Preprints.org

Review

Not peer-reviewed version

Integrating Omics Data and AI for Cancer Diagnosis and Prognosis: A Systematic Review

Yousaku Ozaki , Phil Broughton , Hamed Abdollahi , Homayoun Valafar * , Anna Blenda *

Posted Date: 11 June 2024

doi: 10.20944/preprints202406.0657.v1

Keywords: omics technologies; artificial intelligence; cancer

Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Review

Integrating Omics Data and AI for Cancer Diagnosis and Prognosis: A Systematic Review

Yousaku Ozaki [1], Philip Broughton [1], Hamed Abdollahi [2], Homayoun Valafar [2,*] and Anna Blenda [1,3,*]

[1] Department of Biomedical Sciences, University of South Carolina School of Medicine Greenville
[2] Department of Computer Science and Engineering, USC College of Engineering and Computing
[3] Prisma Health Cancer Institute
* Correspondence: homayoun@cse.sc.edu (H.V.); ablenda@greenvillemed.sc.edu (A.B.)

Abstract: Cancer is one of the leading causes of death, making timely diagnosis and prognosis very important. Utilization of AI (artificial intelligence) enables providers to organize and process patient data in a way that can lead to better overall outcomes. This review paper aims to look at the varying uses of AI for diagnosis and prognosis and clinical utility. PubMed and EBSCO databases were utilized for finding publications from January 1, 2013, to December 22, 2023. Articles were collected using key search terms such as "artificial intelligence" and "machine learning." Included in the collection were studies of the application of AI in determining cancer diagnosis and prognosis using multi-omics data, radiomics, pathomics, clinical and laboratory data. The resulting 89 studies were categorized into eight sections based on the type of data utilized and then further subdivided into two subsections focusing on cancer diagnosis and prognosis, respectively. 8 studies integrated more than one form of omics, namely genomics, transcriptomics, epigenomics, and proteomics. Incorporating AI into cancer diagnosis and prognosis alongside omics and clinical data represents a significant advancement. Given the considerable potential of AI in this domain, ongoing prospective studies are essential to enhance algorithm interpretability and to ensure safe clinical integration.

Keywords: omics technologies; artificial intelligence; cancer

1. Introduction

In 1950, Alan Turing introduced the concept of a thinking machine, marking the birth of artificial intelligence (AI) [1]. Today, AI has seamlessly integrated into our lives through familiar names like Siri, Alexa, and Google Assistant. The impact of AI is profoundly felt in the field of oncology, where it has revolutionized the approach to complex challenges posed by cancer. AI-driven techniques have notably elevated the precision and efficiency of oncologic research, opening doors to personalized cancer treatments. Its applications span across various areas, including cancer image analysis, genomic studies, data mining from medical records, and drug discovery [2–5].

There are two main subsets of AI: machine learning and deep learning [4]. Machine learning is a branch of AI that concentrates on creating computer software or algorithms capable of learning from data to make predictions autonomously, without the need for explicit programming. Three fundamental branches of machine learning are supervised, unsupervised, and reinforcement learnings [6]. Supervised learning trains models on labeled data, enabling the algorithm to learn patterns, like differentiating benign and malignant tumors in medical imaging for cancer detection. Unsupervised learning works on unlabeled data, identifying patterns within, like grouping patients based on genetic similarities for personalized treatment plans in cancer research. Reinforcement learning trains models to make sequential decisions, learning through trial and error, optimizing treatment plans in medicine. Meanwhile, deep learning uses neural networks with multiple layers to learn representations of data and excels in handling unstructured data like images and text [7].

© 2024 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

Convolutional neural networks (CNN) are phenomenal at image recognition tasks such as cancer image analysis. Recurrent neural networks and long short-term memory networks are frequently utilized in sequential data, aiding in genetic sequence analysis and mining medical records.

Considering the technological advances in the collection of multi-omics data over the past decades, their integration into cancer research is paramount to help us better understand this complex disease. Multi-omics includes several "-omics" methodologies, like genomics, transcriptomics, proteomics, epigenomics, and metabolomics, to comprehensively understand biological systems [8]. Each "-omics" field contributes to deeper insights into biological systems and diseases, unraveling various levels of anatomy and molecular and cellular interactions, laying the groundwork for precision medicine and personalized healthcare approaches. Genomics focuses on an organism's complete set of genes, gene sequences, interactions, and functions to understand how variations in genes contribute to traits or diseases. Epigenomics examines chemical modifications and alterations in DNA that regulate gene expression without changing the DNA sequence itself, to understand their impact on gene activity and cellular functions. Transcriptomics examines all RNA transcripts produced by cells or organisms at a given moment to discover levels and variations in gene expression. Proteomics studies all proteins within cells, tissues, or organisms to understand their biological processes. Metabolomics analyzes the roles of small molecules or metabolites within a biological system. Microbiomics analyzes the collective genetic material of microorganisms in specific environments. Radiomics involves extracting and analyzing quantitative data from medical imaging, like CT scans or MRIs, to identify patterns and correlations between imaging features and diseases such as textures [9]. Pathomics analyzes tissue samples at a microscopic level, integrating imaging, pathology, and molecular data to unravel disease mechanisms and assist in diagnosis and treatments [10]. In this review, we included radiomics and pathomics because they augment the comprehensive understanding provided by multi-omics approaches by supplying vital spatial and structural insights at both the tissue and imaging levels.

In the current era of personalized medicine and precision oncology, providers need to tailor treatment for each patient based on diagnosis and prognosis that are derived from enormous amounts of data. AI enables providers to organize and process the data to achieve goals that cannot be done with the human mind alone. Alongside multi-omics, radiomics, pathomics data, and clinical information — encompassing laboratory results and demographic information — plays a pivotal role in predictive modeling and personalized treatment. They offer insights into a patient's physiological status, potential risk factors, and responses to specific interventions, with the aims for tailored cancer management strategies. The field of multi-omics, radiomics, pathomics, and clinical data analysis with AI have exploded in the past decade but these advances have not been comprehensively reviewed. This review paper aims to close the gap by defining the novel scope in following way.

2. Materials and Methods

This systematic review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement [103].

We included randomized controlled trials and cohort studies of application of artificial intelligence in determining cancer diagnosis and prognosis using multi-omics data, radiomics, pathomics, clinical and laboratory data. In this study, the emphasis lies on exploring the potential of AI in cancer prediction and diagnosis, rather than on establishing a direct comparison group or intervention. However, in some studies, a comparison group might involve traditional methods of cancer prediction and diagnosis without AI. ChatGPT was utilized in this study to check spelling and grammar error.

Identification:

PubMed and EBSCO databases were used to search the eligible publications from January 1, 2013, to December 22, 2023. The query terms were "artificial intelligence", "machine learning", "deep learning", "cancer diagnosis", "cancer prognosis", "multi-omics", "genomics", "epigenomics", "transcriptomics", "proteomics", "metabolomics", "microbiomics", "radiomics", "pathomics", and "clinical data." Articles on relevant clinical studies in English were included. The search criteria on

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

PubMed were filtered to include only results with "full text available." On EBSCO, we utilized the "find all my search term" option and included the "also search within full text of the articles" expander. We set result limits to include peer-reviewed, full text, and references available.

Screening:

We used articles from publications with the DOAJ (Directory of Open Access Journals) seal as a measure to prevent the inclusion of articles from predatory journals. The DOAJ is a reputable database that indexes high-quality, open-access scholarly journals. We postulated that using articles from journals with the DOAJ seal would add a layer of quality assurance since the DOAJ employs a stringent review process for journal inclusion. We independently screened the database on December 23, 2023, and reached a consensus under the instructions of a project supervisor (Dr. Anna Blenda). A meta-analysis could not be conducted due to the heterogeneity in the design of these studies. We utilized Zotero, a reference management software, for the systematic screening of articles throughout the review process. In Zotero, articles were screened by title and abstract. Full texts were retrieved. We reviewed the full texts against the inclusion criteria, which is peer-reviewed, scholarly articles evaluating use of AI in cancer diagnosis and prognosis using multi-omics data, radiomics, pathomics, and/or clinical and laboratory data. Following were excluded from this study: 1. duplicates, 2. review articles, 3. systematic reviews, 4. absence of AI implementation, 5. study aims that were not associated with our theme, 6. inappropriate data type, 7. studies that mislabeled LASSO-Cox method as a machine learning, 8. studies with sample size of less than 100, 9. a study protocol, 10. A study that failed to specify the particular machine learning technique employed, and 11. animal studies. To focus on more recent developments and ensuring relevance to current trends of AI, we opted to narrow the time frame of the studies from 2013-2023 to 2020-2023.

3. Results

In the analysis of 89 studies, we found a broad spectrum of AI applications within cancer research (Figure 1). There were 2 articles focusing on genomics data, 21 articles on transcriptomics data, 3 articles on epigenomics data, 1 article each on proteomics and metabolomics data, 8 articles on multiomics data, 30 articles on radiomics data, 3 articles on pathomics data, and 20 articles on clinical data. No article on microbiomics data was found. Among these studies, 35 articles were pertinent to cancer diagnosis, while 54 articles were about cancer prognosis. Figure 2 shows a visual representation of the frequency of top five AI models employed.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

