He, "Estrogen receptor-positive breast cancer survival prediction and analysis of resistance-related genes introduction," *PeerJ*, vol. 9, p. e12202, 2021.
- [5] K. Holli-Helenius, A. Salminen, I. Rinta-Kiikka, I. Koskiuvu, N. Brück, P. Boström, and R. Parkkola, "MRI texture analysis in differentiating luminal A and luminal B breast cancer molecular subtypes—a feasibility study," *BMC Medical Imaging*, vol. 17, p. 69, 2017.
- [6] H. Lindman, F. Wiklund, and K. K. Andersen, "Long-term treatment patterns and survival in metastatic breast cancer by intrinsic subtypes—an observational cohort study in Sweden," *BMC Cancer*, vol. 22, p. 1006, 2022.
- [7] Y. Han, J. Wang, and B. Xu, "Clinicopathological characteristics and prognosis of breast cancer with special histological types: a surveillance, epidemiology, and end results database analysis," *The Breast*, vol. 54, pp. 114–120, 2020.
- [8] M. A. Han, E. C. Hwang, and J. H. Jung, "Prognostic factors of mortality in patients with cancer infected with COVID-19: a systematic review protocol," *BMJ Open*, vol. 13, no. 7, p. e071810, 2023.
- [9] C. Nero, F. Ciccarone, A. Pietragalla, S. Duranti, G. Daniele, G. Scambia, and D. Lorusso, "Adjuvant treatment recommendations in early-stage endometrial cancer: what changes with the introduction of the integrated molecular-based risk assessment," *Frontiers in Oncology*, vol. 11, p. 612450, 2021.
- [10] A. N. Zahari, N. S. Mohamad, and M. H. Mahmud, "Impacts of clinicopathological factors on metabolic parameters of 18F fluorodeoxyglucose PET/CT in the staging of breast cancer," *Journal of Sustainability Science and Management*, vol. 17, no. 12, pp. 166–173, 2022.
- [11] X. Li, L. Liu, G. J. Goodall, A. Schreiber, T. Xu, J. Li, and T. D. Le, "A novel single-cell based method for breast cancer prognosis," *PLoS Computational Biology*, vol. 16, no. 8, pp. 1–20, 2020.
- [12] M. I. Jaber, L. Beziaeva, C. W. Szeto, and S. C. Benz, "Deep learning-based risk stratification for HER2-negative breast cancer patients," *bioRxiv*, p. 10.1101/2021.05.26.445720, 2021.
- [13] M. Garutti, G. Griguolo, A. Botticelli, G. Buzzatti, C. D. Angelis, L. Gerratana, C. Molinelli, V. Adamo, G. Bianchini, L. Biganzoli, G. Curgiliano, M. D. Laurentiis, A. Fabi, A. Frassoldati, A. Genari, M. Scaltriti, F. Perrone, G. Viale, C. Zamagni, A. Zambelli, L. D. Mastro, S. D. Placido, V. Guarneri, P. Marchetti, and F. Puglisi, "Definition of high-risk early hormone-positive HER2-negative breast cancer: a consensus review," *Technology in Cancer Research & Treatment*, vol. 14, no. 8, p. 1898, 2022.
- [14] B. Lu, E. Natarajan, H. R. B. Raghavendran, and U. D. Markanday, "Molecular classification, treatment, and genetic biomarkers in triple-negative breast cancer: a review," *Technology in Cancer Research & Treatment*, vol. 22, p. 15330338221145246, 2023.
- [15] L. Guo, D. Kong, J. Liu, L. Zhan, L. Luo, W. Zheng, Q. Zheng, C. Chen, and S. Sun, "Breast cancer heterogeneity and its implication in personalized precision therapy," *Experimental Hematology & Oncology*, vol. 12, no. 1, p. 3, 2023.
