Computational and Structural Biotechnology Journal 21 (2023) 2454–2470

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Review article

Machine learning in metastatic cancer research: Potentials, possibilities, and prospects

Olutomilayo Olayemi Petinrin [a], Faisal Saeed [b], Muhammad Toseef [a], Zhe Liu [a], Shadi Basurra [b], Ibukun Omotayo Muyide [c], Xiangtao Li [d], Qiuzhen Lin [e], Ka-Chun Wong [a,f,*]

[a] Department of Computer Science, City University of Hong Kong, Kowloon Tong, Kowloon, Hong Kong SAR
[b] DAAI Research Group, Department of Computing and Data Science, School of Computing and Digital Technology, Birmingham City University, Birmingham B4 7XG, UK
[c] College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
[d] School of Artificial Intelligence, Jilin University, Jilin, China
[e] School of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[f] Hong Kong Institute for Data Science, City University of Hong Kong, Kowloon Tong, Kowloon, Hong Kong SAR

ARTICLE INFO

Article history:
Received 17 January 2023
Received in revised form 26 March 2023
Accepted 27 March 2023
Available online 29 March 2023

Keywords:
Cancer metastasis
Data inequality
Deep learning
Early detection
Machine learning
Metastatic cancer

ABSTRACT

Cancer has received extensive recognition for its high mortality rate, with metastatic cancer being the top cause of cancer-related deaths. Metastatic cancer involves the spread of the primary tumor to other body organs. As much as the early detection of cancer is essential, the timely detection of metastasis, the identification of biomarkers, and treatment choice are valuable for improving the quality of life for metastatic cancer patients. This study reviews the existing studies on classical machine learning (ML) and deep learning (DL) in metastatic cancer research. Since the majority of metastatic cancer research data are collected in the formats of PET/CT and MRI image data, deep learning techniques are heavily involved. However, its black-box nature and expensive computational cost are notable concerns. Furthermore, existing models could be overestimated for their generality due to the non-diverse population in clinical trial datasets. Therefore, research gaps are itemized; follow-up studies should be carried out on metastatic cancer using machine learning and deep learning tools with data in a symmetric manner.

Contents

1. Introduction . . . 2454
2. Machine learning for metastatic cancer drug discovery . . . 2455
3. Response to treatment by metastatic cancer patients . . . 2458
4. Early detection of cancer metastasis and determination of survival outcomes . . . 2459
5. Unravelling tumor heterogeneity using machine learning . . . 2461
6. Minorities in metastatic cancer data . . . 2462
7. Repositories for metastatic cancer data . . . 2463
8. Summary and outlook . . . 2465
Funding statement . . . 2466
CRediT authorship contribution statement . . . 2467
Declaration of interest . . . 2467
Acknowledgement . . . 2467
Appendix A Supporting information . . . 2467
References . . . 2467

1. Introduction

According to the American Cancer Society, cancer refers to a group of diseases caused by uncontrollable growth and rapid spread of abnormal cells in various body parts [1]. Cancerous cells are

* Corresponding author at: Department of Computer Science, City University of Hong Kong, Kowloon Tong, Kowloon, Hong Kong SAR
E-mail address: kc.w@cityu.edu.hk (K.-C. Wong).

https://doi.org/10.1016/j.csbj.2023.03.046

developed when normal cells stop functioning as they ought to, progressing rapidly into a tumor, and possbily invade other body parts different from their origin. Even worse, cancer behaves differently according to the organ of origin [2]. In 2020, there were 19.3 million new cases of cancer and about 10 million cancer-related deaths worldwide [3]. It is the second leading cause of death in the USA, with an estimate of 1958,310 new cases and 609,820 Americans expected to die due to cancer in 2023 [4]. In the USA, the adoption of improved targeted therapies has resulted in the steady decline in specific cancer-induced mortality rate to 29% between 1991 and 2017. However, the mortality rate traceable to other cancer types such as liver cancer, pancreatic cancer, uterus cancer, sarcoma and, *de novo* or recurrent metastasis remain stagnant. Unfortunately, most patients in this category die within five years of diagnosis [5,6].

Metastasis is the invasion and spread of abnormal cancer cells to adjoining body parts and organs, with over 90% mortality, making it the primary cause of cancer-induced death [7,8]. It is influenced by several things, including the microenvironment [9] or the resistance of inhibitors in the body [10]. Fig. 1 shows the process of a primary tumor metastasized in the brain. A review by Schneider & Pozzi [11] discusses some factors that either aid or inhibit the occurrence of metastasis. Some of these influencing/contributing factors are natural compounds or chemicals in the human body. An example is the gut microbiome, a microbial signature known to promote cancer development [12]. Although metastasis can occur in any organ of the body, specific body organs such as bone, brain, lungs, lymph nodes and liver are more prone as sites for specific cancer metastasis (see Fig. 2) [13]. For lung cancer patients, the most common pathological subtype of patients with bone metastasis is adenocarcinoma [14,15]. However, the spine, which consists of the cervical, lumbar, and thoracic segments, is the most frequent site of bone metastasis according to Zhang & Gong [15], whereas the rib is the most frequent site according to Zhou et al., [14].[1,2] .

The continuous generation of large and complex datasets in healthcare is a significant enabling factor in the adoption of machine learning. This is because it has a proven potential for analyzing these complex datasets, thus advancing the technological objective of precision medicine in cancer [12]. In Fig. 3 we depict the typical framework of data science techniques that can be categorized under ML for detection and prediction analysis. In the figure, we show that there are different types of data, which must undergo data preprocessing (such as removal of duplicates, detection of outliers and incomplete data, and addressing the issue of missing data) before they are fit for analysis. The data can either be structured or unstructured. Structured data are quantitative and organized; hence, they can be easily analyzed using predictive software. Unstructured data such as images, video, audio, and text, on the other hand, are unorganized. Input data visualization helps to detect outliers, gain prior information about the data, and inform appropriate procedure for subsequent analysis. Feature engineering which involves the extraction and selection of useful features is a key step in the framework. For example, convolution is an efficient way of feature extraction when working with image data in deep learning. In other models, dimension reduction and feature maps are used for eliminating data redundancy. Moving on to model training, data augmentation (particularly in deep learning) is sometimes necessary to increase the robustness of the trained model by introducing it to different data formats. However, a concept that is popularly employed in classical ML for imbalanced data is oversampling. It helps to make sure that there are sufficient samples of each class label in the data during training. Furthermore, an important procedure undertaken to evaluate and tune model performance is cross validation. This technique is used during training for the tuning of hyperparameters toward the model robustness and generalization to new data. Parameters and hyperparameters are unique for different algorithms. Nevertheless, to produce models that are not susceptible to overfitting, parameter tuning should be expertly done.

With metastasis being the leading cause of mortality in cancer patients, ML frameworks aimed at early detection, identification of the specific form of metastasis, and staging can enable proper diagnosis and treatment recommendation [16]. Studies also show that metastasized tumor retains the properties of its primary organ. With the unique molecular signature of metastasized tissues, ML algorithms can be used to identify primary lesion from gene expression data. This is a useful feature for distinguishing between tumor types as a complementary procedure to tumor biopsy [2,17,18]. Furthermore, generated data such as gene expression data are usually high-dimensional, containing heterogenous molecular profile of tumors as features. The manual selection of these features is complex, unless handled with computational tools [19]. In addition to feature selection, the use of machine learning for cancer research ranges from risk assessment, lesion grading and genomics, lesion detection and characterization, imaging, prognosis, staging, therapy response, and other downstream applications [20].

Previous papers have contributed to the review of computer-aided diagnosis of metastasis cancer [21–24,25,26]. While these studies are quite valuable, they seem to be limited in their scope of coverage, with most of them focusing on only one type of cancer and/or one aspect of ML/DL application in cancer metastasis research. In this article, we extend the review of the application of machine learning in metastatic cancer beyond one body organ and data type. We give a comprehensive review of the available recent researches and progress on the usage of ML for the prognosis and detection of metastasis, determination of overall survival, response to treatment, tumor heterogeneity, and the occurrence of racial disparities in the data used for analyses. We took the forms of presentation of the datasets and their peculiarities to specific computational methods into cognizance. The necessary procedures for improving the performance and the limitations of specific methods were also discussed. Finally, based on the potentials and possibilities of ML, we discuss their prospects for metastatic cancer and highlighted new directions in the future application of computational tools to metastatic cancer.

2. Machine learning for metastatic cancer drug discovery

Over the years, the need to provide treatment options for several diseases, especially a mortality-inducing disease like cancer has become paramount. Drug discovery and development is a long process which takes quite a number of years and involves lots of funds. In fact, the launching of a new drug can take up to 15 years and over $1 billion in cost [27]. For the identification of novel targets, the drug design field is being revolutionized using several virtual screening approaches. Virtual screening (VS) is important for the repositioning and repurposing of drug for the optimization and quick characterization of novel drug candidates while speeding up drug discovery [28,29]. Ligand-based virtual screening and structure-based virtual screening are two major areas of VS. With the involvement of 3D visualization in VS, objective insight and ease of manipulation is derived. As the popularity and efficiency of deep learning tools advances, the visualization and analysis of 3D images are easier.

A key technology that falls under the domain of drug discovery is molecular docking. Molecular docking is a widely used tool in VS for the streamlining of search for drug-target interaction, especially

[1] Metastatic Colorectal Cancer May Spread Early in the Disease, Study Finds was originally published by the National Cancer Institute.

[2] Häggström, Mikael (2014). "Medical gallery of Mikael Häggström 2014". WikiJournal of Medicine 1 (2). DOI:10.15347/wjm/2014.008. ISSN 2002–4436. Public Domain.

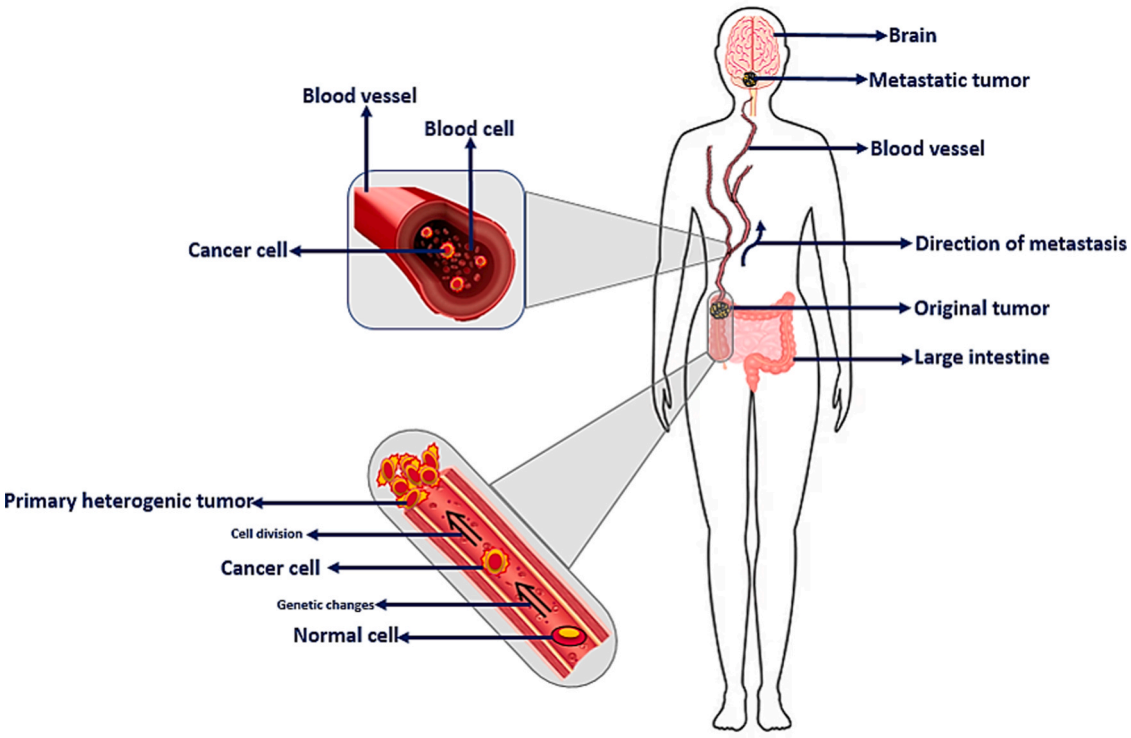

Fig. 1. A depiction of the process of primary tumor metastasized in the brain. The figure was adapted from "Metastatic Colorectal Cancer May Spread Early in the Disease, Study Finds" which was originally published by the National Cancer Institute.

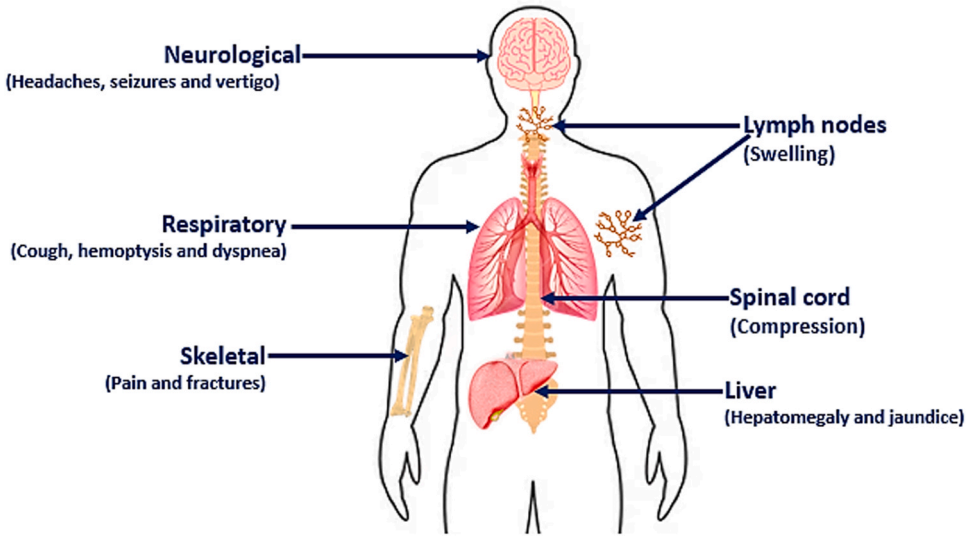

Fig. 2. Common sites and symptoms of metastasis in the body. [Mikael Häggström (25 July 2014). "Medical gallery of Mikael Häggström 2014". *WikiJournal of Medicine* **1** (2).]

when there is a confirmation of the presence of protein 3D structure. Docking aims to predict how a ligand will attach to a receptor in its binding site using either empirical, descriptor-based, knowledge-based or force-field-based scoring functions [30,31]. As a successful *in silico* method, docking predicts the interactions between molecules and targets. However, the shortcomings of docking include conformational and structural flexibility, directional interactions, low accuracy due to assumptions and simplifications in scoring functions [32]. Over time, the involvement of ML in the formulation of the underlying models for docking has increased the accuracy of molecular docking. A key area of its impact is its application in the reliability of scoring functions. The limitation of the traditional scoring functions for reverse docking approach is tackled with ML's ability to enable the scoring functions efficiently discriminate between non-targets and targets [33,34].