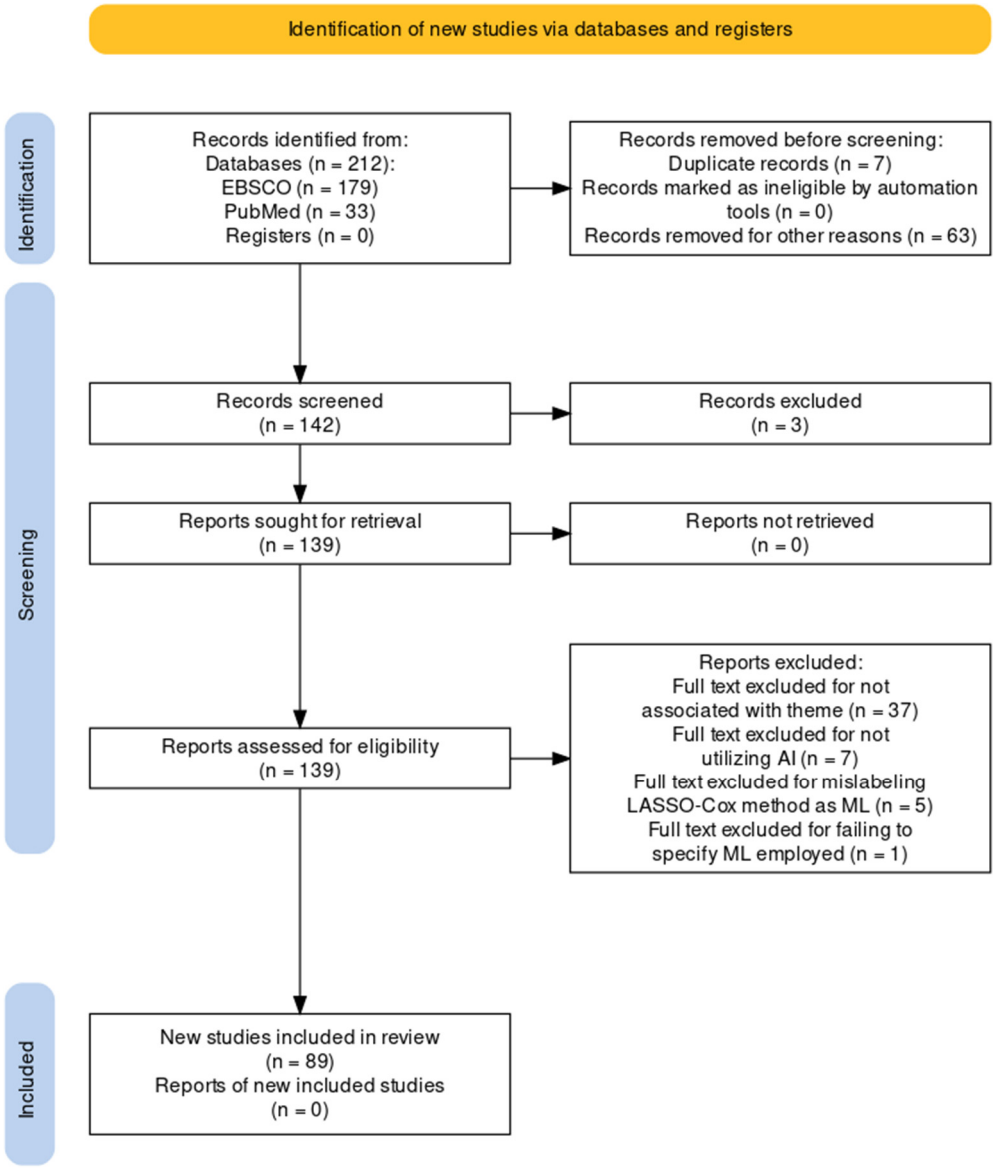

Figure 1. PRISMA flow diagram of the selection of studies to be included in the systematic review.

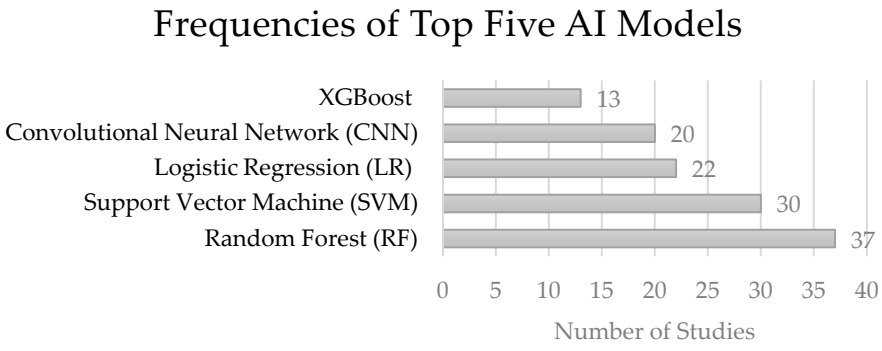

Figure 2. Frequencies of Top Five AI Models.

Random Forest (RF) method was the most prominently employed method. Studies that we reviewed with all data types except for pathomics did use RF. RF is a ML classifier composed of a collection of tree-structured classifiers $\{h(x, Q_k), k = 1,...\}$ where the $\{Q_k\}$ are identically distributed

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

random vectors and each tree casts a unit vote for the most popular class from a dataset [11]. Each decision tree within RF is trained on Q_k, which is a random subset of the training data and features, as illustrated in Figure 3. During prediction, the output of each tree is aggregated to produce the final prediction. This integration of multiple decision trees serves to improve accuracy and robustness of RF.

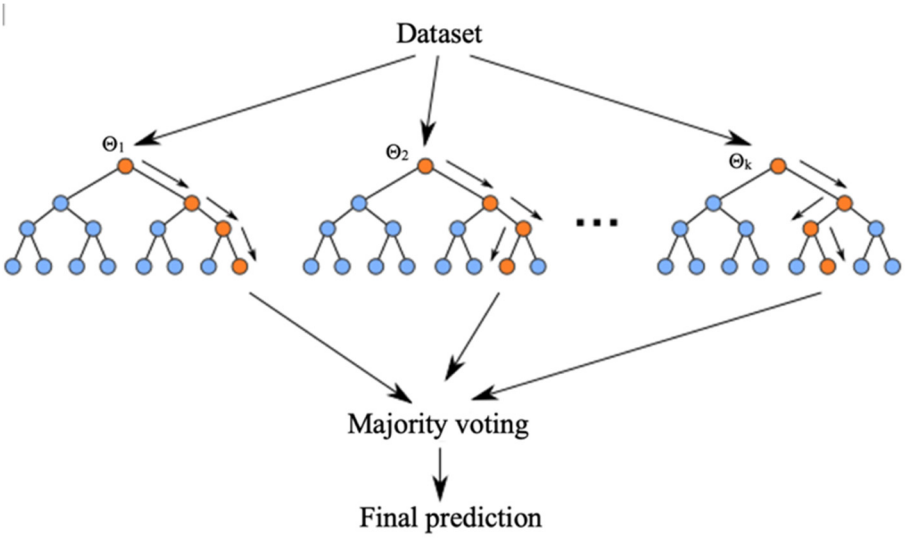

Figure 3. Schematic of Random Forest (RF); This image was adapted from Random Forest diagram from https://levelup.gitconnected.com/random-forest-regression-209c0f354c84.

Since Convolutional Neural Networks (CNNs) was the popular method in radiomics and pathomics data analysis, we will explain how CNNs work. Figure 4 shows a schematic depiction of CNN. CNN is a type of DL model that uses convolutional operations to find important features in input data by overlapping and combining local areas [12]. This helps the network to recognize patterns, even when they are not pre-labeled in the training data. The first step is to extract features from the input image. These features are then combined and reduced in size through pooling before being turned into the final network outputs. The last layers of the CNN connect all the neurons together and act as classifiers by sorting the input into different categories. Finally, the output layer gives the final classification or regression result, often using Softmax to calculate class probabilities.

For clarity and organization, the studies were categorized into eight sections based on the type of data utilized. Each section was further subdivided into two subsections focusing on cancer diagnosis and prognosis, respectively. Within these subsections, pertinent information from the articles was systematically collated into tables, including title, author and year, study aim, modality of AI employed, and outcome or performance. Each table serves as a discrete subset under either cancer diagnosis or prognosis to facilitate efficient referencing and comparison. Articles under each table were then organized based on their study aim.

In evaluating the performance of AI models, it is important to understand several parameters, including accuracy, sensitivity, specificity, area under the curve (AUC), and concordance index (C-index). Accuracy measures the proximity of measurements to their true values. Sensitivity evaluates a model's ability to predict true positives, while specificity assesses the model's capacity to predict true negatives. AUC gives a comprehensive measure of performance across various classification thresholds, calculated as the area under the ROC curve. Then C-index, like AUC, assesses the performance of prediction models, particularly in the context of survival analysis. A C-index closer to 1.0 indicates better predictive performance. In addition to the parameters, various AI algorithms or statistical methods were compared to evaluate the performance of AI.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

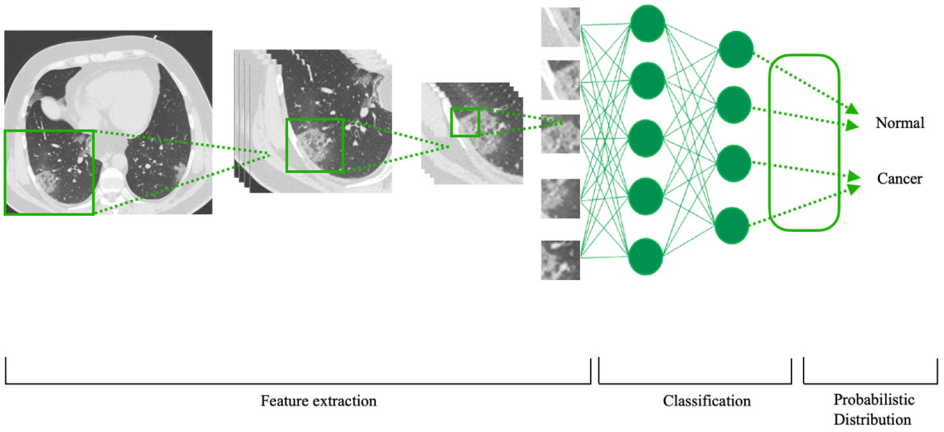

Figure 4. Schematic of Convolutional Neural Network (CNN); This image was adapted from CT image from http://pubs.rsna.org/doi/10.1148/rg.2020200159 and CNN diagram from https://www.geeksforgeeks.org/introduction-deep-learning/.

3.1. Clinical Applications Based on Genomics

3.1.1. Genomics-Based Prediction of Cancer Prognosis

Following studies made notable contributions to the field of genomics by leveraging computational algorithms to predict key genetic patterns and treatment responses in cancer patients.

Table 1. Genomics-based prediction of cancer prognosis.

Outcome/Performance	Modality of AI	Study aim	Author, year
Genomics-based prediction of prognostic biomarker			
8 (14.8%), 10 (16.9%), 17 (18.7%), and 43 (59.7%) cases were predicted to exhibit the *MYC*-trans, *BCL2*-trans, *BCL6*-trans, and MC signatures.	Random forest (RF)	To identify overlapping genetic patterns in DLBCL patients	Zhang et al., 2020 [13].
Genomics-based prediction of treatment responses			
Mean root square error (0.1587) of ANN-SCGP was lowest among other traditional MLs, including RF, SVM, and ANN. Mean root square error assesses the average difference between the predicted values generated by a model and actual values.	ANN with Selective Connection based on Gene Patterns (ANN-SCGP) RF, support vector machine (SVM), ANN, DeepSurv	To predict treatment response to radiotherapy based on gene patterns	Zeng et al., 2022 [14].

3.2. Clinical Applications Based on Transcriptomics

3.2.1. Transcriptomics-Based Prediction of Cancer Diagnosis and Prognosis

Following studies advanced the field of transcriptomics by employing machine learning (ML) and deep learning (DL) methods to analyze gene expression data and identify biomarkers associated with cancer. In terms of cancer prognosis, these studies employed ML methods to identify RNA signatures associated with various aspects of cancer prognosis and treatment response.

Table 2. Transcriptomics-based prediction of cancer diagnosis and prognosis.