- [16] J. C. Wei, A. A. Suriawinata, B. Ren, Y. Zhang, M. Lisovsky, L. J. Vaickus, C. R. Brown, M. J. Baker, N. Tomita, L. Torresani, J. Z. Wei, and S. Hass nanopour, "A petri dish for histopathology image analysis," *Artificial Intelligence in Medicine*, pp. 11–24, 2021.
- [17] E. Schaafsma, B. Zhang, M. Schaafsma, C. Tong, L. Zhang, and C. Cheng, "Impact of oncotype DX testing on ER+ breast cancer treatment and survival in the first decade of use," *Breast Cancer Research*, vol. 23, no. 1, p. 74, 2021.
- [18] A. D. Caluwé, L. Buisseret, P. Poortmans, D. V. Gestel, R. Salgado, C. Sotiriou, D. Larsimont, M. Paesmans, L. Craciun, S. Drisis, C. Vandekerckhove, F. Reynal, I. Veys, D. Eiger, M. Piccart, E. Romano, and M. Ignatiadis, "Neo-CheckRay: radiation therapy and adenosine pathway blockade to increase benefit of immunotherapy and chemotherapy in early stage luminal B breast cancer, a randomized phase II trial," *BMC Cancer*, vol. 21, p. 899, 2021.
- [19] S. Paik, G. Tang, S. Shak, C. Kim, J. B. Baker, W. Kim, M. Cronin, F. L. Baehner, D. Watson, J. Bryant, J. P. Costantino, C. E. Geyer, D. L. Wickerham, and N. Wolmark, "Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer," *Journal of Clinical Oncology*, vol. 24, no. 23, pp. 3726–3734, 2006.
- [20] K. Lee, S. H. Sim, E. J. Kang, J. H. Seo, H. D. Chae, K. S. Lee, J. Y. Kim, J. S. Ahn, Y. H. Im, S. Park, Y. H. Park, and I. H. Park, "The role of chemotherapy in patients with HER2-negative isolated locoregional recurrence of breast cancer: a multicenter retrospective cohort study," *Frontiers in Oncology*, vol. 11, p. 653243, 2021.
- [21] Y. Naoi, R. Tsunashima, K. Shimazu, and S. Noguchi, "The multi-gene classifiers 95GC/42GC/155GC for precision medicine in ER-positive HER2-negative early breast cancer," *Cancer Science*, vol. 112, no. 4, pp. 1369–1375, 2021.
- [22] J. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network," *BMC Medical Research Methodology*, vol. 18, p. 24, 2018.
- [23] Y. He, B. Hu, C. Zhu, W. Xu, X. Hao, B. Dong, X. Chen, Q. Dong, and X. Zhou, "A novel multimodal radiomics model for predicting prognosis of resected hepatocellular carcinoma," *Frontiers in Oncology*, vol. 12, p. 745258, 2022.
- [24] V. Subramanian, T. Syeda-Mahmood, and M. N. Do, "Multimodal fusion using sparse CCA for breast cancer survival prediction," *IEEE International Symposium on Biomedical Imaging*, pp. 1429–1432, 2021.
- [25] R. Vanguri, J. Luo, A. Aukerman, J. V. Egger, C. J. Fong, N. Horvat, A. Pagano, J. d. A. B. Araújo-Filho, L. Geneslaw, H. Rizvi, R. E. Sosa, K. M. Boehm, S. Yang, F. M. Bodd, K. Ventura, T. J.

- Hollmann, M. S. Ginsberg, J. Gao, M. D. Hellmann, J. L. Sauter, and S. P. Shah, "Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer," *Nature Cancer*, vol. 3, no. 10, pp. 1151–1164, 2022.
- [26] S. Deng, Y. Suo, S. Liu, X. Ma, H. Chen, X. Liao, J. Zhang, and W. W. Y. Ng, "MFCSA-CAT: a multimodal fusion method for cancer survival analysis based on cross-attention transformer," *International Conference on Computer Information Science and Artificial Intelligence*, vol. 12566, p. 1256608, 2023.
