HHS Public Access

Author manuscript

Cancer Cell. Author manuscript; available in PMC 2023 November 17.

Published in final edited form as:

Cancer Cell. 2022 October 10; 40(10): 1095–1110. doi:10.1016/j.ccell.2022.09.012.

Artificial Intelligence for Multimodal Data Integration in Oncology

Jana Lipkova[1,2,3], **Richard J. Chen**[1,2,3,4], **Bowen Chen**[1,7], **Ming Y. Lu**[1,2,3,6], **Matteo Barbieri**[1], **Daniel Shao**[1,5], **Anurag J. Vaidya**[1,5], **Chengkuan Chen**[1,2,3], **Luoting Zhuang**[1,2], **Drew FK Williamson**[1,2,3], **Muhammad Shaban**[1,2,3], **Tiffany Y. Chen**[1,2,3], **Faisal Mahmood**[*,1,2,3,8]

[1]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

[2]Cancer Program, Broad Institute of Harvard and MIT, Cambridge, MA

[3]Data Science Program, Dana-Farber Cancer Institute, Boston, MA

[4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA

[5]Harvard-MIT Health Sciences and Technology (HST), Cambridge, MA

[6]Department of Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, MA

[7]Department of Computer Science, Harvard University, Cambridge, MA

[8]Harvard Data Science Initiative, Harvard University, Cambridge, MA

Abstract

In oncology, the patient state is characterized by a whole spectrum of modalities, ranging from radiology, histology, genomics to electronic-health records, each one providing additional insights. Current AI models, however, operate mainly in the realm of single modality, neglecting the broader clinical context, which inevitably diminishes their potential. Integration of different data modalities provides opportunities to increase robustness and accuracy of diagnostic and prognostic models, bringing AI closer to clinical practice. At the same time, AI models are capable of discovering novel patterns within and across modalities suitable for explaining differences in patient outcomes or treatment resistance. The insights gleaned from such models can guide exploration studies and contribute to the discovery of novel biomarkers and therapeutic targets. To support these advances, here we present a synopsis of AI methods and strategies for multimodal data fusion. We outline approaches for AI interpretability and directions for AI-driven exploration through multimodal data interconnections. We examine challenges towards clinical adoption and discuss possible emerging solutions.

[*]**Correspondence:** Faisal Mahmood, 60 Fenwood Road, Hale Building for Transformative Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02445, faisalmahmood@bwh.harvard.edu, www.mahmoodlab.org.

Declaration of interests

The authors declare no competing interests.

1 Introduction

Cancer is a highly complex disease involving a cascade of microscopic and macroscopic changes with mechanisms and interactions that are not yet fully understood. Cancer biomarkers provide insights on the state and course of disease in the form of quantitative or qualitative measurements, which consequently guide patient management. Based on their primary use, biomarkers can be diagnostic, predictive or prognostic. *Diagnostic biomarkers* stand at the first line of cancer detection and diagnosis, including examples such as prostate-specific antigen (PSA) values or neoplastic changes in tissue biopsy. *Predictive biomarkers* determine cancer aggressiveness (*e.g.* grade) or predict treatment response. For instance, microsatellite instability predicts response to immune checkpoint-inhibitor therapy in colorectal cancer (Marcus et al., 2019), whereas KRAS-mutations indicate resistance to anti-EGFR treatment (Van Cutsem et al., 2009). *Prognostic biomarkers* forecast risks associated with clinical outcomes such as survival, recurrence or disease progression. For example, Oncotype-DX assays are used to estimate recurrence scores and survival likelihood in breast cancer (Paik et al., 2004). Despite the vital role of biomarkers, even patients with similar profiles can be met with diverse treatment responses (Shergalis et al., 2018), recurrence rates (Roy et al., 2015) or treatment toxicity (Kennedy and Salama, 2020), while the underlying reasons for such dichotomies are often unknown. Given the increasing incidence rates of cancer, there is a crucial need to identify novel and more specific biomarkers.

The biomarker discovery process typically involves manual examination of potentially informative measurements and their association with clinical endpoints. For instance, the Nottingham grading system in breast cancer was determined through dedicated examination of thousands of histopathology slides, revealing association between morphological features and patient outcome. Although identification of each new biomarker represents a milestone in oncology, this process faces several challenges. Manual assessment is time- and resource-intensive, often without the possibility of translating observations from one cancer model to another. Cancer assessment is often qualitative with substantial inter-rater variability, which hinders reproducibility and contributes to inconsistent outcomes in clinical trials. Given the large complexity of medical data, current biomarkers are almost exclusively unimodal. However, constraining the biomarkers to a single modality can significantly reduce their clinical potential. For instance, glioma patients with similar genetic or histology profile can be met with diverse outcomes caused by macroscopic factors such as tumor location preventing full resection and irradiation or disruption of blood-brain barrier altering the efficacy of drug delivery (Miller, 2002).

Over the last years, AI has demonstrated great performance in many clinically relevant tasks. AI models are able to integrate complementary information and clinical context from diverse data sources to provide more accurate patient predictions (Fig. 1A). The clinical insights identified by successful models can be further elucidated through interpretability methods and quantitative analysis to guide and accelerate the discovery of new biomarkers (Fig. 1C–D). Similarly, AI models can reveal association across diverse modalities, such as relation between certain mutations and specific changes in cellular morphology (Coudray et al., 2018) or association between radiology findings and histology-specific tumor subtypes

(Ferreira-Junior et al., 2020; Hyun et al., 2019) or molecular features (Yan et al., 2021) (Fig. 1B). Such associations can identify accessible or non-invasive alternatives for existing biomarkers to support large scale population screenings or selection of patients for clinical trials (Fig. 1 E,F).

2 AI Methods in Oncology

AI methods can be categorized as supervised, weakly-supervised and unsupervised. To highlight the concepts specific to each strategy we present all methods in the framework of computer vision, as illustrated in Figure 2.

Supervised methods

Supervised methods learn a mapping from input data to predefined labels (*e.g.* cancer/non-cancer) using training samples. This includes *hand-crafted* and *deep-learning (DL) methods.*

Hand-crafted methods take as input a set of predefined features (*e.g.* cell shape or size) extracted from the data before the training, not the data itself. The training is performed with standard machine learning (ML) models, such as random forest (RF), support-vector machine (SVM) or multi-layer perceptron (MLP) (Bertsimas and Wiberg, 2020) (Fig. 2). Since the feature extraction is not part of the learning process, the models typically have simpler architecture, lower computation cost, and may require less training data than DL models. Additional benefit is a high level of interpretability since the predictive features can be related to the data. On the other hand, the feature extraction is time consuming and can translate human bias to the models. Moreover, human perception cannot be easily captured by a set of computer rules, often leading to simpler features. Since the features are usually tailored to the specific disease, the models cannot be easily translated to other tasks or malignancies. Despite the popularity of DL methods, in many applications the hand-crafted methods are sufficient and preferred due to their simplicity and ability to learn from smaller datasets.

Deep-learning methods operate directly on raw data or in the case of large histopathology scans on image patches. Here we focus on convolutional neural networks (CNNs), the most common DL strategy for image analysis. In CNNs the predictive features are not defined, and the model alone must determine which concepts and features are useful for explaining relations between inputs and outputs, using supervision from pixel-level labels. For instance, in Figure 2 each training whole-slide image (WSI) is manually annotated to outline the tumor region. The WSI is then partitioned into rectangular patches and each patch is assigned with a label “cancer” or “no-cancer” determined by the tumor annotation. The majority of CNNs have a similar architecture, consisting of alternating convolutional, pooling and non-linear activation layers, followed by a small number of fully-connected layers. A convolution layer serves as a feature extractor, while the subsequent pooling layer condenses the features into the most relevant ones. The non-linear activation function allows the model to explore complex relations across features. Fully-connected layers then perform the end-task such as cancer classification. The main strength of CNNs is their ability to extract rich feature representations from raw data, resulting in lower pre-processing cost, higher flexibility and often superior performance over hand-crafted

models. The potential limitations come from the model's reliance on pixel-level annotations which are time-intensive and might be affected by inter-rater variability and human bias. Moreover, predictive regions for many clinical outcomes, such as survival or treatment resistance, may be unknown. CNNs are also often criticized for the lack of interpretability, since neither the model's decision mechanisms nor the learned predictive features can (yet) be satisfyingly interpreted at the human-level. Despite these limitations, CNNs come with impressive performance, contributing to widespread usage in many medical applications.

Weakly-supervised methods

Supervised methods operate under strong supervision, where each data point (*e.g.* patch or pixel) is assigned a label. Such strong labels, however, are not routinely created during the clinical workflow and their construction is expensive. Weakly-supervised methods allow to train models with weak, patient-level labels (such as diagnosis or survival), avoiding the need for manual data annotations. The most common weakly-supervised methods include *graph convolutional networks (GCNs)*, *multiple-instance learning (MIL)* and *Vision Transformers (VITs)*.

Graph Convolutional Networks: Graphs can be used to explicitly capture spatial tissue structures and encode relation between objects. A graph is defined by nodes connected by edges. In histology, a node can represent a cell, image-patch or even tissue region. Edges encode spatial relation and interaction between nodes (Zhang et al., 2019). The graph, combined with the patient-level labels, is processed by a GCN (Ahmedt-Aristizabal et al., 2021) which can be seen as a generalization of CNNs that operate on unstructured graphs. In GCNs, feature representations of a node are updated by aggregating information from neighboring nodes. The updated representations then serve as input for the final classifier. In contrast to CNNs and MIL, GCNs incorporate larger context and spatial tissue structure. This can beneficial in tasks where the clinical context span beyond the scope of a single patch (*e.g.* Gleason score). On the other hand, the interdependence of the nodes in GCNs come with higher training costs and memory-requirements, since the nodes cannot be processed independently. As a consequence, contemporary GCNs are relatively shallow and typically build only on sub-samples of the WSI.

In Multiple-instance learning (Carbonneau et al., 2018; Cheplygina et al., 2019), all patches from a given patient are grouped into a "bag" carrying a patient-level label. The labels of individual patches (referred to as instances) are not known. The label of a bag is assumed positive if there is at least one positive instance in the bag. The goal of the model is to predict the bag label and to identify which instances are relevant for the given prediction. The MIL models comprise of three main modules: feature extraction, aggregation, and prediction. The first module is used to embed the image patches into lower-dimensional embeddings to reduce memory requirements and facilitate fast training. This is often achieved through transfer learning (Weiss et al., 2016) and self-supervision (see next section). The patch-level embeddings are aggregated to create patient-level representations which serve as input for the final classification module. There exist several aggregation strategies. For instance, Fig. 2 shows attention-based pooling (Ilse et al., 2018) where two fully-connected networks are used to learn the relative importance of each image patch

toward the model predictions – the so-called attention score. The patch-level representations, weighted by the corresponding attention score, are summed up to build the patient-level representation. The attention-scores can be also used for model interpretability (see Section 4).