Outcome/Performance	Modality of AI	Study aim	Author, year
Transcriptomics-based cancer detection			
11-gene panel was validated with GB for accuracy in distinguishing NSCLC cases from healthy controls. Among the three	Gradient Boosting Machines (GBM), RF, and Linear	To discriminate between non-metastatic NSCLC cases and healthy samples using 11 platelet-genes	Goswami et al., 2020 [15].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

classifiers, GBM offered the highest AUC = 0.97.			
The model trained with 19 primary cancer types achieved the highest performance with 89.67%, 87.32%, and 84.59% accuracy for 6, 8, and 10-way predictions in test samples.	Siamese convolutional neural network (SCNN)	To predict cancer types for primary and metastatic tumors from gene expression data	Mostavi et al., 2021 [16].
Transcriptomics-based classification of malignant vs benign tumors			
Mean root square error (0.1587) of ANN-SCGP was lowest among other traditional MLs, including RF, SVM, and ANN. Mean root square error assesses the average difference between the predicted values generated by a model and actual values.	ANN with Selective Connection based on Gene Patterns (ANN-SCGP) RF, support vector machine (SVM), ANN, DeepSurv	To predict treatment response to radiotherapy based on gene patterns	Carrillo-Perez et al., 2021 [17].
Transcriptomics-based survival prediction			
AUC values of 4 survival groups were all above 90%. The patient groups predicted by the SVM model demonstrated comparable survival outcomes to those clustered by the K-means algorithm.	Combination of K-means clustering and SVM	To evaluate a microRNA-based machine learning survival prediction model	Ding et al., 2021 [18].
The stemness subtype classifier by RF showed good performance in the classification with an AUC of 0.956, and the sensitivity, specificity, and accuracy were 86.15%, 91.03% and 88.9%.	RF	To predict transcriptional stemness indices of lung cancer from RNA expression data	Lai et al., 2023 [19].
RF-approach outperformed traditional prognostic variables like disease stage and cell of origin (COO) in predictive accuracy for DLBCL patients.	RF	To evaluate a new machine learning-based models of survival prediction using transcriptomic and clinical data	Mosquera Orgueira et al., 2020 [20].
Transcriptomics-based recurrence prediction			
SVM-REF and Random Forest analyses selected 66 and 30 lncRNA prognostic signatures, respectively.	RF and Support Vector Machine Recursive Feature Elimination (SVM-RFE)	To evaluate a lncRNA-based signature for predicting HCC early recurrence	Fu et al., 2023 [21].
Prediction of breast cancer recurrence with XGBoost performed better with mRNA data (AUC=0.74) alone compared to mutation alone (AUC=0.62).	XGBoost	To evaluate prognostic utility of genomic mutations to that of gene expression using breast cancer data	Ravkin et al., 2020 [22].
Transcriptomics-based prediction of risk stratification			
SVMR achieved the best classification performance (accuracy = 0.923, sensitivity = 0.927, specificity = 0.919) compared to other classifiers.	Support vector machine with radial kernel (SVMR)	To identify a novel miRNA signature related to tumor stage and prognosis of clear cell renal cell carcinoma patients	Dessie et al., 2021 [23].
Combining differentially spliced and expression levels of RNA yielded the most performant RF-classifier compared to splicing signature only or expression levels only.	RF	To subclassify highly aggressive breast cancers with transcriptomics analysis of alternative splicing events	Villemin et al., 2021 [24].
SWT-CNN outperformed other machine learning algorithms including support vector machine (SVM) and logistic regression (LR). SWT-CNN performed comparably with RF in predicting tumor stages.	Combination of a convolutional neural network with stationary wavelet transform (SWT-CNN)	To stratify the prognostic risk for cancer patients by using SWT-CNN	Zhao et al., 2020 [25].
Transcriptomics-based prediction of prognostic biomarker			
RF exhibited the highest Area Under the Curve (AUC) across all datasets, while SVM demonstrated the highest sensitivity and specificity.	RF, KNN, SVM, naïve bayes (NB), and neural networks (NNET) for feature extraction	To identify transcript biomarkers that could help in early prognosis for HCC	Gupta et al., 2021 [26].
The top 5 significant molecules pinpointed by each machine learning algorithm revealed a single intersecting molecule which is SFN.	Logical regression (LR), SVM, artificial neural network (ANN), RF, and XGBoost	To identify key prognostic molecule with multiple MLs	Li et al., 2021 [27].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

RF ranked top 10 important master genes for two prognostic groups including CCNA2, CBX7, TMEM48, SPC25, GAPDH, WDHD1, PSMD2, ERO1L, DDX52, ARNTL2.	RF	To identify the key prognosis impacting genes and relevant subtypes for lung adenocarcinoma	Lv and Lei., 2020 [28].
A machine learning-based approach identified C5AR1/SYT5 and MSR1/SLC32A1 signatures which were able to discriminate NL IDH-WT gliomas with high sensitivity and specificity in various glioma expression datasets.	K-nearest neighbor (KNN)	To characterize novel biomarker in glioma	Nguyen et al., 2020 [29].
SVM-RFE yielded 72 prognostic features with classification accuracy of 0.934.	SVM-RFE	To evaluate the association between immune infiltration and the prognosis in ovarian cancer	Yan et al., 2020 [30].
The intersection of the top 10 feature lncRNAs obtained from both the XGBoost and Boruta algorithms resulted in eight intersecting lncRNAs.	XGBoost and Boruta algorithm	To identify and explore prognostic biomarkers associated with clear cell renal cell carcinoma	Zhong et al., 2023 [31].
Transcriptomics-based prediction of laterality of cancer			
SVM-RBF classified the different locations by the highest accuracy of 99%. RF classified with high accuracy. NB was not satisfactory.	NB, SVM-RBF, and RF	To identify biomarkers which are associated with specific tumor locations	Hamzeh et al., 2020 [32].
Transcriptomics-based prediction of treatment responses			
RF yielded best results with mean accuracy of 84.1% for 5-FU and 82.3% for GCB.	RF, SVM, LR	To predict treatment response of multiple cancer types to 5-Fluorouracil and Gemcitabine	Clayton et al., 2020 [33].
Cluster 2 exhibited a notably poorer prognosis compared to Cluster 1.	K means clustering	To examine relationships between the effects of platinum-containing drugs with of metabolic genes and FAK activity in advanced ovarian high-grade serous carcinoma	Sato et al. 2022 [34].
KNN derived AUC of 0.72. This model performed better than previously published pan-cancer predictive models for immunotherapy efficacy.	KNN	To predict survival and immunotherapy response with transcriptomic marker from tumor endothelial cells	Wu et al., 2023 [35].

3.3. Clinical Applications Based on Epigenomics

3.3.1. Epigenomics-Based Prediction of Cancer Diagnosis and Prognosis

Following studies contributed to the field of epigenomics by employing various ML techniques to analyze epigenetic data and uncover important insights related to cancer prognosis and mutation detection.

Table 3. Epigenomics-based prediction of cancer diagnosis and prognosis.

Outcome/Performance	Modality of AI	Study aim	Author, year
Epigenomics-based classification of malignant vs benign tumors			
SETRED with SVM base learner performed the best with mean accuracy above 0.95 and AUC for methylation class and family prediction (AUC = 0.73 and 0.94, respectively). The NN model exhibited notably higher balanced accuracy (92.9% and 97.5%) compared to the RF classifier (70.9% and 72.3%).	11 semi-supervised learning models based on SVM, decision tree, and one nearest neighbor 2 supervised classification models: RF and NN	To explore utility of semi-supervised in methylation data	Tran et al., 2022 [36].
Epigenomics-based classification of tumor staging			
Precisions for groups A-D were 0.931, 0.833, 0.577, and 0.414.	RF	To classify neuroblastoma staging with epigenomic data	Sugino et al., 2022 [37].
Epigenomics-based prediction of biomarker for cancer prognosis			
The model yielded a sensitivity of 0.94, specificity of 0.82, and a false negative rate of 0.06.	Binomial logistic regression	To develop and validate a 3-CpG methylation signature to predict *SETD2* mutation status	Javaid et al., 2023 [38].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

3.4. Clinical Applications Based on Proteomics and Metabolomics

3.4.1. Proteomics and Metabolomics-Based Prediction of Cancer Diagnosis

Following studies employed various ML techniques to analyze proteomics and metabolomics data.

Table 4. Proteomics and metabolomics-based prediction of cancer diagnosis and prognosis.

Outcome/Performance	Modality of AI	Study aim	Author, year
Proteomics-based prediction of diagnostic biomarker			
A diagnostic model (RF) incorporating seven factors (CLU, CA19-9, IBIL, GGT, LDL-C, TG, and TBA), showed a high diagnostic utility with AUC: 0.947, sensitivity: 90.3%, specificity: 84.9%.	RF	To evaluate diagnostic performance of proteomic biomarker for cholangiocarcinoma	Gao et al., 2023 [39].
Metabolimics-based cancer prediction			
The XGBoost model showed best predictive power (AUC = 0.81, accuracy = 75.29%, sensitivity = 74%).	XGBoost, SVM, KNN, RF	To predict lung cancer with metabolic data	Guan et al., 2023 [40].

3.5. Clinical Applications Based on Multiomics Data

3.5.1. Cancer Diagnosis and Prognosis Based on Multiomics Data

Following studies significantly advanced the field of multiomics by introducing innovative approaches to integrate diverse data types for cancer research. Multiomics data included genomics, transcriptomics, epigenomics, and proteomics. In terms of cancer prognosis, these studies leveraged various omics data and integrated them with clinical features to predict important outcomes in cancer.

Table 5. Cancer diagnosis and prognosis based on multiomics data.

Outcome/Performance	Modality of AI	Study aim	Author, year
Cancer prediction based on multiomics data			
The AUC of LGDLDA was 0.880, which was 0.034, 0.088, 0.053 and 0.208 higher than that of IDHI-MIRW, NCPLDA, LncDisAP and NCPHLDA, respectively.	LncRNA-Gene-Disease association networks based LncRNA-Disease Association prediction (LGDLDA): base model is a neural network	To identify cancer-related lncRNAs	Yuan et al., 2021 [41].
Subclassification of malignant tumors based on multiomics data			
Mean root square error (0.1587) of ANN-SCGP was lowest among other traditional MLs, including RF, SVM, and ANN. Mean root square error assesses the average difference between the predicted values generated by a model and actual values.	moBRCA-net: base model is a neural network	To evaluate moBRCA-net	Choi and Chae, 2023 [42].
Survival prediction based on multiomics data			
Survival prediction: accuracy of 94% and AUC of 0.98 Drug response prediction: AUC of 0.83 and 0.78 for Docitaxel and Gemcitabine	Neural network-based classifier	To predict survival and drug response for breast cancer patients	Malik et al., 2021 [43].
Autoencoder outperformed 2 statistical methods with C-index 0.92 (PCA and iCluster).	Combination of Autoencoder and SVM	To identify survival subtype of glioma with RNA expression and DNA methylation data	Tian et al., 2022 [44].
DNA methylation and miRNA expression resulted in best performance with C-index of 0.641	Concatenation autoencoder (ConcatAE) and CrossAE	To predict breast cancer survival by integrating multi-omics data	Tong et al., 2020 [45].
Multiomics-based prediction of prognostic biomarker			

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

75 mRNAs identified as prognostic in TCGA cohort. 29 mRNAs identified as prognostic in Autoencoder LIRI-JP dataset.		To identify biomarkers that distinguish prognostic subgroups in liver cancer	Owens et al., 2021 [46].
Multiomics-based prediction of laterality of cancer			
The classification model derived from the 17 gene expressions resulted in an AUC of 0.96.	XGBoost	To identify gene mutation and expression patterns between left-sided and right-sided colon cancer	Jiang et al., 2020 [47].
The accuracies of the RF models were 90%, 70%, and 87% with corresponding area under the curve (AUC) values of 0.9, 0.76, and 0.89 for the human genomic, microbial, and combined feature sets, respectively.	RF	To predict sidedness of colon cancer	Kolisnik et al., 2023 [48].