- [27] T. Wei, X. Yuan, R. Gao, L. J. Johnston, J. Zhou, Y. Wang, W. Kong, Y. Xie, Y. Zhang, D. Xu, and Z. Yu, "Survival prediction of stomach cancer using expression data and deep learning models with histopathological images," *Cancer Science*, vol. 114, no. 2, pp. 690–701, 2022.
- [28] R. J. Chen, M. Y. Lu, W.-H. Weng, T. Y. Chen, D. F. Williamson, T. Manz, M. Shady, and F. Mahmood, "Multimodal co-attention transformer for survival prediction in gigapixel whole slide images," *IEEE/CVF International Conference on Computer Vision*, pp. 3995–4005, 2021.
- [29] L. A. Vale-Silva and K. Rohr, "Long-term cancer survival prediction using multimodal deep learning," *Scientific Reports*, vol. 11, p. 13505, 2021.
- [30] R. Li, X. Wu, A. Li, and M. Wang, "HFBSurv: Hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction," *Bioinformatics*, vol. 38, no. 9, pp. 2587–2594, 2022.
- [31] R. J. Chen, M. Y. Lu, J. Wang, D. F. Williamson, S. J. Rodig, N. I. Lindeman, and F. Mahmood, "Pathomic Fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," *IEEE Transactions on Medical Imaging*, vol. 41, no. 4, pp. 757–770, 2022.
- [32] Z. Lv, Y. Lin, R. Yan, Y. Wang, and F. Zhang, "TransSurv: Transformer-based survival analysis model integrating histopathological images and genomic data for colorectal cancer," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 6, pp. 3411–3420, 2023.
- [33] S. Garg, H. SS, and S. Kumar, "On-device document classification using multimodal features," *ACM India Joint International Conference on Data Science & Management of Data*, p. 203–207, 2021.
- [34] K. Xu, Z. Lin, J. Zhao, P. Shi, W. Deng, and H. Wang, "Multimodal deep learning for social media popularity prediction with attention mechanism," *ACM International Conference on Multimedia*, p. 4580–4584, 2020.
- [35] S. Qiu, X. Cui, Z. Ping, N. Shan, Z. Li, X. Bao, and X. Xu, "Deep learning techniques in intelligent fault diagnosis and prognosis for industrial systems: a review," *Sensors*, vol. 23, no. 3, p. 1305, 2023.
- [36] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, p. 323, 2011.
- [37] P. Bankhead, M. B. Loughrey, J. A. Fernández, Y. Dombrowski, D. G. McArt, P. D. Dunne, S. McQuaid, R. T. Gray, L. J. Murray, H. G. Coleman, J. A. James, M. Salto-Tellez, and P. W. Hamilton, "QuPath: open source software for digital pathology image analysis," *Scientific Reports*, vol. 7, p. 16878, 2017.
- [38] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis," *IEEE International Symposium on Biomedical Imaging*, pp. 1107–1110, 2009.
- [39] C. Mazo, S. Barron, and C. Mooney, "Multi-gene prognostic signatures and prediction of pathological complete response to neoadjuvant chemotherapy in ER-positive, HER2-negative breast cancer patients," *Cancers*, vol. 12, no. 5, p. 1133, 2020.
- [40] P. Blanchette, D. Sivajohanathan, J. M. Bartlett, A. Eisen, H. Feiloter, R. C. Pezo, G. Turashvili, and P. E. Williams, "Clinical utility of multigene profiling assays in early-stage invasive breast cancer: an Ontario Health (Cancer Care Ontario) clinical practice guideline," *Current Oncology*, vol. 29, no. 4, pp. 2599–2616, 2022.
- [41] K. Almstedt, S. Mendoza, M. Otto, M. Battista, J. Steetskamp, A. S. Heimes, S. Krajnak, A. Poplawski, A. Gerhold-Ay, A. Hasenburg, C. Denkert, and M. Schmidt, "Endopredict® in early hormone receptor-positive, HER2-negative breast cancer," *Breast Cancer Research and Treatment*, vol. 182, pp. 137–146, 2020.