Machine learning have also been proven to be effective tools for metastatic cancer drug discovery in the areas of toxicity prediction, drug repositioning, virtual screening, and the prediction of the bioactivity of molecules [35,36]. It aids the virtual screening process, either as a standalone method or an ensemble with other VS methods. These tools have been effective for enhanced similarity search, performance evaluation, and the improvement of scoring functions, thereby improving the drug design and discovery process [37,38]. The chemical space is large, containing billions of chemical structures that needs to be explored [39]. Each compound in the chemical compound datasets used in screening usually contains a

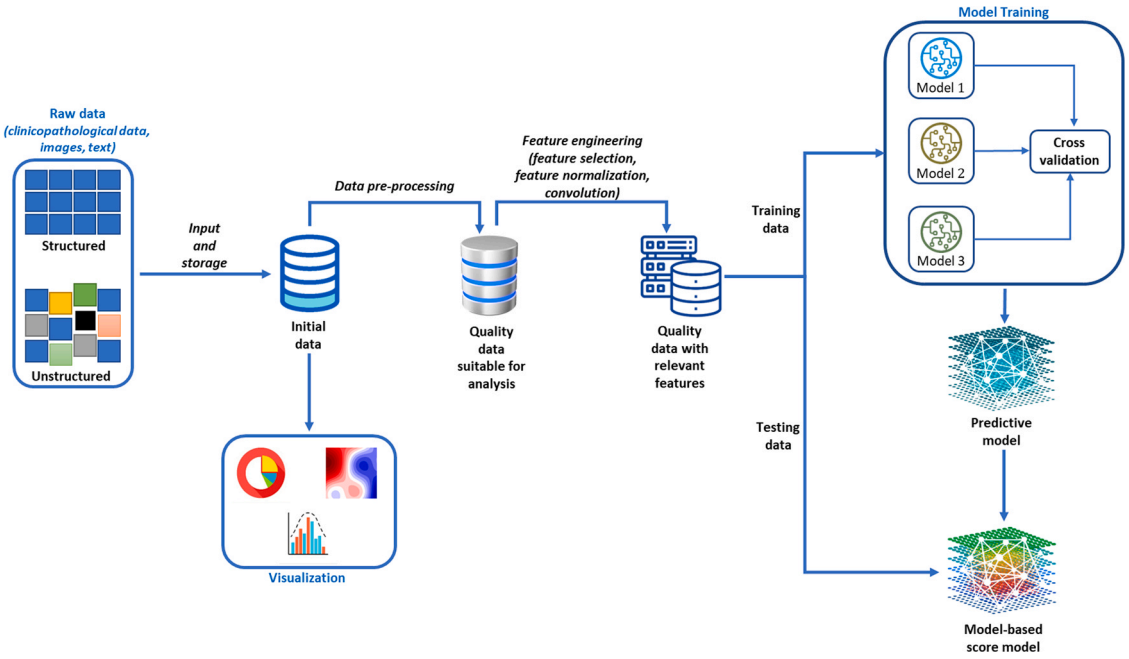

Fig. 3. A Typical Framework of Data Science Techniques that can be categorized under ML for Detection and Prediction Analysis.

large number of files for each compound. The analysis of these datasets, new chemical descriptors, and assessing the scores of docked poses require the use of computation-intensive procedures such as feature ranking/selection, and visualization [40]. Machine learning therefore appear well-suited to these roles, albeit at a heavy computational overhead. In addition, ML algorithms have been used to detect the response of patients and cancer cell lines to new drug and combination of multiple drugs [41].

Meanwhile in structure-based VS, the explicit representation of the receptor in each docking run results in a high expense of computation, making it difficult to incorporate the receptor's flexibility. Combining ensemble docking is a viable alternative where consensus strategies is used to aggregate the scores. Using machine learning classifiers such as gradient boosting trees and logistic regression, Ricci et al., [42] analyzed ensemble docking results with repeated 4-fold cross validation for 30 times. Their results showed the significant performance of the ML methods over the traditional consensus strategies. The ensemble docking result for each protein was however represented as a matrix before analysis. The representation indicates the importance of preprocessing techniques such as feature engineering and representation of different data types in ML. Ensemble of algorithms also utilize the concept of the wisdom of crowds for prediction of drug-drug interaction using more than one algorithm. In Kumar et al., [43], the deep learning architecture used combined algorithms such as convolutional neural network, recurrent neural network, and mixture density network in the prediction of drug synergy in the development of cancer drugs. In the same trend, Sharma & Rani [44] used a modified rotation forest to predict cancer drug sensitivity. Compared with the Cancer Cell Line Encyclopedia (CCLE), and Genomics of Drug Sensitivity in Cancer, the ensemble method improved the prediction of anticancer drug response. However, to address the problems of simultaneously integrating multiple sources of information, novel techniques have been developed. One of these is the combination of multiple kernel learning algorithms with Kronecker regularized least squares into a large drug-target interaction framework [45]. This conceptual approach integrates heterogeneous sources of information into one chemogenomic space which simplifies the prediction of drug-target interactions.

A major challenge during the metastatic cancer drug discovery process is the failure of the lead compound during the trial phase. An example is an ovarian cancer drug known as Olaparib which failed its Phase I trial [46]. Evidently, this must have led to a waste of time and other resources. In addition, the drug development pipeline is impacted by complicated and massive data from microarrays, genomics, proteomics, and clinical trials. Therefore, it provides an opening for the deployment of machine learning. Furthermore, ML can aid cancer drug discovery by identifying targets and hits, optimizing the lead compounds [47], and predicting cancer treatment outcomes during clinical trials [48]. This is seen in its ability to identify drug-cancer cell interactions from *in vitro* databases [49]. A typical example is the aberrant activation of Signal Transducers and Activators of the Transcription 3 (STAT3) which results in oncogenic gene expression for tumor proliferation and metastasis, while its activation in immune cells elevates immunosuppressive factors [50]. For its role in tumor formation and metastasis, STAT3 signalling pathway is a therapeutic target in cancer treatment [51]. Classical ML algorithms such as SVM, KNN, Gaussian naive Bayes, and random forest were used in [52] to classify inactive and active inhibitors for a STAT3 drug-target based on a 10-fold cross validation. The ML-based virtual screening revealed 20 compounds which were active against STAT3, and were docked into STAT3 active site for anticancer drug development. In a similar context, [53] used machine learning molecular docking simulation to determine the structure-activity relationship between eight different essential oils from *Ocimum basilicum* and BRCA1 and BRCA2, which are protein targets for breast cancer. The study used a lazy predict package which contains a suite of machine learning methods to reveal that components from these essential oils can be used in the development of drugs against MCF-7 breast cancer cell line. Furthermore, candidates with effective biological activity against the protein targets were identified, while the inactive ones were detected and excluded to prevent resource wastage from unsuccessful testing.

Due to the high cost and time implications associated with traditional drug discovery methods, researchers have been compelled to embrace more economical, yet powerful techniques of ML. These have been instrumental in metastatic cancer drug sensitivity prediction and the deduction of knowledge from drug-target

interactions, particularly within the framework of precision oncology in cancer treatment therapies. However, a potential limitation to this approach is the quality of data used for analysis. [54]. Moreover, the low signal-to-noise ratio of small-molecule structure-activity data makes prediction of new candidate drugs difficult in most prediction tasks [55]. Certain results also show that ML methods can sometimes outperform complex ones [56]. Whereas, a general performance-enhancing technique in the application of ML in drug discovery is to carry out analysis using combination of methods such as ensemble methods. Furthermore, learning based on an existing deep learning base model (e.g., ImageNet) can also be used to train a more complicated deep learning-based prediction sub-model. This approach is immensely useful in reducing model training time and effort [57].

3. Response to treatment by metastatic cancer patients

Resistance to anticancer drugs is a complex process that can arise from drug target alteration. One of the earliest studies in this area shows that the response of patients to chemotherapy and their survival depend on clinical factors such as the use of combination chemotherapy, previous record of surgery, response to chemotherapy during the early stage of treatment, and nourishment [58]. Apart from these, multi-drug resistance can also cause the release of drugs outside the cell, and a reduction in their absorption or their inactivation. Other factors include tumor heterogeneity, tumor microenvironment, and cancer stem cells. Beyond the cellular level, genetic mechanisms that also contribute to drug resistance include apoptosis pathway blocking, micro-RNA, epigenetic altering, gene amplification, and DNA repair. Other fundamental causes for the reduced efficacy of cancer drug therapies (or drug resistance) could either be a change in chemotherapeutic agent target or drug metabolism [59–61].

Furthermore, the same way antibiotics overdose can induce drug resistance to a specific bacterium in other ailment or microbial infections, resistance to cancer drugs can also occur due to the genetic instability of human cancer cells having high proliferation rate [59]. All these factors contribute to metastatic cancer drug resistance and mortality in some patients [62]. Adverse reaction to the different types of treatment can also speed up the spread of the tumor. This, in addition to resistance to specific treatments, may necessitate the discontinuation of the treatment in question and a switch to another type. One of the challenges associated with metastatic cancer treatment involves recognizing the appropriate treatment that maintains the best quality of life suitable for the patients. Sometimes, this may require discontinuing treatment even before an adverse reaction occurs. A practical study that utilized ML to evaluate this tendency was carried out by Petinrin et al., [63]. In their work, they used an optimized Gaussian process classifier in the prediction of docetaxel treatment discontinuation for metastatic castration-resistant prostate cancer patients. By training the model on a combination of three datasets, appropriate metrics such as Area under Precision-Recall Curve (AUPRC), and Area under Curve (AUC) were used to evaluate the model due to the skewness of the data. The optimized classifier performed better than traditional ML methods in predicting treatment discontinuation for the mitigation of adverse effects. Using random forest for feature ranking, the crucial features that influenced the analysis were seen to be parameters associated with lab record data. These include glucose level, benign or malignant neoplasm, creatinine, blood urea nitrogen value, target lesion, testosterone level, and lymphocytes value. Missing data, which happens to be a common occurrence in health databases, was handled using single imputation method. However, multiple imputation methods are recommended to be more suitable for handling missing data [64]. Regarding patient's response to drug in metastatic and recurrent colorectal cancer patients, Lu et al., [65] conducted a cross-validation study to determine recurrent or metastatic colorectal cancer patients who show sensitivity to FOLFOX (s-FU, leucovorin and oxaliplatin) therapy. Their approach was to apply six parameter-tuned machine learning algorithms to a combination of microarray datasets. The adopted models were based on support vector machine and random forest algorithms, which showed significant performance superiority over other algorithms such as k-nearest neighbor, gradient boosting machine, decision tree, and neural network. In a previous study [66], SVM and random forest have been noted as suitable algorithms for microarray datasets, especially when the genes are selected. The variety of several microarray datasets improved the reliability and universality of the analysis compared to the use of a single microarray dataset. This concept is similar to data augmentation in deep learning where a model trained on augmented data with variation is able to make better prediction. Using cross validation, the optimal parameter(s) for each algorithm can be tuned to improve performance of the models in the determination of each patient's sensitivity to FOLFOX treatment. Different parameters are peculiar to each algorithm. During the training of the model, the best fit parameters are eventually chosen. In addition, cross validation helps to curb overfitting, and improve model generalization.

Moreover, some studies reveal how biological factors affect the responses of patients to treatment. In a multivariate cox regression analysis by Van et al., [67], independent biological factors such as stem cell-ness, proliferation, Epithelial to Mesenchymal Transition (EMT) and DNA repair, radio-sensitivity, tumor acute and chronic hypoxia, and CD8+ T-cell parameters were seen as biomarkers which largely contribute to locoregional control rate in response to radiation. Furthermore, surface markers and transcriptomic phenotype, abundance of neoantigens, suppressive tumor microenvironment, CD8+ and CD4+ T cells, T cells stemness and memory, chimeric antigen receptor (CAR) design and integration site, cytokine production, tumor infiltration, epigenetics (signatures), surface markers and transcriptomic phenotype, CAR methylation, tumor load, antigen escape, immunological clearance, and inflammatory cytokines are noted as biological and molecular factors which can be used to determine the response to adoptive cell therapies [68]. Risk stratification analysis can be conducted based on these biological factors to determine when treatment regimen should be intensified or lessened. Tseng et al., [69] showed that by using machine learning algorithms, patients risk level and survival can be assessed based on the specific cancer type, the occurrence of distant metastasis, or the occurrence of locoregional recurrence. These studies show that models trained on the genetic and clinicopathologic data can effectively inform the classification of patient groups based on the level of risk and survival. Hence, sufficient attention can be placed on patients in the high-risk group to improve their quality of life and longevity.