3.6. Clinical Applications Based on Radiomics

3.6.1. Radiomics-Based Prediction of Cancer Diagnosis and Prognosis

In the field of radiomics, these studies employed ML and DL techniques for various tasks, including classification of malignant versus benign tumors, gene expression prediction, and cancer invasion prediction. In terms of cancer prognosis, these studies achieved several advancements in predictive modeling and prognosis assessment for survival, metastasis prediction, and treatment complications.

Table 6. Radiomics-based prediction of cancer diagnosis and prognosis.

Outcome/Performance	Modality of AI	Study aim	Author, year
Radiomics-based classification of malignant vs benign tumors			
Xception: AUC 0.970 DenseNet169: AUC 0.959 Both DLs outperformed radiologists ($P < 0.05$).	Xception DenseNet169 DenseNet121 NASNetLarge ResNet101v2	To evaluate diagnostic performance of DLs in distinguishing benign vs malignant thyroid calcified nodules	Chen et al., 2023 [49].
Fusion model achieved AUC 0.916 for SCA diagnosis and AUC 0.973 for MCA and IPMN diagnosis.	Fused model: Based on logistic regression (LR) and SVM	To evaluate diagnostic models based on radiomics and deep learning algorithms to differentiate three types of pancreatic cystic neoplasms	Liang et al., 2022 [50].
A total of 180 tumor texture features were extracted from enhanced CT and unenhanced CT.	AK software (Artificial Intelligence Kit V3.0.0.R) by GE Healthcare	To diagnose anterior mediastinal cysts vs thymomas with radiomic features	Liu et al., 2020 [51].
Mean accuracy of 93.25%, a sensitivity of 89.22%, a specificity of 95.82%, and AUC of 0.9629.	RGB: combination of 5 CNNs and 1 GCN	To evaluate RGB model classify benign vs malignant lung nodules	Ma et al., 2023 [52].
The DLR model achieved an AUC of 0.986, 0.978, 0.967, and 0.953 in the training, internal validation, and external validation.	DLR model: based on ResNet50	To evaluate role of deep learning radiomics on contrast-enhanced US in distinguishing pancreatic adenocarcinoma vs chronic pancreatitis	Tong et al., 2022 [53].
CNN model with clinical features achieved the highest AUC 0.819.	CNN and RF	To distinguish benign vs malignant lung nodules in chest CT	Zhang et al., 2022 [54].
The model established by the LR method had the best performance, and the AUC values in the training group and test group were 0.840 and 0.960. AUC of the combined model was 0.940, 0.990 and 0.960 in the training group, test group and external validation group.	Radiomics model: RF, SVM, and LR Combined model: LR	To differentiate pulmonary mucinous adenocarcinoma from tuberculoma based on features from CT images and clinical features	Zhang et al., 2023 [55].
The accuracy of the test set was 0.84.	PB-LNet: Based on ResNext50 and Bidirectional LSTM (BiLSTM)	To classify CT images of lung nodules into six categories based on pathological subtypes	Zhang et al., 2023 [56].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

96 images from the test set without data augmentation were analyzed and the accuracy was 0.89.			
AUCs for lymphoma ranged 0.670 to 0.936 in three testing sets. AUCs for metastatic carcinoma ranged 0.804 to 0.855 in three testing sets.	ResNet50	To accurately diagnose unexplained cervical lymphadenopathy with ultrasound	Zhu et al., 2022 [57].
Radiomics-based prediction of gene expression in malignant tumors			
AUCs of the clinical model (LR) in the testing, internal validation, and external validation sets were 0.794, 0.711, and 0.75. AUCs of the deep models and joint models ranged from 0.939 to 0.993. The predictive performance of the DLRS-Resnet model was inferior to that of the Nomogram-Resnet model ($p < 0.01$).	1 ML model: LR 3 DL models: DLRS-Resnet, DLRS-Inception, and DLRS-Densenet 3 joint models: Nomogram-Resnet, Nomogram-Inception, and Nomogram-Densenet	To predict Ki67 expression in prostate cancer with MRI radiomics	Deng et al., 2023 [58].
ResNet model in the axial direction achieved the higher AUC 0.90 in the testing cohort than coronal or sagittal directions. AUC of radiomics model (RF) in testing cohorts was 0.818.	ResNet and RF	To predict KRAS mutation in colorectal cancer with CT radiomics	He et al., 2020 [59].
Radiomics-based prediction of cancer invasion			
Radiomics model (LR) performed best in training and external data set. Combined model (LR) performed best in testing set.	LR	To predict lymphovascular invasion status in cervical cancer	Li et al., 2021 [60].
Three SVM-based prediction models demonstrated relatively high efficacy in identifying LVI of breast cancer, with AUCs of 79.00%, 80.00% and 79.40% and an accuracy of 71.00%, 80.00% and 75.00% in the validation cohort for AP, SP and CP plane image. Fusion model achieved the highest AUC of 87.90% and an accuracy of 85.00% in the validation cohort.	SVM	To predict the lymphovascular invasion status in breast cancer	Li et al., 2023 [61].
FNN performed the best with CMP demonstrating the highest AUC 0.81.	SVM-RBF, KNN, LR, Linear discriminant analysis (LDA), Forward neural network (FNN)	To predict capsule invasion in renal cell carcinoma	Yang et al., 2022 [62].
The six models showed the certain value of radiomics, with AUCs from 0.642 to 0.701. LR demonstrated best performance.	KNN, LR, Decision tree, Linear-SVM, Gaussian-SVM, Polynomial-SVM	To predict extrathyroidal extension (ETE) in papillary thyroid cancer (PTC) patients	Yu et al., 2022 [63].
XGBoost model demonstrated best performance in both training and testing set with AUC 0.917 and 0.874.	LR, SVM, XGBoost	To predict histological invasiveness of sub-centimeter subsolid pulmonary nodules	Zhang et al., 2023 [64].
Radiomics-based survival prediction			
EN and RF achieved top prognostication performances of AUC = 0.795 and AUC = 0.811. RF prognostication slightly outperformed the EN for the complete and radiochemotherapy cohort.	Elastic Net (EN) RF	To predict survival of squamous cell carcinoma of the head and neck with CT radiomics	Bernatz et al., 2023 [65].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

The overall prediction accuracy for 3-year survival status in training and validation cohort was 92.50% and 85.71%, and the AUC was 0.965 and 0.869.	SVM	To predict survival of unresectable lung cancer patients with CT radiomics	Chen et al., 2022 [66].
RF models built with clinical, CT and PET features outperformed other models with solely clinical, PET or CT features with C-index 0.780 and 0.820 in training and testing set.	RF	To predict survival of colorectal cancer patients with ^{8}F- FDG PET/CT radiomic features	Lv et al., 2022 [67].
For 2- and 5-year survival predictions, ResNet 50 achieved best performance for 2D PET images, while ResNet 34 achieved best performance for 3D PET images. ResNet 34 demonstrated best performance with C-index 0.749.	ResNet50 for 2D PET images ResNet3D34 for 3D PET images	To predict survival of non-small cell lung cancer patients with PET radiomics	Oh et al., 2023 [68].
Radiomics-based metastasis prediction			
The patient-demographic model resulted in an accuracy of 67.31% and 73.08% and AUC of 0.706 and 0.773 for training and testing cohorts. The radiomic-derived model resulted in an accuracy of 81.09% and 79.49% and AUC of 0.882 and 0.825 for training and testing cohorts.	SVM	To predict lymph node metastasis with pre-op CT	Eresen et al., 2020 [69].
MNB outperformed other MLs with AUC, specificity, and accuracy on the testing set of 0.745, 0.900, and 0.778.	XGBoost, LR, Multinomial Naive Bayes (MNB), SVM, Decision Tree, RF, Gradient Boosting Decision Tree (GBDT)	To predict lymph node metastasis in cervical cancer with MRI radiomics	Liu et al., 2023 [70].
XGBoost outperformed other MLs with AUC 0.98, sensitivity 0.75, and specificity 0.94.	Ada Boosting (ADA), Bagging Classifier (BAGC), Bernoulli Naïve Bayes (BNB), Decision Tree, Gaussian Naïve Bayes (GNB), KNN, RF, Stochastic Gradient Descent (SGD), SVM, and XGBoost	To predict lymph node metastasis in extrahepatic cholangiocarcinoma	Tang et al., 2021 [71].
LR demonstrated best performance, achieving an AUC of 0.754.	SVM, KNN, RF, and LR	To predict distant metastasis in esophageal cancer	Zhu et al., 2022 [72].
Radiomics-based prediction of treatment responses			
The axial and coronal combination model in ResNet (AUC = 0.85) demonstrated best performance.	AlexNet, GoogLeNet Inception v3, and ResNet-101	To predict treatment outcome in oropharyngeal squamous cell carcinoma by DLs	Fujima et al., 2021 [73].
The average accuracy of C-SVM, R-SVM, and C-R SVM were 0.712, 0.792, and 0.844, respectively, while the average AUC values were 0.775, 0.804, and 0.877.	SVM	To predict prognosis of downstaging treatment in hepatocellular carcinoma	Wang et al., 2023 [74].
DLRPM exhibited superior prediction performance compared to single-scale prediction models, achieving an AUC of 0.927 in the validation set.	DLRPM: based on SVM	To predict response to chemotherapy in breast cancer patients	Zhang et al., 2023 [75].
Radiomics-based prediction of treatment complications			
Combined model (RF) of radiation dose and radiomics resulted in best performance with AUC 0.9993 and 0.9000 in training and testing set.	ResNet50 for feature extraction RF for classification	To predict radiation pneumonitis after radiotherapy	Huang et al., 2022 [76].
RF achieved AUC range of 0.713 to 0.756.	Linear SVM for feature extraction RF for classification	To predict post-radiation nasopharyngeal necrosis after radiotherapy	Liu et al., 2023 [77].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

The radiomic models (N1, N2, N3) with longitudinal MRI yielded AUCs of 0.872, 0.836, and 0.780 for RTLI prediction.	RF	To predict radiation-induced brain injury after radiotherapy	Zhang et al., 2020 [78].

3.7. Clinical Applications Based on Pathomics

3.7.1. Pathomics-Based Prediction of Cancer Diagnosis and Prognosis

In the field of pathomics, following studies made notable contributions to cancer diagnosis and treatment response prediction by employing CNN models and were able to highlight the potential of pathomic analyses in personalized medicine and treatment optimization for cancer patients.

Table 7. Pathomics-based prediction of cancer diagnosis and prognosis.

Outcome/Performance	Modality of AI	Study aim	Author, year
Pathomics-based prediction of cancer diagnosis			
Google Inception V3 yielded average AUC of 98.06%.	Google Inception V3	To diagnose colorectal cancer with DL on weakly-labeled WSIs	Wang et al., 2021 [79].
Pathomics-based classification of malignant vs benign tumors			
Richer fusion network outperformed other models in the literature with average accuracy of 92.9%.	Richer fusion network: based on Sparse denoising autoencoder and VGG16	To classify benign vs malignant breast lesions with WSIs and EMR	Yan et al., 2021 [80].
Pathomics-based prediction of treatment responses			
VGGNet had best predictive ability and was utilized as a backbone model to identify transcriptomic subtypes and predict therapy response.	AlexNet, GoogLeNet, and VGGNet	To evaluate a CNN model that diagnose ovarian cancer and predict treatment response	Yu et al., 2020 [81].