Vision transformers (Dosovitskiy et al., 2020; Vaswani et al., 2017) are type of attention-based learning. In contrast to MIL, where patches are assumed independent and identically distributed, VITs account for correlation and context among patches. The main components of VITs include positional encoding, self-attention and multi-head self-attention. Positional encoding learns the spatial structure of the image and the relative distance among patches. Self-attention mechanism determines the relevance of each patch while also accounting for the context and contribution from the other patches. Multi-head self-attentions simultaneously deploys multiple self-attention blocks to account for different types of interactions between the patches and combines them into a single self-attention output. A typical VIT architecture is shown in Fig. 2. A WSI is converted into a series of patches, each coupled with positional information. Learnable encoders map each patch and its position into a single embedding vector, referred to as a token. Additional token is introduced for the classification task. The class token together with the patch tokens are fed into the transformer encoder to compute multihead self-attention and output the learnable embeddings of patches and the class. The output class token serves as a slide-level representations used for the final classification. The transformer encoder consists of several stacked identical blocks. Each block includes multi-head self-attention and MLP, along with layer-normalization and residual connections. The positional encoding and multiple self-attention heads allows to incorporate spatial information, increase the context and robustness (Li et al., 2022; Shamshad et al., 2022) of VIT methods over MIL. On the other hand, VIT tend to be more data hungry (Dosovitskiy et al., 2020).

Often it is difficult to know *a priori* which model is most suitable for the given task. While GCNs methods tend to perform better in tasks that require larger spatial context, such as Gleason grade, Bulten *et al.* (Bulten et al., 2022) showed that MIL can predict the Gleason score with clinical-grade accuracy even without the graph-based models. Without proper benchmarks offering head-to-head comparison of methods on the same data and tasks, it is difficult to make strong conclusions.

Weakly-supervised methods offer several benefits. The liberation from manual annotations reduces the cost of data pre-processing and mitigates the bias and inter-rater variability. Consequently, the models can be easily applied to large datasets, diverse tasks and also situations where the predictive regions are unknown. Since the models are free to learn from the entire scan, they can identify predictive features even beyond the regions typically evaluated by pathologists. The great performance demonstrated by weakly-supervised methods suggests that many tasks can be addressed without expensive manual annotations or hand-crafted features.

Unsupervised methods

Unsupervised methods explore structures, patterns and subgroups in data without relying on any labels. This includes *self-supervised* and *fully-unsupervised* strategies.

Self-supervised methods are designed to improve performance of fully/weakly-supervised models, where limited data are available and overfitting might be of concern. Self-supervised methods exploit available un-labeled data to learn high-quality image features and then transfer this knowledge to supervised models. To achieve this, supervised methods such as CNNs are used to solve various pretext tasks (Jing and Tian, 2019) for which the labels are generated automatically from the data. For instance, a patch can be removed from an image and a network is trained to predict the missing part of the image from its surroundings, using the actual patch as label. The patch-prediction has no direct clinical relevance, but it guides the model to learn general-purpose features of image characteristics which can be beneficial for other practical tasks. The early layers of the network usually capture general image features, while the later layers pick features relevant for the pretext task. The later layers are thus excluded, while the early layers serve as feature extractors in the fully/weakly supervised models (*i.e.* transfer learning).

Fully-unsupervised methods search for correlations, structures, and subgroups in the data itself, without having a prescribed task-specific outcome or label. The most common unsupervised methods include clustering and dimensionality reduction. Clustering methods (Rokach and Maimon, 2005) partition data into subgroups such that the similarity within the subgroup and separation between subgroups are maximized. Although the output clusters are not task-specific, they can reveal different cancer subtypes or patient subgroups. The aim of dimensionality reduction is to obtain low-dimensional representation capturing the main characteristics and correlations in the data. Common DL approach include autoencoders. An autoencoder (Hinton and Salakhutdinov, 2006) consists of two neural networks: encoder and decoder. The encoder learns to encode data into a lower-dimensional feature vector similar to feature extraction in CNNs. At the same time, a decoder is trained to reconstruct the original input data from the lower-dimensional vector. Along this process, the encoder learns to identify the most relevant aspects of the data. The encoder can be used as the feature extractor in supervised methods (Sharmay et al., 2021), similar to the encoder obtained through self-supervised learning. There is no guarantee that an unsupervised model will result in a desired output; rather, these models serve as tools for data exploration and feature learning.

3 Multimodal data fusion

The aim of multimodal data fusion is to extract and combine complementary, contextual information across different modalities for better decision making. This is of particular relevance in medicine where similar findings in one modality may have diverse interpretations in combination with other modalities. For instance, mass in lung X-ray scans can be interpreted as pneumonia, COVID-19, or also tumor – depending on accompanying symptoms such as fever, elevated blood count, or biopsy assessment. Similarly for cancer biomarkers, *IDH1* mutation status or histology profile alone are insufficient for explaining the variance in patient outcomes, whereas the combination of both have been recently used to redefine the WHO classification of diffuse glioma (Louis et al., 2016). AI offers automated and objective way for incorporating complementary information and clinical context from diverse data for improved predictions. The models can also utilize supplementary information in modalities; if unimodal data are noisy or incomplete,

supplementing redundant information from other modalities can improve the robustness and accuracy of the predictions. AI-driven data fusion strategies (Baltrušaitis et al., 2018) can be divided as *early, late,* and *intermediate* (see Fig. 3).

Early fusion integrates information from all modalities at the input level before feeding it into a single model. The modalities can be represented as raw data, hand-crafted or deep-features. The joint representation is build through operations such as vector concatenation, element-wise sum, element-wise multiplication (Hadamard Product) or bilinear pooling (Kronecker Product) (Huang et al., 2020a; Ramachandram and Taylor, 2017). In early fusion, only one model is trained which simplify the design process. However, it is assumed that the single model is well suited for all modalities. Early fusion requires a certain level of alignment or synchronization between the modalities. Although this is more obvious in other domains, such as synchronization of audio and visual signals in speech recognition, it is also relevant in clinical settings. If the modalities come from significantly different time points, such as pre- and post-interventions, then early fusion might not be appropriate choice.

Applications of early-fusion include integration of similar modalities such as multimodal, multiview ultrasound images for breast cancer detection (Qian et al., 2021) or fusion of structural CT and/or MRI data with metabolic PET scans for cancer detection (Le et al., 2017), treatment planning (Lipková et al., 2019), or survival prediction (Nie et al., 2019). Other examples include fusion of imaging data with electronic medical records (EMRs) such as integration of dermoscopic images and patient data for skin lesion classification (Yap et al., 2018), or fusion of cervigram and EMRs for cervical dysplasia diagnosis (Xu et al., 2016). Several studies investigate the correlation between changes in gene expression and tissue morphology, integrating genomics data with histology and/or radiology images for cancer classification (Khosravi et al., 2021), survival (Chen et al., 2020b, 2021c) and treatment response (Feng et al., 2022; Sammut et al., 2022) prediction.

Late fusion, also known as decision-level fusion, trains a separate model for each modality and aggregates the prediction from individual models for the final prediction. The aggregation can be performed by averaging, majority voting, Bayes-based rules (Ramanathan et al., 2022), or learned models such as MLP. Late fusion allows to use different model architecture for each modality and does not pose any constrains on data synchronization, making it suitable for systems with large data heterogeneity or modalities from different time points. In case of missing or incomplete data, late fusion retains the ability to make predictions since each model is trained separately and aggregations, such as majority voting, can be applied even if a modality is missing. Similarly, inclusion of a new modality can be performed without the need to retrain the full model. Simple covariates, such as age or gender, are often included through late fusion due to its simplicity (see Fig. 3 B). If the unimodal data do not complement each other or do not have strong interdependencies, late fusion might be preferable thanks to the simpler architecture and smaller number of parameters compared to other fusion strategies. This is also beneficial in situations with limited data. Furthermore, errors from individual models tend to be uncorrelated, resulting in potentially lower bias and variance in late-fusion predictions. In situations when information density varies significantly across modalities, predictions from shared representations can be heavily influenced by the most dominant modality. In late

fusion the contribution from each modality can be accounted for in a controlled manner by setting equal or diverse weights per modality in the aggregation step.

Examples of late fusion include integration of imaging data with non-imaging inputs, such as fusion of MRI scans and PSA-blood tests for prostate cancer diagnosis (Reda et al., 2018), integration of histology scans and patient gender for inferring origin of metastatic tumors (Lu et al., 2021), fusion of genomics and histology profiles for survival prediction (Chen et al., 2021c; Shao et al., 2019), combination of pre-treatment MRI or CT scans with EMRs for chemotherapy response prediction (Joo et al., 2021) and survival estimation (Nie et al., 2016).

Intermediate fusion is a strategy where the loss from the multimodal model propagates back to the feature extraction layer of each modality to iteratively improve feature representations under the multimodal context. For comparison, in early and late fusion the unimodal embeddings are not affected by the multimodal information. Intermediate fusion can combine individual modalities at different levels of abstractions. Moreover, in systems with three and more modalities the data can be fused either all at once (Fig. 3C) or gradually across different levels (Fig. 3D). The intermediate single-level fusion is similar to early fusion, however, in the early fusion the unimodal embeddings are not affected by the multimodal context. Gradual fusion allows to combine data from highly correlated channels at the same level, forcing the model to consider the cross-correlations between specific modalities, followed by fusion with less correlated data in later layers. For instance, in Fig. 3D, genomics and histology data are fused first, to account for the interplay between mutations and changes in the tissue morphology, while the relation with the macroscopic radiology data is considered in the later layer. Gradual fusion has shown improved performance over single-level fusion in some applications (Joze et al., 2020; Karpathy et al., 2014). *Guided fusion* uses information form one modality to guide feature extraction from another modality. For instance, in Fig. 2E, genomics information guides the selection of histology features. The motivation is that different tissue regions might be relevant in the presence of specific mutations. Guided fusion learns co-attention scores that reflect relevance of different histology features in the presence of specific molecular information. The co-attention scores are learned with the multimodal model, where the genomics feature and the corresponding genomics-guided histology features are combined for the final model predictions.

Examples of intermediate fusion include integration of diverse imaging modalities such as fusion of PET and CT scans in lung cancer detection(Kumar et al., 2019), fusion of MRI and ultrasound images in prostate cancer classification (Sedghi et al., 2020), or combination of multimodel MRI scan in glioma segmentation (Havaei et al., 2016). Fusion of diverse multi-omics data was used for cancer subtyping (Liang et al., 2014) or survival prediction (Lai et al., 2020). Genomics data have been used in tandem with histology (Vale-Silva and Rohr, 2021) or mammogram (Yala et al., 2019) images for improved survival prediction. Guided fusion of different radiology modalities was used to improve segmentation of liver lesions (Mo et al., 2020) and anomalies in breast tissue (Lei et al., 2020). EMRs were used to guide feature extraction from dermoscopic (Zhou and Luo, 2021) and mamography (Vo et al., 2021) images to improve detection and classification of lesions. Chen *et al.* (Chen et

al., 2021b) used genomics information to guide selection of histology features for improved survival prediction in multiple cancer types.

There is no conclusive evidence that one fusion type is ultimately better than the others, as each type is heavily data- and task-specific.

4 Multimodal Interpretability

Interpretability is a crucial component of AI development, deployment and validation. With the ability of AI models to discover their own features, there is a worry that the models might use spurious shortcuts for predictions, instead of learning clinically relevant aspects. Such models might fail to generalize when presented with new data or discriminate against underrepresented populations (Banerjee et al., 2021; Chen et al., 2021a). On the other hand, the models can discover novel and clinical relevant insights. Here we present a brief overview of different levels of interpretability used in oncology (Fig. 4), while technical details can be find in the recent review (Arrieta et al., 2020).