Furthermore, study shows that the adherence of patients to treatments such as tamoxifen citrate influences the recurrence and survival rate of patients [70]. Factors such as prior clinical procedures, health care encounters, previous treatments, and comorbidity influence patients' adherence to particular treatments. To examine the impact of treatment adherence within the ML context, Yerrapragada et al., [71] trained machine learning models such as logistic regression, random forest, boosted logistic regression, and feedforward neural network with a data of 3022 patients where 40% were nonadherent patients, and 60% were adherent patients. The trained models were evaluated using area under receiver operating characteristic curve (AUROC). Logistic regression, the best performing model for the data, was used to determine the highest contributing variables. The model showed that age, pre-treatment procedures (such as arterial surgery, radiation oncology, and lymphatic nuclear medicine), therapy (such as antidepressants, beta blocker, and stimulants), and previous diagnoses accounted for the patients'

adherence to treatment. The model however had higher confidence among patients classified as adherent compared to patients classified as nonadherent. The results of these studies show the capability of ML to potentially limit unnecessary invasive procedures while making personalized and adaptive therapy for patients. The genetic makeup of individual patient, and the presence of anti-tumorigenic agents such as cyclooxygenases and lipoxygenases inhibitors makes each patient's response to treatment different [11]. On a similar note, using Convolutional Neural Network and Recurrent Neural Network, Xu et al., [72] predicted the response of non-small-cell lung cancer patients to definitive chemo-radiation treatment. The outcomes of the treatment were either distant metastasis, progression, or local-regional recurrence. Based on the analysis of time series CT images of the patients, deep learning methods were used to improve outcome prediction by integrating the imaging scans at multiple time points.

An area of cancer treatment that provokes keen interest in researchers is the rate of relapse in patients. For instance, studies show that about 70% of colorectal cancer patients with liver metastasis tend to relapse within two years of treatment with surgical resection [73]. Thus, Wei et al., [74] utilized the computational and predictive ability of deep learning to develop a deep learning-based radiomics model to determine the response of colorectal liver metastases patients to chemotherapy treatment. Convolutional layers were used to extract the radiomic features from the CT and MRI images, and the classifier part of the deep learning model effectively predicted response to chemotherapy. Zhu et al., [75] also used deep learning to determine the pathological tumor regression grade response of colorectal cancer liver metastasis patients receiving preoperative chemotherapy treatment. These studies indicate that MRI-based deep learning model could effectively predict pathological response compared to the traditional Response Evaluation Criteria in Solid Tumours and survival outcomes after hepatectomy based on the pre-chemotherapy and post-chemotherapy MRI.

With many available treatment options, the factors such as differences in genetic makeup of different patients and their body's ability to deal with external agents can affect how the body will respond to a chosen treatment. As a basis for pre-operative treatment planning, ML can be utilized to predict the expected response based on specific information. A non-general risk stratification for different treatment options tailored for each patient will enable the physician, patients, and stakeholders to be involved in the treatment process to make informed decision on treatment options without risking expedited mortality. Since ML models are trained based on specific features, the appropriate features which takes cognizance of the factors should be considered in model building. In addition, important features can be extracted during convolution of image data. We show the use of some ML algorithms for the prediction of treatment discontinuation for metastatic cancer patients in the supplementary material. The code for reproducibility can be found on our github page.

4. Early detection of cancer metastasis and determination of survival outcomes

Dissemination of metastasis occur at the early stage of malignant cancer progression, but it often takes years before it is clinically manifested. The occurrence of clinical manifestation indicates closeness to mortality for the vast majority of patients [76]. Cancer research over the years seek to understand the mechanism behind migration of cancer cells and metastasis. The identification and monitoring of biomarkers can give relevant information that help to predict the onset of metastasis, and in essence increase the chances of survival [77]. A study conducted on 295,213 patients revealed that a more significant percentage of the newly diagnosed breast cancer patients had metastasis in the bone compared to the liver, lungs, and brain [78]. This was attributed to the interaction between the osteoblasts (or osteoclasts) and the tumor cells, making breast osteoblast-like cells an early marker for bone metastasis [79]. Additionally, genetic markers are considered strong contributors to how the diagnosis and treatment of breast and bone cancers is approached. The study conducted by Cai et al., [80] provided evidence to this effect. They revealed that the disruption of the miRNA-dependent regulatory axis, which links the tumor suppressor microRNA-124 to the interleukin-11-induced osteolysis, makes microRNA-124 and Interleukin-11 possible prognostic markers. These also play an important role in the identification of new therapeutic targets for early stage breast cancer and advanced stage bone metastatic patients. Furthermore, independent risk factors such as the serum concentration of biomarkers CA-125, alkaline phosphatase and the histopathological type is advised to be noted in diagnosed patients for early detection of bone metastasis [14].

A significant aspect of the application of ML in cancer research is the early detection of cancer cells in the body, and the subsequent prediction of the expected period of survival for affected patients. The use of these statistical methods for the determination of associated risks and life expectancy has become prominent over the years [81]. Furthermore, deep learning is used to identify specific tumor types based on image data. DeepSurv, a deep learning-based algorithm, showed better performance in the recommendation of treatment and prediction of survival outcome of non-small cell lung cancer patients compared to nodes, tumors and metastatic staging systems [82]. MetaCancer is a similar model which applies deep learning to microRNA sequencing (microRNA-Seq), RNA sequencing (RNA-Seq), and DNA methylation data for the identification of metastatic cancer status [83]. The MetaCancer model showed remarkable effectiveness as a preoperative noninvasive tool in diagnosing lymph node metastasis. The data used for the analysis was obtained from magnetic resonance imaging and DL-mined tumor image information of stage IB to IIB cervical cancer patients. Based on the AUC and Kaplain-Meier evaluation metric, the status of the lymph node metastasis and the survival outcome of the patients were deduced [84]. In the same vein, for a preoperative diagnosis of metastatic lymph nodes in rectal cancer patients, Ding et al., [85] reported the use of a faster region-based convolutional neural network nomogram. Unlike solid tumors, there are difficulties in recognizing lymph nodes that have large quantities and minor differences. The model exhibited excellent performance based on reliability, convenience, and the capacity to predict metastasis status and degree. However, the region-based convolutional neural network is based only on MRI images and excludes pathological images. Before creating a nomogram for predicting the presence of lymph node metastasis, algorithms such as logistic regression can be used to find the best-fit model. Peak et al., [86] used this method to detect lymph node metastasis in penile cancer. However, the data used had limited clinicopathological information. In their study, details regarding the lymph node, such as dissection type, and the time between penile tumor surgery and metastasis were not indicated. Essentially, treatment of metastatic cancer is based on factors peculiar to each patient, such as the number of lymph node metastasis.

In addition, due to the low accuracy recorded on the routinely-used preoperative methods in determining the staging/number of lymph node metastasis in gastric cancer patients, Dong et al., [87] implemented a deep learning radiomic nomogram for early detection of metastasis. The radiomics technique converts the medical images into features before selecting the important ones for quantitative analysis. The deep learning radiomic nomogram had a high correlation with the overall survival of locally advanced gastric cancer patients. However, despite the model's predictive performance for N staging, the model showed that the combination of computed tomography (CT) and endoscopic ultrasonography has a higher chance of improving the accuracy of N staging. The derived nomogram can also be used with machine learning methods.

Meanwhile in some cases, metastasis is already present at the time of cancer detection. For instance, it is highly likely that metastasis has already occurred at the time of rectal cancer detection [88]. In tackling the difficulty associated with rectal cancer metastasis due to the wide variation in the overall survival, a study by Zhao et al., [89] revealed a high concordance index of lasso penalized cox proportion hazard regression. However, research by Nicolò et al., [90] on the possibility of a metastatic relapse by breast cancer patients revealed that the mechanistic model built for prediction performed similarly to machine learning models. With a C-index of 0.65 (95% CI, 0.60–0.71), the mechanistic model had comparable performance to examined machine learning algorithms. Nevertheless, a need for the validation of mechanistic model on external datasets is recommended. Notwithstanding, the predictive performance of ML will enhance accurate expectations and clinical decision-making for patients and doctors.

Switching our attention from patients to medical personnel, factors such as psychology, exhaustion, fatigue, and level of expertise can cause a degradation in performance and diagnostic accuracy. Consequent to the fact that human concentration and efficiency can wane over time, Liu et al., [91] trained, tested, and validated some classical ML and DL models for the determination of metastatic auxiliary lymph nodes in breast cancer patients based on contrast-enhanced computed tomography images. The models were trained with 800 image samples. Although specific ML methods such as support vector machine and random forest performed better than some of the examined DL architectures, the overall best performing model, DA-VGG19, is a deep learning-based model. This outcome can be attributed to the improved performance of deep learning when data augmentation is applied to image data. Similar to the way where data augmentation can be a necessary preprocessing step for deep learning, dimensionality reduction can be implemented on a dataset before training with a machine learning algorithm. In Chu et al., [92] where linear regression, support vector machine, decision tree, and k-nearest neighbour were trained based on an initial dataset of oral squamous cell carcinoma patients, the data dimension was reduced with principal component analysis (PCA) and bivariate analysis. In predicting the progression of the disease up to the stage of metastasis, whether the dataset had undergone data reduction or not, affected the performance of each algorithms. Furthermore, the identification of correlated features aided the removal of redundant variables and facilitated better prediction.

The interaction of organs is a major subject of cancer metastasis detection. In fact, bone metastasis is frequent in lung, breast, and prostate cancer. To demonstrate the benefit of computational analysis in the detection of bone metastasis against the frequently used bone scintigraphy, Papandrianos et al., [93] proposed a convolutional neural network (CNN) model in the diagnosis of metastatic breast cancer in the bone using image data of whole-body scans. Compared to previous CNN-based models, the proposed model had superior performance, especially for RGB images compared to the grayscale images. Moreover, sufficient data should be used for deep learning-based models to guarantee accurate outcomes, and techniques like data augmentation should be utilized to make deficient data sources robust. In animals, there are studies which examined the growth of breast cancer bone macrometastasis; which are metastases with tumor cell deposits larger than 2 mm. Prediction was made by integrating imaging parameters from PET/CT and MRI into a neural network. The application of these diagnostics tools in lab animals is a very important testing field for validating their efficacy in living tissues before the full scale adoption in human beings. The essence of these techniques lies in the timing before any symptom or physical abnormalities are observed using the standard imaging methods; the flexible ML model can predetermine the possibility of metastasis based on extracted features such as tissue vascularization and glucose metabolism [94].

Nevertheless, even with an early diagnosis and removal of the primary tumor, a progression to metastasis is still a threat to melanoma patients. It is therefore vital to determine the prognostic biomarkers in these cases. Furthermore, due to the important information contained in the serum about organism's general health status, it is widely regarded as a source of biomarkers. Using machine learning and Kaplan-Meier technique, Mancuso et al., [95] predicted metastasis in early-stage melanoma patients with serological biomarkers together with the histopathological and clinical features of the disease. The algorithm effectively classified patients according to the risk (high or low) of developing metastasis. Factors such as the serum level of interleukin-4, dermcidin, granulocyte-macrophage colony-stimulating factor, and Breslow thickness were identified to contribute to the risk level of metastatic progression. However, a large volume of patient data is recommended to avoid overfitting.

Another important type of metastasis is occult metastasis. They are metastases that are initially undetected during pathological examination. They are different from micrometastasis which are metastases with tumor deposits lesser than 2 mm. The inability to detect the potential spread of cancer leads to grave consequences since early treatment will not be conducted. Therefore, it is essential to uncover occult metastasis before they are clinically evident based on clinical and pathological analyses [96]. However, both early and advanced local diseases are equally susceptible to the risk of occult metastasis [97]. Staging accuracy can be improved based on the early detection of occult metastasis. To reduce the chance of the re-occurrence of a bout with the disease, patients with high risk can be introduced to post-operative treatment. [98]. Variables such as perineural invasion and lymphovascular invasion have been shown to increase the rate of occult metastases [99,100], but their clinical application is limited prior to postoperative pathology [101]. Jiang et al., [102] used a deep neural network for the early identification of clinically occult peritoneal metastasis in gastric cancer patients. The model, built with preoperative CT images, outperformed clinicopathological factors. However, other information such as endoscopic ultrasonography and laparoscopy can be used to improve the specificity and sensitivity of the model. Bur et al., [103] on the other hand, trained four machine learning algorithms namely logistic regression, kernel support vector machine, decision forest, and gradient boosting machine for the detection of occult pathological lymph node metastasis in oral cavity squamous cell carcinoma patients. The data, consisting of five clinical and pathological variables were from a single institution, and missing data were filled using the median single imputation technique. Five-fold cross-validation was employed to avoid overfitting. Based on the four different evaluation metrics, the decision forest performed better in terms of the AUC. Three out of the examined algorithms had the same sensitivity, and gradient boosting had a better specificity. However, all the ML algorithms performed better than models which are based on tumor depth of invasion.

One of the more rare cancer types encountered in diagnosis and therapy is glioblastomas [104]. Despite their rarity, they are regarded as the most common form of brain malignancy and are mostly lethal [105]. To distinguish single brain metastasis from glioblastomas, Bae et al., [106] trained a deep learning model and seven traditional machine learning models using radiomics features from MRI images. Feature selection techniques (such as recursive feature elimination, and LASSO), tree-based method, and parameters optimization were carried out based on tenfold cross-validation. Based on the AUC metric, the DL model performed better than the seven traditional ML methods, namely naive Bayes, k-nearest neighbor, AdaBoost, random forest, RBF-SVM, L-SVM, and LDA. AdaBoost was the best performing model among the conventional machine learning methods. A further comparison with two neuro-radiologists revealed that agreement is better using machine learning classifiers compared to the neuro-

radiologists. This agreement shows that the computational methods better generalize and overcome human bias and errors. In addition, the extraction of relevant radiomic features based on the feature importance ranking improved the performance of the deep learning model.

Moreso, a very powerful technique adopted in ML is the use of ensemble algorithms. They have proven to be better at prediction than using single algorithms. Using somatic mutation data of metastatic breast cancer, Mirsadeghi et al., [107] applied an ensemble of artificial neural network, random forest, and SVM to determine the possible driver genes for metastatic breast cancer. Their work argues that it is crucial to understand the drivers that influence the aggression of cancer cells, and not focus solely on the prognostic biomarkers of different cancer forms. The ensemble method, which is less expensive than bio-molecular techniques, performed better in driver gene prediction than the individual algorithms, based on the aggregated predicted scores from the individual algorithms.