3.8. Clinical Applications Based on Clinical and Laboratory Data

3.8.1. Cancer Diagnosis and Prognosis Based on Clinical and Laboratory Data

In the field of clinical data analysis, following studies showcased the integration of diverse data modalities for cancer prediction and classification by collectively highlighting the potential of integrating clinical and traditional medical data with ML approaches to enhance cancer diagnosis and prognostication. In terms of cancer prognosis, these studies made significant strides in utilizing ML methods for survival prediction, recurrence prediction, and treatment response assessment across various cancer types. These studies collectively demonstrate the effectiveness of ML approaches in leveraging clinical data to predict cancer prognosis, recurrence risk, and treatment outcomes, thus paving the way for personalized cancer management strategies.

Table 8. Cancer diagnosis and prognosis based on clinical and laboratory data.

Outcome/Performance	Modality of AI	Study aim	Author, year
Cancer prediction based on clinical and laboratory data			
CyPath resulted in AUC 0.89, sensitivity of 82.1%, and sensitivity of 87.7% for test set and AUC 0.94 for test set.	CyPath Lung: based on LR	To detect lung cancer in sputum with ML	Lemieux et al., 2023 [82].
XGboost generated the highest AUC value of models, which were 0.915, 0.9529, 0.9557, 0.9614 for diagnosing ASCUS higher, ASC-H higher, LSIL higher, and HSIL higher staged cervical lesions, indicating the acceptable accuracy of the selected diagnostic model.	LR for feature selection Six MLs for classification: Decision Tree, XGBoost, RF, SVM, LR, and Neural net	To predict cervical cancer with HPV screening dataset	Meng et al., 2022 [83].
AutoML had the highest AUC 0.807 of 4 MLs. AutoML had encouraging discriminative power with AUCs of 0.820 in the validation cohort and 0.807 and 0.850 in the two prospective test cohorts.	RF for feature selection AutoML, LR, RF, XGBoost for model establishment	To diagnose prostate cancer with clinical data	Zhang et al., 2023 [84].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

RF model incorporating selected features, exhibited excellent performance in predicting HCC events occurring within 1 year, achieving an AUC of 0.9507. Predictions for the 2-year and 3-year time frames also yielded favorable results, with AUCs of 0.8767 and 0.8307, respectively.	RF	To predict risk of hepatocellular carcinoma in patients with hepatitis C cirrhosis	Zou et al., 2023 [85].
Classification of malignant vs benign tumors based on clinical and laboratory data			
The XGBoost model provided better performance (AUC 0.82) compared with free-to-total PSA ratio (AUC 0.75), total PSA (AUC 0.68) and free PSA (AUC 0.61).	XGBoost	To distinguish benign prostate hyperplasia from prostate cancer using ML	Chen et al., 2023 [86].
Xception CNN showed an AUROC of 0.8741, 0.9199, and 0.8363 for the detection of myeloblasts, promyelocytes, and Auer rods. ENNs resulted in AUCs of 0.8575 and 0.9585 in distinguishing between APL and non-APL AML as well as APL and healthy donors.	XceptionCNN to label cell border Binary ensemble neural nets (ENNs) for classification	To predict acute promyelocytic leukemia from bone marrow smear images	Eckardt et al., 2022 [87].
Xy-SkinNet achieved a 64.75% accuracy rate for its top-ranked diagnosis, surpassing the average performance of dermatologists, which stood at 62.13%.	Xy-SkinNet: based on ResNet and Fast R-CNN	To classify six common skin disease with AI	Huang et al., 2021 [88].
Tumor grading based on clinical and laboratory data			
Incorporating additional non-image information such as cytology and HPV status improved CAIADS' diagnostic performance, with an AUC of 0.712 for LSIL and 0.829 for HSIL and cancer. CAIADS surpassed the diagnostic performance of colposcopists, achieving an AUC of 0.678 for LSIL and 0.777 for HSIL.	Colposcopic Artificial Intelligence Auxiliary Diagnostic System (CAIADS):	To evaluate AI system that diagnose colposcopy images	Xue et al., 2020 [89].
Tumor staging based on clinical and laboratory data			
Neural Network, RF, and NB demonstrated superior classification ability with combined input. Accuracies of Neural network, RF, and NB were 0.767, 0.718, and 0.688, respectively, and the AUCs were 0.793, 0.779, and 0.771.	Decision tree, LR, SVM, RF, Naive bayes (NB), and Neural network	To diagnose lung cancer staging based on tongue images and tumor markers	Shi et al., 2023 [90].
Survival prediction based on clinical and laboratory data			
All six models demonstrated satisfactory predictive performance, with AUC ranging from 0.73 to 0.86. The 3-year model exhibited the highest performance, achieving an AUC of 0.86.	Gradient boosting machine (GBM)	To estimate survival in patients with metastatic prostate cancer	Anderson et al., 2022 [91].
SVM-ELAS performed better than LR-ELAS and CART-ELAS. SVM-ELAS exhibited superior performance with an average AUC of 0.736, demonstrating significant enhancements over SVM-AdaBoost, SVM-Bagging, SVM-SMOTE, and SVM-TomekLinks.	SVM-ELAS, LR-ELAS, CART-ELAS	To predict survival and recurrence in patients with non-small cell lung cancer (NSCLC)	Hu et al., 2022 [92].
The GBM model demonstrated a predictive accuracy for survival with a C-index of 0.751.	GBM	To predict survival in patients with intrahepatic cholangiocarcinoma after liver resection	Ji et al., 2022 [93].
The cause-specific Cox model and PLANN demonstrated the highest performance, closely followed by the Fine-Gray model, RF, and PLANN original.	RF, Partial logistic artificial neural network (PLANN)	To predict survival with data on competing risk	Kantidakis et al., 2023 [94].
The 1-, 3-, and 5-year AUCs were 0.794, 0.849, and 0.872.	RF	To predict survival of patients with urothelial carcinoma	Liu et al., 2023 [95].
A nomogram predicting 1-, 3-, and 5-year survival was created using selected LOFs and HOFs by DeepSurv, demonstrating favorable predictive efficacy for lung cancer patients at 1 and 3 years, with a C-index of 0.744.	DeepSurv: based on a neural network	To predict survival with different features from routine blood test	Luo et al., 2023 [96].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

XGBoost yielded the best outcome with highest AUCs. XGBoost achieved an accuracy of 83% in predicting the mortality rate for Group 1 post-surgical resection and 69% accuracy for Group 2 post-trans arterial chemoembolization (TACE).	Voting ensembles, LR, KNN, Decision Tree, SVM, RF, XG Boost, Light GBM, and Natural Gradient Boosting (NG Boost)	To predict mortality rate with clinical features	Noh et al., 2022 [97].
DeepSurv yielded a C-index of 0.824 using the training cohort, while validation using the test cohort yielded a C-index of 0.821.	DeepSurv	To predict survival with SEER database	Yu et al., 2022 [98].
Recurrence prediction based on clinical and laboratory data			
For 1-year post-NAC, RF outperformed LR with AUC 0.810. For 5-year post-NAC, RF again outperformed LR with AUC 0.829. And for external validation set with SEER database, RF outperformed LR with AUC 0.779.	RF and LR	To predict breast cancer relapse or metastasis with clinical data	Jin et al., 2023 [99].
AdaBoost showed a prediction performance of a sensitivity of 0.673, specificity of 0.807, accuracy of 0.799, AUC of 0.740.	SVM, LR, KNN, NB, RF, gradient boost, AdaBoost, and XGBoost	To predict recurrence in renal cell carcinoma with clinical data	Kim et al., 2022 [100].
Cancer treatment response prediction based on clinical and laboratory data			
The average accuracy from D1 to D3 in predicting outcomes on the test set was 83.21%, with specific accuracies of 83.96% for survival. The optimal DQL model (survival + dysphagia, 2 neural network layers, without radiomics input) demonstrated a 70.4% similarity to physician decisions on the training set and 69.65% on the test set.	Deep Q Learning (DQL): based on neural network	To select treatment and its outcome with clinical data using DQL	Tardini et al., 2022 [101].

4. Discussion

In this review, we presented various AI techniques to utilize multi-omics, radiomics, pathomics, as well as clinical and laboratory data. While some studies focused solely on assessing AI performance using individual data types, a significant proportion incorporated the integration of diverse data types. Studies that purportedly limited to a single data type integrated demographic information into their AI models. This integrated approach is advantageous given the complexity of cancer as a biological phenomenon, consequently bolstering diagnostic and prognostic capabilities.

However, these diverse datasets often comprise a substantial number of features. Some studies have noted overfitting in their models due to the utilization of a larger number of features relative to a smaller sample size [66,96]. This issue is commonly referred to as the 'n << P problem,' where 'n' represents the sample size and 'P' denotes the number of features [102]. Dealing with many features in data can pose challenges when employing AI models, particularly in the context of high dimensionality. One significant challenge associated with high dimensionality is the increased sparsity of data, where information becomes thinly distributed across the feature space. Imagine each piece of data as a dot on a graph. As we add more and more features, the space where these dots exist gets bigger and bigger, making the dots more spread out, or "sparse." Consequently, making accurate predictions becomes challenging unless a substantial number of data points are available. This difficulty is particularly pronounced when analyzing medical data since it often exhibits considerable variation. Hence, researchers take steps to maximize the number of available samples while minimizing the number of features. We observed that many studies have adopted various feature selection and extraction techniques to address this challenge.

Feature selection and extraction can be accomplished by human experts or with computational algorithms. ML methods such as SVM and RF, along with statistical methods including the LASSO-Cox model, were frequently employed for feature selection. Autoencoder, a type of ML algorithm, was a popular method to integrate multi-omics ML data. DL methods were applied more extensively in radiomics data analysis for feature selection and extraction. This preference for DL, particularly CNNs, stems from their efficiency in handling large volumes of data compared to traditional ML or statistical methods. Additionally, CNNs automate the process of feature extraction and classification by identifying patterns and extracting features from images. A limitation of DL lies in its 'black box

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

problem', where it fails to offer interpretations to justify model findings or provide additional clinical insights. Despite this challenge, efforts have been made to demonstrate the importance of features extracted by CNNs. For instance, researchers like Fujima et al. attempted to validate significant radiomic features extracted using CNN through statistical analysis [73]. Unlike DL methods, statistical methods such as the Cox Proportional Hazards (PH) model offer interpretable outcome values. Shapley values derived from the SHapley Additive exPlanations (SHAP) algorithm can interpret outcomes derived from ML methods [86]. Shapley values offer insights into the contributions of features towards specific outcomes.

Another approach to address the 'n << P problem' involves increasing the sample size. Many studies have leveraged data from publicly available datasets such as The Cancer Genome Atlas (TCGA). However, excessive reliance on TCGA data may introduce bias towards the -omics data types present in the TCGA dataset, potentially leading to overfitting of models and resulting in bias and misrepresentation of the outcome. Therefore, initiatives aimed at providing large-scale, multi-modal datasets to the research community are necessary. Moreover, studies that increased number of samples encountered challenges related to imbalanced data. To mitigate this issue, Meng et al. and Hu et al. employed the Synthetic Minority Over-sampling Technique (SMOTE) algorithm, which replicates minority class samples.