In **histopathology**, VIT or MIL can reveal relative importance of each image patch for the model predictions. The attention scores can be mapped to their spatial location to obtain slide-level attention heatmaps as shown in Fig. 4A, where a MIL model was trained to classify cancer suptyes in WSIs. Although no manual annotations were used, the model learned to identify morphology specific for each cancer type and to discriminate between normal and malignant tissues. Class activation methods (CAMs), such as GradCAM (Selvaraju et al., 2017) or GradCAM++ (Chattopadhay et al., 2018), allow to determine importance of the model inputs (*e.g.* pixels) by computing how the changes in the inputs affect the model outputs for each prediction class. GradCAM is often used in tandem with the guided-backpropagation method, so-called Guided-GradCAM (Selvaraju et al., 2016), where the guided-backpropagation determines the pixel-level importance inside the predictive regions specified by the GradCAM. This is illustrated in Fig. 4B, where a CNN was trained to classify cancer subtypes in image patches. For comparison, in the attention methods the importance of each instance is determine during the training while the CAM-based methods are model-agnostic *i.e.* independent on the model training.

In **radiology**, the interpretability methods are similar to those used in histology. The attention-scores can reflect the importance of slides in a 3D scan. For instance, in Fig. 4D. MIL model was trained to predict survival in glioma patients (Zhuang et al., 2022). The model considered the 3D MRI scan as a bag, where the axial slides are modeled as individual instances. Even in the absence of manual annotations, the model has placed high-attention to the slides with tumor, while low-attention was assigned to healthy tissue. CAM-based methods can be consequently deployed to localize the predictive regions within individual slides (Fig. 4F).

Molecular data can be analyzed by Integrated Gradient method (Sundararajan et al., 2017) which computes attribution values indicating how changes in specific inputs impact the model outputs. For the regression tasks, such as survival analysis, the attribution values can reflect the magnitude of the importance as well as the direction of the impact: features

with positive attribution increase the predicted output (*i.e.* higher risk) while features with negative attribution reduce the predictive values (*i.e.* lower risk). At the patient-level, these is visualized as a bar plot, where the y-axis corresponds to the specific features (ordered by their absolute attribution value) and the x-axis shows the corresponding attribution values. At the population-level, the attributions plots depict the distribution of the attributions scores across all subjects. Fig. 4C shows the attribution plots for most genomics features used for survival prediction in glioma patients (Chen et al., 2021c). Other tabular data, such as **hand-crafted features** or values obtained from **EMR** can be interpreted in the same way. EMRs can be also analyzed by natural language processing (NLP) methods, such as Transformers, where the attention scores determine importance of specific words in the text (Fig. 4E.)

In **multimodal models**, the attribution plots can also determine contribution of each modality towards the model predictions. All previously mentioned methods can be used in multimodal models to explore interpretability within each modality. Moreover, shifts in feature importance under unimodal and multimodal settings can be investigated to analyze impact of the multimodal context.

The interpretability methods usually come without any accuracy measures and thus it is important not to over-interpret them. While CAM- or attention-based methods can localize the predictive regions they cannot specify which features are relevant, *i.e.* they can explain *where* but not *why*. Moreover, there is no guarantee that all high-attention/attribution region carry clinical relevance. High scores just mean that the model has considered these regions more important than others.

5 Multimodal data interconnection

The aim of multimodal data interconnection is to reveal associations and shared information across modalities. Such associations can provide new insights on cancer biology and guide discovery of novel biomarkers. Although there are many approaches for data explorations, here we illustrate few possible directions (Fig. 5).

Morphological associates:

Malignant changes often propagate across different scales; oncogenic mutations can affect cell behaviour, which in turn reshape tissue morphology or tumor microenviroment visible in histology images. Consequently, the microscopic changes might impact tumor metabolic activity and macroscopic appearance detectable by PET or MRI scans. The feasibility of AI methods to identify associations across modalities was first demonstrated by Coudray *et al.* (Coudray et al., 2018) who showed that certain mutations in lung cancer can be inferred directly from hematoxylin and eosin (H&E)-stained WSIs. Other studies have followed shortly, predicting the mutation status from WSIs in liver (Chen et al., 2020a), bladder (Loeffler et al., 2021), colorectal (Jang et al., 2020), thyroid cancer (Tsou and Wu, 2019), as well as pan-cancer pan-mutation studies attempting to predict any genetic alternation in any tumor type (Fu et al., 2020; Kather et al., 2020). Additional molecular biomarkers, such as gene expression (Anand et al., 2020; Binder et al., 2021; Schmauch et al., 2020), hormone-receptor status (Naik et al., 2020), tumor mutational burden (Jain and

Massoud, 2020), and microsatellite instability (Cao et al., 2020; Echle et al., 2020) have also been inferred from WSIs (Murchan et al., 2021). In radiology, AI models have predicted *IDH* mutation and *1p/19q* co-deletion status from preoperative brain MRI scans (Bangalore Yogananda et al., 2020; Yogananda et al., 2020), *BRCA1* and *BRCA2* mutational status from breast mammography (Ha et al., 2017) and MRI (Vasileiou et al., 2020) scans, while *EGFR* and *KRAS*-mutation have been detected from CT scans in lung (Wang et al., 2019) and colorectal (He et al., 2020) cancer.

By discovering presence of morphological associates across modalities, AI models can enhance exploratory studies and reduce the search space for possible biomarker candidates. For instance, in Fig. 5A, the AI has revealed that one of the studied mutations can be reliably inferred from WSI. Although the predictive features used by the model might be unknown, interpretability methods can provide additional insights. Attention-heatmaps can reveal tissue regions relevant for the prediction of the specific mutation. Distinct tissue structures and cell types withing high and low-attention regions can be identified, and their properties such as nuclei shape or volume can be further extracted and analyzed. Clustering or dimensionality reduction methods can be deployed to examine the promising features, potentially revealing associations between mutation status and distinct morphological features. The identified morphological associates can serve as cost-efficient biomarker surrogates to support screening in low-to-middle income settings or reveal new therapeutic targets.

Non-invasive alternatives:

Similarly, AI can discover relationships between non-invasive and invasive modalities. For instance, AI models were used to predict histology sutypes or grades from radiomics features in lung (Sha et al., 2019), brain (Lasocki et al., 2015), liver (Brancato et al., 2022) and other cancers (Blüthgen et al., 2021). The predictive image regions can be further analyzed to identify textures and patterns with possible diagnostic values (see Fig. 5B), which in turn can serve as non-invasive surrogates for existing biomarkers.

Outcome associates:

Benefits of personalized medicine are often limited by paucity of biomarkers able to explain dichotomies in patient outcomes. On the other hand, AI models are demonstrating great performance in predicting clinical outcomes, such as survival (Lai et al., 2020), treatment response (Echle et al., 2020), recurrence (Yamamoto et al., 2019), and radiation toxicity (Men et al., 2019) using unimodal and multimodal (Chen et al., 2020b, 2021c; Joo et al., 2021; Mobadersany et al., 2018) data. These works imply the feasibility of AI models to discover relevant prognostic patterns in data, which might be elucidated by interpretability methods. For instance in Fig. 5C, a model is trained to predict survival from histology and genomics data. Attention heatmaps reveal tissue regions related to low and high risk patients groups while the molecular profiles are analyzed through attribution plots. The predictive tissue regions can be further analyzed by examining tissue morphology, cell subtypes or other human-intepretable data characteristics. Tumor-infiltrating lymphocytes can be estimated through co-localization of tumor and immune cells to specify immune hot and cold tumors. Attribution of specific modalities as well as shift in feature importance

in unimodal vs. multimodal data can be explored to determine the influence of multimodal contextualization.

Such exploration studies have already provided new clinical insights. For instance, Geessink *et al.* (Geessink et al., 2019) showed that tumor-stroma ratio can serve as independent prognosticator in rectal cancer, while the ratio of tumor area to metastatic lymph node regions has prognostic value in gastric cancer (Wang et al., 2021). Other morphological features, such as arrangement of collagen fibers in breast histology (Li et al., 2021) or spatial tissue organization in colorectal tissue (Qi et al., 2021) have been identified as possible biomarkers for aggressiveness or recurrence.

Early predictors:

AI can also explore diverse data acquired prior to patient diagnosis to identify potential predictive risk factors. EMRs provide rich information on patient history, medication, allergies or immunization, which might contribute to patient outcome. Such diverse data can be efficiently analyzed by AI models to search for distinct patient subgroups (Fig. 5 D). Identified subgroups can be correlated with different patient outcomes, while attribution plots can identify relevance of different factors at the patient and population level. Recently, Placido *et al.* (Placido et al., 2021) showed the feasibility of AI to identify patients with a higher risk of developing pancreatic cancer by exploration of EMR. Similarly EMRs were used to predict treatment response prediction (Chu et al., 2020) or length of hospital stay (Alsinglawi et al., 2022). The identified novel predictive risk factors can support large-scale population screenings and early preventive care.

Outside of the hospital setting, smartphones and wearable devices offer another great opportunity for real-time and continuous patient monitoring. Changes in the measured values, such as decrease in patient step counts, have been shown as robust predictors of worse clinical outcome, treatment toxicity, and increased risk of hospitalization (Low, 2020). Furthermore, the modern wearable devices are continually expanding their functionality including measurements of temperature, stress levels, blood-oxygen saturation, or electrocardiograms. These measurements can be analyzed in tandem with clinical data to search for risk factors indicating early stages of increased toxicity or treatment resistance, to allow personalized interventions during the course of treatment. Research on personalized monitoring and nanotechnologies is investigating novel directions such as detection of patient measurements in sweat (Xu et al., 2019) or ingestible sensors to monitor medication compliance and drug absorption (Weeks et al., 2018). All these novel devices provide useful insights on the patient state, which could be analyzed in larger clinical context through AI models.

6 Challenges and path towards clinic

The path of AI into clinical practice is still laden with obstacles, many of which are amplified in the presence of multimodal data. While several recent works discus challenges, such as fairness and datasets shifts (Banerjee et al., 2021; Chen et al., 2021a; Cirillo et al., 2020; Howard et al., 2021; Mehrabi et al., 2021; Zhang et al., 2018), limited interpretability (Adebayo et al., 2018; Linardatos et al., 2020; Reyes et al., 2020) or regulatory guidelines

(Cruz Rivera et al., 2020; Topol, 2020; Wu et al., 2021), here we focus on challenges specific to multimodal learning.

Missing data

The challenge of missing data refers to the absence of part of modality or complete unavailability of one or more modalities. The missing data impact both, the model training and deployment since majority of existing AI models cannot handle missing information. Moreover, the need to train models with complete multimodal data significantly constrains the size of the training datasets. For instance, The Cancer Genome Atlas (TCGA) dataset contains over 900 WSIs from glioma patients, however, only 127 cases contain corresponding radiology and genomics information. The incomplete modalities still contain valuable information, and the inability to deploy them pose a significant limitation, especially in fields like medicine where data are precious. Below we discuss two strategies for handling missing data.

Synthetic data generation can be used to complement the missing information. If part of an image is corrupted or if specific mutations are not reported, the missing information can be fabricated from the remaining data. If a whole modality is missing, its synthetic version can be derived from existing similar modalities. For instance, Haan *et. al* (de Haan et al., 2021) trained a supervised model for translation of H&E-stains into special stains, using the special stains as ground truth labels. The model was trained on pairs of perfectly aligned data obtained *e.g* through re-staining of the same slides. If paired data are not available, unsupervised methods such as cycle generative adversarial networks (GANs) (Zhu et al., 2017) can be used. While synthetic data can improve performance of detection and classification methods, they are less suitable for outcome prediction or biomarker exploration, where the predictive features are not well understood and thus there is no guarantee that the synthetic data contain the relevant disease characteristics. Moreover, the algorithms can hallucinate malignant features also into the supposedly normal synthetic images (Cohen et al., 2018) which can further hurt prediction results.