In other works, variants of convolutional neural network, such as graph deep learning have also been used in metastasis prediction based on the goal of the analysis and the data type. The relation graph convolutional neural network used by Xu et al., [108] on image data was essential for the advanced extraction of gene expression features and the construction of the gene regulation network. The model outperformed existing network-based methods based on a ten-fold cross-validation of 1779 data samples. The provision of a higher feature dimension by the feature extraction model enhanced the convolutional neural network's performance. Chereda et al., [109] also utilized the graph convolutional neural network to predict metastatic events in breast cancer patients using gene expression data. Extraction of features from image data, using an appropriate strategy for dealing with missing data, and combining traditional machine learning algorithms can improve the predictive ability of models [110].

Concisely speaking, the identification of biomarkers as indicators which suggest the onset of metastasis enables its early detection. In addition, an increase or decrease in the level of particular clinical lab values beyond certain thresholds can signal a potential risk of a primary tumor becoming metastasized. Since some tumors are hardly detected during pathological examination, computational approach can be a remedy. The extraction of radiomic features, a combination of traditional methods, selection of important features, and tuning of parameters have also been shown as essential factors to consider when developing appropriate models. Furthermore, a combination of methods can be used to extract information from data. Although ML has shown better performance compared to humans in some cases, it is recommended that predictions should be further checked by experts, and prognosis be made based on a high level of agreement between human and computational agents. However, extensive datasets are needed for performance generalization and to establish the effectiveness of the methods discussed.

5. Unravelling tumor heterogeneity using machine learning

Due to the dynamism of cancer, cancer tumor becomes more heterogeneous over the course of disease. This means the tumor mass containing cells with different molecular signatures develop; hence exhibiting different sensitivity to treatment. The existence of these cells between tumors having the same histopathological subtype and within the primary tumor and the secondary tumor is known as inter-tumor and intra-tumor respectively [111,112]. Since the distribution of heterogeneity cuts across multi-omics layers, methods for characterization of tumor heterogeneity should be based on the multi-omics layer instead of individual layers [113]. Cancer detection, treatment, and response to treatment are greatly affected by tumor heterogeneity [114]. The distinctions between the primary and metastasized forms of the same type of tumor in different patients can affect their specific treatment recommendations and responses. In fact, a high level of tumor heterogeneity has a significant influence on survival outcome [115,116]. MRI, CT and/or PET scans are the examples of non-invasive radiological imaging for observing tumor heterogeneity [117]. Single-cell profiling of circulating tumor cells (CTCs) provides a unique perspective on tumor heterogeneity and further contributes to the identification of specific CTCs that contribute to metastasis [118].

There has been minimal success in the use of dynamic contrast-enhanced MRI for distinguishing different primary cancers from metastasis [119]. However, Lang et al., [120] used radiomics analysis to extract texture and histogram features from dynamic contrast-enhanced parametric maps. These were then fed as input to two DL techniques in a comparative study. The techniques adopted were convolutional neural network and convolutional long short term memory networks. The objective of the study was to distinguish other cancers from metastatic lesion in the spine originating in the lung. The use of the convolutional long short term memory network improved the accuracy by 0.1% compared to the convolutional neural network. Moreover, different variants of an algorithm can perform better analysis based on the improvement made to the algorithm. A study by Vera-Yunca et al., [115] reveals that tumor heterogeneity and its derived metrics serve as satisfactory predictors of overall survival of metastatic colorectal cancer patients. Additionally, it could be a factor that improves the prediction of drug efficacy. The authors analysed individual target lesions based on four metastatic colorectal cancer studies to determine the difference in tumor size dynamics since tumor size metrics do not consider tumor heterogeneity. The rule-based classification, cross-correlation analysis and k-means clustering methods applied show that tumor heterogeneity is a useful predictor in the overall survival of metastatic colorectal cancer patients, and hence should be considered in drug design and discovery. The importance of primary tumor location, tumor size, and tumor heterogeneity as predictors of overall survival is further highlighted by the authors' follow-up research [121].

Likewise, an analysis by Wang et al., [122] for the detection of nodal metastasis in lung cancer patients reveals that the tumor size and heterogeneity contribute to the estimation of the deep learning architecture used. Furthermore, using five machine learning methods, Lee et al., [123], by quantifying tumor heterogeneity and angiogenesis properties of MRI images, predicted the molecular subtypes and prognostic biomarkers of breast cancer. Texture and perfusion features were extracted for model training . Based on the AUC evaluation, random forest performed better than decision tree, logistic regression, Naive Bayes, and artificial neural network. Texture irregularity and relative extracellular extravascular space were further revealed as important MRI features in prediction. Daye et al., [124] further establish that statistical tumor heterogeneity MRI profiling helps to improve the prognosis of metastatic cancer patients. In their study based on FOLFOX- or FOLFIRI-based chemotherapy-treated stage IV colon cancer with liver metastasis patients, pathological and standard clinical variables were collected in addition to radiomic features extracted from the metastatic lesions. Image segmentation and texture analysis were carried out before the survival outcome of patients was predicted using a random forest algorithm. The model trained on a combination of pathological, standard clinical and radiomics variables performed better than the models trained on selected variables. In a similar study, the radiomic features derived by the quantification and characterization of the tumor heterogeneity also improved the predictive model by 16% [125]. This suggests that better model performance is more likely when a model is trained on data from several sources, compared to when it is limited to selected sources. Furthermore, the visualization of tumor heterogeneity can aid prediction and consolidate the process of decision making, thereby aiding diagnosis [126,127]. In addition to the visualization of biomarkers

derived from imaging, tumor heterogeneity can be automatically quantified with DL techniques [128]. However, improper and imprecise annotation of training data can affect performance [127].

Despite the utilization of machine learning algorithms in assisting clinicians and patients in the development and choice of personalized therapies, proper assessment and characterization of tumor heterogeneity for the collection of radiomic features before the application of machine learning will help to improve performance [111,129]. This is especially important because of the significant impact tumor heterogeneity has on patients' sensitivity and response to treatment. Nevertheless, this should be done with careful consideration given the fact that high variability exists between patients and cancer types. Moreover, the impact of tumor heterogeneity on patients' survival outcomes suggest that it should be a major consideration in the development of drug design and discovery frameworks. The advancement of machine learning techniques such as extraction of important features, feature selection to curb the "curse of dimensionality", optimization of algorithms, ensembles, and several other advancements will contribute to the ease of tumor heterogeneity characterization. Furthermore, different degrees of accuracy were obtained with different algorithms. This was partly predicated on the inclusion of variables such as tumor texture and perfusion as key features. The studies examined show that they improve the efficacy of metastatic cancer diagnostics using ML techniques. At the same time, a statistical treatment of tumor heterogeneity features was also found to be a useful technique for making the predictive capabilities of models rather robust. Ultimately, a combination of radiomics, clinical and pathological variables was seen to provide superior predictive performance.

6. Minorities in metastatic cancer data

There has been several discussions about the importance of racial consciousness in medical health. In fact, it is not out of place to expect that body composition will be different across racial groups. As a case in point, Asians have been observed to have higher visceral body fat vis-a-vis a lower body-mass index which makes them to be at risk of diabetes [130], while people of sub-Saharan African descent on the other hand generally have a higher muscle mass which can influence renal function [131]. However, it is also widely understood that, in terms of genetics, biologically distinct categories do not exist amongst humans [132,133]. Moreover, beyond genetics and fundamental biological construct, several socio-economic factors and bias can contribute to racial disparities in healthcare and cancer research [134].

African Americans and people of sub-Saharan African descent reportedly have the lowest survival rate of most cancers amongst other races. Although the women sub-group within this population have a lower incidence rate (8%) compared to Caucasian women, they still have a higher rate of mortality (12%) [135]. Besides, there is also a possibility of having less representation of minority cancer patients' documentation in electronic health record. This seriously restricts the identification and treatment of cancer-associated diseases which they might experience [136].

A study by Duma et al., [137] revealed that over the past 14 years, only 31% of 1012 clinical trials report ethnicity. Moreover, there is a decline in the recruitment of minorities in clinical trials. A report on the 2015–2016 global participation of races in clinical trials revealed that only 2.7% of African-American patients were involved in oncology clinical trials, while 4% were enrolled in oncology drug trials [138]. This is quite low compared to the statistic that places them to be about 30% of the total population. In a study on the postoperative outcome in metastatic brain tumor patients, the authors eventually converted the race attribute to a binary attribute (Caucasian and Non-Caucasian) due to the high ratio of Caucasian patients (76.6%) to every other races in the data [139]. The ethnic/racial imbalance in clinical data prompts the need to have a careful design in precision medicine that caters to all [140]. As a matter of urgency, an inclusive approach to developing clinical trials in cancer research needs to be adopted [141]. The importance of this is clearly underscored by the fact that under-represented races are vulnerable to the aggressive effects of cancer due to unfavorable environmental and socio-economic factors. This should begin with the crucial task of identifying the reasons for disparities amongst races and regions in order to facilitate a better understanding of the drivers/mechanism of each metastatic cancer type [142]. However, the variation in geographic locations has been indentified as a factor that influences the risk of cancer in patients. Cancers that are potentially preventable such as cervical cancer, lung cancer and melanoma of the skin, are influenced based on factors such as obesity, smoking, and other factors that are closely related to healthy living. Population behavior in different environments tend to be localized, hence a strong indicator of cancer risk [143]. An example is a community with little or no smoking restriction or disincentive. One concrete study that provided an example of these population behavior-based tendencies was carried out by Tseng et al., [69]. They conducted a risk stratification analysis of patients with oral cavity squamous cell carcinoma on an East Asian population.The objective was to fill the gap of the lack of prognosis for the East Asian population, where the disease is prevalent due to behaviour, culture, and socioeconomic status [144]. Similarly, Jiang et al., [102], in an analysis of gastric cancer patients with occult peritoneal metastasis, observed clear difference in stages at which cancer is detected across different races. These, among others show that disease presentation and prevalence in patients may vary due to their race and place of habitation.

Dong et al., [87] developed a deep learning-based radiomic nomogram for gastric cancer patients with lymph node metastasis. The authors pointed out that their data, which was obtained mostly from Chinese patients and some of Italian descent, is deficient. They attributed the deficiency of the data to the different biology and aetiology associated with people from different races and countries. The issues raised by these studies show that it is vital to have a balanced composition of patient data from different races and countries to improve model performance. Nevertheless, some studies recorded a minimal influence of racial disparity on model performance. One of these was Halabi et al. [145] who reported that patients from different racial backgrounds show similar reactions to the same treatment in their clinical trial. Moreover, ML has been used to highlight the impact of socio-economic factors in cancer care. In the study carried out by Qiao et al., [146], while their model ranked patient demography low in the determination of survival outcome for lung cancer patients, access to care was ranked high. It is worth mentioning that access to care is closely related to racial segregation in some communities. The study further revealed that older unemployed female minorities had less medical care access compared to Caucasian patients.

Similarly, in a study involving the most extensive report of metastatic breast cancer patients with bone-only metastasis, Parkes et al., [147] reported that age and race/ethnicity affects the overall survival of bone-only metastatic breast cancer patients. Furthermore, in a study carried out by Deeb et al., [148] on 21335 metastatic cancer patients with terminal hospitalization between 2010 and 2017, it was revealed that ethnic and racial minorities were more likely to receive high-cost but low-value medical intervention towards the end of their lives. The findings from these studies inform the need to understand the disparities and consider them when building machine learning models that will be robust for deployment across racial and ethnic boundaries. It also necessitates considering other variables affecting patients' health status from minority and underrepresented races. The exorbitant cost of

randomized clinical trails may prove to be prohibitive to researchers' ability to provide an all-inclusive framework for cancer diagnostics and treatment [149]. The data obtained from an otherwise skewed process, being inadequate in capturing ethnic/racial differences in patients, will unfailingly affect the performance of trained models when applied to real world data. Regardless, the critical nature of cancer as a disease compels researchers and clinicians in this field to adhere to best practices in ensuring the maintenance of an ethnically and demographically balanced database for research and model training.

7. Repositories for metastatic cancer data

The availability of public datasets is a key enabler of research activities and impactful data analysis within the cancer research community. However, locating metastatic cancer data may not be a straightforward task in some cases. A key objective of this review is to publicize some of these major data repositories in order to advance their visibility and accessibility in the community. The location and brief description of selected repositories are provided.

The Cancer Genome Atlas (TCGA) is a repository that contains over 2.5 petabytes of genomic, transcriptomic, epigenomic, and proteomic data. It has publicly available metastatic cancer data for access by those in the research community. The purpose of the repository is to improve cancer diagnosis, prevention, and treatment-focused research. Researchers can access the public data via the genomics Data Commons Data Portal using web-based analysis and visualisation tools.

The Human Cancer Metastasis Database (HCMBD) is a user-friendly database for metastatic cancer. It contains metastasis-related transcriptome data and metastatic genomic and genetic data, including copy number alteration and somatic mutation data. It also has pharmacological drug data. The purpose of the database is to enable better diagnosis and treatment of metastatic cancer based on a good understanding of the transcriptomic regulation mechanisms [150]. The database contains 29 cancer types and 45 cancer subtypes, with 38 metastasis sites obtained from over 455 experiments. Based on 7081 published literature, 2183 genes consisting of 1901 protein-coding genes, 203 miRNAs and 24 long non-coding RNAs were curated to annotate the potentially metastasis-related genes.

The National Cancer Database is a nationally recognized database sponsored by the American Cancer Society and the American College of Surgeons. Hospital registry data are sourced for and collected from over 1500 facilities. The database contains over 34 million historical records and over 70% of newly diagnosed cancer cases in America. Based on the data from this database, the quality of care provided for cancer (metastatic) patients can be improved effectively upon further research. A study by Yang et al. [151] reports the prevalence of gaps in data capture and documentation in the National Cancer Database, leading to missing data. However, this can be mitigated by adequately capturing and documenting patients' medical records.