Overall, most AI models examined in this study were focused on tasks such as classification, clustering, and regression. These models have demonstrated promising outcomes and performance; however, they are not currently suitable for use in clinical settings. This limitation arises from the predominantly retrospective nature of the studies, which were often single-center and thus subject to inherent biases and variations. Challenges persist in achieving feature reproducibility, interpretability, and generalization, as well as in ensuring model interpretability. Thus, robust prospective studies are necessary to guarantee the safety and efficacy of AI models. Furthermore, concerted efforts to enhance algorithm interpretability and comprehend human-algorithm interactions will be important for future adoption and safety.

5. Conclusion

After gathering a plethora of studies, we were able to draw several general and specific conclusions. As general conclusions, inclusion of AI and ML approaches in the domain of medicine has helped to advance the science of diagnosis and prognosis. Because the use of AI/ML in medicine has increased precipitously, the future developments will clearly lead to improved diagnosis and treatment, efficiency in healthcare delivery, enhanced patient care, precision medicine, and discovery of new drugs or treatments, to name a few. It is important to note that to fully reap the benefits of AI/ML in healthcare, policies for proper and ethical use and development of AI need to be developed in parallel with the technological developments.

Directly related to the specific topic of cancer, AI/ML are being explored as avenues of alleviating the impact of cancer in our communities. With cancer being one of the leading causes of death, improving diagnosis and prognosis in cancer is an area of medicine that has caught the attention of many physicians and researchers. In this review we have provided a systematic synopsis of some of the most promising AI utilizations and discussed the limitations associated with each method. One contrast that illustrated an interesting dichotomy was related to studies presenting an individual data type versus studies that presented a diverse array of data types, both having their limitations and advantages. It is important to note that one of the primary advantages of utilizing AI/ML methods is their ability to incorporate heterogenous data types such as genomic, proteomic, imaging, electronic records, as well as others. The traditional methods of data analytics have been unable to integrate this diverse set of data.

Many different ML techniques were examined in this review including RF, XGBoost, KNN, SVM, amongst others. One of the more popular models that seemed to stand out above the others were CNN due to its ability to handle large amounts of data. Overall, the different modalities outlined in this review have been found to have varying levels of efficacy in improving diagnosis and prognosis. One of the more important factors going forward is to keep in mind the importance of

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

properly utilizing current methods and future methods so that any process that is implemented will be helpful in improving patient care.

Author Contributions: The research was conceptualized, and methodology developed by A.V.B. and H.V., with contributions from H.A. Validation was conducted by H.A., H.V., A.V.B., and Y.O., while formal analysis was performed by Y.O., P.B., and H.A. Investigation and data curation were led by Y.O., P.B., A.V.B., H.V., and H.A. A.V.B. and H.V. provided resources, supervised the project, and administered the project alongside H.A. Writing and editing were collaborative efforts involving all authors. Visualization was conducted by Y.O. and P.B. Funding acquisition was managed by A.V.B. and H.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research and ACP was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number P20GM103499 (to Homayoun Valafar) and seed funding from Prisma Health (Homayoun Valafar and Anna V. Blenda).

Institutional Review Board Statement: Not applicable since this is a review article.

Informed Consent Statement: Not applicable to this study.

Data Availability Statement: The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. B. Copeland, "Alan Turing and the beginning of AI," *Encyclopedia Britannica*. 2024. Accessed: Feb. 04, 2024. [Online]. Available: https://www.britannica.com/technology/artificial-intelligence/Alan-Turing-and-the-beginning-of-AI
2. D. Göndöcs and V. Dörfler, "AI in medical diagnosis: AI prediction & human judgment," *Artif. Intell. Med.*, vol. 149, p. 102769, Mar. 2024, doi: 10.1016/j.artmed.2024.102769.
3. X. Jiang, Z. Hu, S. Wang, and Y. Zhang, "Deep Learning for Medical Image-Based Cancer Diagnosis," *Cancers*, vol. 15, no. 14, p. 3608, Jul. 2023, doi: 10.3390/cancers15143608.
4. A. Samaras *et al.*, "Artificial intelligence-based mining of electronic health record data to accelerate the digital transformation of the national cardiovascular ecosystem: design protocol of the CardioMining study," *BMJ Open*, vol. 13, no. 4, p. e068698, Apr. 2023, doi: 10.1136/bmjopen-2022-068698.
5. R. Qureshi *et al.*, "AI in drug discovery and its clinical relevance," *Heliyon*, vol. 9, no. 7, p. e17575, Jul. 2023, doi: 10.1016/j.heliyon.2023.e17575.
6. I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, p. 160, May 2021, doi: 10.1007/s42979-021-00592-x.
7. F. M. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A Comprehensive Overview and Comparative Analysis on Deep Learning Models: CNN, RNN, LSTM, GRU." arXiv, Jun. 01, 2023. Accessed: Feb. 04, 2024. [Online]. Available: http://arxiv.org/abs/2305.17473
8. Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome Biol.*, vol. 18, no. 1, p. 83, May 2017, doi: 10.1186/s13059-017-1215-1.
9. C. McCague *et al.*, "Introduction to radiomics for a clinical audience," *Clin. Radiol.*, vol. 78, no. 2, pp. 83–98, Feb. 2023, doi: 10.1016/j.crad.2022.08.149.
10. R. Gupta, T. Kurc, A. Sharma, J. S. Almeida, and J. Saltz, "The Emergence of Pathomics," *Curr. Pathobiol. Rep.*, vol. 7, no. 3, pp. 73–84, Sep. 2019, doi: 10.1007/s40139-019-00200-x.
11. L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
12. Z. Rguibi, A. Hajami, D. Zitouni, A. Elqaraoui, and A. Bedraoui, "CXAI: Explaining Convolutional Neural Networks for Medical Imaging Diagnostic," *Electronics*, vol. 11, no. 11, p. 1775, Jun. 2022, doi: 10.3390/electronics11111775.
13. W. Zhang *et al.*, "Novel bioinformatic classification system for genetic signatures identification in diffuse large B-cell lymphoma.," *BMC Cancer*, vol. 20, no. 1, p. 714, Jul. 2020, doi: 10.1186/s12885-020-07198-1.
14. Z. Zeng *et al.*, "Prediction of radiosensitivity and radiocurability using a novel supervised artificial neural network.," *BMC Cancer*, vol. 22, no. 1, p. 1243, Dec. 2022, doi: 10.1186/s12885-022-10339-3.
15. C. Goswami *et al.*, "Molecular signature comprising 11 platelet-genes enables accurate blood-based diagnosis of NSCLC.," *BMC Genomics*, vol. 21, no. 1, p. 744, Oct. 2020, doi: 10.1186/s12864-020-07147-z.
16. M. Mostavi, Y.-C. Chiu, Y. Chen, and Y. Huang, "CancerSiamese: one-shot learning for predicting primary and metastatic tumor types unseen during model training.," *BMC Bioinformatics*, vol. 22, no. 1, p. 244, May 2021, doi: 10.1186/s12859-021-04157-w.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

17. F. Carrillo-Perez, F. M. Ortuno, A. Börjesson, I. Rojas, and L. J. Herrera, "Performance comparison between multi-center histopathology datasets of a weakly-supervised deep learning model for pancreatic ductal adenocarcinoma detection.," *Cancer Imaging Off. Publ. Int. Cancer Imaging Soc.*, vol. 23, no. 1, p. 66, 2023.
18. D. Ding *et al.*, "Machine learning-based prediction of survival prognosis in cervical cancer.," *BMC Bioinformatics*, vol. 22, no. 1, p. 331, Jun. 2021, doi: 10.1186/s12859-021-04261-x.
19. J. Lai, X. Lin, H. Zheng, B. Xie, and D. Fu, "Characterization of stemness features and construction of a stemness subtype classifier to predict survival and treatment responses in lung squamous cell carcinoma.," *BMC Cancer*, vol. 23, no. 1, p. 525, Jun. 2023, doi: 10.1186/s12885-023-10918-y.
20. A. Mosquera Orgueira *et al.*, "Improved personalized survival prediction of patients with diffuse large B-cell Lymphoma using gene expression profiling.," *BMC Cancer*, vol. 20, no. 1, p. 1017, Oct. 2020, doi: 10.1186/s12885-020-07492-y.
21. Y. Fu *et al.*, "Combining a machine-learning derived 4-lncRNA signature with AFP and TNM stages in predicting early recurrence of hepatocellular carcinoma.," *BMC Genomics*, vol. 24, no. 1, p. 89, Feb. 2023, doi: 10.1186/s12864-023-09194-8.
22. H. D. Ravkin, O. Givton, D. B. Geffen, and E. Rubin, "Direct comparison shows that mRNA-based diagnostics incorporate information which cannot be learned directly from genomic mutations.," *BMC Bioinformatics*, vol. 21, no. 1, p. 196, May 2020, doi: 10.1186/s12859-020-3512-z.
23. E. Y. Dessie, J. J. P. Tsai, J.-G. Chang, and K.-L. Ng, "A novel miRNA-based classification model of risks and stages for clear cell renal cell carcinoma patients.," *BMC Bioinformatics*, vol. 22, no. Suppl 10, p. 270, May 2021, doi: 10.1186/s12859-021-04189-2.
24. J.-P. Villemin, C. Lorenzi, M.-S. Cabrillac, A. Oldfield, W. Ritchie, and R. F. Luco, "A cell-to-patient machine learning transfer approach uncovers novel basal-like breast cancer prognostic markers amongst alternative splice variants.," *BMC Biol.*, vol. 19, no. 1, p. 70, Apr. 2021, doi: 10.1186/s12915-021-01002-7.
25. Y. Zhao *et al.*, "Uncovering the prognostic gene signatures for the improvement of risk stratification in cancers by using deep learning algorithm coupled with wavelet transform.," *BMC Bioinformatics*, vol. 21, no. 1, p. 195, May 2020, doi: 10.1186/s12859-020-03544-z.
26. R. Gupta, J. Kleinjans, and F. Caiment, "Identifying novel transcript biomarkers for hepatocellular carcinoma (HCC) using RNA-Seq datasets and machine learning.," *BMC Cancer*, vol. 21, no. 1, p. 962, Aug. 2021, doi: 10.1186/s12885-021-08704-9.
27. M.-X. Li *et al.*, "Using a machine learning approach to identify key prognostic molecules for esophageal squamous cell carcinoma.," *BMC Cancer*, vol. 21, no. 1, p. 906, Aug. 2021, doi: 10.1186/s12885-021-08647-1.
28. Z. Lv and T. Lei, "Systematical identifications of prognostic meaningful lung adenocarcinoma subtypes and the underlying mutational and expressional characters.," *BMC Cancer*, vol. 20, no. 1, p. 56, Jan. 2020, doi: 10.1186/s12885-019-6462-y.
29. H. D. Nguyen, A. Allaire, P. Diamandis, M. Bisaillon, M. S. Scott, and M. Richer, "A machine learning analysis of a 'normal-like' IDH-WT diffuse glioma transcriptomic subgroup associated with prolonged survival reveals novel immune and neurotransmitter-related actionable targets.," *BMC Med.*, vol. 18, no. 1, p. 280, Oct. 2020, doi: 10.1186/s12916-020-01748-x.
30. S. Yan *et al.*, "Comprehensive analysis of prognostic gene signatures based on immune infiltration of ovarian cancer.," *BMC Cancer*, vol. 20, no. 1, p. 1205, Dec. 2020, doi: 10.1186/s12885-020-07695-3.
31. S. Zhong, S. Chen, H. Lin, Y. Luo, and J. He, "Selection of M7G-related lncRNAs in kidney renal clear cell carcinoma and their putative diagnostic and prognostic role.," *BMC Urol.*, vol. 23, no. 1, pp. 1–19, Nov. 2023, doi: 10.1186/s12894-023-01357-9.
32. O. Hamzeh, A. Alkhateeb, J. Zheng, S. Kandalam, and L. Rueda, "Prediction of tumor location in prostate cancer tissue using a machine learning system on gene expression data.," *BMC Bioinformatics*, vol. 21, no. Suppl 2, p. 78, Mar. 2020, doi: 10.1186/s12859-020-3345-9.
33. E. A. Clayton, T. A. Pujol, J. F. McDonald, and P. Qiu, "Leveraging TCGA gene expression data to build predictive models for cancer drug response.," *BMC Bioinformatics*, vol. 21, no. Suppl 14, p. 364, Sep. 2020, doi: 10.1186/s12859-020-03690-4.
34. M. Sato *et al.*, "Clinical significance of metabolism-related genes and FAK activity in ovarian high-grade serous carcinoma.," *BMC Cancer*, vol. 22, no. 1, pp. 1–11, Jan. 2022, doi: 10.1186/s12885-021-09148-x.
35. Z. Wu, B. Uhl, O. Gires, and C. A. Reichel, "A transcriptomic pan-cancer signature for survival prognostication and prediction of immunotherapy response based on endothelial senescence.," *J. Biomed. Sci.*, vol. 30, no. 1, p. 21, Mar. 2023, doi: 10.1186/s12929-023-00915-5.
36. Q. T. Tran, M. Z. Alom, and B. A. Orr, "Comprehensive study of semi-supervised learning for DNA methylation-based supervised classification of central nervous system tumors.," *BMC Bioinformatics*, vol. 23, no. 1, p. 223, Jun. 2022, doi: 10.1186/s12859-022-04764-1.
37. R. P. Sugino, M. Ohira, S. P. Mansai, and T. Kamijo, "Comparative epigenomics by machine learning approach for neuroblastoma.," *BMC Genomics*, vol. 23, no. 1, pp. 1–12, Dec. 2022, doi: 10.1186/s12864-022-09061-y.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