Drop-out based methods, aims to make models robust to missing information. For instance, Choi *et. al*(Choi and Lee, 2019) proposed EmbraceNet model which can handle incomplete or missing data during training and deployment. The EmbraceNet model probabilistically selects partial information from each modality and combines it into a single representation vector, which then serves as an input for the final decision model. When missing or invalid data is encountered, it is not sampled; instead, other more complete modalities are used to compensate for the missing data. The probabilistic data selection has also a regularization effect, similar to the dropout mechanism. The authors have tested the EmbraceNet method on natural images and videos, whereas its utility in clinical settings, although very promising, is yet to be explored.

Data alignment

To investigate cancer processes across different scales and modalities, a certain level of data alignment is required This might include alignment of: i) similar or ii) diverse modalities.

Alignment of similar modalities typically involves different imaging modalities of the same system. This is usually achieved through image registration, which is formulated as an optimization problem minimizing the difference between the modalities.

In *radiology*, rigid anatomical structures can guide the data alignment. For instance, registration of MRI and PET brain scans is usually achieved with high accuracy, even with simple affine registration, thanks to the rigid skull. The situation is more complex in the presence of motion and deformations, *e.g.* breathing in lung imaging or changes in the body posture between scanning sessions. Alignment of such data usually require deformable registrations using natural or manually placed landmarks for guidance. A particularly challenging situation is the registration of scans between interventions, *e.g.* registration of preoperative and postoperative scans, which exhibit lot of non-trivial changes due to tumor resection, response to treatment, or tissue compression (Haskins et al., 2020).

In *histology*, each stained slide usually come from different tissue cut. Even in consecutive tissue cuts there are substantial differences in the tissue appearance caused by changes in tissue microenviroment or artefacts such as tissue folding, tearing or cutting (Taqi et al., 2018), which all complicate data alignment. Automated registration of histology images is still an open problem (Borovec et al., 2020) and thus many studies deploy non-algorithmic strategies such as clearing and re-staining of the tissue slides (Hinton et al., 2019). A newly emerging direction is stainless imaging, including approaches such as ultraviolet microscopy (Fereidouni et al., 2017), stimulated Raman histology (Hollon et al., 2020) or colorimetric imaging (Balaur et al., 2021).

Alignment of diverse modalities refers to the integration of data from different scales, time points or measurements. Often an acquisition of one modality results in the destruction of the sample, preventing collection of multiple measurements of the same system. For instance, most omics measurements require tissue disintegration which inevitably affects the possibilities to study relations between cell appearance and corresponding gene expressions. Here, cross-modal autoencoders can be used to enable integration and translation between arbitrary modalities. Cross-modal autoencoders (Dai Yang et al., 2021) build a pair of encoder-decoder networks for each modality, where the encoder maps each modality into a lower-dimensional latent space, while the decoder maps it back into the original space. A discriminative objective function is used to ensure that different modalities are matched in the shared latent space. Having the shared latent space, one can combine an encoder of one modality with the decoder of another modality to translate one modality to another one. Yang *et al.* (Dai Yang et al., 2021) demonstrated translations between single-cell chromatin images and RNA-seq data. The feasibility and utility of the cross-modal autoencoders are yet to be tested with more advanced data. However, if proven potent, they hold great potential to address challenges with alignment and harmonization of data from diverse sources.

Transparency and prospective clinical trials

Given the complexity of AI methods, it is possible that their mechanisms will not be fully understood in the near future. However, many aspects in medicine are not fully understood either (Kirkpatrick, 2005). And thus, rather than dwelling on the full opacity of AI methods,

we should advocate for their rigorous validation under randomized clinical trials, same as is done for other medical devices and drugs. Prospective trials will allow to stress-test the models under real-world conditions, compare their performance against standard-of-care practice, estimate how clinicians interact with the AI-tool and find the best way in which the models can enhance, rather than disturb the clinical workflow. In the case of biomarker surrogates discovered by AI methods, regulation paths similar to “me-too” drugs and devices (Aronson and Green, 2020) should be required to ensure comparable-level of performance. Transparency about study design and the used data are necessary to determine the intended use and conditions under which the model performance has been verified to prevent misuse under non-tested conditions. Prospective clinical trials are inevitable to truly demonstrate and quantify the added value of AI models, which will in turn increase trust and motivation of practitioners towards the AI tools.

7 Outlook and Discussion

AI has potential to impact the whole landscape of oncology, ranging from prevention to intervention. AI models can explore complex and diverse data to identify factors related with high risks of developing cancer to support large population screenings and preventive care. The models can further reveal associations across modalities to help identify diagnostic or prognostic biomarkers from easily accessible data to improve patient risk stratification or selection for clinical trials. In a similar way, the models can identify non-invasive alternatives to existing biomarkers to minimize invasive procedures. Prognostic models can predict risk factors or adverse treatment outcomes prior to interventions to guide patient management. Information acquired from personal wearable devices or nanotechnologies could be further analyzed by AI models to search for early signs of treatment toxicity or resistance, with other great application yet to come.

As with any great medical advances, there is a need for rigorous validation and examination under prospective clinical trials to verify the promises made by AI models. The role of AI in advancing the field of oncology is not autonomous, rather it is a partnership between models and human experience that will drive the further progress. AI-models come with limitations and challenges, however, these should not intimidate but rather inspire us. With increasing incidence rates of cancer, it is our obligation to capitalize on benefits offered by AI methods to accelerate discovery and translation of advances into clinic practice to serve patients and health care providers.

Acknowledgements

This work was supported in part by the BWH President’s Fund, National Institute of General Medical Sciences (NIGMS) R35GM138216 (to F.M.), Google Cloud Research Grant, Nvidia GPU Grant Program and internal funds from BWH and Massachusetts General Hospital (MGH) Pathology.

References

Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, and Kim B. Sanity checks for saliency maps. In Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, and Garnett R, editors, Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018.

Ahmedt-Aristizabal D, Armin MA, Denman S, Fookes C, and Petersson L. A survey on graph-based deep learning for computational histopathology. Computerized Medical Imaging and Graphics, page 102027, 2021. [PubMed: 34959100]

Alsinglawi B, Alshari O, Alorjani M, Mubin O, Alnajjar F, Novoa M, and Darwish O. An explainable machine learning framework for lung cancer hospital length of stay prediction. Scientific reports, 12(1):1–10, 2022. [PubMed: 34992227]

Anand D, Kurian NC, Dhage S, Kumar N, Rane S, Gann PH, and Sethi A. Deep learning to estimate human epidermal growth factor receptor 2 status from hematoxylin and eosin-stained breast tissue images. Journal of Pathology Informatics, 11, 2020.

Aronson JK and Green AR. Me-too pharmaceutical products: History, definitions, examples, and relevance to drug shortages and essential medicines lists. British Journal of Clinical Pharmacology, 86(11):2114–2122, 2020. [PubMed: 32358800]

Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion, 58:82–115, 2020.

Balaur E, O'Toole S, Spurling AJ, Mann GB, Yeo B, Harvey K, Sadatnajafi C, Hanssen E, Orian J, Nugent KA, et al. Colorimetric histology using plasmonically active microscope slides. Nature, 598 (7879):65–71, 2021. [PubMed: 34616057]

Baltrušaitis T, Ahuja C, and Morency L-P. Multi-modal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence, 41(2):423–443, 2018. [PubMed: 29994351]

Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen L-C, Correa R, Dullerud N, Ghassemi M, Huang S-C, Kuo P-C, et al. Reading race: Ai recognises patient's racial identity in medical images. arXiv preprint arXiv:2107.10356, 2021.

Bangalore Yogananda CG, Shah BR, Vejdani-Jahromi M, Nalawade SS, Murugesan GK, Yu FF,Pinho MC, Wagner BC, Mickey B, Patel TR, et al. A novel fully automated mri-based deep-learning method for classification of idh mutation status in brain gliomas. Neuro-oncology, 22(3):402–411, 2020. [PubMed: 31637430]

Bertsimas D and Wiberg H. Machine learning in oncology: Methods, applications, and challenges. JCO Clinical Cancer Informatics, 4:885–894, 2020. [PubMed: 33058693]

Binder A, Bockmayr M, Hägele M, Wienert S, Heim D, Hellweg K, Ishii M, Stenzinger A, Hocke A, Denkert C, et al. Morphological and molecular breast cancer profiling through explainable machine learning. Nature Machine Intelligence, 3(4):355–366, 2021.

Blüthgen C, Patella M, Euler A, Baessler B, Martini K, von Spiczak J, Schneiter D, Opitz I, and Frauenfelder T. Computed tomography radiomics for the prediction of thymic epithelial tumor histology, tnm stage and myasthenia gravis. PloS one, 16(12):e0261401, 2021. [PubMed: 34928978]

Borovec J, Kybic J, Arganda-Carreras I, Sorokin DV, Bueno G, Khvostikov AV, Bakas S, Eric I, Chang C, Heldmann S, et al. Anhir: automatic non-rigid histological image registration challenge. IEEE transactions on medical imaging, 39(10):3042–3052, 2020. [PubMed: 32275587]

Brancato V, Garbino N, Salvatore M, and Cavaliere C. Mri-based radiomic features help identify lesionsand predict histopathological grade of hepatocellular carcinoma. Diagnostics, 12(5):1085, 2022. [PubMed: 35626241]

Bulten W, Kartasalo K, Chen P-HC, Ström P, Pinckaers H, Nagpal K, Cai Y, Steiner DF, van Boven H, Vink R, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. Nature medicine, pages 1–10, 2022.

Cao R, Yang F, Ma S-C, Liu L, Zhao Y, Li Y, Wu D-H, Wang T, Lu W-J, Cai W-J, et al. Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in colorectal cancer. Theranostics, 10(24):11080, 2020. [PubMed: 33042271]

Carbonneau M-A, Cheplygina V, Granger E, and Gagnon G. Multiple instance learning: A survey of problem characteristics and applications. Pattern Recognition, 77:329–353, 2018.

Chattopadhay A, Sarkar A, Howlader P, and Balasubramanian VN. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE winter conference on applications of computer vision (WACV), pages 839–847. IEEE, 2018.

Chen M, Zhang B, Topatana W, Cao J, Zhu H, Juengpanich S, Mao Q, Yu H, and Cai X. Classification and mutation prediction based on histopathology h&e images in liver cancer using deep learning. NPJ precision oncology, 4(1):1–7, 2020a. [PubMed: 31934644]

Chen RJ, Lu MY, Wang J, Williamson DF, Rodig SJ, Lindeman NI, and Mahmood F. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE Transactions on Medical Imaging, 2020b.

Chen RJ, Chen TY, Lipkova J, Wang JJ, Williamson DF, Lu MY, Sahai S, and Mahmood F. Algorithm fairness in ai for medicine and healthcare. arXiv preprint arXiv:2110.00603, 2021a.

Chen RJ, Lu MY, Weng W-H, Chen TY, Williamson DF, Manz T, Shady M, and Mahmood F. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4015–4025, 2021b.

Chen RJ, Lu MY, Williamson DF, Chen TY, Lipkova J, Shaban M, Shady M, Williams M, Joo B, Noor Z, et al. Pan-cancer integrative histology-genomic analysis via interpretable multimodal deep learning. arXiv preprint arXiv:2108.02278, 2021c.

Cheplygina V, de Bruijne M, and Pluim JP. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical image analysis, 54:280–296, 2019. [PubMed: 30959445]

Choi J-H and Lee J-S. Embracenet: A robust deep learning architecture for multimodal classification. Information Fusion, 51:259–270, 2019.