Gene Expression Omnibus (GEO) database is a database of the National Center for Biotechnology Information (NCBI) [152]. It is a public functional repository for genomics data. Researchers can submit array and sequence-based data for accessibility to other researchers. Gene expression profiles can easily be downloaded using the available query tools. The repository contains about 4350 datasets and about 4766100 samples.

The Cancer Imaging Archive (TCIA), funded by the Cancer Imaging Program, is a publicly available database of medical images of cancer. The image data are classified according to the disease, and image types, such as CT, MRI, digital histopathology, and image modality. Researchers can submit data, and published analysis results for accessibility to other researchers [153].

The Hartwig Medical Foundation Database is a database that consists of the clinical and genetic data of metastatic cancer patients in the Netherlands. Whole Genome Sequencing is used for the generation of genetic data. According to Priestley et al. [154], it is the largest metastatic whole-genome cancer resource. The purpose of the database is to enable faster discovery of biomarkers and improve existing biomarkers for the effectiveness of treatments. It also aims to allow researchers to understand metastasis development and encourage tumor DNA-based personalized treatment for each patient.

Project Data Sphere (PDS) is a repository that contains different cancer datasets, including metastatic cancer data of various body parts. The aim is to break down the barriers in sharing cancer clinical trial data. It is believed that sharing data and making them easily accessible to other researchers will ultimately benefit the patients who participate in these clinical trials. Data is generated from academic medical centres, biopharmaceutical companies, and government organizations. The free and open-access platform aims to improve the speed of cancer trials, reduce cost, and improve the effectiveness of cancer treatments [63,155,156].

The UNC Breast Cancer Metastatic Database is the database established by the UNC Lineberger Comprehensive Cancer Center to monitor and track the evolution of metastatic cancer, pathology, treatments, and clinical trials. It consists of the record of metastatic breast cancer patients at the UNC Breast Center. Its use of data for research is permissible based on approvals.

The Metastatic Breast Cancer Project Data shares genomic, clinical, molecular and patient-record data of metastatic breast cancer patients via the cBioPortal web-based platform. The goal is to increase the speed of discovering and developing new treatment strategies. The generated data are already cleaned, and the database is updated as new patients are enrolled. The cBioPortal repository contains other metastatic cancer data for research purposes.

Side-Out Foundation Metastatic Breast Cancer Database captures data from studies sponsored by the foundation. The database with clinical trial numbers (NCT01074814, NCT01919749, NCT03195192) contains more than 700 data fields. It consists of NGS-based whole/targeted exome sequencing generated genomic data, RNA microarray or RNA Seq generated transcript analysis data, Reverse Phase Protein Microarray (RPPA) generated phosphoproteomic data. Patients are de-identified, and information such as treatment history, demographics, pathological and clinical information, information about metastatic lesions, and outcome data are collected during the trials.

Other available metastatic cancer data include BIOGPS, a gene annotation web portal which is a repository containing data for lung, breast, brain and bone metastasis; The Southeast Netherlands Advanced Breast Cancer (SONABRE) Registry, a registry with clinical trial government registration number NCT03577197, based on the multi-centre study of advanced and metastatic breast cancer patients in the Netherlands; the Prostate Cancer Registry, containing a multi-centre collection of data of over 3000 metastatic castration-resistant prostate cancer patients from 16 countries. With a clinical government registration number NCT02236637, the data collected includes the treatment, baseline characteristics, survival outcome, and other necessary information [157]; Metastatic Colorectal Cancer Database contains about 1000 patient's data in a structured and centralized way; and Colorectal Liver Metastasis Database (CLIMB), a clinical trial study data collected based on colorectal carcinoma patients with liver metastasis.

Generally, since most metastatic cancer data contain patients' personal data, privacy policies, ethical data collection, and license agreements are usually involved. Usually, requests are subject to scrutiny, review, and assessment to ensure that the data is used according to the respective privacy policies.

Table 1
Summarized analysis of some works of literature utilizing machine learning for metastasis analysis.

REF.	INPUT DATA	MODEL	CANCER TYPE	AIM	PERFORMANCE	LIMITATION
[83]	RNA-Seq, microRNA-Seq, and DNA methylation data	convolutional variational autoencoder	11 cancer types	To distinguish pan-cancer metastasis status	Accuracy = 88.85, Precision = 91.65, Recall = 87.69, F1-score = 90.44, Specificity = 89.61	1. Lack of model interpretation. 2. Limited sample size
[106]	MRI images	Deep Neural Network	Glioblastoma, single brain metastasis	To distinguish single brain metastasis from glioblastoma using radiomic features	AUROC = 0.956 (95% CI, 0.918–0.990), Sensitivity = 90.6% (95% CI, 80.5–100), Specificity = 88.0% (95% CI, 79.0–97.0), Accuracy = 89.0% (95% CI, 82.3–95.8)	1. Variance in acquisition parameters of the MR images used. 2. There is a need for further multiparametric analysis.
[103]	structured	Logistic regression, kernel support vector machine, decision forest, and gradient boosting machine	Oral squamous cell carcinoma with occult lymph node metastasis	To detect occult nodal metastasis	AUC ranged from 0.657 to 0.840	1. Single imputation technique for missing data. 2. Single center data.
[85]	MRI images	Faster Region-Řbased Convolutional Neural Network	Rectal cancer with metastatic lymph node	Preoperative diagnosis of metastatic lymph nodes.	Metastatic Status: AUC = 0.920 Metastasis degree: AUC = 0.886	1. The model is based on MRI images, not pathological images.
[87]	CT images	Deep Convolutional Neural Network, DenseNet-201	Advanced Gastric cancer with metastatic lymph node	Preoperative evaluation of the number of lymph nodes.	C-indexes (Confidence Interval) = 0.822 (0.756–0.887)	1. Region-concentrated data. 2. 2D features were used instead of 3D features.
[94]	MRI, PET/CT images	Neural Network	Breast cancer with bone metastasis	Early detection of micrometastasis	60%− 80%	1. The model was built on animal data.
[102]	CT images	Deep Convolutional Neural Network	Gastric Cancer with Occult peritoneal metastasis	Preoperative noninvasive assessment of occult peritoneal metastasis	Cohort 1: AUC = 0.946 (95% CI, 0.927–0.965), sensitivity = 75.4%, specificity = 92.9%. Cohort2 AUC = 0.920 (95% CI, 0.848–0.992), sensitivity = 87.5%. specificity = 98.2%.	1. Non-diverse population-based dataset. 2. Variance in acquisition parameters of the CT images used. 3. Low accuracy for segmentation.
[91]	CT images	DenseNet, ResNet, ResNeXt, Xception, NASNet, VGG16, VGG19, DA-VGG19, Random Forest, Support Vector Machine	Breast Cancer with Axillary Lymph Node Metastasis	To predict axillary lymph node metastasis in breast cancer patients	DA-VGG19: Accuracy = 0.9088, Sensitivity = 0.9500, Specificity = 0.8675	1. Single center data. 2. 2D analysis instead of 3D analysis.
[65]	Differentially expressed gene	K-nearest neighbour, Support vector machine, Gradient boosting machine, decision tree, random forest, neural network	Recurrent Colorectal Cancer	To predict response to treatment	Sensitivity ranging from 0.800 to 0.900, Specificity ranging from 0.538 to 0.692	1. Lack of subgroup analysis due to small sample size.
[95]	structured	Logistic regression (L2 regularization), support vector machine (radial basis function kernel), decision tree, Gaussian naive Bayes classifier, K-nearest neighbours vote algorithm.	Melanomas	Prediction of metastatic events from prognostic serological biomarkers	Accuracy = 81.85%, Precision = 60.55%, Recall = 78.60%, ROC area = 88.80%	1. A larger population is needed to avoid overfitting.
[107]	RNA-sequencing data	Ensemble of non-linear support vector machine, random forest, and artificial neural network	Metastatic breast cancer	To predict the driver genes in metastatic breast cancer	AUROC = 99.24%	1. The model needs to be validated on clinical trial data.
[93]	whole-body scan images	Convolutional Neural Network	Breast cancer with bone metastasis	To classify breast cancer patients based on the occurrence of metastasis.	Accuracy = 92.50%.	1. Insufficient data. 2. Non-interpretability of the black-box model.
[63]	structured	Particle Swarm Optimized Gaussian Process Classifier	Metastatic castration-resistant prostate cancer	To determine the discontinuation of docetaxel treatment	AUC ranging between 0.6717 and 0.8499, and AUPRCs ranging between 0.1392 and 0.5423.	1. Single imputation technique for missing data.
[82]	structured	Deep Feed-Forward Neural Network	Non-Small-Cell Lung	Survival prediction and treatment recommendation	C-statistic = 0.742; 95% CI, 0.709–0.775	1. Lack of external validation. 2. The model is computationally expensive and difficult to explain.
[69]	structured	Cox proportional hazard regression model (elastic net penalized)	Advanced Oral Cancer	Postoperative Risk stratification for survival	C-index ranging from 0.689 to 0.702	1. Lack of model generalizability due to region-centric data.
[74]	CT images	ResNet10-based deep learning	Colorectal cancer with liver metastasis	To predict response to chemotherapy treatment	AUC: 0.830	1. Single location data. 2. Limited sample size.
[84]	MRI images	Convolutional Neural Network, ResNet18, ImageNet.	Cervical cancer with Lymph Node metastasis	Preoperative non-invasive diagnosis of lymph node metastasis.	AUROC = 0.933	1. Non-generalizability of performance due to limited dataset.
[71]	structured	Logistic regression, boosted logistic regression, random forest, and feedforward neural network.	Metastatic breast cancer	To predict adherence to treatment	AUROC ranging from 0.61 to 0.64	1. Low performance of models.

(*continued on next page*)

Table 1 (*continued*)

REF.	INPUT DATA	MODEL	CANCER TYPE	AIM	PERFORMANCE	LIMITATION
[75]	MRI images	Densely Connected Center-Cropping Convolutional Neural Network	Colorectal cancer with liver metastasis	Evaluation of tumor response to preoperative chemotherapy	AUC ranging from 0.833 to 0.875	1. Limited population

8. Summary and outlook

Cancer is a leading cause of death globally, and there has been a tremendous increase in new cancer cases over the years. More specifically, the primary cause of cancer-induced death is metastasis, which is responsible for over 90% of cancer mortality. Recently, several techniques have been developed to combat the menace of cancer. WIth a significant level of success, a compelling motivation for the application of ML stems from the need to develop non-invasive approaches for preoperative cancer diagnosis. Such an approach reduces treatment cost and time, improving patient quality of life. This study reviews classical machine learning and deep learning adoption in metastatic cancer-related studies. It examines the potential of these computational methods in the early diagnosis of metastasis in cancer patients and the prediction of the survival outcome. It further x-rays their use in the evaluation of patients' response to treatment, the need for discontinuation in case of adverse effects, and the impact of models built on racial/ethnic diverse populations.

In principle, in comparison to classical machine learning, deep learning is mainly used for image data. The feature extraction capability of deep learning aids the development of better and more accurate predictive models. Although there were cases where classical machine learning had better predictive accuracy than deep learning, deep learning remained the overall better technique, especially when the input data were associated with images. However, deep learning's expensive computational power and complexity are important factors to be considered when using the approach. It is also a black-box model, which has limited interpretability. Thus, clinicians tend to be wary of it. Regardless, studies show that these computation tools are useful for the study of metastasis . In addition to the foregoing, another issue with existing data repositories is the lack of ethnic diversity. Along this line of enquiry, we saw that the outcomes of clinical trials had a certain skewness that does not give a complete picture of the state of cancer metastasis across ethnic populations. Consequently, we can foresee that the models built on these biased repositories may not be robust enough for worldwide implementations.

Moreover, explainable AI (XAI) is a budding concept towards navigating the hidden/blurry part of black-box models. Over time, the need to understand, interpret and explain the result of ML models has been recognized. Several approaches exist, and they are recommended for the implementation in metastasis research. One of such approaches used for image data is the Gradient-weighted Class Activation Mapping (grad-CAM) technique [158]. This technique is an improved version of CAM which uses the gradient of the target flowing to the final convolution layer, to produce coarse localization and highlight important regions. As a means of feature selection, [159] used grad-CAM with multilayer perceptron for the extraction of crucial variables . In a related manner, it can visualize specific ECG waves responsible for myocardial infarction [160]. It has also been shown to be better at localizing heatmap patterns compared to CAM and grad-CAM++ in the classification of sclerosis in brain MRI [161]. Due to the prominent use of image data in metastasis research, such tool for DL black-box is encouraged as it has been utilized in similar clinical researches [162–164]. Other tools that can be utilized for image-based model interpretability include D-RISE [165], DeepLIFT [166], and integrated gradients [167]. Two other methods include Local Interpretable Model-Agnostic Explanations (LIME) [168], and Shapley Additive Explanations (SHAP) [169], which is built on the concept of game theory. These methods can either provide global or local explanations. However, to unravel the DL black-box, there is usually a trade-off between performance and the level of explanability/interpretability [170].

As previously mentioned, the application of machine learning in metastasis is an emerging research area. It can be utilized in the

metastatic stage of cancer for various analysis including detection, distinguishing tumor types, treatment recommendation, and other prognosis aimed at longevity and improving life quality. In Table 1, we provide a detailed summary of published research works where ML were used to explore metastatic cancer data, including the resulting model performance and their perceived limitations. Our overall deduction is that many studies tend toward preoperative non-invasive procedures for the diagnosis and treatment of metastatic cancer. From our examination of these works, we can identify the areas of future research. Since most curated data are image-based, we opine that focus should be on developing resource-efficient algorithms. This could be through architectures that require less information extracted from tumor image for prediction. Another approach is to develop hybrid schemes that combine the low-cost throughput of ML with the detail-oriented, more accurate feature extraction feature of DL. Furthermore, an alternative technique that could be adopted is to find a suitable (light weight) replacement for the convolution operation in DL architecture. This is because the convolution kernel requires high computation power for its calculations. Moreover, another persisting issue in the use of DL is the interpretability of its results. Unraveling its black box status will go a long way in providing insight into practical ways of tuning its parameters for better performance.