38. H. Javaid *et al.*, "A role for SETD2 loss in tumorigenesis through DNA methylation dysregulation.," *BMC Cancer*, vol. 23, no. 1, p. 721, Aug. 2023, doi: 10.1186/s12885-023-11162-0.
39. L. Gao *et al.*, "Identification of a novel bile marker clusterin and a public online prediction platform based on deep learning for cholangiocarcinoma.," *BMC Med.*, vol. 21, no. 1, p. 294, Aug. 2023, doi: 10.1186/s12916-023-02990-9.
40. X. Guan *et al.*, "Construction of the XGBoost model for early lung cancer prediction based on metabolic indices.," *BMC Med. Inform. Decis. Mak.*, vol. 23, no. 1, p. 107, Jun. 2023, doi: 10.1186/s12911-023-02171-x.
41. L. Yuan, J. Zhao, T. Sun, and Z. Shen, "A machine learning framework that integrates multi-omics data predicts cancer-related LncRNAs.," *BMC Bioinformatics*, vol. 22, no. 1, p. 332, Jun. 2021, doi: 10.1186/s12859-021-04256-8.
42. J. M. Choi and H. Chae, "moBRCA-net: a breast cancer subtype classification framework based on multi-omics attention neural networks.," *BMC Bioinformatics*, vol. 24, no. 1, p. 169, Apr. 2023, doi: 10.1186/s12859-023-05273-5.
43. V. Malik, Y. Kalakoti, and D. Sundar, "Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer.," *BMC Genomics*, vol. 22, no. 1, p. 214, Mar. 2021, doi: 10.1186/s12864-021-07524-2.
44. J. Tian *et al.*, "Deep learning algorithm reveals two prognostic subtypes in patients with gliomas.," *BMC Bioinformatics*, vol. 23, no. 1, p. 417, Oct. 2022, doi: 10.1186/s12859-022-04970-x.
45. L. Tong, J. Mitchel, K. Chatlin, and M. D. Wang, "Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis.," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, p. 225, Sep. 2020, doi: 10.1186/s12911-020-01225-8.
46. A. R. Owens, C. E. McInerney, K. M. Prise, D. G. McArt, and A. Jurek-Loughrey, "Novel deep learning-based solution for identification of prognostic subgroups in liver cancer (Hepatocellular carcinoma).," *BMC Bioinformatics*, vol. 22, no. 1, p. 563, Nov. 2021, doi: 10.1186/s12859-021-04454-4.
47. Y. Jiang *et al.*, "Discovering the molecular differences between right- and left-sided colon cancer using machine learning methods.," *BMC Cancer*, vol. 20, no. 1, p. 1012, Oct. 2020, doi: 10.1186/s12885-020-07507-8.
48. T. Kolisnik *et al.*, "Identifying important microbial and genomic biomarkers for differentiating right- versus left-sided colorectal cancer using random forest models.," *BMC Cancer*, vol. 23, no. 1, p. 647, Jul. 2023, doi: 10.1186/s12885-023-10848-9.
49. C. Chen *et al.*, "Deep learning approaches for differentiating thyroid nodules with calcification: a two-center study.," *BMC Cancer*, vol. 23, no. 1, p. 1139, Nov. 2023, doi: 10.1186/s12885-023-11456-3.
50. W. Liang *et al.*, "Classification prediction of pancreatic cystic neoplasms based on radiomics deep learning models.," *BMC Cancer*, vol. 22, no. 1, p. 1237, Nov. 2022, doi: 10.1186/s12885-022-10273-4.
51. L. Liu, F. Lu, P. Pang, and G. Shao, "Can computed tomography-based radiomics potentially discriminate between anterior mediastinal cysts and type B1 and B2 thymomas?," *Biomed. Eng. Online*, vol. 19, no. 1, p. 89, Nov. 2020, doi: 10.1186/s12938-020-00833-9.
52. L. Ma, C. Wan, K. Hao, A. Cai, and L. Liu, "A novel fusion algorithm for benign-malignant lung nodule classification on CT images.," *BMC Pulm. Med.*, vol. 23, no. 1, p. 474, Nov. 2023, doi: 10.1186/s12890-023-02708-w.
53. T. Tong *et al.*, "Deep learning radiomics based on contrast-enhanced ultrasound images for assisted diagnosis of pancreatic ductal adenocarcinoma and chronic pancreatitis.," *BMC Med.*, vol. 20, no. 1, p. 74, Mar. 2022, doi: 10.1186/s12916-022-02258-8.
54. R. Zhang *et al.*, "The diagnostic and prognostic value of radiomics and deep learning technologies for patients with solid pulmonary nodules in chest CT images.," *BMC Cancer*, vol. 22, no. 1, p. 1118, Nov. 2022, doi: 10.1186/s12885-022-10224-z.
55. J. Zhang *et al.*, "Radiomics nomogram for preoperative differentiation of pulmonary mucinous adenocarcinoma from tuberculoma in solitary pulmonary solid nodules.," *BMC Cancer*, vol. 23, no. 1, p. 261, Mar. 2023, doi: 10.1186/s12885-023-10734-4.
56. Y. Zhang *et al.*, "PB-LNet: a model for predicting pathological subtypes of pulmonary nodules on CT images.," *BMC Cancer*, vol. 23, no. 1, p. 936, Oct. 2023, doi: 10.1186/s12885-023-11364-6.
57. Y. Zhu *et al.*, "Deep learning radiomics of dual-modality ultrasound images for hierarchical diagnosis of unexplained cervical lymphadenopathy.," *BMC Med.*, vol. 20, no. 1, p. 269, Aug. 2022, doi: 10.1186/s12916-022-02469-z.
58. S. Deng *et al.*, "Deep learning-based radiomic nomograms for predicting Ki67 expression in prostate cancer.," *BMC Cancer*, vol. 23, no. 1, p. 638, Jul. 2023, doi: 10.1186/s12885-023-11130-8.
59. K. He, X. Liu, M. Li, X. Li, H. Yang, and H. Zhang, "Noninvasive KRAS mutation estimation in colorectal cancer using a deep learning method based on CT imaging.," *BMC Med. Imaging*, vol. 20, no. 1, p. 59, Jun. 2020, doi: 10.1186/s12880-020-00457-4.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