Chu J, Dong W, Wang J, He K, and Huang Z. Treatment effect prediction with adversarial deep learning using electronic health records. BMC Medical Informatics and Decision Making, 20(4):1–14, 2020. [PubMed: 31906929]

Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, Gigante A, Valencia A, Rementeria MJ, Chadha AS, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ digital medicine, 3(1):1–11, 2020. [PubMed: 31934645]

Cohen JP, Luck M, and Honari S. Distribution matching losses can hallucinate features in medical image translation. In International conference on medical image computing and computer-assisted intervention, pages 529–536. Springer, 2018.

Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, Moreira AL, Razavian N, and Tsirigos A. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature medicine, 24(10):1559–1567, 2018.

Crosetto N, Bienko M, and Van Oudenaarden A. Spatially resolved transcriptomics and beyond. Nature Reviews Genetics, 16(1):57–66, 2015.

Cruz Rivera S, Liu X, Chan A-W, Denniston AK, and Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the spirit-ai extension. Nature medicine, 26(9):1351–1363, 2020.

Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 486(7403):346–352, 2012. [PubMed: 22522925]

Dai Yang K, Belyaeva A, Venkatachalapathy S, Damodaran K, Katcoff A, Radhakrishnan A, Shivashankar G, and Uhler C. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. Nature Communications, 12(1):1–10, 2021.

de Haan K, Zhang Y, Zuckerman JE, Liu T, Sisk AE, Diaz MF, Jen K-Y, Nobori A, Liou S, Zhang S, et al. Deep learning-based transformation of h&e stained tissues into special stains. Nature communications, 12(1):1–13, 2021.

Dijkstra KK, Voabil P, Schumacher TN, and Voest EE. Genomics-and transcriptomics-based patient selection for cancer treatment with immune checkpoint inhibitors: a review. JAMA oncology, 2(11):1490–1495, 2016. [PubMed: 27491050]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

Echle A, Grabsch HI, Quirke P, van den Brandt PA, West NP, Hutchins GG, Heij LR, Tan X, Richman SD, Krause J, et al. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. Gastroenterology, 159(4):1406–1416, 2020. [PubMed: 32562722]

Feng L, Liu Z, Li C, Li Z, Lou X, Shao L, Wang Y, Huang Y, Chen H, Pang X, et al. Development and validation of a radiopathomics model to predict pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: a multicentre observational study. The Lancet Digital Health, 4 (1):e8–e17, 2022. [PubMed: 34952679]

Fereidouni F, Harmany ZT, Tian M, Todd A, Kintner JA, McPherson JD, Borowsky AD, Bishop J, Lechpammer M, Demos SG, et al. Microscopy with ultraviolet surface excitation for rapid slide-free histology. Nature biomedical engineering, 1(12):957–966, 2017.

Ferreira-Junior JR, Koenigkam-Santos M, Magalhaes Tenorio AP, Faleiros MC, Garcia Cipriano FE, Fabro AT, Näppi J, Yoshida H, and de Azevedo-Marques PM. Ct-based radiomics for prediction of histologic subtype and metastatic disease in primary malignant lung neoplasms. International journal of computer assisted radiology and surgery, 15(1):163–172, 2020. [PubMed: 31722085]

Fu Y, Jung AW, Torne RV, Gonzalez S, Vöhringer H, Shmatko A, Yates LR, Jimenez-Linan M, Moore L, and Gerstung M. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nature Cancer, 1(8):800–810, 2020. [PubMed: 35122049]

Geessink OG, Baidoshvili A, Klaase JM, Bejnordi BE, Litjens GJ, van Pelt GW, Mesker WE, Nagtegaal ID, Ciompi F, and van der Laak JA. Computer aided quantification of intratumoral stroma yields an independent prognosticator in rectal cancer. Cellular Oncology, 42(3):331–341, 2019.

Ha SM, Chae EY, Cha JH, Kim HH, Shin HJ, and Choi WJ. Association of brca mutation types, imaging features, and pathologic findings in patients with breast cancer with brca1 and brca2 mutations. American Journal of Roentgenology, 209(4):920–928, 2017. [PubMed: 28796549]

Haskins G, Kruger U, and Yan P. Deep learning in medical image registration: a survey. Machine Vision and Applications, 31(1):1–18, 2020.

Havaei M, Guizard N, Chapados N, and Bengio Y. Hemis: Hetero-modal image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 469–477. Springer, 2016.

He K, Liu X, Li M, Li X, Yang H, and Zhang H. Noninvasive kras mutation estimation in colorectal cancer using a deep learning method based on ct imaging. BMC medical imaging, 20:1–9, 2020.

Hinton GE and Salakhutdinov RR. Reducing the dimensionality of data with neural networks. science, 313 (5786):504–507, 2006. [PubMed: 16873662]

Hinton JP, Dvorak K, Roberts E, French WJ, Grubbs JC, Cress AE, Tiwari HA, and Nagle RB. A method to reuse archived h&e stained histology slides for a multiplex protein biomarker analysis. Methods and protocols, 2(4):86, 2019. [PubMed: 31731599]

Hollon TC, Pandian B, Adapa AR, Urias E, Save AV, Khalsa SSS, Eichberg DG, D'Amico RS, Farooq ZU, Lewis S, et al. Near real-time intraoperative brain tumor diagnosis using stimulated raman histology and deep neural networks. Nature medicine, 26(1):52–58, 2020.

Horgan RP and Kenny LC. 'omic'technologies: genomics, transcriptomics, proteomics and metabolomics. The Obstetrician & Gynaecologist, 13(3):189–195, 2011.

Howard FM, Dolezal J, Kochanny S, Schulte J, Chen H, Heij L, Huo D, Nanda R, Olopade OI, Kather JN, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nature communications, 12(1):1–13, 2021.

Huang S-C, Pareek A, Seyyedi S, Banerjee I, and Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ digital medicine, 3(1):1–9, 2020a. [PubMed: 31934645]

Huang S-C, Pareek A, Zamanian R, Banerjee I, and Lungren MP. Multimodal fusion with deep neural networks for leveraging ct imaging and electronic health record: a case-study in pulmonary embolism detection. Scientific reports, 10(1):1–9, 2020b. [PubMed: 31913322]

Hyun SH, Ahn MS, Koh YW, and Lee SJ. A machine-learning approach using pet-based radiomics to predict the histological subtypes of lung cancer. Clinical nuclear medicine, 44(12):956–960, 2019. [PubMed: 31689276]

Ilse M, Tomczak J, and Welling M. Attention-based deep multiple instance learning. In International conference on machine learning, pages 2127–2136. PMLR, 2018.

Iv WCS, Kapoor R, and Ghosh P. Multimodal classification: Current landscape, taxonomy and future directions. ACM Computing Surveys (CSUR), 2021.

Jain MS and Massoud TF. Predicting tumour mutational burden from histopathological images using multiscale deep learning. Nature Machine Intelligence, 2(6):356–362, 2020.

Jang H-J, Lee A, Kang J, Song IH, and Lee SH. Prediction of clinically actionable genetic alterations from colorectal cancer histopathology images using deep learning. World Journal of Gastroenterology, 26 (40):6207, 2020. [PubMed: 33177794]

Jing L and Tian Y. Self-supervised visual feature learning with deep neural networks: A survey. arXiv preprint arXiv:1902.06162, 2019.

Joo S, Ko ES, Kwon S, Jeon E, Jung H, Kim J-Y, Chung MJ, and Im Y-H. Multimodal deep learning models for the prediction of pathologic response to neoadjuvant chemotherapy in breast cancer. Scientific reports, 11(1):1–8, 2021. [PubMed: 33414495]

Joze HRV, Shaban A, Iuzzolino ML, and Koishida K. Mmtm: Multimodal transfer module for cnn fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13289–13299, 2020.

Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, and Fei-Fei L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 1725–1732, 2014.

Kather JN, Heij LR, Grabsch HI, Loeffler C, Echle A, Muti HS, Krause J, Niehues JM, Sommer KA, Bankhead P, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. Nature Cancer, 1(8):789–799, 2020. [PubMed: 33763651]

Kennedy LB and Salama AK. A review of cancer immunotherapy toxicity. CA: a cancer journal for clinicians, 70(2):86–104, 2020. [PubMed: 31944278]

Khodadadian A, Darzi S, Haghi-Daredeh S, Eshaghi FS, Babakhanzadeh E, Mirabutalebi SH, and Nazari M. Genomics and transcriptomics: the powerful technologies in precision medicine. International Journal of General Medicine, 13:627, 2020. [PubMed: 32982380]

Khosravi P, Lysandrou M, Eljalby M, Li Q, Kazemi E, Zisimopoulos P, Sigaras A, Brendel M, Barnes J, Ricketts C, et al. A deep learning approach to diagnostic classification of prostate cancer using pathology–radiology fusion. Journal of Magnetic Resonance Imaging, 2021.

Kirkpatrick P. New clues in the acetaminophen mystery. Nature Reviews Drug Discovery, 4(11):883–883, 2005.

Kumar A, Fulham M, Feng D, and Kim J. Co-learning feature fusion maps from pet-ct images of lung cancer. IEEE Transactions on Medical Imaging, 39(1):204–217, 2019.

Lai Y-H, Chen W-N, Hsu T-C, Lin C, Tsao Y, and Wu S. Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. Scientific reports, 10(1):1–11, 2020. [PubMed: 31913322]

Lasocki A, Tsui A, Tacey M, Drummond KJ, Field K, and Gaillard F. Mri grading versus histology: predicting survival of world health organization grade ii–iv astrocytomas. American journal of neuroradiology, 36(1):77–83, 2015. [PubMed: 25104288]

Le MH, Chen J, Wang L, Wang Z, Liu W, Cheng K-TT, and Yang X. Automated diagnosis of prostate cancer in multi-parametric mri based on multimodal convolutional neural networks. Physics in Medicine & Biology, 62(16):6497, 2017. [PubMed: 28582269]

Lei B, Huang S, Li H, Li R, Bian C, Chou Y-H, Qin J, Zhou P, Gong X, and Cheng J-Z. Self-co-attention neural network for anatomy segmentation in whole breast ultrasound. Medical image analysis, 64: 101753, 2020. [PubMed: 32574986]

Li H, Bera K, Toro P, Fu P, Zhang Z, Lu C, Feldman M, Ganesan S, Goldstein LJ, Davidson NE, et al. Collagen fiber orientation disorder from h&e images is prognostic for early stage breast cancer: clinical trial validation. NPJ Breast Cancer, 7(1):1–10, 2021. [PubMed: 33397968]

Li J, Chen J, Tang Y, Landman BA, and Zhou SK. Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. arXiv preprint arXiv:2206.01136, 2022.

Liang M, Li Z, Chen T, and Zeng J. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. IEEE/ACM transactions on computational biology and bioinformatics, 12(4):928–937, 2014.

Linardatos P, Papastefanopoulos V, and Kotsiantis S. Explainable ai: A review of machine learning interpretability methods. Entropy, 23(1):18, 2020. [PubMed: 33375658]

Lipková J, Angelikopoulos P, Wu S, Alberts E, Wiestler B, Diehl C, Preibisch C, Pyka T, Combs SE, Hadjidoukas P, et al. Personalized radiotherapy design for glioblastoma: Integrating mathematical tumor models, multimodal scans, and bayesian inference. IEEE transactions on medical imaging, 38(8):1875–1884, 2019. [PubMed: 30835219]

Loeffler CML, Bruechle NO, Jung M, Seillier L, Rose M, Laleh NG, Knuechel R, Brinker TJ,Trautwein C, Gaisa NT, et al. Artificial intelligence–based detection of fgfr3 mutational status directly from routine histology in bladder cancer: A possible preselection for molecular testing? European Urology Focus, 2021.