Consequently, we summarily show in Fig. 4 the current state of machine learning in metastatic cancer research, and areas that could be considered in future research. The future directions include explainable AI (XAI), federated learning, quantum computing, transfer learning, and the adoption of a diverse population in the data used for training. Federated learning involves the training of algorithms across multiple decentralized servers and devices which hold local data samples without necessarily sharing them. Few papers [171–173] have implemented the use of federated learning for cancer research, and these papers are produced recently, which shows it is a budding area of research. Federated learning applied to real world dataset across multiple centers helps to train models on relatively large datasets, and a diverse population, while protecting patients' privacy [171]. It has also shown superior performance compared to data from single institution [172]. However, according to [173], there is a risk of data leakage during training, which is undesirable. Another important technique is transfer learning. It involves utilizing previously trained models as a starting point for a new model to perform a new task. This method has shown a possibility of good performance with medical image analysis [174], especially with small samples [175] which is common in medical image data. We also see that it has been used in some lymph node metastasis detection studies [174,176,177,175]. Considering the prevalence of image data in metastatic cancer research, a gradual shift towards transfer learning for faster and optimized model training is envisaged. In addition to the aforementioned, quantum machine learning involving the application of quantum computing to machine learning is projected to improve generalization even with few data [178]. This technique can be utilized to speed up computation and improve performance of analysis compared to some classical ML methods [179]. However, standardized quantum datasets are essential for analysis [180].

In summary, the information written above could help medical practitioners make more informed diagnostic decisions, while also helping patients choose suitable treatment options. With the need to address the increase in cancer-induced deaths with less invasive treatment options, the possibilities and prospects we have shown should convince research granting organizations and funding agencies of the continued viability of ML in cancer diagnosis and treatment. A cumulative effort in this direction would go a long way in stemming the continuous waves of cancer mortality.

Funding statement

This research is funded by Data Analytics and Artificial Intelligence (DAAI) research group, School of Computing and Digital Technology, Birmingham City University, UK. This research was substantially sponsored by the research projects (Grant No. 32170654 and Grant No. 32000464) supported by the National Natural Science Foundation of China and was substantially supported by the Shenzhen Research Institute, City University of Hong Kong. This project is substantially funded by the Strategic Interdisciplinary Research Grant of City University of Hong Kong (Project No. 2021SIRG036). The work described in this paper was substantially supported by the grant from the Health and Medical Research Fund, the Food and Health Bureau, The Government of the Hong Kong Special Administrative Region [07181426]. The work described in this paper was partially supported by the grants from City University of Hong Kong (CityU 11203520, CityU 11203221).

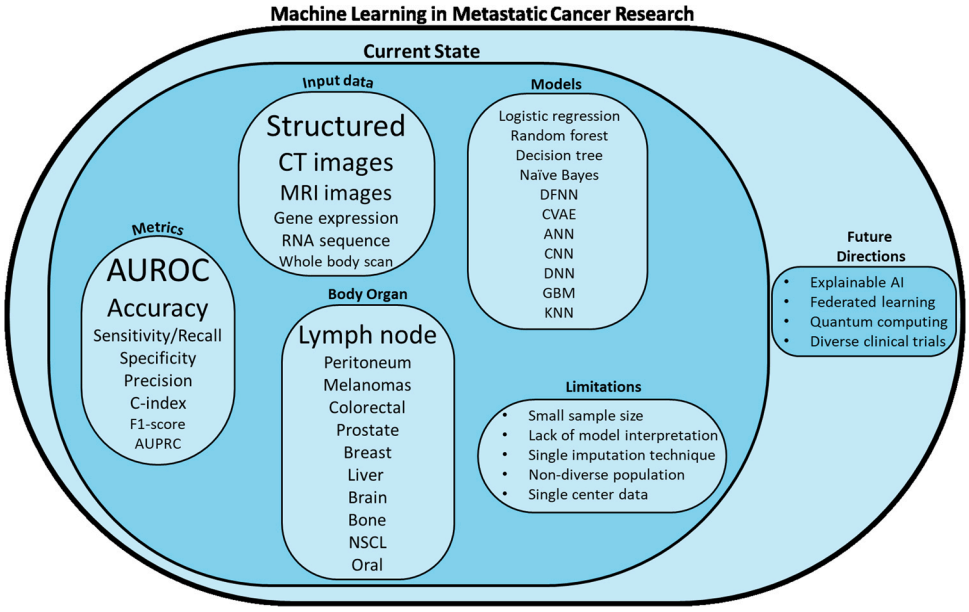

Fig. 4. The current state and future direction of machine learning in metastatic cancer research.

CRediT authorship contribution statement

Olutomilayo Olayemi Petinrin: Conception, Resources, Writing – original draft, Writing – review & editing. **Faisal Saeed**: Writing – review & editing, Fund Acquisition. **Muhammad Toseef**: Writing – original draft. **Zhe Liu**: Writing – original draft. **Shadi Basurra**: Fund acquisition. **Ibukun Omotayo Muyide**: Writing – original draft. **Xiangtao Li**: Visualization. **Qiuzhen Lin**: Visualization. **Ka-Chun Wong**: Supervision, Fund acquisition.

Declaration of interest

None.

Acknowledgement

The authors would like to appreciate Ikeoluwa Ogedengbe, and Emmanuel for their effort towards the revision and proofreading of the manuscript.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.03.046.

References

[1] Rock CL, Thomson C, Gansler T, Gapstur SM, McCullough ML, Patel AV, et al. American cancer society guideline for diet and physical activity for cancer prevention. CA Cancer J Clin 2020;70(4):245–71.
[2] SEER Training Modules, What is Cancer?, u. s. national institutes of health, national cancer institute, [Accessed: 2022–11-28] (2022). ⟨https://training.seer.cancer.gov/disease/cancer/⟩.
[3] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71(3):209–49.
[4] Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin 2023;73(1):17–48.
[5] N. Howlader, A. Noone, M. Krapcho, J. Garshell, D. Miller, S. Altekruse, et al., Seer cancer statistics review, National Cancer Institute 2008 (1975).
[6] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin 2020;70(1):7–30.
[7] Ganesh K, Massagué J. Targeting metastatic cancer. Nat Med 2021;27(1):34–44.
[8] Robinson DR, Wu Y-M, Lonigro RJ, Vats P, Cobain E, Everett J, et al. Integrative clinical genomics of metastatic cancer. Nature 2017;548(7667):297–303.
[9] Wang C, Sandhu J, Ouyang C, Ye J, Lee PP, Fakih M. Clinical response to immunotherapy targeting programmed cell death receptor 1/programmed cell death ligand 1 in patients with treatment-resistant microsatellite stable colorectal cancer with and without liver metastases. JAMA Netw Open 2021;4(8):e2118416.
[10] Place AE, JinHuh S, Polyak K. The microenvironment in breast cancer progression: biology and implications for treatment. Breast Cancer Res 2011;13(6):1–11.
[11] Schneider C, Pozzi A. Cyclooxygenases and lipoxygenases in cancer. Cancer Metastas Rev 2011;30(3):277–94.
[12] Cammarota G, Ianiro G, Ahern A, Carbone C, Temko A, Claesson MJ, et al. Gut microbiome, big data and machine learning to promote precision medicine for cancer. Nat Rev Gastroenterol Hepatol 2020;17(10):635–48.
[13] Riihimäki M, Thomsen H, Sundquist K, Sundquist J, Hemminki K. Clinical landscape of cancer metastases. Cancer Med 2018;7(11):5534–42.
[14] Zhou Y, Yu Q-F, Peng A-F, Tong W-L, Liu J-M, Liu Z-L. The risk factors of bone metastases in patients with lung cancer. Sci Rep 2017;7(1):1–6.
[15] Zhang L, Gong Z. Clinical characteristics and prognostic factors in bone metastases from lung cancer. Med Sci Monit: Int Med J Exp Clin Res 2017;23:4087.
[16] Brodowicz T, Hadji P, Niepel D, Diel I. Early identification and intervention matters: a comprehensive review of current evidence and recommendations for the monitoring of bone health in patients with cancer. Cancer Treat Rev 2017;61:23–34.
[17] Chen S, Zhou W, Tu J, Li J, Wang B, Mo X, et al. A novel xgboost method to infer the primary lesion of 20 solid tumor types from gene expression data. Front Genet 2021;12:632761.
[18] Samani ZR, Parker D, Wolf R, Hodges W, Brem S, Verma R. Distinct tumor signatures using deep learning-based characterization of the peritumoral microenvironment in glioblastomas and brain metastases. Sci Rep 2021;11(1):1–9.
[19] Poturnajova M, Furielova T, Balintova S, Schmidtova S, Kucerova L, Matuskova M. Molecular features and gene expression signature of metastatic colorectal cancer. Oncol Rep 2021;45(4):1.
[20] Cuocolo R, Caruso M, Perillo T, Ugga L, Petretta M. Machine learning in oncology: a clinical appraisal. Cancer Lett 2020;481:55–62.
[21] Albaradei S, Thafar M, Alsaedi A, Van Neste C, Gojobori T, Essack M, et al. Machine learning and deep learning methods that use omics data for metastasis prediction. Comput Struct Biotechnol J 2021;19:5008–18.
[22] Cho SJ, Sunwoo L, Baik SH, Bae YJ, Choi BS, Kim JH. Brain metastasis detection using machine learning: a systematic review and meta-analysis. Neuro-Oncol 2021;23(2):214–25.
[23] Shivakumar N, Chandrashekar A, Handa AI, Lee R. Use of deep learning for detection, characterisation and prediction of metastatic disease from computerised tomography: a systematic review. Postgrad Med J 2021.
[24] Zheng Q, Yang L, Zeng B, Li J, Guo K, Liang Y, et al. Artificial intelligence performance in detecting tumor metastasis from medical radiology imaging: A systematic review and meta-analysis. EClinicalMedicine 2021;31:100669.
[25] Adeoye J, Tan JY, Choi S-W, Thomson P. Prediction models applying machine learning to oral cavity cancer outcomes: a systematic review. Int J Med Inform 2021;154:104557.
[26] Kocher M, Ruge MI, Galldiks N, Lohmann P. Applications of radiomics and machine learning for radiotherapy of malignant brain tumors. Strahlenther und Onkol 2020;196(10):856–67.
[27] Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. Br J Pharmacol 2011;162(6):1239–49.
[28] Hussain W, Rasool N, Khan YD. Insights into machine learning-based approaches for virtual screening in drug discovery: Existing strategies and streamlining through fp-cadd. Curr Drug Discov Technol 2021;18(4):463–72.
[29] Chen B, Garmire L, Calvisi DF, Chua M-S, Kelley RK, Chen X. Harnessing big 'omics' data and ai for drug discovery in hepatocellular carcinoma. Nat Rev Gastroenterol Hepatol 2020;17(4):238–51.
[30] Li H, Sze K-H, Lu G, Ballester PJ. Machine-learning scoring functions for structure-based virtual screening. Wiley Interdiscip Rev: Comput Mol Sci 2021;11(1):e1478.
[31] Sabe VT, Ntombela T, Jhamba LA, Maguire GE, Govender T, Naicker T, et al. Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. Eur J Med Chem 2021;224:113705.
[32] Honarparvar B, Govender T, Maguire GE, Soliman ME, Kruger HG. Integrated approach to structure-based enzymatic drug design: molecular modeling, spectroscopy, and experimental bioactivity. Chem Rev 2014;114(1):493–537.
[33] Shen C, Ding J, Wang Z, Cao D, Ding X, Hou T. From machine learning to deep learning: Advances in scoring functions for protein–ligand docking. Wiley Interdiscip Rev: Comput Mol Sci 2020;10(1):e1429.
[34] Nogueira MS, Koch O. The development of target-specific machine learning models as scoring functions for docking-based target prediction. J Chem Inf Model 2019;59(3):1238–52.
[35] Petinrin OO, Saeed F. Stacked ensemble for bioactive molecule prediction. IEEE Access 2019;7:153952–7.
[36] Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers 2021;25(3):1315–60.
[37] Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today 2018;23(6):1241–50.
[38] Lima AN, Philot EA, Trossini GHG, Scott LPB, Maltarollo VG, Honorio KM. Use of machine learning approaches for novel drug discovery. Expert Opin Drug Discov 2016;11(3):225–39.
[39] Warr WA, Nicklaus MC, Nicolaou CA, Rarey M. Exploration of ultralarge compound collections for drug discovery. J Chem Inf Model 2022;62(9):2021–34.
[40] Glaser J, Vermaas JV, Rogers DM, Larkin J, LeGrand S, Boehm S, et al. High-throughput virtual laboratory for drug discovery using massive datasets. Int J High Perform Comput Appl 2021;35(5):452–68.
[41] Meng C, Hu Y, Zhang Y, Guo F. Psbp-svm: a machine learning-based computational identifier for predicting polystyrene binding peptides. Front Bioeng Biotechnol 2020;8:245.
[42] Ricci-Lopez J, Aguila SA, Gilson MK, Brizuela CA. Improving structure-based virtual screening with ensemble docking and machine learning. J Chem Inf Model 2021;61(11):5362–76.
[43] KumarShukla P, KumarShukla P, Sharma P, Rawat P, Samar J, Moriwal R, et al. Efficient prediction of drug–drug interaction using deep learning models. IET Syst Biol 2020;14(4):211–6.
[44] Sharma A, Rani R. Ensembled machine learning framework for drug sensitivity prediction. IET Syst Biol 2020;14(1):39–46.
[45] Nascimento AC, Prudêncio RB, Costa IG. A drug-target network-based supervised machine learning repurposing method allowing the use of multiple heterogeneous information sources. Computational Methods for Drug Repurposing. Springer; 2019. p. 281–9.
[46] Kaufman B, Shapira-Frommer R, Schmutzler RK, Audeh MW, Friedlander M, Balmaña J, et al. Olaparib monotherapy in patients with advanced cancer and a germline brca1/2 mutation. J Clin Oncol: J Am Soc Clin Oncol 2015;33(3):244–50.
[47] Karwasra R, Khanna K, Singh S, Ahmad S, Verma S. The incipient role of computational intelligence in oncology: Drug designing, discovery, and development. Computational Intelligence in Oncology: Applications in Diagnosis, Prognosis and Therapeutics of Cancers. Springer; 2022. p. 369–84.