- [42] S. Jahn, A. Bösl, O. Tsybrovskyy, C. Gruber-Rossipal, R. Helfgott, F. Fitzal, M. Knauer, M. Babić, Z. Jasarevic, F. Offner, and F. Moinfar, "Clinically high-risk breast cancer displays markedly discordant molecular risk predictions between the mammaprint and endopredict tests," *British Journal of Cancer*, vol. 122, no. 12, pp. 1744–1746, 2020.
- [43] S. P. Somashekhar, S. Zaveri, D. G. Vijay, P. S. Dattatreya, R. Kumar, F. Islahi, and C. Bahl, "Individualized chemotherapy benefit prediction by endopredict in patients with early breast cancer in an Indian cohort," *JCO Global Oncology*, no. 6, pp. 1363–1369, 2020.
- [44] R. Buus, I. Šestak, R. Kronenwett, S. Ferree, C. A. Schnabel, F. L. Baehner, E. Mallon, J. Cuzick, and M. Dowsett, "Molecular drivers of Oncotype DX, Prosigna, EndoPredict, and the Breast Cancer Index: a TransATAC study," *Journal of Clinical Oncology*, vol. 39, no. 2, pp. 126–135, 2021.
- [45] J. Warwick, L. Tabár, B. Viták, and S. W. Duffy, "Time-dependent effects on survival in breast carcinoma," *Cancer*, vol. 100, no. 7, pp. 1331–1336, 2004.
- [46] C. Woo, H. Silberman, S. Nakamura, W. Ye, R. Spoto, W. J. Colburn, J. Waisman, and M. J. Silverstein, "Lymph node status combined with lymphovascular invasion creates a more powerful tool for predicting outcome in patients with invasive breast cancer," *American Journal of Surgery*, vol. 184, no. 4, pp. 337–340, 2002.
- [47] H. Bor, E. N. Maina, B. Nyambega, K. Patel, C. O. Olwal, W. Nalyanya, and Y. Gavamukulya, "The potential of differentiation-related gene-1 (DRG1) as a biomarker for metastasis of estrogen receptor-positive breast cancer," *Journal of Advances in Medicine and Medical Research*, pp. 162–169, 2021.
- [48] S. Naz, M. Siddiqui, A. I. Memon, A. M. Bhatti, Z. I. Hussain, and Iqra, "Analysis of breast cancer receptors status and molecular subtypes among female population," *Pakistan Journal of Medical and Health Sciences*, vol. 17, no. 1, pp. 656–658, 2023.
- [49] M. Kang, H. Song, S. Park, D. Yoo, and S. Pereira, "Benchmarking self-supervised learning on diverse pathology datasets," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3344–3354, 2023.
- [50] R. J. Chen and R. G. Krishnan, "Self-supervised vision transformers learn visual concepts in histopathology," *Annual Conference on Neural Information Processing Systems*, 2021.
- [51] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," *CoRR*, p. 2104.02057, 2021.
- [52] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv*, p. 1807.03748, 2018.
- [53] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, and X. Han, "Transformer-based unsupervised contrastive learning for histopathological image classification," *Medical Image Analysis*, vol. 81, p. 102559, 2022.

- [54] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, “Multimodal deep learning models for early detection of Alzheimer’s disease stage,” *Scientific Reports*, vol. 11, p. 3254, 2021.
- [55] S. Steyaert, Y. L. Qiu, Y. Zheng, P. Mukherjee, H. Vogel, and O. Gevaert, “Multimodal deep learning to predict prognosis in adult and pediatric brain tumors,” *Communications Medicine*, vol. 3, no. 1, p. 44, 2023.