Louis DN, Perry A, Reifenberger G, Von Deimling A, Figarella-Branger D, Cavenee WK, Ohgaki H,Wiestler OD, Kleihues P, and Ellison DW. The 2016 world health organization classification of tumors of the central nervous system: a summary. Acta neuropathologica, 131(6):803–820, 2016. [PubMed: 27157931]

Low CA. Harnessing consumer smartphone and wearable sensors for clinical cancer research. npj Digital Medicine, 3(1):1–7, 2020. [PubMed: 31934645]

Lu MY, Chen TY, Williamson DF, Zhao M, Shady M, Lipkova J, and Mahmood F. Ai-based pathology predicts origins for cancers of unknown primary. Nature, 594(7861):106–110, 2021. [PubMed: 33953404]

Marcus L, Lemery SJ, Keegan P, and Pazdur R. Fda approval summary: pembrolizumab for the treatment of microsatellite instability-high solid tumors. Clinical Cancer Research, 25(13):3753–3758, 2019. [PubMed: 30787022]

Mehrabi N, Morstatter F, Saxena N, Lerman K, and Galstyan A. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6):1–35, 2021.

Men K, Geng H, Zhong H, Fan Y, Lin A, and Xiao Y. A deep learning model for predicting xerostomia due to radiation therapy for head and neck squamous cell carcinoma in the rtog 0522 clinical trial. International Journal of Radiation Oncology* Biology* Physics, 105(2):440–447, 2019. [PubMed: 31201897]

Miller G. Breaking down barriers, 2002.

Mo S, Cai M, Lin L, Tong R, Chen Q, Wang F, Hu H, Iwamoto Y, Han X-H, and Chen Y-W. Multimodal priors guided segmentation of liver lesions in mri using mutual information based graph co-attention networks. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 429–438. Springer, 2020.

Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Vega JEV, Brat DJ, and Cooper LA. Predicting cancer outcomes from histology and genomics using convolutional networks. Proceedings of the National Academy of Sciences, 115(13):E2970–E2979, 2018.

Murchan P, Ó'Brien C, O'Connell S, McNevin CS, Baird A-M, Sheils O, Ó Broin P, and Finn SP. Deep learning of histopathological features for the prediction of tumour molecular genetics. Diagnostics, 11(8):1406, 2021. [PubMed: 34441338]

Naik N, Madani A, Esteva A, Keskar NS, Press MF, Ruderman D, Agus DB, and Socher R. Deep learning-enabled breast cancer hormonal receptor status determination from base-level h&e stains. Nature communications, 11(1):1–8, 2020.

Nie D, Zhang H, Adeli E, Liu L, and Shen D. 3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. In International conference on medical image computing and computer-assisted intervention, pages 212–220. Springer, 2016.

Nie D, Lu J, Zhang H, Adeli E, Wang J, Yu Z, Liu L, Wang Q, Wu J, and Shen D. Multi-channel 3d deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages. Scientific reports, 9(1):1–14, 2019. [PubMed: 30626917]

Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. New England Journal of Medicine, 351(27):2817–2826, 2004. [PubMed: 15591335]

Placido D, Yuan B, Hjaltelin JX, Haue AD, Yuan C, Kim J, Umeton R, Antell G, Chowdhury A, Franz A, et al. Pancreatic cancer risk predicted from disease trajectories using deep learning. BioRxiv, 2021.

Qi L, Ke J, Yu Z, Cao Y, Lai Y, Chen Y, Gao F, and Wang X. Identification of prognostic spatial organization features in colorectal cancer microenvironment using deep learning on histopathology images. Medicine in Omics, page 100008, 2021.

Qian X, Pei J, Zheng H, Xie X, Yan L, Zhang H, Han C, Gao X, Zhang H, Zheng W, et al. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. Nature Biomedical Engineering, 5(6):522–532, 2021.

Ramachandram D and Taylor GW. Deep multimodal learning: A survey on recent advances and trends. IEEE signal processing magazine, 34(6):96–108, 2017.

Ramanathan TT, Hossen M, Sayeed M, et al. Naïve bayes based multiple parallel fuzzy reasoning method for medical diagnosis. Journal of Engineering Science and Technology, 17(1):0472–0490, 2022.

Reda I, Khalil A, Elmogy M, Abou El-Fetouh A, Shalaby A, Abou El-Ghar M, Elmaghraby A, Ghazal M, and El-Baz A. Deep learning role in early diagnosis of prostate cancer. Technology in cancer research & treatment, 17:1533034618775530, 2018. [PubMed: 29804518]

Reyes M, Meier R, Pereira S, Silva CA, Dahlweid F-M, von Tengg-Kobligk H, Summers RM, and Wiest R. On the interpretability of artificial intelligence in radiology: challenges and opportunities. Radiology: artificial intelligence, 2(3), 2020.

Rokach L and Maimon O. Clustering methods. In Data mining and knowledge discovery handbook, pages 321–352. Springer, 2005.

Roy S, Lahiri D, Maji T, and Biswas J. Recurrent glioblastoma: where we stand. South Asian journal of cancer, 4(4):163, 2015. [PubMed: 26981507]

Sammut S-J, Crispin-Ortuzar M, Chin S-F, Provenzano E, Bardwell HA, Ma W, Cope W, Dariush A, Dawson S-J, Abraham JE, et al. Multi-omic machine learning predictor of breast cancer therapy response. Nature, 601(7894):623–629, 2022. [PubMed: 34875674]

Schmauch B, Romagnoni A, Pronier E, Saillard C, Maillé P, Calderaro J, Kamoun A, Sefta M, Toldo S, Zaslavskiy M, et al. A deep learning model to predict rna-seq expression of tumours from whole slide images. Nature communications, 11(1):1–15, 2020.

Sedghi A, Mehrtash A, Jamzad A, Amalou A, Wells WM, Kapur T, Kwak JT, Turkbey B, Choyke P, Pinto P, et al. Improving detection of pro state cancer foci via information fusion of mri and temporal enhanced ultrasound. International journal of computer assisted radiology and surgery, 15(7):1215–1223, 2020. [PubMed: 32372384]

Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, and Batra D. Grad-cam: Why did you say that? arXiv preprint arXiv:1611.07450, 2016.

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, and Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, pages 618–626, 2017.

Sha X, Gong G, Qiu Q, Duan J, Li D, and Yin Y. Identifying pathological subtypes of non-small-cell lung cancer by using the radiomic features of 18f-fluorodeoxyglucose positron emission computed tomography. Translational Cancer Research, 8(5):1741, 2019. [PubMed: 35116924]

Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, and Fu H. Transformers in medical imaging: A survey. arXiv preprint arXiv:2201.09873, 2022.

Shao W, Han Z, Cheng J, Cheng L, Wang T, Sun L, Lu Z, Zhang J, Zhang D, and Huang K. Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis. IEEE transactions on medical imaging, 39(1):99–110, 2019. [PubMed: 31170067]

Sharmay Y, Ehsany L, Syed S, and Brown DE. Histotransfer: Understanding transfer learning for histopathology. In 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), pages 1–4. IEEE, 2021.

Shergalis A, Bankhead A, Luesakul U, Muangsin N, and Neamati N. Current challenges and opportunities in treating glioblastoma. Pharmacological reviews, 70(3):412–445, 2018. [PubMed: 29669750]

Sundararajan M, Taly A, and Yan Q. Axiomatic attribution for deep networks. In International conference on machine learning, pages 3319–3328. PMLR, 2017.

Taqi SA, Sami SA, Sami LB, and Zaki SA. A review of artifacts in histopathology. Journal of oral and maxillofacial pathology: JOMFP, 22(2):279, 2018.

Topol EJ. Welcoming new guidelines for ai clinical research. Nature medicine, 26(9):1318–1320, 2020.

Tsou P and Wu C-J. Mapping driver mutations to histopathological subtypes in papillary thyroid carcinoma: applying a deep convolutional neural network. Journal of clinical medicine, 8(10):1675, 2019. [PubMed: 31614962]

Vale-Silva LA and Rohr K. Long-term cancer survival prediction using multimodal deep learning. Scientific Reports, 11(1):1–12, 2021. [PubMed: 33414495]

Van Cutsem E, Köhne C-H, Hitre E, Zaluski J, Chang Chien C-R, Makhson A, D'Haens G, Pintér T, Lim R, Bodoky G, et al. Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. New England Journal of Medicine, 360(14):1408–1417, 2009. [PubMed: 19339720]

Vasileiou G, Costa MJ, Long C, Wetzler IR, Hoyer J, Kraus C, Popp B, Emons J, Wunderle M, Wenkel E, et al. Breast mri texture analysis for prediction of brca-associated genetic risk. BMC medical imaging, 20(1):1–13, 2020.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, and Polosukhin I. Attention is all you need. Advances in neural information processing systems, 30, 2017.

Vo HQ, Yuan P, He T, Wong ST, and Nguyen HV. Multimodal breast lesion classification using cross-attention deep networks. In 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), pages 1–4. IEEE, 2021.

Wang S, Shi J, Ye Z, Dong D, Yu D, Zhou M, Liu Y, Gevaert O, Wang K, Zhu Y, et al. Predicting egfr mutation status in lung adenocarcinoma on computed tomography image using deep learning. European Respiratory Journal, 53(3), 2019.

Wang X, Chen Y, Gao Y, Zhang H, Guan Z, Dong Z, Zheng Y, Jiang J, Yang H, Wang L, et al. Predicting gastric cancer outcome from resected lymph node histopathology images using deep learning. Nature communications, 12(1):1–13, 2021.

Weeks WA, Dua A, Hutchison J, Joshi R, Li R, Szejer J, and Azevedo RG. A low-power, low-cost ingestible and wearable sensing platform to measure medication adherence and physiological signals. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 5549–5553. IEEE, 2018.

Weiss K, Khoshgoftaar TM, and Wang D. A survey of transfer learning. Journal of Big data, 3(1):1–40, 2016.

Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, and Zou J. How medical ai devices are evaluated: limitations and recommendations from an analysis of fda approvals. Nature Medicine, 27(4):582–584, 2021.

Xu S, Jayaraman A, and Rogers JA. Skin sensors are the future of health care, 2019.

Xu T, Zhang H, Huang X, Zhang S, and Metaxas DN. Multimodal deep learning for cervical dysplasia diagnosis. In International conference on medical image computing and computer-assisted intervention, pages 115–123. Springer, 2016.

Yala A, Lehman C, Schuster T, Portnoi T, and Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology, 292(1):60–66, 2019. [PubMed: 31063083]

Yamamoto Y, Tsuzuki T, Akatsuka J, Ueki M, Morikawa H, Numata Y, Takahara T, Tsuyuki T, Tsutsumi K, Nakazawa R, et al. Automated acquisition of explainable knowledge from unannotated histopathology images. Nature Communications, 10(1):1–9, 2019.