60. X. Li, C. Xu, Y. Yu, Y. Guo, and H. Sun, "Prediction of lymphovascular space invasion using a combination of tenascin-C, cox-2, and PET/CT radiomics in patients with early-stage cervical squamous cell carcinoma.," *BMC Cancer*, vol. 21, no. 1, p. 866, Jul. 2021, doi: 10.1186/s12885-021-08596-9.
61. Y. Li, X. Wu, Y. Yan, and P. Zhou, "Automated breast volume scanner based Radiomics for non-invasively prediction of lymphovascular invasion status in breast cancer.," *BMC Cancer*, vol. 23, no. 1, p. 813, Aug. 2023, doi: 10.1186/s12885-023-11336-w.
62. L. Yang, L. Gao, D. Arefan, Y. Tan, H. Dan, and J. Zhang, "A CT-based radiomics model for predicting renal capsule invasion in renal cell carcinoma.," *BMC Med. Imaging*, vol. 22, no. 1, p. 15, Jan. 2022, doi: 10.1186/s12880-022-00741-5.
63. P. Yu *et al.*, "Extrathyroidal Extension Prediction of Papillary Thyroid Cancer With Computed Tomography Based Radiomics Nomogram: A Multicenter Study.," *Front. Endocrinol.*, vol. 13, p. 874396, 2022, doi: 10.3389/fendo.2022.874396.
64. H. Zhang, S. Wang, Z. Deng, Y. Li, Y. Yang, and H. Huang, "Computed tomography-based radiomics machine learning models for prediction of histological invasiveness with sub-centimeter subsolid pulmonary nodules: a retrospective study.," *PeerJ*, vol. 11, p. e14559, 2023, doi: 10.7717/peerj.14559.
65. S. Bernatz *et al.*, "Radiomics for therapy-specific head and neck squamous cell carcinoma survival prognostication (part I).," *BMC Med. Imaging*, vol. 23, no. 1, p. 71, Jun. 2023, doi: 10.1186/s12880-023-01034-1.
66. N.-B. Chen *et al.*, "CT radiomics-based long-term survival prediction for locally advanced non-small cell lung cancer patients treated with concurrent chemoradiotherapy using features from tumor and tumor organismal environment.," *Radiat. Oncol. Lond. Engl.*, vol. 17, no. 1, p. 184, Nov. 2022, doi: 10.1186/s13014-022-02136-w.
67. L. Lv *et al.*, "Radiomic analysis for predicting prognosis of colorectal cancer from preoperative (18)F-FDG PET/CT.," *J. Transl. Med.*, vol. 20, no. 1, p. 66, Feb. 2022, doi: 10.1186/s12967-022-03262-5.
68. S. Oh, S.-R. Kang, I.-J. Oh, and M.-S. Kim, "Deep learning model integrating positron emission tomography and clinical data for prognosis prediction in non-small cell lung cancer patients.," *BMC Bioinformatics*, vol. 24, no. 1, p. 39, Feb. 2023, doi: 10.1186/s12859-023-05160-z.
69. A. Eresen *et al.*, "Preoperative assessment of lymph node metastasis in Colon Cancer patients using machine learning: a pilot study.," *Cancer Imaging Off. Publ. Int. Cancer Imaging Soc.*, vol. 20, no. 1, p. 30, Apr. 2020, doi: 10.1186/s40644-020-00308-z.
70. S. Liu, Y. Zhou, C. Wang, J. Shen, and Y. Zheng, "Prediction of lymph node status in patients with early-stage cervical cancer based on radiomic features of magnetic resonance imaging (MRI) images.," *BMC Med. Imaging*, vol. 23, no. 1, p. 101, Aug. 2023, doi: 10.1186/s12880-023-01059-6.
71. Y. Tang, C. M. Yang, S. Su, W. J. Wang, L. P. Fan, and J. Shu, "Machine learning-based Radiomics analysis for differentiation degree and lymphatic node metastasis of extrahepatic cholangiocarcinoma.," *BMC Cancer*, vol. 21, no. 1, p. 1268, Nov. 2021, doi: 10.1186/s12885-021-08947-6.
72. C. Zhu, F. Mu, S. Wang, Q. Qiu, S. Wang, and L. Wang, "Prediction of distant metastasis in esophageal cancer using a radiomics-clinical model.," *Eur. J. Med. Res.*, vol. 27, no. 1, p. 272, Dec. 2022, doi: 10.1186/s40001-022-00877-8.
73. N. Fujima *et al.*, "Prediction of the local treatment outcome in patients with oropharyngeal squamous cell carcinoma using deep learning analysis of pretreatment FDG-PET images.," *BMC Cancer*, vol. 21, no. 1, p. 900, Aug. 2021, doi: 10.1186/s12885-021-08599-6.
74. S.-Y. Wang *et al.*, "Predicting the outcomes of hepatocellular carcinoma downstaging with the use of clinical and radiomics features.," *BMC Cancer*, vol. 23, no. 1, p. 858, Sep. 2023, doi: 10.1186/s12885-023-11386-0.
75. J. Zhang *et al.*, "Development and validation of a radiopathomic model for predicting pathologic complete response to neoadjuvant chemotherapy in breast cancer patients.," *BMC Cancer*, vol. 23, no. 1, p. 431, May 2023, doi: 10.1186/s12885-023-10817-2.
76. Y. Huang *et al.*, "Radiation pneumonitis prediction after stereotactic body radiation therapy based on 3D dose distribution: dosiomics and/or deep learning-based radiomics features.," *Radiat. Oncol. Lond. Engl.*, vol. 17, no. 1, p. 188, Nov. 2022, doi: 10.1186/s13014-022-02154-8.
77. T. Liu *et al.*, "Radiomic signatures reveal multiscale intratumor heterogeneity associated with tissue tolerance and survival in re-irradiated nasopharyngeal carcinoma: a multicenter study.," *BMC Med.*, vol. 21, no. 1, p. 464, Nov. 2023, doi: 10.1186/s12916-023-03164-3.
78. B. Zhang *et al.*, "Machine-learning based MRI radiomics models for early detection of radiation-induced brain injury in nasopharyngeal carcinoma.," *BMC Cancer*, vol. 20, no. 1, p. 502, Jun. 2020, doi: 10.1186/s12885-020-06957-4.
79. K. S. Wang *et al.*, "Accurate diagnosis of colorectal cancer based on histopathology images using artificial intelligence.," *BMC Med.*, vol. 19, no. 1, p. 76, Mar. 2021, doi: 10.1186/s12916-021-01942-5.
80. R. Yan *et al.*, "Richer fusion network for breast cancer classification based on multimodal data.," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. Suppl 1, p. 134, Apr. 2021, doi: 10.1186/s12911-020-01340-6.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

81. K.-H. Yu *et al.*, "Deciphering serous ovarian carcinoma histopathology and platinum response by convolutional neural networks.," *BMC Med.*, vol. 18, no. 1, p. 236, Aug. 2020, doi: 10.1186/s12916-020-01684-w.
82. M. E. Lemieux *et al.*, "Detection of early-stage lung cancer in sputum using automated flow cytometry and machine learning.," *Respir. Res.*, vol. 24, no. 1, p. 23, Jan. 2023, doi: 10.1186/s12931-023-02327-3.
83. B. Meng *et al.*, "Establishment of early diagnosis models for cervical precancerous lesions using large-scale cervical cancer screening datasets.," *Virol. J.*, vol. 19, no. 1, p. 177, Nov. 2022, doi: 10.1186/s12985-022-01908-w.
84. H. Zhang *et al.*, "Artificial intelligence for the diagnosis of clinically significant prostate cancer based on multimodal data: a multicenter study.," *BMC Med.*, vol. 21, no. 1, p. 270, Jul. 2023, doi: 10.1186/s12916-023-02964-x.
85. Y. Zou *et al.*, "Accurate prediction of HCC risk after SVR in patients with hepatitis C cirrhosis based on longitudinal data.," *BMC Cancer*, vol. 23, no. 1, pp. 1–11, Nov. 2023, doi: 10.1186/s12885-023-11628-1.
86. G. Chen *et al.*, "Machine learning-based prediction model and visual interpretation for prostate cancer.," *BMC Urol.*, vol. 23, no. 1, p. 164, Oct. 2023, doi: 10.1186/s12894-023-01316-4.
87. J.-N. Eckardt *et al.*, "Deep learning identifies Acute Promyelocytic Leukemia in bone marrow smears.," *BMC Cancer*, vol. 22, no. 1, p. 201, Feb. 2022, doi: 10.1186/s12885-022-09307-8.
88. K. Huang *et al.*, "The Classification of Six Common Skin Diseases Based on Xiangya-Derm: Development of a Chinese Database for Artificial Intelligence.," *J. Med. Internet Res.*, vol. 23, no. 9, p. e26025, Sep. 2021, doi: 10.2196/26025.
89. P. Xue *et al.*, "Development and validation of an artificial intelligence system for grading colposcopic impressions and guiding biopsies.," *BMC Med.*, vol. 18, no. 1, p. 406, Dec. 2020, doi: 10.1186/s12916-020-01860-y.
90. Y. Shi *et al.*, "Machine learning prediction models for different stages of non-small cell lung cancer based on tongue and tumor marker: a pilot study.," *BMC Med. Inform. Decis. Mak.*, vol. 23, no. 1, p. 197, Sep. 2023, doi: 10.1186/s12911-023-02266-5.
91. A. B. Anderson *et al.*, "Machine learning algorithms to estimate 10-Year survival in patients with bone metastases due to prostate cancer: toward a disease-specific survival estimation tool.," *BMC Cancer*, vol. 22, no. 1, p. 476, Apr. 2022, doi: 10.1186/s12885-022-09491-7.
92. D. Hu, H. Zhang, S. Li, H. Duan, N. Wu, and X. Lu, "An ensemble learning with active sampling to predict the prognosis of postoperative non-small cell lung cancer patients.," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 245, Sep. 2022, doi: 10.1186/s12911-022-01960-0.
93. G.-W. Ji, C.-Y. Jiao, Z.-G. Xu, X.-C. Li, K. Wang, and X.-H. Wang, "Development and validation of a gradient boosting machine to predict prognosis after liver resection for intrahepatic cholangiocarcinoma.," *BMC Cancer*, vol. 22, no. 1, p. 258, Mar. 2022, doi: 10.1186/s12885-022-09352-3.
94. G. Kantidakis, H. Putter, S. Litière, and M. Fiocco, "Statistical models versus machine learning for competing risks: development and validation of prognostic models.," *BMC Med. Res. Methodol.*, vol. 23, no. 1, p. 51, Feb. 2023, doi: 10.1186/s12874-023-01866-z.
95. J. Liu, P. Wu, S. Lai, J. Wang, H. Hou, and Y. Zhang, "Prognostic models for upper urinary tract urothelial carcinoma patients after radical nephroureterectomy based on a novel systemic immune-inflammation score with machine learning.," *BMC Cancer*, vol. 23, no. 1, p. 574, Jun. 2023, doi: 10.1186/s12885-023-11058-z.
96. L. Luo *et al.*, "The potential of high-order features of routine blood test in predicting the prognosis of non-small cell lung cancer.," *BMC Cancer*, vol. 23, no. 1, p. 496, Jun. 2023, doi: 10.1186/s12885-023-10990-4.
97. B. Noh *et al.*, "Machine learning-based survival rate prediction of Korean hepatocellular carcinoma patients using multi-center data.," *BMC Gastroenterol.*, vol. 22, no. 1, p. 85, Feb. 2022, doi: 10.1186/s12876-022-02182-4.
98. H. Yu, T. Huang, B. Feng, and J. Lyu, "Deep-learning model for predicting the survival of rectal adenocarcinoma patients based on a surveillance, epidemiology, and end results analysis.," *BMC Cancer*, vol. 22, no. 1, p. 210, Feb. 2022, doi: 10.1186/s12885-022-09217-9.
99. Y. Jin, A. Lan, Y. Dai, L. Jiang, and S. Liu, "Development and testing of a random forest-based machine learning model for predicting events among breast cancer patients with a poor response to neoadjuvant chemotherapy.," *Eur. J. Med. Res.*, vol. 28, no. 1, p. 394, Sep. 2023, doi: 10.1186/s40001-023-01361-7.
100. H. M. Kim *et al.*, "Machine learning-based prediction model for late recurrence after surgery in patients with renal cell carcinoma.," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 241, Sep. 2022, doi: 10.1186/s12911-022-01964-w.
101. E. Tardini *et al.*, "Optimal Treatment Selection in Sequential Systemic and Locoregional Therapy of Oropharyngeal Squamous Carcinomas: Deep Q-Learning With a Patient-Physician Digital Twin Dyad.," *J. Med. Internet Res.*, vol. 24, no. 4, p. e29455, Apr. 2022, doi: 10.2196/29455.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 June 2024 doi:10.20944/preprints202406.0657.v1

102. D. Feldner-Busztin *et al.*, "Dealing with dimensionality: the application of machine learning to multi-omics data," *Bioinformatics*, vol. 39, no. 2, p. btad021, Feb. 2023, doi: 10.1093/bioinformatics/btad021.
103. Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis Campbell Systematic Reviews, 18, e1230. https://doi.org/10.1002/cl2.1230

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.