- [56] M. Althobiti, K. A. El-sharawy, C. Joseph, M. Aleskandarany, M. S. Toss, A. R. Green, and E. A. Rakha, “Oestrogen-regulated protein SLC39A6: A biomarker of good prognosis in luminal breast cancer,” *Breast Cancer Research and Treatment*, vol. 189, no. 3, pp. 621–630, 2021.
- [57] L. Liu, J. Yang, and C. Wang, “Analysis of the prognostic significance of solute carrier (SLC) family 39 genes in breast cancer,” *Bioscience Reports*, vol. 40, no. 8, p. BSR20200764, 2020.
- [58] C. Hogstrand, P. Kille, M. L. Ackland, S. Hiscox, and K. M. Taylor, “A mechanism for epithelial-mesenchymal transition and anoikis resistance in breast cancer triggered by zinc channel ZIP6 and STAT3 (signal transducer and activator of transcription 3),” *Biochemical Journal*, vol. 455, pp. 229–237, 10 2013.
- [59] S. U. Gerold, K. M. Taylor, I. A. Muraina, D. Brethour, T. Nimanon, S. Ziliotto, P. Kille, and C. Hogstrand, “Zinc transporter ZIP10 forms a heteromer with ZIP6 which regulates embryonic development and cell migration,” *Biochemical Journal*, vol. 473, no. 16, pp. 2531–2544, 2016.
- [60] K. M. Taylor, H. E. Morgan, K. Smart, N. M. Zahari, S. Pumford, I. O. Ellis, J. F. Robertson, and R. I. Nicholson, “The emerging role of the LIV-1 subfamily of zinc transporters in breast cancer,” *Molecular Medicine*, vol. 13, no. 7-8, pp. 396–406, 2007.
- [61] A. Sappok and U. Mahlknecht, “Ribavirin restores ESR1 gene expression and tamoxifen sensitivity in ESR1 negative breast cancer cell lines,” *Clinical Epigenetics*, vol. 3, p. 8, 2011.
- [62] W. Xu, M. Marcu, X. Yuan, E. Mimnaugh, C. Patterson, and L. Neckers, “Chaperone-dependent E3 ubiquitin ligase CHIP mediates a degradative pathway for c-ErbB2Neu,” *Cell Biology*, vol. 99, pp. 12847–12852, 2002.
- [63] W. Xia, S. Bacus, P. Hegde, I. Husain, J. Strum, L. Liu, G. Paulazzo, L. Lyass, P. Trusk, J. Hill, J. Harris, and N. L. Spector, “A model of acquired autoresistance to a potent ErbB2 tyrosine kinase inhibitor and a therapeutic strategy to prevent its onset in breast cancer,” *Proceedings of the National Academy of Sciences*, vol. 103, no. 20, pp. 7795–7800, 2006.



[Ewan Millar](#) is a Senior Staff Specialist Histopathologist with NSW Health Pathology at St George Hospital Sydney with expertise in breast cancer pathology and translational research and a strong interest in AI and digital pathology applications.



[Arcot Sowmya](#) is Professor in the School of Computer Science and Engineering, UNSW. Her major research interest is in the area of Machine Learning for Computer Vision and includes learning object models, feature extraction, segmentation and recognition based on computer vision, machine learning and deep learning. In recent years, applications in the broader health area are a focus, including biomedical informatics and rapid diagnostics in the real world. All of these areas have been supported by competitive, industry and government funding.



[Erik Meijering](#) (Fellow, IEEE), is a Professor of Biomedical Image Computing in the School of Computer Science and Engineering. His research focusses on the development of innovative computer vision and machine learning (in particular deep learning) methods for automated quantitative analysis of biomedical imaging data.



[Raktim Kumar Mondol](#) is a PhD candidate in Computer Science and Engineering, specializing in computer vision and bioinformatics. He completed his MEng in Engineering with High Distinction from RMIT University, Australia. Mondol’s research interests include histopathological image analysis, clinical prognosis prediction, and enhancing clinical understanding through the interpretability of computational models.