Yan J, Zhang B, Zhang S, Cheng J, Liu X, Wang W, Dong Y, Zhang L, Mo X, Chen Q, et al. Quantitative mri-based radiomics for noninvasively predicting molecular subtypes and survival in glioma patients. NPJ Precision Oncology, 5(1):1–9, 2021. [PubMed: 33479506]

Yap J, Yolland W, and Tschandl P. Multimodal skin lesion classification using deep learning. Experimental dermatology, 27(11):1261–1267, 2018. [PubMed: 30187575]

Yogananda CGB, Shah BR, Yu FF, Pinho MC, Nalawade SS, Murugesan GK, Wagner BC, Mickey B, Patel TR, Fei B, et al. A novel fully automated mri-based deep-learning method for classification of 1p/19q co-deletion status in brain gliomas. Neuro-oncology advances, 2(Supplement 4):iv42–iv48, 2020.

Zhang BH, Lemoine B, and Mitchell M. Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pages 335–340, 2018.

Zhang S, Tong H, Xu J, and Maciejewski R. Graph convolutional networks: a comprehensive review. Computational Social Networks, 6(1):11, 2019. [PubMed: 37915858]

Zhou L and Luo Y. Deep features fusion with mutual attention transformer for skin lesion diagnosis. In 2021 IEEE International Conference on Image Processing (ICIP), pages 3797–3801. IEEE, 2021.

Zhu J-Y, Park T, Isola P, and Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 2223–2232, 2017.

Zhuang L, Lipkova J, Chen R, and Mahmood F. Deep learning-based integration of histology, radiology, and genomics for improved survival prediction in glioma patients. In Medical Imaging 2022: Digital and Computational Pathology, volume 12039, page 120390Z. SPIE, 2022.

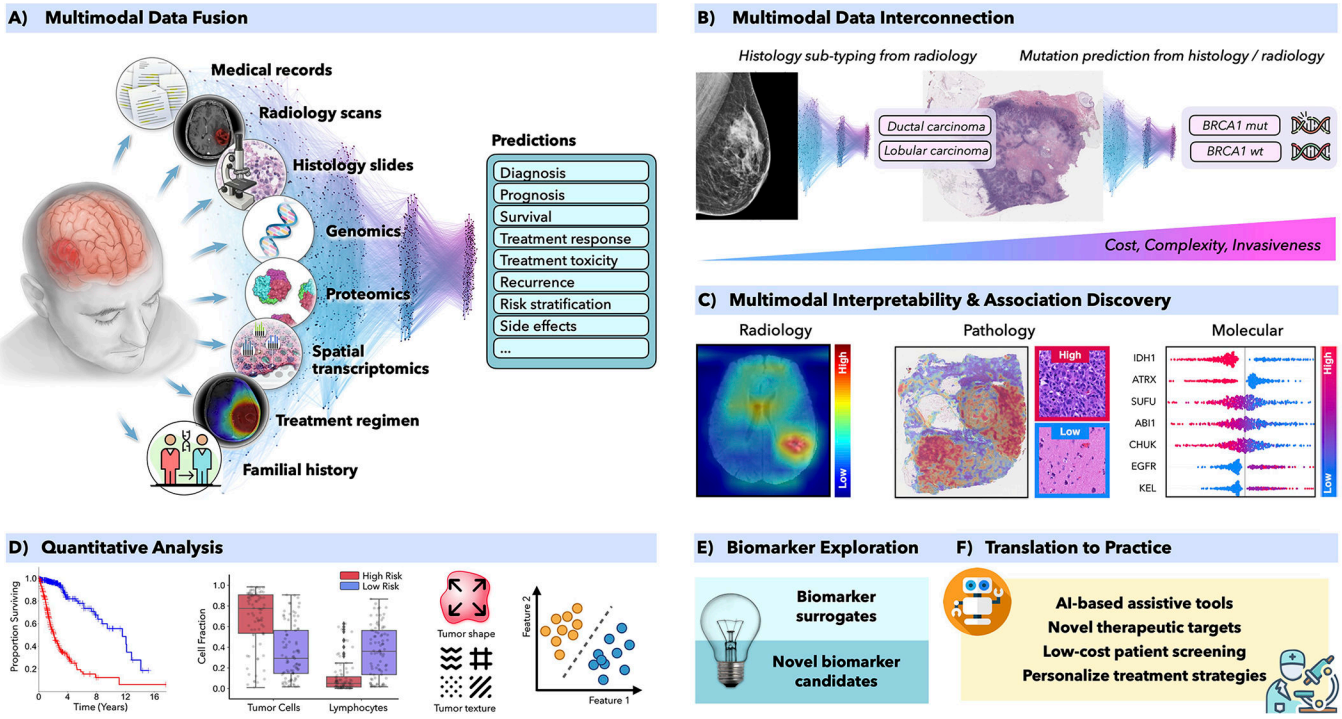

Figure 1. AI-driven multimodal data integration.

A) AI models can integrate complemenatary information and clinical context from diverse data sources to provide more accurate outcome predictions. The clinical insights identified by such models can be further elucidated through **C**) interpretability methods and **D**) quantitative analysis to guide and accelerate the discovery of new biomarkers or therapeutic targets (**E-F**). **B**) AI can reveal novel multimodal interconnestions, such as relation between certain mutations and changes in cellular morphology or associations between radiology findings and histology tumor subtypes or molecular features. Such associations can serve as non-invasive or cost-efficient alternatives to existing biomarkers to support large-scale patient screening (**E-F**)

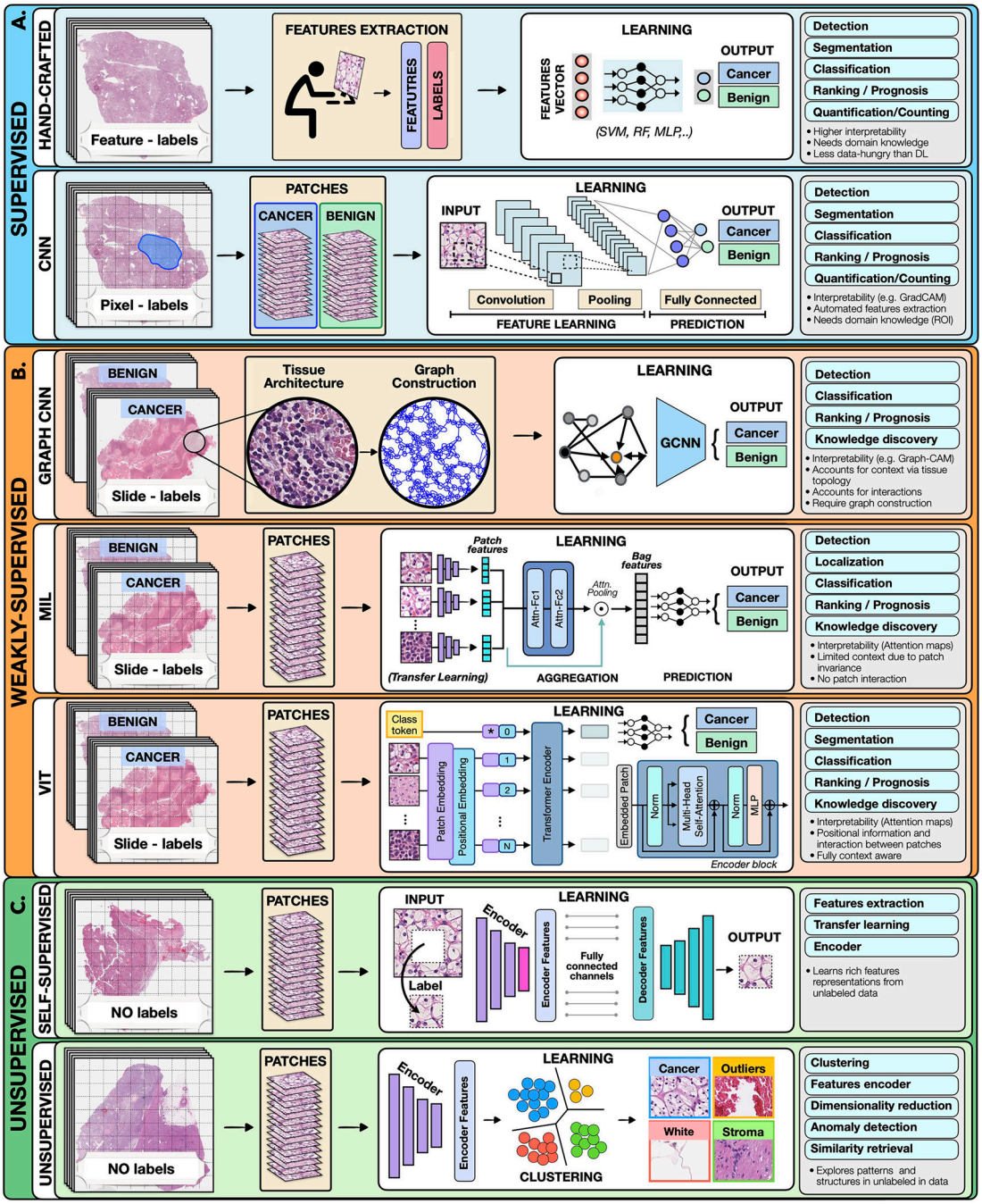

Figure 2. Overview of AI methods.

A) Supervised methods use strong supervision where each data point (*e.g.* feature or image patch) is assigned with a label. **B**) Weakly-supervised methods allow to train model with weak, patient-level labels, avoiding the need for manual annotations. **C**) Unsupervised methods explore patterns, subgroups and structures in unlabelled data. For comparison, all methods are illustrated on binary cancer detection task.

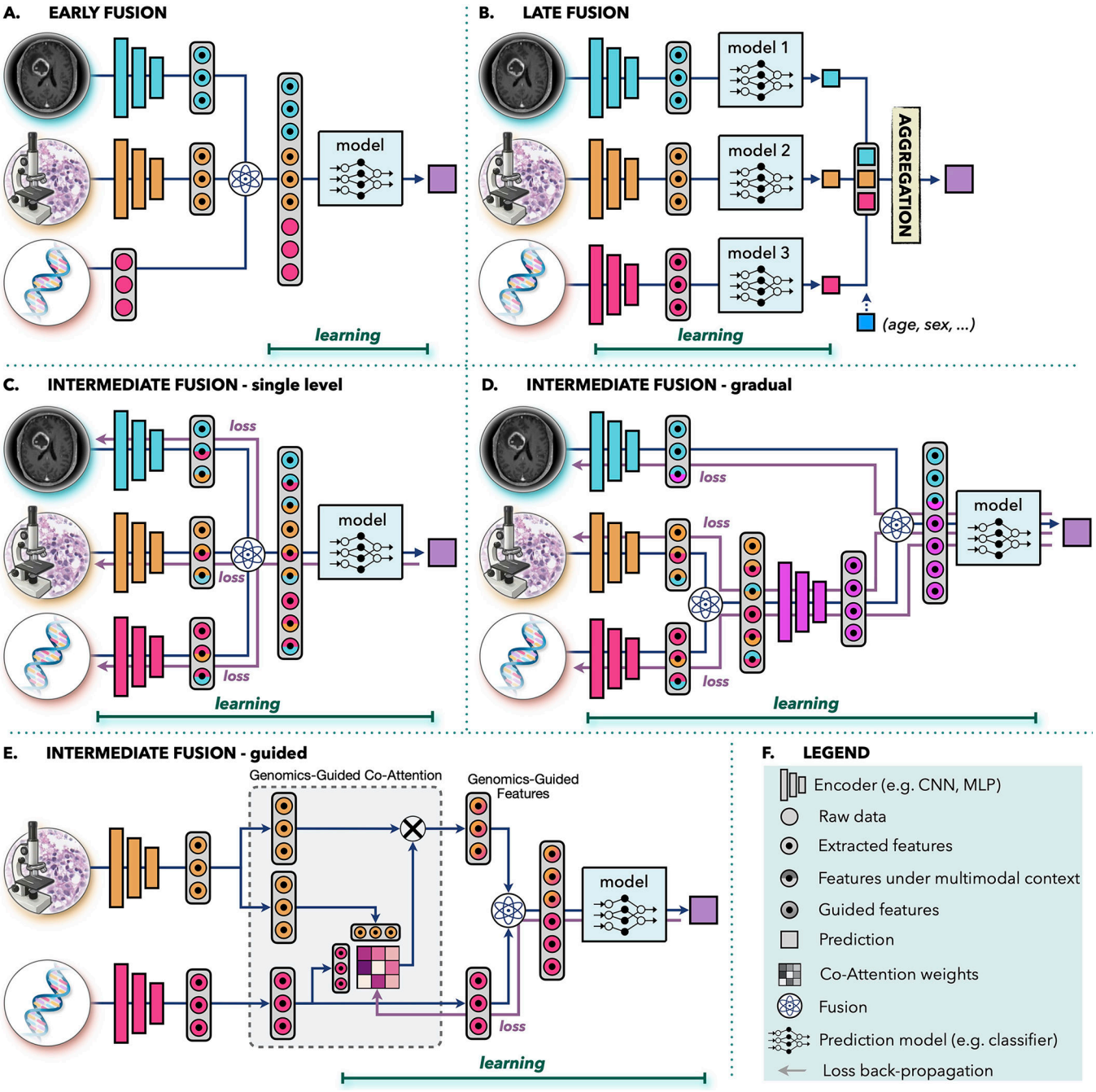

Figure 3. Multimodal data fusion.