[48] Beacher FD, Mujica-Parodi LR, Gupta S, Ancora LA. Machine learning predicts outcomes of phase iii clinical trials for prostate cancer. Algorithms 2021;14(5):147.
[49] Sherman J, Verstandig G, Rowe JW, Brumer Y. Application of machine learning to large in vitro databases to identify drug–cancer cell interactions: azithromycin and klk6 mutation status. Oncogene 2021;40(21):3766–70.
[50] Lee H, Jeong AJ, Ye S-K. Highlighted stat3 as a potential drug target for cancer therapy. BMB Rep 2019;52(7):415.
[51] Zou S, Tong Q, Liu B, Huang W, Tian Y, Fu X. Targeting stat3 in cancer immunotherapy. Mol Cancer 2020;19(1):1–19.
[52] Wadood A, Ajmal A, Junaid M, Rehman AU, Uddin R, Azam SS, et al. Machine learning-based virtual screening for stat3 anticancer drug target. Curr Pharm Des 2022;28(36):3023–32.
[53] Nguyen TK, LeNguyen TN, Nguyen K, Nguyen HVT, Tran LTT, Ngo TXT, et al. Machine learning-based screening of mcf-7 human breast cancer cells and molecular docking analysis of essential oils from Ocimum basilicum against breast cancer. J Mol Struct 2022;1268:133627.
[54] Issa NT, Stathias V, Schürer S, Dakshanamurthy S. Machine and deep learning approaches for cancer drug repurposing Vol. 68 Seminars in cancer biologyElsevier; 2021. p. 132–42. Vol. 68.
[55] Cao L, Guler M, Tagirdzhanov A, Lee Y-Y, Gurevich A, Mohimani H. Moldiscovery: learning mass spectrometry fragmentation of small molecules. Nat Commun 2021;12(1):1–13.
[56] Watson OP, Cortes-Ciriano I, Taylor AR, Watson JA. A decision-theoretic approach to the evaluation of machine learning algorithms in computational drug discovery. Bioinformatics 2019;35(22):4656–63.
[57] Wu Z, Lawrence PJ, Ma A, Zhu J, Xu D, Ma Q. Single-cell techniques and deep learning in predicting drug response. Trends Pharmacol Sci 2020.
[58] Amer MH, Al-Sarraf M, Vaitkevicius VK. Factors that affect response to chemotherapy and survival of patients with advanced head and neck cancer. Cancer 1979;43(6):2202–6.
[59] Mansoori B, Mohammadi A, Davudian S, Shirjang S, Baradaran B. The different mechanisms of cancer drug resistance: a brief review. Adv Pharm Bull 2017;7(3):339.
[60] Zahreddine H, Borden KL. Mechanisms and insights into drug resistance in cancer. Front Pharmacol 2013;4:28.
[61] Alfarouk KO, Stock C-M, Taylor S, Walsh M, Muddathir AK, Verduzco D, et al. Resistance to cancer chemotherapy: failure in drug response from adme to p-gp. Cancer Cell Int 2015;15(1):1–13.
[62] Zhang C, Liu X, Jin S, Chen Y, Guo R. Ferroptosis in cancer therapy: a novel approach to reversing drug resistance. Mol Cancer 2022;21(1):1–12.
[63] Petinrin OO, Li X, Wong K-C. Particle swarm optimized gaussian process classifier for treatment discontinuation prediction in multicohort metastatic castration-resistant prostate cancer patients. IEEE J Biomed Health Inform 2021.
[64] Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol 2017;9:157.
[65] Lu W, Fu D, Kong X, Huang Z, Hwang M, Zhu Y, et al. Folfox treatment response prediction in metastatic or recurrent colorectal cancer patients via machine learning algorithms. Cancer Med 2020;9(4):1419–29.
[66] Statnikov A, Wang L, Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinforma 2008;9(1):1–10.
[67] Van der Heijden M, Essers PB, De Jong MC, De Roest RH, Sanduleanu S, Verhagen CV, et al. Biological determinants of chemo-radiotherapy response in hpv-negative head and neck cancer: a multicentric external validation. Front Oncol 2020;9:1470.
[68] Ferrer G, Álvarez-Errico D, Esteller M. Biological and molecular factors predicting response to adoptive cell therapies in cancer. JNCI: J Natl Cancer Inst 2022.
[69] Tseng Y-J, Wang H-Y, Lin T-W, Lu J-J, Hsieh C-H, Liao C-T. Development of a machine learning model for survival risk stratification of patients with advanced oral cancer. JAMA Netw Open 2020;3(8):e2011768.
[70] Showalter SL, Meneveau MO, Keim-Malpass J, Camacho TF, Squeo G, Anderson RT. Effects of adjuvant endocrine therapy adherence and radiation on recurrence and survival among older women with early-stage breast cancer. Ann Surg Oncol 2021;28(12):7395–403.
[71] Yerrapragada G, Siadimas A, Babaeian A, Sharma V, O'Neill TJ. Machine learning to predict tamoxifen nonadherence among us commercially insured patients with metastatic breast cancer. JCO Clin Cancer Inform 2021;5:814–25.
[72] Xu Y, Hosny A, Zeleznik R, Parmar C, Coroller T, Franco I, et al. Deep learning predicts lung cancer treatment response from serial medical imaging. Clin Cancer Res 2019;25(11):3266–75.
[73] Imai K, Allard M-A, Benitez CC, Vibert E, Cunha AS, Cherqui D, et al. Early recurrence after hepatectomy for colorectal liver metastases: what optimal definition and what predictive factors? oncologist 2016;21(7):887.
[74] Wei J, Cheng J, Gu D, Chai F, Hong N, Wang Y, et al. Deep learning-based radiomics predicts response to chemotherapy in colorectal liver metastases. Med Phys 2021;48(1):513–22.
[75] Zhu H, Xu D, Ye M, Sun L, Zhang X, Li X, et al. Deep learning-assisted magnetic resonance imaging prediction of tumor response to chemotherapy in patients with colorectal liver metastases. Int J Cancer 2021;148(7):1717–30.
[76] Klein CA. Cancer progression and the invisible phase of metastatic colonization. Nat Rev Cancer 2020;20(11):681–94.
[77] Madhavan D, Peng C, Wallwiener M, Zucknick M, Nees J, Schott S, et al. Circulating mirnas with prognostic value in metastatic breast cancer and for early detection of metastasis. Carcinogenesis 2016;37(5):461–70.
[78] Xiao W, Zheng S, Yang A, Zhang X, Zou Y, Tang H, et al. Breast cancer subtypes and the risk of distant metastasis at initial diagnosis: a population-based study. Cancer Manag Res 2018;10:5329.
[79] Scimeca M, Antonacci C, Toschi N, Giannini E, Bonfiglio R, Buonomo CO, et al. Breast osteoblast-like cells: a reliable early marker for bone metastases from breast cancer. Clin Breast Cancer 2018;18(4):e659–69.
[80] Cai W-L, Huang W-D, Li B, Chen T-R, Li Z-X, Zhao C-L, et al. microrna-124 inhibits bone metastasis of breast cancer by repressing interleukin-11. Mol Cancer 2018;17(1):1–14.
[81] Eloranta S, Smedby K, Dickman P, Andersson T. Cancer survival statistics for patients and healthcare professionals–a tutorial of real-world data analysis. J Intern Med 2021;289(1):12–28.
[82] She Y, Jin Z, Wu J, Deng J, Zhang L, Su H, et al. Development and validation of a deep learning model for non-small cell lung cancer survival. JAMA Netw Open 2020;3(6):e205842.
[83] Albaradei S, Napolitano F, Thafar MA, Gojobori T, Essack M, Gao X. Metacancer: a deep learning-based pan-cancer metastasis prediction model developed using multi-omics data. Comput Struct Biotechnol J 2021;19:4404–11.
[84] Wu Q, Wang S, Zhang S, Wang M, Ding Y, Fang J, et al. Development of a deep learning model to identify lymph node metastasis on magnetic resonance imaging in patients with cervical cancer. JAMA Netw Open 2020;3(7):e2011625.
[85] Ding L, Liu G, Zhang X, Liu S, Li S, Zhang Z, et al. A deep learning nomogram kit for predicting metastatic lymph nodes in rectal cancer. Cancer Med 2020;9(23):8809–20.
[86] Peak TC, Russell GB, Dutta R, Rothberg MB, Chapple AG, Hemal AK. A national cancer database-based nomogram to predict lymph node metastasis in penile cancer. BJU Int 2019;123(6):1005–10.
[87] Dong D, Fang M-J, Tang L, Shan X-H, Gao J-B, Giganti F, et al. Deep learning radiomic nomogram can predict the number of lymph node metastasis in locally advanced gastric cancer: an international multicenter study. Ann Oncol 2020;31(7):912–20.
[88] Riihimäki M, Hemminki A, Sundquist J, Hemminki K. Patterns of metastasis in colon and rectal cancer. Sci Rep 2016;6(1):1–9.
[89] Zhao B, Gabriel RA, Vaida F, Lopez NE, Eisenstein S, Clary BM. Predicting overall survival in patients with metastatic rectal cancer: a machine learning approach. J Gastrointest Surg 2020;24(5):1165–72.
[90] Nicolò C, Périer C, Prague M, Bellera C, MacGrogan G, Saut O, et al. Machine learning and mechanistic modeling for prediction of metastatic relapse in early-stage breast cancer. JCO Clin Cancer Inform 2020;4:259–74.
[91] Liu Z, Ni S, Yang C, Sun W, Huang D, Su H, et al. Axillary lymph node metastasis prediction by contrast-enhanced computed tomography images for breast cancer patients based on deep learning. Comput Biol Med 2021;136:104715.
[92] Chu CS, Lee NP, Adeoye J, Thomson P, Choi S. Machine learning and treatment outcome prediction for oral cancer. J Oral Pathol Med 2020;49(10):977–85.
[93] Papandrianos N, Papageorgiou E, Anagnostis A, Feleki A. A deep-learning approach for diagnosis of metastatic breast cancer in bones from whole-body scans. Appl Sci 2020;10(3):997.
[94] Ellmann S, Seyler L, Gillmann C, Popp V, Treutlein C, Bozec A, et al. Machine learning algorithms for early detection of bone metastases in an experimental rat model. JoVE (J Vis Exp) 2020;1(162):e61235.
[95] Mancuso F, Lage S, Rasero J, Díaz-Ramón JL, Apraiz A, Pérez-Yarza G, et al. Serum markers improve current prediction of metastasis development in early-stage melanoma patients: a machine learning-based study. Mol Oncol 2020;14(8):1705–18.
[96] Hawes D, Neville AM, Cote R. Occult metastasis. Biomed Pharmacother 2001;55(4):229–42.
[97] Arain AA, Rajput MSA, Ansari SA, Mahmood Z, Ahmad AN, Dogar MR, et al. Occult nodal metastasis in oral cavity cancers. Cureus 2020;12(11).
[98] Chen S-b, Liu D-t, Huang S-j, Weng H-r, Wang G, Li H, et al. Prognostic value of occult lymph node metastases in patients with completely resected esophageal squamous cell carcinoma. Sci Rep 2020;10(1):1–10.
[99] Fang Q, Li P, Qi J, Luo R, Chen D, Zhang X. Value of lingual lymph node metastasis in patients with squamous cell carcinoma of the tongue. Laryngoscope 2019;129(11):2527–30.
[100] Lop J, Rigó A, Codina A, de Juan J, Quer M, León X. Prognostic significance of extranodal extension in head and neck squamous cell carcinoma cn0 patients with occult metastatic neck nodes. Acta Otorrinolaringol (Engl Ed) 2018;69(3):156–64.
[101] Yang L, Liu F, Wu Y, Fang Q, Zhang X, Du W, et al. Predictive value of occult metastasis and survival significance of metabolic tumor volume determined by pet-ct in ct1-2n0 squamous cell carcinoma of the tongue. Front Oncol 2020;10:2583.
[102] Jiang Y, Liang X, Wang W, Chen C, Yuan Q, Zhang X, et al. Noninvasive prediction of occult peritoneal metastasis in gastric cancer using deep learning. JAMA Netw Open 2021;4(1):e2032269.
[103] Bur AM, Holcomb A, Goodwin S, Woodroof J, Karadaghy O, Shnayder Y, et al. Machine learning to predict occult nodal metastasis in early oral squamous cell carcinoma. Oral Oncol 2019;92:20–5.
[104] Jovčevska I. Genetic secrets of long-term glioblastoma survivors. Bosn J Basic Med Sci 2019;19(2):116.