A.) *Early fusion* builds a joint representation from raw data or features at the input level, before feeding it to the model. **B.**) *Late fusion* trains a separate model for each modality and aggregates the predictions from individual models at the decision level. (**C-E**) In *intermediate fusion*, the prediction loss is propagated back to the feature extraction layer of each modality to iteratively learn improved feature representations under the multimodal context. The unimodal data can be fused at **C.**) single-level or **D.**) gradually in different

layers. **E.**) Information from one modality can be also used to guide feature extraction from another modality. **F.**) Legend of the used symbols.

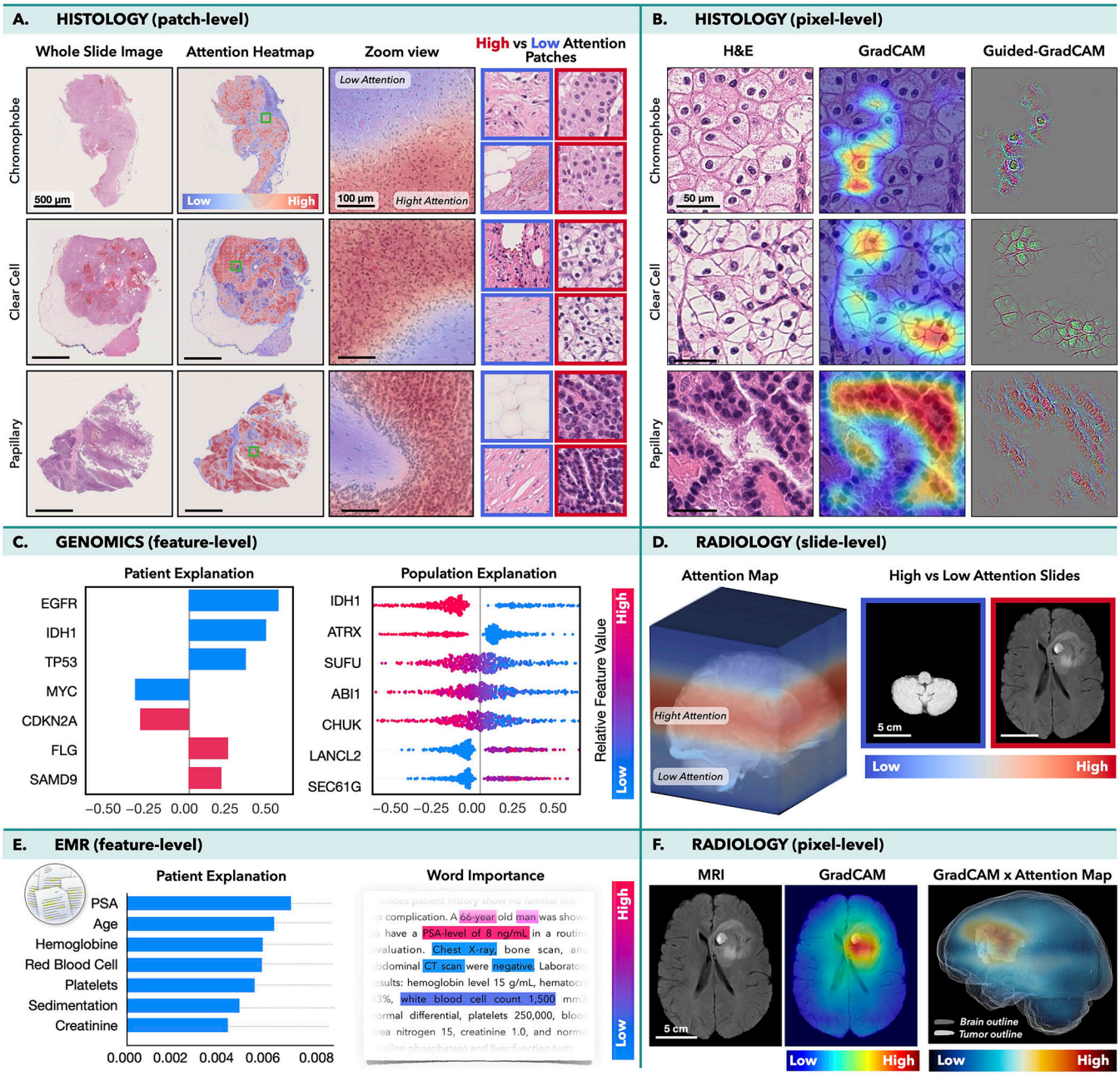

Figure 4. Interpretability methods.

Histology: MIL model was trained to classify subtypes of renal-cell carcinoma in WSIs, while CNN was trained to perform the same task in image patches. **A**) Attention-heatmaps and patches with the lowest/highest attention score. **B.**) GradCAM attributions for each class. ***Radiology:*** MIL model was trained to predict survival from MRI scans using axial slides as individual instances. **D.**) Attention-heatmaps mapped into the 3D MRI scan and slides with the highest/lowest attention. **F.**) GradCAM was used to obtain pixel-level interpretability in each MRI slide. A 3D pixel-level interpretability is computed by weighting the slide-level GradCAM maps by the attention score of the respective slide. Integrated gradient attributions can be used to analyze **C.**) ***Genomics*** or **E.**) ***EMRs***. The

attribution magnitude corresponds to the importance of each feature and direction indicates feature impact towards low (left) vs. high (right) risk. The color specifies the value of the input features: copy number gain/presence of mutation are shown in red, while blue is used for copy number loss/wild-type status. E.) Attention scores can be used to analyse the importance of words in the medical text.

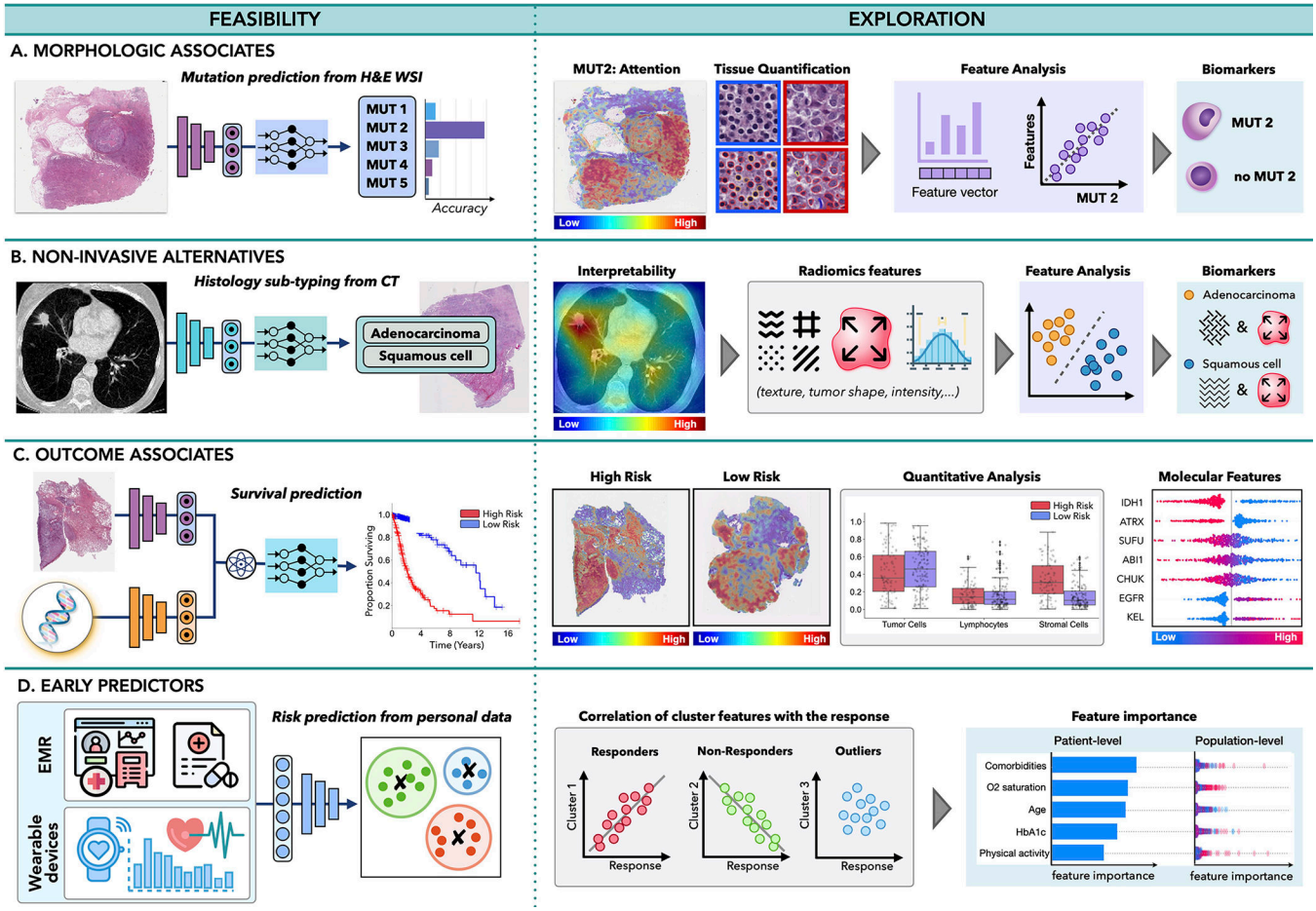

Figure 5. Multimodal data interconnection.

A.) AI can identify associations across modalities, such as feasibility of inferring certain mutations from histology or radiology images or **B.)** relation between non-invasive and invasive modalities, such as prediction of histology subtype from radiomics features. **C.)** The models can uncover associations between clinical data and patient outcome, contributing to discovery of predictive features within and across modalities. **D.)**Information acquired by EMRs or wearable devices can be analyzed to identify risk factors related with cancer onset or uncover patterns related with treatment response or resistance, to support early interventions.