[105] Chaurasia A, Park S-H, Seo J-W, Park C-K. Immunohistochemical analysis of atrx, idh1 and p53 in glioblastoma and their correlations with patient survival. J Korean Med Sci 2016;31(8):1208–14.
[106] Bae S, An C, Ahn SS, Kim H, Han K, Kim SW, et al. Robust performance of deep learning for distinguishing glioblastoma from single brain metastasis using radiomic features: model development and validation. Sci Rep 2020;10(1):1–10.
[107] Mirsadeghi L, Hosseini RH, Banaei-Moghaddam AM, Kavousi K. Earn: an ensemble machine learning algorithm to predict driver genes in metastatic breast cancer. BMC Med Genom 2021;14(1):1–19.
[108] Xu Y, Cui X, Wang Y. Pan-cancer metastasis prediction based on graph deep learning method. Front Cell Dev Biol 2021;9:1133.
[109] H. Chereda, A. Bleckmann, F. Kramer, A. Leha, T. Beissbarth, Utilizing molecular network information via graph convolutional neural networks to predict metastatic event in breast cancer, in: GMDS 2019, 181–186.
[110] Wang Y, Zheng Y, Jia Q, Wang Y. Prediction algorithm of regional lymph node metastasis of rectal cancer based on improved deep neural network. J Med Imaging Health Inform 2021;11(2):370–7.
[111] Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 2018;15(2):81–94.
[112] Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. Br J Cancer 2013;108(3):479–85.
[113] Lee D, Park Y, Kim S. Towards multi-omics characterization of tumor heterogeneity: a comprehensive review of statistical and machine learning approaches. Brief Bioinforma 2021;22(3):bbaa188.
[114] Chao J, Bedell V, Lee J, Li MS, Chu P, Yuan Y-C, et al. Association between spatial heterogeneity within nonmetastatic gastroesophageal adenocarcinomas and survival. JAMA Netw Open 2020;3(4):e203652.
[115] Vera-Yunca D, Girard P, Parra-Guillen ZP, Munafo A, Trocóniz IF, Terranova N. Machine learning analysis of individual tumor lesions in four metastatic colorectal cancer clinical studies: linking tumor heterogeneity to overall survival. AAPS J 2020;22(3):1–12.
[116] Levy-Jurgenson A, Tekpli X, Kristensen VN, Yakhini Z. Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. Sci Rep 2020;10(1):1–11.
[117] Gillies RJ. Cancer heterogeneity and metastasis: Life at the edge. Clin Exp Metastas– 2021:1–5.
[118] Keller L, Pantel K. Unravelling tumour heterogeneity by single-cell profiling of circulating tumour cells. Nat Rev Cancer 2019;19(10):553–67.
[119] Lang N, Su M-Y, Hon JY, Lin M, Hamamura MJ, Yuan H. Differentiation of myeloma and metastatic cancer in the spine using dynamic contrast-enhanced mri. Magn Reson Imaging 2013;31(8):1285–91.
[120] Lang N, Zhang Y, Zhang E, Zhang J, Chow D, Chang P, et al. Differentiation of spinal metastases originated from lung and other cancers using radiomics and deep learning based on dce-mri. Magn Reson Imaging 2019;64:4–12.
[121] Vera-Yunca D, Parra-Guillen ZP, Girard P, Trocóniz IF, Terranova N. Relevance of primary lesion location, tumour heterogeneity and genetic mutation demonstrated through tumour growth inhibition and overall survival modelling in metastatic colorectal cancer. Br J Clin Pharmacol 2022;88(1):166–77.
[122] Wang Y-W, Chen C-J, Huang H-C, Wang T-C, Chen H-M, Shih J-Y, et al. Dual energy ct image prediction on primary tumor of lung cancer for nodal metastasis using deep learning. Comput Med Imaging Graph 2021;91:101935.
[123] Lee JY, Lee K-s, Seo BK, Cho KR, Woo OH, Song SE, et al. Radiomic machine learning for predicting prognostic biomarkers and molecular subtypes of breast cancer using tumor heterogeneity and angiogenesis properties on mri. Eur Radiol 2022;32(1):650–60.
[124] Daye D, Tabari A, Kim H, Chang K, Kamran SC, Hong TS, et al. Quantitative tumor heterogeneity mri profiling improves machine learning–based prognostication in patients with metastatic colon cancer. Eur Radiol 2021;31(8):5759–67.
[125] Jaberipour M, Soliman H, Sahgal A, Sadeghi-Naini A. A priori prediction of local failure in brain metastasis after hypo-fractionated stereotactic radiotherapy using quantitative mri and machine learning. Sci Rep 2021;11(1):1–10.
[126] Hagiwara A, Tatekawa H, Yao J, Raymond C, Everson R, Patel K, et al. Visualization of tumor heterogeneity and prediction of isocitrate dehydrogenase mutation status for human gliomas using multiparametric physiologic and metabolic mri. Sci Rep 2022;12(1):1078.
[127] Truong AH, Sharmanska V, Limback-Stanic C, Grech-Sollars M. Optimization of deep learning methods for visualization of tumor heterogeneity and brain tumor grading through digital pathology. Neuro-Oncol Adv 2020;2(1):vdaa110.
[128] Rubinstein JC, ForoughiPour A, Zhou J, Sheridan TB, White BS, Chuang JH. Deep learning image analysis quantifies tumor heterogeneity and identifies microsatellite instability in colon cancer. J Surg Oncol 2023;127(3):426–33.
[129] He J, Wang Q, Zhang Y, Wu H, Zhou Y, Zhao S. Preoperative prediction of regional lymph node metastasis of colorectal cancer based on 18f-fdg pet/ct and machine learning. Ann Nucl Med 2021;35(5):617–27.
[130] Hsu WC, Araneta MRG, Kanaya AM, Chiang JL, Fujimoto W. Bmi cut points to identify at-risk asian americans for type 2 diabetes screening. Diabetes care 2015;38(1):150–8.
[131] Stevens, Lesley A. and Levey, Andrew S., Frequently asked questions about GFR estimates, [Accessed: 2022–11-28] (2007). ⟨https://www.niddk.nih.gov/health-information/professionals/clinical-tools-patient-management/kidney-disease/laboratory-evaluation/frequently-asked-questions⟩.
[132] Mersha TB, Abebe T. Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. Hum Genom 2015;9(1):1–15.
[133] A. S. of Human Genetics. Ashg denounces attempts to link genetics and racial supremacy. Am J Hum Genet 2018;103:636.
[134] Cerdeña JP, Plaisime MV, Tsai J. From race-based to race-conscious medicine: how anti-racist uprisings call us to act. Lancet 2020;396(10257):1125–8.
[135] Giaquinto AN, Miller KD, Tossas KY, Winn RA, Jemal A, Siegel RL. Cancer statistics for african american/black people 2022. CA: A Cancer J Clin 2022;72(3):202–29.
[136] Chen WC, Boreta L, Braunstein SE, Rabow MW, Kaplan LE, Tenenbaum JD, et al. Association of mental health diagnosis with race and all-cause mortality after a cancer diagnosis: Large-scale analysis of electronic health record data. Cancer 2021.
[137] Duma N, VeraAguilera J, Paludo J, Haddox CL, GonzalezVelez M, Wang Y, et al. Representation of minorities and women in oncology clinical trials: review of the past 14 years. J Oncol Pract 2018;14(1):e1–10.
[138] U. Food, D. Administration, 2015–2016 global participation in clinical trials report, US Food and Drug Administration (2020).
[139] Khalafallah AM, Jimenez AE, Patel P, Huq S, Azmeh O, Mukherjee D. A novel online calculator predicting short-term postoperative outcomes in patients with metastatic brain tumors. J neuro-Oncol 2020;149(3):429–36.
[140] Prasad P, Branch M, Asemota D, Elsayed R, Addison D, Brown S-A. Cardio-oncology preventive care: racial and ethnic disparities. Curr Cardiovasc Risk Rep 2020;14(10):1–14.
[141] Takvorian SU, Haas NB. Use of bone resorption inhibitors in metastatic castration-resistant prostate cancer—20 years later, and the answer is still yes. JAMA Netw Open 2021;4(7):e2117159.
[142] Mahal B. Re: Clinical outcomes in men of diverse ethnic backgrounds with metastatic castration-resistant prostate cancer. Ann Oncol 2020;31(7):829.
[143] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. CA: Cancer J Clin 2021;71(1):7–33.
[144] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a Cancer J Clin 2018;68(6):394–424.
[145] Halabi S, Dutta S, Tangen C, Rosenthal M, Petrylak D, Thompson Jr I, et al. Clinical outcomes in men of diverse ethnic backgrounds with metastatic castration-resistant prostate cancer. Ann Oncol 2020;31(7):930–41.
[146] Qiao EM, Voora RS, Nalawade V, Kotha NV, Qian AS, Nelson TJ, et al. Evaluating the clinical trends and benefits of low-dose computed tomography in lung cancer patients. Cancer Med 2021;10(20):7289–97.
[147] Parkes A, Warneke CL, Clifton K, Al-Awadhi A, Oke O, Pestana RC, et al. Prognostic factors in patients with metastatic breast cancer with bone-only metastases. oncologist 2018;23(11):1282.
[148] Deeb S, Chino FL, Diamond LC, Tao A, Aragones A, Shahrokni A, et al. Disparities in care management during terminal hospitalization among adults with metastatic cancer from 2010 to 2017. JAMA Netw Open 2021;4(9):e2125328.
[149] Coquet J, Bievre N, Billaut V, Seneviratne M, Magnani CJ, Bozkurt S, et al. Assessment of a clinical trial-derived survival model in patients with metastatic castration-resistant prostate cancer. JAMA Netw Open 2021;4(1):e2031730.
[150] Zheng G, Ma Y, Zou Y, Yin A, Li W, Dong D. Hcmdb: the human cancer metastasis database. Nucleic Acids Res 2018;46(D1):D950–5.
[151] Yang DX, Khera R, Miccio JA, Jairam V, Chang E, James BY, et al. Prevalence of missing data in the national cancer database and association with overall survival. JAMA Netw Open 2021;4(3):e211793.
[152] Edgar R, Domrachev M, Lash AE. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic Acids Res 2002;30(1):207–10.
[153] Freymann JB, Kirby JS, Perry JH, Clunie DA, Jaffe CC. Image data sharing for biomedical research—meeting hipaa requirements for de-identification. J Digit Imaging 2012;25(1):14–24.
[154] Priestley P, Baber J, Lolkema MP, Steeghs N, de Bruijn E, Shale C, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature 2019;575(7781):210–6.
[155] Miyashita H, Cruz C, Patel V. Risk factors of skeletal-related events in patients with bone metastatic castration-resistant prostate cancer undergoing treatment with zoledronate. Support Care Cancer 2021:1–4.
[156] Vazquez E, Gouraud H, Naudet F, Gross CP, Krumholz HM, Ross JS, et al. Characteristics of available studies and dissemination of research using major clinical data sharing platforms. Clin Trials 2021;18(6):657–66.
[157] Chowdhury S, Bjartell A, Lumen N, Maroto P, Paiss T, Gomez-Veiga F, et al. Real-world outcomes in first-line treatment of metastatic castration-resistant prostate cancer: the prostate cancer registry. Target Oncol 2020;15:301–15.
[158] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, 618–626.
[159] He T, Guo J, Chen N, Xu X, Wang Z, Fu K, et al. Medimlp: using grad-cam to extract crucial variables for lung cancer postoperative complication prediction. IEEE J Biomed Health Inform 2019;24(6):1762–71.
[160] Jahmunah V, Ng EYK, Tan R-S, Oh SL, Acharya UR. Explainable detection of myocardial infarction using deep learning models with grad-cam technique on ecg signals. Comput Biol Med 2022;146:105550.
[161] Zhang Y, Hong D, McClement D, Oladosu O, Pridham G, Slaney G. Grad-cam helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. J Neurosci Methods 2021;353:109098.

[162] Panwar H, Gupta P, Siddiqui MK, Morales-Menendez R, Bhardwaj P, Singh V. A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images. Chaos, Solitons Fractals 2020;140:110190.
[163] Marmolejo-Saucedo JA, Kose U. Numerical grad-cam based explainable convolutional neural network for brain tumor diagnosis. Mob Netw Appl 2022:1–10.
[164] Kim J-K, Jung S, Park J, Han SW. Arrhythmia detection model using modified densenet for comprehensible grad-cam visualization. Biomed Signal Process Control 2022;73:103408.
[165] V. Petsiuk, R. Jain, V. Manjunatha, V.I. Morariu, A. Mehra, V. Ordonez, et al., Black-box explanation of object detectors via saliency maps, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 11443–11452.
[166] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: International conference on machine learning, PMLR, 2017, 3145–3153.
[167] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, 3319–3328.
[168] M.T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, 1135–1144.
[169] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30.
[170] Petch J, Di S, Nelson W. Opening the black box: the promise and limitations of explainable machine learning in cardiology. Can J Cardiol 2022;38(2):204–13.
[171] duTerrail JO, Leopold A, Joly C, Béguier C, Andreux M, Maussion C, et al. Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. Nat Med 2023;29(1):135–46.
[172] Sarma KV, Harmon S, Sanford T, Roth HR, Xu Z, Tetreault J, et al. Federated learning improves site performance in multicenter deep learning without data sharing. J Am Med Inform Assoc 2021;28(6):1259–64.
[173] Hansen CR, Price G, Field M, Sarup N, Zukauskaite R, Johansen J, et al. Larynx cancer survival model developed through open-source federated learning. Radiother Oncol 2022;176:179–86.
[174] Yu J, Deng Y, Liu T, Zhou J, Jia X, Xiao T, et al. Lymph node metastasis prediction of papillary thyroid carcinoma based on transfer learning radiomics. Nat Commun 2020;11(1):4807.
[175] Li J, Zhou Y, Wang P, Zhao H, Wang X, Tang N, et al. Deep transfer learning based on magnetic resonance imaging can improve the diagnosis of lymph node metastasis in patients with rectal cancer. Quant Imaging Med Surg 2021;11(6):2477.
[176] Kang H, Yang M, Zhang F, Xu H, Ren S, Li J, et al. Identification lymph node metastasis in esophageal squamous cell carcinoma using whole slide images and a hybrid network of multiple instance and transfer learning, Biomedical. Signal Process Control 2023;82:104577.
[177] Khan A, Brouwer N, Blank A, Müller F, Soldini D, Noske A, et al. Computer-assisted diagnosis of lymph node metastases in colorectal cancers using transfer learning with an ensemble model. Mod Pathol 2023;36(5):100118.
[178] Caro MC, Huang H-Y, Cerezo M, Sharma K, Sornborger A, Cincio L, et al. Generalization in quantum machine learning from few training data. Nat Commun 2022;13(1):4919.
[179] Huang H-Y, Broughton M, Mohseni M, Babbush R, Boixo S, Neven H, et al. Power of data in quantum machine learning. Nat Commun 2021;12(1):2631.
[180] Cerezo M, Verdon G, Huang H-Y, Cincio L, Coles PJ. Challenges and opportunities in quantum machine learning. Nat Comput Sci 2022;2(9):567–76.